MICROPHONE ARRAY PROCESSING FOR AUGMENTED LISTENING

BY

RYAN MICHAEL COREY

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Andrew C. Singer, Chair
Professor Jont B. Allen
Assistant Professor Ivan Dokmanić
Associate Professor Paris Smaragdis

Abstract

Modern augmented listening technologies, such as hearing aids, smart headphones, and audio augmented reality platforms, perform poorly in noisy environments with many competing sound sources. This work explores the benefits of large microphone arrays, including novel wearable devices and distributed sensor networks, for augmented listening systems. Perceptually transparent space-time remixing filters can apply separate processing to each sound source to modify the auditory scene perceived by a listener. The design parameters and performance tradeoffs of such filters are described, with particular emphasis on the ways in which augmented listening applications differ from machine listening and telecommunication applications. Theoretical tools are developed for interaural cue preservation, delay-constrained array processing, and dynamic range compression of multiple sources. Several implementation issues are considered, including acoustic channel estimation, the design of wearable microphone arrays, the acoustic effects of the body, and models and algorithms for deformable microphone arrays. Finally, the performance of the listening system is improved by cooperative processing among many distributed devices. The proposed system would dramatically improve the performance of listening devices in noisy environments and enable new listening applications that are impossible with current technology.

To Uncle Bill.

Acknowledgments

To paraphrase legendary drag queen and Champaign-Urbana native Sasha Velour, there is no greater joy in life than to pursue a crazy, over-the-top idea and have a team of brilliant people to help you do it. This project would not have been possible without the support of some incredible people.

I will be forever grateful to my advisor, Andy Singer, for letting me chase the dream of superhuman hearing. One day in early 2015, I mentioned in passing that my hearing aids didn't work very well and I thought I could make them better. He could have shrugged and told me to get back to work on my project, but instead he helped me dive into a field that was new to both of us. The next thing I knew, I was dressing up mannequins in giant hats and doing the macarena to recordings of British people reading the newspaper. No matter how farfetched my ideas, Andy has never said "no, that won't work," only "yes, and what if we did this?"

To make those crazy ideas happen, we needed a team. Over the past two years, I've had the privilege to work with more than twenty talented students (and counting!) in the Illinois Augmented Listening Laboratory. They've built prototypes, designed research tools, collected data sets, run demos, and helped to promote the project to the community. I would especially like to thank Haige Chen, Uriah Jones, Sooraj Kumar, JJ Martinez-Villalba, Manan Mittal, Matt Skarha, Bryce Tharp, and Naoki Tsuda, who worked with the team for multiple terms.

Several of these students contributed directly to the experiments and results presented in this dissertation:

- The artificial ears that were affixed to the mannequins during binaural recordings were made by Benjamin Khachaturian.

- The behind-the-ear earpieces used in nearly every experiment in this work were made by Uriah Jones.
- The smart speaker enclosures used in the distributed array data set were designed by Uriah Jones and Ben Stoehr and assembled by Matt Skarha.
- The distributed array data set of Section 2.4 was collected with significant help from Matt Skarha.
- The wearable microphone data set of Section 2.3 was collected with significant help from Naoki Tsuda.
- The vowel-like test signals used to characterize delay-performance tradeoffs in Chapter 5 were generated by Naoki Tsuda. He also measured the impulse responses used in that experiment.
- The embedded system architecture described in Section 9.1 was developed by JJ Martinez-Villalba, with assistance from many other students.

I am grateful to Professors Jont Allen and Paris Smaragdis, whose courses helped me to get started in the field of audio signal processing. I am also thankful for the team at Arm Research, especially Jesse Beu, Ganesh Dasika, and David Palframan, for inviting me to Austin to explore embedded audio processing. I must especially acknowledge office manager Peggy Wells, who helped to arrange travel, order equipment, and generally make my life easier.

Giant microphone arrays are not free. This project was supported in part by the National Science Foundation Graduate Research Fellowship Program under grant number DGE-1144245; by Systems on Nanoscale Information fabriCs (SONIC), a STARnet Center sponsored by SRC and DARPA; by the Microsoft Research Dissertation Grant; and by the National Science Foundation Partnerships for Innovation program under grant number 1919257.

Finally, I could not have survived graduate school without the love and support of my family and friends. Thank you all so much!

Table of Contents

Chapter 1 Augmented Listening 1
1.1 What Is Augmented Listening? 2
1.2 Listening Technology Today 6
1.3 Microphone Arrays 13
1.4 Array Processing and Human Listeners 20
1.5 Microphone Array Processing for Augmented Listening 26

Chapter 2 Data and Methods 34
2.1 Equipment and Facilities 35
2.2 Experimental Methods 40
2.3 Wearable Microphone Data Set 43
2.4 Distributed Microphone Data Set 47
2.5 Experiments with Human Subjects 50

Chapter 3 Arrays and Spatial Filtering 52
3.1 Notation Used in the Dissertation 54
3.2 Signal Representations 55
3.3 Array Processing System 60
3.4 Models of Signal Propagation 68
3.5 Statistical Filter Design Criteria 74

Chapter 4 Binaural Audio Source Remixing 83
4.1 A Source-Remixing Filter 86
4.2 Performance and Desired Responses 90
4.3 Interaural Cue Preservation 98
4.4 Summary and Future Directions 111

Chapter 5 Delay-Constrained Array Processing 113
5.1 Delay in Listening Devices 113
5.2 Causal Space-Time Filtering 118
5.3 Exact Results for Special Cases 126
5.4 Experimental Results 131
5.5 Delay Constraints and Augmented Listening 138

Chapter 6 Dynamic Range Compression 140
6.1 Dynamic Range Compression of a Single Source 142
6.2 Dynamic Range Compression and Noise 148
6.3 Dynamic Range Compression of Multiple Sources 161

Chapter 7 Time-Varying Space-Time Filters 171
7.1 Time-Varying Methods 172
7.2 Source Activity Mask 179
7.3 High-Low Space-Time Filter 182
7.4 Time-Varying Filters for Augmented Listening 190

Chapter 8 Source-Informed Acoustic Channel Estimation 197
8.1 Acoustic Channel Estimation 198
8.2 In Situ Channel Measurement From Pilot Signals 202
8.3 Channel Estimation from Speech Keywords 207

Chapter 9 Wearable Microphone Arrays 215
9.1 Design and Construction 216
9.2 Acoustic Effects of the Body 224
9.3 Beamforming Performance of Wearable Arrays 227
9.4 Modeling Small Microphone Motion 230
9.5 Motion-Tolerant Processing for Deformable Arrays 238
9.6 Wearable Array Design 243

Chapter 10 Cooperative Processing with Multiple Devices 246
10.1 Room-Scale Microphone Arrays 247
10.2 Cooperative Listening Enhancement with Room-Scale Arrays 253
10.3 Deformable and Asynchronous Arrays 260
10.4 Cooperative Nonlinear Processing for Partially Asynchronous Arrays 266
10.5 Cooperative Processing for Augmented Listening Devices 273

Chapter 11 Developing Augmented Listening Systems 277
11.1 Performance Tradeoffs and Design Principles 278
11.2 The Future of Listening Technology 286
11.3 Broader Applications of This Work 289
11.4 A Telescope for the Ears 292

References 294

Chapter 1

Augmented Listening

Throughout human history, we have used technology to help us do things that our bodies cannot on their own. The wheel let our ancestors move things they could not carry in their hands. Today we use airplanes to go farther and faster than our legs can carry us. Some of mankind's most impactful technological advances are those that augment human senses. The microscope let scientists see the invisible world of microbes with their own eyes, ushering in germ theory and modern medicine, and we can use telescopes to gaze across the cosmos. At human scales, eyeglasses and contact lenses can restore normal vision by compensating for distortion in the eyes.

Humanity has not been nearly as successful at augmenting our sense of hearing. Our greatest engineering triumph in listening technology is arguably the stethoscope. Many modern technologies, like motorcycles and air conditioners, actually make hearing more difficult. Meanwhile, for hundreds of millions of people living with hearing loss, our most advanced hearing aid technology is embarrassingly primitive. I know because I am one of them. While hearing aids can help me to understand a conversation in a quiet room, they are no help at all in a noisy crowd where I need them most. Even people with normal hearing could use help in noisy listening environments like restaurants.

The hearing aid industry has been working for generations to build the hearing equivalent of contact lenses: small, comfortable, virtually invisible devices that can fully restore normal sensory function. Unfortunately, most hearing loss is far more complex than the linear optical distortion that causes nearsightedness. Sensorineural hearing loss is highly nonlinear, and it is likely impossible to make an impaired ear function like a healthy one. This dissertation proposes a more ambitious approach:

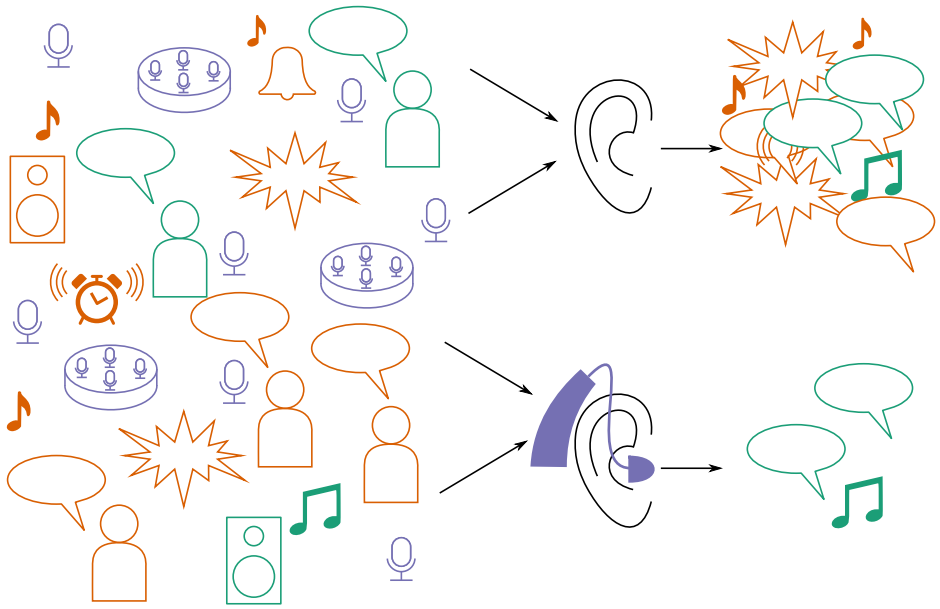

Figure 1.1: Augmented listening technology modifies the auditory scene perceived by the user, for example by suppressing unwanted sounds.

instead of trying in vain to restore normal function to an impaired ear, let us develop technology to give any listener superhuman hearing. That is, if we cannot build contact lenses, let us build a telescope instead.

1.1 What Is Augmented Listening?

Augmented listening refers to any technology that enhances human perception by altering the sounds that people hear, as shown in Figure 1.1. Assistive listening technologies such as hearing aids are an important subset of augmented listening. But the system proposed in this dissertation can do more than just help people with hearing loss to understand speech: augmented listening devices could help both normal-hearing and hearing-impaired listeners to hear better in noisy situations. This technology could also enable new augmented reality applications that deliberately alter the user's perceived auditory scene, for example by introducing artificial sound

sources or replacing one signal with another.

1.1.1 The cocktail party problem

Humans' ability to extract meaning from noisy sound mixtures is known as the *cocktail party phenomenon* [1, 2]. We use knowledge of speech patterns and rules of acoustics to infer which sound events belong with each other and assemble those pieces into actionable information [3]. Humans can also use spatial information to distinguish between sounds coming from different directions [4]; for example, sounds from the left arrive at the left ear before the right ear and they also have greater intensity in the left ear, especially at high frequencies. In addition to time and level differences that locate sounds on a left-right axis, our brains can use the subtle filtering effects of the head and pinna to distinguish between sounds coming from the front and back and from different elevations. We can do this even in reverberant environments where sounds arrive from multiple directions at once. But the cocktail party phenomenon has limits: even normal-hearing people struggle in noisy, reverberant environments with many competing sound sources. Could machines do better?

The challenge of automatically separating different speech signals from a mixture is called the *cocktail party problem*. Because humans can distinguish these signals so well, many researchers developing machine listening algorithms such as automatic speech recognition and audio event detection have tried to imitate the function of the auditory system. Historically, researchers used computational auditory scene analysis [5] to classify sounds based on spectral and temporal patterns, as the auditory system is believed to do. Many speech recognition systems incorporate multiple microphones and use spatial information to help separate sounds [6]. More recently, researchers are applying machine learning to imitate the pattern-matching abilities of the brain [7].

Mimicking the human auditory system is a good approach to designing machine listening systems. It is not enough, however, if we hope to build augmented listening systems that confer superhuman abilities. We cannot expect computer algorithms to extract more meaning than the human brain from the same set of signals. To have any chance of surpassing human abilities, machines must have access to information

that humans do not. It is possible in principle that machine learning algorithms with enormous data sets could build better models of speech and noise than humans could learn by experience. This approach could be useful for speech in a language that is unfamiliar to the user or for public transit systems that have distinctive noise profiles, for example.

However, there is a simpler way for augmented listening systems to have more information than the auditory system: humans have only two ears, but machines can have hundreds of microphones. While it would be challenging to develop algorithms with better-than-human signal models and pattern matching, we can already build machines with spatial resolution that far surpasses that of our ears.

1.1.2 Spatial sound processing

Humans have remarkable spatial hearing abilities despite having just two ears. Still, the ears cannot resolve multiple closely spaced sound sources that are far away. To do that, we would need much larger *sensor arrays*, like those that have long been used in radar, sonar, and telecommunication applications [8, 9]. Arrays can sample signals in space to localize, separate, and enhance sound waves arriving from different directions [10, 11]. Next-generation wireless communication technology uses massive arrays with hundreds of antennas to communicate with multiple users over the same frequency band at the same time [12]. Today, microphone arrays are widely used in speech recognition and teleconferencing applications, where they can isolate the speech signal of one talker from unwanted background noise [13, 14]. Systems with microphone arrays have been shown to perform better than single-microphone systems on speech recognition tasks [6, 15].

Microphone arrays are also used in listening devices [16,17]. Most high-end hearing aids include two microphones in each earpiece, and some can share data between ears for a total of four microphones. Engineers have spent more than 30 years trying to build listening devices with larger arrays [18–21], but none have been commercially successful. Until recently, there were no small, inexpensive microphones that could be built into comfortable wearable devices, nor did embedded processors have the power

to apply complex spatial processing algorithms. Furthermore, as this dissertation will make clear, using microphone arrays for human listening is often more complicated than using them for machine listening applications.

1.1.3 New possibilities for augmented listening

Despite generations of failure, the time has finally come when engineers can build superhuman listening devices. We have all the tools we need: digital microphones smaller than a pea that can be hidden in clothing, accessories, appliances, and furniture; embedded processors with advanced linear algebra accelerators; low-latency, high-throughput, high-concurrency wireless networks; versatile cloud and edge computing systems; and, crucially, a favorable economic and regulatory environment for innovative listening technologies. The challenge before us is to develop the theory, algorithms, and architectures to put these tools together.

This dissertation proposes a new approach to spatial signal processing for augmented listening. A listening device collects sound data from microphones that extend far beyond the ears. These arrays could be mounted on wearable accessories such as eyeglasses or hats, they could be spread across multiple accessories all over the body, or they could even be distributed around the room. Unlike a conventional hearing aid that processes sounds as a mixture or a conventional array device that tries to isolate a single source, the proposed augmented listening system attempts to apply independent processing to each source signal and then recombine them in a perceptually transparent way. This "remixing" process could be subtle—for example, reducing background noise in a restaurant just enough so that the user's conversation partner is intelligible—or profound, such as introducing a new virtual sound source with realistic acoustics or translating multiple speech signals into different languages.

Designing such a system is a daunting task. To enhance human hearing in the complex, noisy, reverberant environments where listeners most need help, we must apply the most advanced acoustic models and spatial processing methods available, but many state-of-the-art algorithms are not designed for human listeners. To use them in a listening system, we must reconcile the linear worlds of acoustics and array

processing with the highly nonlinear processes of human hearing and perception. To use microphones in wearable devices, we must account for the acoustics of the body and for complex motion patterns. To remain perceptually transparent, the output signals must have no more than a few milliseconds of delay and no distortion of interaural cues. Even if these engineering challenges were solved, we know relatively little about what sort of processing should be applied to each sound source to best enhance the listening experience for different users and in different situations.

Microphone-array listening devices are not a new idea; they have tantalized engineers for decades and inspired countless publications, including several other dissertations [18–20, 22–24]. Yet there has never been a commercially successful hearing aid or similar listening device that uses a large microphone array. The primary goal of this work is not to solve particular technical problems or develop new algorithms, although it does do that; instead, it is to understand why past efforts have failed and to explore the challenges that must be overcome to make ambitious augmented listening technology a reality. Some of these challenges, such as interaural cue preservation, are well-studied. Others, like delay, dynamic range compression, and body acoustics, are characterized for the first time in the context of microphone array processing. It is hoped that this dissertation will guide future research so that engineers can soon, at long last, dramatically augment human hearing.

1.2 Listening Technology Today

A few years ago, there were only two kinds of device that would play sound into the ears: headphones, which play back transmitted or recorded sound, and hearing aids, which amplify and enhance sound in the immediate environment. The landscape today is far more complex and changing rapidly. Figure 1.2 shows a few recent listening products.

Figure 1.2: There are many kinds of listening devices on the market, including hearing aids (left), "hearables" (center), and advanced headphones (right).

1.2.1 Listening systems

Let us begin with traditional hearing aids. These devices, produced by only a handful of large companies, cost thousands of dollars and are available only through medical professionals. Audiologists measure the patient's hearing profile using audiograms and other techniques, then customize a prescription hearing aid to fit their needs. While there are fitting guidelines, much of the process is based on experience or trial and error. In the United States, most insurance plans do not cover hearing aids and so there is a large market for lower-cost devices. Many companies sell hearing aids over the internet for a few hundred dollars; these are configured using automated software or an at-home listening test rather than an in-person exam.

Until recently, there was also a category of over-the-counter devices known as personal sound amplification products (PSAP) [25]. Although ostensibly intended for people without hearing loss, they were clearly targeted at people with hearing loss who could not afford prescription hearing aids. These inexpensive devices had few if any customization options. Following the Over-the-Counter Hearing Aid Act of 2017, certain hearing aids can now be sold over the counter; this change will presumably eliminate PSAPs as a distinct product category.

Even before this legal change, consumer electronics companies were developing listening devices with hearing-aid-like features. These smart headphones, which some in the media have dubbed "hearables," amplify and process environmental sounds in some of the same ways as hearing aids [26]. Several companies have also tried to

incorporate advanced augmented reality features such as intelligent noise reduction and language translation, but to date they have had limited success. At the same time, hearing aid companies have raced to incorporate consumer-audio features such as Bluetooth music and call streaming, voice assistants, motion sensors, and even fitness trackers. The previously distinct categories of consumer headphones and medical hearing aids are quickly converging into a broader class of augmented listening devices.

An important emerging product category is augmented reality headsets [27]. While most recent attention has been on visual augmented reality, such as video overlays, many augmented reality headsets also incorporate arrays of microphones that could be used to alter the user's auditory experience. They could impose new virtual sources into the environment for mixed-reality games and remote presence applications. They could also dynamically alter sound sources in real time; the possibilities range from the frivolous—make your friend sound like a chipmunk!—to the profound, such as real-time translation. These bulky headsets have ample computational capabilities and room for dozens of widely spaced microphones, making them attractive platforms for the spatial processing methods developed in this work.

There are relatively few technological differences between a listening device intended to correct hearing loss and one designed to enhance normal hearing. An assistive device would likely provide stronger and more-frequency-selective amplification and more-aggressive dynamic range compression. It might also choose remixing parameters to emphasize intelligibility over naturalness, for example by applying more noise reduction. However, these differences are largely matters of degree. Therefore, this work will rarely distinguish between hearing aids and more general augmented listening devices.

1.2.2 Listening devices

Most hearing aids and other listening devices share a similar set of basic components [28, 29]. One or two microphones capture incoming sound and digitize it. These sounds are manipulated by a digital processor to generate an output signal. This

digital output is converted back to an analog signal and presented to the listener by a transducer known as a *receiver*, which usually sits inside the ear canal. The device also includes a battery and hardware controls such as buttons and knobs. High-end devices usually have wireless capabilities, including low-latency near-field magnetic induction (NFMI) for exchanging data between earpieces and Bluetooth for communicating with mobile devices. Although it is possible in principle for binaural hearing aids to wirelessly share audio data and perform binaural beamforming, few commercial hearing aids do so due to power constraints. Instead, the wireless link is used to synchronize processing settings between the left and right hearing aids several times per second.

Hearing aids come in several form factors. The most popular is a behind-the-ear earpiece, which contains one or two microphones, a processor, and a battery, and connects via a thin tube to a receiver in the ear canal. These are shown on the left in Figure 1.2. Fully in-the-ear hearing aids can be used for some types of hearing loss, but they have stricter size and power constraints. Hearing aid companies have traditionally tried to create discreet products that users can comfortably wear all day and that other people will not easily notice. Meanwhile, today's high-end wireless earbuds, like those in the center in Figure 1.2, are bulky and conspicuous; some even consider them to be status symbols, like an expensive watch or designer eyeglasses. These changing consumer preferences might ease the size and power constraints on traditional hearing aids as well, allowing them to apply more ambitious processing.

1.2.3 Signal processing for listening enhancement

There are many ways that listening devices, especially hearing aids, enhance sound for human listeners [28, 30]. A typical hearing aid processing system is illustrated in Figure 1.3.

Amplification: Hearing aids provide gain to compensate for the reduced sensitivity of hearing-impaired listeners. The cheapest off-the-shelf hearing-aid-like devices, like the second device from the left in Figure 1.2, contain only a mi-

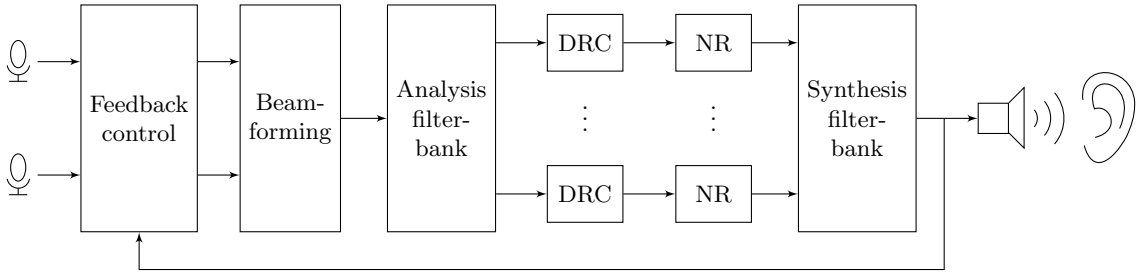

Figure 1.3: Architecture of a typical modern hearing aid.

crophone, a receiver, and an amplifier.

Spectral shaping: The majority of sensorineural hearing loss is more severe at high frequencies than at low frequencies. Hearing aids provide more gain at frequencies where users have more hearing loss.

Feedback control: Because the microphone and receiver are only a few centimeters apart, listening devices use echo cancellation techniques to prevent whistling due to feedback [31]. Although feedback has historically been an important limiting factor for hearing aids, modern feedback control systems perform well. Because the feedback problem is largely solved, it is not addressed in this work.

Dynamic range compression (DRC): Many hearing-impaired listeners have reduced dynamic range compared to normal-hearing listeners. To compensate, all advanced hearing aids perform dynamic range compression, which is a form of automatic gain control [32–34]. The device tracks sound level over time and increases gain when the level is low to improve audibility or decreases gain when the level is high to prevent discomfort. Expensive hearing aids apply compression independently across many frequency bands. Dynamic range compression is arguably the most important, most challenging, and least understood form of signal processing that listening devices apply; compression and its problems are the subject of Chapter 6.

Noise reduction (NR): All listeners, with or without hearing loss, have trouble hearing in loud background noise. Over the years, engineers have devised

countless nonlinear processing methods for reducing background noise while preserving a signal of interest, usually speech [35]. These range from time-varying spectral subtraction and Wiener filtering [36] to time-frequency masks designed by deep-learning classifiers [37]. However, while single-microphone noise reduction methods can improve comfort, they do not help listeners to understand speech better than they otherwise would [38, 39].

Beamforming: Unlike single-microphone noise reduction methods, spatial processing has been shown to improve intelligibility in noise [38]. Many hearing aids do perform simple spatial filtering, known as beamforming, to emphasize sounds from the front and reduce background noise [40]. Beamforming often features prominently in hearing aid marketing materials. However, because the microphones in each earpiece are just a few millimeters apart, and because of the disturbing distortion that aggressive beamforming can introduce (see Chapter 4), hearing aids use fairly conservative spatial processing.

Scene classification: Listening devices decide what kind of environment they are in—a home, a restaurant, a noisy train—and adjust their settings accordingly. They might also adapt to the types of sound sources present in a space. Although this work does not discuss scene classification methods, they would be an essential component of a complete augmented listening system.

Dereverberation: It can be more difficult to hear in strongly reverberant environments. Although the author is not aware of any hearing aids that perform dereverberation, it has been studied extensively by signal processing researchers [41] and it would be a useful feature of an augmented listening system. However, dereverberation is not addressed directly in this work.

Frequency lowering: Some hearing aids have features to perform frequency compression or frequency lowering, which maps higher frequencies onto lower frequencies for users with severe high-frequency hearing loss [42]. It will not be addressed in this work.

1.2.4 Shortcomings of listening technology

Modern listening devices fall far short of their potential. Despite widespread interest from both established electronics companies and new startups over the last few years, there have been no commercially successful "hearables" products with advanced augmented listening features. Furthermore, although hearing loss affects about one in seven adults in the United States, fewer than 30% of people who would benefit from hearing aids actually wear them [43]. There are two generally accepted causes of this low adoption rate: the first is the prohibitive cost of hearing aids, which are rarely covered by insurance in the United States, and the second is poor performance. Anecdotally, the author has never met a single hearing aid user—outside the hearing aid industry—who is satisfied with their hearing aids. A common complaint is about the poor performance of hearing aids in noisy environments. Indeed, a recent National Institute on Deafness and other Communication Disorders (NIDCD) Strategic Plan for hearing and balance research highlights the need to "[i]mprove the performance of traditional (external) hearing aids in background noise and other real-world settings." [43]

There are two reasons that listening devices perform so poorly in noisy environments. First, hearing aids perform nonlinear processing, including dynamic range compression, that does not obey superposition. When applied to mixtures of sound sources rather than single signals, these algorithms do not behave as intended. The distortion caused by compression in multisource environments is well documented in the hearing literature and has been observed in state-of-the-art commercial hearing aids [44–50]. Most hearing aids process signals separately in different frequency bands based on the assumption that signals of interest and unwanted noise have different spectra. However, this is not true in the most challenging listening environments where the dominant noise source is speech.

Second, most listening devices perform what is effectively single-microphone noise reduction. Although earpieces often include multiple microphones, they are too closely spaced to provide meaningful spatial noise reduction. It has been widely observed that single-microphone noise reduction algorithms do not significantly improve

Figure 1.4: Sensor arrays can distinguish between signals arriving from different directions based on time differences of arrival and other signal features.

intelligibility; at best, they can improve listening comfort in some situations [38, 39].

Over the last several years, the hearing aid industry has focused on quality-of-life improvements, such as rechargeable batteries and remote-control apps, and consumer-gadget features such as Bluetooth streaming and heart-rate tracking. Rather than building larger devices that can perform meaningful spatial processing, they have continued racing to build smaller and more discreet devices. Large hearing aid companies appear to be focusing their research efforts on machine learning techniques. While these methods could lead to incremental improvements in performance, they do not provide the device with any information that the ear does not have already; for that, we must use large microphone arrays.

1.3 Microphone Arrays

Microphone arrays, like sensor arrays more broadly, add another dimension to signal processing by sampling signals in space. Array processing systems can manipulate signals across space, time, and frequency to do things that would be impossible with a single sensor.

1.3.1 Spatial signal processing

While modern statistical array processing methods are quite complex [8–11], the basic intuition is simple: Sound waves from different directions will arrive at the different sensors of the array at different times, as shown in Figure 1.4. If the positions of the microphones are known, the system can use time differences of arrival between sensors to calculate the direction of a signal; this process is called *localization.*

A complementary problem, called *beamforming*, is to focus on signals arriving from a target direction and suppress all others [51]. Simple delay-and-sum beamformers, suitable for anechoic environments, apply different delays to the signals from each microphone so that sounds arriving from the target direction interfere constructively and are amplified. Signals from other directions may be amplified or attenuated depending on direction and frequency. More complex filter-and-sum beamformers can be used to control beam patterns and apply constraints to multiple directions of arrival.

Acoustic arrays are rarely used in anechoic environments. Indeed, many of the spaces in which augmented listening systems would be most helpful, such as restaurants, bars, and conference rooms, are strongly reverberant. Sound does not travel in a direct path from each source to each microphone: it bounces off of walls, furniture, and other surfaces. The microphones themselves may also be directional and have nonuniform frequency responses. In these situations, we can still enhance some sound sources and suppress others, but such a system cannot be said to form a "beam." In this work, such processing is referred to as *space-time filtering.* It will be reviewed in Chapter 3.

The related problem of *source separation* deals with extracting multiple signals from a mixture [10]. For example, a common source separation task is to recover the speech signals of individual talkers from a recording of several people talking at once. Space-time filters can be used to perform unmixing, but only if the spatial parameters—such as locations or transfer function vectors—of the sources are known. If they are not, they must be estimated from the mixture itself. This difficult problem, known as *blind source separation* [52, 53], has vexed audio signal process-

ing researchers for many years. Even today, there are no known methods that can blindly separate more than a few competing sound sources in challenging real-world environments.

1.3.2 Trends in microphone array technology

The past few years have been eventful for microphone array technology. Thanks to low-cost, high-performance microelectromechanical-systems (MEMS) microphones, it is now easy to embed multiple microphones into any electronic device [54]. Nearly every audio device larger than a watch now has at least two microphones. Meanwhile, array-equipped smart speakers that perform localization and beamforming have surged in popularity, bringing microphone arrays into millions of homes for the first time.

There has also been significant recent interest from researchers, enthusiasts, and industry in spatial sound capture. High-channel-count microphone arrays, which are often spherical, can be used to capture a sound field with rich spatial information. These spatially encoded recordings can be reproduced with realistic spatial cues in virtual reality applications [26].

By contrast, in the more traditional array processing research fields of audio source separation and enhancement, there has been a trend toward smaller numbers of microphones. If there are fewer microphones than sound sources, then the source signals cannot be perfectly separated using linear time-invariant methods. So-called *underdetermined source separation* techniques rely on special properties of certain natural sound signals, such as time-frequency sparsity, to design nonlinear separation algorithms [55, 56]. For example, a mixture of three or four speech sources can often be separated from a single-microphone recording by splitting it into finely spaced time intervals and frequency bands, then assigning each time-frequency component to a single source [57].

Underdetermined methods became prominent at a time when arrays with more than two microphones were rare. Now that large arrays are more common, these methods are still useful because they can help to estimate acoustic channel param-

eters and they provide extra degrees of freedom to apply additional constraints or improve robustness to errors. They also integrate easily with data-driven compositional models, such as nonnegative matrix factorization [58], and machine learning methods, such as deep neural networks [59].

As machine learning methods have become increasingly popular among audio researchers, array processing risks becoming an afterthought. In speech recognition systems, microphone arrays are often used only for preprocessing the input to a deep neural network [6]. Single-microphone machine learning methods have also attracted significant recent attention from hearing aid researchers [37]. While there have been some promising results, it seems unwise to invest too many resources into single-microphone methods for human listening enhancement: the reason that single-microphone noise reduction techniques do not improve intelligibility is not that the algorithms are not clever enough at identifying noise; it is that humans, even those with hearing loss, already do a good job at separating sources and extracting information amid background noise. Any system that merely preprocesses signals using the same information available to the brain cannot be expected to offer much advantage.

1.3.3 Microphone array listening devices

If we hope to build listening devices that surpass normal human abilities, we should provide them with more information than is normally available to humans. We can do that by adding more acoustic sensors.

Microphone array hearing aids first appeared in the signal processing literature in the late 1980s. A series of dissertations between 1989 and 1994 examined the performance of fixed and adaptive directional beamformers [18–20]. These early arrays were designed to replace the directional microphones that had long been available for hearing aids and that were known to improve intelligibility in noise. Thus, array designs were evaluated primarily on their directivity, that is, on their ability to improve the signal-to-noise ratio for a sound source directly in front of the listener. Fixed analog beamformers used microphone arrays mounted on eyeglasses

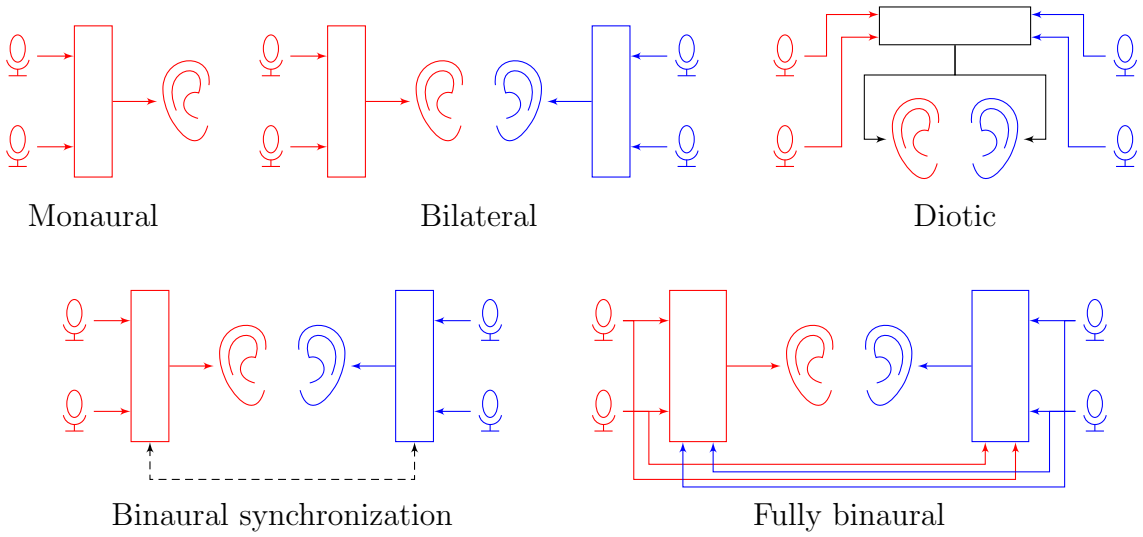

Figure 1.5: There are several possible processing configurations for listening devices on one or two ears.

to increase the directivity of hearing aids and therefore to improve intelligibility in noise [60–62]. Researchers compared simple delay-and-sum beamformers against more complex "superdirective" beamformers as well as digital adaptive beamforming algorithms [63–65]. Adaptive beamformers can provide better performance in reverberant environments, but are more complex [40].

During these early stages of research, most commercial hearing aids relied on analog processing technology. All-digital hearing aids became widespread after the turn of the century, enabling more sophisticated adaptive beamforming technologies [21, 66, 67] and statistical beamforming methods such as the speech-distortion-weighted multichannel Wiener filter [68]. It also became possible for hearing aids on either side of the head to communicate with each other via a wireless link [69]. Whereas most prior research had focused on monaural systems that output processed signals to a single ear, bilateral systems that operate independently in each ear, or diotic systems that output the same signal to both ears, fully binaural hearing aid systems could coordinate processing between the left and right ears [16, 70]. These configurations are compared in Figure 1.5. Optimistic researchers anticipated that

hearing aids would soon wirelessly stream audio data to perform binaural beamforming between the ears. Today, only a few state-of-the-art commercial hearing aids perform fully binaural beamforming; due to power constraints, the majority of wireless hearing aids only synchronize settings between earpieces.

Binaural hearing aids introduced a new challenge for array designers: preserving the listener's spatial awareness. Bilateral hearing aids that operate independently can apply different gains to the signals in each ear, distorting the interaural time and level differences that humans use to localize sounds [22]. Devices can synchronize processing settings to better preserve these cues. However, directional beamformers, even those designed to preserve the spatial cues of a target sound source, distort the cues of all other sounds [71, 72]. Over the last decade, signal processing researchers have developed many new methods to preserve interaural cues in beamformers [23, 73]. Most reduce spatial distortion by intentionally preserving parts of background sources; these binaural background-preserving beamformers anticipate the source-remixing space-time filters that are the focus of Chapter 4.

1.3.4 Scaling up microphone arrays

Most of the microphone array hearing aids studied in the literature have had only a few microphones and covered an area no larger than the head. If we hope to achieve superhuman hearing, these arrays are not nearly large enough.

To understand why, consider again the augmented vision analogy. The optical devices that can dramatically enhance human vision, microscopes and telescopes, are both far larger than the human eye. By the laws of optics, they have to be. There is no pair of eyeglasses that will let us resolve bacteria or the moons of Neptune. Why should we expect to hear a mouse across a busy street using an array no larger than a human head?

Just as larger lens systems can focus on objects that are smaller or farther away, larger microphone arrays have finer spatial resolution. For conventional anechoic beamformers, there is a well-known inverse relationship between the physical extent of the array and the width of the beam. While the space-time filters used in listening

Figure 1.6: Large wearable microphone arrays such as the Sombrearo can provide better spatial resolution than microphones worn on the ears.

devices are more complex to analyze, the same principle applies. The more information they can collect about the sound field, the better they will be able to remix the auditory scene.

One way to increase the spatial diversity of a microphone array is to place microphones on opposite sides of an acoustically opaque object, such as the torso. In Chapter 9, it will be shown that wearable arrays that span the body perform better than those with closely clustered microphones, such as eyeglasses. We can also increase the array aperture using large wearable accessories, such as the "Sombrearo" shown in Figure 1.6. The largest wearable array examined in this work has 80 microphones spread across the entire body.

There have been several efforts over the years to build massive-scale microphone arrays. Over thirty years ago, a 63-microphone analog array was used to isolate talkers in an auditorium [74]. Using digital signal processing, researchers were able to process an array of 512 microphones [75]. A few years later, a 1020-microphone array was used to demonstrate a novel computer architecture [76]. More recently, MEMS microphones have enabled large-scale arrays mounted on printed circuit boards [77, 78].

Rather than build one array with a massive number of microphones, we could also

distribute microphones throughout the environment [79]. Room-scale microphone arrays that surround most of the sound sources in a space are more useful than compact arrays that are surrounded by sound sources. Whereas compact arrays rely mostly on time or phase differences between microphones, distributed arrays also provide amplitude information [80, 81]. They can use triangulation to localize sound sources and can focus on sounds from a region of space, rather than a general direction [82, 83]. However, wirelessly connected devices suffer from bandwidth and latency limitations and from sample rate mismatch [84]. Distributed microphone arrays are the subject of Chapter 10.

1.4 Array Processing and Human Listeners

After more than 30 years of research and development, multiple dissertations, and dozens of crowdfunding campaigns for microphone array listening devices, and despite consistent evidence that spatial noise reduction can improve intelligibility, no large-microphone-array listening device has found success outside of the laboratory. Why not? This section explores reasons that large microphone arrays have not been widely adopted in human listening devices and proposes a new approach to space-time processing for augmented listening.

1.4.1 Humans are not machines

Microphone arrays have been widely successful in many machine listening applications. Spatial source separation algorithms are useful for meeting diarization [85] and acoustic event detection [86], while directional beamformers can improve the performance of automatic speech recognition systems. Arrays are also useful in teleconferencing applications, where they can reduce background noise and help to suppress feedback [87]. The vast majority of studies on microphone array listening devices have used the same type of processing as machine listening systems: the device attempts to completely separate all sound sources in order to focus on one

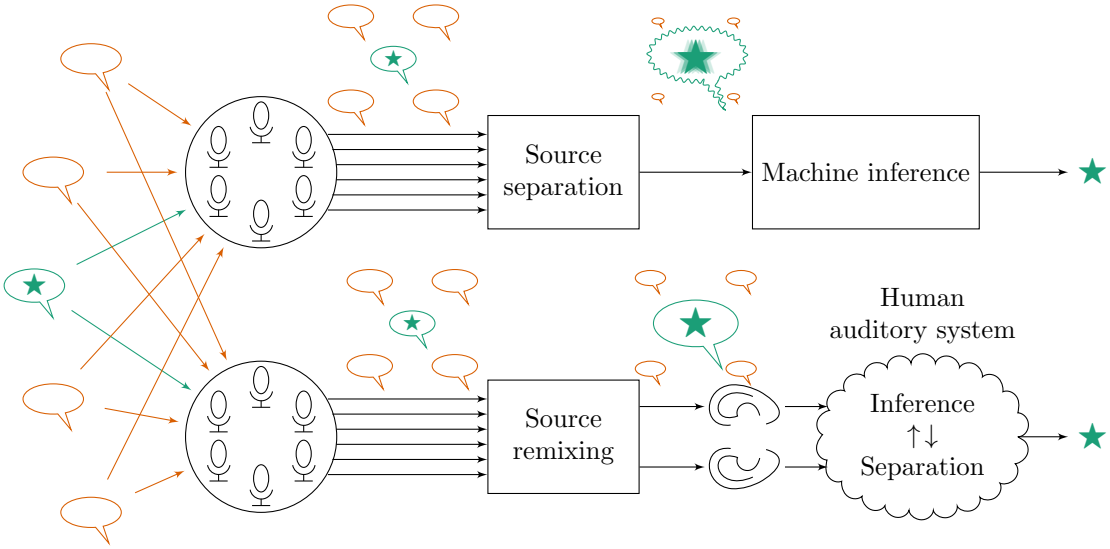

Figure 1.7: In machine listening systems (top), a source separation or beamforming stage performs noise reduction but may introduce distortion. In human augmented listening systems (bottom), the auditory system itself can help to compensate for unwanted noise.

signal of interest. If other sounds are included in the output, it is only to help reduce error sensitivity or spatial cue distortion [88].

This full-separation approach makes sense for machine listening, as illustrated in Figure 1.7: with a high-performance source-separation or beamforming front end, the inference application itself does not need to account for noise. For example, a speech recognition system could use a machine learning model trained on clean speech rather than many different types of background noise. While these machine listening algorithms often try to mimic the human auditory system, they do not enjoy humans' natural robustness to noise. In augmented listening, however, the brain is part of the system, along with its sophisticated source-separation and information-extraction machinery.

The auditory system does not need the listening device to completely isolate a sound source. Even if the device could do so perfectly, it would be disturbing to sit in a crowded restaurant and hear only one person talking. If a device introduces distortion while trying to reduce noise, as it surely would in challenging environments,

it might to more harm than good by confusing the auditory system's natural pattern recognition processes. For example, directional beamformers can distort interaural cues, impeding the brain's ability to distinguish sounds from different directions. Single-microphone noise reduction methods do not distort these cues, but they have their own problems: the artifacts introduced by nonlinear time-frequency masking algorithms (Chapter 7) are often interpreted by the brain as a distinct, unnatural sound stream, sometimes known as "musical noise." Thus, poorly executed source separation may be worse than no source separation at all.

Machine listening systems try to imitate human hearing; augmented listening systems need only supplement it.

1.4.2 Perceptual transparency

To support the auditory system's natural analysis capabilities, the output of the listening device should resemble a real-world sound mixture as closely as possible. An automatic speech recognition algorithm does not need to know the direction of a sound, the acoustics of the room, or even the timbre of the talker's voice in order to produce a transcription; this information can be discarded after the source-separation or noise-reduction stage of a machine listening system. The human brain, however, relies on this information to extract meaning from sound. To ensure that the augmented listening system is natural and comfortable for the user, we must apply human-specific perceptual constraints.

Spectral distortion: Certain types of space-time filters can apply different amounts of gain to different frequencies. Conventional beamformers have flat responses in the target direction, but frequency-dependent attenuation in other directions. Statistical space-time filters designed to minimize squared error apply more gain at frequencies where the target signal is strong and less gain where it is weak, which can introduce spectral coloration. Of course, some frequency-dependent gain may be desirable, for example if the user has high-frequency hearing loss.

Spatial distortion: Conventional beamformers can destroy the spatial cues, such as interaural time and level differences, of non-target sources [72]. Any sound that is not fully removed by the beamformer will appear to come from the target direction, causing a disturbing and potentially dangerous distortion effect. Anecdotally, it sounds like being in a long tunnel. Space-time filters can be designed to preserve spatial cues at the cost of noise reduction [22–24]. Spectral and spatial distortion are discussed in Chapter 4.

Delay: Listening device users hear both processed and unprocessed sounds at the same time. If the delay between these signals is more than a few milliseconds, it can cause disturbing distortion [89, 90]. These effects are most pronounced for the user's own speech because the delay interrupts the auditory feedback path used in speech production. Some delay is introduced by analog-to-digital and digital-to-analog conversion, but the most important source of delay is algorithmic. Frequency-selective processing methods, such as equalizers, filterbanks, and time-frequency masks, require longer delay to achieve finer frequency resolution. Delay constraints are the subject of Chapter 5.

A listening device could meet all three of the above perceptual constraints perfectly by applying pure amplification with no other processing. Such a system would be perfectly transparent, but would not be very useful. Meanwhile, a directional beamformer followed by a single-channel noise reduction algorithm could cause significant spectral distortion, spatial distortion, and delay. Augmented listening systems can strike a balance between these two extremes by remixing sound sources rather than fully separating them.

1.4.3 Sound source remixing

In a sense, the augmented listening problem is easier than many machine listening problems because the human auditory system can do much of the work on its own. The listening device merely needs to help. Rather than source separation or single-target beamforming, augmented listening systems should perform *source remixing*,

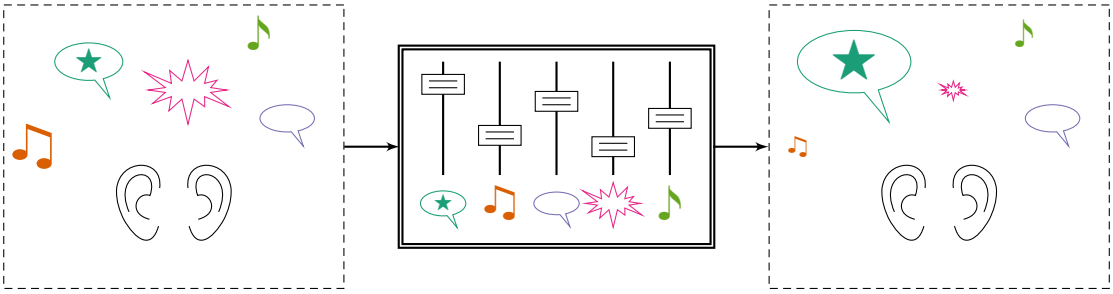

Figure 1.8: Instead of separating sound sources, the proposed augmented listening system remixes them, applying different processing to each.

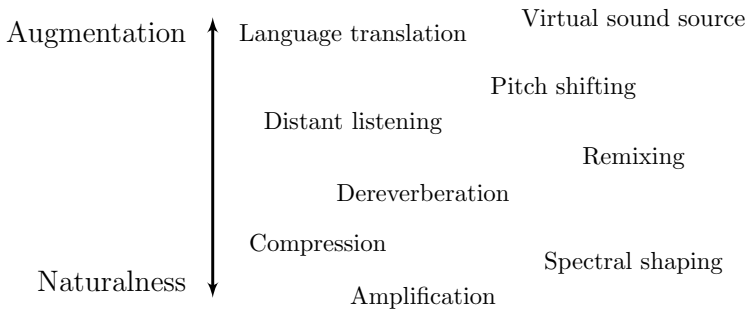

Figure 1.9: Augmented listening systems can apply many different types of processing. More aggressive processing can better enhance human listening abilities but may sound unnatural.

as illustrated in Figure 1.8. The space-time filter applies different processing to each source signal or group of source signals, then recombines the processed signals to form a new mixture. This approach is inspired by sound mixing in music, television, and film studios: each performer or instrument is recorded separately and the resulting signals are carefully processed and combined by an expert mixing engineer. A good mixture ensures that lyrics and dialogue are intelligible but also includes immersive environmental sounds, special effects, and music.

The amount of processing the device should perform depends on the situation and the user's hearing and cognitive abilities. Users with normal hearing or mild loss might prefer no processing at all in quiet environments; the listening device would intervene only in especially challenging circumstances. In other situations, the

device might need to completely remove or replace certain sound sources. Choosing what type of processing to apply and how much to alter the signals is a tradeoff between augmentation and naturalness, as illustrated in Figure 1.9. A similar idea has been proposed in object-based audio systems for television broadcasts [91, 92]: the broadcast includes separate streams for dialog, sound effects, and music, and each listener can adjust the mixture to trade off between intelligibility and immersiveness.

Remixing also has advantages from a signal processing perspective. It will be shown that filters that only slightly alter the levels of sound sources relative to each other introduce less spectral and spatial distortion than more aggressive filters. They may also be less sensitive to parameter estimation errors and require less delay. There are also important advantages of remixing for nonlinear processing such as dynamic range compression. Compression is traditionally applied after beamforming, so that all sounds in a mixture experience the same gain. It has been widely observed, however, that compression performs poorly in background noise [45] and that it can introduce distortion when applied to mixtures of multiple sounds [46]. Applying independent compression to each signal when possible can help to mitigate this distortion [93].

It is not yet understood what type of processing should be applied to each source. How much should we reduce noise to ensure that a conversation partner is intelligible for a particular listener? How much compression should we apply to different musical instruments? How much delay and distortion can the user tolerate in a crowded restaurant? What types of sound does the user care about and which sounds can be safely removed? Ideally, these processing settings would be automatically determined by classification algorithms according to each user's hearing profile and personal preferences. Such algorithms will require new clinical research that is beyond the scope of this work. In the meantime, however, we can address the many engineering challenges of source-remixing augmented listening systems.

1.5 Microphone Array Processing for Augmented Listening

In engineering, as in science more broadly, we often learn the most by studying extreme cases. By building the tallest tower, the fastest plane, or the largest microphone array, we test the limits of current technologies and understand how they could be improved in the future. While there is undoubtedly a need for incremental progress in listening device performance, the goal of this dissertation is to demonstrate dramatic improvements that could change the way we approach listening technology, even if they require impractically elaborate systems. This work describes an ambitious system that, if realized, would empower a listener to independently adjust every sound source in the environment, to hear what they want to hear how they want to hear it.

1.5.1 Proposed system

The proposed augmented listening system is shown in Figure 1.10. The core of the system is a wearable microphone array, which includes a minimum of two microphones, one in or near the left ear and one in or near the right ear. These in-ear microphones allow the system to produce natural-sounding mixtures as they would have been heard by each ear. The wearable array should also include microphones spread across the body. While most prior research on wearable arrays has focused on eyeglasses, necklaces, and small hats, this work will show that sensors should be spread as far apart as possible, including on the acoustically opaque torso, to maximize spatial diversity.

To dramatically enhance human hearing beyond its normal limits, the system must collect information not just from one listening device, but from many microphones spread throughout the environment. Microphones are already abundant in human spaces. Mobile devices, game systems, teleconferencing equipment, and smart appliances often contain multiple microphones. Better yet, the room could be deliberately instrumented with microphones embedded in walls, ceilings, and furniture. Different devices may be used in different ways depending on their individual bandwidth,

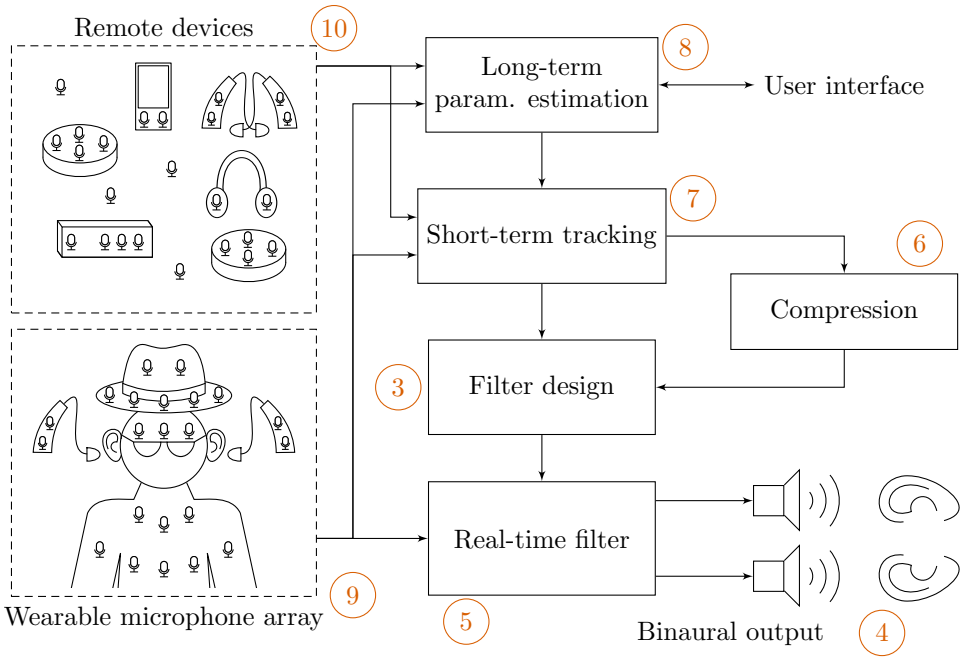

Figure 1.10: A complete augmented listening system brings together many different kinds of audio signal processing. The circled numbers indicate chapters that cover each part of the system.

latency, and synchronization with respect to the listening device. A cooperative listening network might also include other augmented listening devices, allowing users to hear through each other's ears.

An augmented listening system operates on several time scales. At the longest scale, classifiers determine the listener's environment and detect what types of sound are present. Based on user interface controls or an automatic decision-making algorithm driven by perceptual models, the system decides what kind of processing should be applied to each sound source or group of sources.

Source separation and acoustic channel estimation algorithms also work best over long time scales. They would likely use tens of seconds of data to learn where the sound sources are in the room or, more generally, to learn a space-time statistical model for each source. If any wireless devices have severe sample rate offsets or long transmission delays, then their data can only be used at this scale. Large motion, such as a talker or wearable-array user walking across the room, can also be tracked on multiple-second time scales. These computational tasks would likely be delegated to a powerful mobile device, workstation, or cloud service.

Other decisions must be made several times per second. Small motion, such as breathing or gesturing, must be tracked quickly. The spectra of speech signals change on time scales of tens of milliseconds, and many sparsity-based source separation and enhancement methods adapt to these changes. Dynamic range compression algorithms typically react to increases in signal level within a few milliseconds and to decreases in signal level over a few hundred milliseconds. The listening device might also incorporate audio data from distant devices with low-latency digital connections to help track motion and short-term source spectra. Many of these tasks might be executed on the listening device itself, perhaps with some assistance from a more powerful mobile device.

At the shortest time scale, the listening device must process sound signals with an input-to-output delay of a few milliseconds to avoid disturbing distortion or perceptible echoes. To meet this strict delay requirement, the device must perform filtering using its own internal processor. Typically, only microphones with wired or analog wireless connections, such as near-field magnetic induction, can be directly processed

by the space-time filter. Faraway digital wireless devices can be used only if the delay due to transmission is short compared to the acoustic propagation time between them and the listener and if their sample clocks can be synchronized with that of the listening device.

The binaural outputs generated by the space-time filter are carefully designed to preserve the listener's spatial awareness and to minimize unintended spectral distortion. The filters can be adjusted to trade off between distortion, noise reduction, delay, and motion robustness.

1.5.2 Contributions of the dissertation

Many past works on microphone array listening technology have focused on one small piece of the overall system, such as directivity or spatial cues. But to realize truly dramatic performance improvements and translate them to the real world, we must bring together tools from many areas of signal processing: source separation, event classification, causal filtering, nonlinear gain control, and distributed sensing, to name a few. This work takes a broad approach, describing the ways in which different tools fit into the larger system and how established ideas and methods must be adapted for human listening. It also introduces problems that have not been previously addressed in the literature, such as delay-constrained array processing, the design of body-scale wearable microphone arrays, and the effects of body movement on array performance.

This work does not attempt to solve every problem required to build a powerful augmented listening system. There are several missing pieces that will be needed to make the different parts of the system work together and many improvements will be needed to improve reliability and robustness. For example, the time-frequency methods used to implement dynamic range compression, underdetermined source separation, and asynchronous distributed processing do not meet the strict delay constraints of human listening. The results presented here on moving and deformable microphone arrays are first steps toward understanding a problem that will likely take many years to solve. This dissertation also does not include any clinical research. To

show that the proposed augmented listening technologies really can help people to hear better, they must ultimately be evaluated by real people. At this stage, however, the priority is to understand the engineering challenges that must be overcome to make augmented listening systems work better.

Many of the results in this work do not only apply to augmented listening. Results on delay-constrained array processing could apply to many spatial signal processing applications, even outside of audio. The acoustic channel measurement and distributed array processing methods proposed in the latter half of the dissertation are directly applicable to machine listening problems such as speech recognition. Wearable microphone arrays could be used for recording, telecommunication, and machine listening applications. However, the experimental results in this work focus on listening applications.

The technical chapters of the dissertation can be roughly divided into two parts. Chapters 3 through 6 present theoretical and experimental analysis of array signal processing for human listeners, emphasizing the ways in which human augmented listening differs from machine listening. They show that many of the unique constraints of human listening can be met by using larger microphone arrays. The theoretical work in this dissertation builds on the existing literature by framing audio enhancement as a remixing rather than separation problem, by incorporating delay constraints that are usually ignored in array processing, and by developing new theory and methods for dynamic range compression. Chapters 7 through 10 deal with the implementation challenges of realizing a practical augmented listening system in complex dynamic environments and propose novel architectures and algorithms to address these challenges. This work proposes new solutions to well-studied problems, such as a more scalable sparse model for speech mixtures and a resampling-free method for asynchronous arrays, and introduces previously unstudied problems, such as compensating for motion in deformable microphone arrays.

Array signal processing for human listeners

The technical material in the dissertation begins with Chapter 3, which reviews the mathematics of array signal processing. Space-time filter criteria are derived in both the time and frequency domains.

Chapter 4 shows how these filter criteria must be adjusted for human listeners. A weighted-square-error criterion is used to design a source-remixing filter that alters the relative levels of the source signals in the enhanced mixture. The filters must be carefully designed to avoid spectral distortion and to preserve the listener's spatial awareness. Such filters are well-studied for the case of a single target source of interest; here, the analysis is extended to remixing filters, with particular attention to the effect on filter performance of the relative levels of different source signals in the output. This analysis is easiest in the frequency domain.

Chapter 5 moves to the time domain, applying constraints to ensure that listening devices have imperceptible delay. Classic theoretical tools from causal signal processing are applied to characterize the tradeoff between delay and performance for a delay-constrained microphone array processing system. These theoretical tools provide exact expressions for squared-error performance in certain special cases, while new experiments demonstrate delay-performance tradeoffs in a real room.

Whereas the earlier chapters focus on linear time-invariant systems, Chapter 6 introduces nonlinearity in the form of dynamic range compression. Although it is used in almost every hearing aid and many consumer-targeted listening devices, compression is poorly understood and can fail in noisy environments. This chapter presents new mathematical analysis to explain why compression performance degrades in noise and proposes a novel source-specific compression strategy that leverages the spatial diversity of microphone array devices.

Implementation of an augmented listening system

The first half of the dissertation is enough to build an augmented listening system for controlled laboratory conditions. However, the system must be able to deal with

uncertain and constantly changing real-world conditions, especially for large arrays that span multiple devices. Chapter 7 introduces a new set of signal processing tools: time-varying source separation methods based on the short-time Fourier transform. These methods exploit the sparsity of speech and other natural signals to let inference systems do more with limited spatial information. The chapter proposes a simplified time-varying separation method that is more computationally tractable than related state-of-the-art methods, but that can scale well to large arrays.

Chapter 8 discusses the longstanding open problem of acoustic channel estimation. To design space-time filters, either linear or nonlinear, the system must learn how acoustic signals propagate from each source to each microphone. Traditional blind source separation methods do not work well in the challenging environments in which augmented listening devices would be most useful. This chapter proposes several semi-blind methods that use prior knowledge about the sound source signals themselves. For example, a known speech phrase can be used as a pilot signal to estimate channel parameters in keyword-activated listening systems.

Next, Chapter 9 covers the design of wearable microphone arrays, which has been discussed surprisingly little in the microphone array listening device literature. A first-of-its-kind wearable microphone data set is used to study tradeoffs in sensor placement, while a prototype embedded implementation provides insight about practical design challenges. The chapter also considers the previously unaddressed problem of small relative motion between microphones, which would occur in any wearable array. A second-order statistical model is used to characterize the effects of such motion both theoretically and empirically. The chapter compares several motion-robust processing strategies and demonstrates their performance experimentally using a wearable microphone array.

Although large wearable arrays can significantly improve performance compared to conventional earpieces, they do not quite deliver on the promise of superhuman augmented listening abilities. Chapter 10 shows how devices spread throughout the environment can cooperate to provide far greater performance than any listening device could on its own. Between two different large-scale experiments, this chapter combines nearly all the techniques developed in this dissertation. One experi-

ment combines established source separation and channel measurement techniques in a hierarchical architecture suitable for networked listening devices, while another demonstrates a novel partially asynchronous source separation technique for moving, independently clocked devices that does not require explicit tracking or resampling.

Finally, Chapter 11 explains how the tools developed in previous chapters can be combined into a complete augmented listening system. It also outlines the open research problems that must be addressed to achieve superhuman augmented listening.

As noted above, this dissertation is concerned with extreme listening systems unlike any that have been built before. To study such large-scale wearable and distributed arrays, however, we need realistic data. When this project began, suitable data sets simply did not exist. The next chapter describes how the Augmented Listening Laboratory team developed new data sets that let us study wearable and distributed microphone arrays with dozens or hundreds of sensors.

Chapter 2

Data and Methods

A major impediment to research on microphone array listening devices has been a lack of high-quality data sets. Audio data sets, including speech and other sound source recordings and room impulse responses, are crucial for audio researchers. They can be used to conduct controlled experiments to compare different array designs and processing strategies. Standardized data sets are used to compare results between research groups, for example as part of challenges like CHiME [6], REVERB [15], and SiSEC [94].

While there is ample data for binaural head-related transfer functions [95, 96] and one large data set for behind-the-ear earpieces [97], to the best of our knowledge before this work there had not been any public data sets of acoustic measurements for larger wearable microphone arrays. Similarly, while there are several high-channel-count real-world speech data sets [6, 85, 98–100], these are intended primarily for speech recognition applications and do not include ground-truth source recordings or impulse responses. To fill this important gap, this dissertation presents two first-of-their-kind data sets, one for wearable microphones (Section 2.3) and one for massive-scale distributed microphone arrays (Section 2.4). Several smaller data sets were also collected for particular experiments.

To evaluate the performance of listening enhancement systems, we must strike a balance between control and realism in experiments. At one extreme, convolving simulated [101] or measured [102] room impulse responses with prerecorded anechoic sounds [103,104] allows us to simulate arbitrary rooms and sound mixtures and track every reflection of every syllable through the entire system, but algorithms that work well in simulations might not work well in the real world. At the other, audio recorded

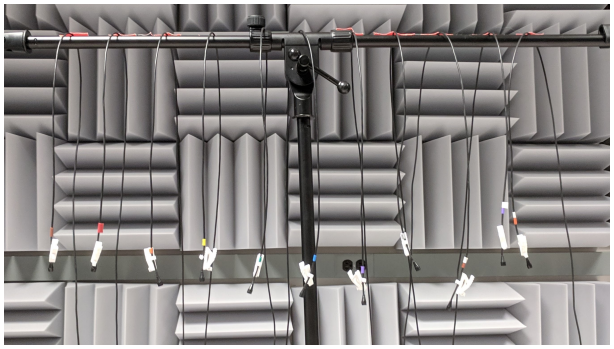

Figure 2.1: An array of omnidirectional lavalier microphones.

from live sound sources by microphones on live humans in uncontrolled environments can validate the real-world performance of processing systems, but no variables can be manipulated and performance can only be evaluated qualitatively.

The experiments in this dissertation try to strike a balance: recordings are made in real rooms, but the sound is from loudspeakers played one-at-a-time rather than from live talkers. These data incorporate real room acoustics, transducer nonlinearities, and environmental noise, but allow software experiments to manipulate the number and intensity of sound sources and the number and placement of microphones. Critically, they also let us measure the amount of each sound source in the output of the listening device so that we can objectively quantify system performance.

2.1 Equipment and Facilities

2.1.1 Microphone arrays

Nearly all the data used in this dissertation were recorded using a set of 16 Countryman B3 omnidirectional lavalier condenser microphones, shown in Figure 2.1. These tiny microphones have a flat frequency response over the range of audible speech frequencies (100 Hz–20 kHz). Unlike most lavalier microphones, this model is hardwired, ensuring that all microphones can be sampled synchronously by the audio

Figure 2.2: Mannequins are better than human subjects at standing perfectly still for long experiments.

interface.

Because augmented listening devices are worn by humans, a realistic data set should use microphones affixed to human subjects. Since working with human subjects requires special care (Section 2.5) and live humans introduce uncontrollable motion and noise into recordings (Chapter 9), only some of the data used in this dissertation was recorded using human subjects. Most wearable-microphone data was captured using a pair of life-size plastic mannequins, shown in Figure 2.2. Because the mannequins have unnaturally small ears that do not support earpieces, custom-made plastic ears were attached to the mannequins' heads. These ears are not intended to have fully realistic head-related transfer functions. In Chapter 9, we consider how well plastic mannequins match the acoustic properties of real humans; there do not appear to be substantial differences in acoustic transparency in the test conditions used in this dissertation.

To simulate the electronic devices that often house microphone arrays, the research team developed two custom array enclosures, shown in Figure 2.3.[1] Every wearable microphone array includes a pair of behind-the-ear earpieces, each of which holds two lavalier microphones. These mimic the most popular style of hearing aids today. A second enclosure, used for the distributed-array data set, imitates the form factor of a smart speaker. While most commercial voice-enabled speakers use low-cost

[1]The author gratefully acknowledges Uriah Jones, Matthew Skarha, and Benjamin Stoehr for their assistance in designing and producing these enclosures.

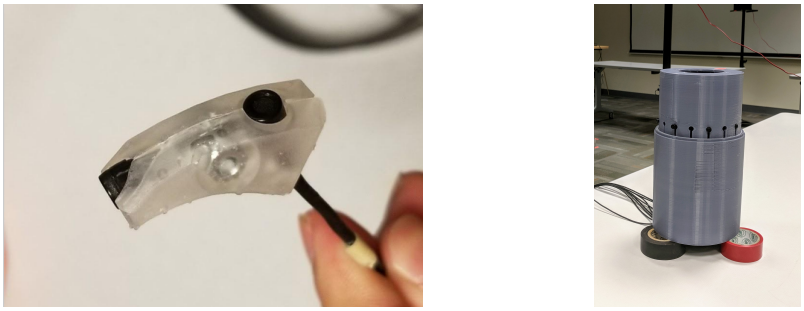

Figure 2.3: Custom enclosures emulate behind-the-ear earpieces (left) and smart speakers (right).

digital MEMS microphones, this prototype records from up to 16 studio-quality microphones. The microphone slots are arranged in a circle with diameter 10 cm.

2.1.2 Playback and recording

To ensure that all audio data uses a common time scale, both playback and recording are performed over wired connections by a 16-input, 10-output digital audio interface. The interface is a Focusrite Scarlett 18i20 with attached Scarlett OctoPre. All data is sampled at 48 kHz and 24 bits unless otherwise indicated. The interface is controlled by a fanless miniature Windows computer using the Reaper digital audio workstation software. This system is capable of simultaneously transmitting to 10 loudspeakers and recording from 16 microphones, all using a common sample clock. The recording system is mounted in a mobile cart that can be used to perform field recordings and live demonstrations, as shown in Figure 2.4.

Prerecorded sound sources are played back over a set of 10 Presonus Eris E3.5 two-way studio monitors, shown in Figure 2.5. These monitors are poor analogues for real human talkers: they have inconsistent frequency responses, relatively strong nonlinearities, and directivity patterns that do not closely resemble those of human talkers. The research team is actively developing full-range loudspeakers designed to mimic the directivity of human talkers. Fortunately, the evaluation methods used in this dissertation are not strongly affected by imperfections in the loudspeakers:

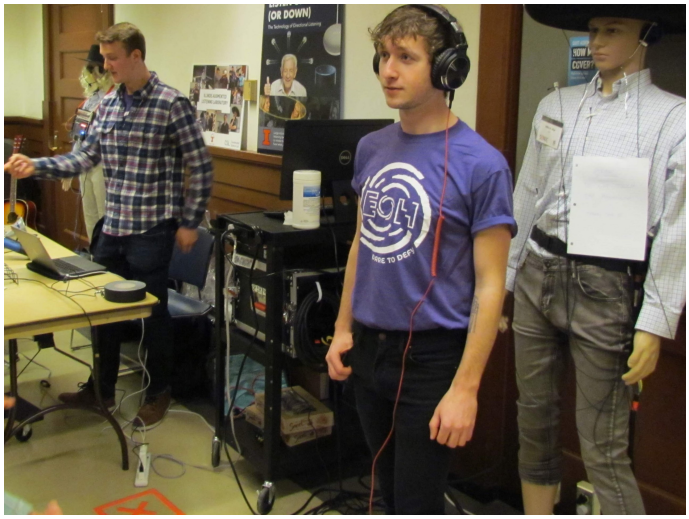

Figure 2.4: A mobile recording cart houses a 16-input, 10-ouput digital audio interface attached to a fanless computer. It can be used for field experiments and live demonstrations.

Figure 2.5: Studio monitors are used to simulate human talkers.

Figure 2.6: The Illinois Augmented Listening Laboratory features an acoustically treated recording space.

processing systems are evaluated by comparing the processed outputs to the recorded inputs, not to the original source data.

2.1.3 Augmented Listening Laboratory

Most of the experiments described in this dissertation were performed at the University of Illinois Augmented Listening Laboratory, an acoustics research space in the Coordinated Science Laboratory. The laboratory, shown in Figure 2.6, is equipped with a variety of microphones, loudspeakers, recording and playback devices, array enclosures, mannequins, wearable accessories, and physical and electronic prototyping equipment used to develop novel audio devices.

The laboratory features a low-reverberation recording space treated with 8" melamine foam wedges, 2" Auralex Studiofoam wedges, and a heavy curtain. The remainder of the laboratory is untreated and contains many smooth, reflective surfaces. Thus, it can be used for both low-reverberation and moderate-reverberation recordings. Although the laboratory is quiet at high frequencies, there is intense low-frequency noise from a mechanical room across the hall. A representative acoustic impulse response and environmental noise spectrum are shown in Figure 2.7. Fortunately,

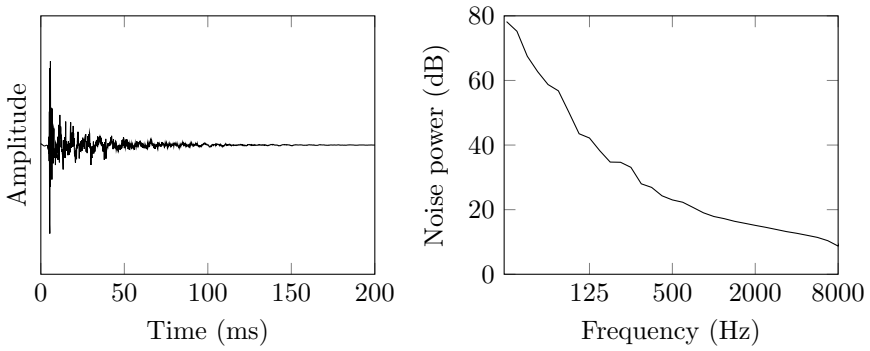

Figure 2.7: Representative room impulse response (left) and background noise spectrum (right) in the Augmented Listening Laboratory.

the noise is mostly concentrated below 100 Hz and can be removed by a highpass filter without adversely affecting speech signals.

2.2 Experimental Methods

2.2.1 Impulse response measurements

An important component of any data set used for array processing is acoustic impulse responses (AIRs) that describe how sound propagates from sources to microphones. These AIRs are useful for two reasons: first, because audio acoustics is quite linear, they can be used to simulate an acoustic mixture using arbitrary sound sources without having to make time-consuming audio recordings. Second, AIRs can be used to derive mathematically optimal space-time filters, as explained in Chapter 3. These ground-truth channel parameters are a useful baseline against which to compare blind source separation methods and channel estimation methods, like those in Chapter 8.

Acoustic impulse responses were measured using repeated linear or exponential sweeps [105, 106] with duration at least ten seconds. Acoustic channels can also be measured using pseudorandom noise signals such as maximum length sequences,

which are almost perfectly temporally uncorrelated [107].

2.2.2 Speech recordings

To ensure that the recorded sound mixtures are as similar as possible to live human speech in the same room, we should use speech samples recorded in an anechoic chamber. Only a few anechoic speech data sets are available. The TIMIT database [103] is widely used, but is not freely available to the public and has a restrictive license. Instead, the more recent experiments in this dissertation use the new VCTK corpus, which is free and has a Creative Commons Attribution license [104]. The corpus consists of recordings of different talkers reading individual sentences taken from British newspapers. To create the source signals used in these experiments, utterances from individual talkers were manually concatenated together with brief gaps between sentences. The resulting speech clips have the pace and cadence of radio news broadcasts. The talkers were chosen to represent a variety of genders, timbres, and accents.

Speech clips were played from the loudspeakers in two ways. First, each speech clip was played back from its corresponding loudspeaker by itself and recorded by the microphones. The background noise of the room was also recorded with no loudspeakers active. An isolated source signal as recorded by the microphones is known as a *source spatial image* or simply *source image* [108]. These source images, or a subset of them at specific microphones of interest, are the desired output of a source separation algorithm. Because it is generally impossible to unambiguously recover the sound produced by a source, separation performance is measured against the sound as received by the array. By recording these images separately and then adding them together to form a simulated mixture, we can control the number and intensity of source signals in the mixture and we can compare the output of a separation or remixing algorithm against an exact ground-truth output. Furthermore, we can separately analyze the effect of the processing system on each source. Input and output source images and the performance metrics that use them will be described mathematically in Chapter 3.

Recording sources one-at-a-time is less realistic than capturing true simultaneous mixtures. Because each recording includes background noise, the sum of the image signals has unnaturally amplified noise. It is also possible that the acoustics of the room could change between recordings, for example if any people or furniture move. To improve the realism of the experiments, recordings were also made with several loudspeakers active simultaneously. This mixture data cannot be used directly to quantify system performance, but it can be used to evaluate it qualitatively. Simultaneous recordings are especially important for moving sources and microphones because motion cannot be exactly reproduced between recordings (Chapter 9).

2.2.3 Large arrays

Much of this dissertation is concerned with large-scale arrays that combine data from dozens or even hundreds of microphones. Such arrays could be realized in practice using small, inexpensive digital MEMS microphones [54] and highly parallel digital signal processing hardware [109]; the Augmented Listening Laboratory research team is developing just such a system. For studio-quality recordings, however, it is expensive and impractical to record from so many microphones simultaneously. Instead, large microphone arrays are simulated by recording source images at one set of microphones at a time, then moving the microphones and repeating the recordings. This process is repeated for as many microphone locations as desired.

To ensure that such recordings are realistic, it is essential that the source signals be produced in exactly the same way during each recording. The loudspeakers must not be moved between recordings and all playback and recording must be performed using the same audio interface to ensure a common timescale between experiments. There are still weaknesses of this method, however: the microphones themselves, including their housing, enclosures, and cables, are moved between recordings, altering the acoustics near the array. For experiments with more than two wearable-array users, the mannequins must also be moved for some recordings, further altering the room acoustics. Each recording captures the same prerecorded source signals produced by the same loadspeakers, but the background noise in the room changes between takes,

making it seem less spatially correlated than it really is. Thus, this method is most appropriate for quiet, well-controlled laboratory environments.

2.2.4 Preprocessing

Although most data was recorded at 48 kHz, such a high sample rate is unnecessary for the speech data used in most experiments in this dissertation. To reduce computational complexity, most experiments use decimated data sampled at 16 kHz.

Because the laboratory and other rooms used to collect data contain strong low-frequency noise, and because the source image method artificially amplifies background noise, the source images are preprocessed to remove this background noise. When microphone impulse responses are used to generate synthetic mixture data, a highpass filter is applied to the system output before analysis to remove unreliable signal components below about 200 Hz.

For speech recordings made with the distributed array, noise is removed using a time-frequency generalized-singular-value-decomposition method similar to [110]. For each source image recording, the short-time-Fourier-transform vectors are prewhitened based on the measured space-time statistics of the background noise in the room. These vectors are projected onto the subspace defined by their four dominant singular vectors, then de-whitened and transformed back to the time domain. Four singular vectors were used instead of one to preserve the diffuse components of the source images and avoid unfairly advantaging rank-1 source separation methods. No denoising was applied to the simultaneous mixture signals used to generate audio examples.

2.3 Wearable Microphone Data Set

In the early phases of this research project, it was difficult to evaluate listening enhancement methods because there was little real-world data available for microphone-array listening devices. Early publications [93, 111, 112] relied on a data set of acous-

Figure 2.8: The human and mannequin subjects have generally similar height and build.

tic impulse responses for behind-the-ear earpieces with three microphones each [97]. Thus, the largest array that could be simulated had six closely spaced microphones. Many ambitious spatial processing methods, such as binaural source remixing (Chapter 4), require much larger arrays. Therefore, a new data set was collected that includes over 8000 acoustic impulse responses for microphones placed all over the body and on several wearable accessories [113]. This data set has been released to the public under a Creative Commons Attribution license and is available on the Illinois Data Bank, an archival service maintained by the University of Illinois Library [114].

The data set includes two subjects, one human and one mannequin, shown in Figure 2.8. The mannequin is 183 cm tall and has a head circumference of 56 cm, while the human is 181 cm tall with a head circumference of 61 cm. The two subjects wore the same button-up shirt for all recordings except the experiments comparing different outerwear, which used the mannequin only.

Test signals were generated from a total of 24 positions in a ring around the subject. Because of limited space in the laboratory, these signals were generated by six loudspeakers in a quarter-ring and the subject was carefully rotated four times to capture all 24 directions of arrival. The loudspeakers sat on stands about 150 cm above the tile floor and 200 cm away from the subject, as shown in Figure 2.9.

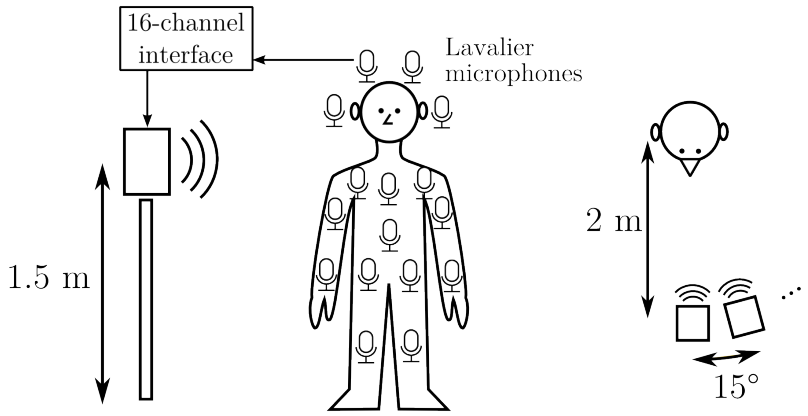

Figure 2.9: Acoustic impulse responses were measured from 24 directions of arrival around each subject. Figure adapted from [113].

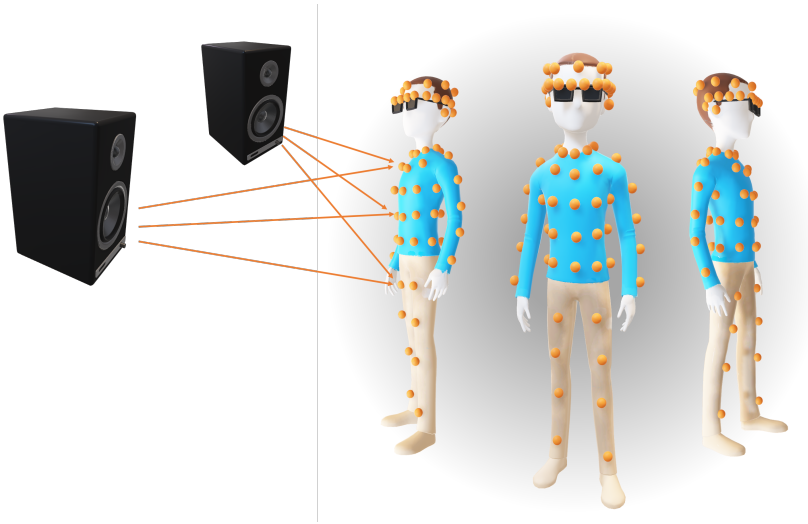

Figure 2.10: Microphones were placed at 80 positions across the body.

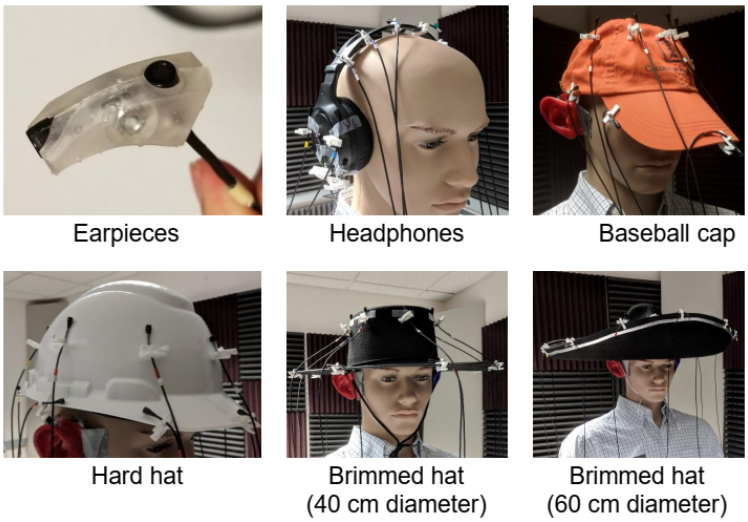

Figure 2.11: A total of 80 microphone positions were measured across several wearable accessories.

The lavalier microphones were placed at 80 positions across the body, as shown in Figure 2.10. One microphone was affixed to each ear using medical tape so that it recorded sound just outside the ear canal. These two microphones are used as the left and right references for all binaural processing experiments throughout the dissertation. Another four microphones were encased in two behind-the-ear earpieces (Figure 2.3). Ten more were clipped to eyeglasses, eight to a headband, and the remaining 56 to the subject's clothing.

Wearable microphone arrays might take the form of wearable accessories, such as headphones, hats, or glasses. To help engineers and designers compare the performance of different wearable accessories, supplemental measurements were taken with five head-mounted accessories, each with sixteen microphones, as shown in Figure 2.11. These included over-the-ear headphones, a baseball cap, a hard hat, a hat with a 40 cm flat brim, and a hat with a 60 cm curved brim. The latter, known as the "Sombrearo," is also featured in several smaller-scale data sets used throughout the dissertation.

Although the author considers wearable microphone arrays to be quite fashionable (see Figure 1.6 from the previous chapter), some users might prefer to hide arrays

Figure 2.12: Recordings were made with several types of clothing covering microphone on the torso.

under clothing. To evaluate the effects of outerwear on array performance, the torso measurements were repeated with the microphones clipped to an undershirt and covered by a cotton t-shirt, a cotton button-up shirt, a cotton sweatshirt, a fleece pullover, a wool coat, and a leather jacket. The outerwear items are shown in Figure 2.12.

The acoustics of the human and mannequin bodies and the performance of different array designs are evaluated in Chapter 9.

2.4 Distributed Microphone Data Set

To meaningfully augment normal human hearing in challenging noisy environments, or to dramatically improve the performance of machine listening systems in those conditions, we must use microphone arrays far larger than would fit in a wearable accessory, even one as large as the Sombrearo. In situations with many competing sound sources, it would be beneficial to use microphones distributed throughout the

Figure 2.13: Smart-speaker enclosures and mannequins were spread throughout a large conference room.

space. Although there has been significant recent research interest in distributed arrays (see Chapter 10), and although there are several real-world speech data sets using distributed arrays [6, 100], there had not been any such data sets that provide the ground-truth acoustic impulse responses and source spatial images necessary for source separation and enhancement research.

To simulate large distributed arrays of wearable and smart-home devices [115], the Augmented Listening Laboratory team collected a large data set using 10 loudspeakers and 160 microphones spread throughout a large, reverberant conference room, as shown in Figure 2.13. The data set includes impulse response measurements, background noise recordings, and 60-second speech recordings. This data set has also been released to the public under a Creative Commons Attribution license on the Illinois Data Bank [116].

The experiment included four wearable arrays of 16 microphones each. Eight microphones were placed in the ears, earpieces, and eyeglasses, and eight were clipped to different positions on the torso. To capture all four listening positions, the two mannequins were placed in two locations each. There were also twelve smart-speaker arrays with eight microphones each (Figure 2.3). The two smart-speaker enclosures were placed in six positions each at the center of twelve tables. Speech signals were produced by ten loudspeakers spread throughout the room, shown in Figure 2.14.

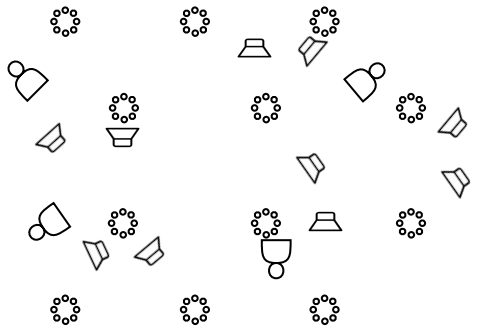

Figure 2.14: Placement of mannequins, circular smart-speaker arrays, and loudspeakers in the large conference room.

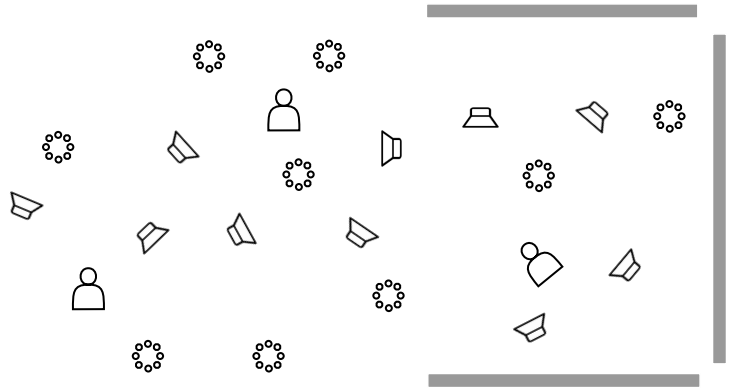

Figure 2.15: Distributed array in the Augmented Listening Laboratory.

The conference room is strongly reverberant with $T_{60} \approx 780$ ms. Each listener is in front of only a few loudspeakers, so that many sources have no direct acoustic path but are instead dominated by reflections from walls and furniture. This challenging acoustic environment reflects the adverse conditions in which augmented listening systems are most needed, but it also makes analysis more difficult. For example, it is difficult to evaluate interaural cues (Chapter 4) when many sources have no direct path. Therefore, many experiments in this dissertation use a smaller data set recorded using the same equipment in the Augmented Listening Laboratory.

This smaller data set used ten loudspeakers, three listener positions and nine smart-speaker positions, for a total of 120 microphone channels. The sources and

arrays were arranged throughout the treated and untreated parts of the laboratory, as shown in Figure 2.15. The data includes impulse responses and 20-second speech clips.

2.5 Experiments with Human Subjects

To study the design of microphone arrays for human listening enhancement, we must conduct experiments with human subjects. Three experiments reported in this dissertation used data collected from human subjects under protocols approved by the University of Illinois Institutional Review Board:

1. The wearable microphone data set described in Section 2.3,
2. Recordings from wearable microphones on a single moving human, which will be described in Chapter 9, and
3. Recordings from wearable microphones on multiple moving humans, which will be described in Chapter 10.

These experiments used human subjects because they could not practically be completed with nonhuman analogues. In the case of the wearable microphone data set, it was unknown whether mannequins are a reasonable acoustic analogue for real humans; in fact, part of the purpose of the experiment was to evaluate their differences. The latter two experiments were to specifically evaluate the impact of human motion, including subtle motion such as breathing and more complex motions such as dancing.

The protocols used to conduct these three experiments have similar procedures and risks. Microphones were clipped onto the clothing of human subjects and sounds were played over loudspeakers at an intensity similar to that of conversational speech. The primary risk for participants in these studies is to their privacy: photographs and video were taken in order to document experiments. This risk was mitigated by obtaining separate and explicit consent to use photographs and videos in research

materials. There was also a privacy risk to non-participants in the vicinity of the experiment. Their privacy could be violated if their speech was accidentally captured by the recording equipment. To mitigate this risk, recordings were performed in a closed laboratory outside of normal business hours. All recordings that captured conversations from neighboring laboratories were excluded from the data set.

Chapter 3

Arrays and Spatial Filtering

Engineers struggle to replicate the function of the human auditory system. In most ways—dynamic range, frequency resolution, source separation, speaker and speech recognition—the human auditory system is a match for even our most advanced sensing and computing technology [7]. There is one area, however, in which modern technology wins handily: spatial processing. Humans have only two ears, and those ears are right next to each other. While these two sensors provide remarkably rich information about the three-dimensional direction of arrival of sound events [4, 117], they cannot compete with arrays of dozens or hundreds of microphones spread around a room. This dissertation is concerned with augmenting natural human hearing using the spatial processing at which machines excel but humans do not.

Listening machines can use microphone arrays to localize, track, separate, and filter sounds in space. When a sound arrives at a single microphone, there is often no way to tell that sound's direction of arrival. When a sound arrives at multiple microphones, however, it will be captured by the nearest microphone first, then the next microphone a few milliseconds later, and so on. A system with many microphones spaced reasonably far apart from each other, known as a *microphone array* [13, 14], can exploit these time differences of arrival between sensors to determine the direction of arrival of one or more sounds. The task of estimating signal direction is called *localization*. It can be used by a conferencing system to point a camera at a person talking, in machine maintenance to pinpoint a part making a sound it shouldn't, or a traffic camera to identify motorcycles violating noise ordinances.

A microphone array, like that shown in Figure 3.1, can also be used to isolate sounds from a particular direction. The signals from different sensors are delayed,

Figure 3.1: An array of microphones or other sensors can be used to process signals spatially.

weighted, or filtered and then added together in such a way that a signal of interest is amplified and other signals cancel out. A processing system that isolates one sound source or focuses on all sounds from a particular direction is called a *beamformer* [8, 9, 51]. Beamformers are increasingly popular in teleconferencing systems, hands-free voice communication [87], distant speech recognition [118], including in popular smart-home devices, and, of course, in listening devices [17]. Many of the array processing methods proposed in this dissertation can apply to these other array applications as well.

Arrays can do more than just beamform, however: a large enough array can be used to separate multiple signals that each arrive along multiple reverberant paths, filter them independently, and recombine them, possibly with multiple output channels for multiple users or ears [10]. Such a system cannot be said to form a "beam", so we will use the more general terms *spatial filtering* and *space-time filtering* for more complex processing methods. To design these filters, we need a mathematical model for how acoustic waves propagate from one or more sound sources—such as talkers, musical instruments, loudspeakers, or noisy appliances—to each of the microphones in an array. We can use these acoustic channel models to design filters that process sounds based on a variety of optimization criteria. In this chapter, we will review the signal representations, channel models, and optimization criteria that will be used throughout the dissertation.

Table 3.1: Subscripts used to indicate signals in different domains/bases.

Notation	Meaning	Mapping
$x(t)$	Continuous-time signal	$\mathbb{R} \to \mathbb{R}$
$x_{\mathrm{d}}[k]$	Discrete-time sequence	$\mathbb{Z} \to \mathbb{R}$
$x_{\mathrm{fb}}[k, b]$	Discrete-time filterbank sequence	$\mathbb{Z} \times \{1, \ldots, B\} \to \mathbb{R}$
$X(\Omega)$	Continuous-time Fourier transform of x	$\mathbb{R} \to \mathbb{C}$
$X_{\mathrm{d}}(\omega)$	Discrete-time Fourier transform of x_{d}	$[-\pi, \pi] \to \mathbb{C}$
$X_{\mathrm{df}}[f]$	Discrete Fourier transform of x_{d}	$\{0, \ldots, F-1\} \to \mathbb{C}$
$X_{\mathrm{tf}}[k, f]$	Short-time Fourier transform of x_{d}	$\mathbb{Z} \times \{0, \ldots, F-1\} \to \mathbb{C}$

3.1 Notation Used in the Dissertation

The following notation will be used throughout the dissertation:

- Continuous-index and mixed-index signals are indexed using ( ) and discrete-index sequences are indexed using [ ].
- Lowercase letters indicate time-domain signals and sequences. Uppercase letters indicate frequency- and STFT-domain signals and sequences.
- When the same letter is used for a signal in different domains/bases, they will be distinguished by subscripts. For reference, commonly used subscripts are tabulated in Table 3.1.
- Vectors and matrices are indicated by bold symbols.
- The complex conjugate is indicated with superscript $*$, the matrix transpose with superscript T and the Hermitian transpose with superscript H.
- The imaginary unit is $j = \sqrt{-1}$.
- Italic e is the base of natural logarithms.
- Bold $\mathbf{e}_n$ is the unit vector with a 1 in position n. For example, $\mathbf{e}_2^T = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}$.

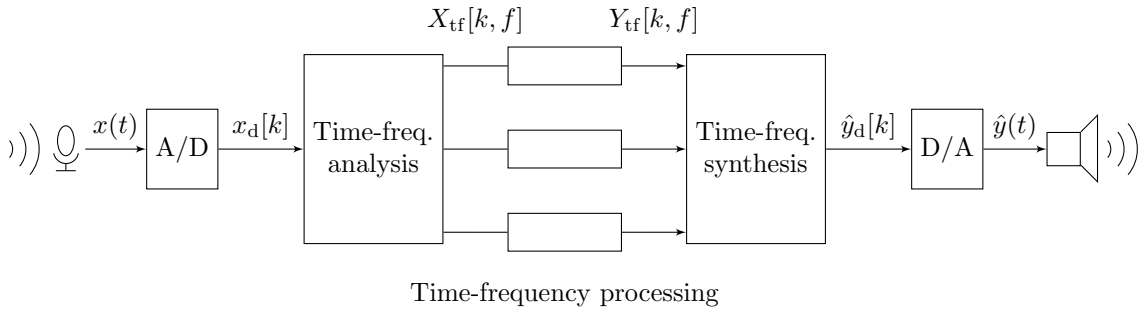

Figure 3.2: A listening system processes signals in continuous-time, discrete-time, and time-frequency representations.

- The all-zero and all-one vectors/matrices are $\mathbf{0}$ and $\mathbf{1}$, respectively, and the identity matrix is $\mathbf{I}$. If it is not clear from context, the size will be indicated by a subscript, e.g. $\mathbf{1}_M$.
- The Kronecker delta function is $\delta[\cdot]$. The Dirac delta is $\delta(\cdot)$. The Heaviside step function is $u(\cdot)$.
- The set of real numbers is $\mathbb{R}$. The set of nonnegative real numbers is $\mathbb{R}^+$. The set of complex numbers is $\mathbb{C}$. The set of integers is $\mathbb{Z}$.
- Statistical expectation is denoted by $\mathbb{E}[\cdot]$ and covariance is denoted by $\mathrm{Cov}(\cdot)$. Expectation and covariance are with respect to all random variables in the argument unless stated otherwise.
- There is no notational distinction between random and nonrandom signals. When signals are modeled as random processes, they will be defined as such in the text.

3.2 Signal Representations

To analyze multimicrophone augmented listening systems, it will be helpful to work with and convert between different representations of signals, as shown in Figure 3.2.

3.2.1 Continuous and discrete time and frequency

The real-world inputs to and outputs from an audio processing system are continuous-time signals. These will be denoted in the format $x(t)$, where t is a real-valued time variable and x is a real-valued signal measuring acoustic pressure or, in practice, an electrical analogue of acoustic pressure.

To analyze the spectral content of signals, and to conveniently model linear time-invariant systems such as room acoustics and many types of signal processing, it will be convenient to work in the frequency domain. If a signal $x(t)$ has finite energy, then it has a continuous-time Fourier transform (CTFT) given by

$$X(\Omega) = \int_{-\infty}^{\infty} x(t) e^{-j\Omega t} \, \mathrm{d}t \tag{3.1}$$

for all real-valued radian frequencies Ω. This dissertation is primarily concerned with frequencies in the audible range, which is often stated as 20 Hz to 20 kHz ($\Omega = 40\pi$ to 40000π).

Nearly all modern listening devices are digital, meaning that they operate on quantized, discrete-time signals generated by analog-to-digital converters. We will ignore the effects of quantization in this dissertation since they are usually orders of magnitude smaller than the effects of the acoustic noise that we hope to address; however, see [119, 120] for the author's recent work on quantized array processing. The discrete-time sequence sampled from $x(t)$ is given by

$$x_{\mathrm{d}}[k] = x(kT_{\mathrm{s}}), \tag{3.2}$$

for all integers k, where T_{s} is the sample period. If $x_{\mathrm{d}}[k]$ has finite energy, then its discrete-time Fourier transform $x_{\mathrm{d}}[k]$ is

$$X_{\mathrm{d}}(\omega) = \sum_{k=-\infty}^{\infty} x_{\mathrm{d}}[k] e^{-j\omega k} \tag{3.3}$$

for real-valued discrete-time frequencies ω.

Most of the experimental data presented in this dissertation is sampled at 48 kHz, so that $T_s = 1/48000$ sec. Because this is well above the Nyquist rate for audible signals, it is assumed that no aliasing occurs. Sampling and reconstruction are therefore assumed to be linear processes, which will allow us to analyze the effects of discrete-time signal processing in continuous time. In particular, we will use the CTFT domain to derive and analyze spatial and space-time filters that are implemented by discrete-time digital processing.

3.2.2 Time-frequency representations

Most interesting audio signals, such as speech and music, have frequency spectra that vary over time. For such sounds, it might not be useful to take a Fourier transform of the entire signal. Instead, many audio applications use time-frequency representations such as filterbanks and the short-time Fourier transform.

Many hearing aids use filterbanks to separate discrete-time signals into different frequency bands [28]. These separated signals are

$$x_{\mathrm{fb}}[k, b] = \sum_{\tau=-\infty}^{\infty} h_{\mathrm{d},b}[\tau] x_{\mathrm{d}}[k - \tau] \tag{3.4}$$

for $b = 1, \ldots, B$, where k is a time index, b is a band index, and $h_{\mathrm{d},b}[\tau]$ is the unit pulse response of the analysis filter for band b. Filterbank signals are real-valued and processed using convolutional filters, just like time-domain signals. Many different types of filter can be used and the bands need not be uniform.

Another popular time-frequency representation is the short-time Fourier transform (STFT) [121–123], given by

$$X_{\mathrm{tf}}[k, f] = \sum_{\tau=-\infty}^{\infty} \mathrm{awin}(kT_{\mathrm{step}} - \tau) x_{\mathrm{d}}[\tau] e^{-j2\pi\tau f/F} \tag{3.5}$$

for frequency indices $f = 0, \ldots, F-1$ and time indices k, where T_{step} is the number of samples between frames and $\mathrm{awin}(\tau)$ is an analysis window function. In the

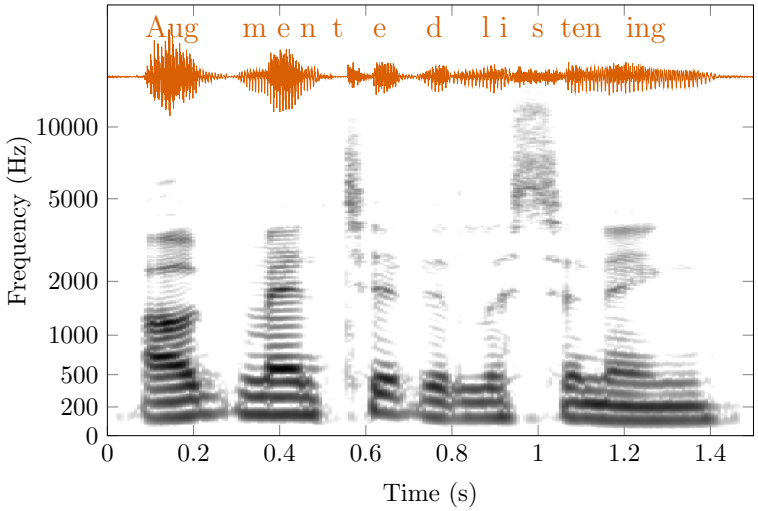

Figure 3.3: A spectrogram is a time-frequency representation of a signal. The time-domain waveform is shown above the spectrogram on the same time scale.

experiments presented here, the analysis window is a raised cosine function, also known as a Hann window:

$$\text{awin}(\tau) = \begin{cases} \frac{1+\cos\frac{2\pi\tau}{L}}{2} & \text{if } -\frac{L}{2} \leq \tau < \frac{L}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{3.6}$$

where L is the window length, which should be less than or equal to F. The STFT breaks the discrete-time signal into length-L frames, which usually overlap, and applies a tapered window to suppress frequency-domain ripples. Finally, the length-F discrete Fourier transform (DFT) is applied to each block to produce the STFT representation.

Unlike the filterbank representation, $X_{\text{tf}}[k, f]$ is complex-valued. It is usually processed via frequency-by-frequency complex multiplication.

The STFT is often used to analyze speech signals. Figure 3.3 shows the magnitude of the STFT of a speech signal—the author saying "augmented listening"—with time on one axis and frequency on the other. This visual representation is known as a *spectrogram*. In many speech processing applications, such as automatic speech

recognition, only the magnitude of the STFT is used and the phase is discarded. In array processing, however, phase information is crucial because it encodes time differences of arrival between nearby sensors.

The STFT can be inverted by taking the inverse DFT of each frame and adding the overlapping blocks:

$$x_{\mathrm{d}}[\tau] = \sum_{k=-\infty}^{\infty} \mathrm{swin}(kT_{\mathrm{step}} - \tau) \sum_{f=0}^{F-1} X_{\mathrm{tf}}[k, f] e^{+j2\pi\tau f/F}, \tag{3.7}$$

where $\mathrm{swin}(\tau)$ is a length-F synthesis window that isolates one period of the inverse DFT. When the analysis window is the Hann window and $T_{\mathrm{step}} = L/2$ (or $L/4$, $L/8$, etc.), the synthesis window is simply a rectangle; the Hann window satisfies the perfect reconstruction condition

$$\sum_{k=-\infty}^{\infty} \mathrm{awin}[k\frac{L}{2} - \tau] = 1 \quad \text{for all } \tau, \tag{3.8}$$

so that the overlap-add step of the inverse STFT reverses the windowing process from the forward STFT.

If the signal is altered before it is reconstructed, care must be taken to avoid temporal aliasing within frames. Because multiplication in the DFT domain is equivalent to circular convolution in the time domain, STFT processing can introduce "wrap-around" errors. Zero-padding can help to mitigate these errors, as can using non-rectangular synthesis windows. Furthermore, it is possible to create time-frequency signals that are not the STFT of any time-domain signal. That is, there exist some time-frequency signals X_{tf} such that $\hat{x}_{\mathrm{d}} = \mathrm{STFT}^{-1}\{X_{\mathrm{tf}}\}$ but $X_{\mathrm{tf}} \neq \mathrm{STFT}\{\hat{x}_{\mathrm{d}}\}$. Finally, although the STFT and inverse STFT are linear operations, STFT-based systems are not necessarily time-invariant because of the windowing process.

3.3 Array Processing System

Consider an audio processing system with J outputs, such as loudspeakers or earpieces, and an array of M microphone inputs, as shown in Figure 3.4.

3.3.1 Signals and sources

Let $x_m(t)$ be the continuous-time acoustic signal received by microphone m at time $t \in \mathbb{R}$ for $m = 1, \ldots, M$. Define the M-dimensional vector $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]^T$. Let $x_{\mathrm{d},m}[k]$, $x_{\mathrm{fb},m}[k, b]$, and $X_{\mathrm{tf},m}[k, f]$ be the discrete-time, filterbank, and STFT representations of $x_m(t)$ and let $\mathbf{x}_{\mathrm{d}}[k]$, $\mathbf{x}_{\mathrm{fb}}[k, b]$, and $\mathbf{X}_{\mathrm{tf}}[k, f]$ be the discrete-time, filterbank, and STFT representations of $\mathbf{x}(t)$.

In source separation [10, 11, 52, 53], an observed signal is assumed to be composed of a certain number of *source* signals. There is no one definition of what constitutes a source: it could be single talker, an air conditioner, or an entire orchestra, for example. To design a listening device, we would like to identify sound sources that humans perceive as a single auditory stream [3]. To design a space-time filter, meanwhile, we would like to choose sets of sounds that propagate according to a common set of equations, such as those from directional acoustic emitters. These perceptual and mathematical notions of "source" are not necessarily the same. As a compromise, in this dissertation we adopt the following functional definition.

Definition 3.1. A *source channel* characterized by signal vector $\mathbf{c}_n(t) \in \mathbb{R}^M$ is a set of sounds that, in the ideal desired processing system, would be processed as a single sound object. The set of all N source channel signals $\mathbf{c}_1(t), \ldots, \mathbf{c}_N(t)$ must completely characterize the observed signal, that is,

$$\mathbf{x}(t) = \sum_{n=1}^{N} \mathbf{c}_n(t). \tag{3.9}$$

This decomposition property also holds in discrete time and in all linear transform representations.

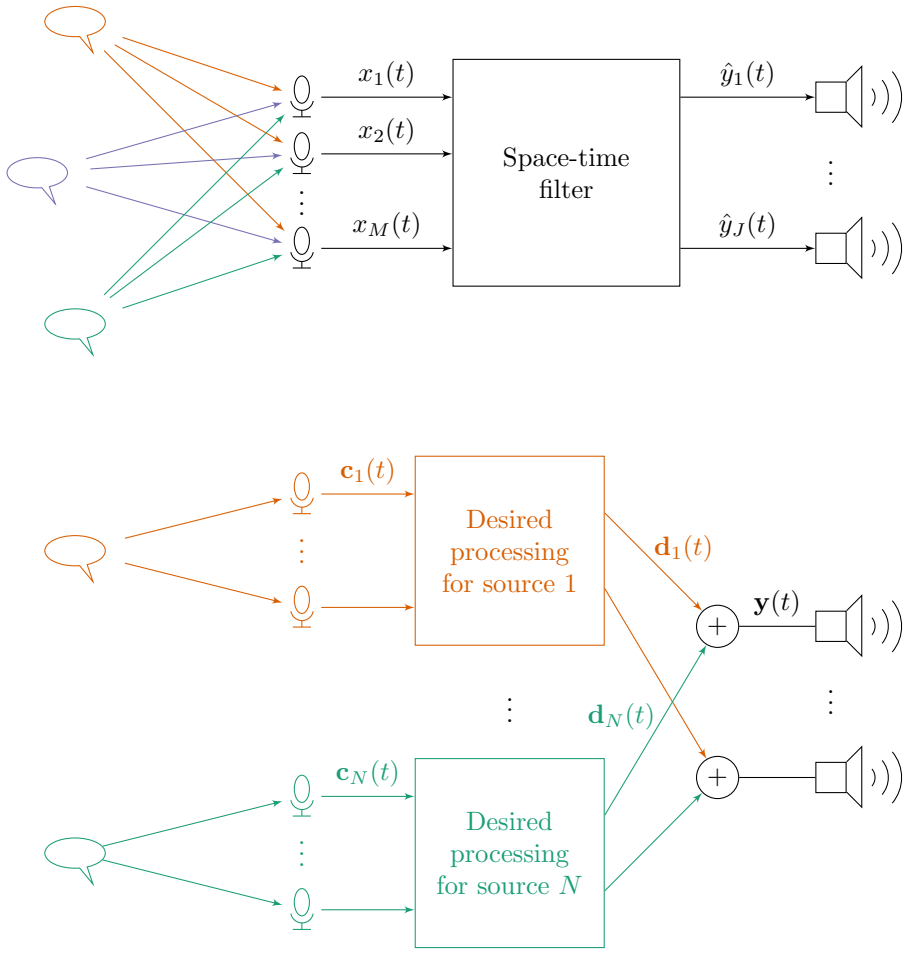

Figure 3.4: Top: Inputs to and outputs from a space-time processing system. Bottom: Desired effect of the remixing filter on source images $\mathbf{c}_1(t), \ldots, \mathbf{c}_N(t)$.

In other words, two sounds are part of a single source channel only if a hypothetical ideal system would apply the same processing to both of them. Each source channel could consist of an individual talker, a group of talkers, all sounds of a particular type (bird chirps, traffic noise, music), or a catchall background noise channel, for example. The source channel signals $\mathbf{c}_1(t), \ldots, \mathbf{c}_N(t)$ are sometimes known as *source spatial images* or simply *source images* since they represent the response of the array to the source signals [124].

To make the concept of a source channel more concrete, let us also define a set of desired output images $\mathbf{d}_n(t) \in \mathbb{R}^J$ for $n = 1, \ldots, N$, where J is the number of outputs of the system. Each $\mathbf{d}_n(t)$ is a processed version of its corresponding $\mathbf{c}_n(t)$. The overall desired output signal $\mathbf{y}(t) \in \mathbb{R}^J$ is

$$\mathbf{y}(t) = \sum_{n=1}^{N} \mathbf{d}_n(t). \tag{3.10}$$

In most of the dissertation, with the exception of Chapter 6, the desired processing will be assumed to be linear and time-invariant, even when the listening system that implements it is not. In this case,

$$\mathbf{d}_n(t) = \int_{-\infty}^{\infty} \mathbf{g}_n(v)\mathbf{c}_n(t-v)\,\mathrm{d}v, \quad n = 1, \ldots, N, \tag{3.11}$$

where each $\mathbf{g}_n(t)$ is a $J \times M$ matrix of desired impulse responses. For example, in a source-remixing system (Chapter 4), each desired impulse response might be a scaled Dirac impulse that changes the relative level of its corresponding signal in the mixture.

Note that the observational model (3.9) does not include a separate "noise" term: any environmental noise, microphone self-noise, quantization error, etc., must be included in the set of source channels. When performing directional beamforming or spatial filtering, it is usually assumed that at least one source channel has an image that is the convolution of a scalar signal with a vector of acoustic impulse responses (Section 3.4.1). When performing statistical space-time filtering, it is usually as-

sumed that all source channel signals are statistically uncorrelated with each other (Section 3.4.2).

3.3.2 Space-time filtering

Let $\hat{\mathbf{y}}(t) = [\hat{y}_1(t), \ldots, \hat{y}_J(t)]^T$ be the J-dimensional vector of system outputs. As with the input signals and desired outputs, these can be decomposed into N source channels.

Definition 3.2. The *output image* $\hat{\mathbf{d}}_n(t) \in \mathbb{R}^J$ is the contribution of source channel n to the system output for $n = 1, \ldots, N$. The set of N output images must completely characterize the output signal such that

$$\hat{\mathbf{y}}(t) = \sum_{n=1}^{N} \hat{\mathbf{d}}_n(t). \tag{3.12}$$

This decomposition property also holds in discrete time and all linear transform representations.

Linear time-invariant processing

The output images are easily defined for a linear time-invariant (LTI) system. In an LTI system, the output is given by

$$\hat{\mathbf{y}}(t) = \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{x}(t-v)\,\mathrm{d}v, \tag{3.13}$$

where $\mathbf{w}(t)$ is a $J \times M$ matrix of impulse responses that completely characterize the system. By linearity, we have

$$\hat{\mathbf{y}}(t) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{c}_n(t-v)\,\mathrm{d}v \tag{3.14}$$

$$= \sum_{n=1}^{N} \hat{\mathbf{d}}_n(t), \tag{3.15}$$

where

$$\hat{\mathbf{d}}_n(t) = \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{c}_n(t-v)\,\mathrm{d}v, \quad n = 1, \ldots, N.$$

If the source images have Fourier transforms, then (3.13) can be expressed in the frequency domain as

$$\hat{\mathbf{Y}}(\Omega) = \mathbf{W}(\Omega)\mathbf{X}(\Omega) \tag{3.16}$$

$$= \sum_{n=1}^{N} \mathbf{W}(\Omega)\mathbf{C}_n(\Omega) \tag{3.17}$$

$$= \sum_{n=1}^{N} \hat{\mathbf{D}}_n(\Omega), \tag{3.18}$$

where $\hat{\mathbf{D}}_n(\Omega) = \mathbf{W}(\Omega)\mathbf{C}_n(\Omega)$ for $n = 1, \ldots, N$. The equivalent relationships hold for LTI processing in discrete time.

Processing in the STFT domain is not necessarily time-invariant, but it can still be linear:

$$\hat{\mathbf{Y}}_{\mathrm{tf}}[k, f] = \mathbf{W}_{\mathrm{df}}[f]\mathbf{X}_{\mathrm{tf}}[k, f] \tag{3.19}$$

$$= \sum_{n=1}^{N} \mathbf{W}_{\mathrm{df}}[f]\mathbf{C}_{\mathrm{tf},n}[k, f] \tag{3.20}$$

$$= \sum_{n=1}^{N} \hat{\mathbf{D}}_{\mathrm{tf},n}[k, f], \tag{3.21}$$

where $\hat{\mathbf{D}}_{\mathrm{tf},n}[k,f] = \mathbf{W}_{\mathrm{df}}[f]\mathbf{C}_{\mathrm{tf},n}[k,f]$ for $n = 1,\ldots,N$ and $\mathbf{W}_{\mathrm{df}}[f] \in \mathbb{C}^{J\times M}$ is a matrix of complex coefficients. If negligible temporal aliasing occurs so that the system is approximately LTI, then $\mathbf{W}_{\mathrm{df}}[f]$ can be thought of as the discrete Fourier transform of a discrete-time impulse response, as suggested by the subscript "df".

Nonlinear processing

Many of the systems studied in this dissertation are nonlinear: the filters vary over time in response to changes in the input signals. Because such systems do not obey superposition, it is not obvious how to relate input signal components to output signal components—altering one input source image could affect all N output images. In these cases, the output image is defined as if the nonlinear system were a linear time-varying system. For example, in the STFT domain,

$$\hat{\mathbf{D}}_{\mathrm{tf},n}[k,f] = \mathbf{W}_{\mathrm{df}}[k,f]\mathbf{C}_{\mathrm{tf},n}[k,f], \quad n = 1,\ldots,N. \tag{3.22}$$

This definition is, in a sense, aspirational: a perceptually transparent listening system *should* satisfy superposition, at least approximately. That is, signals that are perceived as separate sounds before processing should also be perceived as separate sounds after processing. Thus, this definition is useful for listening systems that are working as they should. It fails, however, with certain types of highly nonlinear processing. Aggressive dynamic range compression (Chapter 6) can cause different sounds to modulate each other so much that they perceptually fuse [46]. In an extreme case, fast-acting, many-band compression limiting distorts any input signal into nearly white noise by forcing the signal to have a constant, flat spectrum. Other types of nonlinear processing can introduce new, artificial sounds that are perceived as separate auditory streams. Time-frequency masks (Chapter 7) are notorious for introducing artifacts known as "musical noise" because they can be perceived as musical. Thus, the concept of input and output images should be used with caution in nonlinear systems.

3.3.3 Performance evaluation

Source channels are defined here as groups of sounds that would be processed jointly in an ideal system.

For each source channel $n \in \{1, \ldots, N\}$, the error signal $\text{err}_n(t)$ is given by

$$\text{err}_n(t) = \hat{\mathbf{d}}_n(t) - \mathbf{d}_n(t) \tag{3.23}$$

and the overall error is

$$\text{err}(t) = \hat{\mathbf{y}}(t) - \mathbf{y}(t) \tag{3.24}$$

$$= \sum_{n=1}^{N} \text{err}_n(t). \tag{3.25}$$

The most commonly used error metric in this dissertation will be the signal-to-error ratio, which is also referred to as the signal-to-distortion ratio in the source separation literature [124].

Definition 3.3. The *signal-to-error ratio* (SER) for output signal $\hat{\mathbf{y}}(t)$ and desired signal $\mathbf{y}(t)$ is given by

$$\text{SER} = \frac{\text{mean}_t \, |\mathbf{y}(t)|^2}{\text{mean}_t \, |\hat{\mathbf{y}}(t) - \mathbf{y}(t)|^2}. \tag{3.26}$$

This metric differs slightly from other commonly used metrics such as the signal-to-noise ratio, which compares the power in a desired "target" output image to the power in undesired "noise" output images.

Definition 3.4. The *signal-to-noise ratio* (SNR) for target source channel 1 and noise source channel 2 is given by

$$\text{SNR} = \frac{\text{mean}_t |\hat{\mathbf{d}}_1(t)|^2}{\text{mean}_t |\hat{\mathbf{d}}_2(t)|^2}. \tag{3.27}$$

The SNR does not account for distortion to the target source and does not make sense for remixing filters, which do not distinguish between target and undesired

source channels. However, it is useful when it is difficult to measure the desired output image, as for moving arrays (Chapter 9).

Empirical performance evaluation

In many of the experiments presented in this dissertation, source images are used to directly compute the SER or SNR. Microphones and loudspeakers are placed in fixed positions in the test environment and prerecorded digital source signals $s_{\mathrm{d},1}[k], \ldots, s_{\mathrm{d},N}[k]$ are played one-at-a-time from different loudspeakers. The resulting recordings are the sampled source images $\mathbf{c}_{\mathrm{d},1}[k], \ldots, \mathbf{c}_{\mathrm{d},N}[k]$ plus background noise from the room. The N independent recordings are added together to give the sampled mixture $\mathbf{x}_{\mathrm{d}}[k]$.

Because separate recordings of the source images are available, they can be used to directly calculate the desired output images $\mathbf{d}_{\mathrm{d},n}[k]$ for $n = 1, \ldots, N$ and therefore the desired output $\mathbf{y}_{\mathrm{d}}[k]$. The experimental SER can therefore be computed in discrete time as

$$\mathrm{SER} = \frac{\mathrm{mean}_k \left|\mathbf{y}_{\mathrm{d}}[k]\right|^2}{\mathrm{mean}_k \left|\hat{\mathbf{y}}_{\mathrm{d}}[k] - \mathbf{y}_{\mathrm{d}}[k]\right|^2}. \tag{3.28}$$

Furthermore, the system under test can be applied to each source image recording individually to generate the experimental output images $\hat{\mathbf{d}}_{\mathrm{d},n}[k]$ for $n = 1, \ldots, N$. These can be used to break down the error for each source channel and to compute the experimental SNR. In nonlinear systems, the input-dependent filter parameters are computed from the full mixture, fixed as a linear time-varying filter, and applied separately to each source image. Because the ground-truth source images are available, we do not need to use the cumbersome source signal projections used in the popular BSS-EVAL package [124], which attempts to account for unknown distortion introduced by loudspeakers, microphones, and room acoustics when evaluating separation from simultaneous mixtures.

The downside of recording source images separately is that it artificially amplifies background noise when they are added together. It also skews the error calculations because background noise is included as part of every source channel so that filters

are penalized for removing the noise that they would typically be designed to suppress. Therefore, in many experiments the recordings are denoised using an oracle multichannel noise-reduction filter before processing and the room background noise is recorded separately to be used as its own source channel.

3.4 Models of Signal Propagation

One classic example of array processing is the delay-and-sum beamformer. In an anechoic environment with signals arriving from far away, the signal components at each sensor will be delayed versions of each other, with the delays depending on the direction of arrival of the signal. The delay-and-sum beamformer applies the opposite delays to the signals from each sensor, causing signals from the target direction to add constructively. These classic beamformers are widely used in radio frequency applications. For high-frequency, narrowband signals, time delays can be well approximated by phase shifts and so these beamformers are also called phased arrays.

Array processing is more complicated for audio signals, which span orders of magnitude in frequency content. Although delay-and-sum beamformers are sometimes used, they do not account for the reflections and reverberation that are common in indoor environments where humans spend much of their time. Listening devices must use more sophisticated models of signal propagation and more complex spatial filters to process real-world sounds.

3.4.1 Acoustic impulse responses and transfer functions

Fortunately for designers of spatial sound processing systems, acoustic propagation is mostly linear. If the sound sources and microphones do not move, then an acoustic system can be well modeled by an LTI system. Suppose that each source channel n represents a single directional sound source, such as a talker, instrument, or loudspeaker. Then its source image is the convolution of a source signal $s_n(t)$ with a

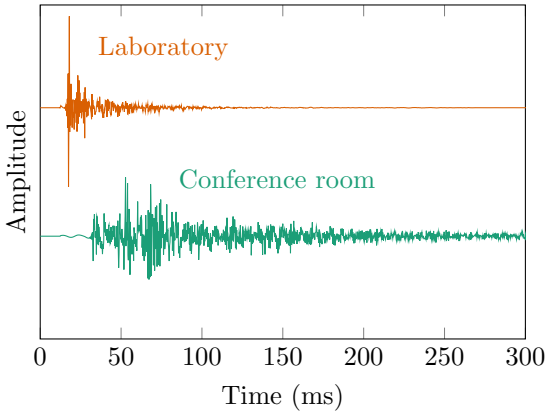

Figure 3.5: Acoustic impulse responses measured in the Illinois Augmented Listening Laboratory and in a large reverberant conference room. The two AIRs are plotted on different amplitude scales.

vector of *acoustic impulse responses* (AIR):

$$\mathbf{c}_n(t) = \int_{-\infty}^{\infty} \mathbf{a}_n(v) s_n(t-v)\,\mathrm{d}v, \quad n = 1, \ldots, N. \tag{3.29}$$

Figure 3.5 shows sampled acoustic impulse responses from two of the data sets used in this dissertation. One is from the acoustically treated Illinois Augmented Listening Laboratory, which has little reverberation. It consists of a strong direct path and a few early reflections. The other is from a large, strongly reverberant conference room in which the loudspeaker and microphone are far apart. It has no discernible direct path at all and a strong reverberant tail.

If $s_n(t)$ and $\mathbf{a}_n(t)$ have Fourier transforms, then (3.29) can be conveniently expressed as a vector-scalar product in the frequency domain:

$$\mathbf{C}_n(\Omega) = \mathbf{A}_n(\Omega) S_n(\Omega), \quad n = 1, \ldots, N, \tag{3.30}$$

where $\mathbf{C}_n(\Omega)$, $\mathbf{A}_n(\Omega)$, and $S_n(\Omega)$ are the continuous-time Fourier transforms of $\mathbf{c}_n(t)$, $\mathbf{a}_n(t)$, and $s_n(t)$, respectively. The vector $\mathbf{A}_n(\Omega)$ is commonly called the *acoustic transfer function* (ATF), although it is really a frequency response. This ATF can

be used to design a matched-filter beamformer [125],

$$\mathbf{W}_{\text{matched}}(\Omega) = \frac{\mathbf{A}_n^H(\Omega)}{\mathbf{A}_n^H(\Omega)\mathbf{A}_n(\Omega)}, \tag{3.31}$$

which projects onto the subspace of $\mathbf{A}_n(\Omega)$, preserving the source signal $S_n(\Omega)$ and suppressing any signal not parallel to $\mathbf{A}_n(\Omega)$. The delay-and-sum beamformer is a special case of (3.31) where the ATF consists of pure time delays.

For multiple directional sources, the overall mixing process is a matrix-vector multiplication:

$$\mathbf{X}(\Omega) = \mathbf{A}(\Omega)\mathbf{S}(\Omega), \tag{3.32}$$

where $\mathbf{A}(\Omega) = [\mathbf{A}_1(\Omega), \ldots, \mathbf{A}_N(\Omega)]$ is the $M \times N$ mixing matrix, $\mathbf{X}(\Omega)$ is the continuous-time Fourier transform of $\boldsymbol{x}(t)$, and $\mathbf{S}(\Omega) = [S_1(\Omega), \ldots, S_N(\Omega)]^T$.

This LTI model of signal propagation, which is sometimes called the rank-1 model because each $\mathbf{c}_n(t)$ has a rank-1 power spectral density matrix when modeled as a random process (Section 3.4.2), allows us to solve directly for the unknown source signals. Specifically, if $M \geq N$ and the mixing matrix has full column rank for all Ω of interest, then the signals can be perfectly separated by any unmixing filter whose Fourier transform $\mathbf{W}(\Omega)$ is a left inverse of $\mathbf{A}(\Omega)$. If the ATFs are known, then we do not need to make any assumptions about the signal statistics or content.

Relative transfer functions

In real-world spatial audio processing applications, it is rarely possible to directly observe the source signals $s_n(t)$ or ATFs $\mathbf{A}_n(\Omega)$. The system observes only the microphones' response to the signals. There is an ambiguity in these observations because $\mathbf{C}_n(\Omega) = S_n(\Omega)\mathbf{A}_n(\Omega) = \left(\frac{S_n(\Omega)}{Q(\Omega)}\right)(Q(\Omega)\mathbf{A}_n(\Omega))$ for any invertible frequency response $Q(\Omega) \neq 0$. It is therefore convenient to define the source signal to be equal to source image at a designated reference microphone, for example $s_n(t) = \mathbf{e}_1^T\mathbf{c}_n(t)$. The channel is then described by the *relative transfer function* (RTF) $\mathbf{A}_n(\Omega)/\mathbf{e}_1^T\mathbf{A}_n(\Omega)$ for $\mathbf{e}_1^T\mathbf{A}_n(\Omega) \neq 0$ [126]. Because the corresponding *relative impulse response* (RIR) is

generally noncausal and infinite in length, RTFs are most useful in frequency-domain and time-frequency-domain processing. In this dissertation, it will typically not be necessary to distinguish between relative and absolute transfer functions and the same notation will be used for both.

Other variations include RTFs with spatially processed references, in which case the reference is itself the output of a space-time filter rather than a specific microphone [127]. Many recent papers have used *relative early transfer functions* (RETF), which represent only the perceptually important direct path and early reflections of a signal [128]. These are useful for performing dereverberation. Although dereverberation is not explicitly addressed in this dissertation, the techniques proposed here could be modified to perform it by splitting the early and reverberant components into two separate source channels and suppressing the latter.

Time-frequency models

The rank-1 model (3.30) is also commonly applied in STFT-based array processing. If the acoustic impulse response is short compared to the DFT length of the STFT, then

$$\mathbf{C}_{\mathrm{tf},n}[k,f] \approx \mathbf{A}_{\mathrm{df},n}[f] S_{\mathrm{tf},n}[k,f], \tag{3.33}$$

where $S_{\mathrm{tf},n}[k,f]$ is the STFT of representation of the sampled source signal and $\mathbf{A}_{\mathrm{df},n}[f]$ is the DFT of the discrete-time impulse response $\mathbf{a}_{\mathrm{d},n}[k]$. The rank-1 model does not hold exactly in the STFT domain because windowing introduces dependencies across time frames and frequency indices. Because multiplication is performed independently for each frequency index f, this approximation is sometimes called the *narrowband model.*

Alternative models explicitly account for between-frequency effects [129] and across-time effects [130], but they will not be used in this dissertation. A third method, called the full-rank covariance model [131], models each source image STFT as a random vector with a full-rank covariance matrix. Statistical processing of source images is the subject of the next section.

3.4.2 Statistical source models

The convolutional model of Section 3.4.1 is a useful starting point for array processing, but it has limitations. Some sources are not directional, that is, their images are not characterized by a single AIR or ATF vector. Diffuse noise, for example, comes from many directions. Furthermore, some mixtures might not be easily separable using their spatial characteristics alone: sources might come from the same direction but have different spectral characteristics, or there might be more sources than sensors. In these cases, we can use information about the signals themselves—their correlation structures or frequency spectra—to help to separate and process the source images.

To leverage the powerful tools of statistical inference in designing space-time filters, the signals must be modeled as random processes. Specifically, each source spatial image $\mathbf{c}_n(t)$ is a random function of time. Throughout the dissertation, these images are assumed to have zero mean and to be statistically uncorrelated with each other, that is, $\mathbb{E}\left[\mathbf{c}_n(t)\right] = \mathbf{0}$ for all n and all t and $\mathbb{E}\left[\mathbf{c}_{n_1}(t_1)\mathbf{c}_{n_2}^T(t_2)\right] = \mathbf{0}$ for all t_1, t_2, and $n_1 \neq n_2$. In the time domain, each source channel is characterized by an $M \times M$ matrix autocorrelation function,

$$\mathbf{r}_{\mathbf{c}_n}(\tau) = \mathbb{E}\left[\mathbf{c}_n(t)\mathbf{c}_n^T(t-\tau)\right], \quad n = 1, \ldots, N. \tag{3.34}$$

Let $\mathbf{R}_{\mathbf{c}_n}(\Omega)$ be the corresponding power spectral density (PSD) matrix, that is, the continuous-time Fourier transform of $\mathbf{r}_{\mathbf{c}_n}(\tau)$. Similarly, let $\mathbf{r}_{\mathbf{x}}(\tau) = \mathbb{E}\left[\mathbf{x}(t)\mathbf{x}^T(t-\tau)\right]$. Since the images are assumed to be uncorrelated with each other, the mixture autocorrelation matrix is given by

$$\mathbf{r}_{\mathbf{x}}(\tau) = \sum_{n=1}^{N} \mathbf{r}_{\mathbf{c}_n}(\tau). \tag{3.35}$$

Let $\mathbf{R}_{\mathbf{x}}(\Omega)$ be the PSD matrix for $\mathbf{x}(t)$, which is similarly equal to the sum of the $\mathbf{R}_{\mathbf{c}_n}(\Omega)$'s. For an LTI processing system, the PSDs of the output signal and output

source images can be computed from the frequency response matrix:

$$\mathbf{R}_{\hat{\mathbf{d}}_n}(\Omega) = \mathbf{W}(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{W}^H(\Omega), \quad n = 1, \ldots, N \tag{3.36}$$

$$\mathbf{R}_{\hat{\mathbf{y}}}(\Omega) = \mathbf{W}(\Omega)\mathbf{R}_{\mathbf{x}}(\Omega)\mathbf{W}^H(\Omega). \tag{3.37}$$

If a particular source channel has an acoustic transfer function $\mathbf{A}_n(\Omega)$, then its power spectral density matrix is given by

$$\mathbf{R}_{\mathbf{c}_n}(\Omega) = R_{s_n}(\Omega)\mathbf{A}_n(\Omega)\mathbf{A}_n^H(\Omega), \tag{3.38}$$

where $R_{s_n}(\Omega)$ is the power spectral density of $s_n(t)$. Such a source channel is said to be rank 1 because its PSD is a rank-1 matrix.

In the STFT domain, the power spectral density is replaced by the covariance of the STFT samples:

$$\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k, f] = \mathrm{Cov}(\mathbf{C}_{\mathrm{tf},n}[k, f]). \tag{3.39}$$

This covariance may or may not vary as a function of the time index k. For stationary directional sources under the narrowband approximation (3.33), this covariance is the outer product of an ATF scaled by a source power:

$$\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[f] \approx R_{S_{\mathrm{tf},n}}[f]\mathbf{A}_{\mathrm{df}}[k]\mathbf{A}_{\mathrm{df}}^H[k], \tag{3.40}$$

where $R_{S_{\mathrm{tf},n}}[f] = \mathrm{Cov}(S_{\mathrm{tf},n}[k, f])$. An advantage of the full-rank covariance model is that it can help to compensate for the limitations of the narrowband approximation for stationary directional sources [131]. We will see in Chapter 9 that it can also help to model small motion of sources or microphones [132].

Statistical models allow the processing system to account for the space-time structure of signals and separate them based on their spectral as well as spatial characteristics. For example, if a target source has ATF $\mathbf{A}_1(\Omega)$ and an undesired noise source has PSD $\mathbf{R}_{\mathbf{c}_2}(\Omega)$, then the unbiased estimator of $s_1(t)$ that minimizes noise power in

the output is the minimum variance distortionless response (MVDR) beamformer,

$$\mathbf{W}_{\mathrm{MVDR}}(\Omega) = \frac{\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)}{\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}. \tag{3.41}$$

Notice that the matched filter beamformer (3.31) is a special case of (3.41) for spatially uncorrelated noise ($\mathbf{R}_{\mathbf{c}_2} \propto \mathbf{I}$). Other statistical filter design criteria are discussed in the next section.

3.5 Statistical Filter Design Criteria

It is not obvious why the audio signals received by an array should be modeled as random processes. After all, what is random about the acoustic waves that carry speech or music? Even if these signals were truly random, they are certainly not stationary, as implied by the use of the autocorrelation function and power spectral density. They also do not have Gaussian distributions, meaning that the linear estimators derived in this section do not truly minimize mean square error.

In array processing, the value of time-domain autocorrelation, frequency-domain power spectral density, and STFT-domain covariance matrices is often not that they accurately capture the temporal characteristics and frequency spectra of sound signals—though this is useful for stationary sources such as appliance or engine noise—it is that they encode the *spatial* structure of the signals. For example, if we wish to design a space-time filter that isolates one directional source and suppresses several other directional sources as well as diffuse noise, we do not need to explicitly constrain its response to each unwanted source; we can simply add the PSD matrices for all the unwanted channels and use the sum in place of $\mathbf{R}_{\mathbf{c}_2}(\Omega)$ in (3.41). In fact, we will do exactly that in Section 3.5.2.

For acoustically small arrays, such as those in conventional hearing aids, and especially in underdetermined scenarios with more sources than sensors, signal statistics provide valuable information that can help to separate sources. Indeed, small arrays can benefit from explicitly nonstationary, non-Gaussian models of source signals, as

explained in Chapter 7. Much of this dissertation, however, is devoted to large arrays with ample spatial diversity and many more sensors than sources. Such arrays rely primarily on spatial information to separate and process source signals: as we will see in Section 3.5.3, when a mixture covariance matrix is dominated by a few low-rank sources, a minimum-mean-square-error filter approaches a linearly constrained inverse filter that does not depend on the signal statistics. Thus, the PSD and covariance matrices are a mathematical convenience for designing filters with the desired spatial properties. In fact, in many of the experiments in this dissertation, filters are designed assuming that all speech signals have identical long-term average spectra, so that they rely entirely on spatial diversity.

Finally, although linear filters are suboptimal estimators for many of the non-Gaussian signals encountered in the real world, their efficiency, ease of implementation, and ease of analysis make them attractive for use in embedded systems such as augmented listening devices. In this section, we will derive several linear time-invariant space-time filters based on the second-order statistics of the source images.

3.5.1 Multichannel Wiener filter

The simplest statistical linear estimator used in array processing is the multichannel Wiener filter (MWF), which is the linear minimum-mean-square-error estimator of the desired output signal $\mathbf{y}(t)$ given the observed input signal $\mathbf{x}(t)$. That is, it solves the optimization problem

$$\mathbf{w}_{\mathrm{MWF}} = \arg\min_{\mathbf{w}} \mathbb{E}\left[\left|\mathbf{y}(t) - \hat{\mathbf{y}}(t)\right|^2\right], \tag{3.42}$$

where $\hat{\mathbf{y}}(t)$ is related to $\mathbf{w}(t)$ and $\mathbf{x}(t)$ by (3.13). By the orthogonality principle, the $\mathbf{w}(t)$ that minimizes (3.42) satisfies [133]

$$\mathbf{0} = \mathbb{E}\left[(\mathbf{y}(t) - \hat{\mathbf{y}}(t))\,\mathbf{x}^T(t-\tau)\right], \quad \tau \in \mathbb{R} \tag{3.43}$$

$$= \mathbb{E}\left[\left(\mathbf{y}(t) - \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{x}(t-v)\,\mathrm{d}v\right)\mathbf{x}^T(t-\tau)\right], \quad \tau \in \mathbb{R} \tag{3.44}$$

$$= \mathbf{r}_{\mathbf{yx}}(\tau) - \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{r}_{\mathbf{x}}(\tau - v)\,\mathrm{d}v, \quad \tau \in \mathbb{R}, \tag{3.45}$$

where

$$\mathbf{r}_{\mathbf{yx}}(\tau) = \mathbb{E}\left[\mathbf{y}(t)\mathbf{x}^T(t-\tau)\right] \tag{3.46}$$

is the cross-correlation between the desired and observed signals. Because the source images are assumed to be uncorrelated, (3.11) can be used to decompose the cross-correlation in terms of the source images $\mathbf{c}_n(t)$ and desired responses $\mathbf{g}_n(t)$ for source channels $1, \ldots, N$:

$$\mathbf{r}_{\mathbf{yx}}(\tau) = \sum_{n=1}^{N} \mathbb{E}\left[\mathbf{d}_n(t)\mathbf{c}_n^T(t-\tau)\right] \tag{3.47}$$

$$= \sum_{n=1}^{N} \int_{-\infty}^{\infty} \mathbf{g}_n(v)\mathbf{r}_{\mathbf{c}_n}(\tau - v)\,\mathrm{d}v. \tag{3.48}$$

Assuming for now that the filter is allowed to be noncausal, the Wiener-Hopf equation (3.45) can be written in the frequency domain as

$$\mathbf{R}_{\mathbf{yx}}(\Omega) = \mathbf{R}_{\hat{\mathbf{y}}\mathbf{x}}(\Omega) \tag{3.49}$$

$$\sum_{n=1}^{N} \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega) = \sum_{n=1}^{N} \mathbf{W}(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega), \tag{3.50}$$

for all Ω of interest, where each $\mathbf{G}_n(\Omega)$ is the continuous-time Fourier transform of the desired response $\mathbf{g}_n(t)$ and $\mathbf{W}(\Omega)$ is the continuous-time Fourier transform of

$\mathbf{w}(t)$. The solution is the frequency-domain noncausal MWF:

$$\mathbf{W}_{\text{MWF}}(\Omega) = \left(\sum_{n=1}^{N} \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\right)\left(\sum_{n=1}^{N} \mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1} \tag{3.51}$$

$$= \sum_{n=1}^{N} \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{R}_{\mathbf{x}}^{-1}(\Omega). \tag{3.52}$$

To illustrate the behavior of the MWF, consider again the example of a single rank-1 source $\mathbf{c}_1(t)$ and a full-rank noise source $\mathbf{c}_2(t)$. In this case, $\mathbf{G}_1(\Omega) = \mathbf{e}_1^T$ and $\mathbf{G}_2(\Omega) = \mathbf{0}$ for all Ω so that the filter attempts to isolate the target signal referenced to microphone 1 and remove the noise. The MWF is

$$\mathbf{W}_{\text{MWF}}(\Omega) = R_{s_n}(\Omega)\mathbf{e}_1^T\mathbf{A}_1(\Omega)\mathbf{A}_1^H(\Omega)\left(R_{s_1}(\Omega)\mathbf{A}_1(\Omega)\mathbf{A}_1^H(\Omega) + \mathbf{R}_{\mathbf{c}_2}(\Omega)\right)^{-1}. \tag{3.53}$$

Using the Woodbury identity to rearrange terms,

$$\mathbf{W}_{\text{MWF}}(\Omega) = \mathbf{e}_1^T\mathbf{A}_1(\Omega)\frac{R_{s_1}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)}{1 + R_{s_1}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}. \tag{3.54}$$

Unlike with the MVDR beamformer (3.41), a signal component parallel to $\mathbf{A}_1(\Omega)$ will be scaled down in order to reduce the noise component in that direction. The amount of this attenuation may vary as a function of frequency. Thus, while this filter achieves the minimum mean square error between the desired and actual output, it also distorts the spectrum of the signal of interest.

3.5.2 Linear constraints

To avoid distorting the signals within certain source channels, we can use *linearly constrained* filters. These also minimize mean square error, but with the additional requirement that exactly the desired processing be applied to a certain subspace of

input signals [9]:

$$\begin{aligned}\mathbf{w}_{\mathrm{LC}}(t) = \arg\min_{\mathbf{w}} \mathbb{E}\left[|\hat{\mathbf{y}}(t) - \mathbf{y}(t)|^2\right] \\ \text{s.t.} \quad \mathbf{W}(\Omega)\mathbf{A}_{\mathrm{c}}(\Omega) = \mathbf{Const}(\Omega),\end{aligned} \tag{3.55}$$

where $N_{\mathrm{c}} \leq M$ is the number of constraints, $\mathbf{A}_{\mathrm{c}}(\Omega) \in \mathbb{C}^{M \times N_{\mathrm{c}}}$ defines the constrained subspace, and $\mathbf{Const}(\Omega) \in \mathbb{C}^{J \times N_{\mathrm{c}}}$ contains the constraint values. Typically, such constraints are used to require error-free processing for a subset of the source channels that have rank-1 models:

$$\begin{aligned}\mathbf{w}_{\mathrm{LC}}(t) = \arg\min_{\mathbf{w}(t)} \sum_{n=N_{\mathrm{c}}+1}^{N} \mathbb{E}\left[\left|\hat{\mathbf{d}}_n(t) - \mathbf{d}_n(t)\right|^2\right] \\ \text{s.t.} \quad \hat{\mathbf{d}}_n(t) = \mathbf{d}_n(t) \quad \text{for } n = 1, \ldots, N_{\mathrm{c}}.\end{aligned} \tag{3.56}$$

Often, the linear constraints are applied to a set of desired source channels and the other source channels are unwanted noise, so that $\mathbf{d}_n(t) = 0$ for $n > N_{\mathrm{c}}$. This type of LC filter is known as the linearly constrained minimum variance (LCMV) filter:

$$\begin{aligned}\mathbf{W}_{\mathrm{LCMV}}(\Omega) = \arg\min_{\mathbf{W}} \operatorname{trace}\left(\mathbf{W}(\Omega)\mathbf{R}_{\mathrm{noise}}(\Omega)\mathbf{W}^H(\Omega)\right) \\ \text{s.t.} \quad \mathbf{W}(\Omega)\mathbf{A}_n(\Omega) = \mathbf{G}_n(\Omega)\mathbf{A}_n(\Omega) \quad \text{for } n = 1 \ldots, N_{\mathrm{c}}.\end{aligned} \tag{3.57}$$

If a solution exists, it is given by

$$\begin{aligned}\mathbf{W}_{\mathrm{LCMV}}(\Omega) = \Big[\mathbf{G}_1(\Omega)\mathbf{A}_1(\Omega) \cdots \mathbf{G}_{N_{\mathrm{c}}}(\Omega)\mathbf{A}_{N_{\mathrm{c}}}(\Omega)\Big] \\ \cdot \left(\mathbf{A}_{\mathrm{c}}^H(\Omega)\mathbf{R}_{\mathrm{noise}}^{-1}(\Omega)\mathbf{A}_{\mathrm{c}}\right)^{-1} \mathbf{A}_{\mathrm{c}}^H(\Omega)\mathbf{R}_{\mathrm{noise}}^{-1}(\Omega),\end{aligned} \tag{3.58}$$

where $\mathbf{A}_{\mathrm{c}}(\Omega) = [\mathbf{A}_1(\Omega) \cdots \mathbf{A}_{N_{\mathrm{c}}}(\Omega)]$ and $\mathbf{R}_{\mathrm{noise}}(\Omega) = \sum_{n=N_{\mathrm{c}}+1}^{N} \mathbf{R}_{\mathbf{c}_n}(\Omega)$.

A special case for $N_c = 1$ is the MVDR beamformer:

$$\mathbf{W}_{\text{MVDR}}(\Omega) = \mathbf{G}_1(\Omega)\mathbf{A}_1(\Omega)\frac{\mathbf{A}_1^H(\Omega)\left(\sum_{n=2}^{N}\mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1}}{\mathbf{A}_1^H(\Omega)\left(\sum_{n=2}^{N}\mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1}\mathbf{A}_1(\Omega)}. \tag{3.59}$$

These linearly constrained filters do not depend on the temporal correlations of the constrained source channels and the signals in these channels need not be modeled as random processes. The filters do, however, make use of the statistics of the non-constrained channels.

Let us consider again the example of a rank-1 target source and a full-rank noise source with $\mathbf{G}_1(\Omega) = \mathbf{e}_1^T$ and $\mathbf{G}_2 = \mathbf{0}$. The MVDR beamformer that minimizes noise power subject to a distortionless constraint on the target is

$$\mathbf{W}_{\text{MVDR}}(\Omega) = \mathbf{e}_1^T\mathbf{A}_1(\Omega)\frac{\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)}{\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}. \tag{3.60}$$

This filter applies unity gain to all signals parallel to $\mathbf{A}_1(\Omega)$. As a result, its squared error is larger than that of the MWF. Compare the filters defined by (3.54) and (3.60). The single-target MWF performs the same projection operation as the MVDR, but also scales the signal amplitude. In fact, the two beamformers are parallel, and the single-target MWF can be written as an MVDR beamformer followed by a scalar Wiener filter:

$$\mathbf{W}_{\text{MWF}}(\Omega) = \frac{R_{s_n}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}{1 + R_{s_n}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}\mathbf{W}_{\text{MVDR}}(\Omega). \tag{3.61}$$

3.5.3 Distortion weights

Clearly, there is a tradeoff between the spectral distortion affecting a target source and the overall squared error of the output. The MVDR and MWF beamformers represent two extremes. There is a more general class of weighted least-squares filters that occupy the space in between. The *speech-distortion-weighted multichannel*

Wiener filter (SDW-MWF) [68] solves the single-target optimization problem

$$\mathbf{w}_{\text{SDWMWF}} = \arg\min_{\mathbf{w}} \lambda_1 \mathbb{E}\left[\left|\hat{\mathbf{d}}_1(t) - \mathbf{d}_1(t)\right|^2\right] + \mathbb{E}\left[\left|\hat{\mathbf{d}}_2(t)\right|^2\right], \tag{3.62}$$

where $\lambda_1 \geq 0$ is a *distortion weight* that trades off between target distortion and noise reduction. Note that the notation here differs from that in the original formulation [68], where the distortion weight was applied to the noise term rather than the distortion term of the cost function.

This optimization problem can solved in the frequency domain using the PSDs of the source images:

$$\mathbf{W}_{\text{SDWMWF}}(\Omega) = \arg\min_{\mathbf{W}} \lambda_1 (\mathbf{W} - \mathbf{G}_1(\Omega))\, \mathbf{R}_{\mathbf{c}_1}(\Omega)\, (\mathbf{W} - \mathbf{G}_1(\Omega))^H + \mathbf{W}\mathbf{R}_{\mathbf{c}_2}(\Omega)\mathbf{W}^H \tag{3.63}$$

$$= \lambda_1 \mathbf{G}_1(\Omega)\mathbf{R}_{\mathbf{c}_1}(\Omega)\left(\lambda_1 \mathbf{R}_{\mathbf{c}_1}(\Omega) + \mathbf{R}_{\mathbf{c}_2}(\Omega)\right)^{-1}. \tag{3.64}$$

For a rank-1 target source and full-rank noise source, the filter can be written

$$\mathbf{W}_{\text{SDWMWF}}(\Omega) = \mathbf{G}_1(\Omega)\mathbf{A}_1(\Omega)\frac{\lambda_1 R_{s_n}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)}{1 + \lambda_1 R_{s_n}(\Omega)\mathbf{A}_1^H(\Omega)\mathbf{R}_{\mathbf{c}_2}^{-1}(\Omega)\mathbf{A}_1(\Omega)}. \tag{3.65}$$

If $\lambda_1 = 1$, then (3.65) reduces to the MWF (3.54). If $R_{s_n}(\Omega) > 0$, then in the limit as $\lambda_1 \to \infty$, we have the MVDR beamformer (3.60).

The MSDW-MWF

This dissertation adopts a more general version of the SDW-MWF, the multiple speech-distortion-weighted multichannel Wiener filter (MSDW-MWF) [134]:

$$\mathbf{w}_{\text{MSDW-MWF}} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \lambda_n \mathbb{E}\left[\left|\hat{\mathbf{d}}_n(t) - \mathbf{d}_n(t)\right|^2\right]. \tag{3.66}$$

In the frequency domain, the optimization criterion is

$$\mathbf{W}_{\text{MSDW-MWF}}(\Omega) = \arg\min_{\mathbf{W}} \sum_{n=1}^{N} \lambda_n \left(\mathbf{W} - \mathbf{G}_n(\Omega)\right) \mathbf{R}_{\mathbf{c}_n}(\Omega) \left(\mathbf{W} - \mathbf{G}_n(\Omega)\right)^H \tag{3.67}$$

and the solution is

$$\mathbf{W}_{\text{MSDW-MWF}}(\Omega) = \left(\sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega) \mathbf{R}_{\mathbf{c}_n}(\Omega)\right) \left(\sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1}. \tag{3.68}$$

The distortion weights $\lambda_1, \ldots, \lambda_N$ scale the power spectral densities of their respective source channels, causing the filter to prioritize those channels with larger weights. If $\lambda_n = 1$ for all n, we have the ordinary MWF (3.54).

Large-distortion-weight limit

In the limit as $\lambda_n \to \infty$ for some or all rank-1 sources, we have an LC filter with linear constraints applied to those sources. For example, to obtain the LCMV beamformer (3.58), let $\mathbf{R}_{\mathbf{c}_n}(\Omega) = R_{s_n}(\Omega)\mathbf{A}_n(\Omega)\mathbf{A}_n^H(\Omega)$ for $n = 1, \ldots, N_\text{c}$, let $\lambda_1 = \cdots = \lambda_{N_\text{c}} = \lambda_\text{c}$ and $\lambda_{N_\text{c}+1}, \ldots, \lambda_N = 1$, let $\mathbf{A}_\text{c} = [\mathbf{A}_1 \cdots \mathbf{A}_{N_\text{c}}]$, let $\mathbf{R}_\text{c}(\Omega) = \text{diag}(R_{s_1}(\Omega), \ldots, R_{s_{N_\text{c}}}(\Omega))$, let $\mathbf{G}_n(\Omega) = \mathbf{0}$ for $n = N_\text{c}+1, \ldots, N$, and let $\bar{\mathbf{R}}_\text{noise}(\Omega) = \sum_{n=N_\text{c}+1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega)$. Omitting the frequency variable Ω for brevity, the MSDW-MWF (3.68) can be written as

$$\mathbf{W} = [\mathbf{G}_1\mathbf{A}_1 \cdots \mathbf{G}_{N_\text{c}}\mathbf{A}_{N_\text{c}}] \, \lambda_\text{c}\mathbf{R}_\text{c}\mathbf{A}_\text{c}^H \left(\lambda_\text{c}\mathbf{A}_\text{c}\mathbf{R}_\text{c}\mathbf{A}_\text{c}^H + \bar{\mathbf{R}}_\text{noise}\right)^{-1}. \tag{3.69}$$

Applying the Woodbury identity and combining terms,

$$\begin{aligned}
\mathbf{W} &= [\mathbf{G}_1\mathbf{A}_1 \cdots \mathbf{G}_{N_\mathrm{c}}\mathbf{A}_{N_\mathrm{c}}]\, \lambda_\mathrm{c}\mathbf{R}_\mathrm{c}\mathbf{A}_\mathrm{c}^H \\
&\quad \cdot \left(\bar{\mathbf{R}}_\mathrm{noise}^{-1} - \bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\left(\lambda_\mathrm{c}^{-1}\mathbf{R}_\mathrm{c}^{-1} + \mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\right)^{-1}\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\right) && (3.70) \\
&= [\mathbf{G}_1\mathbf{A}_1 \cdots \mathbf{G}_{N_\mathrm{c}}\mathbf{A}_{N_\mathrm{c}}] \\
&\quad \cdot \left(\lambda_\mathrm{c}\mathbf{R}_\mathrm{c} - \lambda_\mathrm{c}\mathbf{R}_\mathrm{c}\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\left(\lambda_\mathrm{c}^{-1}\mathbf{R}_\mathrm{c}^{-1} + \mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\right)^{-1}\right)\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1} && (3.71) \\
&= [\mathbf{G}_1\mathbf{A}_1 \cdots \mathbf{G}_{N_\mathrm{c}}\mathbf{A}_{N_\mathrm{c}}]\left(\lambda_\mathrm{c}^{-1}\mathbf{R}_\mathrm{c}^{-1} + \mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\right)^{-1}\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}. && (3.72)
\end{aligned}$$

In the limit for large distortion weights on source channels 1 through N_c, we have

$$\begin{aligned}
\lim_{\lambda_\mathrm{c}\to\infty}\mathbf{W}_\mathrm{MSDW-MWF} &= [\mathbf{G}_1\mathbf{A}_1 \cdots \mathbf{G}_{N_\mathrm{c}}\mathbf{A}_{N_\mathrm{c}}]\left(\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1}\mathbf{A}_\mathrm{c}\right)^{-1}\mathbf{A}_\mathrm{c}^H\bar{\mathbf{R}}_\mathrm{noise}^{-1} && (3.73) \\
&= \mathbf{W}_\mathrm{LCMV}. && (3.74)
\end{aligned}$$

This limiting argument also applies to overdetermined mixtures with $N \leq M$ rank-1 channels and negligible full-rank noise. Although the mixture PSD $\mathbf{R}_\mathbf{x}$ is not invertible in such a system, its column space includes the ATFs of interest and so we can still analyze the MWF or MSDW-MWF by interpreting them as linearly constrained unmixing filters.

Because it is such a general LTI filter—many commonly used filters are special cases—the MSDW-MWF will form the basis of the proposed source-remixing system, which is the subject of the next chapter.

Chapter 4

Binaural Audio Source Remixing

Because the spatial processing power of microphone arrays can far exceed that of human listeners, engineers have long tried to use microphone arrays to help humans hear better in noisy environments [17]. Modern high-end hearing aids typically include two microphones in each earpiece, which can be used to perform directional beamforming. Because these microphones are so closely spaced, however, they can provide only a few decibels of noise reduction. Some new hearing aids can share data between earpieces, allowing them to leverage a total of four microphones on both sides of the head. Even these devices, however, provide modest improvements at best. The larger arrays used in distant audio capture and machine listening applications could achieve higher spatial resolution than a pair of earpieces. However, there are important differences between human and machine listening applications and they require different array processing strategies.

In most array processing applications, from antenna arrays to smart speakers, the goal is to steer a beam toward a single target, removing as much interference from other sources as possible. Many microphone array listening devices reported in the engineering literature have been designed to do the same thing: isolate a single target source, often a talker, and suppress all others. Human listeners are not the same as wireless modems or automatic speech recognition algorithms, however, and this approach has several drawbacks:

- Humans do not always want to listen to only one sound source. For example, a user might want to listen to background music during a conversation or switch their attention between multiple talkers.

- The listening device usually does not know which sound a human wishes to hear. There have been recent efforts to select a target based on eye tracking or neural sensing, but even if they successfully identified the attended source, these methods would inhibit the listener's ability to quickly switch attention.
- Fully suppressing all background noise would be unnatural and disturbing—imagine watching someone's lips moving but hearing no sound come out! It could also be dangerous if the listener failed to hear an alarm, crash, or scream.
- Designing such a beamformer requires precise knowledge of the acoustic channel parameters, such as the source channel AIRs or correlation matrices. An aggressive background-suppressing beamformer would be highly sensitive to parameter mismatch, especially if the target and noise sources are close to each other (Section 4.2.3).
- Any background noise that is not removed by the beamformer will be spectrally distorted (Section 4.3.1). Removing only some spectral features of an interfering sound could be worse than not removing it at all because the unnatural distortion could disrupt the auditory system's source separation process [3].
- Any residual background sources will lose the interaural cues that allow listeners to localize them (Section 4.3.1). It would seem as though every sound in the room were coming from the same direction.

In this chapter, we will explore a different approach to array processing for listening devices: *source remixing*. Instead of isolating one source and removing all others, the system applies different filters to different sources, as illustrated in Figure 4.1. Some sources might be amplified and others attenuated. A device could also apply different spectral filters to speech and music, for example.

The choice of how many source channels to preserve and what processing to apply to them are open questions, ones better addressed by clinical researchers than by engineers. It is known that there is a tradeoff between intelligibility, or how well a listener can understand the single source to which they are attending, and immersiveness, or how aware the listener is of the overall acoustic scene [92]. Different listeners

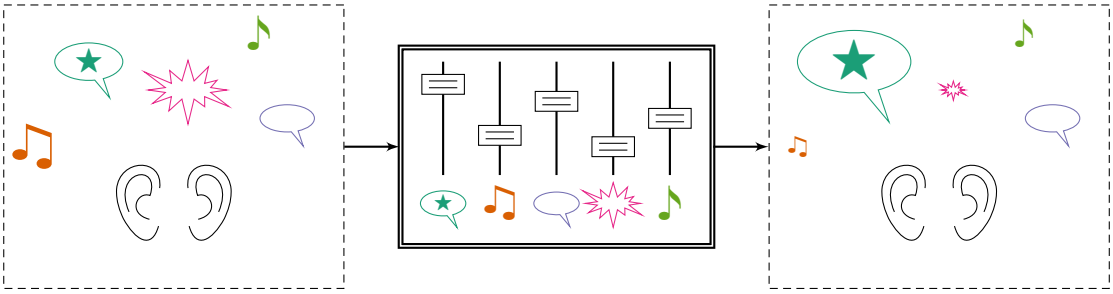

Figure 4.1: A remixing space-time filter applies different processing to each source channel, much like a mixing engineering processes recording tracks.

will likely choose different levels of intelligibility versus immersiveness depending on their individual hearing and cognitive abilities and personal preferences. The processing applied to a source will also depend on what kind of source it is: stationary noise sources, such as air conditioners, can probably be removed completely, while a distant human talker shouting the listener's name should presumably be amplified. Learning-based acoustic event detection and classification algorithms will surely help to identify and score the importance of different sources.

While we await clinical research to understand the potential benefits of remixing for listeners, we can characterize the *engineering* benefits of remixing for filter design. For system designers, source remixing is advantageous over complete separation because it eases the design constraints on the space-time filters. A filter that is only asked to remove part of a source has an easier job than one that fully suppresses it. This chapter describes these quantifiable advantages of remixing for linear time-invariant filters. Building on the extensive literature on spatial and spectral distortion in single-target binaural filters, this work explicitly analyzes the relationship between desired responses and filter distortion for source-remixing filters. In Section 4.2, it will be shown that remixing has advantages for squared error, spectral distortion, and sensitivity. Remixing filters are also essential if we hope to preserve the interaural cues of background sources and therefore the user's spatial awareness, as described in Section 4.3.

4.1 A Source-Remixing Filter

A source-remixing space-time filter attempts to alter the observed mixture to apply a different two-output LTI filter in every source channel. For input source image $\mathbf{c}_n(t)$, the desired output source image is

$$\mathbf{d}_n(t) = \int_{-\infty}^{\infty} \mathbf{g}_n(v)\mathbf{c}_n(t-v)\,\mathrm{d}v \tag{4.1}$$

for $n = 1, \ldots, N$, where $\mathbf{g}_n(v)$ is a $2 \times M$ matrix of impulse responses. The $J = 2$ output channels correspond to the left and right ears. By convention, output channel 1 is to the left ear and output channel 2 is to the right. If $\mathbf{c}_n(t)$ has a Fourier transform, then in the frequency domain

$$\mathbf{D}_n(\Omega) = \mathbf{G}_n(\Omega)\mathbf{C}_n(\Omega), \quad n = 1, \ldots, N, \tag{4.2}$$

and the ideal output of the listening device is

$$\mathbf{y}(t) = \sum_{n=1}^{N} \mathbf{d}_n(t) \tag{4.3}$$

$$\mathbf{Y}(\Omega) = \sum_{n=1}^{N} \mathbf{D}_n(\Omega) \tag{4.4}$$

$$= \sum_{n=1}^{N} \mathbf{G}_n(\Omega)\mathbf{C}_n(\Omega). \tag{4.5}$$

While the desired responses $\mathbf{G}_1, \ldots, \mathbf{G}_N$ can be any $2 \times M$ matrices of frequency responses, a perceptually transparent listening device should apply the same processing to the source images at the left and right ears. If microphone 1 is in the left

ear and microphone 2 is in the right ear then the response matrices take the form

$$\mathbf{G}_n(\Omega) = \begin{bmatrix} G_n(\Omega) & 0 & 0 & \cdots & 0 \\ 0 & G_n(\Omega) & 0 & \cdots & 0 \end{bmatrix} \tag{4.6}$$

$$= G_n(\Omega) \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix}, \tag{4.7}$$

for $n = 1, \ldots, N$, where $G_n(\Omega)$ is a scalar desired response for source channel n. Processing the sounds at the two ears identically ensures that they retain their original interaural cues, as discussed in Section 4.3.

4.1.1 Weighted squared error cost function

Recall from the previous chapter that the filter output is given by

$$\hat{\mathbf{y}}(t) = \int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{x}(t-v)\,\mathrm{d}v \tag{4.8}$$

$$= \sum_{n=1}^{N} \underbrace{\int_{-\infty}^{\infty} \mathbf{w}(v)\mathbf{c}_n(t-v)\,\mathrm{d}v}_{\hat{\mathbf{d}}_n(t)}. \tag{4.9}$$

To design a filter that estimates this desired output, assume that the source channels are wide-sense stationary random processes that are mutually uncorrelated. That is, $\mathbb{E}\left[\mathbf{c}_{n_1}(t_1)\mathbf{c}_{n_2}(t_2)\right] = 0$ for all $n_1 \neq n_2$ and all t_1, t_2. We will apply the MSDW-MWF [134], which minimizes the weighted squared-error cost function

$$\mathcal{J}_{\mathrm{MSDW-MWF}} = \mathbb{E}\left[\sum_{n=1}^{N} \lambda_n \left|\hat{\mathbf{d}}_n(t) - \mathbf{d}_n(t)\right|^2\right], \tag{4.10}$$

where λ_n is a distortion weight that controls the relative importance of source channel n. If $\lambda_n = 0$, source n is ignored completely. In the limit as $\lambda_n \to \infty$, the filter has a linear constraint on source n. If $\lambda_n = 1$ for all n, we have the standard MWF.

The optimization problem (4.10) has different solutions depending on the constraints on $\mathbf{w}$. If the space-time filter is allowed to be noncausal, then it can be readily derived and analyzed using linear algebra in the continuous-time frequency domain. Our mathematical analysis in this chapter will use this noncausal formulation (Section 4.1.2). In implementation, however, the filter must be causal with an imperceptibly short delay and must be realizable with finite computational resources. Thus, the experiments presented in this chapter will use a causal finite-impulse-response discrete-time filter (Section 4.1.3).

4.1.2 Remixing MSDW-MWF

If the filter is allowed to be noncausal, then the cost function (4.10) can be minimized by independently minimizing the diagonal entries of the weighted error PSD matrix,

$$\mathcal{J}(\Omega) = \sum_{n=1}^{N} \lambda_n \left(\mathbf{W}(\Omega) - \mathbf{G}_n(\Omega)\right) \mathbf{R}_{\mathbf{c}_n}(\Omega) \left(\mathbf{W}(\Omega) - \mathbf{G}_n(\Omega)\right)^H, \tag{4.11}$$

for all frequencies Ω of interest. The solution, which will henceforth be denoted as $\mathbf{W}$ with no subscript, is

$$\mathbf{W}(\Omega) = \left(\sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega) \mathbf{R}_{\mathbf{c}_n}(\Omega)\right) \left(\sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega)\right)^{-1} \tag{4.12}$$

$$= \sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega) \mathbf{R}_{\mathbf{c}_n}(\Omega) \bar{\mathbf{R}}_{\mathbf{x}}^{-1}(\Omega) \tag{4.13}$$

where $\bar{\mathbf{R}}_{\mathbf{x}}(\Omega) = \sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega)$ is the weighted observed signal PSD [134].

4.1.3 Discrete-time implementation

While the noncausal continuous-time filter is easy to analyze, it is not realizable, especially in a listening device that must have imperceptible latency. Let $\mathbf{w}_{\mathrm{d}}[\tau] \in$

$\mathbb{R}^{2\times M}$ for $\tau = 0, \ldots, L-1$ be a causal, finite-length, discrete-time filter such that the output is

$$\hat{\mathbf{y}}_{\mathrm{d}}[k] = \sum_{\tau=0}^{L-1} \mathbf{w}_{\mathrm{d}}[\tau]\mathbf{x}_{\mathrm{d}}[k-\tau]. \tag{4.14}$$

The filter is chosen to be an estimator of the delayed signal $\mathbf{y}_{\mathrm{d},\alpha}[k] = \mathbf{y}_{\mathrm{d}}[k-\alpha]$, where α is a time delay, in samples, that should be imperceptible to the listener. The tradeoffs involved in choosing α are the subject of Chapter 5.

Causal discrete-time MWF

The discrete-time equivalent of the Wiener-Hopf equations (3.45) for a finite-length filter is the system of *normal equations*

$$\mathbb{E}\left[\hat{\mathbf{y}}_{\mathrm{d}}[k]\mathbf{x}_{\mathrm{d}}^T[k-\ell]\right] = \mathbb{E}\left[\mathbf{y}_{\mathrm{d},\alpha}[k]\mathbf{x}_{\mathrm{d}}^T[k-\ell]\right], \quad \ell = 0, \ldots, L-1. \tag{4.15}$$

Using $\mathbf{w}_{\mathrm{d}}$ to compute $\hat{\mathbf{y}}_{\mathrm{d}}$ and $\mathbf{g}_{\mathrm{d}}$ to compute $\mathbf{y}_{\mathrm{d}}$, we have

$$\sum_{\tau=0}^{L-1} \mathbf{w}_{\mathrm{d}}[\tau]\mathbb{E}\left[\mathbf{x}_{\mathrm{d}}[k-\tau]\mathbf{x}_{\mathrm{d}}^T[k-\ell]\right] = \sum_{\tau=-\infty}^{\infty} \mathbf{g}_{\mathrm{d},n}[\tau-\alpha]\mathbb{E}\left[\mathbf{c}_{\mathrm{d},n}[k-\tau]\mathbf{c}_{\mathrm{d},n}^T[k-\ell]\right], \tag{4.16}$$

for $\ell = 0, \ldots, L-1$. Assuming for simplicity of notation that $\mathbf{g}_{\mathrm{d},n}$ has length less than or equal to $L-\alpha$, the system can be expressed as a matrix equation:

$$\begin{bmatrix}\mathbf{w}_{\mathrm{d}}[0] & \cdots & \mathbf{w}_{\mathrm{d}}[L-1]\end{bmatrix} \sum_{n=1}^{N} \underline{\mathbf{r}}_{\mathbf{c}_n} = \sum_{n=1}^{N} \begin{bmatrix}\mathbf{g}_{\mathrm{d},n}[-\alpha] & \cdots & \mathbf{g}_{\mathrm{d},n}[L-1-\alpha]\end{bmatrix} \underline{\mathbf{r}}_{\mathbf{c}_n}, \tag{4.17}$$

where

$$\underline{\mathbf{r}}_{\mathbf{c}_n} = \begin{bmatrix} \mathbf{r}_{\mathbf{c}_n}[0] & \mathbf{r}_{\mathbf{c}_n}[1] & \cdots & \mathbf{r}_{\mathbf{c}_n}[L-1] \\ \mathbf{r}_{\mathbf{c}_n}[-1] & \mathbf{r}_{\mathbf{c}_n}[0] & & \vdots \\ \vdots & & \ddots & \mathbf{r}_{\mathbf{c}_n}[1] \\ \mathbf{r}_{\mathbf{c}_n}[1-L] & \cdots & \mathbf{r}_{\mathbf{c}_n}[-1] & \mathbf{r}_{\mathbf{c}_n}[0] \end{bmatrix}, \quad n = 1, \ldots, N. \tag{4.18}$$

The solution to (4.17) is the discrete-time causal MWF, which minimizes the overall mean square error between $\mathbf{y}_\mathrm{d}$ and $\hat{\mathbf{y}}_\mathrm{d}$ for a given filter length L.

Causal discrete-time MSDW-MWF

To find the discrete-time finite-length MSDW-MWF, let us scale the correlation matrices by the distortion weights $\lambda_1, \ldots, \lambda_N$:

$$\begin{bmatrix}\mathbf{w}_\mathrm{d}[0] & \cdots & \mathbf{w}_\mathrm{d}[L-1]\end{bmatrix} \sum_{n=1}^{N} \lambda_n \underline{\mathbf{r}}_{\mathbf{c}_n} = \sum_{n=1}^{N} \lambda_n \begin{bmatrix}\mathbf{g}_\mathrm{d}[-\alpha] & \cdots & \mathbf{g}_\mathrm{d}[L-1-\alpha]\end{bmatrix} \underline{\mathbf{r}}_{\mathbf{c}_n}. \tag{4.19}$$

Finally, the length-L discrete-time causal MSDW-MWF with delay α is given by

$$\begin{bmatrix}\mathbf{w}_\mathrm{d}[0] & \cdots & \mathbf{w}_\mathrm{d}[L-1]\end{bmatrix} = \sum_{n=1}^{N} \lambda_n \begin{bmatrix}\mathbf{g}_\mathrm{d}[-\alpha] & \cdots & \mathbf{g}_\mathrm{d}[L-1-\alpha]\end{bmatrix} \underline{\mathbf{r}}_{\mathbf{c}_n} \bar{\underline{\mathbf{r}}}_{\mathbf{x}}^{-1} \tag{4.20}$$

where $\bar{\underline{\mathbf{r}}}_{\mathbf{x}} = \sum_{n=1}^{N} \lambda_n \underline{\mathbf{r}}_{\mathbf{c}_n}$ is the weighted stacked correlation matrix.

Because the stacked correlation matrices have block Toeplitz structures, both the MWF and the MSDW-MWF can be efficiently computed using the block Levinson recursion. All experiments in this chapter are performed using the discrete-time causal MWF with a generous filter length spanning 256 ms and delay of 16 ms. The impact of delay on filter performance and on listener perception is the subject of Chapter 5.

4.2 Performance and Desired Responses

The performance of a source-remixing filter depends on the choice of desired source channel filters $\mathbf{g}_1(t), \ldots, \mathbf{g}_N(t)$. Consider the limiting case in which $\mathbf{g}_1 = \mathbf{g}_2 = \cdots = \mathbf{g}_N$, that is, the same processing is to be applied to every source channel. Then we can achieve error-free performance by choosing $\mathbf{w} = \mathbf{g_1}$. It would also have zero additional delay ($\alpha = 0$), no spectral distortion, and no interaural cue distortion.

This transparent filter would not depend on the source statistics, so it would also have no sensitivity whatsoever to parameter mismatch. At the other extreme, we will show that a single-target beamformer is quite sensitive to mismatch, can have large delay, and destroys the interaural cues of background sources.

A source-remixing filter should fall somewhere in between these two extremes. In this section, we examine the impact of the desired responses on several measures of filter performance. More transparent filters, that is, those that alter the input signals less, are more robust and more easily implemented than filters that perform more aggressive processing.

4.2.1 Spectral distortion

Let us begin by considering the distortion applied to a directional source. For the remainder of this chapter, all analysis is performed in the frequency domain and the frequency variable Ω is omitted for brevity. First, note that the remixing filter (4.13) can be written

$$\mathbf{W} = \sum_{m=1}^{N} \lambda_m \mathbf{G}_m \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \tag{4.21}$$

$$\mathbf{W} = \mathbf{G}_n - \mathbf{G}_n \left(\sum_{m=1}^{N} \lambda_m \mathbf{R}_{\mathbf{c}_m} \right) \bar{\mathbf{R}}_{\mathbf{x}}^{-1} + \sum_{m=1}^{N} \lambda_m \mathbf{G}_m \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \tag{4.22}$$

$$\mathbf{W} = \mathbf{G}_n - \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_n - \mathbf{G}_m \right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1}. \tag{4.23}$$

If source channel n has ATF $\mathbf{A}_n$, then the response of the filter to source signal n is

$$\mathbf{W}\mathbf{A}_n = \mathbf{G}_n \mathbf{A}_n - \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_n - \mathbf{G}_m \right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n. \tag{4.24}$$

The second term represents the deviation from the desired result, $\mathbf{G}_n \mathbf{A}_n$. It depends on the difference in desired responses between sources, the spectra and distortion weights of the other sources, and the spatial separation of source channel n from the

other channels. For example, if a source channel m has ATF $\mathbf{A}_m$ and source signal power R_{s_m}, then $\mathbf{R}_{\mathbf{c}_m} = R_{s_m}\mathbf{A}_m\mathbf{A}_m^H$ and the corresponding term in the summation in (4.24) becomes

$$\lambda_m R_{s_m}\mathbf{A}_m^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n(\mathbf{G}_n - \mathbf{G}_m)\mathbf{A}_m. \tag{4.25}$$

Distortion is most severe when there is another source close to the target (so that $\mathbf{A}_m^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n$ is large) with strong power (so that R_{s_m} is large) and strict distortion constraints (so that λ_m is large) but with a very different desired response (so that $\mathbf{G}_n - \mathbf{G}_m$ is large). If a source is easily separated from all others (so that $\mathbf{A}_m^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n$ is small), for example because the array is large and well-positioned relative to the sources, then it does not cause much distortion in the output. If all source channels have similar desired responses, then the filter will not introduce much distortion to any of them.

An important special case is a channel for which $\mathbf{G}_n = \mathbf{0}$. If a filter is designed to completely remove a source channel—as a conventional beamformer would be for all but one source—then the filter response for that channel is

$$\mathbf{W}\mathbf{A}_n = \sum_{m=1}^{N} \lambda_m \mathbf{G}_m \mathbf{R}_{\mathbf{c}_m}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n. \tag{4.26}$$

The residual power of the background signal will be concentrated at frequencies at which the background source is difficult to separate from the target source(s). For small arrays, this often means that high frequencies will be removed while low frequencies will remain, altering the perceived timbre of the background source.

The distortion applied to source channel n also depends on the distortion weight λ_n, which affects the $\bar{\mathbf{R}}_{\mathbf{x}}^{-1}$ term in (4.24). In particular, in the limit for large distortion weights we have

$$\lim_{\lambda_n \to \infty} \mathbf{W}\mathbf{A}_n = \mathbf{G}_n\mathbf{A}_n. \tag{4.27}$$

However, increasing the distortion weight of one source channel to reduce its distortion will increase the distortion on all other channels. We can observe this effect by analyzing the overall error of the system.

4.2.2 Squared error

Now consider the total error PSD at the output of the source-remixing filter:

$$\mathbf{R}_{\text{err}} = \sum_{n=1}^{N} (\mathbf{G}_n - \mathbf{W})\, \mathbf{R}_{\mathbf{c}_n} (\mathbf{G}_n - \mathbf{W})^H . \tag{4.28}$$

This expression is not easily simplified because the MSDW-MWF is not a minimum-mean-square-error estimator. However, the *weighted* squared-error cost function $\mathcal{J}$ (4.11) can be expressed in terms of pairs of sources. Using (4.23) and the fact that the weighted error is orthogonal to the estimate, we have

$$\mathcal{J} = \sum_{n=1}^{N} \lambda_n (\mathbf{G}_n - \mathbf{W})\, \mathbf{R}_{\mathbf{c}_n} (\mathbf{G}_n - \mathbf{W})^H \tag{4.29}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m (\mathbf{G}_n - \mathbf{G}_m)\, \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} (\mathbf{G}_n - \mathbf{W})^H \tag{4.30}$$

$$= \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m (\mathbf{G}_n - \mathbf{G}_m)\, \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{G}_n^H . \tag{4.31}$$

Binaural listening device

If the desired responses $\mathbf{G}_1, \ldots, \mathbf{G}_N$ are real-valued, then

$$\mathcal{J} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left[(\mathbf{G}_n - \mathbf{G}_m)\, \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{G}_n^T + (\mathbf{G}_m - \mathbf{G}_n)\, \mathbf{R}_{\mathbf{c}_n} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_m} \mathbf{G}_m^T \right] \tag{4.32}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m (\mathbf{G}_n - \mathbf{G}_m)\, \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} (\mathbf{G}_n - \mathbf{G}_m)^T . \tag{4.33}$$

In a perceptually transparent listening device with $\mathbf{G}_n = G_n \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix}^T$ for $n = 1, \ldots, N$, the weighted squared errors at the left and right ears are

$$\mathcal{J}^{\text{left}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left| G_n - G_m \right|^2 \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_1 \tag{4.34}$$

$$\mathcal{J}^{\text{right}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \lambda_n \lambda_m \left| G_n - G_m \right|^2 \mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_2. \tag{4.35}$$

The error can be decomposed in terms of pairs of source channels. The error contribution of each source pair depends on the difference in desired responses between channels and on the spatial similarity between the two sources. Source pairs that are easily separated or that have similar processing applied to them contribute little error. Distortion weights can be used to shift error between sources.

Multichannel Wiener filter

The lowest overall squared error is achieved with $\lambda_1 = \cdots = \lambda_N = 1$, which yields the MWF. In that case, the error PSD is

$$\mathbf{R}_{\text{err}} = \sum_{n=1}^{N} \sum_{m=1}^{M} \left(\mathbf{G}_n - \mathbf{G}_m \right) \mathbf{R}_{\mathbf{c}_m} \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{G}_n^H. \tag{4.36}$$

For real-valued scalar desired responses,

$$R_{\text{err}}^{\text{left}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \left| G_n - G_m \right|^2 \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_1 \tag{4.37}$$

$$R_{\text{err}}^{\text{right}} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} \left| G_n - G_m \right|^2 \mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{R}_{\mathbf{c}_n} \mathbf{e}_2. \tag{4.38}$$

4.2.3 Sensitivity to parameter mismatch

So far, we have assumed perfect knowledge of the channel parameters $\mathbf{R}_{\mathbf{c}_n}$ or $\mathbf{A}_n$ for $n = 1, \ldots, N$. In real systems, however, these parameters must be estimated from observed data, for example using the methods of Chapter 8. All parameter estimation methods are prone to error. A good filter to should be robust against errors in parameter estimates.

For directional beamformers, which can be evaluated using SNR, a popular measure of sensitivity to parameter mismatch is the norm of the beamformer vector, $\mathbf{W}\mathbf{W}^H$ where $\mathbf{W}$ is a row vector. To see why, consider the output SNR gain of an MVDR beamformer:

$$\mathrm{SNR}^{\mathrm{out}} = \frac{\mathbf{W}\mathbf{R}_{\mathbf{c}_1}\mathbf{W}^H}{\mathbf{W}\mathbf{R}_{\mathbf{c}_2}\mathbf{W}^H}. \tag{4.39}$$

In [135], sensitivity was defined as the fractional change in SNR due to a small random change in the target source statistics:

$$\begin{aligned}
\text{Sensitivity} &= \frac{\frac{\partial}{\partial\epsilon}\mathrm{SNR}^{\mathrm{out}}|_{\epsilon=0}}{\mathrm{SNR}^{\mathrm{out}}} && (4.40)\\
&= \frac{\frac{\partial}{\partial\epsilon}\mathbf{W}(\mathbf{R}_{\mathbf{c}_1} + \epsilon R_{s_1}\mathbf{I})\mathbf{W}^H}{\mathbf{W}\mathbf{R}_{\mathbf{c}_1}\mathbf{W}^H} && (4.41)\\
&= \frac{R_{s_1}\mathbf{W}\mathbf{W}^H}{\mathbf{W}\mathbf{R}_{\mathbf{c}_1}\mathbf{W}^H} && (4.42)\\
&= \mathbf{W}\mathbf{W}^H. && (4.43)
\end{aligned}$$

The final step follows from the distortionless property of MVDR beamformers. Unfortunately, this popular definition does not extend to source-remixing beamformers, which generally do not have unity-gain constraints and are not evaluated using SNR.

Let us instead define sensitivity as the change in weighted squared error due to a small random change in the statistics of each source.

Definition 4.1. The *sensitivity* of a remixing filter to an offset in the statistics of

source channel n is given by

$$\text{Sensitivity}_n = \frac{\partial}{\partial \epsilon} \mathcal{J}|_{\epsilon=0}, \tag{4.44}$$

where $\mathcal{J}$ is computed by fixing $\mathbf{W}$ and replacing $\mathbf{R}_{\mathbf{c}_n}$ with $\mathbf{R}_{\mathbf{c}_n} + \epsilon R_{s_n} \mathbf{I}$.

Using the error expression (4.11), we find that

$$\text{Sensitivity}_n = \frac{\partial}{\partial \epsilon} \lambda_n (\mathbf{G}_n - \mathbf{W})(\mathbf{R}_{\mathbf{c}_n} + \epsilon R_{s_1} \mathbf{I})(\mathbf{G}_n - \mathbf{W})^H \tag{4.45}$$

$$= \lambda_n R_{s_n} \left| \mathbf{G}_n - \mathbf{W} \right|^2, \quad n = 1, \ldots, N. \tag{4.46}$$

For the MSDW-MWF beamformer, we can apply (4.23) to find that

$$\text{Sensitivity}_n = \lambda_n R_{s_n} \left| \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_n - \mathbf{G}_m \right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \right|^2, \quad n = 1, \ldots, N. \tag{4.47}$$

The sensitivity to parameter errors also depends upon the pairwise differences between desired responses. Less aggressive remixing yields more robust filters.

4.2.4 Remixing experiments

The performance of the proposed source-remixing space-time filter was evaluated using a subset of the low-reverberation distributed array data set (Section 2.4). There were five speech sources placed around the Augmented Listening Laboratory in both the acoustically treated and untreated parts of the room. A sixth source channel contained spatially uncorrelated speech-shaped Gaussian noise about 20 dB below the level of the speech sources. The desired responses were frequency-invariant scalar gains:

$$[G_n(\Omega)]_{n=1}^6 = \left[1.0^u \quad 0.2^u \quad 0.4^u \quad 0.6^u \quad 0.8^u \quad 0.1^u \right], \tag{4.48}$$

where the exponent u was varied from 0 (fully transparent) to 1 (aggressive remixing) to show the effect of the desired responses on filter performance. The SER reported in the figures is averaged over the left and right outputs.

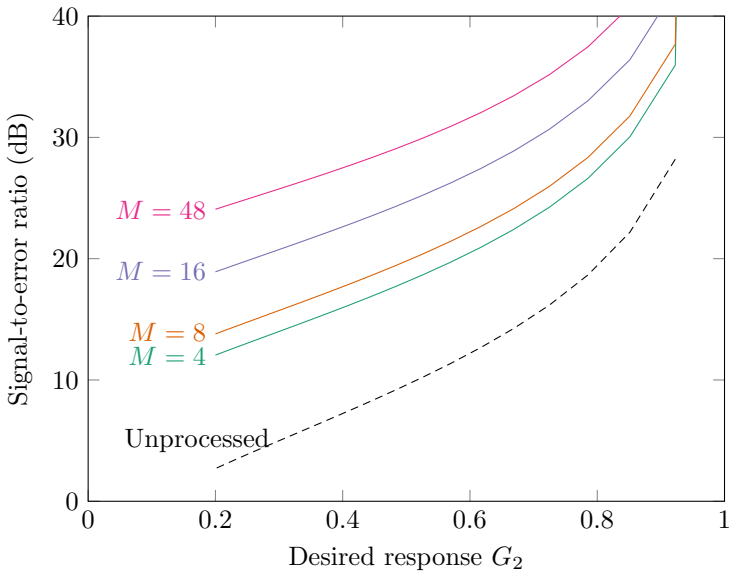

Figure 4.2: Remixing performance for different array configurations.

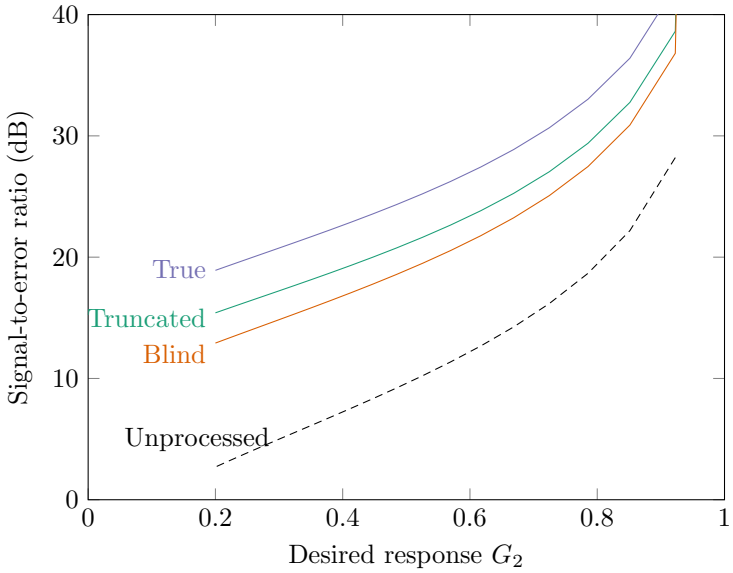

Figure 4.3: Remixing performance for different channel estimates.

Figure 4.2 shows the SER performance of the MWF as a function of $G_2 = 0.2^u$ for different array configurations. The smallest array ($M = 4$) contains two microphones in each earpiece. An 8-microphone array covers the head, including earpieces and glasses, while a 16-microphone array also includes the torso and arms. Finally, a distributed 48-microphone array includes 16-microphone arrays on three listeners in different parts of the room. The results show that larger microphone arrays outperform smaller microphone arrays by similar amounts for most combinations of desired responses.

Now let us consider the robustness of the source-remixing filters to errors in channel estimation. Figure 4.3 shows the performance of the 16-microphone array using three different models for the directional sources: the ground-truth autocorrelation matrices of the test data, measured acoustic impulse responses truncated to 64 ms, and source channel autocorrelation functions estimated using the cooperative blind source separation method described in [115] and Chapter 10. Somewhat surprisingly, the SER curves are roughly parallel: parameter mismatch appears to impose a nearly constant penalty, in dB, on the SER for most sets of desired responses. Further mathematical analysis will be required to understand this result.

4.3 Interaural Cue Preservation

When designing remixing filters that preserve the listener's awareness of multiple sound sources, we must take care to preserve the spatial characteristics of those signals [23, 24]. Humans use interaural cues, especially *interaural phase differences* (IPDs) and *interaural level differences* (ILDs), to localize sounds. As shown in Figure 4.4, sounds arriving from the left will reach the left ear before they reach the right ear. At low frequencies, this time difference produces a frequency-dependent phase shift between the left and right ears. At higher frequencies, sounds from the left will also be more intense in the left ear than in the right because they are attenuated by the head. The opposite is true for sounds arriving from the right.

In addition to the IPDs and ILDs that encode left/right position, the direction-

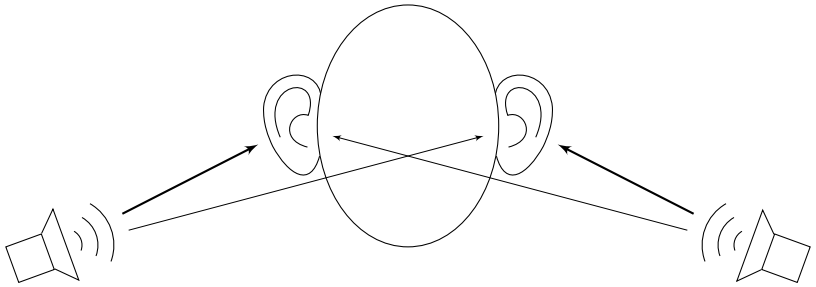

Figure 4.4: Humans localize sounds using interaural time and level differences.

dependent spectral shaping effects of the head, shoulders, and pinna, known as head-related transfer functions (HRTF), help humans to distinguish directions of arrival in three dimensions [117].

Interaural cues are not only useful for localization: they also help the auditory system to assign sound events in a mixture to the correct sources and to separate competing sounds from different directions. This benefit is known as *spatial release from masking* and is a crucial component of the cocktail party phenomenon [2, 3]. The farther an interference source is from the sound source of interest, the less it affects the intelligibility of the target.

Because humans can use interaural cues to naturally separate sound sources, it is important that listening devices preserve these cues. Otherwise, our attempts to improve intelligibility in noise might actually make it worse. Spatial filters do not automatically preserve interaural cues, and the beamformers used in teleconferencing and speech recognition applications typically do not account for spatial cue distortion. In these systems, the array processing algorithm needs to know where the sound sources are, but spatial information is irrelevant for later processing. In a listening application, sound sources are spatially processed twice: once by the listening device and again by the human listener. Therefore, they should retain their interaural cues at the output of the listening device.

In this section, we will show why conventional single-target beamformers distort spatial cues and describe several previously proposed spatial filters that better preserve a listener's spatial awareness. We will analyze the effects of the source-remixing

space-time filter on interaural cues and demonstrate its performance using real-world audio recordings from a wearable microphone array.

4.3.1 Binaural beamformers

Mathematically, the ILD and IPD can be defined in terms of the interaural transfer function:

Definition 4.2. The input and output *interaural transfer functions* (ITF) for source channel n are given by

$$\mathrm{ITF}_n^{\mathrm{in}} = \frac{\mathbf{e}_2^T \mathbf{C}_n}{\mathbf{e}_1^T \mathbf{C}_n} \tag{4.49}$$

$$\mathrm{ITF}_n^{\mathrm{out}} = \frac{\mathbf{e}_2^T \hat{\mathbf{D}}_n}{\mathbf{e}_1^T \hat{\mathbf{D}}_n}. \tag{4.50}$$

For full-rank source channels, the direction of the vector $\mathbf{C}_n$ is not fixed and the ITF is therefore signal-dependent. The ITF is more meaningful for source channels that are well characterized by a single ATF $\mathbf{A}_n$. Then we have

$$\mathrm{ITF}_n^{\mathrm{in}} = \frac{\mathbf{e}_2^T \mathbf{A}_n}{\mathbf{e}_1^T \mathbf{A}_n} \tag{4.51}$$

$$\mathrm{ITF}_n^{\mathrm{out}} = \frac{\mathbf{e}_2^T \mathbf{W}\mathbf{A}_n}{\mathbf{e}_1^T \mathbf{W}\mathbf{A}_n}. \tag{4.52}$$

The ILD and IPD can both be derived directly from the logarithm of the ITF.

Definition 4.3. The *interaural level difference* (ILD), in decibels, is the magnitude of the ITF on a decibel scale:

$$\mathrm{ILD}_n = 20 \log_{10} |\mathrm{ITF}_n| \tag{4.53}$$

$$= \frac{20}{\ln 10} \mathrm{Real} \left[\ln \mathrm{ITF}_n \right]. \tag{4.54}$$

Definition 4.4. The *interaural phase difference* (IPD), in radians, is the angle of the ITF:

$$\begin{aligned} \mathrm{IPD}_n &= \angle \mathrm{ITF}_n && (4.55) \\ &= \mathrm{Imag}\left[\ln \mathrm{ITF}_n\right]. && (4.56) \end{aligned}$$

Diotic space-time filter

Let us begin our analysis with a diotic space-time filter that presents the same processed signal to both ears. That is $\mathbf{e}_2^T\mathbf{W} = \mathbf{e}_1^T\mathbf{W}$. The output ITF is

$$\begin{aligned} \mathrm{ITF}_n^{\mathrm{out}} &= \frac{\mathbf{e}_2^T\mathbf{W}\mathbf{C}_n}{\mathbf{e}_1^T\mathbf{W}\mathbf{C}_n} && (4.57) \\ &= 1 && (4.58) \end{aligned}$$

for all $n = 1, \ldots, N$. Thus, $\mathrm{ILD}_n^{\mathrm{out}} = 0$ and $\mathrm{IPD}_n^{\mathrm{out}} = 0$ for all source channels. Such a beamformer destroys all spatial cues of all sources. Each source would sound like it is coming from inside the listener's head.

Binaural single-target beamformer

To preserve the listener's spatial awareness, a binaural listening device should produce different outputs at the left and right ears. Early cue-preserving binaural beamformers were designed to isolate a single target source and preserve its spatial cues [71, 72].

Assume that the target is source channel 1 and all other channels are unwanted

noise. From (4.7), the desired responses should be

$$\mathbf{G}_1 = \begin{bmatrix} G_1 & 0 & 0 & \cdots & 0 \\ 0 & G_1 & 0 & \cdots & 0 \end{bmatrix} \tag{4.59}$$

$$= G_1 \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix}, \tag{4.60}$$

and $\mathbf{G}_2 = \cdots = \mathbf{G}_N = 0$. The single-target SDW-MWF with this desired response is

$$\mathbf{W} = \lambda_1 \mathbf{G}_1 \mathbf{R}_{\mathbf{c}_1} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \tag{4.61}$$

$$= \lambda_1 G_1 \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix} \mathbf{R}_{\mathbf{c}_1} \bar{\mathbf{R}}_{\mathbf{x}}^{-1}. \tag{4.62}$$

For a rank-1 target source with ATF $\mathbf{A}_1$, the output ITF is

$$\mathrm{ITF}_1^{\mathrm{out}} = \frac{\lambda_1 G_1 R_{s_1} \mathbf{e}_2^T \mathbf{A}_1 \mathbf{A}_1^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_1}{\lambda_1 G_1 R_{s_1} \mathbf{e}_2^T \mathbf{A}_1 \mathbf{A}_1^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_1} \tag{4.63}$$

$$= \frac{\mathbf{e}_2^T \mathbf{A}_1}{\mathbf{e}_1^T \mathbf{A}_1} \tag{4.64}$$

$$= \mathrm{ITF}_1^{\mathrm{in}}. \tag{4.65}$$

This beamformer perfectly preserves the interaural cues of the target source.

Now consider the response to any other source image $\mathbf{C}_n$ that is not fully removed by the left beamformer ($\mathbf{e}_1^T \mathbf{W} \mathbf{C}_n \neq 0$):

$$\mathrm{ITF}_n^{\mathrm{out}} = \frac{\lambda_1 G_1 R_{s_1} \mathbf{e}_2^T \mathbf{A}_1 \mathbf{A}_1^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n}{\lambda_1 G_1 R_{s_1} \mathbf{e}_1^T \mathbf{A}_1 \mathbf{A}_1^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{C}_n} \tag{4.66}$$

$$= \frac{\mathbf{e}_2^T \mathbf{A}_1}{\mathbf{e}_1^T \mathbf{A}_1} \tag{4.67}$$

$$= \mathrm{ITF}_1^{\mathrm{in}}. \tag{4.68}$$

Every signal that is not fully removed by the beamformer is perceived as coming

from the direction of the target source [72]. This spatial distortion is disturbing to listeners. In the author's experience, it sounds like listening in a tunnel.

Binaural filters with constraints on background sources

To address this tunnel effect, researchers proposed several filters that intentionally preserve background sources with their original interaural cues. The simplest of these, which has been called the binaural SDW-MWF-η filter [88, 136, 137], mixes the beamformer output with the unprocessed input:

$$\mathbf{W}_{\mathrm{SDW-MWF}-\eta} = (1-\eta)\mathbf{W}_{\mathrm{SDW-MWF}} + \eta \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix}. \tag{4.69}$$

This approach perfectly preserves the interaural cues of the target while improving the cues of the background sources, as we will show in the next section. It was shown in listening tests that the modified filter partially restored spatial release from masking that was damaged by a conventional beamformer [137].

Later proposals explicitly constrained the response of the filter to one or more background sources. For a binaural listening device with M total microphones, we can perfectly preserve the binaural cues of up to M spatially distinct rank-1 source channels by applying distortionless constraints to them, that is, by creating an LCMV beamformer [138]. For such a source channel, the output of the binaural LCMV beamformer is, by constraint,

$$\mathbf{W}_{\mathrm{LCMV}}\mathbf{A}_n = G_n \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix} \mathbf{A}_n, \quad n = 1, \ldots, n^*. \tag{4.70}$$

Thus, its output ITFs are identical to the input ITFs for those sources. In [139], the desired responses $G_1, \ldots, G_{n^*}$ of the constrained background sources were chosen to maximize overall SNR for the target source.

A drawback of the binaural LCMV beamformer for conventional hearing aids is that it requires more microphones than a single-target beamformer. Adding more mi-

crophones to each earpiece is costly. Furthermore, because those microphones would be so close to each other, the LCMV constraint matrix would be poorly conditioned and therefore highly sensitive to parameter mismatch and diffuse noise. To relax these constraints without adding more microphones, it was proposed in [73, 140] to reduce the number of constraints by constraining only the ITF itself and not the absolute response in each ear. This beamformer, known as the joint binaural LCMV (JBLCMV), solves the optimization problem

$$\mathbf{W}_{\text{JBLCMV}} = \arg\min_{\mathbf{W}} \mathbf{W}\mathbf{R}_{\text{noise}}\mathbf{W}^T \tag{4.71}$$

$$\text{s.t. } \frac{\mathbf{e}_2^T\mathbf{W}\mathbf{A}_n}{\mathbf{e}_1^T\mathbf{W}\mathbf{A}_n} = \frac{\mathbf{e}_2^T\mathbf{A}_n}{\mathbf{e}_1^T\mathbf{A}_n}, \quad n = 1, \ldots, n^*. \tag{4.72}$$

A drawback of the JBLCMV is that it allows arbitrary spectral distortion of each source as long as the same distortion is applied in each ear. For mixtures of many sources, the ITF constraints can be satisfied only by a transparent filter, that is, by $\mathbf{W} \propto \begin{bmatrix}\mathbf{e}_1 & \mathbf{e}_2\end{bmatrix}^T$. Such a filter would apply no spatial processing at all.

For sparse signals such as speech, it is likely that only one or two sound sources are audible at a given time and frequency. Thus, a device can preserve the interaural cues of more sources than microphones using a time-varying, nonlinear filter that constrains only the active source(s) for each time-frequency sample. A time-varying system proposed by the author [112] is described in Chapter 7. Another proposal used a time-frequency mask [141]. With these time-varying methods, care must be taken not to introduce temporal artifacts such as musical noise.

The SDW-MWF-η, binaural LCMV, JBLCMV, and other LTI space-time filters proposed in the literature all improve the listener's spatial awareness by intentionally preserving a portion of the background sources that a conventional beamformer would remove. It seems that the more transparent the listening system—that is, the more similar the desired responses of the source channels—the less the distortion of the sources' interaural cues. We can explicitly study this relationship by analyzing the proposed source-remixing filter.

4.3.2 Interaural cues for the source-remixing filter

Now let us analyze the interaural cue distortion of the proposed source-remixing MSDW-MWF. The spatial transparency of the filter depends on the spatial diversity of the array, the choice of desired responses for the different source channels, and the distortion weights.

Suppose that a source channel n can be well characterized by an ATF $\mathbf{A}_n$. Then, substituting (4.23) into (4.52), the interaural cues at the output are

$$\mathrm{ITF}_n^{\mathrm{out}} = \frac{\mathbf{e}_2^T \mathbf{W}\mathbf{A}_n}{\mathbf{e}_1^T \mathbf{W}\mathbf{A}_n} \tag{4.73}$$

$$= \frac{\mathbf{e}_2^T \mathbf{G}_n \mathbf{A}_n + \mathbf{e}_2^T \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_m - \mathbf{G}_n\right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{\mathbf{e}_1^T \mathbf{G}_n \mathbf{A}_n + \mathbf{e}_1^T \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_m - \mathbf{G}_n\right) \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}. \tag{4.74}$$

If the desired response matrices satisfy (4.7), then

$$\mathrm{ITF}_n^{\mathrm{out}} = \frac{G_n \mathbf{e}_2^T \mathbf{A}_n + \sum_{m=1}^{N} \lambda_m \left(G_m - G_n\right) \mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{G_n \mathbf{e}_1^T \mathbf{A}_n + \sum_{m=1}^{N} \lambda_m \left(G_m - G_n\right) \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}. \tag{4.75}$$

For rank-1 sources, the matrix products in the summations simplify to

$$\mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n = R_{s_m} (\mathbf{A}_m^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n) \mathbf{e}_2^T \mathbf{A}_m \tag{4.76}$$

$$\mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n = R_{s_m} (\mathbf{A}_m^H \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n) \mathbf{e}_1^T \mathbf{A}_m. \tag{4.77}$$

The ratio of these products is the input ITF of source channel m. Thus, the output ITF can be thought of as a mixture of the ITFs of the individual source channels [88]. From (4.75), we can find several conditions under which the presence of source channel m does not affect the interaural cues of source channel n in the filter output:

1. λ_m or R_{s_m} is zero so that the filter ignores channel m,
2. $\mathbf{A}_n$ is parallel to $\mathbf{A}_m$ so that the ITFs are the same,
3. $\mathbf{A}_n$ is perfectly spatially separable from $\mathbf{R}_{\mathbf{c}_m}$ so that $\mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n = \mathbf{0}$,

4. $G_m = G_n$ so that the filter need not separate channels m and n, or

5. $\lambda_n \to \infty$ so that both the left and right output images are distortion-free.

The interaural cues of a target source are most impacted by powerful interference sources that are far enough from the source of interest that they have different interaural cues but near enough that they are difficult to separate from it. The penalty can be reduced by making the desired processing more similar between source channels or by increasing the distortion weight of the target. However, increasing the distortion weights for one source will affect the interaural cues for all other sources. Spatial distortion can also be improved by using a larger array that is better able to separate the source channels and process them independently.

First-order approximation

The errors in the ILD and IPD are the real and imaginary parts, respectively, of the logarithm of $\mathrm{ITF}_n^{\mathrm{out}}/\mathrm{ITF}_n^{\mathrm{in}}$. If G_n and $\mathrm{ITF}_n^{\mathrm{in}}$ are nonzero, then the ITF error can be written

$$\ln \frac{\mathrm{ITF}_n^{\mathrm{out}}}{\mathrm{ITF}_n^{\mathrm{in}}} = \ln \frac{1 + \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \frac{\mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{\mathbf{e}_2^T \mathbf{A}_n}}{1 + \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \frac{\mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{\mathbf{e}_1^T \mathbf{A}_n}}. \tag{4.78}$$

Using the first-order approximation $\ln(1+u) \approx u$ for both the numerator and denominator, the logarithmic ITF error is

$$\ln \frac{\mathrm{ITF}_n^{\mathrm{out}}}{\mathrm{ITF}_n^{\mathrm{in}}} \approx \sum_{m=1}^{N} \lambda_m \frac{G_m - G_n}{G_n} \left(\frac{\mathbf{e}_2^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{\mathbf{e}_2^T \mathbf{A}_n} - \frac{\mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_m} \bar{\mathbf{R}}_{\mathbf{x}}^{-1} \mathbf{A}_n}{\mathbf{e}_1^T \mathbf{A}_n} \right). \tag{4.79}$$

If every source channel were well approximated by a rank-1 PSD $\mathbf{R}_{\mathbf{c}_n} \approx R_{s_n}\mathbf{A}_n\mathbf{A}_n^H$ for $n = 1, \ldots, N$, then the ILD and IPD errors would be

$$\Delta\text{ILD}_n \approx \frac{20}{\ln 10}\sum_{m=1}^{N}\lambda_m R_{s_m}\text{Real}\left[\frac{G_m - G_n}{G_n}\mathbf{A}_m^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n\left(\frac{\mathbf{e}_2^T\mathbf{A}_m}{\mathbf{e}_2^T\mathbf{A}_n} - \frac{\mathbf{e}_1^T\mathbf{A}_m}{\mathbf{e}_1^T\mathbf{A}_n}\right)\right] \quad (4.80)$$

$$\Delta\text{IPD}_n \approx \sum_{m=1}^{N}\lambda_m R_{s_m}\text{Imag}\left[\frac{G_m - G_n}{G_n}\mathbf{A}_m^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n\left(\frac{\mathbf{e}_2^T\mathbf{A}_m}{\mathbf{e}_2^T\mathbf{A}_n} - \frac{\mathbf{e}_1^T\mathbf{A}_m}{\mathbf{e}_1^T\mathbf{A}_n}\right)\right]. \quad (4.81)$$

Thus, spatial distortion depends on the power and distortion weight of each interfering source, the relative difference in desired responses between sources, the spatial separability of the sources, and the difference in interaural cues between source channels.

Next, let us consider several special cases that illustrate the effect of the remixing filter on interaural cues.

Single target source

Suppose that $G_1 = 1$ for a directional source with ATF $\mathbf{A}_1$ and $G_2 = G_3 = \cdots = G_N = 0$, that is, that the filter is a beamformer directed at a particular source. Then for any source channel n, the output cues are

$$\text{ITF}_n^{\text{out}} = \frac{\lambda_1 R_{s_1}\mathbf{e}_2^T\mathbf{A}_1\mathbf{A}_1^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n}{\lambda_1 R_{s_1}\mathbf{e}_1^T\mathbf{A}_1\mathbf{A}_1^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n} \quad (4.82)$$

$$= \frac{\mathbf{e}_2^T\mathbf{A}_1}{\mathbf{e}_1^T\mathbf{A}_1} \quad (4.83)$$

$$= \text{ITF}_1^{\text{in}}. \quad (4.84)$$

This is the "tunnel effect" described in the previous section.

Identical desired responses

Now consider a system that only has two desired responses: G_1 for source channel 1 with ATF $\mathbf{A}_1$ and $G_2 = G_3 = \cdots = G_N$ for all other source channels. This is the SDW-MWF-η beamformer [88, 136, 137], which adds a fraction of the unprocessed input signal to the output of a single-target beamformer. For the target source, (4.75) becomes

$$\begin{aligned}
\mathrm{ITF}_1^{\mathrm{out}} &= \frac{G_1\mathbf{e}_2^T\mathbf{A}_1 + \sum_{m=2}^{N}\lambda_m\left(G_2 - G_1\right)\mathbf{e}_2^T\mathbf{R}_{\mathbf{c}_m}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n}{G_1\mathbf{e}_1^T\mathbf{A}_1 + \sum_{m=2}^{N}\lambda_m\left(G_2 - G_1\right)\mathbf{e}_1^T\mathbf{R}_{\mathbf{c}_m}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n} && (4.85)\\
&= \frac{G_1\mathbf{e}_2^T\mathbf{A}_1 + \left(G_2 - G_1\right)\mathbf{e}_2^T\left(\sum_{m=2}^{N}\lambda_m\mathbf{R}_{\mathbf{c}_m}\right)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n}{G_1\mathbf{e}_1^T\mathbf{A}_1 + \left(G_2 - G_1\right)\mathbf{e}_1^T\left(\sum_{m=2}^{N}\lambda_m\mathbf{R}_{\mathbf{c}_m}\right)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n} && (4.86)\\
&= \frac{G_1\mathbf{e}_2^T\mathbf{A}_1 + \left(G_2 - G_1\right)\mathbf{e}_2^T\left(\bar{\mathbf{R}}_{\mathbf{x}} - \lambda_1\mathbf{R}_{\mathbf{c}_1}\right)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n}{G_1\mathbf{e}_1^T\mathbf{A}_1 + \left(G_2 - G_1\right)\mathbf{e}_1^T\left(\bar{\mathbf{R}}_{\mathbf{x}} - \lambda_1\mathbf{R}_{\mathbf{c}_1}\right)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n} && (4.87)\\
&= \frac{\mathbf{e}_2^T\mathbf{A}_1}{\mathbf{e}_1^T\mathbf{A}_1} && (4.88)\\
&= \mathrm{ITF}_1^{\mathrm{in}}. && (4.89)
\end{aligned}$$

For an interference source with ATF $\mathbf{A}_n$, $n > 1$, we have

$$\begin{aligned}
\mathrm{ITF}_n^{\mathrm{out}} &= \frac{G_2\mathbf{e}_2^T\mathbf{A}_n + \sum_{m=1}^{N}\lambda_m\left(G_m - G_2\right)\mathbf{e}_2^T\mathbf{R}_{\mathbf{c}_m}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n}{G_2\mathbf{e}_1^T\mathbf{A}_n + \sum_{m=1}^{N}\lambda_m\left(G_m - G_2\right)\mathbf{e}_1^T\mathbf{R}_{\mathbf{c}_m}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n} && (4.90)\\
&= \frac{G_2\mathbf{e}_2^T\mathbf{A}_n + \lambda_1 R_{s_1}\left(G_1 - G_2\right)\mathbf{A}_1^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n\mathbf{e}_2^T\mathbf{A}_1}{G_2\mathbf{e}_1^T\mathbf{A}_n + \lambda_1 R_{s_1}\left(G_1 - G_2\right)\mathbf{A}_1^H\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{A}_n\mathbf{e}_1^T\mathbf{A}_1}. && (4.91)
\end{aligned}$$

The cues of the target source are not distorted, but the cues of the background sources are corrupted by those of the target. As observed in [88], if the gain applied to the background noise is the same as that applied to the target, then the interaural cues of all sources are preserved perfectly.

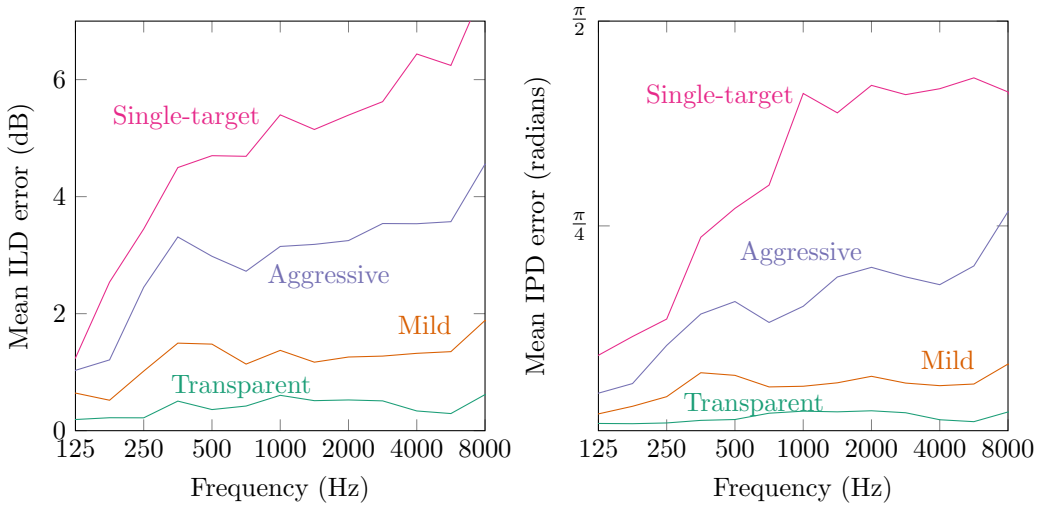

Figure 4.5: Interaural cue preservation for a binaural source-remixing filter with different desired responses.

4.3.3 Experiments

The interaural cue preservation of the source-remixing filter was analyzed using the same experimental setup as in Section 4.2.4. The experimental ITFs of the five directional sources were measured in the STFT domain using their sample cross-correlations [88]:

$$\mathrm{ITF}_n^{\mathrm{in}}[f] = \frac{\sum_k \mathbf{e}_1^T \mathbf{C}_{\mathrm{tf},n}[k,f] \mathbf{C}_{\mathrm{tf},n}^H[k,f] \mathbf{e}_2}{\sum_k \mathbf{e}_1^T \mathbf{C}_{\mathrm{tf},n}[k,f] C_{\mathrm{tf},n}^H[k,f] \mathbf{e}_1} \tag{4.92}$$

$$\mathrm{ITF}_n^{\mathrm{out}}[f] = \frac{\sum_k \mathbf{e}_1^T \hat{\mathbf{D}}_{\mathrm{tf},n}[k,f] \hat{\mathbf{D}}_{\mathrm{tf},n}^H[k,f] \mathbf{e}_2}{\sum_k \mathbf{e}_1^T \hat{\mathbf{D}}_{\mathrm{tf},n}[k,f] \hat{\mathbf{D}}_{\mathrm{tf},n}^H[k,f] \mathbf{e}_1}, \tag{4.93}$$

for $n = 1, \ldots, 5$. The experimental ILD and IPD errors were computed from the ITFs using the absolute values of (4.53) and (4.55), respectively, and averaged over the five directional sources.

Figure 4.5 shows the performance of earpieces with $M = 4$ total microphones for four sets of target responses:

1. A transparent filter with unity gain on the five speech sources and 20 dB

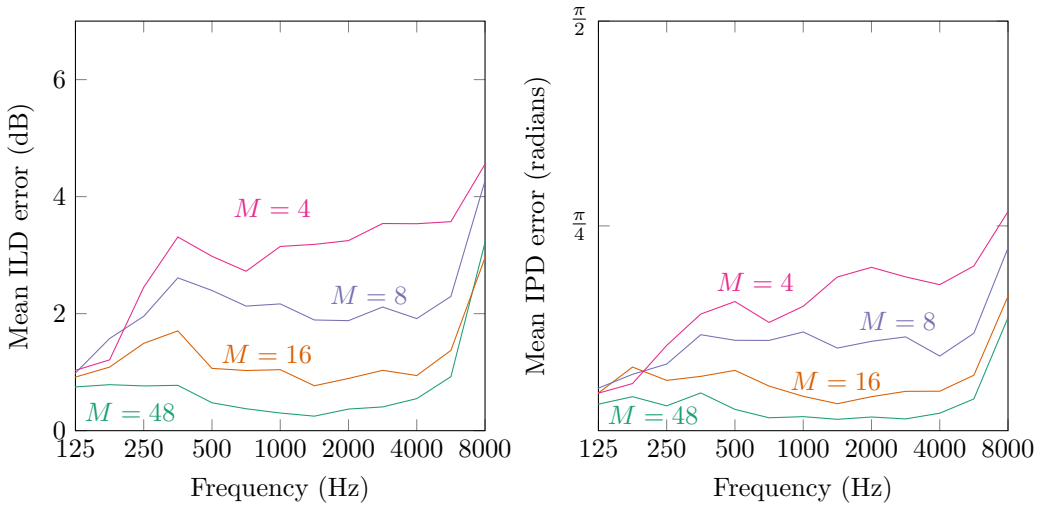

Figure 4.6: Interaural cue preservation for a binaural source-remixing filter with different microphone configurations.

attenuation on the noise channel,

2. A mild remixing filter with gains 1.0, 0.8, 0.7, 0.6, and 0.5 on the speech channels and 20 dB attenuation on the noise channel,

3. An aggressive remixing filter with gains 1.0, 0.4, 0.3, 0.2, and 0.1 on the speech channels and 20 dB attenuation on the noise channel, and

4. A single-target beamformer with $G_1 = 1$ and $G_2 = \cdots = G_6 = 0$.

As expected, the transparent filter has negligible interaural cue distortion, the single-target beamformer severely distorts the background sources, and the two remixing filters fall in between. The distortion is relatively mild at frequencies below a few hundred hertz; these wavelengths are much larger than a human head and so the ILD and IPD of all sources are close to zero.

A larger wearable array should be able to apply complex remixing to more sources than a small earpiece-based array can. Figure 4.6 shows the ILD and ITD distortion for the "aggressive" remixing responses with arrays of different sizes. The four-microphone earpiece array does not have enough degrees of freedom to preserve the

interaural cues of all five directional sources. The 8-microphone head-mounted array does better, and the 16-microphone whole-body array produces little distortion in any of the source channels.

4.4 Summary and Future Directions

The design of source-remixing filters requires a tradeoff between audio enhancement—removing and altering different sound sources to improve intelligibility—and robustness. Filters that alter the signal less, that is, that apply similar desired responses to the different source channels, cause less spectral distortion and less interaural cue distortion and do not require as much accuracy in estimating channel parameters. They also sound more immersive and natural to the listener.

A full understanding of this tradeoff will require new clinical research. The choice of desired responses will depend on the nature of the sources, the preferences of the individual, and the characteristics of the environment. In most cases, listening devices should likely adopt a "do no harm" approach: alter the sources as little as possible while achieving a desired perceptual outcome. For example, in a speech enhancement mode, a hearing aid might reduce background noise just enough to ensure intelligibility according to the listener's individual hearing profile. Applying more processing than necessary—that is, using very different G_n's—risks introducing unnecessary spectral and spatial distortion, delay, and sensitivity to the system. In quiet environments in which the auditory system can fully separate all the sound sources of interest, a listening device might not apply any spatial processing at all.

This principle of avoiding distortion might explain the conservative array processing in modern commercial hearing aids: with their limited spatial diversity, they cannot provide meaningful spatial gain without also causing the tunnel effect or other perceptible distortion. In the author's experience with recent "hearables" products that perform binaural beamforming, strong spatial processing does sound disturbingly unnatural, but can make speech intelligible in noisy environments where it would otherwise be impossible to communicate. Space-time remixing filters with

large wearable arrays could provide the advantages of both approaches: they have enough spatial resolution to meaningfully suppress strong background noise, but they have enough degrees of freedom to ensure that those attenuated noise sources sound natural. Mathematically, large arrays produce smaller values of the spatial correlation term $\mathbf{R}_{\mathbf{c}_n}\bar{\mathbf{R}}_{\mathbf{x}}^{-1}\mathbf{R}_{\mathbf{c}_m}$, which appears in some form in the expressions for spectral distortion, interaural cue distortion, squared error, and sensitivity. To improve remixing performance even further in challenging environments, we can use time-varying filters that track changes in source spectra on syllabic time scales (Chapter 7). We can also use distributed arrays (Chapter 10) that extend human sensory capabilities beyond the body.

There is one critical perceptual constraint that was not analyzed in this chapter: delay. It too depends on the choice of desired responses for different source channels and on the spatial separability of source images. It is the subject of the next chapter.

Chapter 5

Delay-Constrained Array Processing

5.1 Delay in Listening Devices

In the last chapter, we saw that there are important differences between the space-time filters used for human listening and the conventional beamformers used in machine listening and telecommunication applications. While the preservation of interaural cues in array processing has received significant attention from signal processing researchers—it was the subject of three recent doctoral theses!—there is another design constraint that remains largely unexplored: delay. Unlike most other audio processing applications, listening devices have severe constraints on delay, that is, on the time between when a sound event reaches the microphone and when it is reproduced by the receiver.

This chapter, which is an extension of the author's work in [142], explores the tradeoffs between delay and performance for array-based listening devices. How much performance must we give up to maintain imperceptible delay? Can large arrays be used to apply more powerful filters than small arrays for a given delay constraint? Finally, what do these tradeoffs look like in real rooms with wearable and distributed microphone arrays?

5.1.1 Effects of delay on human listeners

When humans listen through hearing aids or other listening devices, the output of the listening device is not the only signal they perceive. As shown in Figure 5.1,

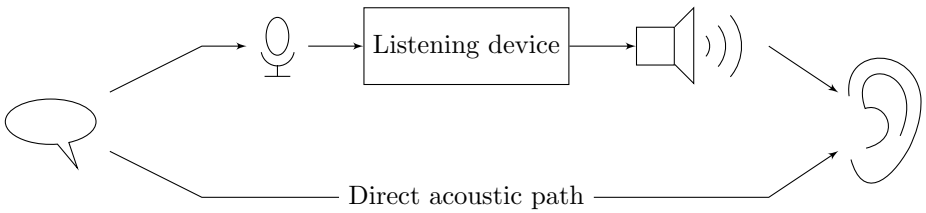

Figure 5.1: Listening devices introduce delay into processed sound signals. The ear receives a mixture of processed and unprocessed sound.

the processed output of the device is mixed with sound that enters the ear directly. Depending on the delay between them, this mixing can result in spectral distortion or even perceptible echoes.

Delay is most critical for the user's own speech, part of which is transmitted through the skull rather than the ear canal. A delay in the listener's own speech, called *delayed auditory feedback*, can interfere with speech production. One study showed that normal-hearing listeners can notice delays in their own speech as short as 3–5 ms [90] and find delays of 10 ms objectionable. Other studies have found that listeners with simulated hearing loss can tolerate longer delays on their own speech; with mild hearing loss, users rated delays of 20–30 ms disturbing and speech production was affected after 30 ms [89, 143]. Similar results were observed for subjects with real hearing loss, with disturbance decreasing with degree of hearing loss at most frequencies [144].

For external sound sources, the relative intensity of the processed and unprocessed sounds depends on the sound source and listening device. A study with closed-fitting devices, which physically block sound from entering the ear canal, found that normal-hearing listeners find delays of 9 ms disturbing and that speech comprehension is affected after 15 ms [145]. Many modern hearing aids, especially for users with mild to moderate hearing loss, have open fittings: the ear tip includes a vent that allows external acoustic pressure waves to propagate directly to the ear canal. Open-fitting aids are more comfortable but cannot provide as much gain due to feedback through the vent. It was found that with open-fitting hearing aids, delays as short as 3 to 5 ms were considered disturbing [146].

These studies have shown that distortion due to delay is most severe when direct and processed sounds have similar intensity. Thus, greater delay is tolerated with greater amplification [146]. Similarly, users with mild and severe low-frequency hearing loss are less disturbed by delay than users with moderate low-frequency loss, for whom direct and processed sounds would have similar intensity [144]. There appears to be no significant effect on delay tolerance from reverberation or dynamic range compression [143]. Users find delay more disturbing when it varies across frequency, as it might when hearing aids provide more amplification at high frequencies or process sound using nonuniform filterbanks [145]. These nonuniform delays can alter the timbre of the sound, distort temporal cues such as stop consonants, and interfere with spectral grouping in auditory scene analysis.

In addition to the direct and processed sound signals, the user also perceives visual information, such as moving lips, that the brain attempts to synchronize with sound information. Audio-visual fusion is affected by delays greater than about 80 ms [147].

There remains much to be discovered about the perceptual effects of delay in listening devices. For example, could listeners tolerate more delay in noisy environments in exchange for greater noise reduction? These open problems are discussed further in Section 5.5.

5.1.2 Sources of delay

The delays introduced by a listening device fall into two categories. The first, which could be called *hardware delay*, depends on the implementation details of the electronic processing system. Nearly all listening devices today use digital signal processing: a continuous-time signal is captured by a microphone, sampled, quantized, and transmitted to a digital processor. There is delay associated with this analog-to-digital conversion process, usually less than 1 ms. A similar delay is required to convert the processed digital signal back to an analog electrical signal that drives the receiver. There is also a processing delay required to execute the computer instructions that analyze the digitized input signal and produce a digital output signal. The amount of this delay depends on the algorithms used and the capabilities of the

processor. Using more powerful hardware can reduce hardware delay.

The second type of delay is *algorithmic delay*, which is a mathematical property of the space-time filter itself. To estimate the output $\mathbf{y}(t)$ at time t, a filter would benefit from knowledge of both present and future values of $\mathbf{x}(t)$. But real-world systems must be *causal*: they can only use information about the present and past, not about the future. For LTI space-time filters, causality requires that

$$\mathbf{w}(\tau) = 0 \quad \text{for all } \tau < 0. \tag{5.1}$$

Because a real-world system cannot observe future values of the input, it instead estimates past values of the output. That is, $\hat{\mathbf{y}}(t)$ is not an estimate of $\mathbf{y}(t)$ but of $\mathbf{y}(t-\alpha)$, where α is the algorithmic delay of the filter. In this work, it is assumed that the delay α is uniform across frequency, although in real devices delay is sometimes frequency-dependent.

While hardware delay can be reduced by using more advanced electronic circuits, algorithmic delay is a fundamental property of the estimation problem. The more information is available to the filter, the better it can perform estimation. Filters with longer delays have finer frequency resolution, for example. The tradeoff between delay and performance is especially pronounced if the listening device uses time-frequency representations. Filterbanks with more bands require more delay in order to separate signals into narrower frequency ranges. Similarly, STFTs with finer frequency resolution use longer frame sizes, which introduce greater delays.

Because hardware delay is not fundamental to the listening-enhancement inference problem and is often negligible compared to algorithmic delay in advanced listening devices, we will not explicitly consider hardware delay in this dissertation. In this chapter, we will characterize the fundamental tradeoff between algorithmic delay and squared-error performance for LTI space-time filters.

5.1.3 Low-delay processing for listening devices

Most of the analysis of Chapters 3 and 4 was performed in the continuous-time frequency domain. Filters derived in the frequency domain are noncausal in general: they are allowed to look at the entire past and future of a signal in order to make the best possible estimate. Noncausal filters are easy to derive and analyze in the frequency domain, but they cannot be implemented in practice.

Similarly, much of the literature on microphone array processing, including later chapters of this dissertation, process signals in the STFT domain [10, 11]. STFT-domain filters are designed from covariance matrices in much the same way that CTFT-domain filters are designed from power spectral densities. They can therefore be thought of as noncausal filters. In practice, the minimum algorithmic delay of any STFT-based algorithm is equal to its frame size: the system must capture an entire frame of samples before it can compute the first sample of the corresponding output frame. For speech separation, typical STFT frame sizes are 50–60 ms [57, 108]. This choice maximizes the sparsity of speech signals in the time-frequency domain, making them easier to separate and recognize. However, it is an order of magnitude larger than the delay tolerated by a human listener.

Signal processing researchers have proposed methods to reduce the algorithmic delay of filters used in hearing aids. Several authors have proposed nonlinear-phase filterbanks for low-delay hearing aid processing [148, 149]. There has also been some recent work to reduce the latency of time-frequency-domain single-microphone audio enhancement methods, including nonnegative matrix factorization [150] and deep neural networks [37, 151, 152]. However, the algorithmic delay of these methods is still fundamentally limited by the frame size of the STFT.

Despite the importance of delay to human listeners, most recent research on microphone array processing for listening devices has used frequency-domain and time-frequency-domain beamformers, which can have unacceptably long delays. However, causal microphone array processing has long been studied in the context of echo cancellation and dereverberation [41, 153–155]. This chapter takes a similar approach, formulating the space-time filter optimization problems of the previous chapter in

the time domain and imposing a causality constraint to explicitly control delay.

5.2 Causal Space-Time Filtering

In this chapter, delay-constrained filtering is posed as a causal estimation problem: Given the observed signals from the infinite past to time t, what is the desired output at time $t - \alpha$? Such problems are well understood in the scalar case [133, 156, 157], while in the multivariate case we have cumbersome theoretical tools [158, 159] but little useful insight.

To impose a delay constraint, we must modify the remixing filter from Chapter 4 in two ways. First, introduce a delay α to the desired processing response:

$$\mathbf{g}_n^{(\alpha)}(\tau) = \mathbf{g}_n(\tau - \alpha), \quad n = 1, \ldots, N. \tag{5.2}$$

Denote the corresponding desired output by

$$\mathbf{y}_\alpha(t) = \sum_{n=1}^{N} \int_{-\infty}^{\infty} \mathbf{g}_n(v - \alpha)\mathbf{c}_n(t - v)\,\mathrm{d}v \tag{5.3}$$

and the filter output by

$$\hat{\mathbf{y}}_\alpha(t) = \int_{-\infty}^{\infty} \mathbf{w}_\alpha(v)\mathbf{x}(t - v)\,\mathrm{d}v \tag{5.4}$$

where $\mathbf{w}_\alpha(\tau)$ is the filter designed to estimate $\mathbf{y}_\alpha(t)$ from $\mathbf{x}(t)$. Second, require that the space-time filter be causal:

$$\mathbf{w}_\alpha(\tau) = 0 \quad \text{for } \tau < 0. \tag{5.5}$$

Note that $\mathbf{y}_\alpha(t) = \mathbf{y}_0(t - \alpha)$, but the same is not true of $\mathbf{w}_\alpha(t)$ in general.

Although α is described as a delay, it is possible for α to be negative. Such a filter would be a space-time linear predictor. We could implement such a system

using microphones that are closer to the source than the listener is; for example, a long-distance telephone call has $\alpha \ll 0$. In augmented listening, one would typically choose α to be a small positive number, such as a few milliseconds.

5.2.1 Causal multichannel Wiener filter

Let us begin with the traditional least-squares problem: let $\lambda_n = 1$ for all $n = 1, \ldots, N$ so that $\mathbf{w}_\alpha(\tau)$ is a causal multichannel Wiener filter. Assume that $\mathbf{x}(t)$ is a zero-mean, wide-sense stationary random process. To find the linear minimum-mean-square-error filter satisfying the causality constraint (5.5), we must solve the causal Wiener-Hopf equation [156]:

$$\mathbf{r}_{\mathbf{y}_\alpha \mathbf{x}}(\tau) = \int_0^\infty \mathbf{w}_\alpha(v) \mathbf{r}_{\mathbf{x}}(\tau - v)\, \mathrm{d}v, \quad 0 < \tau < \infty, \tag{5.6}$$

where $\mathbf{r}_{\mathbf{y}_\alpha \mathbf{x}}(\tau) = \mathbb{E}\left[\mathbf{y}_\alpha(t)\mathbf{x}^T(t-\tau)\right]$. The mean square error of the resulting filter is

$$\begin{aligned} \mathcal{J}(\alpha) &= \mathbb{E}\left[|\mathbf{y}_\alpha(t) - \hat{\mathbf{y}}_\alpha(t)|^2\right] && (5.7)\\ &= \mathrm{trace}\left(\mathbf{r}_{\mathrm{err}}(0)\right) && (5.8)\\ &= \mathrm{trace}\left(\mathbf{r}_{\mathbf{y}}(0) - \int_0^\infty \mathbf{r}_{\mathbf{y}_\alpha \mathbf{x}}(t)\mathbf{w}_\alpha^T(t)\mathrm{d}t\right). && (5.9) \end{aligned}$$

Our goal is to find $\mathbf{w}_\alpha(\tau)$ and show how $\mathcal{J}(\alpha)$ depends on the spatial and spectral correlation structure of the signals.

If $\mathbf{x}(t)$ were an uncorrelated noise process so that $\mathbf{r}_{\mathbf{x}}(\tau) = \delta(\tau)\mathbf{I}$, then (5.6) would be trivial: $\mathbf{w}_\alpha(\tau) = \mathbf{r}_{\mathbf{y}_\alpha \mathbf{x}}(\tau)$ for $\tau \geq 0$. Since it is not, proceed by first whitening the input signal. To begin, decompose $\mathbf{R}_{\mathbf{x}}(\Omega)$ into its *spectral factors* [158]:

$$\mathbf{R}_{\mathbf{x}}(\Omega) = \mathbf{F}(\Omega)\mathbf{F}^H(\Omega), \tag{5.10}$$

where $\mathbf{F}(\Omega) \in \mathbb{C}^{M \times M}$ is the frequency response of a causal filter that has a causal inverse. These factors are guaranteed to exist wherever $\mathbf{R}_{\mathbf{x}}(\Omega)$ is invertible, although

they may be difficult to compute [160]. Now, $\mathbf{F}^{-1}(\Omega)$ is the frequency response of a causal whitening filter for $\mathbf{x}(t)$. Denote the whitened signal by $\mathbf{z}(t)$. The causal multichannel Wiener filter $\tilde{\mathbf{w}}_\alpha(\tau)$ that estimates $\mathbf{y}_\alpha(t)$ from $\mathbf{z}(t)$ is the solution to

$$\mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau) = \int_0^\infty \tilde{\mathbf{w}}_\alpha(v)\mathbf{r}_{\mathbf{z}}(\tau - v)\,\mathrm{d}v, \quad 0 < \tau < \infty \tag{5.11}$$

$$= \int_0^\infty \tilde{\mathbf{w}}_\alpha(v)\mathbf{I}\delta(\tau - v)\,\mathrm{d}v, \quad 0 < \tau < \infty \tag{5.12}$$

$$= \tilde{\mathbf{w}}_\alpha(\tau), \quad 0 < \tau < \infty, \tag{5.13}$$

where $\mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau) = \mathbb{E}\left[\mathbf{y}_\alpha(t)\mathbf{z}^T(t-\tau)\right]$. The filter is therefore

$$\tilde{\mathbf{w}}_\alpha(\tau) = \begin{cases} \mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau), & \text{if } \tau \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.14}$$

The causal Wiener filter is then

$$\mathbf{w}_\alpha(\tau) = \int_0^\infty \mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau - v)\mathbf{f}(v)\,\mathrm{d}v, \quad \tau > 0. \tag{5.15}$$

Since the whitening operation is invertible, the minimum mean square error for estimating $\mathbf{y}_\alpha(t)$ from $\mathbf{x}(t)$ using $\mathbf{w}_\alpha$ is the same as the minimum mean square error for estimating $\mathbf{y}_\alpha(t)$ from $\mathbf{z}(t)$ using $\tilde{\mathbf{w}}_\alpha$:

$$\mathcal{J}(\alpha) = \mathrm{trace}\left(\mathbf{r}_{\mathbf{y}}(0) - \int_0^\infty \mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau)\tilde{\mathbf{w}}_\alpha^T(\tau)\,\mathrm{d}\tau\right). \tag{5.16}$$

Substituting (5.14) into (5.16), the error covariance is given by

$$\mathbf{r}_{\text{err}}(0) = \mathbf{r}_{\mathbf{y}}(0) - \int_0^\infty \mathbf{r}_{\mathbf{y}_\alpha\mathbf{z}}(\tau)\mathbf{r}^T_{\mathbf{y}_\alpha\mathbf{z}}(\tau)\,\mathrm{d}\tau \tag{5.17}$$

$$= \mathbf{r}_{\mathbf{y}}(0) - \int_0^\infty \mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau-\alpha)\mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau-\alpha)\,\mathrm{d}\tau \tag{5.18}$$

$$= \mathbf{r}_{\mathbf{y}}(0) - \int_{-\alpha}^\infty \mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau)\mathbf{r}^T_{\mathbf{y}_0\mathbf{z}}(\tau)\,\mathrm{d}\tau \tag{5.19}$$

$$= \underbrace{\mathbf{r}_{\mathbf{y}}(0) - \int_{-\infty}^\infty \mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau)\mathbf{r}^T_{\mathbf{y}_0\mathbf{z}}(\tau)\,\mathrm{d}\tau}_{\text{Error of noncausal filter}} + \underbrace{\int_{-\infty}^{-\alpha} \mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau)\mathbf{r}^T_{\mathbf{y}_0\mathbf{z}}(\tau)\,\mathrm{d}\tau}_{\text{Causality penalty}}. \tag{5.20}$$

The first term is the error covariance of a noncausal MWF. Its trace will be denoted $\mathcal{J}_{\min}$. Time-reversing the integral in the second term, the causal multichannel Wiener filter error is

$$\mathcal{J}(\alpha) = \mathcal{J}_{\min} + \text{trace}\int_\alpha^\infty \mathbf{r}^T_{\mathbf{z}\mathbf{y}_0}(\tau)\mathbf{r}_{\mathbf{z}\mathbf{y}_0}(\tau)\,\mathrm{d}\tau. \tag{5.21}$$

For a single-output filter ($J = 1$), the error penalty would be the cross-correlation energy $|\mathbf{r}_{\mathbf{z}y_0}(\tau)|^2$ for $\tau > \alpha$. Notice that as $\alpha \to -\infty$, the error covariance approaches $\mathbf{r}_{\mathbf{y}}(0)$, the covariance of the target signal; that is, the best the system can do is guess. As α increases, the error decreases monotonically until it approaches $\mathcal{J}_{\min}$, the error of the noncausal filter.

5.2.2 Causal MSDW-MWF

Now let us consider the performance of the causal source-remixing MSDW-MWF from the previous chapter. As before, assume that the source images $\mathbf{c}_n(t)$, $n = 1, \ldots, N$, are wide-sense-stationary zero-mean random processes that are uncorrelated with each other. Each source image has matrix correlation function $\mathbf{r}_{\mathbf{c}_n}(\tau) = \mathbb{E}\left[\mathbf{c}_n(t)\mathbf{c}_n^T(t-\tau)\right]$. The sum of these source correlations is the overall input correlation $\mathbf{r}_{\mathbf{x}}(\tau)$, which is assumed to have full rank.

We seek a linear time-invariant filter $\mathbf{w}_\alpha(\tau)$ that minimizes the weighted cost

function

$$\mathcal{J}_{\text{MSDW-MWF}}(\alpha) = \sum_{n=1}^{N} \lambda_n \mathbb{E}\left[\left|\mathbf{d}_{n,\alpha}(t) - \hat{\mathbf{d}}_{n,\alpha}(t)\right|^2\right]. \tag{5.22}$$

The derivation proceeds slightly differently from that for the causal MWF because the spectral factors of the weighted PSD are not whitening filters in general. Consider the distortion-weighted cost function in the frequency domain:

$$\begin{aligned}\mathcal{J}_{\text{DW}}(\Omega) &= \sum_{n=1}^{N} \lambda_n \left(\mathbf{W}(\Omega) - \mathbf{G}_n(\Omega)\right) \mathbf{R}_{\mathbf{c}_n}(\Omega) \left(\mathbf{W}(\Omega) - \mathbf{G}_n(\Omega)\right)^H &(5.23)\\ &= \mathbf{W}(\Omega)\bar{\mathbf{R}}_{\mathbf{x}}(\Omega)\mathbf{W}^H(\Omega) - \mathbf{W}(\Omega)\bar{\mathbf{R}}^H_{\mathbf{yx}}(\Omega) - \bar{\mathbf{R}}_{\mathbf{yx}}(\Omega)\mathbf{W}^H(\Omega) + \bar{\mathbf{R}}_{\mathbf{y}}(\Omega), &(5.24)\end{aligned}$$

where the weighted PSD matrices are

$$\bar{\mathbf{R}}_{\mathbf{x}}(\Omega) = \sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{c}_n}(\Omega), \tag{5.25}$$

$$\bar{\mathbf{R}}_{\mathbf{yx}}(\Omega) = \sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega), \text{ and} \tag{5.26}$$

$$\bar{\mathbf{R}}_{\mathbf{y}}(\Omega) = \sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{G}_n^H(\Omega). \tag{5.27}$$

To proceed, we must find a spectral factorization of the weighted signal PSD:

$$\bar{\mathbf{R}}_{\mathbf{x}}(\Omega) = \mathbf{F}(\Omega)\mathbf{F}^H(\Omega), \tag{5.28}$$

where as before $\mathbf{F}(\Omega) \in \mathbb{C}^{M\times M}$ is the frequency response of a causal filter that has a causal inverse. Then, completing the square in (5.24), we have

$$\mathcal{J}_{\mathrm{DW}}(\Omega) = \left(\mathbf{W}(\Omega)\mathbf{F}(\Omega) - \bar{\mathbf{R}}_{\mathbf{yx}}(\Omega)\mathbf{F}^{-H}(\Omega)\right)\left(\mathbf{W}(\Omega)\mathbf{F}(\Omega) - \bar{\mathbf{R}}_{\mathbf{yx}}(\Omega)\mathbf{F}^{-H}(\Omega)\right)^{H} + \underbrace{\bar{\mathbf{R}}_{\mathbf{y}}(\Omega) - \bar{\mathbf{R}}_{\mathbf{yx}}(\Omega)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}(\Omega)\bar{\mathbf{R}}_{\mathbf{yx}}^{H}(\Omega)}_{\text{Noncausal minimum weighted error}}. \tag{5.29}$$

If $\mathbf{W}$ is allowed to be noncausal, then the first term can be set to zero by choosing

$$\mathbf{W}(\Omega) = \bar{\mathbf{R}}_{\mathbf{yx}}(\Omega)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}(\Omega) \tag{5.30}$$

$$= \sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\bar{\mathbf{R}}_{\mathbf{x}}^{-1}(\Omega), \tag{5.31}$$

for all Ω of interest. This is the noncausal MSDW-MWF from the previous chapter.

For a causal filter with delay α, the cost function cannot be minimized independently for different frequencies. Instead, consider the overall error

$$\mathcal{J}(\alpha) = \int_{-\infty}^{\infty} \mathrm{trace}\left(\mathcal{J}_{\mathrm{DW}}(\Omega)\right) \frac{\mathrm{d}\Omega}{2\pi} \tag{5.32}$$

$$= \mathcal{J}_{\min} + \int_{-\infty}^{\infty} \left|\mathbf{W}_{\alpha}(\Omega)\mathbf{F}(\Omega) - \bar{\mathbf{R}}_{\mathbf{y}_{\alpha}\mathbf{x}}(\Omega)\mathbf{F}^{-H}(\Omega)\right|^2 \frac{\mathrm{d}\Omega}{2\pi} \tag{5.33}$$

$$= \mathcal{J}_{\min} + \int_{-\infty}^{\infty} \left|\left[\mathbf{W}_{\alpha}(\Omega)\mathbf{F}(\Omega) - \bar{\mathbf{R}}_{\mathbf{y}_{\alpha}\mathbf{x}}(\Omega)\mathbf{F}^{-H}(\Omega)\right]_{+}\right|^2 \frac{\mathrm{d}\Omega}{2\pi} + \int_{-\infty}^{\infty} \left|\left[\bar{\mathbf{R}}_{\mathbf{y}_{\alpha}\mathbf{x}}(\Omega)\mathbf{F}^{-H}(\Omega)\right]_{-}\right|^2 \frac{\mathrm{d}\Omega}{2\pi}, \tag{5.34}$$

where $[\cdot]_+$ and $[\cdot]_-$ denote the causal and anticausal parts, respectively. Because $\mathbf{W}$ and $\mathbf{F}$ are both constrained to be causal, we can set the causal term to zero by choosing

$$\mathbf{W}_{\alpha}(\Omega) = \left[\bar{\mathbf{R}}_{\mathbf{y}_{\alpha}\mathbf{x}}(\Omega)\mathbf{F}^{-H}(\Omega)\right]_{+}\mathbf{F}^{-1}(\Omega) \tag{5.35}$$

$$= \left[e^{-j\Omega\alpha}\sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{F}^{-H}(\Omega)\right]_{+}\mathbf{F}^{-1}(\Omega). \tag{5.36}$$

The anticausal term of (5.34) is the excess weighted error due to causality:

$$\Delta\mathcal{J}(\alpha) = \int_{-\infty}^{\infty} \left| \left[\bar{\mathbf{R}}_{\mathbf{y}_\alpha \mathbf{x}}(\Omega)\mathbf{F}^{-H}(\Omega)\right]_- \right|^2 \frac{\mathrm{d}\Omega}{2\pi} \tag{5.37}$$

$$= \int_{-\infty}^{\infty} \left| \left[e^{-j\Omega\alpha} \sum_{n=1}^{N} \lambda_n \mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{F}^{-H}(\Omega) \right]_- \right|^2 \frac{\mathrm{d}\Omega}{2\pi}. \tag{5.38}$$

Let $\mathbf{r}_{\mathbf{d}_n\mathbf{z}}(\tau)$ be the inverse CTFT of $\mathbf{G}_n(\Omega)\mathbf{R}_{\mathbf{c}_n}(\Omega)\mathbf{F}^{-H}(\Omega)$. Then by Parseval's relation,

$$\mathcal{J}(\alpha) = \mathcal{J}_{\min} + \int_{-\infty}^{0} \left| \sum_{n=1}^{N} \lambda_n \mathbf{r}_{\mathbf{d}_n\mathbf{z}}(\tau - \alpha) \right|^2 \mathrm{d}\tau \tag{5.39}$$

$$= \mathcal{J}_{\min} + \int_{-\infty}^{-\alpha} \left| \sum_{n=1}^{N} \lambda_n \mathbf{r}_{\mathbf{d}_n\mathbf{z}}(\tau) \right|^2 \mathrm{d}\tau, \tag{5.40}$$

$$= \mathcal{J}_{\min} + \int_{-\infty}^{-\alpha} \left| \bar{\mathbf{r}}_{\mathbf{y}_0\mathbf{z}}(\tau) \right|^2 \mathrm{d}\tau, \tag{5.41}$$

$$= \mathcal{J}_{\min} + \int_{\alpha}^{\infty} \left| \bar{\mathbf{r}}_{\mathbf{z}\mathbf{y}_0}(\tau) \right|^2 \mathrm{d}\tau, \tag{5.42}$$

where $\bar{\mathbf{r}}_{\mathbf{y}_0\mathbf{z}}(\tau) = \sum_{n=1}^{N} \lambda_n \mathbf{r}_{\mathbf{d}_n\mathbf{z}}(\tau)$. Notice that this expression reduces to (5.21) when $\lambda_n = 1$ for $n = 1, \ldots, N$.

5.2.3 Performance of delay-constrained remixing filters

Clearly, the achievable performance of the causal filter depends on the shape of the cross-correlation function between the whitened input and the desired output: $\mathbf{r}_{\mathbf{y}_0\mathbf{z}}(\tau)$ for the MWF or $\bar{\mathbf{r}}_{\mathbf{y}_0\mathbf{z}}(\tau)$ for the MSDW-MWF. Unfortunately, it is not obvious how to characterize this function in the time domain.

To make these error expressions more concrete, consider the case where the desired responses are all scalar gains with a single output channel:

$$\mathbf{g}_n(\tau) = g_n \delta(\tau)\mathbf{e}_1^T, \quad n = 1, \ldots, N. \tag{5.43}$$

Then the weighted cross-correlation function is

$$\bar{\mathbf{r}}_{y_0\mathbf{z}}(\tau) = \sum_{n=1}^{N} \lambda_n g_n \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau). \tag{5.44}$$

The weighted error is

$$\mathcal{J}(\alpha) = \bar{r}_y(0) - \int_{-\alpha}^{\infty} \bar{\mathbf{r}}_{y_0\mathbf{z}}(\tau)\bar{\mathbf{r}}_{y_0\mathbf{z}}^T(\tau)\,\mathrm{d}\tau \tag{5.45}$$

$$= \sum_{n=1}^{N} \lambda_n g_n^2 \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n}(0)\mathbf{e}_1 - \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\lambda_m g_n g_m \int_{-\alpha}^{\infty} \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau)\mathbf{r}_{\mathbf{c}_m\mathbf{z}}^T(\tau)\mathbf{e}_1\,\mathrm{d}\tau. \tag{5.46}$$

To write the first term in a form similar to the second, observe that

$$\mathbf{r}_{\mathbf{c}_n}(0) = \int_0^{\infty} \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau)\mathbf{f}^T(\tau)\,\mathrm{d}t \tag{5.47}$$

$$= \sum_{m=1}^{N} \lambda_m \int_0^{\infty} \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau)\mathbf{r}_{\mathbf{c}_m\mathbf{z}}^T(\tau)\,\mathrm{d}\tau, \tag{5.48}$$

where $\mathbf{f}(\tau)$ is the inverse CTFT of the spectral factor $\mathbf{F}(\Omega)$. Therefore, if $\alpha > 0$, we have

$$\mathcal{J}(\alpha) = \sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\lambda_m g_n(g_n - g_m) \int_{-\alpha}^{\infty} \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau)\mathbf{r}_{\mathbf{c}_m\mathbf{z}}^T(\tau)\mathbf{e}_1\,\mathrm{d}\tau \tag{5.49}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \lambda_n\lambda_m (g_n - g_m)^2 \int_{-\alpha}^{\infty} \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n\mathbf{z}}(\tau)\mathbf{r}_{\mathbf{c}_m\mathbf{z}}^T(\tau)\mathbf{e}_1\,\mathrm{d}\tau. \tag{5.50}$$

This form of the error expression closely resembles the noncausal MSDW-MWF error

from (4.35). In fact, in the limit as $\alpha \to \infty$, we have

$$\mathcal{J}_{\min} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m (g_n - g_m)^2 \int_{-\infty}^{\infty} \mathbf{e}_1^T \mathbf{r}_{\mathbf{c}_n \mathbf{z}}(\tau) \mathbf{r}_{\mathbf{c}_m \mathbf{z}}^T(\tau) \mathbf{e}_1 \, \mathrm{d}\tau \tag{5.51}$$

$$= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m (g_n - g_m)^2 \int_{-\infty}^{\infty} \mathbf{e}_1^T \mathbf{R}_{\mathbf{c}_n}(\Omega) \bar{\mathbf{R}}_{\mathbf{x}}^{-1}(\Omega) \mathbf{R}_{\mathbf{c}_m}(\Omega) \mathbf{e}_1 \frac{\mathrm{d}\Omega}{2\pi}, \tag{5.52}$$

which is the total error from (4.35). The penalty due to causality is

$$\Delta\mathcal{J}(\alpha) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m (g_n - g_m)^2 \int_{\alpha}^{\infty} \mathbf{e}_1^T \mathbf{r}_{\mathbf{z}\mathbf{c}_n}^T(\tau) \mathbf{r}_{\mathbf{z}\mathbf{c}_m}(\tau) \mathbf{e}_1 \, \mathrm{d}\tau. \tag{5.53}$$

From this form of the expression, we can see that the required delay of the filter increases if there is a pair of sources with different gains for which $-\mathbf{e}_1^T \mathbf{r}_{\mathbf{z}\mathbf{c}_n}^T(\tau) \mathbf{r}_{\mathbf{z}\mathbf{c}_m}(\tau) \mathbf{e}_1$ has significant energy for large positive values of τ. Unfortunately, it is difficult to say much in general about these cross-correlation functions. We can, however, derive meaningful exact expressions for $\mathcal{J}(\alpha)$ for certain special cases.

5.3 Exact Results for Special Cases

Despite the complexity of multivariate spectral factorization, it is possible to find exact expressions for $\mathcal{J}(\alpha)$ in certain special cases. To simplify the already unwieldy calculations in this section, set $\lambda_n = 1$ for $n = 1, \ldots, N$. Distortion weights can always be incorporated by scaling the corresponding source correlations.

5.3.1 Plane wave in uncorrelated noise at a uniform linear array

Consider an M-input, single-output listening device with $N = 2$ source channels. The target source is a plane wave arriving at a uniform linear array with time difference

of arrival τ,

$$\mathbf{c}_1(t) = \begin{bmatrix} s_1(t) \\ s_1(t-\tau) \\ \vdots \\ s_1(t-(M-1)\tau) \end{bmatrix}, \tag{5.54}$$

where $s_1(t)$ is temporally uncorrelated. The second source is spatially and temporally uncorrelated noise with power spectral density $\sigma^2 > 0$:

$$\mathbf{R}_{\mathbf{c}_2}(\Omega) = \sigma^2 \mathbf{I}. \tag{5.55}$$

Assume that the desired response reproduces the plane wave alone at the first microphone:

$$\mathbf{G}_1(\Omega) = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \text{ and} \tag{5.56}$$

$$\mathbf{G}_2(\Omega) = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix} \text{ for all } \Omega. \tag{5.57}$$

The input power spectral density is given by

$$\mathbf{R}_{\mathbf{x}}(\Omega) = \begin{bmatrix} \sigma^2+1 & e^{+j\Omega\tau} & \cdots & e^{+j\Omega(M-1)\tau} \\ e^{-j\Omega\tau} & \sigma^2+1 & & e^{+j\Omega(M-2)\tau} \\ \vdots & & \ddots & \vdots \\ e^{-j\Omega(M-1)\tau} & e^{-j\Omega(M-2)\tau} & \cdots & \sigma^2+1 \end{bmatrix}. \tag{5.58}$$

A convenient spectral factor of $\mathbf{R}_{\mathbf{x}}$ is the lower triangular matrix

$$\mathbf{F}(\Omega) = \begin{bmatrix} b_1(\sigma^2+1) & 0 & \cdots & 0 \\ b_1 e^{-j\Omega\tau} & b_2(\sigma^2+2) & & 0 \\ \vdots & & \ddots & \\ b_1 e^{-j\Omega(M-1)\tau} & b_2 e^{-j\Omega(M-2)\tau} & & b_M(\sigma^2+M) \end{bmatrix}, \tag{5.59}$$

where

$$b_m = \sqrt{\frac{\sigma^2}{(\sigma^2+m)(\sigma^2+m-1)}}, \quad m = 1, \ldots, M. \tag{5.60}$$

Meanwhile, the cross-correlation between the input and desired output is

$$\mathbf{R}_{\mathbf{x}y_0}(\Omega) = \begin{bmatrix} 1 \\ e^{-j\Omega\tau} \\ \vdots \\ e^{-j\Omega(M-1)\tau} \end{bmatrix}. \tag{5.61}$$

The whitened cross-correlation function is therefore

$$\mathbf{r}_{\mathbf{z}y_0}(t) = \begin{bmatrix} b_1 \\ b_2\delta(t-\tau) \\ \vdots \\ b_M\delta(t-(M-1)\tau) \end{bmatrix}, \tag{5.62}$$

which yields an output mean square error from (5.21) of

$$\mathcal{J}(\alpha) = \frac{\sigma^2}{\sigma^2+M} + \sum_{m=1}^{M} b_m^2 u\left((m-1)\tau - \alpha\right) \tag{5.63}$$

$$= \frac{\sigma^2}{\sigma^2 + \sum_{m=1}^{M} u\left(\alpha - (m-1)\tau\right)}. \tag{5.64}$$

The error is reduced for each microphone that the plane wave reaches within time α of reaching the reference microphone, as shown in Figure 5.2. Notice that when $\tau < 0$, that is, when the source signal reaches the other microphones before the reference microphone, it is possible to achieve near-minimum-mean-squared-error performance with $\alpha < 0$, even though the signal is temporally unpredictable. Because the spatial correlation of the plane wave is so strong, it is spatially predictable. This example illustrates one of the ways in which large arrays, and especially distributed arrays, can help to reduce delay in a listening device: when some microphones are closer to

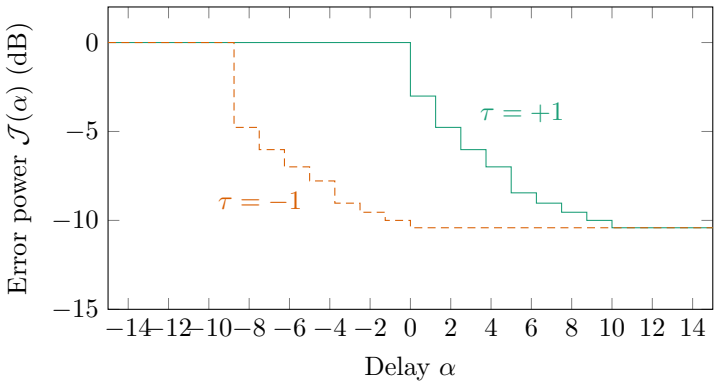

Figure 5.2: Mean square error as a function of delay for a plane wave incident on a uniform linear array of ten elements in uncorrelated noise. The time scale is arbitrary.

the sound sources than the listener's ears are, they can help to predict the signals before they reach the ears.

5.3.2 Two plane waves in uncorrelated noise at two microphones

Next consider two temporally uncorrelated plane waves incident on $M = 2$ microphones with time differences of arrival τ_1 and $\tau_2 \neq \tau_1$, again with uncorrelated background noise. Suppose that we wish to isolate $\mathbf{c}_1(t)$ and suppress the second plane wave $\mathbf{c}_2(t)$ and the noise $\mathbf{c}_3(t)$. That is, $\mathbf{g}_1(t) = \mathbf{e}_1^T \delta(t)$ and $\mathbf{g}_2 = \mathbf{g}_3 = \mathbf{0}$. Then we have

$$\mathbf{R}_{\mathbf{x}}(\Omega) = \begin{bmatrix} 2+\sigma^2 & e^{+j\Omega\tau_1} + e^{+j\Omega\tau_2} \\ e^{-j\Omega\tau_1} + e^{-j\Omega\tau_2} & 2+\sigma^2 \end{bmatrix} \quad \text{and} \tag{5.65}$$

$$\mathbf{R}_{\mathbf{x}y_0}(\Omega) = \begin{bmatrix} 1 \\ e^{-j\Omega\tau_1} \end{bmatrix}. \tag{5.66}$$

The determinant of $\mathbf{R}_{\mathbf{x}}(\Omega)$ can be written

$$\det \mathbf{R}_{\mathbf{x}}(\Omega) = \gamma^{-1} \left| 1 - \gamma e^{-j\Omega(\tau_1 - \tau_2)} \right|^2, \tag{5.67}$$

where γ is a scalar that depends only on σ^2. Thus, $\mathbf{F}^{-1}(\Omega)$ always includes a term of the form

$$\left(1-\gamma e^{+j\Omega|\tau_1-\tau_2|}\right)^{-1}, \tag{5.68}$$

which results in an infinite-duration $\mathbf{r}_{\mathbf{z}y_0}(t)$ and exponential decay of $\mathcal{J}(\alpha)$.

We can apply (5.21) to find an exact expression for the squared error of this filter, which takes different forms depending on the signs of τ_1 and τ_2:

$$\mathcal{J}(\alpha)=\begin{cases}\mathcal{J}_{\min}+\frac{u(t_0-\alpha)+q_1^2\gamma u(t_1-|\tau_1-\tau_2|-\alpha)+q_2^2 f(t_1)}{\sigma^2+2}, & \text{if } \tau_1\tau_2>0\\ \mathcal{J}_{\min}+\sqrt{\gamma}u(t_0-\alpha)+f(t_0-|\tau_1|)+\gamma f(t_1), & \text{if } \tau_1\tau_2\le 0,\end{cases} \tag{5.69}$$

where

$$t_0=\min(0,\tau_1), \tag{5.70}$$

$$t_1=\max(0,\tau_1,\tau_2,\tau_1-\tau_2), \tag{5.71}$$

$$f(t)=\gamma^{1+2\max(0,\lfloor(\alpha-t)/|\tau_2-\tau_1|\rfloor+1)}/(1-\gamma^2),\ \text{and} \tag{5.72}$$

$$(q_1,q_2)=\begin{cases}(0,0), & \text{if } |\tau_1|=|\tau_2|\\ (\sigma^2+1,\gamma\sigma^2+\gamma-1), & \text{if } |\tau_1|<|\tau_2|\\ (1,\sigma^2+1-\gamma) & \text{if } |\tau_1|>|\tau_2|.\end{cases} \tag{5.73}$$

The shape of the delay-error curve depends on the directions of arrival of the two plane waves and which microphones they reach first. Figure 5.3(a) shows these four cases. As in the earlier example, the signal can be predicted even with $\alpha<0$ when the target reaches another microphone before the reference microphone (far/near and far/far). However, after all sources have reached all microphones, the error decays faster as a function of α when the sources are closely spaced (near/near and far/far). This is somewhat counterintuitive since more widely spaced sources are generally easier to separate. However, because of the denominator polynomial from (5.68), $\mathcal{J}(\alpha)$ decays as roughly $\gamma^{2\alpha/|\tau_2-\tau_1|}$, so more narrowly spaced sources have lower error for large α.

Figure 5.3(b) shows the same scenario for speech-shaped sources. The temporal

predictability of the sources both reduces the overall error and smooths the delay-error curves.

5.4 Experimental Results

The results for the exact examples in Section 5.3 suggest that the delay-performance $\mathcal{J}(\alpha)$ curve for an array should depend on several factors:

1. The aperture of the array around the listener, because sources will reach remote microphones before the listener. A larger aperture should shift the error curve to the left.

2. The spatial separability of the signals, which depends on both the number of sources and microphones and their geometry. A larger array should shift the error curve downward.

3. The temporal separability of the signals, which depends on their spectra. Spectrally distinct sources can be separated without an array, but the delay depends on the spectral distance between them.

The exact results in the previous section are for infinite-bandwidth signals in an anechoic environment with isotropic sensors. To determine whether the predicted trends apply in real rooms, several experiments were conducted using wearable and distributed microphone arrays. The experiments in this section use the discrete-time causal MWF from Section 4.1.3, where the delay α was given in samples instead of in seconds.

5.4.1 Separation based on spatial diversity

First, the causal MWF was used to separate speech and speechlike signals recorded in the Augmented Listening Laboratory with microphone arrays of varying aperture. These experimental results were reported in [142]. The experimental setup is shown

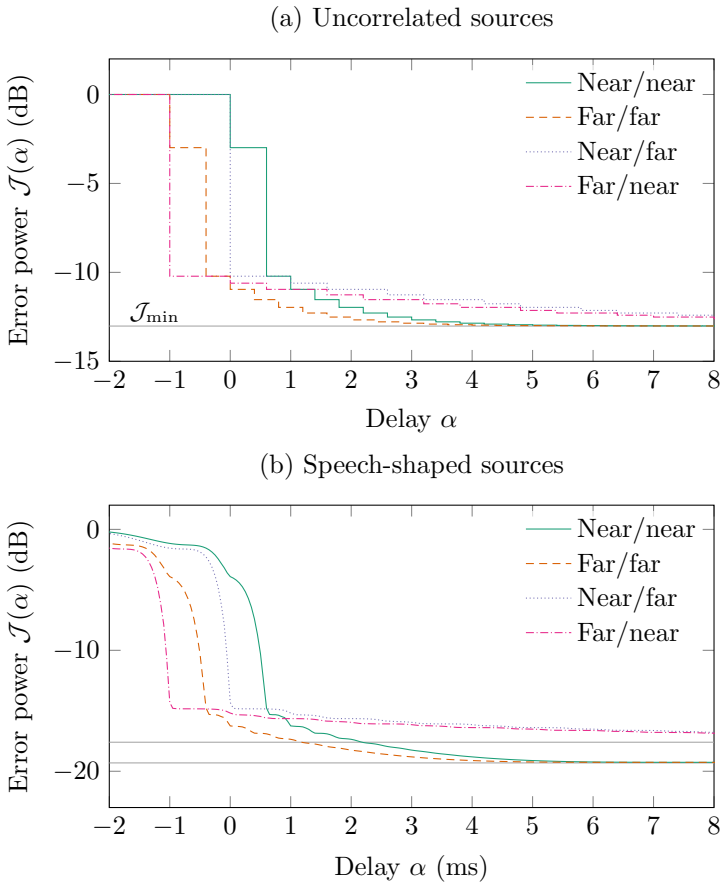

Figure 5.3: Error power as a function of delay for a plane-wave target source and plane-wave interferer incident upon a pair of microphones in uncorrelated noise. The legend indicates the direction of arrival of the target/interference sources relative to the reference microphone. (a) Uncorrelated sources. (b) Speech-shaped sources. Figure adapted from [142].

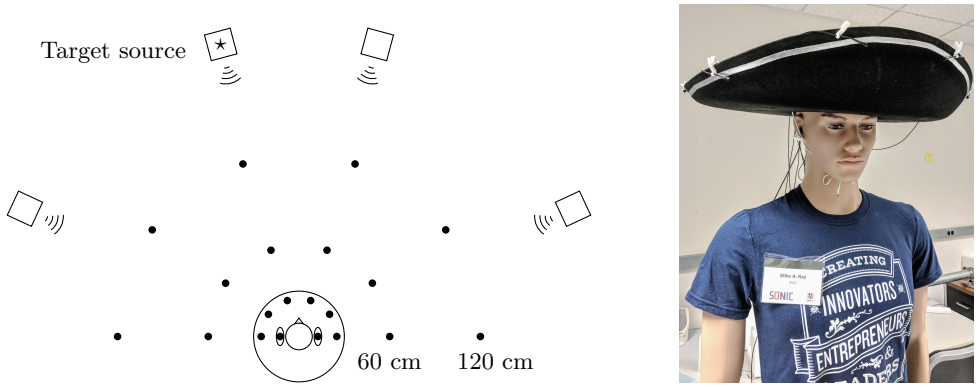

Figure 5.4: Left: Recording setup. Circles are microphones and squares are loudspeakers. Right: Hat-mounted microphone array. Figure reproduced from [142].

in Figure 5.4. Each array contains eight microphones: two in the ears of a mannequin head and six spread in a circle of radius 30 cm (on a hat), 60 cm (on stands), or 120 cm (on stands). The reference microphone is that in the left ear. Discrete-time filters were designed based on measured acoustic impulse responses to isolate the source indicated by the $\star$ and attenuate all others. To ease the computational burden of time-domain processing, the signals were downsampled to 16 kHz. The filters had length $L = 2048$ samples (128 ms).

Figure 5.5 shows the results for four 20 s speech clips from the modified VCTK corpus [104]. The space-time filters were designed assuming that the four sources were stationary with the same long-term average spectrum. That is, the filters used only the spatial diversity between sources and not spectral differences between them. As the radius of the array increases, the error curves move downward and to the left. Note that the SER improves abruptly when the sound reaches the microphones, and increases very little after $\alpha = 0$. That is because the sources are modeled as spectrally identical, so the space-time filter is separating sources primarily in space rather than time.

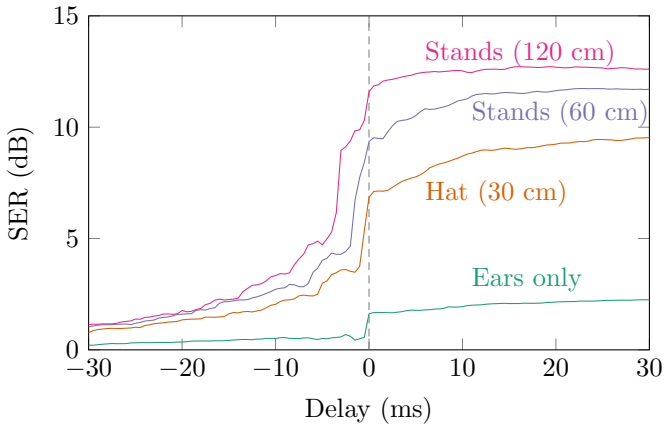

Figure 5.5: SER performance as a function of delay for the laboratory speech isolation experiment shown in Figure 5.4. Figure adapted from [142].

5.4.2 Separation based on space-time statistics

To evaluate the effects of temporal predictability, the experiment was repeated with four synthetic stationary sounds designed to imitate different vowel sounds from different talkers. The filter was designed based on windowed autocorrelation sequences since the synthesized sounds, produced by the Vocaloid music software, are actually periodic and deterministic. The source isolation performance of the filter is shown in Figure 5.6. Since the filter can now separate the signals both spatially and spectrally, the overall performance is better, especially for the two-microphone array.

This experiment may be relevant to the time-varying methods discussed in Chapter 7, which generate different filters in different time frames according to the changing short-time spectra of sound signals, especially speech. Filters that use spectral diversity between sources, like those in Figure 5.6, require far more delay than the spatial-only filters of Figure 5.5. The plot suggests, however, that microphone array listening devices could offer substantial benefits for filter delay: the eight-microphone wearable array has the same performance at $\alpha = 0$ ms as the binaural array at $\alpha = 10$ ms, which would be perceptible to some listeners. Further research is required to realize these delay benefits in time-varying methods.

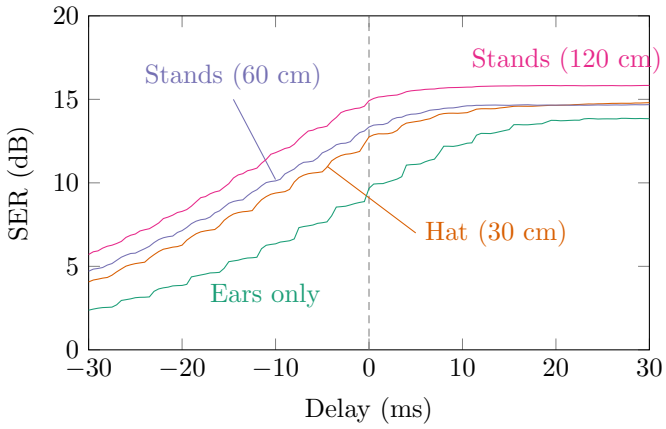

Figure 5.6: SER performance as a function of delay for synthetic speechlike sounds. Figure adapted from [142].

5.4.3 Distributed array

In the largest array characterized in [142], the distance from the farthest microphone to the listener was less than two meters. Furthermore, there was relatively little reverberation. In larger rooms, sound may propagate for tens of milliseconds from a source to a listener and may reverberate for hundreds of milliseconds. To understand delay-performance tradeoffs in larger rooms, several source-remixing space-time filters were designed using the large conference room data set of Section 2.4. The sources were the five loudspeakers closest to the listener plus additive speech-shaped, spatially uncorrelated Gaussian noise. The 128 ms filter length used in the laboratory was found to be a performance bottleneck in this large, reverberant room, so the filter length was doubled to 256 ms ($L = 4096$ at 16 kHz).

Figure 5.7 shows source-remixing performance as a function of delay for the "aggressive" remixing profile from the previous chapter (gains of 1.0, 0.1, 0.2, 0.3, and 0.4 for the talkers and 0.1 for the noise source). The filters were designed using identical long-term average speech spectra for all sources, so the sources cannot be separated based on their temporal spectra. Several different arrays were evaluated: the ears alone ($M = 2$), a wearable array with earpieces and eyeglasses ($M = 8$), the wearable array plus the ears of the other three listeners in the room ($M = 14$), and

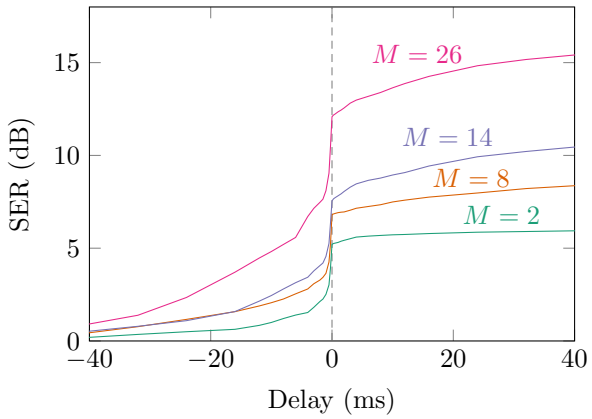

Figure 5.7: SER performance as a function of delay for a source-remixing filter with five speech sources and additive noise in a large, reverberant conference room with distributed microphone arrays.

the wearable array plus one microphone from each other listener and smart-speaker array in the room ($M = 26$).

All arrays show steep improvement between negative and positive delay, that is, when sound reaches the ears. Because the filters assume identical spectra between sources, it is impossible to temporally predict the output. The larger arrays, however, can predict the signals spatially. The 26-microphone distributed array, which includes microphones near each source, can effectively predict the sound at the ears a few milliseconds in advance, which could be valuable in a system with significant hardware delay.

The 2- and 8-microphone wearable arrays, which have microphones a few centimeters apart, benefit little from delay larger than a few milliseconds. The distributed arrays, meanwhile, show improving performance even after tens of milliseconds as sounds reach more distant microphones and reverberate around the room. There was no such benefit in Figure 5.5 because the sound sources in that experiment surrounded the microphone arrays and the room had little reverberation.

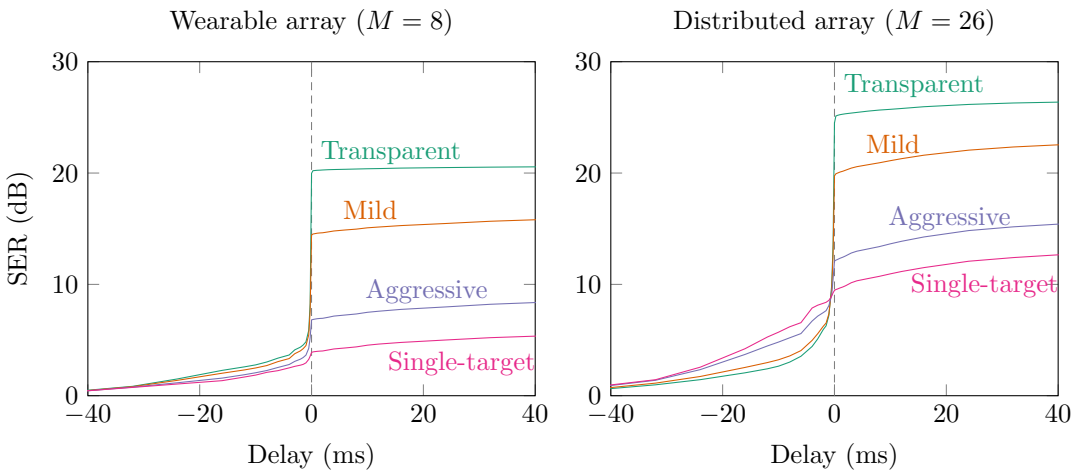

Figure 5.8: SER performance of a source-remixing filter in a large reverberant conference room for different sets of desired responses.

5.4.4 Source remixing

In Section 5.2.3, an equation was derived for the delay-performance curve as a function of the desired responses $\mathbf{g}_1, \ldots, \mathbf{g}_N$ of a remixing filter. The shape of $\mathcal{J}(\alpha)$ depends on the temporal concentration of energy in the pairwise cross-correlations between the generative matrix functions that predict the source spatial images from the whitened input—not exactly an intuitive result! To understand the impact of the desired responses on delay-performance curves in a realistic augmented-listening scenario, the distributed-array experiment was repeated with different combinations of scalar desired responses. These are the same responses as in Section 4.3.3: transparent (1, 1, 1, 1, 1, 0.1), mild (1, 0.5, 0.6, 0.7, 0.8, 0.1), aggressive (1, 0.1, 0.2, 0.3, 0.4, 0.1), and single-target (1, 0, 0, 0, 0, 0).

Figure 5.8 shows the performance-delay results for these different remixing responses with a small wearable array and with a large distributed array. For positive delays, the results are consistent with those in the previous sections: the performance of the small array is mostly flat after a few milliseconds, while the large array continues to improve; the distance between SER curves for different sets of responses is relatively constant as a function of delay. For negative delays, however, the ordering

of the curves is reversed: the single-target beamformer performs best and the transparent filter performs worst. In this experiment, the "target" source is farther from the listener than the "interference" sources, but all sources are located near at least one remote microphone. The single-target beamformer therefore has more advance information about its desired output than the transparent filter does.

5.5 Delay Constraints and Augmented Listening

Unlike hardware delay, which can be reduced by using more-expensive electronics, algorithmic delay is a seemingly inescapable fact of mathematics: to produce a better estimate, a filter needs more information about a signal, including its future values. To provide this information in a single-microphone device, we must introduce a delay to the filter. Array processing offers an escape hatch: instead of increasing the filter's temporal delay, we can increase its spatial aperture. Remote microphones can provide information about "future" values of signals by observing them near their sources, several milliseconds before they reach the user's ears. They can also reduce the need for this future information by separating signals spatially rather than spectrally.

The theoretical and experimental results in this chapter suggest that, in order to achieve good performance with imperceptible delay, microphone-array listening devices should have apertures as large as possible. A small wearable device, even if it has many more microphones than a conventional earpiece, observes sound signals only a millisecond or two before the ears. A distributed array, meanwhile, could capture them tens of milliseconds in advance. A listening device assisted by dozens of other devices spread throughout a room could, in principle, enhance sounds with zero total delay.

Could such a system be implemented in practice? The analysis and experiments in this section assume stationary sound sources, linear time-invariant acoustic channels, synchronous sampling, and linear time-invariant processing. The real world is not so well-behaved. In the following chapters, we will see how nonlinear and time-varying filters can help to deal with perceptual nonlinearity (Chapter 6), signal nonstation-

arity (Chapter 7), source and array motion (Chapter 9), and asynchronicity between devices (10). These chapters all use the short-time Fourier transform, which has a fixed algorithmic delay. To realize the delay benefits of large arrays in real-world systems, new methods must be developed that combine the sparsifying properties of the STFT with the low delay of time-domain filters. The synthetic-vowel experiments of Section 5.4.2 suggest that such methods should be possible. Indeed, spectral separation methods stand to benefit significantly from large arrays because their performance depends strongly on delay.

There are also important unanswered questions about delay and human perception that could inform the design of augmented listening systems. Most of the research that has been reported in the literature has focused on comprehension of a single talker in a controlled environment. We know that the maximum tolerable delay for a listener depends on the style of listening device (closed or open) and on the listener's hearing ability. There are many factors that have not yet been studied, including:

1. Can listeners tolerate more delay in noisy environments? That is, can the benefits of noise reduction outweigh the distortion caused by delay?

2. Can listeners tolerate more or less delay for distant sound sources than for nearby sound sources?

3. Does delay have different impacts on perception for different types of sound, such as speech and nonspeech?

4. Are there detrimental perceptual effects if different sound sources are delayed by different amounts?

With more detailed knowledge of the effects of delay on perception and intelligibility, a listening system could choose a delay or set of delays that balances the benefits of processing with the detrimental impact of delay for a given listener, source, and environment.

Chapter 6

Dynamic Range Compression

Many hearing aid users, including the author, have observed that in loud noise, they hear better with their unaided ears than with hearing aids. In the environments where they need the most help, their assistive devices make the problem worse. This behavior is caused, at least in part, by a form of nonlinear processing used in every advanced hearing aid and known as *dynamic range compression* (DRC) [32–34]. As the name implies, DRC systems map the wide dynamic range of real-world sounds to the narrower dynamic range of a listener, processing system, or storage medium, as shown in Figure 6.1. Compression makes loud sounds quieter and quiet sounds louder; thus, it is a form of time-varying, input-dependent nonlinear amplification. Because it is nonlinear, DRC does not obey the superposition principle and can cause distortion when applied to a mixture of multiple signals. While this distortion has been documented extensively by clinical researchers, there has been little work on DRC in the signal processing literature. This work introduces new mathematical tools to characterize the effect of noise on DRC and describes a novel approach to compression, introduced by the author in [93], that attempts to modify the dynamic range of each source channel independently.

Dynamic range compression should be familiar to anyone who has listened to recorded music. Mixing engineers use compression to increase the perceived loudness of certain musical styles [161]. So-called side-chain compression can be used to lower the loudness of one track when another is active, for example to "duck" music under vocals. In certain genres, especially electronic music, side-chain compression is widely used to produce intentional distortion effects. This unnatural "pumping" sound is entertaining on a Daft Punk album, but is an unwelcome distraction when it happens

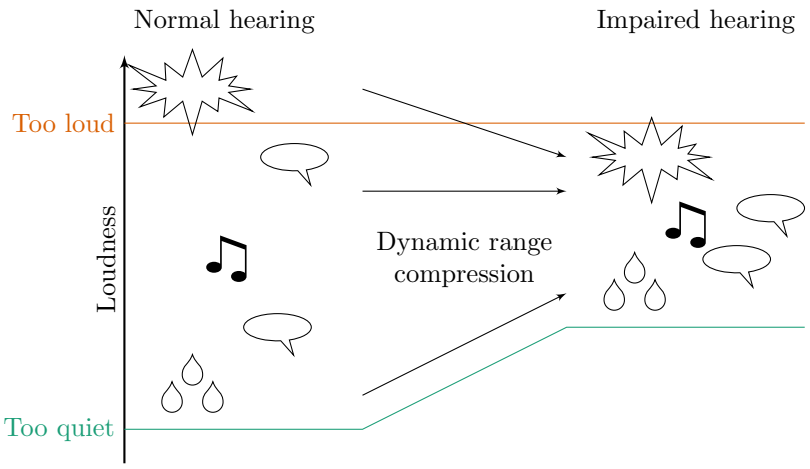

Figure 6.1: Dynamic range compression (DRC) maps the wide dynamic range of an input signal onto a narrower output dynamic range.

in hearing aids.

Hearing aids use DRC in a different way: to mimic the function of a healthy ear. The healthy human hearing system is highly nonlinear, performing automatic gain control within different frequency bands as it translates mechanical pressure waves to nerve impulses [7]. Listeners with hearing loss often have reduced dynamic range compared to normal-hearing listeners: the lower threshold of hearing is higher, but the upper threshold of pain remains the same. In certain types of hearing loss, damage to the outer hair cells of the cochlea results in a *recruitment* effect that can be modeled as a dynamic range expander—the opposite of a compressor—causing perceived loudness to increase more quickly as a function of intensity than it normally would [33]. Hearing aids compensate for reduced dynamic range and for recruitment by applying compression in different frequency bands.

In previous chapters, we have derived listening systems as the solutions to inverse problems. We can treat DRC the same way. Figure 6.2 shows two different inverse problems to which DRC is a solution. The model on the left is motivated by the auditory system: a dynamic range expander models the recruitment phenomenon in an impaired ear, and the DRC system in a listening device compensates for it. The model on the right is reversed: there is an imaginary desired sound source that has a

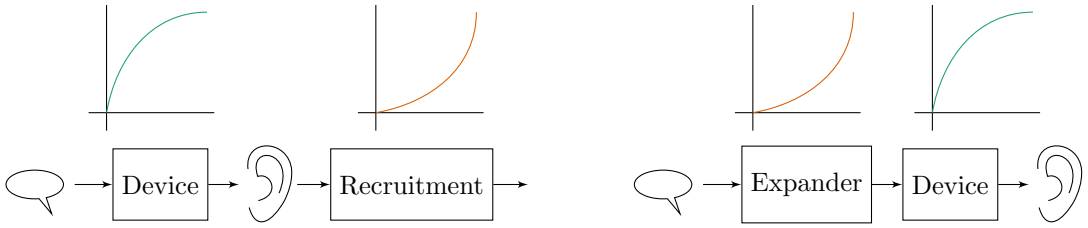

Figure 6.2: Dynamic range compression can be viewed as the solution to either of two inverse problems: an expander on the listener side, modeling loudness recruitment due to hearing loss (left), or an expander on the source side, modeling a source with undesired dynamic range (right).

comfortable dynamic range, but it is made too loud by an imaginary expander. The DRC system undoes this expansion to recover the comfortable source. For a single sound source, these models are largely equivalent. An advantage of the model on the left is that it explicitly models hearing impairment so that the system attempts to restore normal function. The model on the right is not directly motivated by human hearing, but it offers a significant advantage when there are multiple sources present, as we will see in Section 6.3.

6.1 Dynamic Range Compression of a Single Source

A dynamic range compression system has several key parts, shown in Figure 6.3. First, the signal may or may not be split into different frequency bands. Within each band, an envelope detector tracks the amplitude or power of the signal, with nonlinear behavior that helps the system react quickly to sudden loud sounds. A compression function maps the amplitude or power of the input to the desired amplitude or power of the output. Finally, a time-varying amplifier applies gain or attenuation to bring the signal to the desired output level.

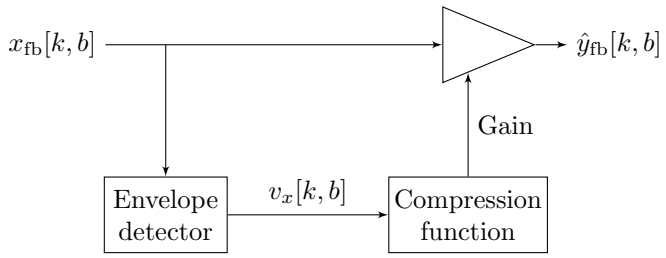

Figure 6.3: A dynamic range compression system consists of an envelope detector, a compression function, and a time-varying amplifier.

6.1.1 Filterbank

A key difference between music compression and hearing aid compression is that music compression is often applied to the signal as a whole or to a small number of frequency bands [161]. Hearing aid DRC systems typically operate independently, or at least partially independently, in several different frequency bands [34].[1] There are a few reasons that hearing aids use multiple bands:

1. Hearing aid DRC systems attempt to mimic the gain control functions of a healthy ear. The nonlinearities of the ear have been observed to operate roughly independently on tones separated by more than the so-called critical bandwidth, which varies with frequency [33]. This can be crudely modeled as a set of DRC systems operating on the bands of a nonuniform filterbank.

2. If desired signals and undesired noise are in different frequency bands, then a change in the level of one signal will not cause distortion in the other. This across-source modulation effect is the subject of Section 6.2. Because many real-world sounds have broad spectra, however, it would be more reliable to use spatial source separation methods, as proposed in Section 6.3.

3. The number of channels in a hearing aid is featured prominently in hearing aid marketing materials, and more expensive hearing aids typically have more

[1] In the hearing aid industry and literature, these are called "channels." Here I use "bands" to avoid confusion with microphone channels or source channels.

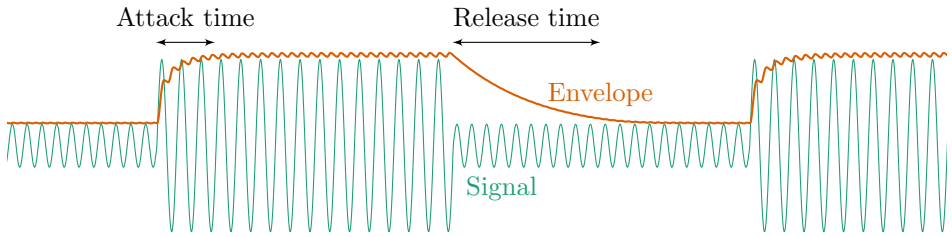

Figure 6.4: An envelope detector tracks the level of the signal over time. It typically responds faster to increases in level (attack time) than to decreases in level (release time).

channels. However, it is a matter of controversy whether there is any clinical benefit to using more than a few channels [32, 162].

Let $x_{\mathrm{fb}}[k, b]$ for $b = 1, \ldots, B$ be the B-band filterbank representation of the sequence $x_{\mathrm{d}}[k]$. These bands may be uniform or nonuniform and may or may not overlap. The filterbanks used in hearing aids typically mimic the critical bands of the cochlea. These are roughly linearly spaced at low frequencies and logarithmically spaced at higher frequencies [7].

6.1.2 Envelope detection

The *envelope detector* is responsible for tracking the level of the signal over time, as illustrated in Figure 6.4. Level is typically defined in terms of either amplitude or power; here we use power. As shown in the figure, a DRC envelope detector typically reacts more quickly to increases in signal level than to decreases. The quick reduction in gain following level increases, known as the *attack time*, protects listeners from sudden loud sounds. The slow increase in gain following decreases, characterized by the *release time*, prevents excessive distortion during brief pauses in speech. The attack and release times are defined in ANSI S3.22 [163] to be the times required for the output envelope to change by 31 dB following a large change in input level. In hearing aids, attack times are typically just a few milliseconds while release times vary from tens to hundreds of milliseconds [164]. Note that in a

multiband DRC system, the time constants must be carefully chosen based on the bandwidth of the filterbank. If the gains within each band change too quickly, they can generate unwanted out-of-band distortion products [165].

Let $v_x[k,b] \in \mathbb{R}^+$ be the output of the envelope detector applied to $x_{\text{fb}}[k,b]$. Envelope detection can be implemented in several ways. In the idealized model of Section 6.2, we will consider the envelope to be the instantaneous variance of a random process. In real systems, the envelope detector typically performs nonlinear smoothing of the signal level. A representative envelope detector, which is used in the experiments presented in this chapter, is the following nonlinear single-tap recursive filter [28, 161]:

$$v_x[k,b] = \begin{cases} \beta_{\text{a}} v_x[k-1,b] + (1-\beta_{\text{a}}) \left| x_{\text{fb}}[k,b] \right|^2, & \text{if } \left| x_{\text{fb}}[k,b] \right|^2 > v_x[k-1,b] \\ \beta_{\text{r}} v_x[k-1,b] + (1-\beta_{\text{r}}) \left| x_{\text{fb}}[k,b] \right|^2, & \text{otherwise,} \end{cases} \tag{6.1}$$

for $b = 1, \ldots, B$, where β_{a} and β_{r} are constants that determine the attack and release times.

6.1.3 Time-varying gain

A dynamic range compression system controls dynamic range using a time-varying gain in the signal path. Let $v_y[k,b]$ be the envelope of the desired output sequence $y_{\text{fb}}[k,b]$ in band b. It is given by the compression function $\mathcal{C}_b$, which relates input power to output power:

$$v_y[k,b] = \mathcal{C}_b \left(v_x[k,b] \right), \quad b = 1, \ldots, B. \tag{6.2}$$

Assume that $v_x[k,b] > 0$ for all k, b. The desired filter unit pulse response at time index k and band b is then

$$g_{\text{fb}}[\ell; k, b] = \sqrt{\frac{v_y[k,b]}{v_x[k,b]}} \delta[\ell], \quad b = 1, \ldots, B, \tag{6.3}$$

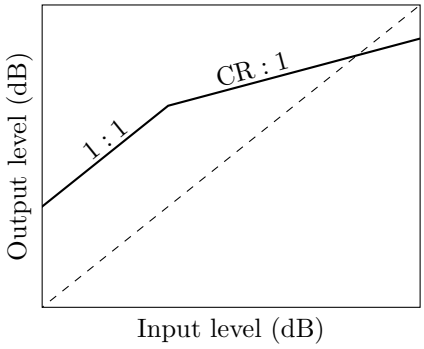

Figure 6.5: The compression function relates the level of the input signal to the level of the output signal. The slope of the compression curve on a log-log scale is the inverse of the compression ratio.

so that the output,

$$y_{\mathrm{fb}}[k,b] = \sum_{\ell=-\infty}^{\infty} g_{\mathrm{fb}}[\ell;k,b]x_{\mathrm{fb}}[k-\ell,b], \quad b = 1,\ldots,B, \tag{6.4}$$

has the desired envelope. Since there is only one source channel in a conventional DRC system, $w_{\mathrm{fb}}[\ell;k,b] = g_{\mathrm{fb}}[\ell;k,b]$ and $\hat{y}_{\mathrm{fb}}[k,b] = y_{\mathrm{fb}}[k,b]$.

6.1.4 Compression function

The change in dynamic range of a signal processed by a DRC system in band b is determined by the compression function $\mathcal{C}_b(v)$. The function may or may not be different in different bands. A typical "knee-shaped" compression function is shown in Figure 6.5. It features a linear region in which gain is constant with input level and a compressive region characterized by a compression ratio (CR), which is the inverse of the slope on a log-log scale. For example, in a 3:1 compression system, the output level increases by 1 dB for every 3 dB increase in input level. For a compressor with constant compression ratio CR, the compression function has the form

$$\mathcal{C}(v) = g_0^2 v^{1/\mathrm{CR}}, \tag{6.5}$$

where g_0^2 is a constant gain factor that shifts the compression curve vertically. Thus, in a 3:1 compression region, the output level will be proportional to the cube root of the input level.

While these power-law compression functions are commonly used, many other compression functions are possible. Cascaded feedback systems can be used to design smoothly curved compression functions with roughly logarithmic shapes [7], while digital signal processing can be used to implement arbitrary functions. To help with our analysis, let us adopt the following definition of a compression function:

Definition 6.1. A function $\mathcal{C}(v)$ is a *compression function* if it is concave, nonnegative, and nondecreasing for all $v > 0$.

The amount of compression applied to a signal is often described by the compression ratio. Though the CR is usually used to describe compression curves that are piecewise linear on a log-log scale (corresponding to piecewise power-law compression functions), it can be extended to more general compression functions. Since the CR can be infinite—in a limiter, for example—it is more convenient to work with its inverse, defined here as the compression slope.

Definition 6.2. For all points v at which a compression function $\mathcal{C}(v)$ is differentiable, the *compression slope* $\mathrm{CS}(v)$ is the slope of $\mathcal{C}(v)$ on a log-log scale:

$$\mathrm{CS}(v) = \frac{\mathrm{d}}{\mathrm{d}u} \ln \mathcal{C}(e^u)|_{u=\ln v} \tag{6.6}$$

$$= \frac{\mathcal{C}'(e^u)}{\mathcal{C}(e^u)} e^u|_{u=\ln v} \tag{6.7}$$

$$= \frac{\mathcal{C}'(v)}{\mathcal{C}(v)} v. \tag{6.8}$$

For example, if $\mathcal{C}(v) = g_0^2 v^\alpha$, then $\mathrm{CS}(v) = \alpha$.

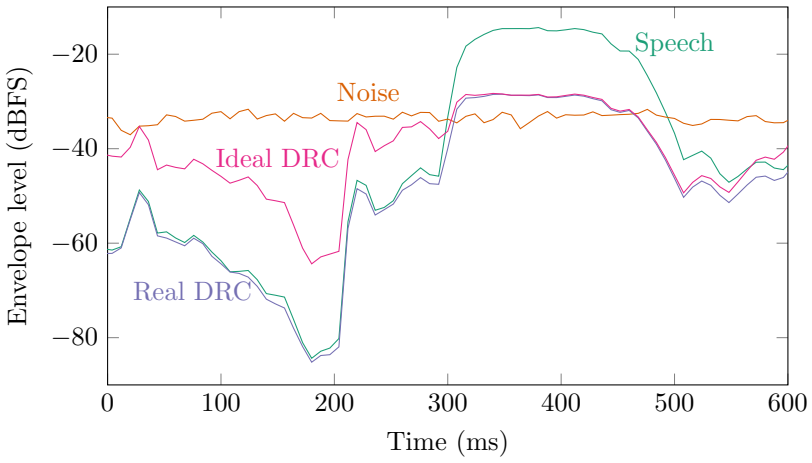

Figure 6.6: When one compressor is applied to a mixture of multiple sources, a change in the level of one source causes distortion in the envelopes of the other sources. Figure adapted from [93].

6.2 Dynamic Range Compression and Noise

Dynamic range compression in hearing aids is useful for improving intelligibility in a quiet room. Unfortunately, the DRC systems used today do not work well when there are multiple signals present.

It has been widely observed that noise has adverse effects on the performance of dynamic range compression [44–46, 48–50]. For example, consider a single speech source in stationary background noise. At low signal-to-noise ratios, the envelope detector follows the noise rather than the signal of interest, rendering the compressor ineffective [45], as shown in Figure 6.6. Meanwhile, at high signal-to-noise ratios, compression amplifies the quieter noise sounds and attenuates the louder speech sounds, reducing the overall signal-to-noise ratio at the output [45, 48, 49]. These effects have been observed in recent commercial hearing aids and shown to adversely affect speech comprehension [48].

Performance is also impacted by multiple sources of speech or other highly non-stationary sounds. A sudden increase in the level of one sound causes the gain applied to all sounds to decrease. This distortion has been called *co-modulation* [44]

or *across-source modulation* [46] in the hearing literature. Anecdotally, the author's hearing aids produce severe distortion in response to loud applause, which is a series of wideband impulsive sounds at irregular intervals.

Although the mechanisms behind compression reduction, signal-to-noise ratio reduction, and across-source modulation appear obvious intuitively, they have never been rigorously analyzed mathematically. In this section, we will prove that, for an idealized DRC system, noise and compression interact in two adversarial ways. First, noise reduces the compression applied to a signal of interest and, second, compression reduces the output signal-to-noise ratio for a rapidly varying source in slowly varying noise. We will also consider the interaction between two time-varying sources. The results in this section are reported here for the first time.

6.2.1 The effective compression function

Distortion between signals in mixtures is caused by the nonlinear behavior of dynamic range compression. There are two sources of nonlinearity in DRC systems: the peak-tracking behavior of the envelope detector, that is, the different speeds of adaptation to increases and decreases in signal level, and the level-dependent gain defined by the compression function. While the dynamics of the envelope tracker are important, here we restrict our attention to the compression function that is the defining feature of DRC.

Consider a scalar sequence $x_{\mathrm{d}}[k]$ that is a mixture of two source channels, $x_{\mathrm{d}}[k] = c_{\mathrm{d},1}[k] + c_{\mathrm{d},2}[k]$. For simplicity of notation, we restrict our attention to a single band and omit the band index b in this section; the analysis presented here applies to compression in general and does not depend on filterbank structure. To facilitate mathematical analysis of the compression function applied to such a mixture, assume that the envelope obeys the superposition property $v_x[k] = v_{c_1}[k] + v_{c_2}[k]$. This is true, for example, if $c_{\mathrm{d},1}[k]$ and $c_{\mathrm{d},2}[k]$ are uncorrelated zero-mean random processes and $v_x[k]$ is a linear transformation of the sequence $\mathbb{E}\left[x_{\mathrm{d}}^2[k]\right] = \mathbb{E}\left[c_{\mathrm{d},1}^2[k]\right] + \mathbb{E}\left[c_{\mathrm{d},2}^2[k]\right]$. Note that this idealized envelope detector does not exhibit the peak-tracking behavior of (6.1).

The gain applied at each sample time k is determined by the mixture level $v_x[k]$:

$$g_{\mathrm{d}}[\ell; k] = \sqrt{\frac{\mathcal{C}(v_x[k])}{v_x[k]}}\delta[\ell]. \tag{6.9}$$

$$= \sqrt{\frac{\mathcal{C}(v_{c_1}[k] + v_{c_2}[k])}{v_{c_1}[k] + v_{c_2}[k]}}\delta[\ell]. \tag{6.10}$$

Thus, the envelope of the output component due to source channel 1 is

$$v_{\hat{d}_1}[k] = \frac{\mathcal{C}(v_{c_1}[k] + v_{c_2}[k])}{v_{c_1}[k] + v_{c_2}[k]} v_{c_1}[k]. \tag{6.11}$$

We can therefore define the effective compression function $\hat{\mathcal{C}}(v_1|v_2)$ for a target source with envelope v_1 in the presence of an interfering source with envelope v_2.

Definition 6.3. The *effective compression function* $\hat{\mathcal{C}}(v_1|v_2)$ applied to v_1 in the presence of v_2 is given by

$$\hat{\mathcal{C}}(v_1|v_2) = \frac{\mathcal{C}(v_1 + v_2)}{v_1 + v_2} v_1, \tag{6.12}$$

where $\mathcal{C}(v)$ is the compression function applied to the mixture envelope $v_1 + v_2$.

The output envelopes for the two channels are then $v_{\hat{d}_1}[k] = \hat{\mathcal{C}}(v_{c_1}[k]|v_{c_2}[k])$ and $v_{\hat{d}_2}[k] = \hat{\mathcal{C}}(v_{c_2}[k]|v_{c_1}[k])$. We can understand the behavior of DRC in the presence of noise by analyzing this effective compression function. First, let us prove that that for a broad class of compression functions, $\hat{\mathcal{C}}(v_1|v_2)$ is concave in v_1 and convex in v_2.

Theorem 6.1. *If $\mathcal{C}(v)$ is concave and nonnegative and $\mathcal{C}(v)/v$ is convex for all $v > 0$, then $\hat{\mathcal{C}}(v_1|v_2)$ is concave in v_1 and convex in v_2.*

Proof. Starting with the definition and letting $v_1 = \lambda x + (1-\lambda)y$,

$$\hat{\mathcal{C}}(\lambda x + (1-\lambda)y | v_2) = \frac{\mathcal{C}(\lambda x + (1-\lambda)y + v_2)}{\lambda x + (1-\lambda)y + v_2}(\lambda x + (1-\lambda)y) \tag{6.13}$$

$$= \mathcal{C}(\lambda(x+v_2) + (1-\lambda)(y+v_2)) - v_2 \frac{\mathcal{C}(\lambda(x+v_2) + (1-\lambda)(y+v_2))}{\lambda(x+v_2) + (1-\lambda)(y+v_2)} \tag{6.14}$$

$$\geq \lambda \mathcal{C}(x+v_2) + (1-\lambda)\mathcal{C}(y+v_2) - v_2 \left(\lambda \frac{\mathcal{C}(x+v_2)}{x+v_2} + (1-\lambda)\frac{\mathcal{C}(y+v_2)}{y+v_2} \right) \tag{6.15}$$

$$= \lambda \frac{\mathcal{C}(x+v_2)}{x+v_2} x + (1-\lambda)\frac{\mathcal{C}(y+v_2)}{y+v_2} \tag{6.16}$$

$$= \lambda \hat{\mathcal{C}}(x|v_2) + (1-\lambda)\hat{\mathcal{C}}(y|v_2). \tag{6.17}$$

Therefore $\hat{\mathcal{C}}(v_1|v_2)$ is concave in v_1.

Similarly, letting $v_2 = \lambda x + (1-\lambda)y$,

$$\hat{\mathcal{C}}(v_1 | \lambda x + (1-\lambda)y) = \frac{\mathcal{C}(v_1 + \lambda x + (1-\lambda)y)}{v_1 + \lambda x + (1-\lambda)y} v_1 \tag{6.18}$$

$$= \frac{\mathcal{C}(\lambda(v_1+x) + (1-\lambda)(v_1+y))}{\lambda(v_1+x) + (1-\lambda)(v_1+y)} v_1 \tag{6.19}$$

$$\leq \lambda \frac{\mathcal{C}(v_1+x)}{v_1+x} v_1 + (1-\lambda)\frac{\mathcal{C}(v_1+y)}{v_1+y} v_1 \tag{6.20}$$

$$= \lambda \hat{\mathcal{C}}(v_1|x) + (1-\lambda)\hat{\mathcal{C}}(v_1|y). \tag{6.21}$$

Therefore $\hat{\mathcal{C}}(v_1|v_2)$ is convex in v_2. □

The condition that $\mathcal{C}(v)/v$ be convex is satisfied by many smooth compression functions, including the popular power-law compression function (6.5) with CR ≥ 1. It is violated by some piecewise functions. Note that if $\mathcal{C}$ is a linear function, that is, a fixed gain, then $\hat{\mathcal{C}}(v_1|v_2)$ is linear in v_1 and does not vary with v_2. For strictly concave compression functions, such as cube-root or logarithmic compression, the output power of one source decreases as the power of the other source increases.

This behavior is the origin of the across-source modulation effect in DRC systems.

6.2.2 Noise reduces effective compression ratio

It has been observed in the hearing aid literature that DRC is less effective in noisy environments [45]. This phenomenon is readily apparent for negative signal-to-noise ratios: the gain of the system is determined primarily by the noise and the dynamics of the signal are largely ignored. However, it can be shown that any level of noise in the mixture will reduce the compression applied to the signal of interest.

From the effective compression function, we can find the effective compression slope that the system applies to each source.

Definition 6.4. If $\hat{\mathcal{C}}(v_1|v_2)$ is differentiable with respect to v_1, then the *effective compression slope* $\hat{\mathrm{CS}}(v_1|v_2)$ is given by

$$\hat{\mathrm{CS}}(v_1|v_2) = \frac{\partial}{\partial u}\hat{\mathcal{C}}(e^u|v_2)|_{u=\ln v_1} \tag{6.22}$$

$$= \frac{\frac{\partial}{\partial v_1}\hat{\mathcal{C}}(v_1|v_2)}{\hat{\mathcal{C}}(v_1|v_2)}v_1. \tag{6.23}$$

Let us now prove that this effective compression slope is larger (less compressive) than the nominal compression slope $\mathrm{CS}(v_1 + v_2)$, which implies that noise reduces the effective compression ratio of a DRC system.

Theorem 6.2. *If a compression function $\mathcal{C}(v)$ is differentiable at $v = v_1 + v_2$, then its effective compression slope satisfies*

$$\hat{\mathrm{CS}}(v_1|v_2) \geq \mathrm{CS}(v_1 + v_2), \tag{6.24}$$

with equality if the compression function is linear or if $v_2 = 0$.

Proof. First, because $\mathcal{C}(v)$ is concave and nonnegative for $v > 0$, we have

$$\mathcal{C}(v) - v\mathcal{C}'(v) \geq 0 \tag{6.25}$$

for all v at which $\mathcal{C}$ is differentiable, with equality if $\mathcal{C}$ is linear.

Let $v_x = v_1 + v_2$. From the definition of the effective compression slope,

$$\hat{\mathrm{CS}}(v_1|v_2) = \frac{\frac{\partial}{\partial v_1}\hat{\mathcal{C}}(v_1|v_2)}{\hat{\mathcal{C}}(v_1|v_2)} v_1 \tag{6.26}$$

$$= \frac{v_x}{\mathcal{C}(v_x)} \left(\frac{\mathcal{C}'(v_x)v_1 + \mathcal{C}(v_x)}{v_x} - \frac{\mathcal{C}(v_x)v_1}{v_x^2} \right) \tag{6.27}$$

$$= \frac{\mathcal{C}'(v_x)}{\mathcal{C}(v_x)} v_1 + 1 - \frac{v_1}{v_x} \tag{6.28}$$

$$= \frac{\mathcal{C}'(v_x)}{\mathcal{C}(v_x)} v_x - \frac{\mathcal{C}'(v_x)}{\mathcal{C}(v_x)} v_2 + \frac{v_2}{v_1 + v_2} \tag{6.29}$$

$$= \mathrm{CS}(v_x) + \frac{v_2}{v_x \mathcal{C}(v_x)} \left(\mathcal{C}(v_x) - v_x \mathcal{C}'(v_x) \right) \tag{6.30}$$

$$\geq \mathrm{CS}(v_x) \tag{6.31}$$

with equality if $\mathcal{C}$ is linear. It is clear from (6.30) that equality also holds if $v_2 = 0$. □

Equation (6.30) illustrates that the impact of noise on the compression slope is stronger at lower signal-to-noise ratios, as one would expect. This analysis does not account for the nonlinear dynamics of the envelope detector, however. To verify that the predicted relationship holds with a realistic DRC system, an experiment was conducted using a speech signal from the VCTK database [104] in varying levels of white Gaussian noise. The DRC system has a constant compression ratio of 3:1, an attack time of 10 ms, a release time of 50 ms, and 6 mel-spaced filterbank bands, which are roughly linearly spaced at low frequencies and logarithmically spaced at high frequencies.

Figure 6.7 shows theoretical and experimental effective compression functions for the speech signal, that is, values of $v_{\hat{d}_1}[k, b]$ plotted against $v_{c_1}[k, b]$. Each effective compression function transitions from linear gain to the nominal compression function near the noise level: speech signal components that are well above the noise level are compressed correctly, while those below the noise level are amplified with constant gain determined by the noise level. The empirical envelope points do not

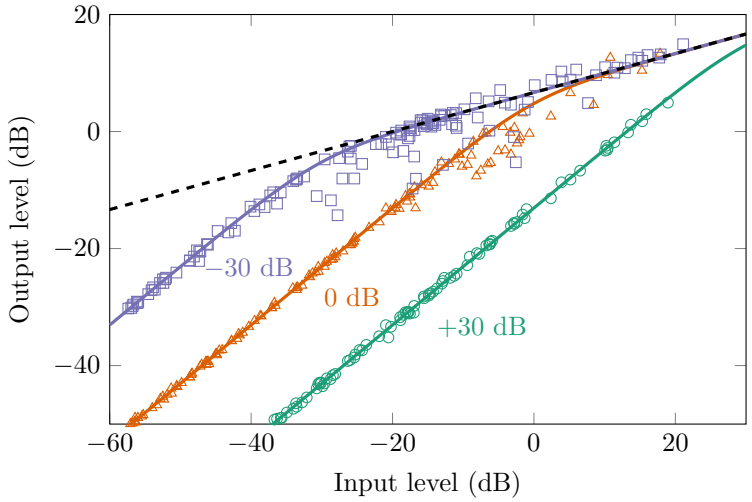

Figure 6.7: Effective compression function of a speech source in varying amounts of white Gaussian noise using a 3:1 compressor. The dashed curve is the nominal compression function, the solid curves are the theoretical effective compression functions, and the plotted points are samples of empirical input and output envelopes. The labels show the noise power; the average level of the speech signal is 0 dB.

lie exactly on the predicted effective compression functions because the nonlinear envelope detector does not obey the superposition property. However, the points do follow the curves closely, especially at low levels where the noise dominates and the effective compression function is nearly linear.

This effect has been documented in the hearing aid literature as well. In [45], the effective compression ratio applied to speech was found to decrease in strong background noise.

6.2.3 Compression reduces signal-to-noise ratio for constant-envelope noise

The detrimental interaction between compression and noise goes both ways: not only does noise reduce the effectiveness of compression, compression can make noise worse. This effect can be motivated by comparing dynamic range compression with classic noise reduction algorithms [35]. In many noise reduction algorithms, such as spectral subtraction and Wiener filtering, time frames with large signal levels are assumed to be dominated by a signal of interest, such as speech, while periods with low levels are considered noise. Gain is increased for high-level intervals and reduced for low-level intervals, resulting in a dynamic range *expansion* system. A compression system, meanwhile, amplifies the quieter periods and attenuates the louder signal of interest, reducing the overall average signal-to-noise ratio.

To characterize this effect mathematically, let us adopt a commonly used assumption from the speech enhancement literature: that the signal of interest is highly nonstationary while the unwanted noise is mostly stationary on the time scale of the algorithm. That is, let $v_{c_1}[k]$ vary arbitrarily with time and fix $v_{c_2}[k] = \bar{v}_2$ for all k. Since the envelopes are proportional to signal power, the average input signal-to-noise is given by

$$\mathrm{SNR}_{\mathrm{in}} = \frac{\mathrm{mean}_k v_{c_1}[k]}{\mathrm{mean}_k v_{c_2}[k]}, \tag{6.32}$$

and the average output signal-to-noise ratio is given by

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\mathrm{mean}_k v_{\hat{d}_1}[k]}{\mathrm{mean}_k v_{\hat{d}_2}[k]} \tag{6.33}$$

$$= \frac{\mathrm{mean}_k \hat{\mathcal{C}}(v_{c_1}[k]|v_{c_2}[k])}{\mathrm{mean}_k \hat{C}(v_{c_2}[k]|v_{c_1}[k])}. \tag{6.34}$$

If the compression function were linear, such as a static gain, the input and output SNRs would be identical. For a concave compression function with convex gain function, we can prove that the output SNR is lower than the input SNR.

Theorem 6.3. *If $\mathcal{C}(v)$ is a compression function and $\mathcal{C}(v)/v$ is convex for all $v > 0$, $v_{c_1}[k] > 0$ for all k, and $v_{c_2}[k] = \bar{v}_2 > 0$ for all k, then*

$$\mathrm{SNR}_{\mathrm{out}} \leq \mathrm{SNR}_{\mathrm{in}} \tag{6.35}$$

with equality if $v_{c_1}[k]$ is constant or $\mathcal{C}$ is linear.

Proof. Since $v_{c_2}[k]$ is fixed, the output SNR can be written

$$\mathrm{SNR}_{\mathrm{out}} = \frac{\mathrm{mean}_k \hat{\mathcal{C}}(v_{c_1}[k]|\bar{v}_2)}{\mathrm{mean}_k \hat{\mathcal{C}}(\bar{v}_2|v_{c_1}[k])}. \tag{6.36}$$

The numerator is the mean over k of a concave function of $v_{c_1}[k]$. By Jensen's inequality, we have

$$\mathrm{mean}_k \hat{\mathcal{C}}(v_{c_1}[k]|\bar{v}_2) \leq \hat{\mathcal{C}}(\mathrm{mean}_k v_{c_1}[k]|\bar{v}_2), \tag{6.37}$$

with equality when $\mathcal{C}$ is linear or $v_{c_1}[k]$ is constant. Similarly, the denominator is the mean of a convex function of $v_{c_1}[k]$. Again applying Jensen's inequality,

$$\mathrm{mean}_k \hat{\mathcal{C}}(\bar{v}_2|v_{c_1}[k]) \geq \hat{\mathcal{C}}(\bar{v}_2|\mathrm{mean}_k v_{c_1}[k]), \tag{6.38}$$

with equality when $\mathcal{C}$ is linear or $v_{c_1}[k]$ is constant. Let $\bar{v}_1 = \mathrm{mean}_k v_{c_1}[k]$. Since the

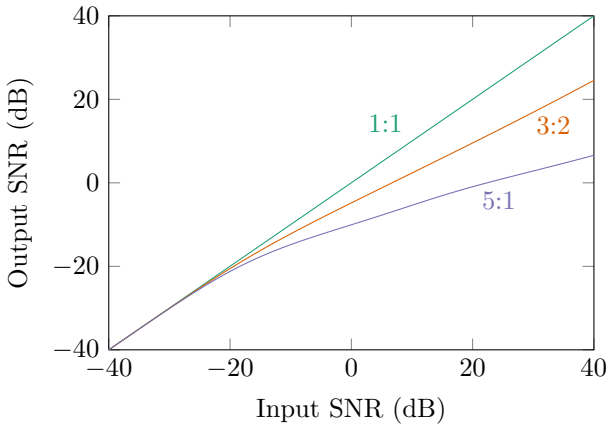

Figure 6.8: Signal-to-noise ratio at the output of a DRC system for mixtures of speech and white Gaussian noise at different compression ratios.

numerator and denominator are both strictly positive, we have

$$\begin{aligned}
\mathrm{SNR}_{\mathrm{out}} &\leq \frac{\hat{\mathcal{C}}(\bar{v}_1|\bar{v}_2)}{\hat{\mathcal{C}}(\bar{v}_2|\bar{v}_1)} && (6.39)\\
&= \frac{\bar{v}_1\mathcal{C}(\bar{v}_1+\bar{v}_2)/(\bar{v}_1+\bar{v}_2)}{\bar{v}_2\mathcal{C}(\bar{v}_1+\bar{v}_2)/(\bar{v}_1+\bar{v}_2)} && (6.40)\\
&= \frac{\bar{v}_1}{\bar{v}_2} && (6.41)\\
&= \mathrm{SNR}_{\mathrm{in}}, && (6.42)
\end{aligned}$$

with equality when $\mathcal{C}$ is linear or $v_{c_1}[k]$ is constant. □

To demonstrate this effect in a realistic compression system, a speech signal from the VCTK database was mixed with varying amounts of white Gaussian noise. The mixtures were compressed using a knee-shaped compression function like that shown in Figure 6.5 with different compression ratios in the compression region. The attack time was 10 ms, the release time was 50 ms, and there were 6 mel-spaced filterbank bands. The input and output SNRs were computed using the time-domain signals rather than the filterbank envelopes used in the proof.

The results are shown in Figure 6.8. Since this noise has a steady envelope, it

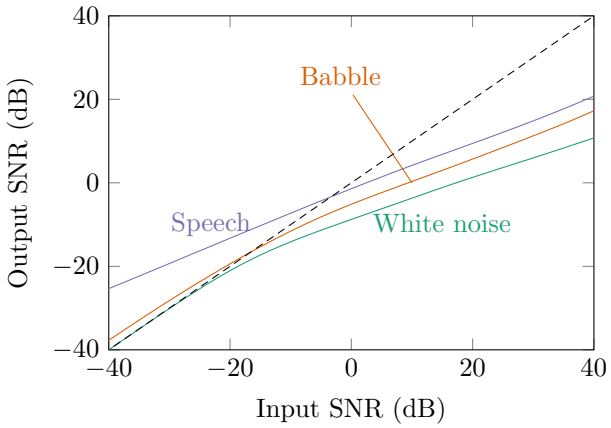

Figure 6.9: Signal-to-noise ratio at the output of a 3:1 DRC system for mixtures of speech and different types of noise.

should closely match the results of Theorem 6.3. At low input SNRs, the effective compression function is more linear and the SNR is largely unchanged. At high input SNRs, the output SNR is reduced by the compressors, with higher compression ratios causing greater reduction in SNR. This effect of greater SNR reduction at greater compression ratios has also been observed in the hearing aid literature [45].

6.2.4 Compression and time-varying noise

Theorem 6.3 only applies to constant-envelope noise. Other noise types may exhibit higher or lower SNR when compressed. If the target and noise sources both vary significantly over time but have different average levels, it is reasonable to expect that the louder source would receive less gain on average and the quieter source would receive more gain on average, pushing the long-term ratio between them closer to one. Mathematical analysis of this interaction is more complex than in the constant-envelope-noise case; however, we can study the empirical effects of compression on different types of noise.

Figure 6.9 shows the output SNR as a function of input SNR for different types of noise with the same 3:1 compressor. The other parameters are the same as in

the previous section. The output SNR for white Gaussian noise is always lower than the input SNR, as predicted by Theorem 6.3. For babble noise—a mixture of several talkers from the VCTK dataset—the SNR is slightly improved at low input SNRs. It is still worse than the input at high input SNRs, though better than white noise. When the noise source is another speech signal comparable to the target, there is substantial SNR improvement at low SNRs and degradation at high SNRs: the compressor makes the two average source powers more similar to each other.

These results align well with experiments that have been conducted in commercial hearing aids. In [48], the input-output SNR curve of a commercial hearing aid was measured for speech in stationary noise, modulated noise, and other speech. These measured curves correspond closely to those in Figure 6.9 for white noise, babble, and speech. These effects were shown to be stronger at higher compression ratios.

The impact of compression on intelligibility for nonstationary noise remains unclear. Some studies have suggested that compression improves intelligibility for modulated noise but not for constant-envelope noise [32,50], perhaps because of improvements in average SNR. However, this SNR improvement comes at a price: the sources modulate each other's envelopes, distorting temporal patterns that may contribute to intelligibility. Two signal envelopes that are jointly compressed become negatively correlated; this across-source modulation effect has been shown to negatively impact intelligibility in normal-hearing listeners [46].

6.2.5 Empirical metrics of compression performance

The results above use an idealized envelope detector based on statistical expectation. It cannot be implemented in the real world. To quantify the interaction between compression and noise in realistic experiments, we will adopt two empirical metrics from the hearing aid literature and one from the speech enhancement literature.

The effective compression ratio (ECR) [45] measures the overall average compression ratio applied to a source by comparing the dynamic range of the output to the dynamic range of the input. This metric is useful for compression functions with regions of constant compression ratio, such as power-law functions.

Definition 6.5. The *effective compression ratio* ECR_n between an input envelope $v_{c_n}[k,b]$ and output envelope $v_{\hat{d}_n}[k,b]$ is given by

$$\mathrm{ECR}_n = \mathrm{mean}_b \left\{ \frac{\log\left(\max_k v_{c_n}[k,b] / \min_k v_{c_n}[k,b]\right)}{\log\left(\max_k v_{\hat{d}_n}[k,b] / \min_k v_{\hat{d}_n}[k,b]\right)} \right\}, \tag{6.43}$$

where the max and min operations exclude the largest and smallest 5% of envelope samples. If the compression curve does not have a constant compression ratio, then envelope samples that fall outside the constant-ratio range of the curve are also excluded.

Note that the envelope detector used to compute the ECR can be different from the envelope detector of the DRC system. Generally, the ECR is smaller than the nominal compression ratio because of the smoothing effects of the envelope detector.

Next, the across-source modulation coefficient (ASMC) [46] measures the degree to which two source channels alter each other's envelopes.

Definition 6.6. The *across-source modulation coefficient* $\mathrm{ASMC}_{n,p}$ between source channels n and p is given by

$$\mathrm{ASMC}_{n,p} = \mathrm{mean}_b \left\{ \mathrm{corr}_k \left\{ \log \max\left(v_{\hat{d}_n}[k,b], \frac{\bar{v}_n[b]}{20} \right), \log \max\left(v_{\hat{d}_p}[k,b], \frac{\bar{v}_p[b]}{20} \right) \right\} \right\}, \tag{6.44}$$

where corr_k is the sample correlation coefficient across time indices and $\log \bar{v}_n[b] = \mathrm{mean}_k v_{\hat{d}_n}[k,b]$.

The across-source modulation coefficient was found to be correlated with intelligibility for human listeners [46].

In addition to these empirical metrics from the hearing aid literature, we can adopt a commonly used speech enhancement metric known as log-spectral distortion (LSD), which is more perceptually relevant than squared error [166].

Definition 6.7. The *log-spectral distortion* of the output envelope $v_{\hat{d}_n}$ with respect

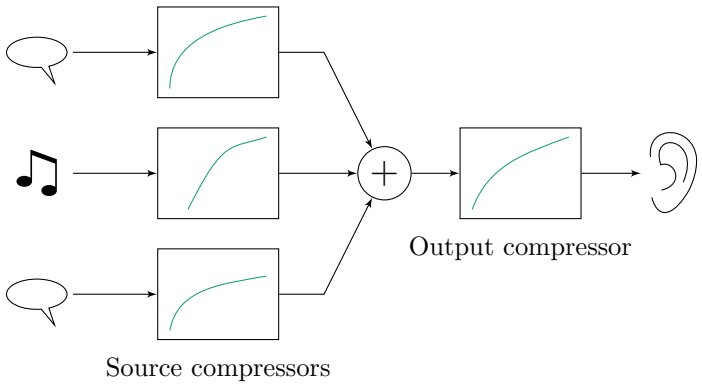

Figure 6.10: The proposed multiple-source compression system aims to apply different compression settings to each source channel.

to the desired envelope $v_{\hat{d}_n}$ is given by

$$\mathrm{LSD}_n = \mathrm{mean}_{k,b} \left\{ \left| 10 \log_{10} v_{d_n}[k,b] - 10 \log_{10} v_{\hat{d}_n}[k,b] \right| \right\}. \tag{6.45}$$

6.3 Dynamic Range Compression of Multiple Sources

The across-source modulation effects described above occur when two or more sources are processed by the same nonlinear system. If instead each source was processed by a separate DRC system, as shown in Figure 6.10, these interactions could be avoided. As a further advantage, each source could be compressed in a different way: for example, music could be processed with longer release times than other signals and background noise could be limited to a barely audible level. The source-specific DRC systems could be followed by an overall DRC that keeps the mixture at a comfortable level.

It should be emphasized that the proposed system is not directly motivated by models of human hearing; there is, to the best of the author's knowledge, no evidence that healthy human ears apply separate compression to each sound source. This approach has never been studied clinically and it is unclear whether or how much it would benefit hearing aid users. However, there are reasons to expect that

it would be beneficial for both hearing-impaired and normal-hearing users. In music mixing, compression is typically applied to individual musical instruments or tracks instead of or in addition to overall mixes. Advanced modern hearing aids change their compression settings in different environments and for different source types, such as speech and music, suggesting that different source types should be compressed in different ways. These hearing aids also often have adaptive features that alter compression settings when there are multiple sources present. For example, it was recently proposed to change compression time constants within time-frequency bins based on a speech presence classifier [167]. The system applies fast-acting compression to speech and slow-acting compression to noise, thus improving the effective compression ratio.

There is a need for clinical research into how hearing-impaired and normal-hearing listeners respond to independent compression of multiple sources and how source-specific compression should be combined with overall-mixture compression. In the meantime, we can develop signal processing methods to implement a multisource compression system using a multimicrophone listening device. This section describes a multimicrophone, multisource compression system proposed by the author in [93] and presents new experiments based on the wearable array data set.

6.3.1 Multisource compression in multimicrophone listening enhancement

While there have been single-microphone approaches proposed for applying independent compression to different sources [167], it would be advantageous to use the source-separating power of multimicrophone systems. Consider again the inverse problems illustrated by Figure 6.2 and imagine that there are multiple sound sources mixed together. When the expander is on the listener side, as is typically assumed for hearing aid signal processing, the resulting DRC system applies the same compression to every source. The model on the right allows us to imagine different expanders on every source, as illustrated in Figure 6.11. This figure also includes an expander on the listener side, allowing us to model recruitment as well as the ranges

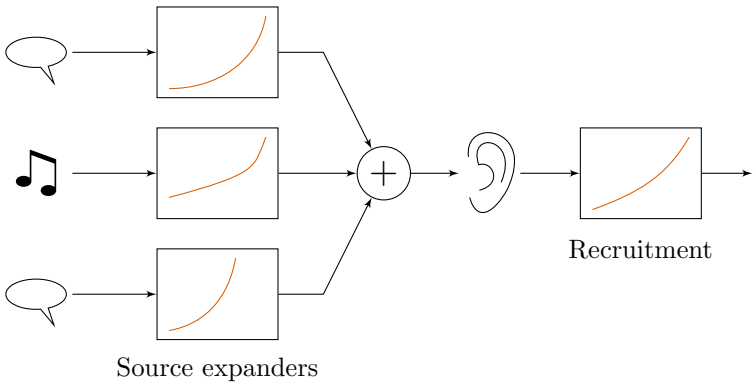

Figure 6.11: The notional multisource compression system solves the inverse problem shown above. Each source is associated with a different dynamic range expander. The listener may also have an expander modeling recruitment.

of individual sources.

Our goal is to reproduce the mixture as it would have been perceived with no expanders. That is, each desired source image $\mathbf{d}_{\mathrm{d},n}[k]$ is the clean source image $\mathbf{c}_{\mathrm{d},n}[k]$ compressed by its source-specific compressor as perceived by the left and right ears. The proposed multisource, multimicrophone compression system is shown in Figure 6.12. We can recover the desired compressed source images using a time-varying version of the space-time filter introduced in previous chapters.

Let $\mathbf{c}_{\mathrm{fb},n}[k,b]$ be the filterbank representation of the source image for source channel n in band b for $n = 1,\ldots,N$ and $b = 1,\ldots,B$. Let $\mathcal{V}(\cdot)$ be a measure of instantaneous power for the M-dimensional source image that satisfies $\mathcal{V}(a\mathbf{c}) = a^2\mathcal{V}(\mathbf{c})$, such as $\mathcal{V}(\mathbf{c}) = \mathbf{c}^T\mathbf{c}$, $\mathcal{V}(\mathbf{c}) = \left|\mathbf{e}_1^T\mathbf{c}\right|^2$, or $\mathcal{V}(\mathbf{c}) = \max\left(\left|\mathbf{e}_1^T\mathbf{c}\right|^2, \left|\mathbf{e}_2^T\mathbf{c}\right|^2\right)$. The latter might be useful for binaural hearing aids in which the gain is synchronized in the left and right earpieces. The experiments presented here use the power in the left ear only. Using the recursive envelope detector from (6.1), the true envelope associated with source channel n is

$$v_{\mathbf{c}_n}[k,b] = \begin{cases} \beta_{\mathrm{a},n} v_{\mathbf{c}_n}[k-1,b] + (1-\beta_{\mathrm{a},n})\mathcal{V}\left(\mathbf{c}_{\mathrm{fb},n}[k,b]\right), & \text{if } \mathcal{V}\left(\mathbf{c}_{\mathrm{fb},n}[k,b]\right) > v_{\mathbf{c}_n}[k-1,b] \\ \beta_{\mathrm{r},n} v_{\mathbf{c}_n}[k-1,b] + (1-\beta_{\mathrm{r},n})\mathcal{V}\left(\mathbf{c}_{\mathrm{fb},n}[k,b]\right), & \text{otherwise,} \end{cases} \tag{6.46}$$

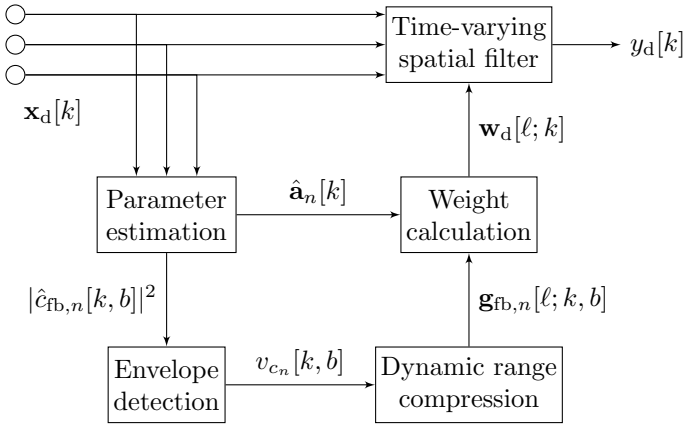

Figure 6.12: The proposed multisource compression system estimates the envelopes of each source and adjusts the target gains of the spatial filters accordingly.

for $n = 1, \ldots, N$ and $b = 1, \ldots, B$. Notice that the attack and release constants can be different for each source channel. A listener would likely prefer different compression parameters for speech and music, for example. In practice, the true envelopes are not available and must be estimated from the mixture using source separation methods; the experiments in the next section use the output powers of fixed MVDR beamformers for each source channel.

Now, the desired source channel output envelopes are

$$v_{\mathbf{d}_n}[k,b] = \mathcal{C}_{b,n}\left(v_{\mathbf{c}_n}[k,b]\right), \quad b = 1, \ldots B \text{ and } n = 1, \ldots, N. \tag{6.47}$$

The compression functions can also be different for each source channel. To preserve interaural cues in a binaural listening device, we should apply the same gains to the left and right outputs. Thus, the desired unit pulse response matrices are

$$\mathbf{g}_{\mathrm{fb},n}[\ell;k,b] = \sqrt{\frac{v_{\mathbf{d}_n}[k,b]}{v_{\mathbf{c}_n}[k,b]}}\delta[\ell]\begin{bmatrix}\mathbf{e}_1^T\\ \mathbf{e}_2^T\end{bmatrix}, \quad b = 1, \ldots B \text{ and } n = 1, \ldots, N. \tag{6.48}$$

The desired output of the listening device is then

$$\mathbf{y}_{\mathrm{fb}}[k,b] = \sum_{n=1}^{N} \sum_{\ell=-\infty}^{\infty} \mathbf{g}_{\mathrm{fb},n}[\ell;k,b]\mathbf{c}_{\mathrm{fb},n}[k-\ell,b] \tag{6.49}$$

$$= \sum_{n=1}^{N} \mathbf{d}_{\mathrm{fb},n}[k,b], \quad b = 1,\ldots B. \tag{6.50}$$

Unlike in the scalar case, it will generally be impossible to achieve this desired output exactly. Instead, the desired gains are used to design a time-varying filter $\mathbf{w}_{\mathrm{d}}[\ell;k]$ that yields the output

$$\hat{\mathbf{y}}_{\mathrm{d}}[k] = \sum_{\ell=-\infty}^{\infty} \mathbf{w}_{\mathrm{d}}[\ell;k]\mathbf{x}_{\mathrm{d}}[k-\ell] \tag{6.51}$$

$$= \sum_{n=1}^{N} \hat{\mathbf{d}}_{\mathrm{d}}[k]. \tag{6.52}$$

The filter coefficients must be chosen so that $\hat{\mathbf{y}}_{\mathrm{d}}[k] \approx \mathbf{y}_{\mathrm{d}}[k]$. In this work, we use the MSDW-MWF based on using the time-varying target gains computed from the compression function:

$$\mathbf{w}_{\mathrm{d}}[\ell;k] = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \lambda_n \mathbb{E}\left[\left(\hat{\mathbf{d}}_{\mathrm{d}}[k] - \mathbf{d}_{\mathrm{d}}[k]\right)^2\right], \tag{6.53}$$

where λ_n are speech distortion weights. One could imagine other filters that explicitly trade off between compression performance, noise reduction, distortion, and other criteria; these are a subject for future work.

The implementation presented here uses a MSDW-MWF in the short-time Fourier transform domain. The output at each time index k and frequency index f is given by:

$$\hat{\mathbf{Y}}_{\mathrm{tf}}[k,f] = \mathbf{W}_{\mathrm{df}}[k,f]\mathbf{X}_{\mathrm{tf}}[k,f]. \tag{6.54}$$

The space-time filter coefficients are recomputed for every time-frequency frame

Table 6.1: Multisource compression results. For SER and ECR, higher numbers are better. For LSD, lower is better. For ASMC, numbers closer to zero are better. Table adapted from [93].

	Two speakers				Speech and noise				Five speakers			
	SER	ECR	ASMC	LSD	SER	ECR	ASMC	LSD	SER	ECR	ASMC	LSD
Ideal DRC	∞	1.76	0.01	0.0	∞	1.76	0.05	0.0	∞	1.62	0.01	0.0
Conventional DRC	11.0	1.27	−0.20	9.7	−2.7	1.36	−0.61	11.8	−2.7	1.11	−0.10	12.1
Ground-truth MWF	16.3	1.73	−0.02	6.3	3.9	1.36	−0.47	12.6	6.8	1.24	−0.03	10.5
Ground-truth LCMV	16.4	1.75	-0.02	4.3	0.09	1.51	-0.68	8.0	1.5	1.57	-0.11	6.0
Mismatch MWF	9.2	1.61	-0.07	6.6	3.1	1.32	−0.44	13.6	−1.1	1.23	−0.04	11.3
Mismatch LCMV	8.9	1.61	-0.07	4.6	0.3	1.48	-0.67	9.0	-4.3	1.36	-0.09	7.3

based on the target gains:

$$\mathbf{W}_{\mathrm{df}}[k,f] = \left(\sum_{n=1}^{N} \lambda_n \mathbf{G}_{\mathrm{df},n}[k,f]\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k,f]\right)\left(\sum_{n=1}^{N} \lambda_n \mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k,f]\right)^{-1}, \tag{6.55}$$

where $\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k,f] = \mathrm{Cov}(\mathbf{C}_{\mathrm{tf},n}[k,f])$ and $\mathbf{G}_{\mathrm{df},n}[k,f]$ is the discrete Fourier transform of $\mathbf{g}_{\mathrm{d},n}[\ell;k]$. The source covariance matrices are computed using time-invariant rank-one transfer function models scaled by the estimated source envelopes:

$$\hat{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[k,f] = \hat{v}_{\mathbf{C}_{\mathrm{tf},n}}[k,f]\hat{\mathbf{A}}_{\mathrm{df}}[f]\hat{\mathbf{A}}_{\mathrm{df}}^{H}[f]. \tag{6.56}$$

6.3.2 Experiments with different mixture types

In [93], a multisource dynamic range compression system was implemented using the STFT-domain filter described above and its performance was compared to that of a conventional compressor using the metrics from Section 6.2.5 as well as SER (3.26). The simulations were performed using measured impulse responses from binaural behind-the-ear earpieces with three microphones each ($M = 6$) in a reverberant courtyard [97]. Those impulse responses were convolved with 20 second speech clips from the TIMIT corpus [103] and processed by six algorithms: an ideal multisource compressor with a priori knowledge of the source channel signals, a conventional

compressor applied to the mixtures at the two reference microphones, time-varying MWF ($\lambda_n = 1$) and LCMV ($\lambda_n \to \infty$) filters designed using ground truth acoustic impulse responses, and time-varying filters designed using mismatched impulse responses, which were measured on the same head dummy but in a different room. The last experiment is the most realistic since it accounts for channel estimation errors. For all multichannel filters, the source channel envelopes and the magnitudes of the time-frequency covariance matrices were estimated from the outputs of time-invariant MVDR filters. All experiments were repeated for 100 combinations of speech clips.

Table 6.1 shows results for three compression problems. In the first, two speech sources were each to be independently compressed with a target ratio of 5:1. Both the ground truth and mismatched beamformers did well at separating these sources, which were far apart from each other in the room. All multimicrophone systems outperformed the conventional compressor on all three envelope metrics, although the mismatched beamformer had a lower SER.

In the second scenario, a single speech source was to be compressed at 5:1 and an approximately isotropic speech-shaped noise source was to be removed completely. Because there was only one directional source, acoustic channel mismatch had little effect on beamformer performance. The MWF beamformers improved SER substantially, but at the cost of target source distortion, which adversely affected the envelope metrics. The LCMV reduced noise less, but better compressed the target source. The ASMC is less meaningful for this experiment because the noise was not compressed.

In the final scenario, one foreground speech source was lightly compressed at 3:1 with +20 dB gain and four background speech sources were heavily compressed at 6:1 with no gain. Because there are five sources and six closely spaced microphones, the SER is sensitive to parameter mismatch. As before, the MWF was more effective at separating the signals and reducing across-source modulation effects while the LCMV produced more accurately compressed envelopes, as measured by the LSD and ECR.

The experiments show that multimicrophone processing can improve compression

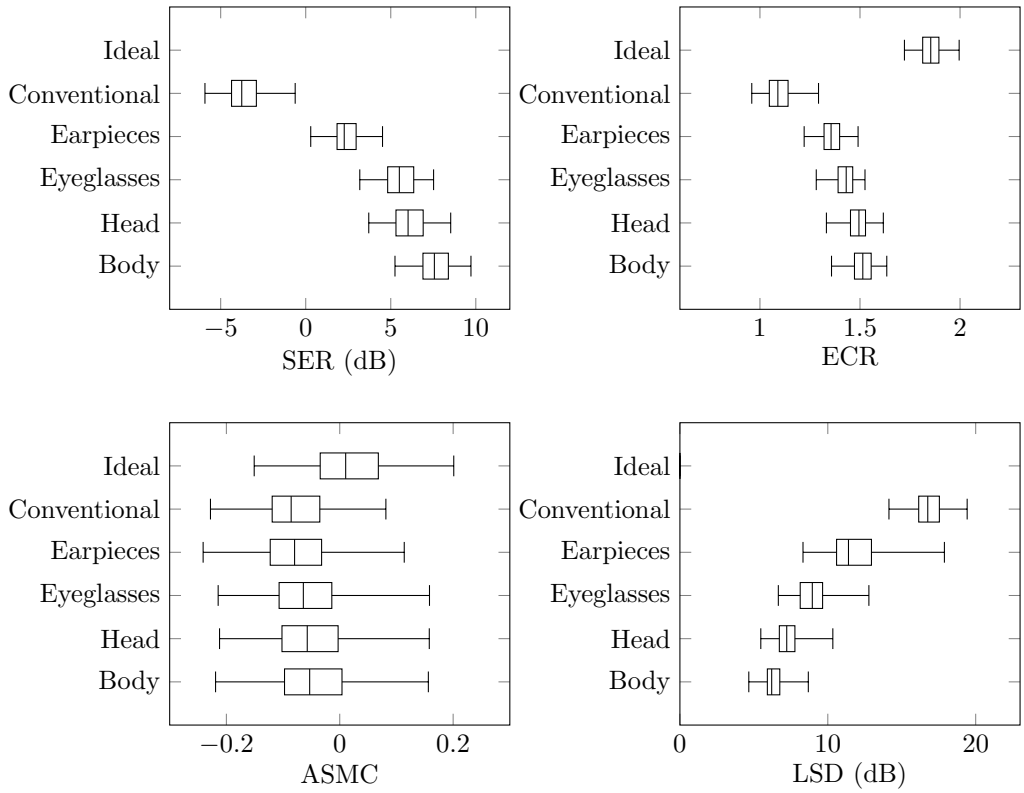

Figure 6.13: Multisource compression performance using different wearable microphone arrays. The plots show the quartiles of the metrics over 100 trials with different source configurations.

performance in the presence of multiple sources. For all experiments, at least one system achieved better noise reduction and compression performance than the conventional single-microphone system. Compression performance appears to be sensitive to the choice of distortion weights, with large weights prioritizing compression fidelity and small weights favoring noise reduction.

6.3.3 Experiments with different array configurations

The six-microphone earpiece array demonstrated in the previous section can improve performance with simple mixtures of two sources, but it did not work as well with five.

To realize the benefits of multisource compression for more challenging mixtures, we need to use larger arrays. The five-source experiment was repeated using the new wearable microphone data set [113] (Section 2.3). To simulate parameter mismatch, the impulse responses used to design the filters were truncated versions of those used to generate the mixtures.

Because the more powerful array is better able to separate sources, in this updated experiment the source covariance matrices were not updated between STFT frames as in (6.56); that is, the sources were separated based on their spatial characteristics alone. The target gains, however, were varied from frame to frame to implement the compression function. The source signals were taken from the VCTK corpus and randomly assigned to $N = 5$ of 24 directions of arrival for each of 100 trials. They were processed by either a conventional scalar compressor, a 6-microphone array comprising the ears and earpieces, a 16-microphone array on the ears and eyeglasses, a 32-microphone array including earpieces, glasses, headband, and collar, and a 64-microphone array covering the entire upper body.

The results are shown in Figure 6.13. The performance on all metrics improves as the number of microphones increases. The greatest performance improvement is between the conventional compressor and the earpiece array. There are diminishing returns as the array size increases. It appears that the performance bottleneck in this algorithm is the envelope estimation step. Because this experiment used a beamformer to recover the signal envelopes, any residual interference in the beamformer output contributes to across-source modulation. More sophisticated envelope-tracking algorithms will likely improve performance significantly.

6.3.4 Future directions

The analysis and experiments presented here show that dynamic range compression does not work well in noisy environments and that spatial signal processing can help improve its performance. There is more work to be done to characterize the interaction between noise and compression, especially with time-varying signals and nonlinear envelope-detection dynamics. The proposed multisource compression

system can also be improved with new source separation algorithms designed to estimate smooth envelopes rather than the signals themselves. The beamformers could also be implemented in the filterbank rather than STFT domain.

The multisource compression framework must also be studied from a clinical and psychoacoustic perspective. If all sound sources could be compressed separately, what kind of compression settings should be applied to each source? The sources need not be compressed completely independently. In music production, the compression gain applied to one source is sometimes based on the envelope of a second source:

$$v_{\mathbf{d}_n}[k,b] = \frac{\mathcal{C}_{b,n,p}\left(v_{\mathbf{c}_p}[k,b]\right)}{v_{\mathbf{c}_p}[k,b]} v_{\mathbf{c}_n}[k,b]. \tag{6.57}$$

This technique, known as *side-chain compression*, can be used to "duck" instruments when they would interfere with vocals or to create distortion in response to drum hits. A related technique was recently proposed to maintain a prescribed dialogue-to-background ratio in television broadcasts [91].

In general, we can compute target envelope powers for each source channel based on the envelopes of all N source channels, and possibly based on other information such as scene classification. For example, in a cocktail party setting, we might dynamically alter mixing ratios so that the listener hears as much background speech as possible while ensuring the intelligibility of a conversation partner. Ideally, such a time-varying objective system would be based on psychoacoustic models and personalized for each user.

The ability to compress multiple sources independently may be the single greatest benefit of large arrays for listening devices. Arrays will allow listening systems to reliably apply different compression characteristics to different sources, helping to improve the quality, intelligibility, and comfort of the resulting mixture. By applying compression independently, they will avoid the harmful distortion effects caused by across-source modulation. With further development, multisource compression could dramatically transform the way listening devices are designed and allow listeners to hear better even in challenging noisy environments.

Chapter 7

Time-Varying Space-Time Filters

The theory of space-time filtering developed in Chapters 3, 4, and 5 was for linear time-invariant filters. Such filters are appropriate when the inference problem is stationary. In the real world, however, acoustic sources and environments are constantly changing. We have already seen one example of a time-varying audio enhancement problem: in Chapter 6, the desired responses vary as a function of time to alter the dynamic range of the processed signals. In this chapter, we consider the problem in which the temporal statistics of the source signals change over time. A third source of nonstationarity, motion of sources and microphones, will be introduced in Chapter 9.

This chapter provides an overview of time-varying methods from the audio source separation and enhancement literature and considers their application to augmented listening. Because most of these methods are designed for small numbers of microphones and sound sources, they do not scale well in challenging environments or benefit from the spatial diversity afforded by the large arrays considered in this work. The author previously proposed two new models that can scale better with large numbers of sources and microphones: a hypothesis-testing source activity detector [168] and a high-low model for source activity [112]. New analysis considers the performance of binary masks and the proposed nonlinear spatial filters for the source-remixing application, including interaural cue preservation.

7.1 Time-Varying Methods

Most signals to which humans would care to listen, such as speech, music, animal sounds, and alarms, are nonstationary, meaning that their spectral distributions change over time. In conversational speech, for example, the spectrum of the speech signal changes several times per second. This variation is what allows sound to convey useful information, but it also makes audio signals more difficult to analyze mathematically. While we could take the Fourier transform of an entire sentence of speech, it would not accurately represent the spectral content at any given moment. Long-term average spectra are useful for large arrays that can process signals based on spatial diversity alone. In single-microphone systems and smaller arrays, however, we must use time-varying methods.

To analyze nonstationary sounds, engineers use time-frequency representations, especially the short-time Fourier transform (STFT). Recall from Section 3.2.2 that the STFT is a sequence of discrete Fourier transforms of overlapping windowed frames:

$$\mathbf{X}_{\mathrm{tf}}[k, f] = \sum_{\tau=-\infty}^{\infty} \mathrm{awin}(kT_{\mathrm{step}} - \tau)\mathbf{x}_{\mathrm{d}}[\tau]e^{-j2\pi\tau f/F}, \tag{7.1}$$

where $\mathrm{awin}(\tau)$ is a tapered analysis window.

For the remainder of this chapter, all signals will be in the STFT domain and all filters will be in the discrete frequency domain. The subscripts tf and df will be omitted.

7.1.1 Sparsity and W-disjoint orthogonality

The STFT allows us to visualize the time-varying statistics of speech and other natural sounds. But the STFT is more than just a visualization tool: it turns out that the STFT is a sparse basis for these sound signals. That is, when we take the STFT of a speech signal, most of the energy of the signal is concentrated in a small fraction of time-frequency samples. The degree of sparsity depends on the frame length used for the transform: for typical speech signals, the STFT is most sparse

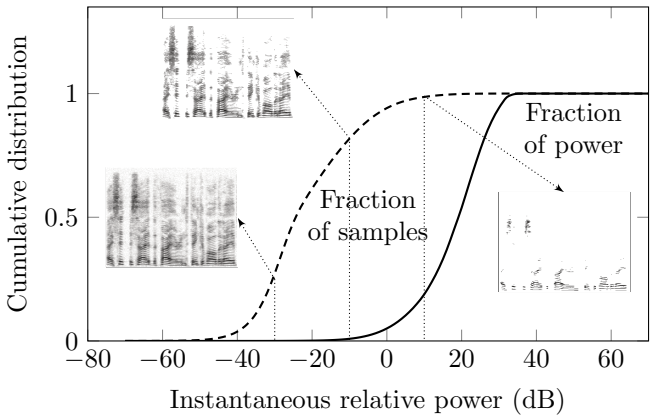

Figure 7.1: A small fraction of time-frequency indices contain most of the energy in a speech signal. Each inset shows the spectrogram with lower-energy samples removed. Figure adapted from [168].

with a frame length of about 60 ms [57, 169].

Figure 7.1 shows the concentration of energy in the STFT of a speech recording. For this sample, if all time-frequency samples with instantaneous power less than 10 dB below the long-term average are removed (top inset), the remaining 18% of samples contain 99% of the signal energy and the speech signal has perceptual quality comparable to that of the original. The strongest 1% of signal samples (bottom right inset) contain 77% of the signal energy. The signal reconstructed from these samples is severely distorted but still intelligible.

Sparsity is a useful property in its own right: for example, the STFT is often used for feature extraction in machine learning methods such as speech recognition. But parsimony alone is not the reason that nearly every speech separation algorithm uses the STFT. Different speech signals, especially if they are from different talkers, tend to concentrate their energy in different time-frequency samples. This property, known as *W-disjoint orthogonality* [57, 170], ensures that for speech mixtures with a few different talkers, it is rare that more than one speech signal has significant energy at the same time and the same frequency.

W-disjoint orthogonality can be stated mathematically as follows. Let $S_n[k, f]$ be the STFT of source signal n for source channels $n = 1, \ldots, N$. Then at every

time-frequency index $[k, f]$ there exists an $n^*[k, f] \in \{1, \ldots, N\}$ such that

$$\left|S_{n^*[k,f]}[k, f]\right|^2 \gg \left|S_n[k, f]\right|^2, \quad n \neq n^*[k, f]. \tag{7.2}$$

Assuming that the acoustic channels are not too different between source channels, we can then approximate the observed mixture by the source image of the dominant source channel at each time-frequency index:

$$\mathbf{X}[k, f] \approx \mathbf{C}_{n^*[k,f]}[k, f]. \tag{7.3}$$

7.1.2 Time-frequency masks

The observation that speech signals overlap little in the STFT domain gave rise to the method that has dominated audio source separation research for more than a decade: the *time-frequency mask*. The idea follows directly from (7.3): to recover most of the energy of source image n, keep all the observed signal samples $\mathbf{X}[k, f]$ for which $n^*[k, f] = n$ and discard the rest:

$$\mathbf{C}_n[k, f] \approx \begin{cases} \mathbf{X}[k, f], & \text{if } n^*[k, f] = n, \\ 0, & \text{otherwise.} \end{cases} \tag{7.4}$$

Binary mask

This estimation procedure can also be interpreted as a time-varying scalar STFT-domain filter:

$$\hat{\mathbf{C}}_n[k, f] = W_n[k, f]\mathbf{X}[k, f], \tag{7.5}$$

where the mask $W_n[k, f]$ is given by

$$W_n[k, f] = \begin{cases} 1, & \text{if } n^*[k, f] = n \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

Because this scalar filter takes values of either 0 or 1, it is known as a binary mask.

The *ideal binary mask* [171], generated from prior knowledge of the source images, is widely used as an oracle estimator to benchmark blind source separation algorithms [169]. It works well as long as the sources truly are sparse and there are only a few sources in the mixture. In practice, however, the mask must be estimated by a classifier. When the classifier makes errors, each estimated source signal is mixed with haphazardly distributed time-frequency samples from other sources. These erroneous samples do not necessarily obey the temporal or harmonic structure of their original source signal, so they can produce unnatural and disturbing distortion.

Soft mask

Time-frequency masks need not be binary. A *soft mask* is one that takes values between 0 and 1. Such masks can achieve higher performance with fewer artifacts, but they cannot be computed with a simple hard-decision classification. One popular choice is the estimated posterior probability:

$$W_n[k,f] = \Pr\left(n^*[k,f] = n \mid \mathbf{X}[k,f]\right), \quad n = 1,\ldots,N. \tag{7.7}$$

Another is the scalar Wiener filter, which uses estimates $\hat{R}_{S_n}[k,f]$ of the instantaneous source variance:

$$W_n[k,f] = \frac{\hat{R}_{S_n}[k,f]}{\sum_{m=1}^{N} \hat{R}_{S_m}[k,f]}, \quad n = 1,\ldots,N. \tag{7.8}$$

These soft masks are similar to the binary mask in that the masks for the different source channels sum to unity.

7.1.3 Classification methods

Much of the audio source separation research from the past decade has focused on methods to separate speech mixtures using time-frequency masks. These methods

differ primarily in how they generate masks.

Single-microphone classification methods

Time-frequency masks are often applied to single-microphone source separation. With no spatial information, these classifiers must rely exclusively on the properties of the sound signals themselves. If the sources are dissimilar, such as different musical instruments, they can be separated using compositional models, which decompose the magnitude spectra into different components that are assigned to different sources [58]. For example, nonnegative matrix factorization decomposes an $F \times K$ matrix of STFT magnitudes into two low-rank matrices of spectral patterns and temporal activation sequences [172,173]. Data-driven methods, meanwhile, use classifiers trained on large data sets to assign source labels to each time-frequency sample. Recently, many researchers have applied deep neural networks to classify time-frequency samples [37,152,174]. Deep learning methods have also been combined with nonnegative matrix factorization [59]. In the recent Signal Separation Evaluation Campaign (SiSEC 2018), learning-based methods outperformed oracle binary mask estimators in separating vocal and instrument tracks from music recordings [175].

Multimicrophone classification methods

Compositional models are useful for separating different types of signals: speech from music, for example, or guitar from drums. While there have been promising recent results in separating multiple speech signals using elaborate machine learning models, in general it is difficult to separate two similar types of signals using single-microphone methods. If a microphone array is available, and if the two similar signals arrive from different directions, then spatial information can be used to classify the dominant source at each time-frequency index.

The celebrated degenerate unmixing estimation (DUET) method, one of the first methods to use binary masks for source separation, is inspired by the human auditory system: it assigns each time-frequency index to a single source channel based on phase

and level differences between a pair of microphones [57, 176]. It has been extended to multiple microphones using subspace methods [177, 178] and pairwise time and level differences [179, 180]. The masks can also be soft, meaning that they apply a gain between 0 and 1 rather than a binary value.

Masks and array processing

Time-frequency-masking methods, even those that use multiple microphones for classification, do not take full advantage of the spatial diversity afforded by microphone arrays. A simple way to combine spatial processing with mask-based processing is to apply a scalar time-frequency mask to the output of a time-invariant beamformer. For example, a single-target time-varying multichannel Wiener filter can be implemented using a time-invariant MVDR beamformer and a time-varying scalar Wiener postfilter. Binary masks are also a popular choice for postfiltering.

The W-disjoint orthogonality property only applies to mixtures of a few sources. If there are many sources present, or if not all the source signals are sparse, then there may be more than one active source at each time-frequency index. With arrays, the sparse model can be extended to account for more than one active source at each time-frequency index [111, 181–184].

7.1.4 The local Gaussian model

Time-frequency masks are useful for mixtures of small numbers of sparse sources, but they do not work as well when there is significant overlap between source spectra. They also do not benefit from all the degrees of freedom afforded by large microphone arrays. Instead of assigning each time-frequency index to a single source, we can leverage sparsity by incorporating time-varying source statistics into the design of a space-time filter. This method is often referred to as the *local Gaussian model* [11].

Under the local Gaussian model, the covariance $\mathbf{R}_{\mathbf{C}_n}[k, f]$ of each source image STFT is an unknown parameter, which could be random or nonrandom, and the

time-frequency samples of each source image are *conditionally* complex Gaussian random variables:

$$\mathbf{C}_n[k,f] \mid \mathbf{R}_{\mathbf{C}_n}[k,f] \sim \mathcal{CN}\left(\mathbf{0}, \mathbf{R}_{\mathbf{C}_n}[k,f]\right). \tag{7.9}$$

These samples are also usually assumed to be *conditionally* independent of each other across time, frequency, and source channels. This conditionality is important because the marginal distribution of most interesting sounds in the STFT domain is not Gaussian and the time-frequency samples of a given source image are not independent of each other.

If the time-varying covariance matrices are known, then the minimum-weighted-mean-square-error estimate of the desired output is the MSDW-MWF designed using those covariances:

$$\hat{\mathbf{Y}}[k,f] = \sum_{n=1}^{N} \lambda_n \mathbf{G}_n[k,f] \mathbf{R}_{\mathbf{C}_n}[k,f] \left(\sum_{m=1}^{N} \lambda_m \mathbf{R}_{\mathbf{C}_m}[k,f] \right)^{-1} \mathbf{X}[k,f]. \tag{7.10}$$

It would seem that we have not gained much by applying the local Gaussian model: at each time-frequency index, instead of estimating the MN complex values of $\mathbf{C}_n[k,f]$ for $n = 1, \ldots, N$, we must estimate the M^2N complex values of $\mathbf{R}_{\mathbf{C}_n}[k,f]$. Therefore, the local Gaussian model is often combined with constraints on the possible values of $\mathbf{R}_{\mathbf{C}_n}$. For example, if the signal due to source channel n is relatively spatially coherent and if the sound source and microphones do not move, then $\mathbf{R}_{\mathbf{C}_n}[k,f]$ can be modeled as the product of a fixed spatial covariance matrix $\bar{\mathbf{R}}_n[f]$ and a time-varying source signal variance $R_{S_n}[k,f]$ [131]:

$$\mathbf{R}_{\mathbf{C}_n}[k,f] = R_{S_n}[k,f] \bar{\mathbf{R}}_n[f], \quad n = 1, \ldots, N. \tag{7.11}$$

Now we need only estimate N nonnegative parameters at each time-frequency index.

These parameters can be estimated using models of the underlying source signals $S_n[k,f]$. For example, compositional models such as nonnegative matrix factorization [173] are useful for mixtures of dissimilar sources, such as speech and music.

7.2 Source Activity Mask

Binary time-frequency masks are usually used to assign each time-frequency index to a single source channel according to the W-disjoint orthogonality model. The masks created for each source channel are then disjoint. With this method, mixtures of few sources will have dense masks applied to each source, while mixtures of many sources will have sparse masks. As illustrated in Figure 7.1, the density of the time-frequency mask affects the distortion applied to the target source. There is a tradeoff between distortion and interference rejection. As with the speech distortion weights introduced to space-time filters (Section 3.5.3), we can make this tradeoff explicit by applying a tuning parameter to our time-frequency mask.

To better control the tradeoff between interference and distortion, and to apply masks to mixtures with more than a few competing sound sources, the author proposed replacing the dominant-source classifier with a source activity detector [168], similar to the voice activity detectors often used in speech recognition [185]. That is, instead of asking "which source is dominant at $[k, f]$?" we can ask "is source n active or inactive at $[k, f]$?" The ideal mask acts as a *source activity detector*, that is, it takes the value 1 when the signal is large and 0 when it is small:

$$W_n[k, f] = \begin{cases} 1, & \text{if } |S_n[k, f]|^2 \geq \gamma[f] \\ 0, & \text{otherwise,} \end{cases} \tag{7.12}$$

where γ is a tuning parameter that determines the tradeoff between distortion and interference rejection. The advantage of this definition is that it is independent of the other sources in a mixture, so it is suitable for mixtures of large numbers of sources. Based on informal listening tests, $\gamma[f]$ should be on the order of the long-term average power of the signal at each frequency. Larger values provide greater interference rejection but more distortion of the source signal.

7.2.1 Generalized likelihood ratio test

To generate a source activity mask, we can no longer use the N-state classifiers introduced in the previous section. Instead, for each source channel n for which we wish to generate a mask, we must solve a hypothesis testing problem:

$$\mathcal{H}_1 : \mathbf{X}[k,f] = \mathbf{C}_n[k,f] + \mathbf{Z}[k,f] \tag{7.13}$$

$$\mathcal{H}_0 : \mathbf{X}[k,f] = \mathbf{Z}[k,f], \tag{7.14}$$

where $\mathbf{Z}[k,f] = \sum_{m \neq n} \mathbf{C}_m[k,f]$ contains all other interference and noise signals. If the statistics of $\mathbf{Z}[k,f]$ were known, then we could use a generalized likelihood ratio test, which treats $\mathbf{C}_n[k,f]$ as a nonrandom unknown parameter. The test statistic is

$$T_n(\mathbf{X}[k,f]) = \log \frac{\sup_{\mathbf{C} \neq \mathbf{0}} \Pr(\mathbf{X}[k,f] | \mathbf{C}_n[k,f] = \mathbf{C})}{\Pr(\mathbf{X}[k,f] | \mathbf{C}_n[k,f] = 0)}. \tag{7.15}$$

Using a rank-1 model for $\mathbf{C}_n[k,f]$ and a full-rank Gaussian model for $\mathbf{Z}[k,f]$, the likelihood ratio is

$$T_n(\mathbf{X}[k,f]) = \frac{1}{2} \frac{\left| \mathbf{A}_n^H[f] \mathbf{R}_{\mathbf{Z}}^{-1}[f] \mathbf{X}[k,f] \right|^2}{\mathbf{A}_n^H[f] \mathbf{R}_{\mathbf{Z}}^{-1}[f] \mathbf{A}_n[f]}. \tag{7.16}$$

The mask is given by

$$W_n[k,f] = \begin{cases} 1, & \text{if } T_n(\mathbf{X}[k,f]) \geq \bar{\gamma}[f] \\ 0, & \text{otherwise,} \end{cases} \tag{7.17}$$

where $\bar{\gamma}[f]$ is a tuning parameter that controls the tradeoff between probability of missed detection (target distortion) and probability of false alarm (interference). It is related to $\gamma[f]$ from (7.12) but also considers the noise distribution.

Notice that the statistic is the signal-to-noise ratio at the output of an MVDR beamformer in the direction of $\mathbf{A}_n$. For a large array that can perfectly suppress interference sources, the mask will be quite dense. For smaller arrays, it will be more conservative. Figure 7.2 shows the receiver operating characteristic (ROC) curve parameterized by $\bar{\gamma}$ for a speech signal in stationary noise.

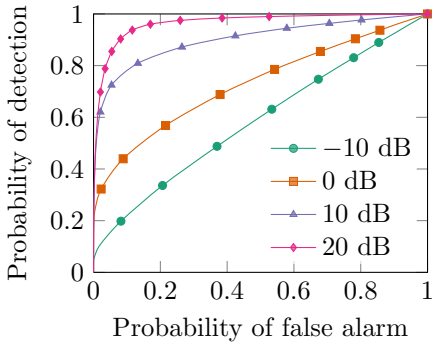

Figure 7.2: Experimental ROC curve for detection of a speech signal in white Gaussian noise at different input SNRs. Figure adapted from [168].

7.2.2 Nonstationary hypothesis test

The generalized likelihood ratio test is appropriate for detecting a nonstationary signal in stationary noise with known statistics. In speech mixtures, however, both the target signal and interference signals are nonstationary. Therefore, the author proposed to use multiple hypothesis tests, one for each interference source channel [168]. Each test compares the hypothesis that target source channel n and interference source channel m are both present against the hypothesis that only interference channel m is present:

$$\mathcal{H}_{1,m} : \mathbf{X}[k,f] = \mathbf{C}_n[k,f] + \mathbf{C}_m[k,f] + \mathbf{Z}_0[k,f] \tag{7.18}$$

$$\mathcal{H}_{0,m} : \mathbf{X}[k,f] = \mathbf{C}_m[k,f] + \mathbf{Z}_0[k,f], \tag{7.19}$$

where $\mathbf{Z}_0[k,f]$ is diffuse stationary noise that ensures the mixture covariance has full rank. To pass the overall hypothesis test, the sample must pass all $N-1$ of these pairwise hypothesis tests for $m \neq n$. That is, the test statistic is

$$T_n(\mathbf{X}) = \frac{1}{2} \min_{m \neq n} \frac{\left|\mathbf{A}_n^H[f](\mathbf{R}_{\mathbf{C}_m}[f] + \mathbf{R}_{\mathbf{Z}_0}[f])^{-1}[f]\mathbf{X}[k,f]\right|^2}{\mathbf{A}_n^H[f](\mathbf{R}_{\mathbf{C}_m}[f] + \mathbf{R}_{\mathbf{Z}_0}[f])^{-1}\mathbf{A}_n[f]}. \tag{7.20}$$

The outcome of the hypothesis test will be most affected by source channels m that are difficult to separate from the target channel n.

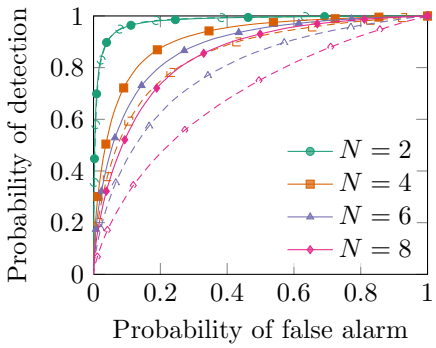

Figure 7.3: ROC curves for source activity detection in speech mixtures. The dashed curves show the generalized likelihood ratio test (7.16) and the solid curves show the nonstationary hypothesis test (7.20). Figure adapted from [168].

Figure 7.3 shows the ROC curves for this nonstationary hypothesis test compared to a conventional generalized likelihood ratio test for mixtures of different numbers of speech signals. For small numbers of sources, the two methods have similar performance, but for mixtures of many sources, the nonstationary hypothesis test outperforms the stationary test. Further results and experimental details are available in [168].

7.3 High-Low Space-Time Filter

Time-frequency masks and the full-rank local Gaussian model represent opposite extremes in terms of separation power and computational complexity. A time-varying space-time filter can apply all its spatial degrees of freedom and also take advantage of signal sparsity to improve separation performance. Such filters are unsuitable for arrays with many microphones, however, because they must estimate too many parameters and perform expensive computations to calculate an entirely new filter at each time-frequency index. Masks, meanwhile, must make only one decision at each time-frequency index and they are trivial to apply to the signal, but they do not scale well to large numbers of sources and microphones.

In this section, a compromise method is proposed. The *high-low model*, like the

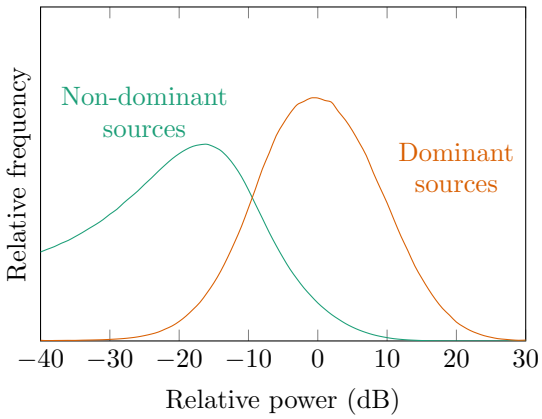

Figure 7.4: Distribution of dominant and non-dominant instantaneous time-frequency sample powers for a mixture of eight quasi-anechoic speech signals. The mean overall power of the signal is 0 dB.

W-disjoint orthogonality model, has only one state parameter at each time-frequency index. However, it still includes all N signals at all indices, whether they are active or not. Thus, when the sources are easily separated spatially, for example if the array is large, it reduces to a conventional space-time filter. When they are more difficult to separate spatially, it behaves more like a mask. Variations of this model have appeared in several previous works by the author [111, 112, 186].

7.3.1 High-low model

The high-low model is similar to other local Gaussian models in that it assumes that the covariance of source image time-frequency sample $\mathbf{C}_n[k, f]$ is the product of a time-invariant spatial covariance matrix $\bar{\mathbf{R}}_n[f]$ and a time-varying scalar source variance $R_{S_n}[k, f]$:

$$\mathbf{R}_{\mathbf{C}_n}[k, f] = R_{S_n}[k, f]\bar{\mathbf{R}}_n[f], \quad n = 1, \ldots, N. \tag{7.21}$$

However, under the high-low model, $R_{S_n}[k, f]$ can take only one of two values:

$$R_{S_n}[k, f] = \begin{cases} R_{\text{high},n}[f], & \text{if } n^*[k, f] = n, \\ R_{\text{low},n}[f], & \text{otherwise,} \end{cases} \tag{7.22}$$

where $R_{\text{high},n}[f]$ and $R_{\text{low},n}[f]$ are the assumed source spectra when source channel n is dominant and non-dominant, respectively.

The high-low model is a generalization of the "on-off" model implicit in binary masks. The behavior of the filter depends on the relative values of $R_{\text{low},n}[f]$ and $R_{\text{high},n}[f]$. If $R_{\text{low},n}[f] = 0$ for all n, then it is equivalent to the W-disjoint orthogonality model (7.3). If $R_{\text{high},n}[f] = R_{\text{low},n}[f]$ for all n, then the source images are modeled as wide-sense stationary and the resulting filter is time-invariant.

In the experiments presented in this work, $R_{\text{high},n}[f]$ and $R_{\text{low},n}[f]$ are 20 dB apart for all speech sources. This choice is based on empirical experiments with mixtures of speech signals. Figure 7.4 shows the histogram of instantaneous time-frequency sample power relative to long-term average power for dominant and non-dominant sources in a mixture of eight speech signals from the VCTK corpus [104]. That is, the curve on the right shows the histogram of $\left|S_{n^*[k,f]}[k, f]\right|^2 / \text{mean}_k \left|S_n[k, f]\right|^2$ and the curve on the left shows the histogram of $\left|S_n[k, f]\right|^2 / \text{mean}_k \left|S_n[k, f]\right|^2$ for $n \neq n^*[k, f]$.

The method can be easily adapted for more than one simultaneous "high" source, but a previous study found only marginal improvement using two sources instead of one [111] at a cost of much higher computational complexity. Such a model would be most useful for large spaces with many distributed sources and arrays, in which the system would need to use more complex filtering methods anyway.

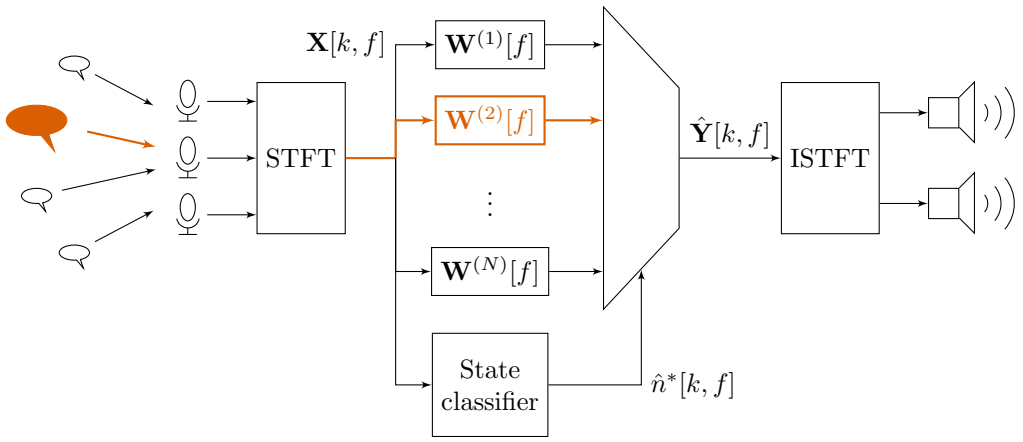

Figure 7.5: Using the high-low model, the time-varying space-time filter switches between several time-invariant filters.

7.3.2 Discrete-state space-time filter

The STFT-domain MSDW-MWF for the local Gaussian model with scalar source variances and time-invariant desired responses is

$$\mathbf{W}[k,f] = \sum_{n=1}^{N} \lambda_n R_{S_n}[k,f]\mathbf{G}_n[f]\bar{\mathbf{R}}_n[f] \left(\sum_{n=1}^{N} \lambda_n R_{S_n}[k,f]\bar{\mathbf{R}}_n[f] \right)^{-1}. \tag{7.23}$$

This filter can take one of N values depending on the "high" source:

$$\mathbf{W}[k,f] = \mathbf{W}^{(n^*[k,f])}[f] \quad \text{where} \tag{7.24}$$

$$\mathbf{W}^{(n)}[f] = \left(\lambda_n R_{\text{high},n}[f]\mathbf{G}_n[f]\bar{\mathbf{R}}_n[f] + \sum_{m\neq n} \lambda_m R_{\text{low},m}[f]\mathbf{G}_m[f]\bar{\mathbf{R}}_m[f] \right) \times \left(\lambda_n R_{\text{high},n}[f]\bar{\mathbf{R}}_n[f] + \sum_{m\neq n} \lambda_m R_{\text{low},m}[f]\bar{\mathbf{R}}_m[f] \right)^{-1}. \tag{7.25}$$

Thus, the filtering algorithm could be implemented by switching between N time-invariant filters based on the output of a classifier, as shown in Figure 7.5. Notice that the high and low variances are always multiplied with the speech distortion

weights. They play a similar role in allocating the degrees of freedom of the array toward particular sources; in fact, an earlier version of the high-low model used high and low distortion weights rather than high and low variances [112]. The two formulations are mathematically identical and differ only in interpretation.

The advantage of the high-low filter is that it devotes its degrees of freedom to the sources that are most active, but it does not completely ignore inactive sources. To illustrate this behavior, let us return to our favorite running example from Chapter 3, the single-target beamformer with $\bar{\mathbf{R}}_1[f] = \mathbf{A}_1[f]\mathbf{A}_1^H[f]$, $\mathbf{G}_1[f] = \mathbf{e}_1^T$ and $\mathbf{G}_2[f] = \cdots = \mathbf{G}_N[f] = \mathbf{0}$. If source channel 1 is the dominant source channel at index $[k, f]$, that is, if $n^*[k, f] = 1$, we have

$$\mathbf{W}[k,f] = \lambda_1 R_{\text{high},1}[f]\mathbf{e}_1^T\mathbf{A}_1[f]\mathbf{A}_1^H[f]\left(\lambda_1 R_{\text{high},1}[f]\mathbf{A}_1[f]\mathbf{A}_1^H[f] + \sum_{n=2}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f]\right)^{-1} \tag{7.26}$$

$$= \mathbf{e}_1^T\mathbf{A}_1[f]\frac{\mathbf{A}_1^H[f]\left(\sum_{n=2}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f]\right)^{-1}}{\lambda_1^{-1}R_{\text{high},1}^{-1}[f] + \mathbf{A}_1^H[f]\left(\sum_{n=2}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f]\right)^{-1}\mathbf{A}_1[f]} \tag{7.27}$$

$$\approx \mathbf{e}_1^T\mathbf{A}_1[f]\frac{\mathbf{A}_1^H[f]\left(\sum_{n=2}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_1[f]\right)^{-1}}{\mathbf{A}_1^H[f]\left(\sum_{n=2}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_1[f]\right)^{-1}\mathbf{A}_1[f]}. \tag{7.28}$$

Because the assumed power of source image 1 is much higher than that of the others, the filter is approximately an MVDR beamformer.

Now suppose that $n^*[k, f] = 2$ so that an interference source is dominant. Then we have

$$\mathbf{W}[k,f] = \frac{\mathbf{e}_1^T\mathbf{A}_1[f]\mathbf{A}_1^H[f]\left(\lambda_2 R_{\text{high},2}[f]\bar{\mathbf{R}}_2[f] + \sum_{n=3}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f]\right)^{-1}}{\lambda_1^{-1}R_{\text{low},1}^{-1}[f] + \mathbf{A}_1^H[f]\left(\lambda_2 R_{\text{high},2}[f]\bar{\mathbf{R}}_2[f] + \sum_{n=3}^{N}\lambda_n R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f]\right)^{-1}\mathbf{A}_1[f]}. \tag{7.29}$$

If the noise source is diffuse so that $\bar{\mathbf{R}}_2[f]$ has full rank, then

$$\mathbf{W}[k,f] \approx \mathbf{e}_1^T \mathbf{A}_1[f] \frac{\lambda_1 R_{\text{low},1}[f]}{\lambda_2 R_{\text{high},2}[f]} \mathbf{A}_1^H[f] \bar{\mathbf{R}}_2^{-1}[f], \tag{7.30}$$

and the filter strongly attenuates the signal, much like a mask. If, however, the interference source is also directional, the filter can direct a null toward source channel 2 without strongly attenuating source channel 1.

For a large array with $M > N$, if all of the nonstationary source channels have rank-1 models and if the diffuse noise is much weaker than the directional sources, the discrete-state MSDW-MWF degenerates to a time-invariant LCMV beamformer. This result follows the same derivation as the large-distortion-weight limit from Section 3.5.3. In this case, the state of each source channel is irrelevant because they can be perfectly separated using spatial information.

At the other extreme, if $M = 1$, the discrete-state filter becomes

$$W[k,f] = \begin{cases} \frac{\lambda_1 R_{\text{high},1}[f]}{\lambda_1 R_{\text{high},1}[f] + \sum_{n=2}^{N} \lambda_n R_{\text{low},n}[f]}, & \text{if } n^*[k,f] = 1 \\ \frac{\lambda_1 R_{\text{low},1}[f]}{\lambda_{n^*[k,f]} R_{\text{high},n^*[k,f]}[f] + \sum_{n \neq n^*[k,f]} \lambda_n R_{\text{low},n}[f]} & \text{otherwise.} \end{cases} \tag{7.31}$$

This filter is a soft mask that is close to 1 when the target source is dominant and close to 0 when it is not.

These limiting cases show that the high-low model scales well with different array sizes. In underdetermined or single-microphone mixtures, it takes advantage of source sparsity and resembles a time-frequency mask. If there is ample spatial diversity so that sources can be separated by their spatial characteristics alone, then it does not need to use sparsity and resembles a linearly constrained time-invariant filter. Between these two extremes, it takes advantage of signal sparsity without ignoring inactive sources.

7.3.3 Maximum likelihood state estimation

The high-low filter could be used with any classifier, including state-of-the-art compositional models and machine-learning methods. However, since the focus of this dissertation is on array processing, the experiments in this chapter use a multimicrophone classifier that does not require complex models of the source signals.

Let us treat $n^*[k,f]$ as an unknown parameter to be estimated. Under the local Gaussian high-low model, the log-likelihood of the observation $\mathbf{X}[k,f]$ given $n^*[k,f] = n^*$ is

$$\begin{aligned}\ln p_{n^*}[k,f] = &-\mathbf{X}^H[k,f]\left(R_{\text{high},n^*}[f]\bar{\mathbf{R}}_{n^*}[f] + \sum_{m\neq n^*} R_{\text{low},m}[f]\bar{\mathbf{R}}_m[f]\right)^{-1}\mathbf{X}[k,f] \\ &- \ln\det\left(\pi R_{\text{high},n^*}[f]\bar{\mathbf{R}}_{n^*}[f]\sum_{m\neq n^*} R_{\text{low},m}[f]\bar{\mathbf{R}}_m[f]\right), \quad n^* = 1,\ldots,N.\end{aligned} \tag{7.32}$$

The maximum likelihood state estimate $\hat{n}^*[k,f]$ is the n^* that maximizes (7.32) at each $[k,f]$.

A simpler alternative uses the rank-1 on-off model to compute a maximum likelihood estimate [178]:

$$\begin{aligned}\ln p_{n^*}[k,f] = &-\frac{\left\|\mathbf{X}[k,f] - \frac{\mathbf{A}_{n^*}[f]\mathbf{A}_{n^*}^H[f]}{\mathbf{A}_{n^*}^H[f]\mathbf{A}_{n^*}[f]}\mathbf{X}[k,f]\right\|^2}{R_{\text{high},n^*}[f]} \\ &- \ln\det\left(\pi R_{\text{high},n^*}\mathbf{A}_{n^*}[f]\mathbf{A}_{n^*}^H[f]\right),\end{aligned} \tag{7.33}$$

for $n^* = 1,\ldots,N$. If the source channels have similar high and low variances, then this is effectively a nearest-neighbor classifier based on the RTFs.

Table 7.1: Computational complexity for each time-frequency sample.

	Filter multiplications	Estimator/classifier
Static filter	M	None
Binary mask	1	Hard-decision
Soft mask	N	Soft-decision
Discrete-state filter	M	Hard-decision
LGM filter	NM^2	Soft-decision

7.3.4 Computational complexity

Time-varying methods can offer better remixing performance for challenging mixtures, but they can also be computationally demanding. Table 7.1 compares the computational requirements of several STFT-domain remixing methods. The short-time Fourier transform and its inverse can be efficiently implemented using the fast Fourier transform, which has complexity $F \log F$.

A time-invariant space-time filter performs M complex multiply-accumulates to generate each time-frequency output sample. A binary mask is trivial to implement: it multiplies each sample by $G_{n^*[k,f]}[f]$. A soft mask requires N complex multiplications for each output sample. The discrete-state filter has the same amortized complexity as the time-invariant filter because the N possible sets of filter coefficients can be computed in advance and only one is used for each sample. Local Gaussian models that allow arbitrary source variances have the highest filtering complexity because the filter itself must be recomputed for each sample.

Most of the complexity of time-varying methods comes from the classifiers and estimators used to update the filter or mask coefficients. The binary mask and discrete-state classifier are easiest to implement because they make hard decisions. The maximum likelihood classifier (7.32) requires N quadratic multiplications with an $M \times M$ matrix. The nearest-neighbor classifier is less demanding. Soft masks and local-Gaussian-model-based filters require estimating the posterior probabilities of source activity or the instantaneous variance of each source channel. The composi-

tional models, expectation-maximization algorithms, or machine-learning classifiers that are typically used with these methods tend to dominate the complexity of the system.

Complex time-varying methods are most appropriate for systems with few microphones. They become prohibitively expensive for large M. However, in systems with large spatial diversity, there is less need for sparse signal models.

7.4 Time-Varying Filters for Augmented Listening

Time-varying methods were developed for underdetermined systems with more sources than sensors. Because not all sources must be considered at all time-frequency indices, time-varying methods can use their available degrees of freedom more efficiently. In underdetermined source separation, these are used to apply constraints to more source channels. In augmented listening systems with extra perceptual constraints, those degrees of freedom can also be applied to reducing spectral and spatial distortion, improving robustness to parameter mismatch and motion, and, perhaps, to reducing delay. These advantages could be useful even in nominally overdetermined scenarios with large wearable microphone arrays. This section is an extension of the author's work in [112] with new mathematical analysis.

While time-varying methods are attractive for augmented listening applications, they must be applied with caution: they are more computationally demanding than time-invariant methods and they can cause disturbing distortion when they perform poorly. Time-frequency methods also have inherently large delay due to the frame-based processing of the STFT.

7.4.1 Source remixing with time-frequency masks

There have been many proposals to apply time-frequency masks to listening devices. Setting aside the issue of algorithmic delay, how effective are time-frequency masks in a source-remixing application?

Let $W_n[k, f]$ be any scalar mask designed to isolate source channel n from the mixture. Applying the desired remixing responses, the overall array output is

$$\hat{\mathbf{Y}}[k, f] = \sum_{n=1}^{N} W_n[k, f]\mathbf{G}_n[f]\mathbf{X}[k, f]. \tag{7.34}$$

Thus, the overall space-time filter at each time-frequency index is

$$\mathbf{W}[k, f] = \sum_{n=1}^{N} W_n[k, f]\mathbf{G}_n[f]. \tag{7.35}$$

Spectral distortion

For any source image sample $\mathbf{C}_n[k, f]$, the output image is

$$\hat{\mathbf{D}}_n[k, f] = \sum_{m=1}^{N} W_m[k, f]\mathbf{G}_m[f]\mathbf{C}_n[k, f]. \tag{7.36}$$

For a binary mask, the processing applied to every source image is the desired response of the dominant source:

$$\hat{\mathbf{D}}_n[k, f] = \mathbf{G}_{n^*[k,f]}[f]\mathbf{C}_n[k, f], \quad n = 1, \ldots, N. \tag{7.37}$$

The overall error covariance is then

$$\mathbf{R}_{\text{err}}[k, f] = \sum_{n=1}^{N} \left(\mathbf{G}_{n^*[k,f]}[f] - \mathbf{G}_n[f]\right) \mathbf{R}_{\mathbf{C}_n}[k, f] \left(\mathbf{G}_{n^*[k,f]}[f] - \mathbf{G}_n[f]\right)^H. \tag{7.38}$$

If the classifier works perfectly, then this filter should still be perceived as transparent by the listener: in the auditory system, the strongest sound source "masks" other sounds at the same time and frequency, so the processing applied to those weaker sources will not be perceived. If the classifier makes errors, however, the distortion could be severe. A recent perceptual study has found that source-remixing algorithms

for music have higher perceptual quality—fewer artifacts and less distortion—when the gains applied to the different sources are similar [187].

Spatial distortion

Now let us consider the spatial distortion of a binaural source-remixing mask with $\mathbf{G}_n[f] = G_n[f] \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix}^T$ for $n = 1, \ldots, N$. For any source image sample $\mathbf{C}_n[k, f]$, the output image is

$$\hat{\mathbf{D}}_n[k, f] = \sum_{m=1}^{N} W_m[k, f] \mathbf{G}_m[f] \mathbf{C}_n[k, f] \tag{7.39}$$

$$= \left(\sum_{m=1}^{N} W_m[k, f] G_m[f] \right) \begin{bmatrix} \mathbf{e}_2^T \mathbf{C}_n[k, f] \\ \mathbf{e}_1^T \mathbf{C}_n[k, f] \end{bmatrix}. \tag{7.40}$$

Because the same scalar processing is applied at both ears, there is no distortion to the interaural transfer function:

$$\mathrm{ITF}_n^{\mathrm{out}}[k, f] = \frac{\mathbf{e}_2^T \mathbf{C}_n[k, f]}{\mathbf{e}_1^T \mathbf{C}_n[k, f]} = \mathrm{ITF}_n^{\mathrm{in}}[k, f]. \tag{7.41}$$

Scalar time-frequency masks are perfectly spatially transparent! However, they can introduce severe spectral distortion if the classifier does not work well or if there are so many sound sources that W-disjoint orthogonality does not apply.

7.4.2 Spectral and spatial distortion of the high-low filter

Now let us consider the spectral and spatial distortion of the high-low filter. Consider the response of the filter to rank-1 source channel n. From (4.24), the response is

$$\mathbf{W}[k, f] \mathbf{A}_n[f] = \mathbf{G}_n[f] \mathbf{A}_n[f] - \sum_{m=1}^{N} \lambda_m \left(\mathbf{G}_n[f] - \mathbf{G}_m[f] \right) R_{S_m}[k, f] \bar{\mathbf{R}}_m[f] \bar{\mathbf{R}}_{\mathbf{x}}^{-1}[k, f] \mathbf{A}_n[f], \tag{7.42}$$

where

$$\bar{\mathbf{R}}_{\mathbf{x}}[k,f] = \sum_{n=1}^{N} \lambda_n R_{S_n}[k,f] \bar{\mathbf{R}}_n[f]. \tag{7.43}$$

This filter can achieve less spectral distortion than a scalar mask because $\bar{\mathbf{R}}_{\mathbf{x}}^{-1}[k,f]$ adapts to block the dominant source at each time and frequency.

There is a similar effect on the interaural cues. If $\mathbf{G}_n[f] = G_n[f] \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix}^T$ for all n, then from (4.75) we have

$$\mathrm{ITF}_n^{\mathrm{out}}[k,f] = \frac{G_n[f]\mathbf{e}_2^T\mathbf{A}_n[f] + \sum_{m=1}^{N} \lambda_m \left(G_m[f] - G_n[f]\right)\mathbf{e}_2^T R_{S_m}[k,f]\bar{\mathbf{R}}_m[f]\bar{\mathbf{R}}_{\mathbf{x}}^{-1}[k,f]\mathbf{A}_n[f]}{G_n[f]\mathbf{e}_1^T\mathbf{A}_n[f] + \sum_{m=1}^{N} \lambda_m \left(G_m[f] - G_n[f]\right)\mathbf{e}_1^T R_{S_m}[k,f]\bar{\mathbf{R}}_m[f]\bar{\mathbf{R}}_{\mathbf{x}}^{-1}[k,f]\mathbf{A}_n[f]}. \tag{7.44}$$

The discrete-state filter adapts to block the interaural cues of the active source from mixing with those of the inactive sources, providing lower spatial distortion than a fixed space-time filter for sparse signals.

7.4.3 Experiments with underdetermined mixtures

The high-low model was applied to binaural source remixing in [112]. The time-varying filter was tested with up to eight speech sources using unity distortion weights, the high-low source model, and a nearest-neighbor source activity classifier [178]. The performance of the proposed method was compared with that of a time-invariant multichannel Wiener filter, a binary mask using the same nearest-neighbor classifier, and the interaural-cue-constrained filtering method ("JBLCMV") proposed in [138] and [73], which constrains the interaural transfer functions ($\mathbf{e}_2^T\hat{\mathbf{D}}_n(\Omega)/\mathbf{e}_1^T\hat{\mathbf{D}}_n(\Omega)$) but not the spectral distortion ($\mathbf{D}_n(\Omega) - \hat{\mathbf{D}}_n(\Omega)$) of the source images.

The desired processing response was unity gain for a target source channel, 20 dB attenuation for all other speech channels, and complete attenuation of a diffuse noise channel. The 20-second speech samples were taken from the TIMIT database [103] and convolved with impulse responses from binaural hearing aid earpieces [97]. All

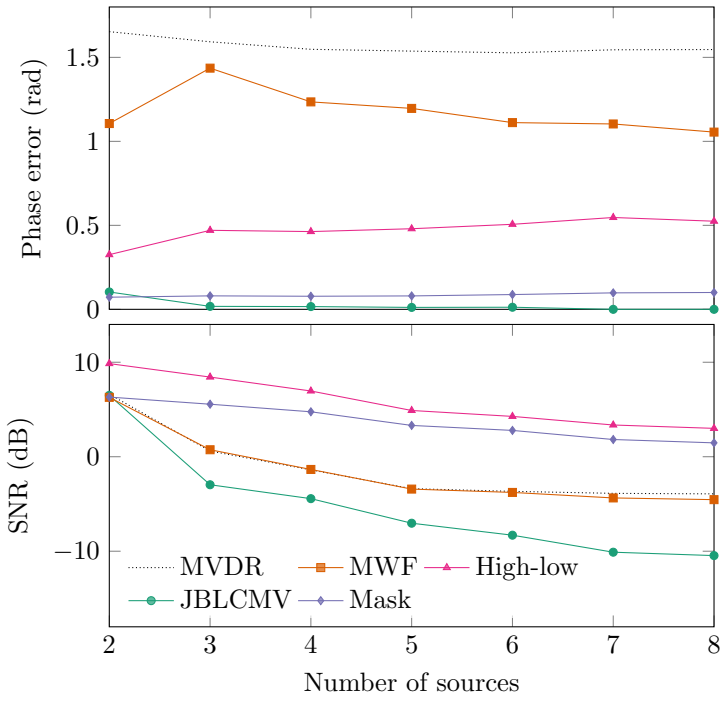

Figure 7.6: Comparison of different binaural source remixing methods with $M = 4$ microphones in an anechoic environment. Figure reproduced from [112].

results are averaged over 50 trials with randomly selected speech samples and approximately isotropic white noise. The experimental IPD and ILD error are weighted by the ground truth source powers in each time-frequency bin to avoid penalizing distortion during speech pauses.

Figure 7.6 shows the performance of the tested methods as a function of the number of speech sources in an anechoic environment using ground truth channel parameters. The top plot shows the average interaural phase difference error averaged over the background sources. The interaural level difference error is not shown as it has the same shape. The MWF has the largest error since it does not explicitly constrain spatial cue distortion. The interaural-cue-constrained filter has essentially zero interaural phase difference error. The mask, which does not perform spatial projection, also has low error. The proposed method falls in between. Notably, the error does not increase significantly with the number of sources.

Table 7.2: Comparison of binaural remixing methods for five speech sources in a reverberant environment. The filters in the top half were designed using ground truth impulse responses and the filters in the bottom half used erroneous impulse responses. Table adapted from [112].

	Filter	Foreground SER (dB)	Background SER (dB)	Background IPD err (rad)	Background ILD err (dB)
True params	Static MWF	12.4	−0.8	0.86	5.08
	JBLCMV	15.4	−8.9	0.46	3.21
	Binary mask	13.8	3.2	**0.10**	**0.82**
	High-low	**17.6**	**5.5**	0.45	2.90
Wrong params	Static MWF	3.7	−8.6	1.23	7.20
	JBLCMV	0.9	−15.5	1.16	6.80
	Binary mask	**12.9**	**1.9**	**0.11**	**0.93**
	High-low	8.6	−0.5	0.96	5.95

The lower plot shows the signal-to-noise ratio (SNR) for the first source channel, which would be around 15 dB for an ideal remixing filter. Here the interaural-cue-constrained filter performs the worst: for large N, the only filter that simultaneously satisfies interaural transfer function constraints on all sources is a passthrough filter; that is, for large N it does nothing at all. The two time-varying filters, the mask and the high-low filter, leverage the sparsity of the sources to achieve the highest performance. The high-low filter performs slightly better than the mask because it can filter in space; with a larger array, it would likely perform even better.

This experiment also suggested that the high-low filter causes less spectral distortion and is more robust to channel mismatch. Table 7.2 shows the signal-to-error ratios for the foreground and background source output images and the average IPD and ILD error of the background sources. The source images were generated using impulse responses from a reverberant courtyard ($T_{60} \approx 900$ ms). The results on the top half of the table use these ground truth impulse responses, while the results on the bottom half use filters designed with anechoic measurements with similar angles of arrival. The proposed method achieves the lowest spectral distortion. The interaural-cue-constrained method severely distorts the spectra of the background

sources even though it preserves their interaural cues. The performance of the proposed method degrades less with erroneous parameters compared to the other two spatial filtering methods, although it is less robust to error than the mask.

7.4.4 Role of time-varying methods in augmented listening

Time-varying methods were originally developed for applications where there are only one or a few microphones. Conventional hearing aids are certainly one such application: they have only two or four closely spaced microphones. In noisy environments with many sound sources—where their users need them the most—they cannot hope to process so many sources with time-invariant methods. As demonstrated in the previous section, nonstationary models and time-varying algorithms can leverage sparsity to apply more aggressive processing to more sources than would be possible with a time-invariant filter.

Are clever nonstationary models and time-varying source separation algorithms enough to dramatically improve the performance of conventional hearing aids in challenging conditions? Could they enhance normal human hearing to superhuman levels? Of course not; if we give our system the same information available to the ears, we would be lucky to even come close to human performance. The premise of this work is that to realize dramatic improvements in listening systems, we must use massive-scale microphone arrays. Massive arrays would have ample degrees of freedom to separate sources and apply perceptual constraints; why, then, do we need time-varying methods?

As we will see in the following chapters, time-varying methods are useful for more than just underdetermined mixtures. The real world is nonstationary: talkers walk around, wearable microphone arrays bend and twist, and sample clocks drift over time. Time-varying methods, especially the local Gaussian model, can help to address these real-world challenges.

Chapter 8

Source-Informed Acoustic Channel Estimation

To design space-time filters that can spatially separate and remix sound sources, we need to know the parameters of the acoustic channel, that is, how sounds propagate from a source to each microphone of an array. These channel parameters take different forms depending on the type of processing for which they will be used. The simplest directional beamformers use a far-field anechoic model: each sound source is characterized by a single direction of arrival, which determines the source signal's time differences of arrival between microphones. Most environments in which humans need help hearing, such as restaurants and convention halls, are strongly reverberant, so space-time filters are designed using M-dimensional acoustic-transfer-function (ATF) or relative-transfer-function (RTF) vectors that characterize more complex acoustic paths including reflections, reverberation, and frequency-selective devices and materials. The most general time-invariant channel model applied in this dissertation is the full-rank spatial covariance model, which uses an M^2-dimensional correlation or power spectral density matrix for each source channel.

Blind acoustic channel estimation is a difficult task, even in controlled environments where nothing moves and the number of sound sources and geometry of the microphone array are known. In more realistic conditions, with unknown numbers of moving, nonstationary sources and unknown array geometry, it is virtually impossible to learn the acoustic channel parameters blindly. While there is reason for optimism that blind source separation and acoustic channel estimation performance can improve with larger arrays and new machine-learning methods, we should look beyond blind methods to build practical augmented listening systems. This chapter reviews established approaches to acoustic channel estimation and introduces two

new methods, based on the author's work in [188] and [189], that exploit varying degrees of prior knowledge about the source signals to improve performance.

8.1 Acoustic Channel Estimation

8.1.1 Direct measurement

The most reliable way to learn the parameters of an acoustic channel—and the only way to learn non-relative ATFs—is to measure them directly. If $s_n(t)$ is a known source signal from rank-1 source channel n, then the source image is given by

$$\mathbf{c}_n(t) = \int_{-\infty}^{\infty} \mathbf{a}_n(v) s_n(t-v)\,\mathrm{d}v, \tag{8.1}$$

where $\mathbf{a}_n(v)$ is the acoustic impulse response vector of the channel. If it were possible to observe the source image $\mathbf{c}_n(t)$ directly, then the acoustic transfer function could be readily computed in the frequency domain:

$$\mathbf{A}_n(\Omega) = \frac{\mathbf{C}_n(\Omega)}{S_n(\Omega)}, \tag{8.2}$$

for all Ω at which $S_n(\Omega)$ is nonzero.

If we are able to choose the probe signal $s_n(t)$, we should use a signal that covers all frequencies of interest. Popular choices include linear and exponential sweeps [105, 106] and pseudorandom noise [107].

In practice, the source signal $s_n(t)$ is rarely known exactly. Fortunately, the absolute ATFs are not necessary for building augmented listening systems that are referenced to the ears; we only need the relative transfer functions. If the noise-free source image $\mathbf{c}_n(t)$ is known, then the RTF relative to microphone 1 is given by

$$\frac{\mathbf{C}_n(\Omega)}{\mathbf{e}_1^T \mathbf{C}_n(\Omega)}, \tag{8.3}$$

for all Ω at which the reference microphone spectrum is nonzero.

8.1.2 Subspace methods

Direct measurement of ATFs or RTFs requires that we have access to the true source images $\mathbf{c}_1, \ldots, \mathbf{c}_N$. But if we knew those signals, we would not need to design a space-time filter in the first place. We need the images to estimate the ATFs and the ATFs to estimate the images, a chicken-and-egg problem. Thus, we must make additional assumptions about the source signals in order to estimate the acoustic channel parameters.

Consider a mixture of a single high-power rank-1 target source and low-power diffuse background noise. In the frequency domain, the power spectral density of the mixture is

$$\mathbf{R}_{\mathbf{x}}(\Omega) = R_{s_1}(\Omega)\mathbf{A}_1(\Omega)\mathbf{A}_1^H(\Omega) + \mathbf{R}_{\mathbf{c}_2}(\Omega). \tag{8.4}$$

If the signal-to-noise ratio is large, then $\mathbf{A}_1$ should be parallel to the principal eigenvector of $\mathbf{R}_{\mathbf{x}}$. Let $\mathbf{U}(\Omega)$ be the principal eigenvector, that is, the solution to

$$\mathbf{R}_{\mathbf{x}}(\Omega)\mathbf{U}(\Omega) = \lambda\mathbf{U}(\Omega) \tag{8.5}$$

for the largest value of λ. Then the estimated RTF is

$$\hat{\mathbf{A}}_1(\Omega) = \frac{\mathbf{U}(\Omega)}{\mathbf{e}_1^T\mathbf{U}(\Omega)}. \tag{8.6}$$

We can obtain a better estimate of the RTF by explicitly accounting for the channel statistics using the *covariance whitening* method [110, 190]. Let $\mathbf{U}(\Omega)$ be the generalized eigenvector that satisfies

$$\mathbf{R}_{\mathbf{x}}(\Omega)\mathbf{U}(\Omega) = \lambda\mathbf{R}_{\mathbf{c}_2}(\Omega)\mathbf{U}(\Omega) \tag{8.7}$$

for the largest value of λ. Then the RTF estimate is

$$\hat{\mathbf{A}}_1(\Omega) = \frac{\mathbf{R}_{\mathbf{c}_2}(\Omega)\mathbf{U}(\Omega)}{\mathbf{e}_1^T\mathbf{R}_{\mathbf{c}_2}(\Omega)\mathbf{U}(\Omega)}. \tag{8.8}$$

Of course, this method requires that we obtain an estimate of the noise covariance $\mathbf{R}_{\mathbf{c}_2}(\Omega)$. For a nonstationary speech target in stationary noise, this can be obtained using a classifier that labels time-frequency samples as predominantly speech or noise [126].

8.1.3 Blind source separation

If the mixture is more complicated than a single speech source in low-level noise, then we must add yet more assumptions about the source signals in order to separate them and identify the channel. The problem of separating similar signals based on assumptions about their structure or statistics is known as *blind source separation* (BSS) [11, 53, 191].

Many BSS techniques rely on the time-frequency sparsity and orthogonality of speech signals, as described in Chapter 7. Classification-based methods like DUET and its variants often use clustering algorithms to assign spatial features to different sources [178, 180, 182, 192]. Other methods rely on specific non-Gaussian probability distributions or other sparsity assumptions [55, 56].

Another class of BSS algorithms relies on statistical independence between source signals. *Independent component analysis* (ICA) separates convolutional mixtures in the STFT domain by assuming statistical independence between the signals in different source channels [193–195]. This independence, which is a stronger condition than uncorrelatedness, can be enforced using higher-order instantaneous statistics, temporal correlations, and non-Gaussianity. ICA uses iterative updates within each frequency band to produce an unmixing filter, $\mathbf{W}_{\mathrm{df}}[f] \in \mathbb{C}^{N\times M}$, that estimates a set

of statistically independent scalar source signals:

$$\begin{bmatrix} \hat{S}_1[k,f] \\ \vdots \\ S_N[k,f] \end{bmatrix} = \mathbf{W}_{\mathrm{df}}[f]\mathbf{X}_{\mathrm{tf}}[k,f]. \tag{8.9}$$

ICA has several weaknesses for the augmented listening application. The ICA update rule operates on a square $N \times N$ unmixing filter. If $M > N$, then the data must be projected onto a lower-dimensional subspace, for example using principal component analysis; the algorithm therefore does not take full advantage of spatial diversity for large arrays. It also has a frequency-domain scale ambiguity, meaning that it estimates an arbitrarily filtered version of the signals. To recover the RTFs, we can take the pseudoinverse of $\mathbf{W}$ and normalize it so that its first entry is 1. Finally, because it operates on each frequency band independently, ICA suffers from a permutation ambiguity between frequencies. That is, the same source signal might be assigned to different outputs for different frequency indices.

8.1.4 Independent vector analysis

To avoid the permutation ambiguity in ICA, a related method called *independent vector analysis* (IVA) performs iterative updates jointly across frequency bands. The blind source separation experiments in this dissertation, which appear in Chapter 10, use a variant called auxiliary-function IVA (AuxIVA) [196]. It uses a set of simple update rules. At each iteration, a weighted covariance matrix estimate is computed as

$$\hat{\mathbf{R}}_n[f] = \mathrm{mean}_k \frac{G'(P_n[k])}{P_n[k]} \mathbf{X}_{\mathrm{tf}}[k,f]\mathbf{X}_{\mathrm{tf}}^H[k,f], \quad \text{where} \tag{8.10}$$

$$P_n[k] = \sqrt{\sum_{f=0}^{F-1} \left|\mathbf{e}_n^T \mathbf{W}_{\mathrm{df}}[f]\mathbf{X}_{\mathrm{tf}}[k,f]\right|^2}, \tag{8.11}$$

for $n = 1, \ldots, N$, where G is a contrast function and $G'(p) = \frac{\mathrm{d}}{\mathrm{d}p}G(p)$. In [196] and the implementation used in this dissertation, $G(P) = P$, so $\hat{\mathbf{R}}_n[f]$ is the sample covariance scaled by the root-mean-square filter output power across frequencies. The unmixing matrix is updated according to

$$\mathbf{e}_n^T \mathbf{W}_{\mathrm{df}}[f] \leftarrow \mathbf{e}_n^T \left(\mathbf{W}_{\mathrm{df}}[f] \hat{\mathbf{R}}_n[f] \right)^{-1} \tag{8.12}$$

$$\mathbf{e}_n^T \mathbf{W}_{\mathrm{df}}[f] \leftarrow \frac{\mathbf{e}_n^T \mathbf{W}_{\mathrm{df}}[f]}{\sqrt{\mathbf{e}_n^T \mathbf{W}_{\mathrm{df}}[f] \hat{\mathbf{R}}_n[f] \mathbf{W}_{\mathrm{df}}^H[f] \mathbf{e}_n}}, \tag{8.13}$$

for $n = 1, \ldots, N$. These iterations are repeated until convergence.

In an augmented listening system designed to reproduce sounds as received by the ear, the converged filter must be normalized so that $\mathbf{W}_{\mathrm{df}}[f]\mathbf{A}_{\mathrm{df},n}[f] \approx \mathbf{e}_1^T \mathbf{A}_{\mathrm{df},n}[f]$ for all n of interest. The unmixing filter can also be used to estimate the relative transfer functions, which can then be used to design other space-time filters.

Although the performance of blind source separation algorithms is improving steadily, there are no known algorithms that can reliably separate more than a few sources in complex real-world environments. To learn the channel parameters we need for space-time filtering, we cannot rely on purely passive estimation techniques. Instead, we can actively gather information on the acoustic channel.

8.2 In Situ Channel Measurement From Pilot Signals

To characterize the achievable performance of microphone array augmented listening devices, it would be helpful to bypass the acoustic channel estimation bottleneck. If the blind acoustic channel estimation problem were solved tomorrow, how well could augmented listening devices work? So far, the results presented in this dissertation have used either synthetic mixtures or carefully controlled experiments in which sound sources are collected one at a time to form a set of ground truth data [108, 169]. However, experiments would be more realistic if the acoustic channels could be measured in the field and in real time. To perform these real-life experiments, we

can use pilot signals emitted from beacons placed on or near the sound sources, as proposed by the author in [188].

8.2.1 Wearable beacons

In wireless communication systems, transmission channels are never blindly estimated: they are measured using known sequences of symbols. Likewise, we can measure an acoustic channel by transmitting a known signal over it. Suppose that a sound source of interest, assumed to be source channel 1, has relative impulse response $\mathbf{a}_1(t)$. Source channel 2 is the pilot, which has relative impulse response $\mathbf{a}_2(t)$. Source channel 3 contains all other signals and is treated as noise. If the beacon is located next to the source of interest, then $\mathbf{a}_1(t) \approx \mathbf{a}_2(t)$. Note that because these are relative rather than absolute impulse responses, the impulse response of the transducer itself, that is, the spectral distortion it applies to the pilot signal, is irrelevant. The observed sampled signal is

$$\mathbf{x}(t) = \int_{v=-\infty}^{\infty} \left(\mathbf{a}_1(v)s_1(t-v) + \mathbf{a}_2(t)s_2(t-v)\right) \mathrm{d}v + \mathbf{c}_3(t) \tag{8.14}$$

$$\approx \int_{v=-\infty}^{\infty} \mathbf{a}_1(v) \left(s_1(t-v) + s_2(t-v)\right) \mathrm{d}v + \mathbf{c}_3(t). \tag{8.15}$$

If the pilot signal is uncorrelated with the noise channel, then the relative impulse response can be estimated using cross-correlation. The pilot signal should occupy the same spectrum as the signal of interest to ensure that the acoustic channel is identifiable. The accuracy of the estimate depends on the signal-to-noise ratio, the length of the impulse response, and the length of the pilot signal. The relative impulse response is noncausal in general and must be windowed to a reasonable length. Figure 8.1 shows the signal-to-error ratio achieved by an MVDR filter designed using relative impulse responses of different lengths, which were estimated using linear sweeps of different duration. The mixtures were simulated using TIMIT speech [103] and behind-the-ear earpiece impulse responses [97]. For best performance, the impulse response should be long compared to the reverberation time of the room and the

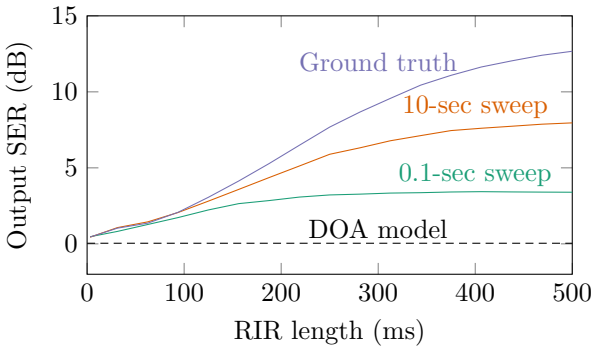

Figure 8.1: Simulated signal-to-error ratio for an MVDR beamformer designed using different impulse response estimates. There were three simulated speech sources in a reverberant office ($T_{60} = 300$ ms) with an overall input SNR of about 0 dB. Figure reproduced from [188].

sweep should be long compared to the impulse response.

8.2.2 Real-room measurements

To validate the beacon idea in a real room, a plastic mannequin was fitted with a small battery-powered stereo loudspeaker, as shown in Figure 8.2. Several other loudspeakers were spread around the then-untreated laboratory to play back noise. The data was recorded by a circular MEMS microphone array designed to imitate the microphone layout of a popular commercial smart speaker. First, the AIRs were measured using ten-second sweeps in quiet. Then, one channel of the loudspeaker played speech from the TIMIT database and one channel played a 100 ms sweep repeated every second. The pilot and speech sources were just a few centimeters apart. Noise images from the other loudspeakers were recorded separately and mixed with the speech-and-pilot recording.

Three single-target MVDR beamformers were designed based on the different channel estimates:

1. The AIRs measured from the pilot signal alone,

2. The AIRs estimated from the pilot signal in the noisy mixture, and

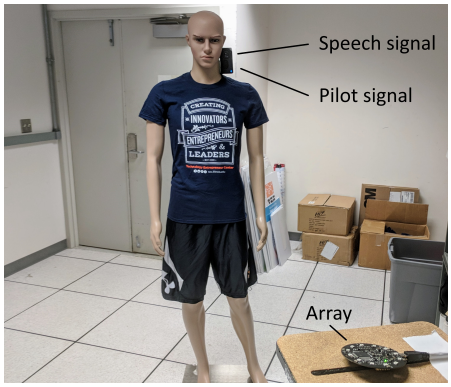

Figure 8.2: Speech and pilot signals were generated by a portable loudspeaker affixed to a plastic mannequin and recorded using a circular MEMS microphone array. Figure reproduced from [188].

Table 8.1: Image response (dB) for MVDR beamformers.

	Pilot signal	Speech signal	Noise signal
Measured AIRs	-0.2	-0.5	-9.9
Estimated AIRs	$+1.3$	$+1.9$	-6.1
Anechoic AIRs	-17.8	-24.8	-35.9

3. An anechoic model based on time differences of arrival.

The results are shown in Table 8.1. Rather than overall SNR or SER, the table shows the change in power of each source image:

$$\text{Image response}_n = 10 \log_{10} \frac{\sum_k \left| \hat{d}_{\mathrm{d},n}[k] \right|^2}{\sum_k \left| \mathbf{e}_1^T \mathbf{c}_{\mathrm{d},n}[k] \right|^2}, \quad n = 1, \ldots N. \tag{8.16}$$

The MVDR beamformer should have a response of 0 dB for the target source. The beamformer designed from the measured pilot signals did have a nearly distortionless response for the pilot and speech sources and attenuated the background noise by more than 9 dB. The system designed using the in situ pilot signals distorted the target signals by about 2 dB and provided slightly less noise reduction. The anechoic model, which a beamformer might use if it could not measure the reverberant AIRs of the channel, did a good job attenuating noise but also severely distorted the target signal.

The results show that it would be advantageous to use pilot signals emitted by beacons worn by talkers to calibrate augmented listening systems. Such a system would be impractical for most real-world applications, however.

8.2.3 Inaudible pilot signals

Because audible pilot signals are annoying and disruptive, the beacon method is only suitable for laboratory experiments. It could be applied to real listening systems if the pilot signals were inaudible, for example, if they were in the near-ultrasonic range. Unfortunately, if the AIR is allowed to be any signal, then inaudible pilot signals provide no information whatsoever about the channel in the audible range. If we hope to use inaudible pilot signals, then we must adopt a parametric model of the channel. For example, we can use them to measure time differences of arrival.

In Chapter 9, near-ultrasonic pseudorandom noise signals from five loudspeakers are used to track the positions of microphones in a deformable array as they move. A similar method could be used to estimate time-difference-of-arrival parameters of an

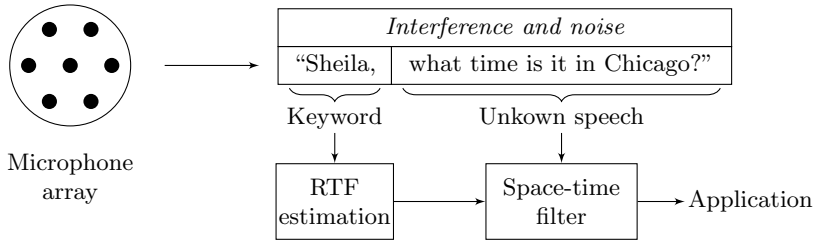

Figure 8.3: In a keyword-activated system, acoustic channel parameters are measured using the keyword utterance as a pilot signal. Figure adapted from [189].

acoustic channel, which could be used to design a crude filter. The results of Table 8.1 show that anechoic models are not suitable for reverberant environments, but they can be useful for initializing convolutional blind source separation algorithms [197].

8.3 Channel Estimation from Speech Keywords

It is not always possible to measure acoustic channel parameters using a deterministic source that is colocated with a sound source of interest. What if we could use the unknown source signal itself as a pilot? For speech signals, we often know a great deal about the possible values that the signal can take. The human vocal system can only produce a limited set of sounds [198]. If we restrict our attention to an individual language, the set of valid speech signals is even narrower. In fact, in certain applications, the system has prior knowledge of the specific word that was uttered.

In this section, we consider a channel estimation method, first proposed by the author in [189], designed for keyword-activated systems. In these systems, which include voice assistants embedded in array-equipped electronic devices, every user command begins with a known word or phrase, such as "Alexa," "Cortana," "OK Google," or "Hey Siri". The activating phrase, known as a keyword or hotword, is followed by a question or command to which the system must respond. Although they often contain microphone arrays, voice-activated devices struggle to understand commands in noisy and reverberant environments. If the system had a good estimate

of the RTF between the user and the array, it could better isolate the user's speech from unwanted noise.

Keyword-activated devices have an advantage over other machine listening systems because part of the signal of interest is known in advance. The proposed acoustic channel estimation method, shown in Figure 8.3, uses the activating keyword as a pilot signal to learn the relative transfer function between the user and the array. This RTF is used to design a single-target beamformer that isolates the rest of the user's question or command from background noise.

8.3.1 System overview

In a typical voice-assistant system, a microphone array continuously captures data and applies an on-device keyword spotting algorithm, usually a machine-learning classifier, that detects utterances of the keyword [199]. When the keyword is detected, the rest of the processor is activated or data is transmitted to a cloud service for further analysis. In this work, it is assumed that the keyword-spotting algorithm works perfectly and that it identifies the time interval in which the keyword is uttered.

The proposed system is developed in the time-frequency domain. Let $\mathbf{C}_{\text{tf},1}[k, f]$ be the STFT-domain source image of the target source, including the keyword. Let $\mathbf{C}_{\text{tf},2}[k, f]$ be the STFT of all other signals, which are treated as unwanted noise. In this proof-of-concept experiment, speech is enhanced by a time-invariant minimum-power-distortionless-response (MPDR) beamformer:

$$\mathbf{W}_{\text{df}}[f] = \frac{\mathbf{e}_1^T \mathbf{A}_{\text{df},1}[f] \mathbf{A}_{\text{df},1}^H \mathbf{R}_{\mathbf{X}_{\text{tf}}}^{-1}[f]}{\mathbf{A}_{\text{df},1}^H[f] \mathbf{R}_{\mathbf{X}_{\text{tf}}}^{-1}[f] \mathbf{A}_{\text{df},1}[f]}. \tag{8.17}$$

The MPDR beamformer is identical to the MVDR beamformer if the channel is estimated correctly, but is known to be more sensitive to channel estimation errors. This is a liability in practice, but it is useful for evaluating the performance of acoustic channel estimation algorithms. It also does not require explicit estimation of $\mathbf{R}_{\mathbf{C}_{\text{tf},2}}[f]$, which is difficult to measure because voice assistants are not permitted

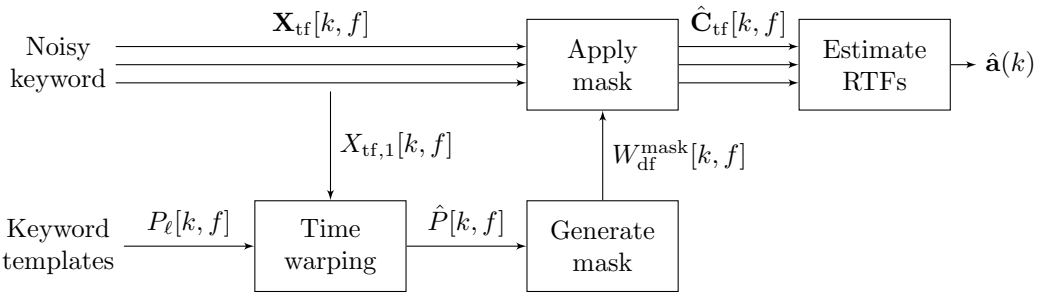

Figure 8.4: The proposed algorithm uses a template database to fit a time-frequency mask to the keyword utterance, then applies the mask to the array recording to isolate the keyword from background noise. The RTFs are computed from the estimated keyword image. Figure adapted from [189].

to capture audio before the keyword is uttered. Instead, the beamformer uses the sample covariance of $\mathbf{X}_{\mathrm{tf}}[k,f]$ over the keyword interval.

The goal is to estimate the discrete-frequency relative transfer function $\mathbf{A}_{\mathrm{df},1}[f]$ at all frequencies f for which there is important speech content. The proposed RTF estimation algorithm, shown in Figure 8.4, works as follows:

1. The keyword-spotting algorithm identifies the keyword and the time interval in which it is uttered.

2. A pattern-matching algorithm identifies the closest match to the uttered keyword from a database of keyword utterances.

3. The STFT of the nearest template is time-warped to align with the STFT of the recorded utterance.

4. The time-warped template is used to design a time-frequency mask.

5. The mask is applied to each of the M microphone recordings to remove background noise.

6. The masked signal is used to estimate the RTF from the talker to the array.

8.3.2 Template matching

Template matching has been used for decades in small-vocabulary speech recognition [200]. One channel of the observed recording is compared against a set of templates $P_\ell[k, f]$, $\ell = 1, \ldots, L$, from a database. Because the same word can be uttered at different rates, each template STFT is stretched or compressed in time to better match the recording. The best-fitting template index $\hat{\ell} \in \{1, \ldots, L\}$ and the corresponding time-warping pattern $\hat{\kappa}[k]$ can be found by solving the optimization problem

$$\min_{\ell, \kappa[k]} \sum_k \text{Cost}\left(\mathbf{e}_1^T \mathbf{X}_{\text{tf}}[k, 0], \ldots, \mathbf{e}_1^T \mathbf{X}_{\text{tf}}[k, F-1]; P_\ell[\kappa[k], 0], \ldots, P_\ell[\kappa[k], F-1]\right), \tag{8.18}$$

where $\kappa[k]$ is constrained to be nondecreasing. In the experiments presented here, the cost function is Euclidean distance between Mel frequency cepstral coefficients of each pair of frames. The optimization problem (8.18) can be solved using dynamic programming [200]. The warped template is given by

$$\hat{P}[k, f] = P_{\hat{\ell}}[\hat{\kappa}[k], f], \quad k = 1, \ldots, K. \tag{8.19}$$

Next, the warped template is used to generate a time-frequency mask:

$$W_{\text{df}}^{\text{mask}}[k, f] = \begin{cases} 1, & \text{if } \left|\hat{P}[k, f]\right| > \gamma[f] \\ 0, & \text{otherwise,} \end{cases} \tag{8.20}$$

where $\gamma[f]$ is a frequency-dependent tuning parameter that trades off noise reduction for target signal distortion. Here, γ was chosen so that about 10% of time-frequency samples are preserved within each frequency band.

8.3.3 Relative transfer function estimation

Next, the time-frequency mask is applied to each microphone recording to remove background noise:

$$\hat{\mathbf{C}}_{\mathrm{tf},1}[k,f] = W_{\mathrm{df}}^{\mathrm{mask}}[k,f]\mathbf{X}_{\mathrm{tf}}[k,f], \quad k = 1,\ldots,K. \tag{8.21}$$

If the speech keyword is sufficiently sparse, then $\hat{\mathbf{C}}_{\mathrm{tf},1}[k,f]$ should be approximately parallel to the target RTF:

$$\hat{\mathbf{C}}_{\mathrm{tf},1}[k,f] \approx \mathbf{A}_{\mathrm{df},1}[f]S_{\mathrm{tf},1}[k,f] \tag{8.22}$$

From these masked microphone signals, we compute the sample spatial covariance of the source images:

$$\hat{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f] = \mathrm{mean}_k \hat{\mathbf{C}}_{\mathrm{tf},1}[k,f]\hat{\mathbf{C}}_{\mathrm{tf},1}^H[k,f]. \tag{8.23}$$

The estimated RTF $\hat{\mathbf{A}}_{\mathrm{df},1}[f]$ is the principal eigenvector of $\hat{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f]$. This method is similar to covariance whitening, except that the mask takes the place of the whitening filter.

8.3.4 Experiments

The experiments for the keyword-based acoustic channel estimation algorithm did not use the laboratory equipment described in Chapter 2. Instead, acoustic impulse responses were measured in a real living room ($T_{60} \approx 400$ ms) using an array of $M = 7$ MEMS microphones designed to imitate the array layout of a popular commercial smart speaker. Four loudspeakers were placed on a sofa, chair, table, and television stand two meters away from the array, which rested on a coffee table, as shown in Figure 8.5.

The test signals were generated by concatenating ten-second speech samples from the TIMIT database [103] with keywords from the crowdsourced Google spoken

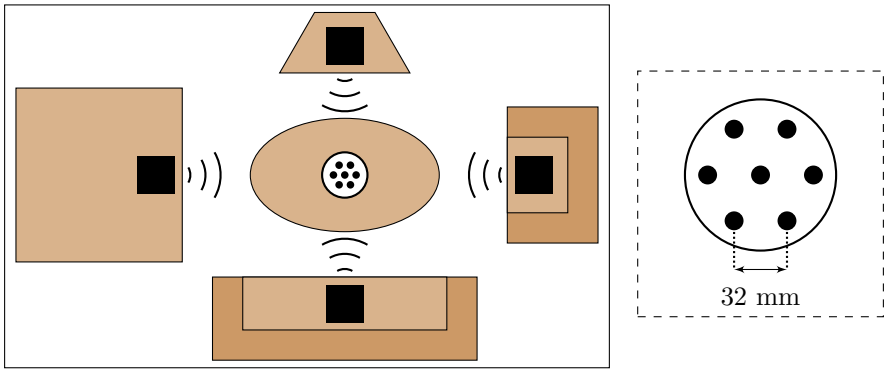

Figure 8.5: A smart-speaker-like microphone array was used to capture sound from four sources in the living room of a small apartment. Figure reproduced from [189].

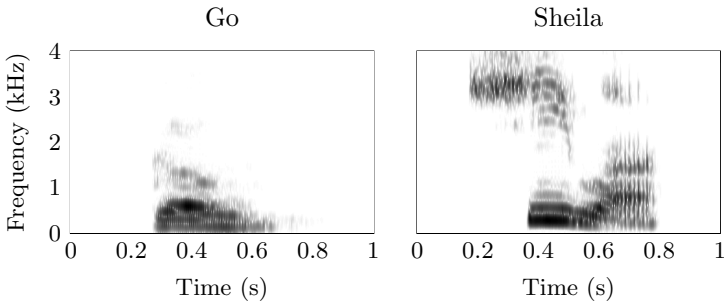

Figure 8.6: Spectrograms of two keyword utterances. Figure reproduced from [189].

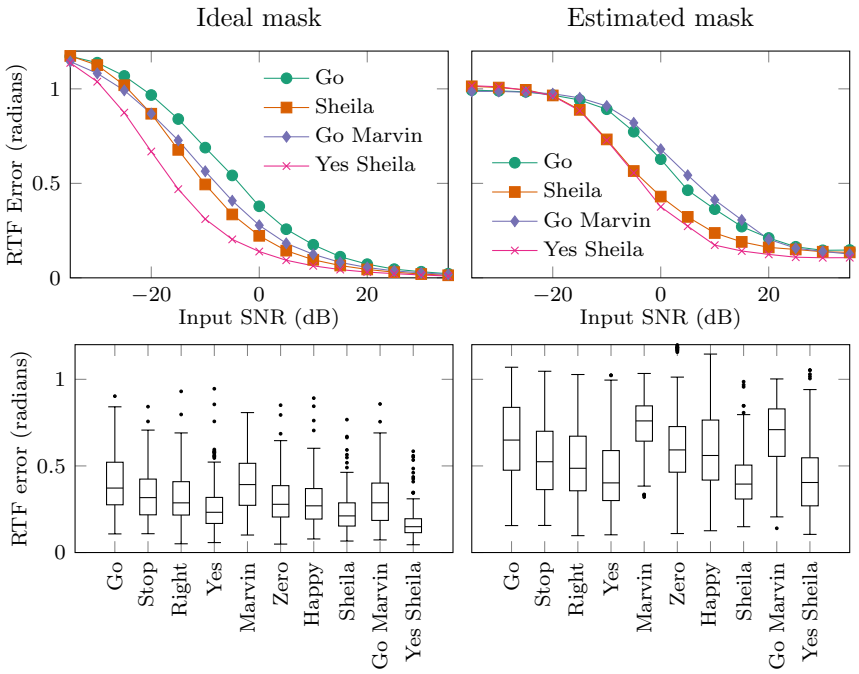

Figure 8.7: RTF vector estimation error, in radians, for different keywords. Top: RTF error versus input SNR. Bottom: RTF error at 0 dB SNR. Figure adapted from [189].

commands data set [201]. The spectrograms of two keyword samples are shown in Figure 8.6. The speech signals were convolved with the four measured impulse responses and added to background noise recorded in the living room, which was primarily appliance and ventilation noise. Note that the keyword is spoken by a different talker in a different environment than the rest of the sample. While the TIMIT samples are anechoic, the keywords were generated by thousands of talkers in different environments, some with strong noise and distortion. A set of $L = 500$ keyword utterances with relatively high perceptual quality were manually selected to form the template-matching database. Another 100 were used as a test set.

Figure 8.7 shows the angle between true and estimated RTF vectors, averaged across all frequencies, for different keywords. The plots on the left show RTFs estimated using the ground-truth ideal binary mask, while the plots on the right show

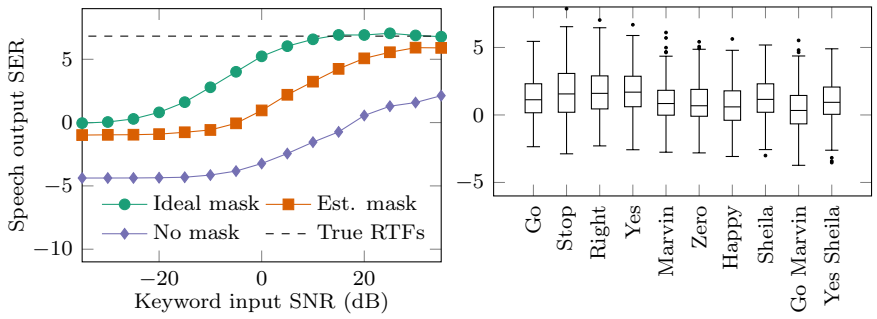

Figure 8.8: Left: Median output SER of the speech query versus input SNR of the keyword "Yes Sheila". Right: Output SER of the speech query at 0 dB keyword input SNR. Figure adapted from [189].

RTFs estimated using the proposed method. Longer keywords provide better estimates than short keywords. There seems to be a significant benefit from keywords that contain sibilants, such as "yes" and "Sheila," presumably because they provide more information about high frequencies.

Next, consider the performance of the overall system. Figure 8.8 shows the performance of the MPDR beamformer designed using the estimated RTFs. The proposed method is compared against both an oracle binary mask, which selects the strongest sound source at each time-frequency index, and no mask; the latter method simply selects the dominant eigenvector of the observed mixture at each frequency. The proposed method provides a benefit of roughly 20 dB over the fully blind method. The oracle mask performs better still, suggesting that there is room for improvement in the pattern-matching algorithm. There is less variation between keywords in the overall performance of the system, perhaps because most speech energy is concentrated at low frequencies where all the keywords have significant support.

These experiments show that known speech can improve the performance of keyword-activated machine listening systems. What about human augmented listening? It should be possible to apply the same principle to listening devices under certain conditions. For example, a headset could detect someone saying the user's name and automatically focus attention on them, while using the utterance as a pilot signal to learn the acoustic channel.

Chapter 9

Wearable Microphone Arrays

To build listening systems with better-than-human spatial perception, we must use microphone arrays that extend beyond the human ears. Humans have two acoustic sensors; thanks to recent advances in microphone and embedded-processing technology, listening devices could soon have hundreds. The previous chapters have explained the signal processing advantages of large microphone arrays and explored the unique challenges of array processing for listening devices. To actually build such a system, however, we must answer some basic questions: How many microphones do we need? Where should we put them? Can they go under clothes? What kind of microphone should we use? What if the user moves around?

It is surprising that these questions have gone unaddressed for so long; engineers have been building wearable microphone arrays for many years. A 1992 study of beamforming hearing aids compared the performance of microphone arrays on different parts of the body, but used only two microphones [64]. A five-microphone array mounted on eyeglasses—a perennially popular form factor—was introduced the following year [61]. New eyeglass-based designs continue to appear regularly in the literature [202] and are a constant presence on crowdfunding websites. In 2001, Widrow and Luo built a six-microphone array worn on the chest [21]. Helmet-mounted arrays with up to 32 microphones have been proposed for military applications [203, 204]. A dissertation from 2009 considered the design of head-mounted microphone arrays based on acoustic modeling of the head [203].

Until recently, large wearable arrays were laboratory curiosities; the technology did not exist to build practical wearable arrays with more than a few microphones. It is now feasible—or at least plausible—to build wearable microphone arrays with dozens

or even hundreds of microphones. New digital microelectromechanical-systems (MEMS) microphones are inexpensive, energy-efficient, and smaller than a pea. They can communicate directly with embedded processing hardware, which is now powerful enough to perform complex array processing in real time. MEMS microphone arrays are already being integrated into commercial wearable devices such as headphones and watches. However, these wearable arrays have been designed to fit into existing devices wherever convenient.

How would we design a wearable microphone array with the best possible spatial signal processing performance? For the last two years, the Augmented Listening Laboratory team has been developing prototypes of wearable microphone arrays with more ambitious designs, from discreet vests that can be hidden under a shirt to the enormous Sombrearo. These prototypes use digital MEMS microphones that could eventually be embedded into sleek accessories and a programmable logic platform that can perform real-time augmented-listening processing. They have also created a first-of-its-kind database of acoustic impulse responses for wearable microphones. These proof-of-concept prototypes and the new data set can help researchers to understand design constraints, tradeoffs, and best practices for wearable microphone arrays.

This chapter is based on the author's work on wearable microphone array design in [113] and on motion-robust array processing in [132].

9.1 Design and Construction

We begin our discussion of wearable arrays with the nuts and bolts—or rather, with the clocks and amplifiers—of microphone array design. An augmented listening device needs several components, as shown in Figure 9.1: microphones to capture sound, electronics to amplify and digitize the microphone signals, a processor to enhance those signals, and a pair of receivers to play them back to the listener.

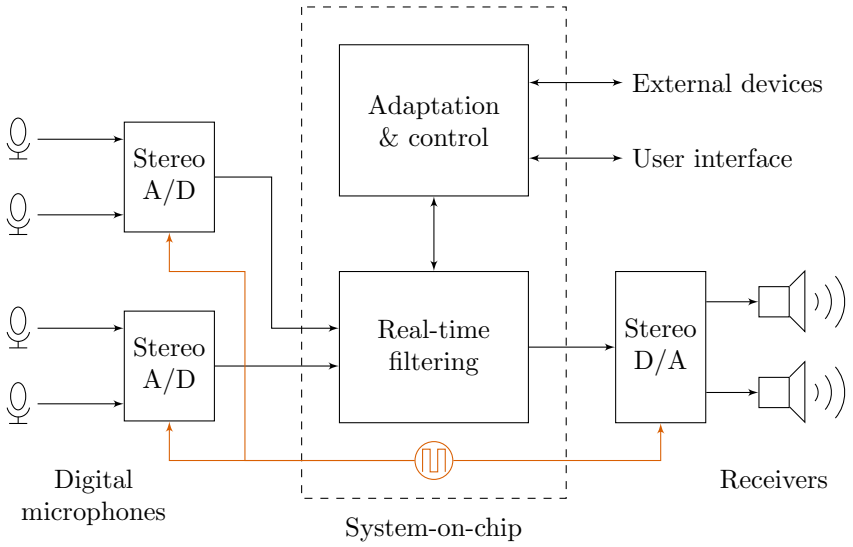

Figure 9.1: System architecture of an array-based digital augmented listening device.

9.1.1 Microphones

Today, engineers are spoiled for choice when selecting microphones for audio enhancement systems. Figure 9.2 shows a few of the different types of microphone available in the Augmented Listening Laboratory, from a large-diaphragm condenser microphone used for high-quality vocal recording to the tiny digital MEMS microphones found in nearly all modern consumer electronics.

Studio-quality microphones can be divided into two categories: condenser microphones, which are sensitive but fragile and require a power supply, and dynamic microphones, which are passive and durable but less responsive to quiet sounds. Dynamic microphones are generally used for live sound applications, such as concerts, while condenser microphones are used in the controlled environment of the studio. Consumer audio devices, such as mobile phones and headsets, historically used inexpensive low-voltage electret condenser microphones. Today, however, they are dominated by microelectromechanical-systems (MEMS) microphones.

Analog MEMS microphones have similar electroacoustical properties to traditional consumer-grade condenser microphones, but are much smaller and have lower cost

Figure 9.2: From left to right: A large-format cardioid condenser microphone, a dynamic vocal microphone, an omnidirectional lavalier microphone, and a MEMS microphone.

and power requirements. These microphones can be reflow-soldered directly onto printed circuit boards alongside other electronics in a compact device. Newer digital MEMS microphones have built-in analog-to-digital converters and can interface directly with digital processors. While most of the experiments in this dissertation use studio-quality condenser microphones, a commercial device would almost certainly use digital MEMS microphones. Indeed, most of the recent large embedded microphone arrays reported in the literature are based on MEMS microphones [54, 77, 78, 109].

All microphone types are available with different directivity patterns. In voice capture applications, most microphones are directional: they amplify sound from the direction of the talker or performer and attenuate sound from other directions. The directivity profile of a directional microphone varies with frequency. The vast majority of microphones used in the array processing literature, meanwhile, are omnidirectional: they have similar responses to signals from all directions at all frequencies in their range. Omnidirectional microphones are attractive for array processing because the frequency and directional responses of the array can be altered using signal processing.

As we will show in Section 9.2, the directivity of a microphone changes when it is worn by a human or mannequin. The body, especially the torso, blocks sound from the opposite side, making the microphone more directional than it would be in free space. While this directivity would harm the beam pattern of a delay-and-

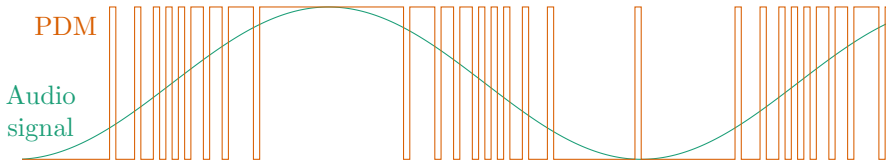

Figure 9.3: Embedded devices often transmit audio in pulse density modulation (PDM) format.

sum beamformer designed for anechoic far-field signals, it is actually an advantage for statistical space-time filters. Different microphones capture different information about the sound field, which is helpful when there are sound sources surrounding the body. Although only omnidirectional microphones are used in this dissertation, it is plausible that directional microphones could offer similar performance benefits.

9.1.2 Sensing architecture

A typical signal path is shown in Figure 9.1. The analog signals captured by the microphones must be sampled and quantized. The analog-to-digital converters could be in dedicated parts or they could built into the processor or the microphones themselves. The processed signals are converted back to analog and then presented to the user by a pair of transducers known as *receivers*. High-end hearing aids use balanced armature receivers, which provide excellent performance in a tiny package, but are more expensive than the transducers found in many consumer headphones.

To perform array processing, which depends strongly on phase relationships between microphone signals, it is critical that all audio inputs and outputs share a common sample clock. A difference of just a few parts per million in the sample rate between microphones would seriously harm the performance of an array [205]. Thus, all microphones attached to a device should be wired to the same clock signal generated by a single crystal. Digital MEMS microphones require two clock signals from the processor: a sample clock and a bit clock.

Modern digital microphones provide data in one of three formats, all of which support transmitting more than one microphone signal on the same wire:

Pulse code modulation (PCM): each audio sample is a binary-coded integer. Under the ubiquitous Inter-IC Sound (I2S) protocol, the bits of the left and right samples are transmitted during the high and low periods of the sample clock signal.

Pulse density modulation (PDM): the transmitted bits are the output of a noise-shaping one-bit analog-to-digital converter [206]. In audio devices, the over-sampling ratio is usually 64. Higher-amplitude signals have more frequent 1's and lower-amplitude signals have more frequent 0's, as shown in Figure 9.3. The PDM signals must be decimated by the processor to recover the PCM signals used for signal processing.

Time division multiplexing (TDM): a form of PCM with more than two channels per wire. Samples from different channels are interleaved in time. Unlike stereo PCM, which usually follows the I2S standard, there is no standard protocol for multichannel.

Despite the lack of standardization, TDM digital microphones are an attractive choice for large-scale microphone arrays because many microphones can be connected to a single port on the processor.

In most commercial products that use digital MEMS microphone arrays, such as smart speakers, mobile phones, and gaming systems, the microphones are all mounted to the same printed circuit board. In a large-scale wearable array, the microphones would be much farther apart. Care must be taken to ensure that clock and data signals are preserved over long wire connections.

9.1.3 Signal processing

Once the audio signals from the microphones have been digitized and transmitted to the processor, they must be processed to produce a pair of output signals. Embedded audio systems vary widely in their computational requirements. Since audio sample rates are relatively slow compared to modern microprocessor clock speeds, it

is possible to build simple audio systems using low-power microcontrollers. More sophisticated processing, such as many-channel space-time filtering and time-frequency algorithms, require more powerful processors. High-end hearing aids typically use custom integrated circuits. A mid-range off-the-shelf hearing-aid chip might include a digital signal processor core, a set of dedicated hardware filterbanks, a few microcontrollers, analog amplifiers, and wireless communication interfaces.

No hearing aids or consumer audio devices on the market today contain as many microphones as the arrays considered in this dissertation. To build systems that can support dozens of microphones, we need a highly parallel architecture. The undergraduate students in the Augmented Listening Laboratory have developed an architecture based on the Intel DE-1 SoC system-on-chip platform, which includes a field programmable gate array (FPGA) and a mobile-class applications processor. The programmable logic supports a large number of I2S microphone interfaces and a set of finite-impulse-response filters that process the received signals and produce low-latency stereo outputs. The data is also stored in memory accessible to the applications processor, which can interface with the user or other devices to update filter parameters. The architecture, shown in Figure 9.4, draws on the open-source Pyramic project [109].

It remains to be seen whether a mobile-class processor can perform the high-dimensional processing required for powerful augmented listening. Fortunately, most of the algorithms described in this dissertation rely on repeated multiplication of large matrices. That is, they require the same type of hardware as deep neural networks, for which the computing industry is racing to build efficient processors. Between continued improvements to digital microphone technology and the increasing availability of highly parallel embedded processing, it will soon be possible to perform sophisticated signal processing on large microphone arrays using a battery-powered wearable device.

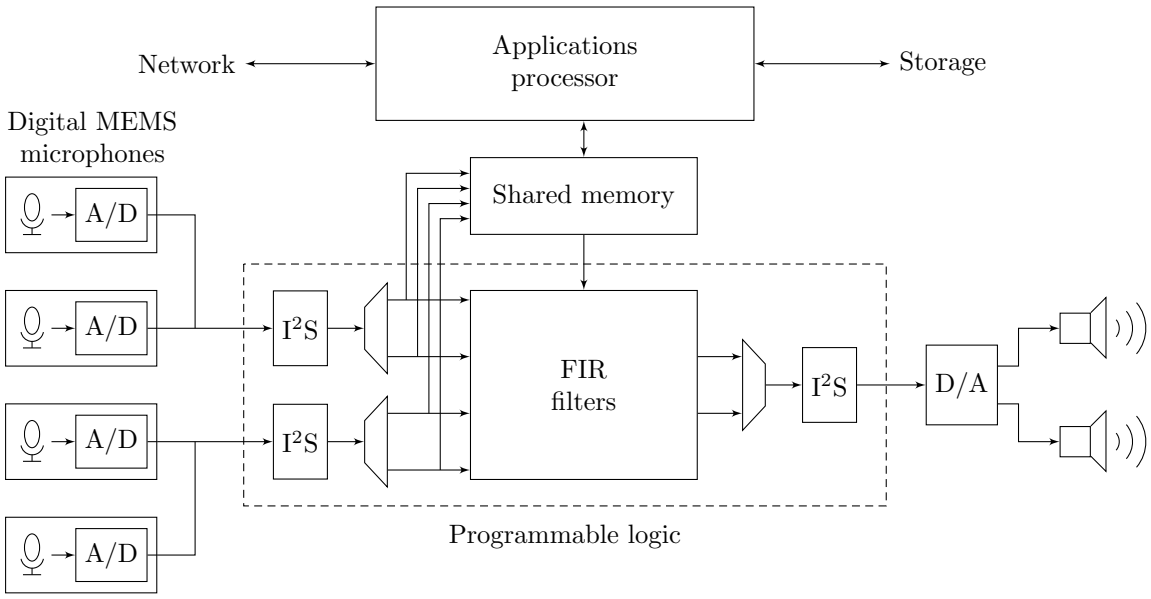

Figure 9.4: Processing architecture of a prototype listening device.

Figure 9.5: Wearable microphone array prototypes.

9.1.4 Array form factors

One of the most important decisions in designing a wearable microphone array is where to place the microphones. The best-studied sensor arrays are uniform linear arrays, which can be easily analyzed mathematically for narrowband signals arriving from far away in an anechoic environment. Thus, some of the earliest microphone arrays proposed for hearing aids were linear arrays placed on eyeglasses [61]. Of course, audio signals have wide bandwidth and a listener's face is certainly not an anechoic environment, so there is no need to use a linear form factor for a wearable microphone array. Circular arrays are also popular because they have many axes of symmetry that simplify beamformer design. They are often used in tabletop teleconferencing devices and smart speakers. Because the human head is roughly circular, several researchers have designed arrays that encircle the head on helmets [203, 204, 207].

Wearable-array form factors have been motivated as much by convenience and aesthetics as by signal-processing performance. Hearing-aid manufacturers sacrifice performance and battery life to squeeze their products into tiny, discreet packages. These companies believe, perhaps correctly, that consumers are embarrassed to wear hearing aids. If it really is important that listening devices be discreet, then we should consider array designs that can be concealed under clothing.

The Augmented Listening Laboratory design team has experimented with several wearable-array designs, as shown in Figure 9.5. The first working prototype was on a construction helmet. The team is also developing a pair of headphones and a vest that can be worn under clothing. However, these proof-of-concept prototypes do not address the fundamental question of where microphones should be placed to maximize spatial signal processing performance. To do that, the team used studio-quality lavalier microphones to measure wearable microphone impulse responses all over the body. We can use this data to study performance tradeoffs and formulate design guidelines for wearable microphone arrays.

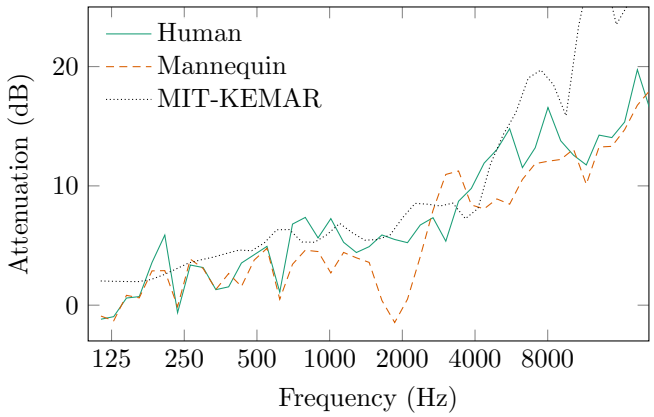

Figure 9.6: Interaural level differences from the wearable microphone-database and from an anechoic HRTF database [96]. Figure adapted from [113].

9.2 Acoustic Effects of the Body

To design high-performance wearable microphone arrays, we must understand the acoustic effects of the body. The wearable-microphone database [113] (Section 2.3) includes measurements for 80 locations on the body and an additional 80 positions on various wearable accessories from 24 angles of arrival. Measurements were performed on both a human subject and a plastic mannequin. We can use this data to understand the acoustic effects of the body and of clothing on wearable microphones.

9.2.1 Head-related transfer functions

The acoustics of some parts of the human body have been studied extensively: the head and ears. *Head-related transfer functions* (HRTF) describe how sound propagates into the ear canal from different directions of arrival [117]. The shaping effects of the pinna allow humans to localize sound not only on a left-right axis, but also front-back and up-down. There are large databases of head-related transfer functions captured in anechoic conditions [95, 96].

Figure 9.6 shows the average interaural level differences between the microphones in the left and right ears for contralateral sources. The human and mannequin data

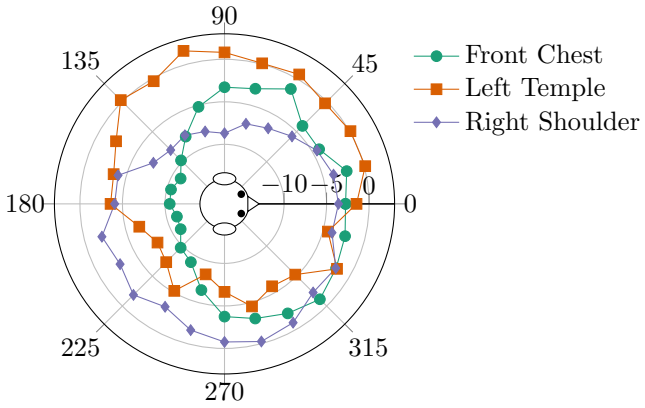

Figure 9.7: Acoustic transfer function magnitude in dB relative to a free-space microphone response for sources from different directions for several wearable microphones. Figure adapted from [113].

from the wearable-microphone data set is compared with that from the MIT-KEMAR database [96]. Although the wearable-microphone data set was not collected in a fully anechoic chamber, the interaural level difference is similar at low and middle frequencies. The plastic mannequin head is slightly more transparent than the human head, but is likely similar enough to be useful for proof-of-concept experiments.

9.2.2 Attenuation by the body

We can also study transfer-function-magnitude differences across other parts of the body. Because the body blocks sound transmission, especially at high frequencies, it causes omnidirectional microphones to have directional responses. Figure 9.7 shows the effective directivity, averaged across all frequencies, of microphones on the chest, head, and shoulder. The head, a popular location for wearable arrays, provides the least directivity. The most effective is the torso, which attenuates contralateral sound by more than 10 dB.

These attenuation effects are frequency-dependent. Figure 9.8 compares the attenuation of the chest and torso as a function of frequency. The body has little effect at low frequency, but significantly attenuates higher-frequency sound. The figure also

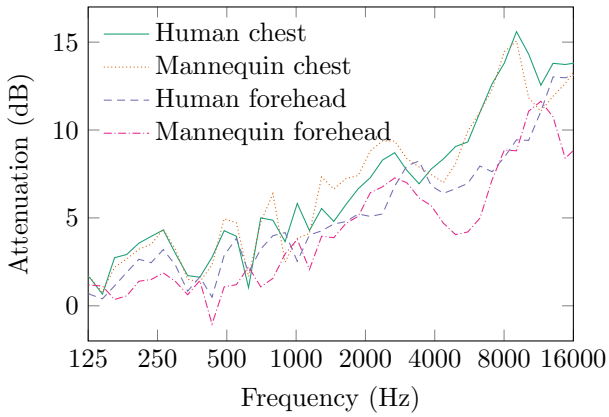

Figure 9.8: Attenuation by the body for sources and microphones on opposite sides of the torso. Figure adapted from [113].

compares the attenuation effects of the human and mannequin subjects. The subjects' bodies, especially the torsos, cause similar attenuation. These results suggest that inexpensive plastic mannequins are suitable acoustic substitutes for real human users in wearable-microphone experiments.

9.2.3 Effects of clothing

Some users might prefer to conceal a wearable microphone array under clothing. To assess the performance of a concealed microphone array, measurements of the 16 microphones on the middle and upper torso were repeated with several types of outerwear. The attenuation due to clothing is shown in Figure 9.9. The cotton t-shirt, cotton dress shirt, fleece pullover, and cotton sweatshirt cause little attenuation except at high frequencies. A microphone array worn under these articles would likely function well, especially because large arrays are most useful at lower frequencies that cannot be resolved by small head-mounted arrays. The wool coat and leather jacket, however, cause severe attenuation above 500 Hz. A wearable array would likely not be useful underneath heavy winter outerwear.

Research on head-related transfer functions has shown that clothing, especially eyeglasses and hats, also affects acoustic transfer functions to the ear. So do large

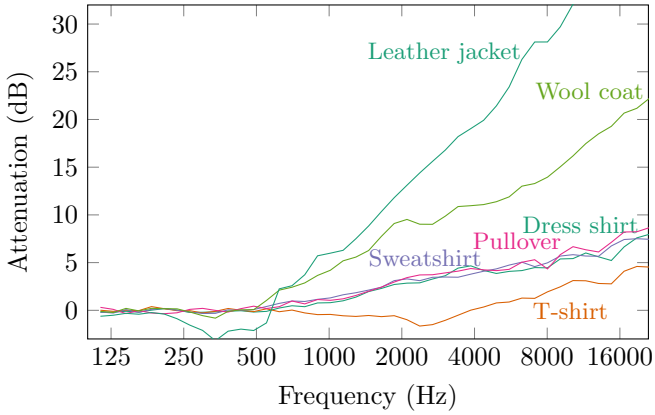

Figure 9.9: Attenuation due to clothing for the 16 microphones on the middle and upper torso. Figure adapted from [113].

curly hairstyles. However, these differences do not appear to affect human localization abilities [117, 208–210].

9.3 Beamforming Performance of Wearable Arrays

To compare the performance of different wearable-microphone-array designs, the measured impulse responses were used to simulate mixtures of six VCTK speech sources and spatially uncorrelated noise. The impulse responses were windowed to 32 ms to simulate parameter estimation errors and the windowed responses were used to design MVDR beamformers, referenced to the left ear, for each source channel. For each array, the experiment was repeated for 100 combinations of six randomly selected directions of arrival.

9.3.1 Number of microphones

Wearable microphone arrays can support many more microphones than conventional earpieces. Figure 9.10 shows the mean SNR improvement over the 100 trials of MVDR beamformers using different numbers of microphones. Every array includes

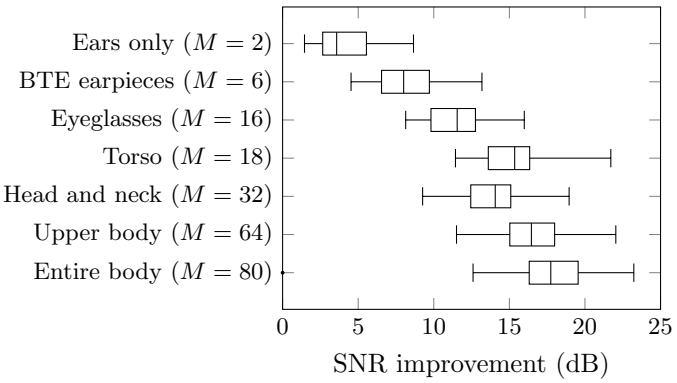

Figure 9.10: Performance of MVDR beamformers with different numbers of microphones. Figure adapted from [113].

the two microphones in the ears. Arrays with more microphones tend to outperform arrays with fewer microphones, but microphone location also matters. The 18-microphone array on the ears and torso outperforms an array of nearly twice as many microphones on the head and neck. The torso blocks more high-frequency sound than the head, providing more spatial diversity, and the microphones are spread farther apart, helping to separate low frequencies.

Figure 9.11 shows mean SNR improvement as a function of frequency for a few of the tested arrays. The wearable arrays are most effective at high frequencies, where wavelengths are small compared to the spacing between microphones and the body effectively blocks contralateral sound sources. The limiting factor in overall performance appears to be low-frequency sound, which is difficult to separate even with a body-scale array.

9.3.2 Microphone placement

Clearly, the placement of microphones affects the performance of the array. Figure 9.12 compares different array designs that all use $M = 18$ microphones: two in the ear and sixteen on different body parts or wearable accessories. Small accessories worn on the head, including the baseball cap and headphones, perform worst. This

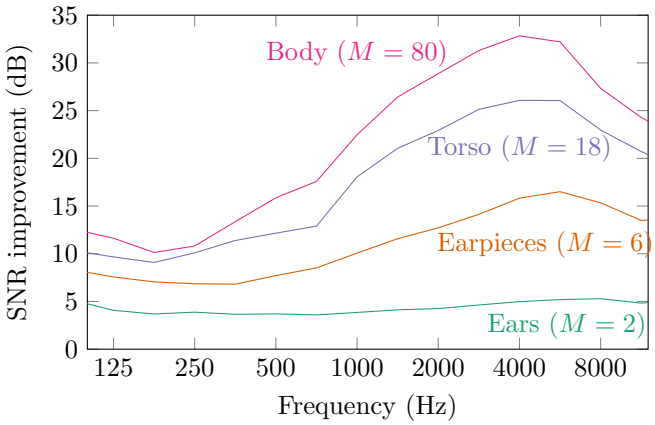

Figure 9.11: Mean SNR improvement versus frequency for MVDR beamformers with different numbers of microphones.

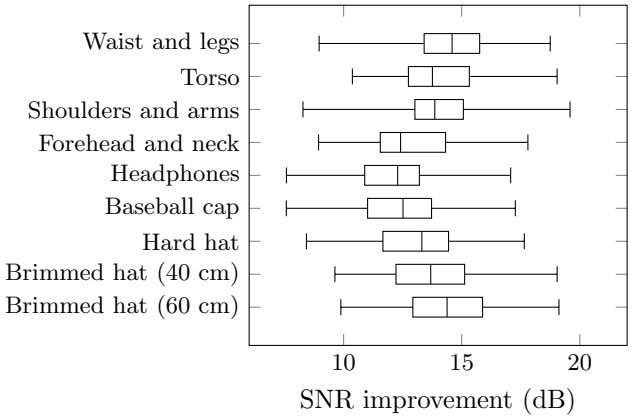

Figure 9.12: Performance of 18-microphone arrays with different form factors. Figure adapted from [113].

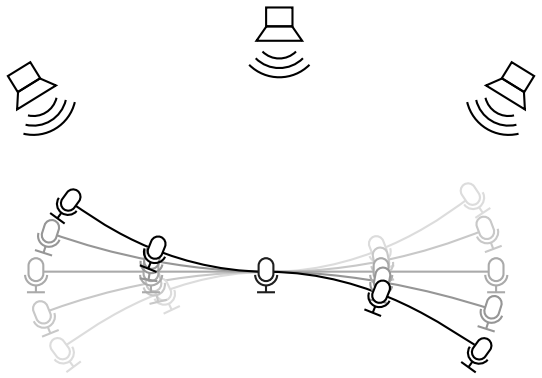

Figure 9.13: In deformable microphone arrays, such as wearable arrays, microphones can move relative to each other. Figure reproduced from [132].

is bad news for wearable-array designers who would like to add more microphones to existing audio accessories. The enormous Sombrearo, shown in the bottom row, has excellent performance. Similar performance can be achieved without an ostentatious accessory by spreading microphones across a large area on the user's body.

9.4 Modeling Small Microphone Motion

Experiments with wearable microphone arrays suggest that for best performance, microphones should be spread across the body. Arrays that cover the head, torso, arms, and legs provide the best spatial diversity for source separation and remixing. There is a serious problem with these body-scale arrays, however: humans move! As the user walks, looks around, gestures, or even just breathes, the microphones will move not only relative to the sound sources, but also relative to each other. This intra-array motion changes the phase differences that are critical to all forms of array processing.

9.4.1 Deformable microphone arrays

A *deformable array* is one in which the sensors can move relative to each other [132], as illustrated in Figure 9.13. This deformation is in contrast to the rigid motion of, say, an eyeglass-mounted array, and to motion of sound sources. Deformable arrays are more difficult to model than rigid arrays because they have more degrees of freedom in their motion.

Tracking and isolating moving sources, while still a challenging problem, has been studied extensively. Most methods combine a technique for localizing sound sources, such as steered response power or multiple signal classification, with a tracking algorithm, such as a Kalman or particle filter [211–216]. Sparse signal models (Chapter 7) can help to improve performance in noisy mixtures [217–220]. Others have applied blind source separation algorithms that adapt over time [221, 222].

There has been much less work on deformable microphone arrays. Robotics researchers have studied microphone arrays mounted on movable structures. In [223], a set of movable arms adaptively repositioned microphones to improve their beamforming performance. In [224], microphones along a hose-shaped robot were used to estimate its posture as it moved. Moving wearable arrays were studied in [186] as part of an asynchronous distributed array, which is discussed in Chapter 10. This part of the chapter is based on [132], which elaborates on the time-invariant full-rank model of motion introduced briefly in [186].

When microphones exhibit large motion, such as when a wearable-array user walks across a room, we have little choice but to explicitly track their positions. But what about smaller motion, such as nodding or breathing? Is there a way to compensate for small deformations using time-invariant methods rather than computationally expensive and error-prone tracking algorithms?

9.4.2 Statistical model of deformation

Small changes in sensor position resemble acoustic channel estimation errors, which have been studied in the context of robust beamforming. As explained in Section

4.2.3, parameter mismatch can be modeled as a random, uncorrelated perturbation to the power spectral density matrix of a source channel. Such errors can be addressed using derivative constraints [225], norm constraints [135], or distortionless constraints within a region of space around the target [83, 226]. These methods amount to widening the beam pattern to improve robustness at the expense of noise reduction.

A similar tradeoff should apply to deformation: if the motion were small enough that the array could still tell different source channels apart, then there would be no need to explicitly track the microphone positions. How small is small enough? Let us consider the effects of motion on the second-order statistics of the array signals, that is, on the covariance matrices used to design linear time-invariant filters.

Since we will be comparing time-varying and time-invariant methods, we will analyze signals in the STFT domain. Assume that the motion is slow enough that the effects of Doppler can be neglected and the microphone positions are approximately constant within each time frame. The state of the microphones in each time frame k is denoted $\theta[k] \in \mathcal{X}$, where $\mathcal{X}$ is a set of states that represent different microphone positions and orientations. While $\mathcal{X}$ should properly be thought of as a continuous set, in the state-tracking implementation used in these experiments it is discretized into a manageable number of states.

To simplify our analysis, let us ignore the nonstationarity of the signals themselves; that is, assume that any variations in the STFT source-image covariance matrices $\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k, f]$ are due to motion alone. Nonstationary signal models are the subject of Chapter 7 and are applied to deformable arrays in Chapter 10. Let $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f; \theta]$ be the source-image covariance matrix for source channel n in array state θ. Then the time-varying source-image covariance matrices are given by

$$\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k, f] = \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f; \theta[k]], \quad n = 1, \ldots, N. \tag{9.1}$$

The long-term average statistics of the source channels, denoted $\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f]$, can be computed if we have a prior distribution p_θ on θ:

$$\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f] = \int_{\mathcal{X}} p_\theta(\theta) \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f; \theta] \, \mathrm{d}\theta, \quad n = 1, \ldots, N. \tag{9.2}$$

In the experiments presented here, long-term average statistics are computed directly from sample statistics of training data, so that the state space and its prior distribution are never explicitly defined.

9.4.3 Quantifying the effects of deformation

To analyze the effects of deformation, we would like to know how different two sets of source channel statistics are from each other. That is, how different is $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f;\theta_1]$ from $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f;\theta_2]$? Are they more different from each other than $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f;\theta_1]$ is from $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},2}}[f;\theta_1]$? If so, we might need to explicitly track the state of the array as it moves. Under the rank-1 model, we could simply measure the angle between the acoustic transfer function vectors $\mathbf{A}_{\mathrm{df},1}[f;\theta_1]$ and $\mathbf{A}_{\mathrm{df},1}[f;\theta_2]$. But the rank-1 model does not apply to all source channels, and the ensemble covariance matrix $\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f]$ surely has rank greater than 1. Instead, consider the Kullback-Leibler divergence between two zero-mean complex Gaussian distributions with covariances $\mathbf{R}_1$ and $\mathbf{R}_2$ [227]:

$$\mathrm{Div}(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{2}\left[\mathrm{trace}\left(\mathbf{R}_1\mathbf{R}_2^{-1} - \mathbf{I}\right) - \ln\frac{\det\mathbf{R}_1}{\det\mathbf{R}_2}\right]. \tag{9.3}$$

The term $\mathbf{R}_1\mathbf{R}_2^{-1}$ is familiar from our analysis of the squared-error performance of full-rank multichannel Wiener filters in Chapter 4. For non-parallel rank-1 covariance matrices, the divergence is infinite. If the two matrices are identical, it is zero. Although the source-image STFTs do not have Gaussian distributions in general, the Gaussian divergence is a useful way to quantify how different two distributions are from each other for the purposes of linear least-squares estimation.

9.4.4 Effect of deformation on a far-field array

Consider a set of N far-field source signals incident on an array of M isotropic sensors. Suppose for simplicity that the source signals all have power $R_{S_{\mathrm{tf},n}}[k,f] = 1$. Under the narrowband model, the STFT covariance matrices of a rigid (nonmoving) array

are

$$\mathbf{R}^{\text{rigid}}_{\mathbf{C}_{\text{tf},n}}[f] = \mathbf{A}^{\text{rigid}}_{\text{df},n}[f](\mathbf{A}^{\text{rigid}}_{\text{df},n}[f])^H \tag{9.4}$$

$$= \begin{bmatrix} e^{-j\Omega_f \tau_{n,1}} \\ \vdots \\ e^{-j\Omega_f \tau_{n,M}} \end{bmatrix} \begin{bmatrix} e^{+j\Omega_f \tau_{n,1}} & \dots & e^{+j\Omega_f \tau_{n,M}} \end{bmatrix}, \tag{9.5}$$

for $n = 1, \dots, N$, where Ω_f is the continuous-time frequency corresponding to discrete frequency index f and $\tau_{n,m}$ is the time delay of arrival for source n and channel m.

If the microphone positions are perturbed so that each $\tau_{n,m}$ is shifted by an amount $\Delta_{n,m}(\theta)$, then the new acoustic transfer functions are

$$\mathbf{A}_{\text{df},n}[f;\theta] = \begin{bmatrix} e^{-j\Omega_f(\tau_{n,1}+\Delta_{n,1}(\theta))} \\ \vdots \\ e^{-j\Omega_f(\tau_{n,M}+\Delta_{n,M}(\theta))} \end{bmatrix}, \quad n, = 1, \dots, N, \tag{9.6}$$

so that the (m_1, m_2) entry of each covariance matrix is

$$\mathbf{R}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f;\theta] = e^{-j\Omega_f(\tau_{n,m_1}-\tau_{n,m_2}+\Delta_{n,m_1}(\theta)-\Delta_{n,m_2}(\theta))} \tag{9.7}$$

$$= \mathbf{R}^{\text{rigid}}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f] e^{-j\Omega_f(\Delta_{n,m_1}(\theta)-\Delta_{n,m_2}(\theta))} \tag{9.8}$$

for $n = 1, \dots, N$. If the offsets $\Delta_{n,m}(\theta)$ have independent and identical Gaussian distributions with zero mean and variance σ^2, then using the moment-generating function for Gaussian random variables, the (m_1, m_2) entry of the ensemble covariance matrix for source n is given by

$$\bar{\mathbf{R}}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f] = \mathbb{E}_\theta \left[\mathbf{R}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f;\theta] \right] \tag{9.9}$$

$$= \mathbf{R}^{\text{rigid}}_{\mathbf{C}_{\text{tf},n},m_1,m_2} \mathbb{E}\left[e^{-j\Omega_f(\Delta_{n,m_1}(\theta)-\Delta_{n,m_2}(\theta))} \right] \tag{9.10}$$

$$= \begin{cases} \mathbf{R}^{\text{rigid}}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f], & \text{if } m_1 = m_2, \\ \mathbf{R}^{\text{rigid}}_{\mathbf{C}_{\text{tf},n},m_1,m_2}[f] e^{-\Omega_f^2 \sigma^2}, & \text{if } m_1 \neq m_2. \end{cases} \tag{9.11}$$

Because all off-diagonal elements are scaled by the same amount, we have

$$\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f] = e^{-\Omega_f^2\sigma^2}\mathbf{R}^{\mathrm{rigid}}_{\mathbf{C}_{\mathrm{tf},n}}[f] + (1 - e^{-\Omega_f^2\sigma^2})\mathbf{I}. \tag{9.12}$$

Substituting (9.12) into (9.3) and applying the Sherman-Morrison formula, it can be shown that the Gaussian divergence between two far-field source-image covariances with independent and identically distributed Gaussian offsets is

$$\mathrm{Div}(\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},1}}[f], \bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},2}}[f]) = \frac{M^2 - \left|(\mathbf{A}^{\mathrm{rigid}}_{\mathrm{df},1}[f])^H \mathbf{A}^{\mathrm{rigid}}_{\mathrm{df},2}[f]\right|^2}{2\left(e^{\Omega_f^2\sigma^2} - 1\right)\left(e^{\Omega_f^2\sigma^2} - 1 + M\right)}. \tag{9.13}$$

From this expression, we can see that the ensemble second-order statistics of the two source channels become more similar to one another as

1. their unperturbed acoustic transfer functions become more similar, for example because the sound sources are closer together,
2. the uncertainty σ^2 due to motion increases, and
3. the frequency Ω_f increases.

Deformation should have little impact if $\Omega_f\sigma$ is small, that is, if the amount of motion is small compared to a wavelength. At low audible frequencies where wavelengths are meters long, the deformation of a wearable array should have virtually no effect. At high audible frequencies where wavelengths are just a few centimeters, small motion could strongly degrade performance.

9.4.5 Experimental measurements

The effects of motion were analyzed experimentally using two deformable microphone arrays, shown in Figure 9.14. The first comprised twelve lavalier microphones hanging on cables from a pole as it was rotated back and forth by hand. The microphones swing by several millimeters relative to each other, but the overall motion is

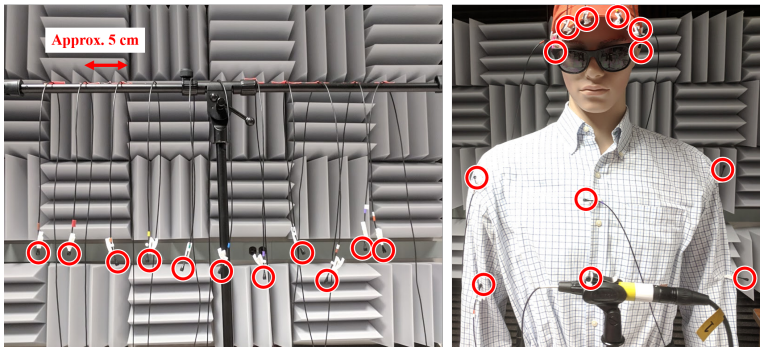

Figure 9.14: Two deformable arrays were measured in the laboratory: a set of 12 microphones hanging from a pole and a 12-microphone wearable array. Figure reproduced from [132].

one-dimensional. The second was a wearable array on a human subject. Twelve microphones were affixed to the ears, chest, shoulders, and elbows. The subject moved in different patterns. In order from most to least motion, the patterns are:

1. Dancing: moving the hips, arms, and head, but not the legs (because the floor squeaks),

2. Gesturing: moving the head and arms to simulate an animated conversation, and

3. Standing still: making a concerted effort to stand as still as possible.

A fourth experiment was conducted with the same wearable array configuration on a mannequin. Unlike the human subject, the mannequin can stand perfectly still without breathing.

Five loudspeakers were positioned around the arrays about 45° apart in a half-circle. To track the state $\theta[k]$ over time, each loudspeaker continuously emitted a different deterministic near-ultrasonic pilot signal. Time differences of arrival were computed between each loudspeaker and microphone and the vectors of these time differences were clustered into several discrete states. This tracking procedure allows us to directly measure the true state of the array rather than estimate it from the data that we are trying to process.

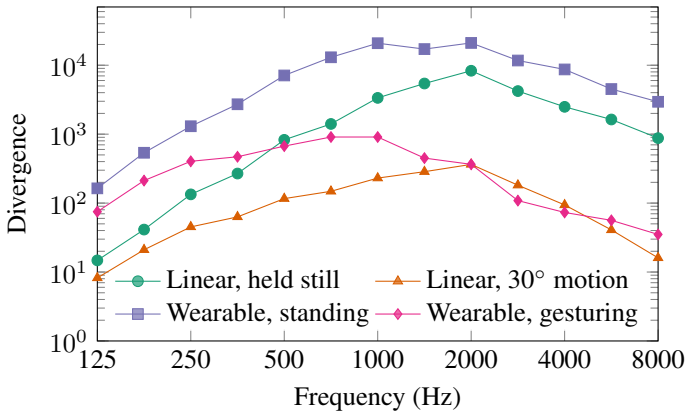

Figure 9.15: Gaussian divergence between different source channels for the two arrays. Figure adapted from [132].

To calibrate the arrays, bandlimited noise was played sequentially from each loudspeaker as the microphones were moved in consistent patterns. The ensemble covariance $\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f]$ for each source channel n was computed from the sample covariance over the full recording. The state-dependent covariance matrices $\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f;\theta]$ were computed using the measured states for each time frame.

Figure 9.15 shows the average Gaussian divergence between ensemble covariances $\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f]$ for the central loudspeaker and the four other loudspeakers. The wearable array provides better spatial diversity than the linear array because it covers a large area and includes the acoustically opaque human torso. Motion has little effect at 125 Hz because the microphones move much less than one wavelength. The penalty due to motion increases at higher frequencies, as predicted.

Figure 9.16 compares the Gaussian divergence between different pairs of covariance matrices for the linear array. The smallest divergence, shown by the curve with triangular markers, is between the ensemble covariance matrices of different source channels, $\mathrm{Div}(\bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f], \bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},3}}[f])$. The plot shows the average for the four outer loudspeakers with respect to the central loudspeaker (channel 3). The curve with circular markers is the divergence between two states for the central source: $\mathrm{Div}(\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},3}}[f;\theta_1], \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},3}}[f;\theta_2])$, where θ_1 and θ_2 represent opposite ends of the range of motion, about 90° apart. The curve with square markers shows the average diver-

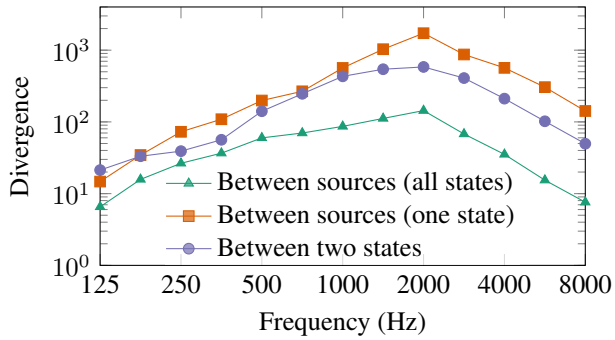

Figure 9.16: Divergence between sources and states for the hanging linear array. Figure adapted from [132].

gence between the outer and central sources in a single state, $\mathrm{Div}(\tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f;\theta], \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},3}}[f;\theta])$. As expected, it is easier to distinguish between different source channels when the state of the array is known than when it is unknown. Furthermore, the states are more different from each other for a single source than the sources are on average, suggesting that state tracking is necessary for this amount of motion.

9.5 Motion-Tolerant Processing for Deformable Arrays

9.5.1 Time-invariant and time-varying methods

Deformation has different effects on the second-order statistics of a mixture signal depending on the amount of motion and on frequency. In this section, two processing methods are compared: a time-varying filter that tracks the state of the array and a time-invariant filter that accounts for all possible states. Both are multichannel Wiener filters designed to isolate a single source channel; the signal-to-error-ratio results are averaged across the five speech sources produced by the loudspeakers.

Given a state estimate $\hat{\theta}[k]$, which in this experiment is provided by measurements of near-ultrasonic pilot signals, the time-varying STFT-domain MWF for source

channel n is

$$\mathbf{W}_{\mathrm{df},n}[k,f] = \mathbf{e}_1^T \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f;\hat{\theta}[k]] \left(\sum_{m=1}^{N} \tilde{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},m}}[f;\hat{\theta}[k]] \right)^{-1}. \tag{9.14}$$

Meanwhile, the time-invariant MWF is given by

$$\mathbf{W}_{\mathrm{df},n}[f] = \mathbf{e}_1^T \bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},n}}[f] \left(\sum_{m=1}^{N} \bar{\mathbf{R}}_{\mathbf{C}_{\mathrm{tf},m}}[f] \right)^{-1}. \tag{9.15}$$

The time-invariant filter is far less computationally complex and would not introduce artifacts from rapid time variation. However, the ensemble second-order statistics of the source channels are only useful if the amount of deformation is small.

9.5.2 Evaluation of moving-array performance

Deformable arrays present unique challenges in quantifying array performance. For most other experiments in this dissertation, we can generate realistic audio mixtures by recording sources separately and then adding their source spatial images. We can then compare the processed output of a filter against a ground-truth recording. This method works because the microphones and loudspeakers are in the same positions throughout the experiment. With a deformable array, however, the microphones cannot move in the same way for two or more recordings. Therefore, we must take a different approach.

To provide a consistent ground-truth target signal, the deformable arrays were supplemented with a nonmoving microphone used as the reference ($m = 1$). The experiments used eleven sets of recordings: five 20-second clips of pseudorandom noise, one from each loudspeaker, were used to measure the second-order statistics of the source channels. Five 20-second VCTK speech clips were then played one-at-a-time and recorded at the nonmoving reference microphone; these were used as the ground-truth output images $\mathbf{d}_1, \ldots, \mathbf{d}_5$. Finally, the five speech samples were played back simultaneously from the loudspeakers; it is this mixture data that is actually

processed by the filters.

The human subject attempted to move in similar patterns throughout the experiment so that the set of states and time spent in each state was similar between recordings. However, each recording has a unique motion pattern for the non-fixed microphones. Because this modified experiment is especially sensitive to ambient noise, the loudspeakers were amplified to a higher level than usual and the human subject was provided with extra hearing protection during the recordings.

Beamformer performance is evaluated using the signal-to-error ratio improvement over the unprocessed signal. The SER improvement is computed in the time-frequency domain and averaged over the five speech sources:

$$\Delta\mathrm{SER}[f] = \frac{1}{5}\sum_{n=1}^{5} 10\log_{10}\frac{\sum_k \left|\mathbf{e}_1^T\mathbf{X}_{\mathrm{tf}}[k,f] - D_{\mathrm{tf},n}[k,f]\right|^2}{\sum_k \left|\mathbf{W}_{\mathrm{df},n}[k,f]\mathbf{X}_{\mathrm{tf}}[k,f] - D_{\mathrm{tf},n}[k,f]\right|^2}. \tag{9.16}$$

To evaluate performance qualitatively, the wearable-array experiment was repeated using binaural beamformers referenced to the left and right ears. Because the ears move differently during every recording, there is no binaural ground truth recording against which to compare the processed output.

Because the nonmoving reference microphone is relatively far from the arrays, it is likely that the quantitative results reported here are worse than they would be with a reference microphone in the deformable array. Researchers may need to develop new methods to better measure performance with moving and deformable arrays.

9.5.3 Time-varying filtering with a linear array

For the swinging linear array of hanging microphones, the pilot signals provide time-difference-of-arrival vectors that correspond to different angles of rotation; thus, it is straightforward to define and track a set of states for different positions. Figure 9.17 compares the beamforming performance of the time-varying and time-invariant processing methods. The performances of the two methods are identical when the pole is not moving. The time-varying algorithm outperforms the time-invariant beam-

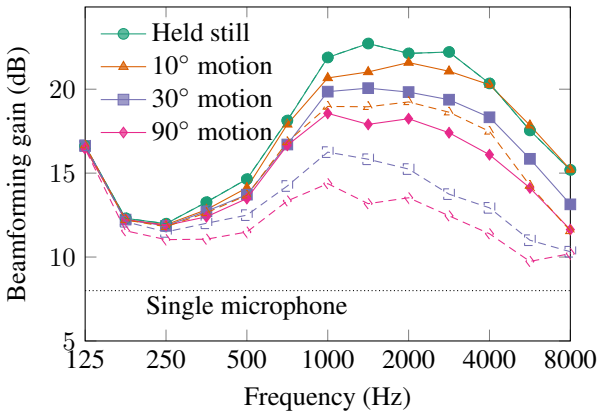

Figure 9.17: Time-invariant (dashed curves) and time-varying (solid curves) beamforming performance with a swinging linear array. Figure adapted from [132].

former at all other motion levels. The penalty due to deformation increases with the range of motion. Low frequencies, for which the array provides little benefit anyway, are largely unaffected by motion. At the highest tested frequencies, any motion at all destroys the performance of the time-invariant beamformer. The time-varying method also suffers at high frequencies, but still provides some beamforming gain.

The motion-tracking experiment was also performed with a wearable array on a human subject. Unfortunately, even with the aid of ultrasonic pilot signals, the motion of the human was too complex to track reliably. Even simple motions require a huge number of states so that the covariances in each state could not be reliably estimated. Further research will be necessary to apply explicit motion tracking to wearable microphone arrays. However, a time-varying method was successfully applied to a distributed array of *multiple* moving humans; it is described in Chapter 10.

9.5.4 Time-invariant filtering with a wearable array

Although it was not possible to explicitly track motion in the wearable-array experiment, it provides some insight about the viability of time-invariant processing for deformable microphone arrays. The time-invariant beamformer (9.15) accounts for

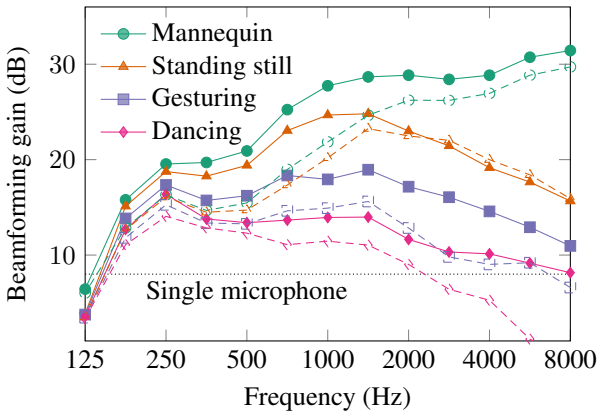

Figure 9.18: Time-invariant beamforming performance with a wearable array with different amounts of motion. The solid curves show the proposed full-rank covariance model and the dashed curves show a conventional rank-1 beamformer. Figure adapted from [132].

motion by incorporating uncertainty about the array state into a full-rank covariance matrix. Figure 9.18 compares the performance of the proposed full-rank-covariance model of deformation to a conventional rank-1 beamformer, which was computed from the principal eigenvectors of the empirical covariance matrices.

The full-rank model outperforms the rank-1 model even for the nonmoving mannequin. The full-rank covariance matrix might better model parameter estimation errors and reverberation that are not captured by the narrowband rank-1 model [131]. At higher frequencies, the full-rank model does appear to help compensate for the uncertainty due to motion. However, these benefits are not especially impressive for gesturing and dancing: at the highest tested frequencies, the full-rank model merely matches the performance of a single-microphone Wiener filter, which would provide about 8 dB of gain for this five-source mixture. The rank-1 beamformer, meanwhile, performs worse than no beamformer at all.

These experiments demonstrate that motion is a serious problem for wearable microphone arrays. While low frequencies are mostly unaffected and full-rank models can help compensate for deformation at mid-range frequencies, even tiny motion due to breathing is enough to seriously harm performance at high frequencies. To ensure

good high-frequency performance, at least part of the wearable array should have rigidly connected, closely spaced microphones.

9.6 Wearable Array Design

We can draw several conclusions from the experiments in this chapter. The performance of a wearable microphone array depends on several factors:

Separation between microphones: When microphones are farther from each other, the phase responses of different source channels tend to be more different, making the signals easier to separate and remix.

Body coverage: When microphones are on opposite sides of the body, the magnitude responses of different source channels tend to be more different, especially at high frequencies.

Deformation: When microphones move relative to each other, as they do if they are on different parts of the body, the high-frequency performance of the array suffers.

Array designers are faced with contradictory criteria. To ensure good high-frequency performance, they should make sure to have microphones surrounding the body, especially the torso. However, those microphones would not be rigidly connected to each other, which would impair high-frequency performance.

Perhaps the challenge of deformation explains the persistent popularity of microphone-array eyeglasses: although they have small aperture and provide little magnitude diversity, they are rigidly connected to the ears. Eyeglasses or headphones should perform well for frequencies above several kilohertz if there are just a few sound sources in front of the listener. To process lower frequencies and more challenging mixtures, however, an augmented-listening system would need a larger array. The better-performing Sombrearo is also rigidly connected to the head and provides much greater separation between microphones than eyeglasses do, making it an excellent

proof of concept. However, a 60 cm hat may not be practical or socially acceptable in all listening environments.

A possible solution is to use several compact, rigid subarrays placed on different parts of the body, perhaps underneath clothing. The individual subarrays would have good high-frequency performance for sources on the same side of the body, while the collective deformable array could better process lower frequencies and separate sources on different sides of the body. There are challenges with this distributed approach, however. The microphones within each subarray would not move relative to each other, but they could move relative to other subarrays and, critically, the listener's ears. A linear time-invariant filter referenced to the ears could not use high-frequency data from the other subarrays directly. Furthermore, the different subarrays would need to communicate with each other, either over a wired connection that could be cumbersome for wearers or over a wireless connection that could introduce synchronization issues. Distributed processing with multiple moving and asynchronous subarrays is the subject of Chapter 10.

Much remains to be learned about wearable arrays. How do the acoustics of the body depend on variables such as body shape and size? An expanded wearable-microphone data set should include multiple human subjects as well as more source directions and acoustic environments. Researchers would also benefit from experimental data on the array-wearer's own speech, which is difficult to simulate using mannequins and loudspeakers. We must also consider non-speech noises generated by the body itself, which may have relatively high levels for microphones in certain locations on the body. Although clothing does not seriously attenuate external sounds, it might generate severe noise when it rubs against microphones. The challenging problem of deformable-microphone-array processing remains largely unexplored. Finally, of course, there is the question of what wearable-array designs users would be willing to wear and to be seen wearing.

It is anyone's guess what the wearable microphone arrays of tomorrow will look like. They could be built into discreet vests worn under clothing or embedded in other wearable electronics such as eyeglasses, headphones, and watches. It is even possible that after this dissertation is published, giant microphone-covered hats will

become the latest fashion trend. Ideally, augmented-listening systems will aggregate data from a combination of conventional earpieces, electronic accessories, purpose-built wearable arrays, and external devices, including both deformable and rigid arrays, to provide the best possible spatial diversity and therefore the best listening experience.

Chapter 10

Cooperative Processing with Multiple Devices

This dissertation has shown that large microphone arrays have many benefits for augmented listening: with greater spatial diversity, we can process more sources with less distortion and lower delay, preserve the listener's spatial awareness, and compensate for the detrimental effects of nonlinear processing in noise. Large wearable microphone arrays can perform better than conventional earpieces for mixtures of several sources. Wearable arrays may not be enough, however, in the most challenging environments where there are many sound sources, where the parameters of the acoustic channel are difficult to estimate, and where the arrays may deform and move over time. If multiple devices spread throughout the environment could cooperate and share data, the distributed system would have far more information about the sound sources than any one device. A room-scale array could surpass the normal listening abilities of the human auditory system, delivering augmented listening experiences that would be impossible with a single device.

This chapter reviews recent work on distributed and asynchronous array processing and describes a hierarchical approach to cooperative processing using listening devices and other microphones in a space. This framework is realized in two sets of algorithms and original experiments that highlight different challenges: one, proposed by the author in [115], uses a massive-scale array to perform source separation and estimate acoustic channel parameters which are then used by a wearable device for real-time audio enhancement. Another, proposed by the author in [186], shares parameters between devices that may move relative to each other and have slightly different sample clocks.

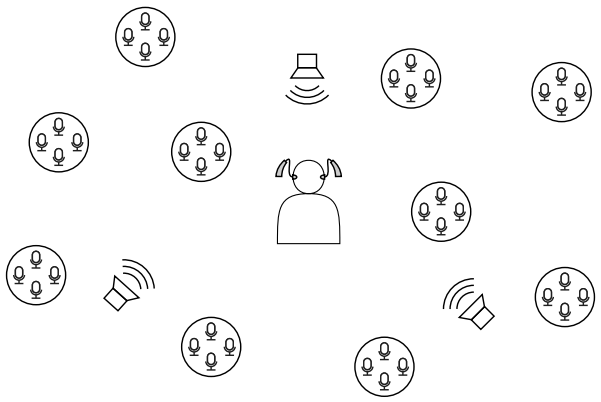

Figure 10.1: Multiple devices distributed throughout an environment could cooperate to track, separate, and enhance sound sources.

10.1 Room-Scale Microphone Arrays

Today, human environments are filled with microphone-equipped devices. Mobile phones, smart appliances, telecommunication equipment, security systems, and wearable devices often contain multiple microphones. Many of these microphones—particularly those in wearable devices worn by distant talkers—would be much closer to sound sources than a user's wearable array would be. Rather than the sound sources surrounding the microphone array, the microphones would be interspersed with the sound sources, as shown in Figure 10.1, providing far more spatial resolution. Such arrays can be used to triangulate source positions or to "spotform" and isolate sounds from a region of space rather than a direction of arrival [82, 83].

To demonstrate the potential performance of a massive-scale array in a challenging environment, consider the distributed-array data set described in Section 2.4. Ten loudspeakers and a total of 160 microphones are spread throughout a large, reverberant conference room across 4 mannequin listeners and 12 tabletop devices. Figure 10.2 shows the performance of single-target multichannel Wiener filters based on different array configurations. Each causal binaural filter is computed from ground-truth source-image statistics. The earpieces alone are nearly useless and even a large-area wearable array offers only a few decibels of improvement. Reasonable per-

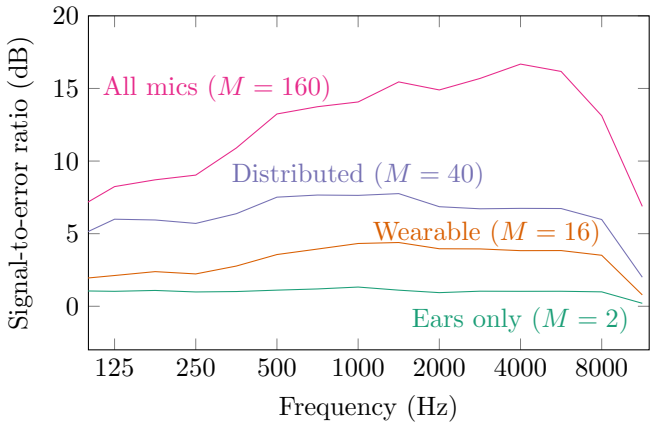

Figure 10.2: Average single-target SER for different array configurations using the distributed array data set from Section 2.4.

formance can be obtained with 40 microphones spread throughout the room. The full 160-microphone array successfully blocks unwanted speech sources so that the limiting factor is diffuse background noise.

This experiment performed coherent space-time filtering using microphones from many distributed devices. However, a listening device might not be able to use the signals from remote arrays directly for real-time filtering. Depending on the wireless protocols used to transmit the data and the distance of the remote array, the signals might or might not arrive in time to be used in a causal filter. Furthermore, each device likely uses its own sample clock circuit and the samples rates of these clocks could differ from each other by a few samples per second—enough to distort the intermicrophone phase differences that are critical for array processing. Finally, the devices may move relative to each other. The most useful information about speech signals would come from microphones worn by human talkers, which would necessarily have independent clocks, wireless connections, and motion ambiguity. Other devices, such as appliances, might have fixed locations and more-reliable network connections. In this chapter, we explore cooperative processing methods for combining information from these heterogeneous devices to improve the performance of augmented listening systems.

10.1.1 Distributed processing

Distributed microphone array processing, along with distributed sensing more broadly, have been studied extensively in recent years. Many researchers have proposed distributed computing methods for networks of devices [79, 228]. In a decentralized system, that is, one without a centralized processing hub, each device must perform a different part of the overall calculation. In these problems, there is typically limited bandwidth between devices—for example, because they are part of a wireless sensor network—so that they must choose what data to transmit. Bandwidth-constrained beamforming for wireless binaural hearing aids was considered in [70]. A more general method, suitable for large arrays, applies linear dimensionality reduction so that each device only transmits as many audio signals as there are target sources [229,230]. Crucially, each device need not estimate every individual source signal, only a subspace that contains those source signals, and each device might have a different set of target sources.

In a distributed array, some devices are closer to some sound sources than others. Thus, a distributed system can assign different sources to different devices to improve the efficiency of source separation, beamforming, or other spatial processing. In [231], a clustering algorithm is used to determine which microphones are near which sources. In [232], sources are assigned to individual devices within a network and separated using independent component analysis. In [233], devices cooperate across a network to calculate ICA updates. Clustering-based methods take advantage of large differences in source image magnitude between widely separated devices [81, 234]. An advantage of magnitude features is that different devices do not require perfectly synchronized sample rates.

10.1.2 Asynchronous arrays

One important obstacle to cooperative array processing is synchronization. Every analog-to-digital converter is driven by a sample clock, which is typically derived from a crystal oscillator. In audio devices, clocks are usually accurate within several parts

per million [205,235,236], which is accurate enough for nearly all audio applications—except array processing. Array signal processing methods, including beamforming, localization, and source separation, rely on time differences of arrival on the scale of fractions of a millisecond to distinguish sources from different directions. Sample rate mismatch between devices can cause these time differences of arrival to drift, harming the performance of the filter or other spatial processing algorithm. Distributed arrays made up of microphones with different sample clocks are known as *asynchronous arrays*.

Most research into asynchronous array processing has focused on estimating sample rate differences between devices and resampling recorded signals onto a common time scale. Small sample rate offsets can be approximated by time-varying linear phase shifts in the STFT domain [205] and compensated using opposite phase shifts [205]. Other methods model phase drift of intermicrophone coherence and apply time-domain resampling [84, 236]. The method used for the baseline experiments in this chapter splits the difference, estimating sample rate offsets in the STFT domain but correcting for them in the time domain [235].

Resampling-based methods are suitable for offline processing of prerecorded signals, but they are difficult to apply to real-time processing. Furthermore, sample rate estimation is strongly sensitive to motion [205, 237], as will be shown in Section 10.3. There are a few proposed methods that do not require sample rate estimation. In [80, 81], clustering methods and nonnegative matrix factorization are used to generate time-frequency masks, which do not depend on phase differences between devices. In the method proposed in [186] and described in Section 10.4, a sparsity-based nonlinear filtering algorithm uses phase information only within synchronous subarrays, then aggregates likelihood features between devices.

10.1.3 Hierarchical processing

A further challenge in cooperative array processing is latency. As explained in Chapter 5, listening devices must process sound within just a few milliseconds to avoid disturbing distortion. A key conclusion of that chapter is that large-scale distributed

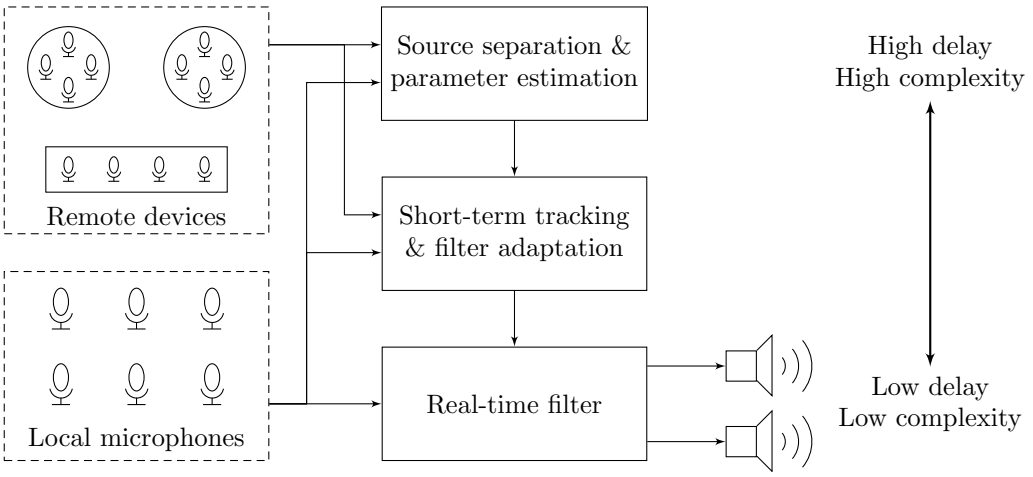

Figure 10.3: A cooperative processing system operates on multiple scales of distance, time, and computational complexity.

arrays can reduce delay—possibly to zero!—by observing source signals several milliseconds before they arrive at the listener's ears. However, this benefit requires a low-latency wireless link. The Bluetooth standard that is widely used for consumer audio devices today has latency of several tens of milliseconds, making it unsuitable for such a network. Future wireless technologies may have much lower latency. In practice, different remote devices will likely have different latencies with respect to the listener's device. Similarly, some devices might have synchronous sample clocks and others might not, and different devices may have different bandwidths and computational capabilities.

To address this heterogeneity in the cooperative listening system, we can use a hierarchical processing architecture, shown in Figure 10.3. The microphones in the network are divided into two categories: local microphones that can be used for synchronous causal filtering, for example those within the listening device, and remote microphones that provide useful information but cannot be used for filtering because they have high-latency or low-bandwidth connections or because of relative motion or sample rate mismatch. Only local microphones are used to perform filtering and generate the output presented to the user. The remote microphones are used to learn about the parameters of the system and generate or adapt the local filters.

Processing tasks are also divided into a hierarchy according to time scales and computational demands. The space-time filters that generate the output have strict delay constraints and must be implemented on the listening device itself. Fortunately, these filters are not computationally intensive. In the laboratory prototype described in Section 9.1, binaural outputs are generated by finite impulse response filters implemented in programmable logic. This processing level might also include filterbanks or Fourier analysis and synthesis for time-frequency processing.

At the other extreme, blind source separation or acoustic channel estimation algorithms operate on time scales of tens of seconds. They may be too computationally demanding to run on an embedded device, but because the channel parameters do not vary too quickly, they could be executed on an external device or even on a cloud service. These long-term parameter-estimation algorithms can aggregate data from the entire distributed array. They might also integrate non-audio data from cameras or other sensors and interface with the users of the system.

Depending on the algorithms used, there may also be a third intermediate level of the cooperative processing hierarchy. This level operates on time scales of tens to hundreds of milliseconds so that it can track short-term changes in signal spectra, apply dynamic range compression, or track small motion in deformable microphone arrays. It is responsible for computing and updating the filter coefficients used by the delay-constrained space-time filter. It could be implemented on the listening device itself or an external processor. This level may or may not have access to data from the full array, depending on the capabilities of the devices.

The remainder of this chapter describes two examples of cooperative processing that highlight different challenges. In Section 10.2, we consider a fully synchronous, nonmoving array in a large, reverberant room. The local device performs causal linear time-invariant filtering using only its local microphones. The local filters are designed using channel estimates from noncausal blind source separation algorithms and data from the entire array. This system includes the lowest and highest levels of the hierarchy in Figure 10.3. In Section 10.4, we consider a different problem: the acoustic channel parameters are known and the devices can share complete data in real time, but they have uncertain sample rate offsets and the arrays can move

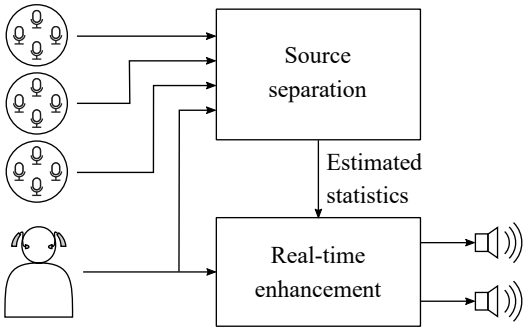

Figure 10.4: Remote devices are used to perform source separation and to estimate slowly varying acoustic channel parameters. Only local microphones are used for real-time listening enhancement. Figure adapted from [115].

relative to each other. Data from remote arrays cannot be used directly for space-time filtering because of phase ambiguities, so information from these asynchronous devices is aggregated to determine the state of a nonlinear filter based on the high-low model. This second problem illustrates the lower and middle levels of the hierarchy. Further work will be required to integrate these methods and demonstrate a complete end-to-end cooperative listening system.

10.2 Cooperative Listening Enhancement with Room-Scale Arrays

Human environments are increasingly filled with network-connected, microphone-equipped devices. In this section, we consider a large, reverberant room with a large number of talkers but an even larger number of microphone arrays. These distributed devices can cooperate to achieve better spatial-processing performance than any one device alone. This section is based on the author's work in [115].

10.2.1 Hierarchical source separation and enhancement

A typical room might have a combination of nonmoving audio devices, such as smart speakers and telecommunication equipment, and moving devices, such as mobile phones and wearable accessories. Fixed devices are more useful for sound source localization, separation, and tracking because they can leverage prior information about their microphones' locations and the acoustics of the environment. However, a listening device must rely primarily on its local microphones for filtering to achieve imperceptible delay and to preserve interaural cues.

A cooperative listening enhancement system can combine the strengths of local and remote devices, as shown in Figure 10.4. All microphones in the network are used to estimate acoustic channel parameters, in this case the discrete-time source-image correlation matrices $\mathbf{r}_{\mathbf{c}_{\mathrm{d},n}}[\ell]$ for $n = 1, \ldots, N$, for the microphones of the local array. These matrices are used to compute binaural discrete-time causal multichannel Wiener filters (Section 4.1.3).

10.2.2 Source separation with a room-scale array

A room-scale distributed array has the advantage that many microphones are much closer to the sources than the listener is. These nearby microphones enjoy a higher signal-to-noise ratio and direct-to-reverberant ratio. Some source separation methods, such as [231,232], leverage this spatial diversity by assigning different sources to different subsets of microphones. Here, these nearby microphones are used as reference signals. Let m_n^* be the microphone closest to sound source n for $n = 1, \ldots, N^*$, where $N^* \leq N$ is the number of directional sound sources. Assume that the distributed array is large enough that each source has a unique nearest microphone. Let $\tilde{S}_{\mathrm{tf},n}[k, f] = \mathbf{e}_{m_n^*}^T \mathbf{C}_{\mathrm{tf},n}[k, f]$ be the source signal as received by the nearest microphone for directional sources $n = 1, \ldots, N^*$.

Let $\hat{S}_{\mathrm{tf},n}[k, f]$ be an estimate of the nearest-microphone source image for $n = 1, \ldots, N^*$. These estimates are generated using all the microphones of the distributed array. Performing blind source separation on such a large scale remains an open

problem. The experiments in the next section compare three methods:

1. The unprocessed mixture at the nearest microphone, $\hat{S}_{\mathrm{tf},n}[k,f] = \mathbf{e}_{m_n^*}^T \mathbf{X}_{\mathrm{tf}}[k,f]$,
2. The output of the AuxIVA blind source separation method [196] (Section 8.1.4) initialized with the nearest-microphone estimate, and
3. An ideal linear unmixing filter designed from the ground-truth source images.

10.2.3 Local array statistics

Using these STFT-domain reference signal estimates, the second-order statistics of the source images at the local array can be estimated:

$$\hat{R}_{\tilde{S}_{\mathrm{tf},n}}[f] = \mathrm{mean}_k |\hat{S}_{\mathrm{tf},n}[k,f]|^2 \tag{10.1}$$

$$\hat{\mathbf{R}}_{\mathbf{X}_{\mathrm{tf,local}}\tilde{S}_{\mathrm{tf},n}}[f] = \mathrm{mean}_k \mathbf{X}_{\mathrm{tf,local}}[k,f]\hat{S}^*_{\mathrm{tf},n}[k,f] \tag{10.2}$$

$$\hat{\mathbf{R}}_{\mathbf{X}_{\mathrm{tf,local}}}[f] = \mathrm{mean}_k \mathbf{X}_{\mathrm{tf,local}}[k,f]\mathbf{X}^H_{\mathrm{tf,local}}[k,f]. \tag{10.3}$$

These statistics provide narrowband rank-1 models for the directional sources. Specifically, the discrete-frequency transfer function for source channel n relative to the nearest microphone to source n (not relative to an in-ear microphone) is given by

$$\hat{\mathbf{A}}_{\mathrm{df,local},n}[f] = \hat{\mathbf{R}}_{\mathbf{X}_{\mathrm{tf,local}}\tilde{S}_{\mathrm{tf},n}}[f]\hat{R}^{-1}_{\tilde{S}_{\mathrm{tf},n}}[f]. \tag{10.4}$$

At this point, the sampled relative impulse responses $\hat{\mathbf{a}}_{\mathrm{d,local},n}[k]$ can be windowed to exclude late reverberation.

The relative (early) transfer functions and empirical source statistics are used to estimate the rank-1 discrete-frequency power spectral density matrices of the local source images:

$$\hat{\mathbf{R}}_{\mathbf{c}_{\mathrm{d,local},n}}[f] = \hat{\mathbf{A}}_{\mathrm{df,local},n}[f]\hat{R}_{\tilde{S}_{\mathrm{tf},n}}[f]\hat{\mathbf{A}}^H_{\mathrm{df,local},n}[f], \quad n = 1,\ldots,N. \tag{10.5}$$

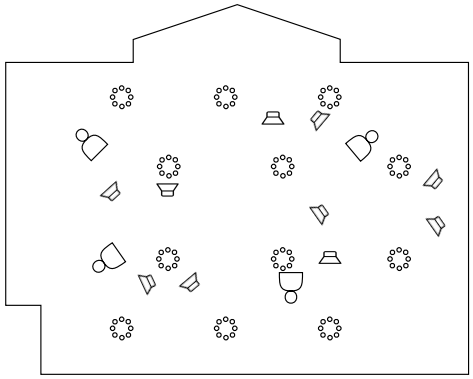

Figure 10.5: A distributed array of 160 microphones is spread among 10 loudspeakers in a large, reverberant conference room.

The statistics of the remaining diffuse noise sources are assumed to be measured in advance, for example when the room is empty.

Applying the inverse discrete Fourier transform yields length-F estimates of the discrete-time correlation matrices $\mathbf{r}_{\mathbf{c}_{\mathrm{d,local},n}}[\tau]$ for $n = 1, \ldots, N$. These in turn are used to compute the causal discrete-time filter $\mathbf{w}_{\mathrm{d}}$, derived in Section 4.1.3, that estimates the desired output from the local microphone signals $\mathbf{x}_{\mathrm{d,local}}[k]$:

$$\hat{\mathbf{y}}[k] = \sum_{\ell=0}^{L} \mathbf{w}_{\mathrm{d}}[\ell]\mathbf{x}_{\mathrm{d,local}}[k-\ell]. \tag{10.6}$$

10.2.4 Experimental setup

The cooperative source separation and listening enhancement system is demonstrated using the large distributed array data set of Section 2.4. The local microphones are the 16 microphones on a wearable array, which covers the ears, eyeglasses, torso, and wrists. The full distributed array has up to 160 microphones spread across 12 smart-speaker enclosures and 4 wearable arrays. There are 10 loudspeakers spread throughout the room facing different directions. Each loudspeaker has at least one microphone-array device within about one meter in front of it, as shown in Figure 10.5. The loudspeakers play VCTK speech samples from different talkers.

Mixtures were generated with a variable number of speech source images plus a recording of diffuse background noise in the room, which was primarily from the ventilation system. The noise is stronger than the speech sources below 100 Hz and about 20 dB weaker than them at higher frequencies. Source separation and acoustic channel estimation were performed using a 16-second mixture. The source separation algorithm has prior knowledge of the total number of sources and the index of the nearest microphone to each source, but no other information about the source or microphone geometry.

The estimated source statistics were used to design causal FIR enhancement filters with length 128 ms and delay 16 ms. The relative early impulse responses were windowed to length 32 ms; changes in the REIR length do not appear to have a strong effect on performance. The resulting enhancement filters were tested on a mixture of different 16-second speech samples from the same talkers as the training data.

Because the filters are designed to enhance only the early parts of the source images and exclude late reverberation, it is difficult to compute a ground-truth target output for this experiment. Therefore, performance is evaluated using signal-to-noise ratio improvement, denoted ΔSNR, for each of $2N^*$ single-target binaural beamformers, one for each ear and each directional source:

$$\mathrm{SNR}_{n,j}^{\mathrm{in}} = 10\log_{10}\frac{\sum_k \left|\mathbf{e}_j^T \mathbf{c}_{\mathrm{d},n}[k]\right|^2}{\sum_k \left|\sum_{m\neq n} \mathbf{e}_j^T \mathbf{c}_{\mathrm{d},m}[k]\right|^2} \tag{10.7}$$

$$\mathrm{SNR}_{n,j}^{\mathrm{out}} = 10\log_{10}\frac{\sum_k \left|\mathbf{e}_j^T \hat{\mathbf{d}}_{\mathrm{d},n}[k]\right|^2}{\sum_k \left|\sum_{m\neq n} \mathbf{e}_j^T \hat{\mathbf{d}}_{\mathrm{d},m}[k]\right|^2} \tag{10.8}$$

$$\Delta\mathrm{SNR}_{n,j} = \mathrm{SNR}_{n,j}^{\mathrm{out}} - \mathrm{SNR}_{n,j}^{\mathrm{in}}, \tag{10.9}$$

for $j = 1, 2$ and $n = 1, \ldots, N^*$. The experiment was repeated for each of the four listeners for a total of $8N^*$ performance measurements for each set of sources.

Table 10.1: Source separation performance (dB SNR improvement).

		Ideal			IVA		
Array	M	$N^* = 4$	7	10	$N^* = 4$	7	10
Reference	10	13	12	10	5	4	3
Wearable	64	25	26	23	8	7	3
Tabletop	96	23	23	21	7	6	5
All mics	160	23	24	23	8	7	6

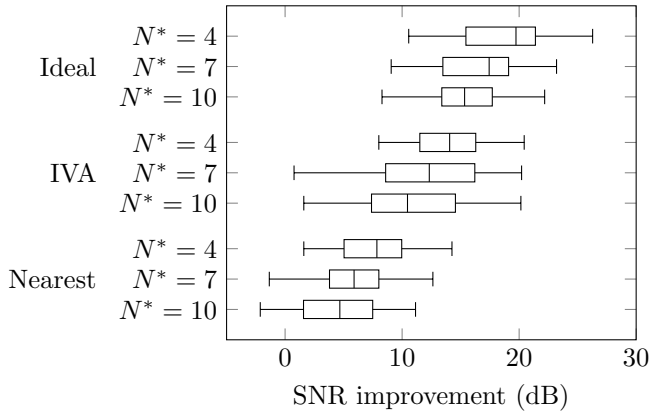

Figure 10.6: Listening enhancement performance of the local causal space-time filters designed using different acoustic channel estimation methods. The quartile statistics are shown over $8N^*$ source-ear combinations. Figure adapted from [115].

10.2.5 Experimental results

First, consider the performance of the source separation algorithm at estimating $\tilde{s}_n(t)$. Table 10.1, which is adapted from [115], shows the average SNR improvement of $\hat{s}_n$ compared to the unprocessed mixture at the reference microphone for different array sizes. The source separation algorithm used either the 10 nearest microphones only, the 64 wearable microphones only, the 96 tabletop microphones only, or all 160 microphones together. There is a clear benefit to using arrays with $M \gg N$ for both the ideal and blind source separation methods.

Figure 10.6 shows the SNR improvement of the causal 16-microphone enhancement filters for different channel-estimation methods and different numbers of direc-

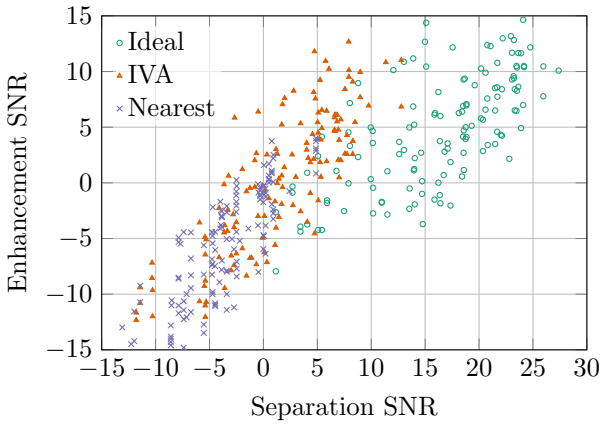

Figure 10.7: Relationship between source separation performance and listening enhancement performance. Each point represents a combination of one sound source, one ear, and one acoustic channel estimation method. Figure adapted from [115].

tional sources. The statistics are shown over all $8N^*$ source-ear combinations. For each experiment, the lower SNR-improvement values generally correspond to distant source-listener pairs and higher SNR-improvement values are for sources close to and facing the listener.

Using the unprocessed remote microphone signals as references, the listening devices can improve SNR by around 5–10 dB. Using blind source separation to help design the filters provides about 5 dB further improvement. An ideal room-scale separation filter provides another 5 dB, showing that there is room for improvement in the blind source separation algorithm.

The listening enhancement filters are designed to isolate the reference signals identified by the source separation algorithm. The system's enhancement performance is therefore limited by the performance of the separation algorithm. Figure 10.7 shows the SNR performance of the causal space-time listening filter as a function of the SNR performance of the source separation algorithm. Each point represents one combination of the ten target source signals and eight ears. There is a clear positive relationship between the separation SNR and enhancement SNR. There are diminishing returns for the ideal signal estimate; the limiting factors might be reverberation

and noise.

These experiments show that, given a good estimate of the reverberant channel parameters, wearable microphone arrays can perform high-quality delay-constrained spatial processing even in large, reverberant rooms with many distant sound sources. Furthermore, remote microphones can help to estimate these channel parameters even if they are not used directly in the resulting filters. However, more work is required to develop scalable blind source separation techniques that work with large numbers of sources and microphones.

10.3 Deformable and Asynchronous Arrays

The large-scale experiment in the previous section shows that distributed microphone arrays can enhance human listening even in the most challenging environments with strong reverberation and many competing sound sources. However, the arrays used in that experiment were all perfectly synchronized and did not move. To realize the benefits of cooperative array processing, we must also account for motion and synchronization issues. Although they seem to be unrelated problems, synchronization and motion both result in a relative phase ambiguity between microphones and so they have related, though not identical, solutions.

10.3.1 Sample rate offsets in the time-frequency domain

If the devices in a distributed microphone array are connected wirelessly, then each device most likely has its own sample clock generated by an internal circuit and crystal. Suppose that the nominal sample period is T_s, so that the ideal sampled signal $\tilde{x}_{\mathrm{d},m}$ at each microphone m would be

$$\tilde{x}_{\mathrm{d},m}[k] = x_m(kT_\mathrm{s}), \quad m = 1, \ldots, M. \tag{10.10}$$

In reality, each microphone has a true sample period T_m for $m = 1, \ldots, M$. The actual sampled signals are therefore

$$x_{\mathrm{d},m}[k] = x_m(kT_m), \quad m = 1, \ldots, M. \tag{10.11}$$

The nominal and true samples are related to each other by a Doppler shift, which warps the time axis and therefore the frequency spectrum. Let $X_{\mathrm{tf},m}[k, f]$ and $\tilde{X}_{\mathrm{tf},m}[k, f]$ be the STFTs of $x_{\mathrm{d},m}$ and $\tilde{x}_{\mathrm{d},m}$, respectively. Because of the sample rate offset, the k and f indices of the two STFTs correspond to different time intervals and different continuous-time frequencies. If the offset is large, then STFT-domain array processing is effectively impossible.

If the offset is small, as offsets between audio devices typically are, then the effect can be modeled linearly in the STFT domain [205,235]. Assume that the sample rate offset is much smaller than the analysis bandwidth of the discrete Fourier transform, that is, $|\frac{1}{T_m} - \frac{1}{T_\mathrm{s}}| \ll \frac{2\pi}{FT_\mathrm{s}}$, so that each frequency index $f \in \{0, \ldots, F-1\}$ corresponds to nearly the same continuous-time frequency for all microphones and the continuous-time duration of the STFT windows differ by much less than a sample period between microphones. Suppose that the time scales are coarsely synchronized—say, within a few samples—so that the continuous-time intervals corresponding to each STFT window index k are nearly the same for each microphone. Then the sampled and nominal-scale STFTs are approximately related by a phase shift

$$X_{\mathrm{tf},m}[k, f] \approx e^{j\alpha_m[k,f]} \tilde{X}_{\mathrm{tf},m}[k, f], \quad m = 1, \ldots, M. \tag{10.12}$$

If the STFT-domain samples of the observed signals are regarded as zero-mean random variables that do not depend on the α's, then the cross-correlation between microphone signals is

$$\begin{aligned} R_{X_{\mathrm{tf},m} X_{\mathrm{tf},\ell}}[k, f] &= \mathbb{E}\left[X_m[k, f] X_\ell^*[k, f]\right] && (10.13) \\ &\approx \mathbb{E}\left[e^{j(\alpha_m[k,f] - \alpha_\ell[k,f])} \tilde{X}_{\mathrm{tf},m}[k, f] \tilde{X}_{\mathrm{tf},\ell}^*[k, f]\right] && (10.14) \\ &= \mathbb{E}\left[e^{j(\alpha_m[k,f] - \alpha_\ell[k,f])}\right] R_{\tilde{X}_{\mathrm{tf},m} \tilde{X}_{\mathrm{tf},\ell}}[k, f]. && (10.15) \end{aligned}$$

The next step depends on whether we view the α's as deterministic parameters that can be measured or as random parameters that contribute to the uncertainty in the system.

10.3.2 Estimating sample rate offsets

The time-scaling effect of a small sample rate offset can be modeled as time-varying time shifts of STFT frames. Assuming a common time origin at $k = 0$, the phase shift $\alpha_m[k, f]$ is approximately linear [205, 235]:

$$\alpha_m[k, f] \approx \frac{2\pi k f T_{\text{step}} T_{\text{s}}}{F} \left(\frac{1}{T_{\text{s}}} - \frac{1}{T_m} \right), \tag{10.16}$$

where T_{step} is the step size, in samples, between frames and F is the length of the discrete Fourier transform.

In the most-studied version of the asynchronous array processing problem, the sources and microphones do not move and the source signals have at least approximately stationary long-term statistics. Then from (10.15) and (10.16), the cross-correlation between microphone signals is

$$R_{X_{\text{tf},m} X_{\text{tf},\ell}}[k, f] \approx e^{-j \frac{2\pi k f T_{\text{step}} T_{\text{s}}}{F} \Delta_{m,\ell}} R_{\tilde{X}_{\text{tf},m} \tilde{X}_{\text{tf},\ell}}[f], \tag{10.17}$$

where $\Delta_{m,\ell} = \frac{1}{T_m} - \frac{1}{T_\ell}$.

The phase of this cross-correlation varies predictably as a function of time and frequency. Thus, it can be estimated from the sample cross-correlation between observed microphone signals. In [84], $\Delta_{m,\ell}$ is derived from the rate of change of the phase of the coherence between microphone signals. In [205], the STFT samples are modeled as Gaussian random variables and $\Delta_{m,\ell}$ is obtained by a maximum likelihood estimator.

The resampling experiments presented here use the two-step correlation maximization method of [235], which combines ideas from [205] and [84]. First, the sample rate offset is coarsely estimated by applying compensatory linear phase shifts in the

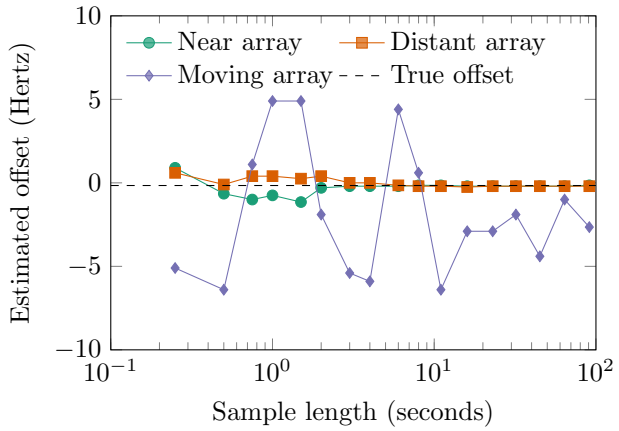

Figure 10.8: Estimated sample rate offset of a handheld recorder relative to a studio interface. The offset is estimated using the method of [235]. Figure adapted from [186].

STFT domain and maximizing the sample correlation coefficient:

$$\hat{\Delta}_{m,\ell}^{\text{coarse}} = \arg\max_{\Delta} \text{mean}_f \text{Corr}_k \left(X_m[k,f], X_\ell[k,f] e^{-j\frac{2\pi k f T_{\text{step}} T_{\text{s}}}{F}\Delta} \right). \tag{10.18}$$

Because the STFT-domain offset model (10.17) is only an approximation, the signal is resampled in the time domain using Lagrange interpolation [84]. The estimation process is then repeated once more on the resampled signal to obtain a fine estimate of the sample rate offset.

10.3.3 Sample rate offsets and relative motion

In a distributed array of devices that rarely move, such as the tabletop smart-speaker enclosures from the distributed array data set, it should be possible to reliably estimate the devices' relative sample rates by analyzing long recordings of background noise. That is an important advantage of distributed room-scale arrays for cooperative augmented listening: the sensor network can include fixed microphone arrays with known geometry.

The task is more difficult for deformable arrays, that is, arrays in which devices

can move relative to each other. The time-stretching effect caused by a difference in sample rates is mathematically identical to the time-stretching effect of constant-velocity motion, either of the microphones or of the sound sources. It has been observed that sample rate offset estimation is more difficult for moving sound sources [205, 235]. It is also challenging if the microphones move relative to each other, as they would in a wearable microphone array [186].

To demonstrate the impact of relative motion, the estimation algorithm of [235] was evaluated using two independently clocked microphone arrays in a cocktail-party scenario with several speech sources. One array of lavalier microphones was sampled by the studio interface. The second device was a handheld stereo recorder with its own internal clock. First, the recorder was placed next to the other microphone array. Next, it was placed on the other side of the room. Finally, it was carried around the room in a haphazard path at walking speed. Figure 10.8 shows the estimated sample rate offset as a function of sample length. The algorithm converges with just a few seconds of data if the microphones do not move, regardless of their locations. However, it fails to converge at all when the handheld array is moving.

It was recently proposed to estimate sample rate offsets with moving sound sources by identifying time frames during which the sources are stationary and using only those frames for estimation [237]. A similar approach could be applied to deformable microphone arrays, especially if some of the microphones are known to be fixed. However, cooperative array processing would be easier if the system could avoid estimating sample rate offsets at all.

10.3.4 Second-order statistics of asynchronous arrays

Suppose that the relative phase between two microphones m and ℓ is unknown, either because of sample rate offsets or because the microphones can move relative to each other. The most pessimistic model of this phase uncertainty is that the phase offsets $\alpha_m[k, f]$ and $\alpha_\ell[k, f]$ are independent random variables uniformly distributed on $[0, 2\pi]$. This model is probably overly pessimistic at low frequencies, but is reasonable for deformable microphone arrays at high frequencies, as demonstrated in Chapter

9. It is also useful for real-time embedded systems that do not have the time or computational resources to estimate and compensate for sample rate offsets.

If the α's are independent uniform random variables, then from (10.15), the cross-correlation between the sampled STFTs is

$$R_{X_{\text{tf},m}X_{\text{tf},\ell}}[k,f] = \mathbb{E}\left[e^{j(\alpha_m[k,f]-\alpha_\ell[k,f])}\right] R_{\tilde{X}_{\text{tf},m}\tilde{X}_{\text{tf},\ell}}[k,f] \tag{10.19}$$

$$= 0 \tag{10.20}$$

for $m \neq \ell$. That is, the unknown phase offsets cause the microphone signals to become uncorrelated with each other.

This relationship also applies to the correlation matrices of source images. If every microphone pair had a random phase offset, then every $\mathbf{R}_{\mathbf{C}_{\text{tf},n}}[k,f]$ would become a diagonal matrix. Consider a monaural-output MSDW-MWF designed using diagonal source-image covariance matrices:

$$\mathbf{W}_{\text{df}}[k,f] = \sum_{n=1}^{N} \lambda_n G_{\text{df},n}[k,f]\mathbf{e}_1^T\mathbf{R}_{\mathbf{C}_{\text{tf},n}}[k,f]\left(\sum_{m=1}^{N}\lambda_m\mathbf{R}_{\mathbf{C}_{\text{tf},m}}[k,f]\right)^{-1} \tag{10.21}$$

$$= \frac{\sum_{n=1}^{N}\lambda_n G_{\text{df},n}[k,f]\mathbf{e}_1^T\mathbf{R}_{\mathbf{C}_{\text{tf},n}}[k,f]\mathbf{e}_1}{\sum_{n=1}^{N}\lambda_n\mathbf{e}_1^T\mathbf{R}_{\mathbf{C}_{\text{tf},n}}[k,f]\mathbf{e}_1}\mathbf{e}_1^T. \tag{10.22}$$

This is simply a scalar weighted Wiener filter applied to the first microphone; it does not use the other microphones at all.

If the sound signals had Gaussian marginal distributions, then independent random phase offsets would make the microphone signals statistically independent and therefore remote microphones would provide no information at all. Fortunately, most of the sounds about which listeners might care, such as speech, are not Gaussian. While random phase offsets make the microphone signals uncorrelated and therefore useless to a linear estimator, the signals are not independent: they are, after all, mixtures of the same source signals. Even if the phases are unreliable, we can obtain useful information from the signal magnitudes. To do so, however, we must apply nonlinear methods.

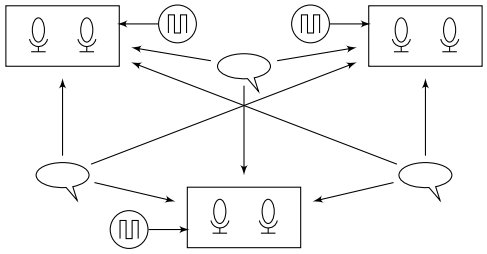

Figure 10.9: In a partially asynchronous array, each device has multiple microphones that share a common sample clock. Figure adapted from [186].

10.4 Cooperative Nonlinear Processing for Partially Asynchronous Arrays

It is often difficult or impossible to estimate the relative sample rates of different devices. It is also computationally expensive to correct these small sample rate offsets. What if we could use information from remote microphones without estimating sample rate offsets and without resampling the recorded signals? If we cannot rely on phase information, then such methods must be nonlinear. Researchers have previously proposed nonlinear asynchronous source separation methods that rely on STFT magnitudes. For example, in [80], nonnegative matrix factorization is applied to the magnitudes of signals from microphones near the different sources. For sparse signals, the devices can share statistics that are used to determine source activity [186, 234].

Most asynchronous processing methods proposed in the literature, both resampling-based and magnitude-based, assume that each individual microphone has an independent clock. However, most modern electronic devices contain at least two microphones. These devices form *partially asynchronous arrays* [186] that include groups of microphones, each with a common sample clock, as shown in Figure 10.9. Thus, while there are phase uncertainties between groups, phase-coherent processing can be applied within each group.

The method desribed here was originally developed by the author to compensate for sample rate offsets [186], but it also applies to relative motion. The microphones within each group are either rigidly connected or only slightly deformable, but the

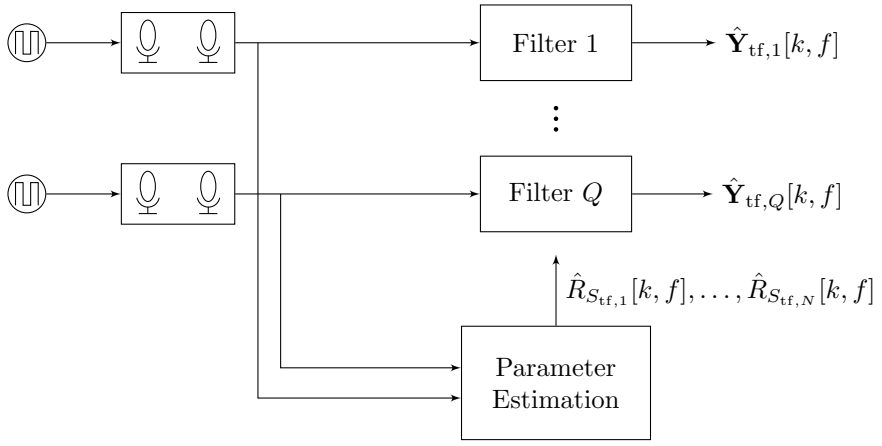

Figure 10.10: A hierarchical asynchronous source separation system estimates time-varying source spectra using the entire distributed array, but separation filters act on local microphones only. Figure adapted from [186].

groups can move relative to each other. In the experiments presented here, the distributed array suffers from both sample rate offsets and relative motion.

10.4.1 Hierarchical time-frequency processing with the local Gaussian model

To combine phase-coherent information from the listener's local array with asynchronous information from remote arrays, let us apply the spatially stationary local Gaussian model from Section 7.1.4:

$$\mathbf{R}_{\mathbf{C}_{\mathrm{tf},n}}[k,f] = R_{S_{\mathrm{tf},n}}[k,f]\bar{\mathbf{R}}_n[f], \quad n = 1,\ldots,N, \tag{10.23}$$

where $R_{S_{\mathrm{tf},n}}[k,f]$ is the time-varying scalar source power, which is the same for all devices, and $\bar{\mathbf{R}}_n[f]$ is the long-term spatial correlation matrix for source channel n. Because of random phase offsets, $\bar{\mathbf{R}}_n[f]$ is block diagonal. Under the local Gaussian model, therefore, the received signals are conditionally independent between devices given $R_{S_{\mathrm{tf},1}},\ldots,R_{S_{\mathrm{tf},N}}$.

To account for deformation of the arrays, each block of $\bar{\mathbf{R}}_n[f]$ is allowed to have

full rank. This model is useful for motions that have small effects on the relative phases between the microphones within each device but large effects on the relative phases between devices. In the experiments presented here, $\bar{\mathbf{R}}_n[f]$ will be estimated using either training data or a blind source separation method.

To estimate the time-varying covariance matrices at each $[k, f]$, the cooperative system must find an estimate $\hat{R}_{S_{\text{tf},n}}[k, f]$ of the shared power parameter for each source channel. The time-varying covariances are used to design a time-varying space-time filter

$$\mathbf{W}_{\text{df}}[k, f] = \sum_{n=1}^{N} \lambda_n \hat{R}_{S_{\text{tf},n}}[k, f] \mathbf{G}_n[k, f] \bar{\mathbf{R}}_n[f] \left(\sum_{n=1}^{N} \lambda_n \hat{R}_{S_{\text{tf},n}}[k, f] \bar{\mathbf{R}}_n[k, f] \right)^{-1}. \tag{10.24}$$

Because $\bar{\mathbf{R}}_n$ is block diagonal for all n, if $\mathbf{G}_n$ is nonzero only for rows corresponding to local microphones—which would be the case for an interaural-cue-preserving listening device—then $\mathbf{W}_{\text{df}}$ is also nonzero only for local microphones. That is, remote microphones are not directly processed by the space-time filters; instead they are used to estimate parameters that are used by the local filters, as shown in Figure 10.10.

10.4.2 Source activity classification with the high-low model

Remote microphones are used to estimate the instantaneous source spectra $\hat{R}_{S_{\text{tf},n}}[k, f]$. There are many possible approaches to estimating these spectra, including compositional models and learning-based classifiers. To take advantage of the distributed array, however, we should use a spatial classifier. In [234], the instantaneous spectra are estimated using the expectation maximization algorithm based on posterior probability estimates computed by each device. Here we describe a similar approach using the high-low model introduced in Section 7.3 [186].

Under the local Gaussian model with high and low source states, the log-likelihood

Table 10.2: Signal-to-error ratios, in dB, of several source-image-estimation methods on the SiSEC ASY data set. Table adapted from [186].

Filter	Sampling	$N = 3$	$N = 4$
Unprocessed		-3.0	-5.0
Local array only		0.7	0.3
Time-invariant MWF	Not resampled	2.1	0.1
Time-invariant MWF	Resampled	8.2	2.9
Cooperative method	Not resampled	5.5	2.2
Cooperative method	Resampled	5.5	2.2

of state n^* for sample $[k, f]$ is given by

$$\ln \Pr(\mathbf{X}_{\text{tf}}[k,f]|n^*) = -\mathbf{X}_{\text{tf}}^H[k,f] \left(R_{\text{high},n^*}[f]\bar{\mathbf{R}}_{n^*}[f] + \sum_{n \neq n^*} R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f] \right)^{-1} \mathbf{X}_{\text{tf}}[k,f] - \ln \det \left(\pi R_{\text{high},n^*}[f]\bar{\mathbf{R}}_{n^*}[f] + \pi \sum_{n \neq n^*} R_{\text{low},n}[f]\bar{\mathbf{R}}_n[f] \right). \tag{10.25}$$

Because each $\bar{\mathbf{R}}_n[f]$ is block-diagonal, that is, because the signals at each device are modeled as conditionally independent given n^*, the overall log-likelihood can be written as a sum of individual log-likelihoods for each local array. Thus, each device need only transmit log-likelihood statistics for each source channel, not full audio data. The likelihoods can be used to either select a single state with maximum likelihood or compute a posterior probability to weight the state estimates.

10.4.3 Experimental results on the SiSEC ASY data set

The proposed cooperative asynchronous source separation method was applied to the 2018 Signal Separation Evaluation Campaign (SiSEC) asynchronous source separation (ASY) task [175]. Four talkers were recorded using four handheld recorders, each with two microphones and an independent sample clock. Because the talkers

do not move and processing is performed offline, it is possible to estimate the sample rates of the devices and resample the signals to a common time scale.

The proposed partially asynchronous method is compared against a baseline method that does perform resampling. It combines a few algorithms used by participants in SiSEC ASY 2015 [94], which used the same data set as SiSEC ASY 2018. First, the sample rate offsets are estimated using two-stage correlation maximization [235] and the signals are resampled using Lagrange interpolation [84]. Next, the resampled signals are separated using offline independent vector analysis [196]. The same offline blind source separation algorithm is used to estimate the acoustic channel parameters used for the cooperative nonlinear method. While it would be difficult to perform sample rate estimation and resampling in real time on an augmented listening device, those tasks could be performed by a fusion center that can store several seconds or minutes of data.

The estimated acoustic channel parameters were used to design separate eight-output, single-target filters to estimate the source images for each source channel, as required by the SiSEC criteria. That is, the filter targeting source n has $\mathbf{G}_n[f] = \mathbf{I}$ for that source channel and zero for all others. Table 10.2 shows the signal-to-error ratio (in SiSEC terminology, this would be the "signal-to-distortion ratio," or SDR) achieved by the baseline and proposed algorithms, with and without a resampling step. The resampled time-invariant filter performs best out of all methods because it can perform fully coherent processing on all eight microphones. However, it performs poorly if the signals are not resampled. The cooperative nonlinear method has slightly worse performance than the resampled eight-microphone filter, but because it does not rely on phase relationships between recording devices, it is unaffected by the sample rate offset. Although it requires synchronization for the offline channel estimation step, the online part of the filter can operate directly on the sampled data without performing costly interpolation.

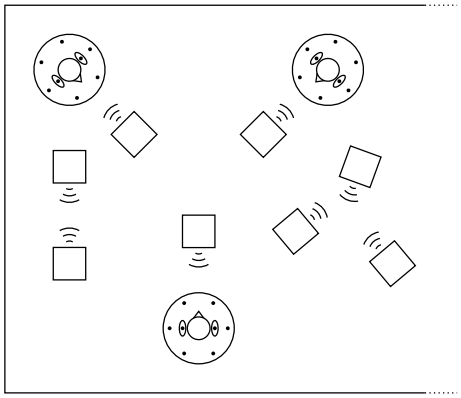

Figure 10.11: Three moving human listeners with wearable microphone arrays cooperate to separate sound from eight loudspeakers. Figure adapted from [186].

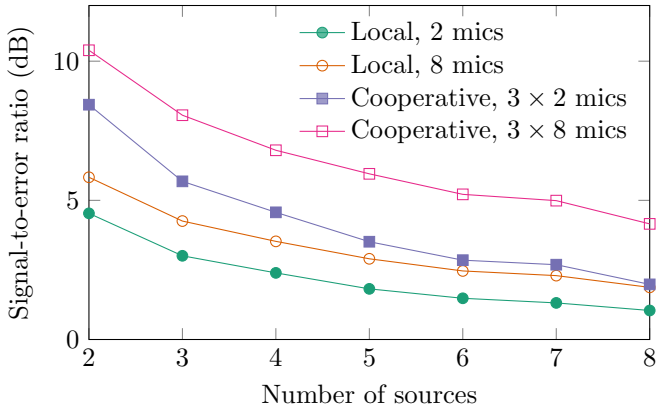

Figure 10.12: Single-target binaural enhancement performance using three moving wearable arrays. Figure adapted from [186].

10.4.4 Experimental results with moving wearable arrays

Partially asynchronous processing is also useful when devices can move relative to each other, as they would in any system involving wearable arrays. These arrays would typically have different sample rates but, unlike in the SiSEC ASY data set, their motion would make them difficult to synchronize. Furthermore, the phase ambiguities caused by deformation are similar to those caused by sample rate offsets. To demonstrate the potential of nonlinear cooperative methods for multiple moving arrays, the proposed algorithm was evaluated in a cocktail-party scenario with up to eight simultaneous speech sources and three moving human listeners. Because there are no known blind source separation or acoustic channel estimation methods suitable for moving microphones, the acoustic channel parameters were measured using training data in a similar way to the deformable-array experiments in Chapter 9.

The experimental setup is shown in Figure 10.11. Eight loudspeakers produced 20-second speech clips from the VCTK data set. Each listener wore an eight-microphone array with one sensor in each ear and six around the 60-cm brim of the Sombrearo. During each recording, the listeners slowly nodded and turned their heads back and forth. The three listeners were simulated by a single human subject who moved to a new location in between recordings; thus, the data do not fully reflect the acoustic effects of the bodies of three human listeners. The source images were recorded independently and without a fixed external reference microphone. The human subject attempted to move in a similar pattern throughout the experiment, but the motion patterns do differ between source images. Although the resulting mixtures are physically impossible, because the source images are modeled as uncorrelated, their statistics can still be described by the ensemble spatial covariance matrices $\bar{\mathbf{R}}_n[f]$. These matrices were estimated using 5 seconds of speech data for each source. The remaining 15 seconds were used to create the test mixtures. Simulated sample rate offsets of ± 0.3 Hz were applied to two of the three arrays.

As in the SiSEC experiment, the system generated single-target source-image estimates for each source channel. However, because this experiment uses listening de-

vices, images were only estimated for the left and right ears, not for the microphones on the hat. The results are shown in Figure 10.12 for four filters: a time-invariant filter using the ears only ($M = 2$), a time-invariant filter using a single wearable array ($M = 8$), a cooperative nonlinear filter using the ears from all three listeners ($M = 6$), and a cooperative filter using all 24 microphones. For mixtures of many sources, the ears alone can do little better than guess. The full hat performs slightly better. However, the large number of sources and the motion of the array make it difficult for any one array to separate the sources.

The cooperative method performs better than the static filters, even with a smaller total number of microphones, because the microphones spread around the room can provide better spatial information. For example, the listener in the upper left corner of Figure 10.11 would have trouble distinguishing between the three sources on the right side of the diagram. However, the other two listeners are better positioned to decide which of those sources is active at a given time and frequency. By pooling their observations, they can better understand the state of the sound sources. Furthermore, because the listeners move independently, their redundant observations can resolve ambiguities caused by motion. The time-varying method based on the high-low model lets each individual array make the most of its limited degrees of freedom while benefiting from the spatial diversity of the full array.

10.5 Cooperative Processing for Augmented Listening Devices

This chapter presented different approaches to cooperative array processing that highlight the advantages and challenges of distributed arrays. Room-scale arrays can outperform individual listening devices, even large wearable arrays, because they are spread around and among sound sources. In the conference-room experiment, microphones in front of each sound source act as high-SNR, low-reverberation references that help to estimate channel parameters and separate sources. In the moving-array experiment, devices spread around the room have different abilities to resolve clusters

Table 10.3: Types of cooperative processing.

Latency	Deformation	Synchronization	Bandwidth	Cooperative method
High	–	–	–	Parameter estimation only (Section 10.2)
Low	Fixed	Synchronous	High	Fully coherent processing (Section 10.1)
			Low	Distributed processing algorithms (Section 10.1.1)
		Asynchronous	High	Offset estimation and resampling (Section 10.3)
			Low	Cooperative nonlinear processing (Section 10.4)
	Moving	–	–	

of sound sources, so they can pool their observations to make better decisions. The distributed devices also provide redundancy that helps to reduce uncertainty due to motion.

Cooperative array signal processing is more complex than space-time filtering with a single static array. Different processing tasks require different amounts of data, computational resources, and delay and may be distributed over multiple devices with different capabilities. In both the conference-room and SiSEC experiments, the acoustic channel parameters were measured using offline, computationally expensive blind source separation algorithms that aggregate data from the full array. This step would best be implemented by a high-bandwidth, high-compute fusion center, possibly hosted by a cloud service. Meanwhile, each listening device produces its own output signal from causal, low-delay filters acting on its local microphones but incorporates information from the rest of the array.

The amount and type of cooperation used to perform listening enhancement depend on the capabilities of the system: there are different cooperative processing strategies for different combinations of latency, bandwidth, synchronization, and deformation, as shown in Table 10.3. If the remote devices cannot provide real-time data to the listening device, as in the conference-room experiment of Section 10.2,

then those microphones can be used only to estimate slowly varying acoustic channel parameters.

If the latency of the communication link is lower than the propagation time of the acoustic signal, then recorded data could be used directly by the listening device. In a low-latency, high-bandwidth, fully synchronous, nonmoving system, all microphones would be considered local and the space-time filter could operate on the entire room, as in the fully coherent experiment in the introduction of Section 10.1. If bandwidth is limited, then the distributed array processing algorithms of 10.1.1 can be used to compress the transmitted data. If the devices use different sample rates but do not move and have ample computational capabilities, then the sample rates can be estimated and the signals resampled, as in the asynchronous array processing literature summarized in Section 10.3. If the devices cannot be resampled or if they move relative to each other, then they can share statistical information used to infer the states of sparse source signals, as proposed in Section 10.4.

Cooperative array processing is most useful if it can take advantage of microphone arrays already present in a space. These will necessarily have different capabilities and can be used in different ways to assist a listening device. For example, infrastructural devices such smart appliances, security systems, and teleconferencing equipment have fixed locations, known geometry, and high bandwidth, and their sample rates could be synchronized with each other. These arrays are therefore ideal for localization and blind source separation. Wearable and mobile devices would be more difficult to synchronize, suffer from motion and deformation, and may have limited bandwidth and computational capabilities; but, because they are typically much closer to human talkers, they would provide valuable information about speech signals. Further research is required to understand the advantages and challenges of processing speech from a talker who is wearing a microphone array.

The work in this chapter has touched on only a few of the possibilities of cooperative array processing. Room-scale microphone arrays should dramatically improve the performance of blind source separation algorithms in noisy and reverberant environments, but highly scalable BSS algorithms remain to be developed. Distributed arrays, especially those with a combination of fixed and moving devices, could be

the key to tracking the state of deformable devices. Furthermore, to realize the benefits of cooperative processing in public spaces, devices from different vendors and different users must share data with each other. There are therefore important compatibility and privacy issues that must be addressed.

In a few years, it may be possible for an augmented listening device user to walk into a crowded room and connect with dozens of other listening devices and infrastructural arrays already in the space. With access to hundreds of distributed microphones, the listener could focus on any sound or combination of sounds they choose, even if those sounds come from the opposite end of the room. The user could listen to signals that would be hopelessly inaudible to the unaided ear or to a conventional listening device. It could be cooperative processing, rather than wearable arrays, machine-learning models, or clever sparse algorithms, that will finally enhance human hearing far beyond its normal limits.

Chapter 11

Developing Augmented Listening Systems

In the first half of this dissertation, we saw that large microphone arrays can help listening devices to meet the unique demands of human listening enhancement applications. Larger arrays have more degrees of freedom with which to preserve the spectral balance and interaural cues of multiple sound sources in complex mixtures. They can apply independent dynamic range compression to different source channels, reducing the distortion effects that plague commercial hearing aids. Large-aperture microphone arrays can achieve the same level of performance with lower delay than smaller devices, and distributed microphone arrays can enhance sound with zero delay by capturing sound well before it reaches a listener.

With recent advances in technology, it is now possible to build large wearable microphone arrays that cover the entire body, not just the head. As audio devices proliferate in human environments, there are ample microphones that could be incorporated into a distributed array. However, as explained in the second half of the dissertation, large-scale wearable and distributed arrays are difficult to realize in practice. They suffer from sample rate mismatch and relative motion, which create phase uncertainty and interfere with spatial processing. Therefore, remote microphones cannot be used directly by a linear time-invariant filter. Instead, distributed devices can be used for parameter estimation and tracking or for nonlinear cooperative processing. Both of these methods assist the listening device in performing real-time filtering with its own local microphones.

This chapter summarizes the major results of the dissertation and synthesizes them into a set of design principles, research priorities, and next steps that will help us to realize better-than-human performance in augmented listening technology.

11.1 Performance Tradeoffs and Design Principles

Any engineering system is characterized by design tradeoffs. This dissertation has characterized several signal processing tradeoffs, while there are many perceptual tradeoffs that must be studied by hearing specialists.

11.1.1 The listening experience

When we use binoculars to look at something far away, we cannot see over as wide an area; we trade field of view for magnification. Similarly, a listening device can be used to amplify distant sounds at the expense of other sounds. When we define the desired responses of a remixing filter or multisource compressor, we make a tradeoff between augmentation and naturalness.

At one extreme is a single-target beamformer that isolates one sound source of interest as much as possible; this approach was used in early microphone array hearing aid work in the 1990s and early 2000s. If the goal is to improve the intelligibility of one talker in a noisy environment, this system will achieve that, but at a cost. It circumvents the auditory system's natural scene analysis abilities and destroys potentially useful information about the environment. At the other extreme is pure amplification, which preserves information from other sounds but helps very little in noisy environments.

This dissertation proposed a compromise: apply different processing to different sound sources depending on the user's listening objectives. The listener could explicitly adjust the desired processing, for example using a slider that trades off between augmentation and naturalness. With new perceptual research, engineers could design algorithms that tune the processing automatically in different environments based on the listener's hearing profile and preferences.

The user might also explicitly choose between different listening modes. For example, some modes might be:

Everyday enhancement: This mode emphasizes transparency and comfort. Most sounds are minimally processed. Unimportant background noise, such as an

airplane engine or air conditioner, is attenuated. Compression is applied only to limit uncomfortably loud sounds.

Conversation mode: This mode attenuates background noise just enough so that a target talker remains intelligible. This mode could include side-chain compression that ducks noise sources when the target talker is active.

Bubbleforming: Suitable for a crowded restaurant or a conference poster session, this mode preserves all sound sources within a meter or two of the listener and attenuates most others.

Augmented reality: Virtual sound sources are added to the environment or certain sound sources are replaced, for example by translating them to a different language. Other sounds in the environment are preserved.

The specific desired responses applied to each source channel would be determined based on a classifier that decides what kind of signal it is and whether the user should care about it. The system will also use customized perceptual models. For example, hearing-impaired listeners require a higher signal-to-noise ratio for speech to be intelligible.

Some settings might be adjusted automatically using artificial intelligence features. For example, a distant talker who would otherwise be attenuated could gain the listener's attention by saying their name. A speech recognition and natural language processing system could also analyze the suppressed speech signals to determine if they might contain important information, such as a shouted warning.

11.1.2 Signal processing design

Performance objectives

When designing space-time filters for microphone array listening devices, we must trade off between several performance objectives. Each objective can be improved

at the expense of the others, with overall performance depending on the available degrees of freedom of the system.

Spectral distortion: A change in the frequency spectrum of sound sources, especially background sources, due to frequency-selective processing. In underdetermined mixtures, less distortion of one source implies more distortion of other sources.

Spatial distortion: A change in the interaural cues of sound sources. Constraints on spectral distortion also help to reduce spatial distortion, but the reverse is not true.

Across-source modulation: When performing dynamic range compression, the degree to which a change in the level of one source affects the level of other sources. It is large when signals are compressed jointly.

Error sensitivity: Degradation in filter performance due to erroneous estimates of acoustic channel parameters or signal statistics. Beamformers that are more directive are also more sensitive to error.

Motion sensitivity: Degradation in filter performance due to deformation or motion of the microphones in an array. It is mathematically similar to error sensitivity.

Delay: Total hardware and algorithmic delay between a sound arriving at the ear and being played back through the receiver. Greater frequency selectivity requires greater algorithmic delay.

Artifacts: Nonlinear processing can introduce disturbing artifacts, sometimes called musical noise, when filters change too quickly.

Design parameters

We can trade off between these different performance metrics by adjusting the design parameters of the filter:

Desired responses: If the desired responses to different sources are more similar to each other, most of the performance metrics listed above will improve, but the system might provide less benefit to the user.

Distortion weights: The distortion weights of the MSDW-MWF directly trade off between spectral and spatial distortion of different source channels.

Filter lag: The estimation lag designed into the causal space-time filter directly tunes the delay of the system. Larger lag provides better squared-error performance.

Model order: A full-rank spatial covariance or power-spectral-density model improves sensitivity to error and deformation compared to a rank-one model but does not distinguish as strongly between different source channels. A motion-tracking filter can either use a few states with broad statistics or many states with narrow statistics.

Sparsity: In nonlinear systems, the threshold level of the source activity detector or the ratio between high and low variance states of the high-low model can be used to adjust the degree of nonlinearity in the system. Both methods automatically become more linear as the spatial diversity of the array improves.

Providing additional information

These tradeoffs are quite restrictive for conventional listening devices with few microphones. To improve all of the listed performance objectives at once, we must provide the system with more information by adding more microphones.

We can add more microphones to the listening device by using wearable microphone arrays that cover the body. With more spatial degrees of freedom, the filter can apply distortion constraints to more sound sources or increase the rank of each source's statistical model. Microphones spread across the body provide amplitude diversity that is less affected by motion and modeling errors.

We can use cooperative processing between multiple devices to further improve spatial resolution. Remote microphones are especially effective at reducing delay because they can capture sound sources before they reach the listener. They also help to reduce acoustic parameter estimation errors because they enjoy higher signal-to-noise ratios and direct-to-reverberant ratios for distant sound sources. Nonlinear cooperative processing can help to compensate for uncertainty due to relative motion and sample rate mismatch.

11.1.3 Listening device design

There are further tradeoffs in the design of the listening device itself, including between aesthetics and performance. The experiments in this work show that larger microphone arrays offer better performance. Large wearable devices, such as the Sombrearo, provide better spatial resolution than smaller wearables like eyeglasses. Conventional hearing aid earpieces, designed to be discreet, cannot perform meaningful spatial processing. The industry's emphasis on invisibility appears to be out of step with consumer trends, especially among young technology enthusiasts who would be the target market for an augmented listening system designed for normal-hearing people. These consumers happily wear bulky, ostentatious audio gadgets; an elegant and stylish wearable array might appeal to them, even if it is quite large.

Microphones can also be spread across the body. Because the body is effective at blocking sound, sensors spread around the torso provide excellent spatial diversity. Experiments suggest that these microphones could be placed under clothing, which may be more cosmetically acceptable than external accessories for some users. However, this work did not address potential problems due to noise when clothing rubs against hidden microphones.

Microphones spread across the body are also vulnerable to relative motion. A body-scale deformable array is only useful at low frequencies, where acoustic wavelengths are much longer than the scale of motion between microphones. One solution is to design several rigid microphone arrays worn on different parts of the body. Earpieces, eyeglasses, and hats move rigidly with the head and so can be used directly

for space-time filtering referenced to the ears. Microphone arrays embedded in vests, belts, watches, and other wearable accessories might move relative to the head, but could be used directly for filtering at low frequencies. At high frequencies, they can be used as part of a nonlinear cooperative algorithm. To support cooperative processing with other listening devices, a wearable array could include a directional microphone pointed toward the user's mouth.

A large wearable listening device could have far more computational power than the tiny chips embedded in traditional hearing aids. Statistical space-time filtering is highly parallel and would benefit from specialized computing hardware, but could be implemented on a digital signal processor with many signal inputs. To ensure that sample clocks remain synchronized and to reduce latency, microphones should be physically wired to the processor wherever possible. Short-range analog wireless technology such as near field magnetic induction could also be used for wearable devices.

11.1.4 System architecture

To describe the design of the overall listening system, let us return to the system architecture proposed in Chapter 1, which is reproduced in Figure 11.1. Having analyzed each piece in depth throughout the dissertation, we can now describe how the pieces fit together.

Cooperative processing network

To realize the most ambitious listening experiences proposed above, such as sound source replacement or bubbleforming, the system must have far greater spatial diversity than could be achieved with a wearable array. We have seen that cooperative processing between devices can provide that diversity. The role of each device in the distributed network depends on its bandwidth, latency, motion, and synchronization. Some common devices that might be included are:

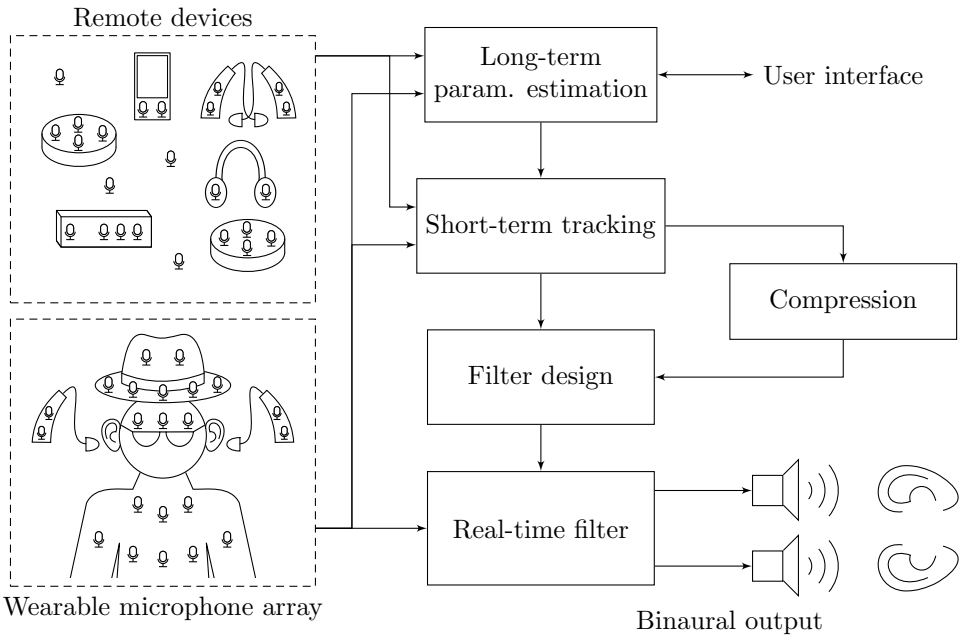

Figure 11.1: The proposed augmented listening system.

Other listening devices: An obvious starting point for cooperative processing is to connect multiple wearable listening devices together. Each device would have similar specifications and capabilities, ensuring compatibility. A major advantage of wearables for cooperative processing is that they are positioned directly next to speech signals. A wearable array that includes microphones on the face or chest would have an excellent signal-to-noise ratio and direct-to-reverberant ratio for the wearer's speech. Signals from remote wearable arrays are the most difficult to process, however, because wearable devices are wireless—meaning that they must have mismatched sample rates—and because humans move constantly.

Existing devices: An opportunistic system could leverage audio devices already in a space, such as smart speakers, appliances, gaming and conferencing systems, intercoms, and security cameras. While these devices may be wireless, they move infrequently, and therefore it would be possible to learn their sample

rates over time. Their positioning within the environment is arbitrary, but could also be learned over time to improve performance, especially for sound source localization and tracking.

Infrastructure: To achieve exceptional augmented listening performance, a space could be deliberately instrumented with hundreds of microphones. For example, small microphone-array tiles could be embedded into walls, ceilings, and furniture at deliberately chosen locations. In this setup, the sensors could be hard-wired to a powerful central processing node, allowing the system to perform fully synchronous source separation with known array geometry.

In a cooperative system, existing devices or infrastructural arrays would be best suited to perform source separation and to track moving sound sources, while wearables would provide perceptual transparency to listeners.

Parameter estimation

The distributed array and powerful computing resources are used to track slowly varying system parameters such as acoustic channel statistics and to identify and classify sound sources. Rapidly varying parameters could be tracked either by the listening device itself or by a remote device with a high-bandwidth, low-latency connection to the listening device.

In a fully integrated system, short-term parameter estimation would unify the source activity classifier or detector used in sparse methods with the envelope detector used in dynamic range compression. Short-term spectral estimates would be used both to allocate the filter's degrees of freedom most efficiently and to determine the time-varying desired processing to be applied to different sources. Further work is required to unify these two types of processing. In particular, compression systems typically have attack times much shorter than the sparsity-maximizing STFT window size of 60 ms; the release times, meanwhile, are much longer.

These short-term parameter estimates are used to update the coefficients of a causal space-time filter. Because the STFT has too much delay for real-time listening

enhancement, new methods are required to incorporate time-frequency sparsity into delay-constrained systems.

11.2 The Future of Listening Technology

Although hearing aids are a well-established technology, the broader category of augmented listening is just beginning to emerge. Listening devices could change radically in the future as they both incorporate and influence future technologies.

11.2.1 Missing pieces and future features

Some of the methods developed in this dissertation are not yet fully compatible with each other. In particular, further research is required to implement time-varying methods with strict delay constraints. The results of Chapter 5 suggest that large arrays should be able to operate with lower delay because they have less need for spectral selectivity. Ideally, time-varying methods would automatically adjust the degree of sparsity that they model and therefore the delay that they require based on the available spatial resolution of the array.

New perceptual models are needed to help set parameters including target delay, compression ratios, and the relative gain applied to each source channel. These models would need to account for the user's hearing profile and the acoustic environment. For example, it is likely that hearing-impaired listeners in loud noise can tolerate more delay but require more noise suppression than normal-hearing listeners in quiet. Similarly, detailed models of intelligibility could be used to implement side-chain compression that attenuates noise only at times and frequencies where it would mask a signal deemed more important.

This dissertation contained no analysis or experiments involving the listener's own speech. Poor processing of the listener's speech, such as a long time delay, can impair speech production. In a cooperative array, wearable devices could help to capture speech and transmit information to remote listeners. It is difficult to study own-

speech processing in the laboratory because mannequins do not have loudspeakers, loudspeakers do not have ears, and real humans cannot be used for repeatable experiments. The Augmented Listening Laboratory team is developing head-shaped loudspeakers with ears that could be used for controlled own-speech experiments.

An augmented listening system will need to identify sound sources and decide which source or sources the user wishes to hear, perhaps using machine learning algorithms. For example, a natural extension of the keyword-based acoustic channel measurement technique in Chapter 8 is a customizable keyword-spotting algorithm that detects the user's name and focuses on that speech source. A user might also configure a listening device to always amplify—or suppress—the speech of a specific talker.

11.2.2 Role of future technologies

There are several emerging technologies that could impact the design and performance of augmented listening systems. The most immediately applicable may be augmented reality (AR) systems, which can provide the listening system with non-audio information about the user's environment. Object detection and motion tracking could be used to follow talkers as they move. Even better, the motion tracking that is essential for visual AR and VR systems could be used to track the position of a head-mounted microphone array relative to sound sources. In fact, some augmented reality headsets already include microphone arrays.

Many other technologies will help to build better cooperative array processing systems. Next-generation wireless systems should support large numbers of nearby devices with greater bandwidth and lower latency than today's wireless networks. These could facilitate truly massive-scale distributed microphone arrays with low enough latency to be used in delay-constrained listening devices. Similarly, high-speed, low-latency networks will allow the listening device to offload intensive processing to powerful cloud or edge computing devices. Distributed sensing and computing would enable demanding source separation techniques that would be impossible on an earpiece.

Augmented listening systems would also benefit from new microphone technologies. Digital MEMS microphones already have time-division-multiplexing features that permit several microphones to be daisy-chained together and connected to a single port. In the future, thin-film microphone arrays could be embedded into walls or tabletops [238] to create instrumented rooms. Wearable devices would also benefit from yet-to-be-invented microphones that could be embedded directly into fabric.

Finally, while machine learning alone is not enough to improve human hearing, new advances in machine learning could complement the spatial information provided by large microphone arrays. For example, learning-based models of speech can improve the performance of sparse source activity classifiers. Acoustic event classification and natural language processing can help a listening device to decide which sound sources might be of interest to a listener. Learning-based algorithms could also use feedback from a listener to customize processing parameters for them.

11.2.3 Research priorities

This work has identified several open research problems that must be addressed in order to realize powerful augmented listening systems. First, as explained above, signal processing researchers must develop ways to implement time-frequency processing in delay-constrained systems and to integrate it with dynamic range compression.

Another urgent problem for wearable devices is to design array processing methods that are robust to deformation. In Chapter 9, it was proposed to model small motion using a full-rank spatial covariance matrix. However, there are many other possible approaches to designing time-invariant filters that are robust to deformation. For larger motion, we will need time-varying methods that explicitly track the relative positions of sources and microphones. Large cooperative arrays—which could include fixed devices with known geometry—will likely help to track larger motion. In augmented reality platforms or instrumented rooms, the system could take advantage of multimodal data, for example from video, lidar, or inertial sensors.

Large distributed arrays could also help with arguably the greatest impediment to superhuman augmented listening: learning acoustic channel parameters. While many

array signal processing researchers are devoting research resources to bandwidth-constrained distributed computing algorithms, there are no known blind source separation algorithms that can scale to leverage hundreds of microphones, even with unlimited bandwidth and perfect synchronization. The new data sets developed as part of this work may help other researchers to investigate massive-scale source separation. Researchers should also consider network latency alongside bandwidth as a critical constraint in wireless sensor networks for delay-constrained applications.

One topic that was not addressed in this work, but that will be critical to real-world success of augmented listening systems, is privacy. The audio remixing system developed here assumes that a user might want to listen to any sound source, even from a great distance. There are obvious privacy concerns in building a system that can listen from across a room. Even if the goal is to suppress those distant sounds, the system would need access to potentially sensitive signals from remote devices. A cooperative listening system would be more practical if there were privacy-preserving inference algorithms [239] that could perform acoustic channel estimation without having direct access to signal content.

To understand what types of processing should be applied by the listening device, new clinical studies must be performed. System designers need to understand tradeoffs between delay and intelligibility or quality in noisy and reverberant environments. While most intelligibility studies focus on a single talker in noise, the proposed remixing system could preserve multiple sound sources; little is known about how humans can attend to multiple talkers in noisy conditions, and it is not clear how such a study would be designed. Finally, of course, the proposed system must eventually be validated using clinical trials.

11.3 Broader Applications of This Work

Some of the challenges addressed in this work, such as dynamic range compression and interaural cue preservation, are unique to human augmented listening. Most, however, are not. Some challenges, such as delay constraints and deformation, are es-

pecially pronounced for augmented listening but are relevant to many signal processing applications. Many of the topics addressed in the second half of the dissertation, including time-varying methods and channel estimation, are problems facing spatial audio capture in general rather than listening technology specifically. The results presented here could be useful in many applications beyond augmented listening.

11.3.1 Machine listening

Machine listening algorithms perform a similar function as the human auditory system: they extract actionable information from audio recordings. Thus, any processing that can help a human listener to hear better in noisy environments would also be useful for machine listening. The primary difference is that machine listening algorithms do not have the intrinsic scene analysis abilities of the human brain, so systems that perform automatic speech recognition or sound event classification often rely on single-target spatial filters rather than perceptually transparent remixing filters. These filters do not have the stringent delay, distortion, dynamic range, and spatial constraints that listening filters do.

One method in this work, acoustic channel estimation from speech keywords, is specifically designed for speech recognition applications. It can be used by keyword-activated voice assistants to improve performance in noisy or reverberant conditions where today's products do not function well. The method could be extended to work with customizable keywords or large-vocabulary speech recognition.

Array-based machine listening systems could benefit from many of the nonlinear processing methods proposed in the second half of the dissertation. The source activity detector and high-low model of Chapter 7 offer scalable alternatives to the binary masks used by many single-channel source separation algorithms. The cooperative asynchronous source separation system proposed in Chapter 10 would benefit any distributed processing system that relies on ad hoc arrays of devices with different sample rates. Unlike many state-of-the-art asynchronous methods, it does not require precise sample rate estimation or resampling.

11.3.2 Array signal processing

Array signal processing is used for many non-audio applications. Arrays could be made of antennas or ultrasonic transducers, for example. Spatial audio processing differs from other types of array signal processing in some ways: audio signals have much wider relative bandwidth than most radio frequency signals, and many interesting sound signals such as speech exhibit time-frequency sparsity. While some nonlinear source separation methods are tailored to audio, the results related to linear time-invariant filters can apply to any sensor array.

The theoretical results on delay-performance tradeoffs in Chapter 5 apply to any space-time filter, regardless of the signal type. Delay-constrained spatial filtering could be relevant to ultra-low-latency wireless communication systems, for example. The analysis of deformable microphone arrays in Chapter 9 could apply to any sensor array on a flexible substrate, such as antenna arrays in bendable mobile devices or wearable accessories.

The cooperative processing systems described in Chapter 10 could apply to a variety of wireless sensor networks. Mobile, smart-home, and internet-of-things devices are rapidly proliferating in homes, workplaces, and public spaces. Next-generation wireless networks will eliminate many of the bandwidth and latency constraints that limit these networks today, but they will still be subject to uncertainty due to motion and synchronization. These networks could benefit from partially asynchronous cooperative processing.

11.3.3 Wearable technology

The market for wearable electronics has expanded rapidly over the last several years. While there have been more failures than successes in the marketplace, wearable audio devices have been consistently popular. Devices that can capture and analyze the acoustic scene around a user have many applications beyond augmented listening.

Wearable microphone arrays could be used to do high-resolution spatial-audio recordings. Dense microphone arrays, often spherical, are used to capture and spa-

tially encode sound for virtual reality experiences. Wearable arrays with many microphones could capture similar data, and they have an advantage over standalone arrays: they include microphones in or near the ears, allowing them to capture real binaural audio.

Microphone arrays are already integrated into some augmented reality headsets. They are used to localize and analyze sound events and to associate them with physical objects. Larger wearable arrays would allow AR systems to more precisely localize and separate sounds. The wearable microphone array data set should help AR researchers and developers to devise new applications that take advantage of high-resolution spatial sound capture.

11.4 A Telescope for the Ears

Engineers have long dreamed of building devices that can augment human hearing the way that microscopes and telescopes augment human vision. Yet we have never been able to improve upon the remarkable capabilities of the human ear. Using large microphone arrays with far greater spatial resolution than the ears, superhuman hearing may finally be possible.

Early microphone arrays failed because they were not much larger than the head, so they provided little information that was not already available to the ears. They also tried to apply single-target beamforming methods that are poorly suited to the auditory system. The space-time filters proposed here can supplement rather than replace humans' natural auditory scene analysis functions, providing a more natural listening experience. Filter designers can use the theoretical tools developed in this work to analyze tradeoffs among delay, distortion, and dynamic range.

To improve the performance of listening systems, device designers must move beyond conventional hearing aid earpieces—and researchers must move beyond eyeglasses! The design principles discussed in this work will help designers to develop novel form factors with microphones spread all across the body. The wearable listening devices of tomorrow could be discreet vests worn under clothing or ostentatious

fashion accessories protruding from the body. The wearable microphone data set collected as part of this work will help engineers and designers to simulate and analyze these new devices.

To achieve dramatically better performance than current technology, listening devices can cooperate with each other and with other devices in the environment. While much more work is required, the experiments presented here show that room-scale microphone arrays are useful in challenging environments where a single listening device would be hopelessly outmatched. The data set collected as part of this work will help source separation researchers to develop scalable algorithms for massive-scale arrays.

The goal of this work was to show that by taking a radically different approach to listening technology, we can provide listening experiences that conventional devices never could. Microphone arrays can do more than just beamform; they can seamlessly alter a listener's auditory experience. There is much more work to be done before superhuman augmented listening technology is a part of daily life, but it is hoped that this work will inspire researchers and engineers to reimagine what listening devices can be. If we change the way that we approach listening technology, then we can change the way that humans experience the world.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica United with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT press, 1994.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT press, 1997.

[5] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Analysis and outcomes," *Computer Speech & Language*, vol. 46, pp. 605–626, 2017.

[7] R. F. Lyon, *Human and Machine Hearing*. Cambridge University Press, 2017.

[8] D. H. Johnson and D. E. Dudgeon, *Array Signal Processing: Concepts and Techniques*. Prentice Hall, 1993.

[9] H. L. Van Trees, *Optimum Array Processing*. Wiley, 2004.

[10] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[11] E. Vincent, T. Virtanen, and S. Gannot, *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[12] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[13] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer, 2008.

[14] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2013.

[15] K. Kinoshita, M. Delcroix, S. Gannot, E. A. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 7, 2016.

[16] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*, S. Haykin and K. R. Liu, Eds. Wiley, 2008, pp. 269–302.

[17] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[18] P. M. Peterson, "Adaptive array processing for multiple microphone hearing aids," Ph.D. dissertation, Massachusetts Institute of Technology, 1989.

[19] W. Soede, "Improvement of speech intelligibility in noise: Development and evaluation of a new directional hearing instrument based on array technology," Ph.D. dissertation, TU Delft, 1990.

[20] J. E. Greenberg, "Improved design of microphone-array hearing aids," Ph.D. dissertation, Massachusetts Institute of Technology, 1994.

[21] B. Widrow and F.-L. Luo, "Microphone arrays for hearing aids: An overview," *Speech Communication*, vol. 39, no. 1-2, pp. 139–146, 2003.

[22] T. Van den Bogaert, "Preserving binaural cues in noise reduction algorithms for hearing aids," Ph.D. dissertation, KU Leuven, June 2008.

[23] D. Marquardt, "Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques," Ph.D. dissertation, Carl von Ossietzky University of Oldenburg, 2016.

[24] A. Koutrouvelis, "Multi-microphone noise reduction for hearing assistive devices," Ph.D. dissertation, Delft University of Technology, 2018.

[25] S. K. Mamo, N. S. Reed, C. L. Nieman, E. S. Oh, and F. R. Lin, "Personal sound amplifiers for adults with hearing loss," *American Journal of Medicine*, vol. 129, no. 3, pp. 245–250, 2016.

[26] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja, "Assisted listening using a headset: Enhancing audio perception in real, augmented, and virtual environments," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 92–99, 2015.

[27] R. Ranjan and W.-S. Gan, "Natural listening over headphones in augmented reality using adaptive filtering techniques," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 11, pp. 1988–2002, 2015.

[28] J. M. Kates, *Digital Hearing Aids*. Plural Publishing, 2008.

[29] H. Dillon, *Hearing Aids*. Hodder Arnold, 2008.

[30] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 2915–2929, 2005.

[31] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Wiley, 2005.

[32] P. E. Souza, "Effects of compression on speech acoustics, intelligibility, and sound quality," *Trends in Amplification*, vol. 6, no. 4, pp. 131–165, 2002.

[33] J. B. Allen, "Amplitude compression in hearing aids," in *MIT Encyclopedia of Communication Disorders*, R. Kent, Ed. MIT Press, 2003, pp. 413–423.

[34] J. M. Kates, "Principles of digital dynamic-range compression," *Trends in Amplification*, vol. 9, no. 2, pp. 45–76, 2005.

[35] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[36] Y. Ephraim and D. Malah, "Speech enhancement using optimal non-linear spectral amplitude estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 8, pp. 1118–1121, Apr. 1983.

[37] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[38] H. Luts, K. Eneman, J. Wouters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, and H. Puder, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1491–1505, 2010.

[39] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 47–56, 2011.

[40] J. E. Greenberg and P. M. Zurek, *Microphone-Array Hearing Aids*. Berlin: Springer Berlin Heidelberg, 2001, pp. 229–253.

[41] P. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Springer Science & Business Media, 2010.

[42] D. Glista, S. Scollie, M. Bagatto, R. Seewald, V. Parsa, and A. Johnson, "Evaluation of nonlinear frequency compression: Clinical outcomes," *International Journal of Audiology*, vol. 48, no. 9, pp. 632–644, 2009.

[43] National Institute on Deafness and Other Communication Disorders, "NIDCD strategic plan 2017-2021," 2017. [Online]. Available: https://www.nidcd.nih.gov/about/strategic-plan/2017-2021

[44] M. A. Stone and B. C. Moore, "Side effects of fast-acting dynamic range compression that affect intelligibility in a competing speech task," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2311–2323, 2004.

[45] P. E. Souza, L. M. Jenstad, and K. T. Boike, "Measuring the acoustic effects of compression amplification on speech in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 41–44, 2006.

[46] M. A. Stone and B. C. Moore, "Quantifying the effects of fast-acting compression on the envelope of speech," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1654–1664, 2007.

[47] M. A. Stone and B. C. Moore, "Effects of spectro-temporal modulation changes produced by multi-channel compression on intelligibility in a competing-speech task," *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1063–1076, 2008.

[48] G. Naylor and R. B. Johannesson, "Long-term signal-to-noise ratio at the input and output of amplitude-compression systems," *Journal of the American Academy of Audiology*, vol. 20, no. 3, pp. 161–171, 2009.

[49] J. M. Alexander and K. Masterson, "Effects of WDRC release time and number of channels on output SNR and speech recognition," *Ear and Hearing*, vol. 36, no. 2, p. e35, 2015.

[50] P. Reinhart, P. Zahorik, and P. Souza, "The role of modulation characteristics on the interaction between hearing aid compression and signal-to-noise ratio," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3438–3438, 2016.

[51] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE AASP magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[52] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[53] S. Makino, Ed., *Audio Source Separation*. Springer, 2018.

[54] E. P. Zwyssig, "Speech processing using digital MEMS microphones," Ph.D. dissertation, The University of Edinburgh, 2013.

[55] P. D. O'grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.

[56] R. Gribonval and S. Lesage, "A survey of sparse component analysis for blind source separation," in *European Symposium on Artificial Neural Networks*, 2006, pp. 323–330.

[57] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[58] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[59] S. Wisdom, T. Powers, J. Pitton, and L. Atlas, "Deep recurrent NMF for speech separation by unfolding iterative thresholding," in *IEEE Worksohp on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[60] W. Soede, A. J. Berkhout, and F. A. Bilsen, "Development of a directional hearing instrument based on array technology," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 785–798, 1993.

[61] W. Soede, F. A. Bilsen, and A. J. Berkhout, "Assessment of a directional microphone array for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 94, no. 2, pp. 799–808, 1993.

[62] R. Stadler and W. Rabinowitz, "On the potential of fixed arrays for hearing aids," *The Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1332–1342, 1993.

[63] P. M. Peterson, N. I. Durlach, W. M. Rabinowitz, and P. M. Zurek, "Multimicrophone adaptive beamforming for interference reduction in hearing aids." *Journal of Rehabilitation Research and Development*, vol. 24, no. 4, pp. 103–110, 1987.

[64] J. E. Greenberg and P. M. Zurek, "Evaluation of an adaptive beamforming method for hearing aids," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1662–1676, 1992.

[65] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *The Journal of the Acoustical Society of America*, vol. 99, no. 5, pp. 3138–3148, 1996.

[66] F.-L. Luo, J. Yang, C. Pavlovic, and A. Nehorai, "Adaptive null-forming scheme in digital hearing aids," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1583–1590, 2002.

[67] J.-B. Maj, L. Royackers, J. Wouters, and M. Moonen, "Comparison of adaptive noise reduction algorithms in dual microphone hearing aids," *Speech Communication*, vol. 48, no. 8, pp. 957–970, 2006.

[68] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Speech distortion weighted multichannel Wiener filtering techniques for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Springer, 2005, pp. 199–228.

[69] V. Hamacher, U. Kornagel, T. Lotter, and H. Puder, "Binaural signal processing in hearing aids: Technologies and algorithms," in *Advances in digital speech transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. Wiley, 2008, ch. 14, pp. 401–429.

[70] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed MWF-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[71] T. J. Klasen, M. Moonen, T. Van den Bogaert, and J. Wouters, "Preservation of interaural time delay for binaural hearing aids through multi-channel wiener filtering based noise reduction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[72] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[73] A. I. Koutrouvelis, R. C. Hendriks, J. Jensen, and R. Heusdens, "Improved multi-microphone noise reduction preserving binaural cues," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 460–464.

[74] J. Flanagan, J. Johnston, R. Zahn, and G. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[75] H. F. Silverman, W. R. Patterson III, and J. L. Flanagan, "The huge microphone array," *IEEE Concurrency*, vol. 6, no. 4, pp. 36–46, 1998.

[76] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, "LOUD: A 1020-node modular microphone array and beamformer for intelligent computing spaces," Massachusetts Institute of Technology, Tech. Rep., 2004.

[77] I. Hafizovic, C.-I. C. Nilsen, M. Kjølerbakken, and V. Jahr, "Design and implementation of a MEMS microphone array system for real-time speech acquisition," *Applied Acoustics*, vol. 73, no. 2, pp. 132–143, 2012.

[78] J. Tiete, F. Domínguez, B. d. Silva, L. Segers, K. Steenhaut, and A. Touhafi, "SoundCompass: a distributed MEMS microphone array-based sensor for sound source localization," *Sensors*, vol. 14, no. 2, pp. 1918–1949, 2014.

[79] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, 2011, pp. 1–6.

[80] H. Chiba, N. Ono, S. Miyabe, Y. Takahashi, T. Yamada, and S. Makino, "Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 203–207.

[81] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Location feature integration for clustering-based speech separation in distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 354–367, 2014.

[82] M. Taseska and E. A. Habets, "Informed spatial filtering for sound extraction using distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 7, pp. 1195–1207, 2014.

[83] M. Taseska and E. A. Habets, "Spotforming: Spatial filtering with distributed arrays for position-selective sound acquisition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1291–1304, 2016.

[84] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[85] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[86] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[87] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Wiley, 2004.

[88] B. Cornelis, S. Doclo, T. Van dan Bogaert, M. Moonen, and J. Wouters, "Theoretical analysis of binaural multimicrophone noise reduction techniques," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 342–355, 2010.

[89] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[90] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[91] J.-M. Jot, B. Smith, and J. Thompson, "Dialog control and enhancement in object-based audio systems," in *Audio Engineering Society Convention*, 2015.

[92] B. Shirley, M. Meadows, F. Malak, J. Woodcock, and A. Tidball, "Personalized object-based audio for hearing impaired TV viewers," *Journal of the Audio Engineering Society*, vol. 65, pp. 293–303, 2017.

[93] R. M. Corey and A. C. Singer, "Dynamic range compression for noisy mixtures using source separation and beamforming," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[94] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation (LVA ICA)*, 2015, pp. 387–395.

[95] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001, pp. 99–102.

[96] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[97] H. Kayser, S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, p. 6, 2009.

[98] S. Renals, T. Hain, and H. Bourlard, "Interpretation of multiparty meetings: The AMI and AMIDA projects," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 2008, pp. 115–118.

[99] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - A corpus of multi-party conversational speech in noisy environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[100] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The Sheffield wargames corpus," in *Interspeech*. ISCA, 2013.

[101] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[102] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Interntional Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 313–317.

[103] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Web Download, Philadelphia, 1993.

[104] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[105] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention*. Audio Engineering Society, 2000.

[106] S. Müller and P. Massarani, "Transfer-function measurement with sweeps," *Journal of the Audio Engineering Society*, vol. 49, no. 6, pp. 443–471, 2001.

[107] D. D. Rife and J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *Journal of the Audio Engineering Society*, vol. 37, no. 6, pp. 419–444, 1989.

[108] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[109] J. Azcarreta Ortiz, "Pyramic array: An FPGA based platform for many-channel audio acquisition," M.S. thesis, Universitat Politècnica de Catalunya, 2016.

[110] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[111] R. M. Corey and A. C. Singer, "Nonstationary source separation for underdetermined speech mixtures," in *Asilomar Conference on Signals, Systems, and Computers*, 2016, pp. 934–938.

[112] R. M. Corey and A. C. Singer, "Underdetermined methods for multichannel audio enhancement with partial preservation of background sources," in *IEEE Worksohp on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[113] R. M. Corey, N. Tsuda, and A. C. Singer, "Acoustic impulse response measurements for wearable audio devices," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[114] R. M. Corey, N. Tsuda, and A. C. Singer, "Wearable microphone impulse responses," 2018. [Online]. Available: https://doi.org/10.13012/B2IDB-1932389_V1

[115] R. M. Corey, M. D. Skarha, and A. C. Singer, "Cooperative audio source separation and enhancement using distributed microphone arrays and wearable devices," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019.

[116] R. M. Corey, M. D. Skarha, and A. C. Singer, "Massive distributed microphone array dataset," 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-6216881_V1

[117] K. A. Riederer, "HRTF analysis: Objective and subjective evaluation of measured head-related transfer function," Ph.D. dissertation, Helsinki University of Technology, 2005.

[118] K. Kumatani, T. Arakawa, K. Yamamoto, J. McDonough, B. Raj, R. Singh, and I. Tashev, "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2012, pp. 1–10.

[119] R. M. Corey and A. C. Singer, "Spatial sigma-delta signal acquisition for wideband beamforming arrays," in *International ITG Workshop on Smart Antennas (WSA)*. VDE, 2016, pp. 1–7.

[120] R. M. Corey and A. C. Singer, "Wideband source localization using one-bit quantized arrays," in *IEEE International Worksho on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2017, pp. 1–5.

[121] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 3, pp. 235–238, 1977.

[122] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[123] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[124] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[125] E.-E. Jan and J. Flanagan, "Sound capture from spatial volumes: Matched-filter processing of microphone arrays having randomly-distributed sensors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1996, pp. 917–920.

[126] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[127] S. Markovich-Golan, S. Gannot, W. Kellermann, S. Markovich-Golan, S. Gannot, and W. Kellermann, "Combined LCMV-TRINICON beamforming for separating multiple speech sources in noisy and reverberant environments," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 2, pp. 320–332, 2017.

[128] O. Schwartz, S. Gannot, and E. A. Habets, "Multi-microphone speech dereverberation and noise reduction using relative early transfer functions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 240–251, 2014.

[129] Y. Avargel and I. Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[130] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1420–1434, 2009.

[131] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[132] R. M. Corey and A. C. Singer, "Motion-tolerant beamforming with deformable microphone arrays," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[133] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I.* Wiley, 2004.

[134] S. Markovich-Golan, S. Gannot, and I. Cohen, "A weighted multichannel Wiener filter for multiple sources scenarios," in *IEEE Convention of Electrical & Electronics Engineers in Israel*, 2012.

[135] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[136] T. J. Klasen, T. Van den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Transactions on Signal Processing*, vol. 55, no. 4, pp. 1579–1585, 2007.

[137] T. Van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 360–371, 2009.

[138] E. Hadad, S. Doclo, and S. Gannot, "The binaural LCMV beamformer and its performance analysis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 543–558, 2016.

[139] D. Marquardt, E. Hadad, S. Gannot, and S. Doclo, "Optimal binaural lcmv beamformers for combined noise reduction and binaural cue preservation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 288–292.

[140] E. Hadad, D. Marquardt, S. Doclo, and S. Gannot, "Theoretical analysis of binaural transfer function MVDR beamformers with interference cue preservation constraints," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 12, pp. 2449–2464, 2015.

[141] J. Thiemann, M. Müller, D. Marquardt, S. Doclo, and S. van de Par, "Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 12, 2016.

[142] R. M. Corey, N. Tsuda, and A. C. Singer, "Delay-performance tradeoffs in causal microphone array processing," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[143] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production," *Ear and Hearing*, vol. 23, no. 4, pp. 325–338, 2002.

[144] M. A. Stone and B. C. Moore, "Tolerable hearing-aid delays. IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," *Ear and Hearing*, vol. 26, no. 2, pp. 225–235, 2005.

[145] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. III. Effects on speech production and perception of across-frequency variation in delay," *Ear and Hearing*, vol. 24, no. 2, pp. 175–183, 2003.

[146] M. A. Stone, B. C. Moore, K. Meisenbacher, and R. P. Derleth, "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," *Ear and Hearing*, vol. 29, no. 4, pp. 601–617, 2008.

[147] M. McGrath and Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," *The Journal of the Acoustical Society of America*, vol. 77, no. 2, pp. 678–685, 1985.

[148] J. M. Kates and K. H. Arehart, "Multichannel dynamic-range compression using digital frequency warping," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 3003–3014, 2005.

[149] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 437807, 2009.

[150] T. Barker, T. Virtanen, and N. H. Pontoppidan, "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 241–245.

[151] G. Naithani, T. Barker, G. Parascandolo, L. Bramsl, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 71–75.

[152] G. Naithani, J. Nikunen, L. Bramslow, and T. Virtanen, "Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 386–390.

[153] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1053–1065, 2007.

[154] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, pp. 85–88.

[155] B. Schwartz, S. Gannot, and E. A. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 2, pp. 394–406, 2015.

[156] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, 1949.

[157] H. W. Bode and C. E. Shannon, "A simplified derivation of linear least square smoothing and prediction theory," *Proceedings of the IRE*, vol. 38, no. 4, pp. 417–425, 1950.

[158] N. Wiener and P. Masani, "The prediction theory of multivariate stochastic processes, II," *Acta Mathematica*, vol. 99, no. 1, pp. 93–137, 1958.

[159] E. Wong and J. Thomas, "On the multidimensional prediction and filtering problem and the factorization of spectral matrices," *Journal of the Franklin Institute*, vol. 272, no. 2, pp. 87–99, 1961.

[160] V. Kucera, "Factorization of rational spectral matrices: A survey of methods," in *International Conference on Control*, 1991, pp. 1074–1078.

[161] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design—A tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 399–408, 2012.

[162] J. M. Kates, "Understanding compression: Modeling the effects of dynamic-range compression in hearing aids," *International Journal of Audiology*, vol. 49, no. 6, pp. 395–409, 2010.

[163] American National Standards Institute, "Specification of hearing aid characteristics (ANSI S3.22-1996)," 1996.

[164] L. M. Jenstad and P. E. Souza, "Quantifying the effect of compression hearing aid release time on speech acoustics and intelligibility," *Journal of Speech, Language, and Hearing Research*, vol. 48, no. 3, pp. 651–667, 2005.

[165] D. V. Anderson, "A modulation view of audio processing for reducing audible artifacts," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5474–5477.

[166] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[167] T. May, B. Kowalewski, and T. Dau, "Signal-to-noise-ratio-aware dynamic range compression in hearing aids," *Trends in Hearing*, vol. 22, pp. 1–12, 2018.

[168] R. M. Corey and A. C. Singer, "A hypothesis testing approach for real-time multichannel speech separation using time-frequency masks," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[169] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[170] S. Rickard and Ö. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 529–532.

[171] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*. Springer, 2005, pp. 181–197.

[172] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[173] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[174] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.

[175] A. Liutkus, F.-R. Stoter, and N. Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation (LVA ICA)*, 2018.

[176] S. Rickard, "The DUET blind source separation algorithm," in *Blind Speech Separation*. Springer, 2007, pp. 217–241.

[177] T. Melia and S. Rickard, "Underdetermined blind source separation in echoic environments using desprit," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–19, 2006.

[178] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[179] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[180] M. Kühne, R. Togneri, and S. Nordholm, "A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation," *Signal Processing*, vol. 90, no. 2, pp. 653–669, 2010.

[181] J. Rosca, C. Borss, and R. Balan, "Generalized sparse signal mixing model and application to noisy blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2004, pp. iii–877.

[182] S. Winter, H. Sawada, S. Araki, and S. Makino, "Overcomplete BSS for convolutive mixtures based on hierarchical clustering," in *Independent Component Analysis and Blind Signal Separation (ICA)*. Springer Berlin/Heidelberg, 2004, pp. 652–660.

[183] M. Togami, T. Sumiyoshi, and A. Amano, "Sound source separation of overcomplete convolutive mixture using generalized sparseness," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[184] A. Aissa-El-Bey, N. Linh-Trung, K. Abed-Meraim, A. Belouchrani, and Y. Grenier, "Underdetermined blind separation of nondisjoint sources in the time-frequency domain," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 897–907, 2007.

[185] H.-G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1. IEEE, 1995, pp. 153–156.

[186] R. M. Corey and A. C. Singer, "Speech separation using partially asynchronous microphone arrays without resampling," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[187] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *Audio Engineering Society Convention*, 2017.

[188] R. M. Corey and A. C. Singer, "Real-world evaluation of multichannel audio enhancement using acoustic pilot signals," in *Asilomar Conference on Signals, Systems, and Computers*, 2017.

[189] R. M. Corey and A. C. Singer, "Relative transfer function estimation from speech keywords," in *International Conference on Latent Variable Analysis and Signal Separation (LVA ICA)*. Springer, 2018, pp. 238–247.

[190] S. Markovich-Golan and S. Gannot, "Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 544–548.

[191] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "Convolutive blind source separation methods," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 1065–1094.

[192] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l1-norm minimization," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 81–81, 2007.

[193] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[194] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[195] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2004.

[196] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[197] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[198] J. Bryan, "A sensorimotor basis of speech communication," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, May 2019.

[199] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4087–4091.

[200] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[201] P. Warden, "Speech commands: A public dataset for single-word speech recognition," 2017. [Online]. Available: http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz

[202] D. Y. Levin, E. A. Habets, and S. Gannot, "Near-field signal acquisition for smartglasses using two acoustic vector-sensors," *Speech Communication*, vol. 83, pp. 42–53, 2016.

[203] P. W. Gillett, "Head mounted microphone arrays," Ph.D. dissertation, Virginia Tech, 2009.

[204] P. Calamia, S. Davis, C. Smalt, and C. Weston, "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection," in *IEEE Worksohp on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[205] S. Miyabe, N. Ono, and S. Makino, "Blind compensation of interchannel sampling frequency mismatch for ad hoc microphone array based on maximum likelihood estimation," *Signal Processing*, vol. 107, pp. 185–196, 2015.

[206] R. Schreier and G. Temes, *Understanding Delta-Sigma Data Converters*. Wiley, 2005.

[207] M. V. Scanlon, "Helmet-mounted acoustic array for hostile fire detection and localization in an urban environment," in *Unattended Ground, Sea, and Air Sensor Technologies and Applications*, vol. 6963. International Society for Optics and Photonics, 2008, p. 69630D.

[208] G. Wersényi and A. Illényi, "Differences in dummy-head HRTFs caused by the acoustical environment near the head," *Electronic Journal of Technical Acoustics*, vol. 1, pp. 1–15, 2005.

[209] B. E. Treeby, J. Pan, and R. M. Paurobally, "The effect of hair on auditory localization cues," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3586–3597, 2007.

[210] G. Wersényi and J. Répás, "Comparison of HRTFs from a dummy-head equipped with hair, cap, and glasses in a virtual audio listening task over equalized headphones," in *Audio Engineering Society Convention*, 2017.

[211] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, 2001, pp. 3021–3024.

[212] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[213] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[214] J. Traa and P. Smaragdis, "Multichannel source separation and tracking with RANSAC and directional statistics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2233–2243, 2014.

[215] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 8, pp. 1408–1423, 2016.

[216] J. Nikunen, A. Diment, and T. Virtanen, "Separation of moving sound sources using multichannel NMF and acoustic tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 281–295, 2018.

[217] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[218] X. Zhong and J. R. Hopgood, "Time-frequency masking based multiple acoustic sources tracking applying Rao-Blackwellised Monte Carlo data association," in *IEEE Workshop on Statistical Signal Processing*, 2009, pp. 253–256.

[219] S. Markovich-Golan, S. Gannot, and I. Cohen, "Subspace tracking of multiple sources and its application to speakers extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 201–204.

[220] T. Higuchi, N. Takamune, T. Nakamura, and H. Kameoka, "Underdetermined blind separation and tracking of moving sources based ONDOA-HMM," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3191–3195.

[221] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2003.

[222] J. Málek, Z. Koldovskỳ, and P. Tichavskỳ, "Semi-blind source separation based on ICA and overlapped speech detection," in *International Conference on Latent Variable Analysis and Signal Separation (LVA ICA)*, 2012, pp. 462–469.

[223] H. Barfuss and W. Kellermann, "An adaptive microphone array topology for target signal extraction with humanoid robots," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014, pp. 16–20.

[224] Y. Bando, T. Mizumoto, K. Itoyama, K. Nakadai, and H. G. Okuno, "Posture estimation of hose-shaped robot using microphone array localization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 3446–3451.

[225] M. Er and A. Cantoni, "Derivative constraints for broad-band element space antenna array processors," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, no. 6, pp. 1378–1393, 1983.

[226] Y. R. Zheng, R. A. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 478–488, 2004.

[227] B. C. Levy, *Principles of Signal Detection and Parameter Estimation*. Springer, 2008.

[228] A. Bertrand, "Signal processing algorithms for wireless acoustic sensor networks," Ph.D. dissertation, KU Leuven, Belgium, May 2011.

[229] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks—Part I: Sequential node updating," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5277–5291, 2010.

[230] A. Bertrand and M. Moonen, "Distributed node-specific LCMV beamforming in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 233–246, 2011.

[231] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2010.

[232] Y. Hioka and W. B. Kleijn, "Distributed blind source separation with an application to audio signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 233–236.

[233] F. Nesta and M. Omologo, "Cooperative Wiener-ICA for source localization and separation by distributed microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1–4.

[234] M. Souden, K. Kinoshita, M. Delcroix, and T. Nakatani, "Distributed microphone array processing for speech source separation with classifier fusion," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012.

[235] L. Wang and S. Doclo, "Correlation maximization-based sampling rate offset estimation for distributed microphone arrays," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 571–582, 2016.

[236] D. Cherkassky and S. Gannot, "Blind synchronization in wireless acoustic sensor networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 651–661, 2017.

[237] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 785–789.

[238] J. Sanz-Robinson, L. Huang, T. Moy, W. Rieutort-Louis, Y. Hu, S. Wagner, J. C. Sturm, and N. Verma, "Large-area microphone array for audio source separation based on a hybrid architecture exploiting thin-film electronics and CMOS," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 979–991, 2015.

[239] P. Kairouz, "The fundamental limits of statistical data privacy," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2016.