1 Introduction

… things inanimate have mov'd,
And, as with living Souls, have been inform'd,
By Magick Numbers and persuasive Sound.
—William Congreve (1697) *The Mourning Bride*

The ear is a most complex and beautiful organ. It is the most perfect
acoustic, or hearing instrument, with which we are acquainted,
and the ingenuity and skill of man would be in vain exercised to imitate it.
—John Frost (1838), *The Class Book of Nature: Comprising Lessons on the Universe, the Three Kingdoms of Nature, and the Form and Structure of the Human Body*

Would it truly be in vain to exercise our ingenuity to imitate the ear? It would have been, in the 1800s—but now we are beginning to do so, using the "magick" of numbers. Machines imitating the ear already perform useful services for us: answering our queries, telling us what music is playing, locating gunshots, and more. By imitating ears more faithfully, we will be able to make machines hear even better. The goal of this book is to teach readers how to do so.

Understanding how humans hear is the primary strategy in designing machines that hear. Like the study of vision, the study of human hearing is ancient, and has enjoyed impressive advances in the last few centuries. The idea of *machines* that can see and hear also dates back more than a century, though the computational power to build such machines has become available only in recent decades. It is now, as they say in the computer business, a simple matter of programming. Well, not quite—there is still work to be done to firm up our understanding of sound analysis in the ear, and yet more to be done to understand the enormous capabilities of the human brain, and to abstract these understandings to better support machine hearing. So let's get started.

Humans tend to take hearing for granted. We are so aware of what's going on around us, largely by extracting information from sound, yet so unable to describe or appreciate how we do it. Can we make machines do as well at interpreting their world, and ours, through sound? We can, if we leverage scientific knowledge of how humans process sound.

Being able to produce and analyze sound waves is a prerequisite to developing a better understanding of hearing. Early progress in the field was made with the help of analytical instruments such as Helmholtz's resonators and recording devices, like the waveform drawing device in Figure 1.1, and controlled sound production instruments such as Seebeck's siren, shown in Figure 1.2. Representing such waves as electrical

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

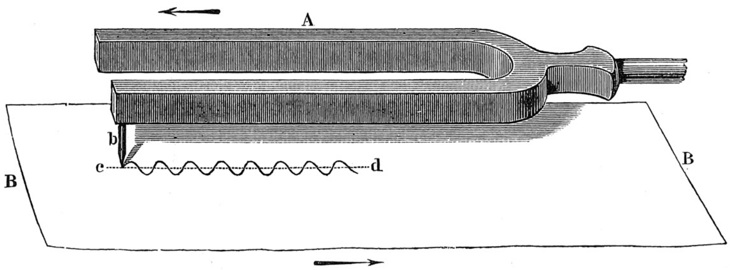

Figure 1.1 Helmholtz explained the idea of a sound's waveform via this diagram of a tuning fork with a stylus point attached, drawing its vibration on a moving piece of paper.

signals has been routine since the invention of the telephone. We now have a myriad of machines that help us generate, compress, communicate, store, reproduce, and modify sound signals, in ways tuned to how we hear. For most of these applications, though, the machines remain "deaf," in that they get very little meaning out of the sounds they process.

What if you had a device at home, always listening to what's going on? Could it tell what interesting things it heard while you were out? Could it tell you the refrigerator sounds like it's wearing out? Would it understand if you asked it a question? Could it find you some music to listen to if you described your mood? Could it listen to you and determine your mood itself? Could it say where a mouse might be hiding because it heard it run there? Could it distinguish between normal household sounds and an anomaly in the dead of night? Could it also be your intelligent answering machine, and tell you who called, and why, based on hearing their voice? Of course it could.

Who might make such a machine? What crazy functionality might they give a machine that could hear and understand sounds? Have we chosen the best path through the complex web of theories about hearing? Can we do better on some tasks by modifying the approach? What advances in the study of human hearing might we discover while trying to put our theories to the test of real use? These are the kinds of ideas and questions about sound and hearing that have been going around in my head for decades—and that we are getting some answers on recently. I've worked on spatial effects in music and games, and on machines to synthesize and recognize speech and music, and on other fun things to do with sound. Where most others deal with sounds by various conventional or ad hoc methods, I keep coming back to how the ear would do it—and this approach has proved fruitful.

There is enough known about how the ear and hearing work that we have gotten serious about putting this knowledge to practical uses. Starting with the anatomy, we model the structure and function of the ear and the auditory nervous system; using physiological and psychophysical techniques, we figure out what the brain gets from the ear, and how it deals with the information to perform meaningful tasks. Then we program computing machines to do similarly, based on this knowledge. In essence, we mimic the biology.

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

Figure 1.2 A make-it-yourself acoustic siren, much like August Seebeck's, as shown by Alfred M. Mayer (1878). The spinning disk, driven from a crank via string and pulleys, interrupts a stream of air from the tube to make waves of sound pressure that we hear as a tone. Different tones can be made by moving the tube to a different row of holes, or by changing the disk to one with a different pattern of holes. August Seebeck and Hermann von Helmholtz were among the nineteenth-century scientists who used such devices in their research that contributed to connecting the physical and perceptual properties of musical tones to the mechanisms of human hearing—though their theories were somewhat in opposition to each other.

Today we have access to massive quantities of sound, to analyze, organize, index, and learn from. The soundtracks of YouTube videos alone have hundreds of millions of hours of sound, and so far our computers are rather ignorant of what those soundtracks are trying to communicate. Imagine what value there might be in having our machines just listen to them and understand. Speech, music, laughing babies, sounds of interesting events, activities, places, and personalities—it's all there to be discovered, categorized, indexed, summarized, remembered, and retrieved.

The full scope of machine hearing will reveal itself as people discover that it is relatively easy to have machines understand sounds of all sorts, and people find

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

imaginative uses for such machines. Elephant infrasound hearing and bat ultrasound hearing and echolocation suggest that the same basic strategies have been put to many purposes by other mammals. We might include other sonic applications—such as medical imaging—that use sound waves but don't rely on anything about sound perception. At Schlumberger Research in the 1980s, we experimented with hearing techniques applied to the analysis of underground sonic waves. Any far-out infrasound through ultrasound applications that can benefit from the use of techniques like those evolved by humans fall within the scope of what we're trying to teach via this book.

As we get more people engaged in machine hearing, there will be more good ideas and more things we can take on. The potential is enormous, and the scope broad.

1.1 On Vision and Hearing *à la* David Marr

The pioneering vision scientist David Marr was a big influence on my approach to modeling hearing. When I visited him at MIT in 1979 to show him what I was working on, he was very encouraging of the approach. Twisting his words, from vision to hearing, illustrates how his thinking influenced mine:

> What does it mean to hear? The plain man's answer (and Aristotle's, too) would be, to know what is where by listening. In other words, hearing is the *process* of discovering from sounds what is present in the world and where it is.
>
> Hearing is therefore, first and foremost, an information processing task, but we cannot think of it just as a process. For if we are capable of knowing what is where in the world, our brains must somehow be capable of representing this information—in all its profusion of color and form, beauty, motion, and detail.—modified from *Vision*, David Marr (1982)

I honor Marr's introduction to his ground-breaking book *Vision* in the quotation above, having changed *see* to *hear*, *looking* to *listening*, *vision* to *hearing*, and *images* to *sounds*. I've left the last phrase unchanged, as I believe that *"color and form, beauty, motion, and detail"* is a much more apt description of what our brains extract and represent about sound than the usual more pedestrian properties of *loudness, pitch,* and *timbre*.

Marr's computational and representational approach to vision helped to define the vibrant field of computer vision, or machine vision as it's also called, more than thirty years ago. My book is motivated by the feeling that something along these lines is still needed in the hearing field. It's a daunting challenge to try to live up to David Marr, even if I've had a few extra decades to prepare, but it's time to give it a shot.

Compared to other mammals, humans have put vision to some very special applications, like reading written language, and analogously have put hearing to use in spoken language and in music. These pinnacle applications should not exclusively drive the study of vision and hearing, however, and perhaps are best addressed only after low-level preliminaries are well understood, and more general applications are under control. Therefore, we focus on these more general and lower-level aspects, and on broader applications of hearing, as Marr focused on the more general aspects of vision. At the end, we come back and touch on applications in speech and music.

David Mellinger (1991) should be credited with helping drive this approach via his dissertation, pointing out that "Advances in machine vision have long stemmed from a physiological approach where researchers have been heavily influenced by Marr's computational theory. Perhaps the same transfer will begin to happen more in machine hearing." But this transfer has been incomplete, so we need to drive it some more.

Martin Cooke (1993) has provided an excellent review of Marr's approach to vision and its influence on work in speech and hearing. Marr's identification of three levels at which the sensory system is to be understood—*function, process,* and *mechanism*, also described as *computation*, *algorithm*, and *implementation*—certainly does help us organize our study of hearing. In an interesting twist, Peter Dallos (1973) used a similar division of concerns into function, mode of operation, and anatomy to describe the auditory periphery, before Marr's work. His scheme is still used this way and credited in current hearing books (Yost, 2007), as shown in Figure 1.3.

Cooke reviews several applications of Marr's levels and principles to speech processing, but provides relatively little connection to hearing. The repurposing of Marr's *primal sketch* concept into a *speech sketch*, by Green and Wood (1986), points up a disconnect: Marr didn't go from primitive images directly to reading, and we shouldn't go from primitive sound representations straight to speech; *primal* should imply a much lower level. A sketch is a "sparsified" version of an image, which may be used as part of a feature extraction strategy at the input to a learning system, as described in Section 25.7.

In vision, objects and images must be analyzed at many different scales. Referring to Marr, Andy Witkin (1983) said, "The problem of scale has emerged consistently as a fundamental source of difficulty, because the events we perceive and find meaningful vary enormously in size and extent. The problem is not so much to eliminate fine-scale noise, as to separate events at different scales arising from distinct physical processes." In hearing, we have the same issue, especially in the temporal dimension, where sounds have periodicities and structure on all time scales.

The idea of an "auditory primal sketch" has been introduced by Neil Todd (1994) as a way to represent the rhythm and temporal structure of music and speech. I had published a related idea on multiscale temporal analysis, as part of a speech recognition approach (Lyon, 1987). Both of these are based on Witkin's scale-space filtering, which was descended from Marr. Both fall far short of a comprehensive framework for machine hearing, but help to inspire some of the sorts of representations that we will be working with.

Albert Bregman (1990), in his book *Auditory Scene Analysis: The Perceptual Organization of Sound*, discusses how aspects of hearing are valued from an evolutionary perspective, yielding certain advantages of hearing over vision. The auditory system evolved in a context in which better understanding of meaning from an auditory scene—better answers to *what* and *where*—led to a better chance of survival. When I refer to *human hearing* in my title, I mean to include the cortical-level processing

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

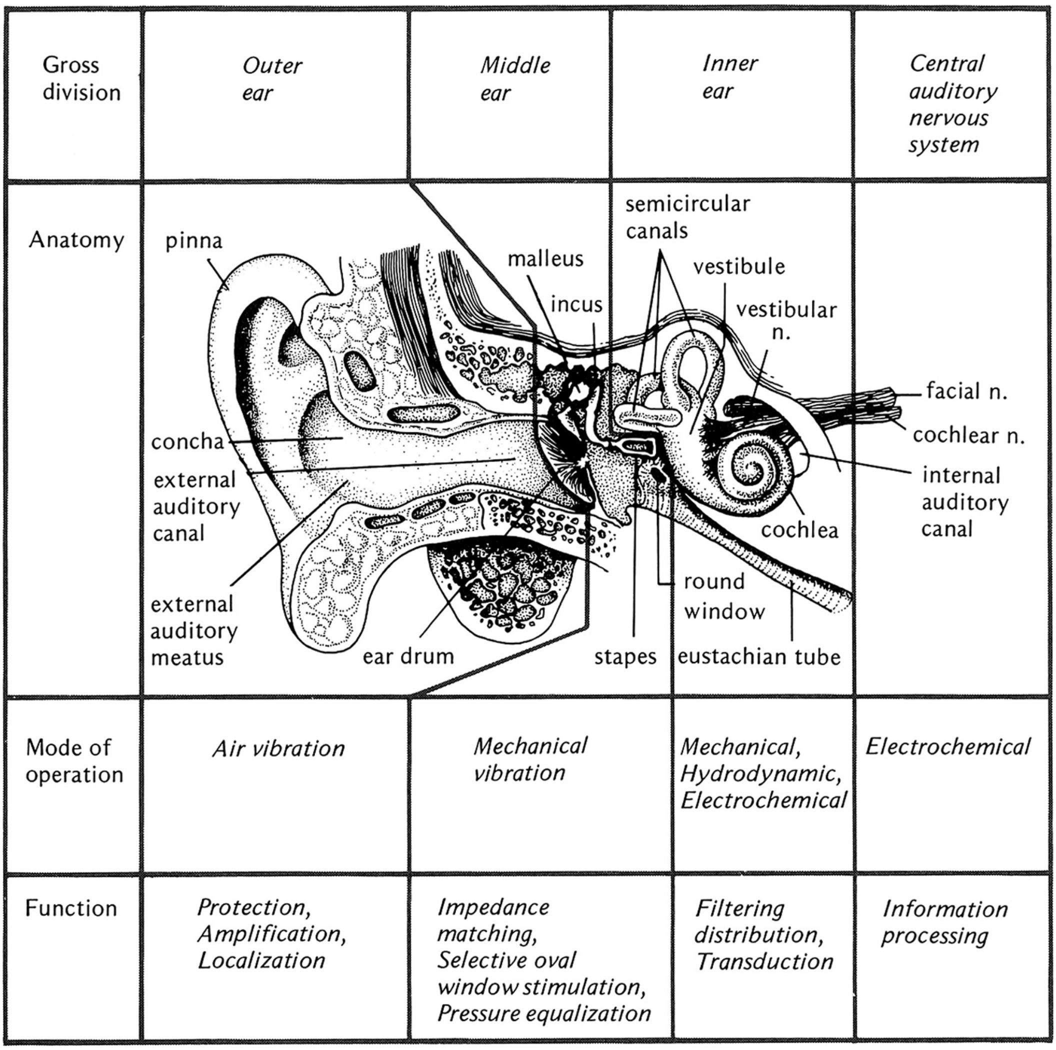

Figure 1.3 Ear diagram by Yost (2007). While the anatomy and modes of operation are important, we are most interested in emulating the *function*, described in the bottom row. The *information processing* in the central nervous system—the bit where meaning is extracted—is the part that remains most open to exploration and speculation. [Figure 6.1 (Yost, 2007) reproduced with permission of Wiliam A. Yost.]

systems that have evolved to handle speech, music, and other big-brain functions; but I do not mean to diminish the importance of the lower levels of auditory processing—in the ear, the brainstem, and the midbrain—that underlie the exquisite hearing capabilities of our pets (and pests), and that form the basis for robust representations of sound from which actionable information can be extracted. Even animals that don't normally use speech can learn to reliably recognize their own names, and discriminate them against other speech sounds; for example, Shepherd (1911) taught four raccoons that their names were Jack, Jim, Tom, and Dolly.

We can question Marr's insistence that a symbolic representation or *description* be generated (Hacker, 1991). Some approaches to machine hearing systems successfully use representations that remain completely abstract and nameless until the final output—the information that the system is trained to extract—with intermediate steps being subsumed in the learning system. Other approaches will use explicit and named concepts, such as objects, events, musical instruments, notes, talkers, and so forth, that artificial intelligence systems can reason with. Different theories of mind, or different computational frameworks that we have available, will bias our machine hearing applications one way or the other. We are not yet in a position to say which way is likely to be more fruitful for any given area, and hope to encourage exploration in all such directions.

Comments on hearing's analogy with vision are not new. For example, in 1797, the effect of auditory masking on sensitivity was observed and compared to visual masking effects in "annotations" on Perrole's "Philosophical Memoir" on sound transmission (Perrole, 1797):

> Sounds seem more intense, and are heard to a greater distance, by night than by day. . . . It is a practical question of some importance to ascertain whether this difference may arise from the different state of the air, the greater acuteness of the organ, or the absence of the ordinary noises produced in the day. By attentive listening to the vibrations of a clock in the night, and remarking the difference between the time when no other noise was heard, and when a coach passed along, it has appeared clear to me that this difference arises from the greater or less stillness only, and that no voluntary effort or attention can render the near sound much more audible, while another noise acts upon the organ. In this situation the ear is nearly in the state of the eye, which cannot perceive the stars in the day time, nor an object behind a candle.

In that memoir, Perrole also introduced the term *timbre* from the French to explain what he meant by *tone* in English: "The tone (*timbre*) was changed in the water in a striking manner." This "catch-all" term, as it has been called, captures everything about what a sound "sounds like," except for its pitch and loudness—sort of like *texture* in vision, which captures much of what shape, size, and brightness don't. It is the job of our machine hearing systems to map timbre (along with pitch and loudness and direction, and their evolution and rhythm over time) into useful information about what the sound represents, be it speech, music, environmental noises, or evidence of mundane or exceptional events.

1.2 Top-Down versus Bottom-Up Analysis

Top-down processing evaluates sensory evidence in support of hypothesized interpretations (meaning), while bottom-up processing converts sensory input to ever-higher-level representations that drive interpretation. Real systems are not necessarily at either extreme, but the distinction can be useful.

Marr says, with respect to general-to-specific (or coarse-to-fine) stereo matching approaches (Marr, 1982),

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

Nomenclature: What to Call This Endeavor

The terms *computer vision* and *machine vision* are in wide use, not quite interchangeably, the former having a more computer-science connotation, and the latter a more industrial or applications connotation. Terms like *computer hearing*, *computational hearing*, and *computer listening* seem awkward to me, especially since I spent a lot of years building analog electronic models of hearing, probably not qualifying as computers. And what about *listening* or *audition* as a better analogy to *vision*? Several of these terms have overloaded meanings: we can convene a hearing, or perform in an audition, or plant listening devices. The term *machine listening* is sometimes used, but mostly in connection with music listening and performance.

The term *machine hearing* has a strong history at Stanford's computer music lab, CCRMA. In their 1992 progress report, Bernard Mont-Reynaud (1992) wrote a section on machine hearing, which noted that "The purpose of this research is to design a model of Machine Hearing and implement it in a collection of computer programs that capture essential aspects of human hearing including source formation and selective attention to one source (the 'cocktail party problem') without tying the model closely to speech, music, or other domain of sound interpretation."

We hope that by calling the space of computer applications of sound analysis *machine hearing*, following Mont-Reynaud, we will leverage this good name and good direction, and help the field build around a good framework, as Marr did with what we refer to as *machine vision*.

> This type of approach is typical of the so-called top-down school of thought, which was prevalent in machine vision in the 1960s and early 1970s, and our present approach was developed largely in reaction to it. Our general view is that although some top-down information is sometimes used and necessary, it is of only secondary importance in early visual processing.

Here we totally agree. Although I have nothing but respect for the strong case for the power of top-down information and expectations in human hearing (Slaney, 1998; Huron, 2006), and though there are prominent "descending" pathways at all levels of the auditory nervous system (Schofield, 2010), my understanding is that the more extensive and complex feedback is within the cortical levels of the central nervous system, and that early audition, like early vision, is best conceived as a modular set of mostly feed-forward bottom-up processing modules. There is feedback, to be sure, but its function can often be treated as secondary, as Marr says. At some levels, feedback may be about parameter learning and optimization; from cortex to thalamus, top-down projections may be about attention. These are important, but not where we start, especially in "early" layers as Marr says.

In the mammalian brain, these early hearing modules include the periphery (the ear) as well as auditory structures in the brainstem and midbrain, and maybe even some stages of cortical processing, such as primary auditory cortex. These levels were successful stable subsystems long before the evolution of the big neocortex that led to

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

speech and music. The "near decomposability" condition (Simon, 1981) is what allows complex systems to evolve. That's why we rely so much on data from bottom-up experiments in animals to help us understand human hearing; we accept that the amazing abilities of humans evolved on top of these stable mammalian subsystems, which are themselves not so different from reptilian, bird, and even fish auditory systems.

Like Marr, we are partly reacting to an overreliance on top-down information in sound processing systems. For example, automatic speech recognition (ASR) systems have been gradually improved over the years by reliance on larger and more complex language models and by statistical models that can capture complex prior distributions, while their front-end processing remains relatively stagnant, stuck with spectro-temporal approaches that have no way to improve in terms of robustness to noise and interference, since they don't represent the aspects of sound that help our auditory systems tease sound mixtures apart. Such problems demand that we understand hearing better, and build systems that can hear and understand multiple sounds at once; how else can we expect a speech recognizer to give us a transcript of a boisterous meeting? Of course, good prior distributions from top-down information will continue to play an important role, too.

Is the auditory system *complex*? Herb Simon (1981) characterizes a complex system this way:

> In such systems, the whole is more than the sum of the parts, not in an ultimate, metaphysical sense, but in the important pragmatic sense that, given the properties of the parts and the laws of their interaction, it is not a trivial matter to infer the properties of the whole.

I think this applies to the auditory system as a whole, when the cortex is included, especially in a living organism in which the auditory system is interacting with visual, motor, and other systems, with strong top-down and feedback effects. But for the various bottom-up modules of lower-level auditory processing, perhaps the system is merely *complicated*, but not so complex that we can't describe its function, and its process, in terms of its mechanisms. I think this is how Marr saw early vision, too. Otherwise, it would be hard to be optimistic about our ability to assemble machines to do similar jobs.

1.3 The Neuromimetic Approach

A strategic element of our machine hearing approach is to respect the representation of sounds on the auditory nerve, which involves both a *tonotopic* (arranged by frequency) organization and detailed temporal structure, as extracted by the rather nonlinear inner ear. At this level, the approach can be said to be *neuromimetic* (Jutten et al., 1988), or *neuromorphic* (Mead, 1990), in the sense that we may be building a copy of a complicated neural system, mimicking its function—or mimicking its structure when we can't quite describe the function. In the neuromorphic case, copying the structure of the neural system, the expectation is that the structure will have an appropriate *emergent*

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

behavior and therefore a useful information-processing function. Here *emergent* means that the behavior is not explicitly designed in, but *emerges* from the simpler behaviors of the lower-level elements as a consequence of the structural pattern of interconnection of those elements (Bar-Yam, 1997).

This neuromimetic approach is somewhat distinct from the Marr approach, but sometimes a useful supplement. When a system built this way is found to have a useful function due to its emergent behavior, it can sometimes be further analyzed, and the important parts of its function abstracted, described, and reengineered more efficiently. I believe we are part of the way through this process with neuromimetic hearing front ends. At the level of the cochlea, for example, the function is largely understood, but the description is still as much structural as functional. We do not have the clean separation of function, process, and mechanism that Marr recommended, but we do have a structure for which we can understand the function.

Beyond the cochlea, we still have a mixed structural and functional view, though it is somewhat speculative, of what the function is—the little "information processing" box in the lower right corner of Bill Yost's diagram, Figure 1.3, is where we ultimately extract meaning. We have pretty good ideas from physiological data about what kinds of auditory images are formed in the brainstem. The main thing we use that is neuromorphic is the very idea of an auditory image: a neural pathway with two spatial dimensions, like the optic nerve from the retina, projecting a time-varying pattern to a two-dimensional sheet of cortical tissue, the primary auditory cortex, for further processing.

An early proponent of a neuromimetic, or *bionic*, approach to machine hearing systems was John L. Stewart (1963), who published a number of reports, papers, patents, and a book on the topic in the 1960s and 1970s. He explains the reasoning behind this approach (Stewart, 1979):

> The model becomes an intermediary—a surrogate reality. . . . It is my belief that effective explanations for the traits of living organisms demand the construction of models which behave as do their living counterparts. For, in no other way can the research be disciplined to produce an effective holistic theory!

Stewart (1979) anticipated much of our current approach, including a cochlear transmission-line analog with nonlinearities, a "neural-like analyzer" stage following the cochlea (Stewart, 1966), and the idea of efferent (feedback) adaptation to conditions, via coupled frequency-dependent gain control (Stewart, 1967).

1.4 Auditory Images

In our approach to hearing, we incorporate the notion of an *auditory image*: a presumed representation developed in the subcortical parts of the auditory nervous system (cochlea, brainstem, and midbrain), projecting to primary auditory cortex in the same way that the retinal image projects to primary visual cortex. This approach brings together the strategies of Marr with the two-dimensional neural circuits of the *place*

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

theory of sound localization of Jeffress (1948) and the *duplex theory of pitch perception* of Licklider (1951).

A *spectrogram* is a picture of sound on a time–frequency plane. But this two-dimensional image is not what we call an auditory image, as it has too few dimensions to be analogous to the image that the eye sends to cortex. In the spectrogram, one axis is time, and there is only one other axis (frequency, mapped to spatial location). To make auditory images, we develop one more dimension, to map to a spatial axis orthogonal to the frequency axis, resulting in a movie-like representation, an image that changes with time. This added spatial dimension can represent direction (lateral or azimuth direction of arrival of a sound) in a binaural auditory image like those of Jeffress, or can represent pitch period and other temporal texture as in Licklider's duplex images. But these are just examples, not the limit of what an auditory image might be.

A possible next (cortical processing) step is to reduce the auditory image to a *sketch*, or line drawing, as Marr does, but that is not the only approach.

Our study of hearing will necessarily involve a lot of function, process, and mechanism to arrive at auditory images, corresponding mostly to levels below primary auditory cortex. This complicated architecture is a bit different from the vision case, where the information starts as an optical image that makes a 2-D response image on the retina, and further processing is mostly in cortex. Even in secondary and later levels of auditory and visual cortex, much of the mammalian brain's processing is about what and where, and only humans, with huge areas of more highly evolved cortex, implement the much higher levels of interpretation that support language and music (Rijntjes et al., 2012).

Marr was very much in touch with the developing sciences of visual psychology and visual neurophysiology, which informed his approach, especially at the level of multiscale edge analysis in visual cortex, on which he modeled his primal sketch. Similarly, our approach to machine hearing draws on the fields of auditory psychology and physiology, where so much is known about many levels of hearing, and where I've been so lucky to know and interact with so many of the great scientists over the last several decades. Part of our goal with this book is to help these fields in return, by providing a conceptual framework in which much of their detailed knowledge can find a place, and be better understood and promulgated in terms of signal processing, information extraction, and sound understanding.

The physiological data informing this approach are from animal studies, in mammals, birds, reptiles, and other groups. Most of the auditory brainstem and midbrain was already stable before the mammals split off from the reptiles, so studies in many animals contribute to our understanding of human hearing, and are included in our scope. For example, the notion of auditory images as a representation of objects in space, as extracted from binaural (two-ear) signals, has been well developed to describe the function and organization of the auditory nervous system in the barn owl (Konishi, 1995). We humans may not swoop down and catch mice in the dark, but we do have an auditory spatial sense that's not so different from that of the barn owl, using very similar structures in our brains.

1.5 The Ear as a Frequency Analyzer?

At the functional level of description, it can be hard to say what the ear is doing. A traditional view is that the *cochlea* in the inner ear acts as a *Fourier analyzer* or *frequency analyzer* (Gold and Pumphrey, 1948; Plomp, 1964). We believe that as a top-level functional description, that's often misleading. One goal of this book is to help displace this view with a better description of the kind of information the ear sends to the brain.

In the late nineteenth century, it was not unusual to find statements such as "the function of the cochlea is to determine the pitch of the sound" (Draper, 1883), or "the function of the cochlea is to receive and appreciate musical sounds" (Murché, 1884). Generally, the cochlea was interpreted as a frequency analyzer. A few interpretations were a bit broader, with statements like "the function of the cochlea is to appreciate the *qualities* of sounds" (Bale, 1879).

The simple frequency view was largely derived from Helmholtz (1863), though his book on the subject was much more thoughtful than these simplifications. He did address function head-on, but his book was about connecting hearing to music, so he can't be faulted for describing the function in relation to musical tones:

> Hence the ear does not distinguish the different forms of wave in themselves, as the eye distinguishes the different vibrational curves. The ear must be said rather to decompose every wave form into simpler elements according to a definite law. It then receives a sensation from each of these simpler elements as from an harmonious tone. By trained attention the ear is able to become conscious of each of these simpler tones separately. And what the ear distinguishes as different qualities of tone are only different combinations of these simpler sensations.

This phase-blind frequency-analysis view of hearing had originally been articulated by Georg Ohm (1843), inspired by Joseph Fourier's 1822 finding that periodic functions could be described as sums of sinusoids. While the idea does have some merit as a model of hearing, it is also easily found to disagree with various experiments, so has often been regarded as a half-truth, or sometimes worse, as in this statement by W. Dixon Ward (1970):

> For years musicians have been told that the ear is able to separate any complex signal into a series of sinusoidal signals—that it acts as a Fourier analyzer. This quarter-truth, known as Ohm's Other Law, has served to increase the distrust with which perceptive musicians regard scientists, since it is readily apparent to them that the ear acts in this way only under very restricted conditions.

Ohm's and Helmholtz's view of hearing as Fourier analysis, and the confusion of frequency with pitch, continued to permeate, if not dominate, thinking about hearing in the early twenty-first century, even though problems with the approach had been repeatedly demonstrated, and arguments against it published continually over a century and a half.

August Seebeck (1841), using his acoustic siren, demonstrated several effects that were hard to explain in Ohm's model. In fact, Ohm published his law in response to Seebeck's first paper in 1841, and they engaged in a back-and-forth in print for a number of

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

years. Helmholtz later sided with Ohm, and tried to explain Seebeck's results in his book (Helmholtz, 1863) in a way that would resuscitate Ohm's point of view. These disputes have been frequently recounted (Scripture, 1902; Jungnickel and McCormmach, 1986; Cahan, 1993; Beyer, 1999), so we don't need to go into detail here. Heller (2013) has a particularly cogent discussion of the evolution of the thinking of Seebeck, Ohm, and Helmholtz, as influenced by Fourier's mathematics (and it is a great undergraduate-level book on sound and hearing in general).

Many modern papers and books sidestep the description at a functional level, with sections entitled "the function of the cochlea" typically describing lots of phenomena, process, and mechanism, but with very little commitment to an idea of function. Statements of function are sometimes made, but are kept very general and conservative, such as "The primary function of the cochlea is hearing" (Van De Water and Staecker, 2006), and "The function of the cochlea is to convert the vibration of sound into nerve impulses in the auditory nerve" (Cook, 2001), and "the essential function of the cochlea can be conceptualized as a transduction process" (Phillips, 2001). Some invoke the traditional Fourier analyzer concept, as in "Its principal role is to perform a real-time spectral decomposition of the acoustic signal in producing a spatial frequency map" (Dallos, 1992).

In a very few cases, we find a bit about capturing the quality of sound and something about temporal properties, as in "The main function of the cochlea is to translate auditory events into a pattern of neural impulses that precisely reflects the nature and timing of the sound stimulus" (Probst et al., 2006). This concept is better, especially in being tied to general properties of the sound instead of to narrower musical properties based on frequency. We need this kind of more general functional thinking if we're going to process arbitrary real-world sounds—the kinds of sounds for which hearing evolved, long before music and speech came along.

An important function of the cochlea that is often missed in functional characterizations has recently been given first-class status: loudness compression. Jont Allen (2001) says:

> The two main roles of the cochlea are to separate the input acoustic signal into overlapping frequency bands, and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell.

Allen's conceptualization of function is a much better starting place, and explains part of why nonlinearities are so important in hearing. A proper focus on function will be key to our progress in machine hearing. In support of the function "to separate the input acoustic signal into overlapping frequency bands," we discuss the progression from Fourier analysis, to short-time Fourier analysis, to linear bandpass filterbanks; and in support of the function "to compress the large acoustic intensity range," compressive nonlinear filterbanks. We further connect filterbanks to filter-cascade structures, to make a more realistic relationship of the filtering function to the underlying mechanisms. Part II of the book develops the necessary systems theory, and Part III applies these concepts to develop good computational models of cochlear function.


14 *TRATTATO DI MUSICA.*

te intuonati. Si ſentirà un terzo ſuono affatto di-
ſtinguibile, e ſarà il ſottopoſto ſegnato in note
chiuſe muſi- cali.

Figure 1.4 Tartini’s 1754 publication of his observation of *un terzo suono*, a third sound, shown as filled notes below the first two sounds playing on violins or horns—among the earliest recognitions of a nonlinear effect in hearing. The note pitches that Tartini illustrated represent the ratios 4:5:2, 5:6:2, 3:4:2, 5:8:2, and 3:5:2 ($f_1 : f_2 : f_3$, for f_1 being the pitch of the lower played sound and f_2 being the pitch of the upper one, and f_3 being the pitch of the low third tone). The third-tone pitch corresponds to the quadratic intermodulation product $f_2 - f_1$, or the cubic intermodulation product $2f_1 - f_2$, and/or an octave above one of those. As Helmholtz (1863) remarked of these observations, “It is very easy to make a mistake of an octave. This has happened to the most celebrated musicians and acousticians. Thus it is well known that Tartini, who was celebrated as a violinist and theoretical musician, estimated all combinational tones an octave too high.” Sorge’s 1745 observation of c″ and a″ making an f would be 3:5:1, with *den dritten Klang*, a third-order (cubic) distortion product, at $2f_1 - f_2$.

1.6 The Third Sound

The importance of nonlinearity is not yet well integrated into the typical understanding of the functions and processes of hearing. One of the earliest phenomena to bring the problem to the attention of scientists was the *third sound*, observed by Sorge (1745) (*den dritten Klang*) and by Tartini (1754) (*un terzo suono*). This third sound is a low-pitch tone heard when two other tones are sustained, for example by two horn players; pitches of such tones are illustrated in Figure 1.4. It turns out to be usually a pitch equal to the difference of the pitches of the first two tones or of some of their harmonics, and is what we call a *combination tone*, a *difference tone*, or a *distortion product*.

We’ll see that there are good reasons for the existence of several types of nonlinearities in hearing, and for modeling them in machine hearing systems. But before we tackle nonlinearity, we have to understand what linear systems are, and how such systems give rise to sinusoidal analysis. We’ll cover the theory of linear and nonlinear systems in Part II, and apply them in subsequent parts of the book.

1.7 Sound Understanding and Extraction of Meaning

We conceptualize the machine hearing space as *sound understanding*, or *information extraction*, or *extraction of meaning*, in a very general sense. Here *understanding* signifies extraction of actionable information, as is sometimes implied in *speech understanding* systems as distinguished from *speech recognition* systems. That is, it means that from a sound we are able to provide useful information for some practical application.

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

It's not just humans and machines that do this—my dog is pretty good at processing sounds, too. If her practical application is to greet someone at the front door, she gets the information she needs from the sound of either a knock or the doorbell. For the application of when to eat, she recognizes the sound of her dish being set down. She's pretty clever about learning the sound cues for when she'll be taken for a walk, and other things she cares about. Does she *understand* sounds? Yes—in the same sense that humans do, and that machine hearing systems do: from sounds, she extracts what she needs to know.

If we can make machines hear half as well as my dog does, that will be progress. Humans are involved because we want to build up to where we can replicate a human's ability to extract information from speech, music, video soundtracks, and the everyday environment that humans live in. And humans provide a wealth of psychophysical experimental data that can be leveraged in the design of machine hearing systems.

Winnie-the-Pooh has introspected on the extraction of meaning from sound (Milne, 1926):

> "That buzzing-noise means something. You don't get a buzzing-noise like that, just buzzing and buzzing, without it meaning something. If there's a buzzing-noise, somebody's making a buzzing-noise, and the only reason for making a buzzing-noise that *I* know of is because you're a bee. . . . And the only reason for being a bee that I know of is making honey. . . . And the only reason for making honey is so as *I* can eat it."

How did Pooh interpret a "buzzing-noise" as indicating the availability of honey? We interpret this question as having two main parts: first, analyzing and representing sound in such a way that this "buzzing-sound" is distinguishable from other sounds; and second, learning one or more decision functions that address the question of when and where food might be available, based on the sound present. The connection from "buzzing-noise" to food is probably the result of a fairly opaque learned decision function, in a brain or a hearing machine; Pooh's semilogical chain of reasoning should probably be regarded as a *post-hoc* rationalization of the decision, not an explanation of how the decision was arrived at. It seems likely that at this level of abstraction, humans and other mammals probably perform such functions about the same way as Milne's anthropomorphized fictional characters do.

When decisions are reached, and those decisions are useful, then we can say that meaning, or information, has been extracted from the sound. Sometimes the meaning is more indirect, as by inference from the linguistic content carried by words in the sound of speech. In speech recognition, we can say that meaning has been extracted when the recovered word sequence serves to further the successful execution of a task.

1.8 Leveraging Techniques from Machine Vision and Machine Learning

At the applications end of machine hearing, there are many overlaps of problems, and techniques, with other fields. Therefore, we have many opportunities to leverage techniques that have been developed in those fields. In particular, machine vision and

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

machine learning, especially as applied to problems in situations involving both images and sound, whether live or recorded, give us a good set of tools to apply. Leveraging these much larger fields is a key part of our strategy in trying to bring the field of machine hearing forward.

The machine vision field gives us a number of successful feature extraction approaches, and trainable system structures, some of which will map well into hearing problems. In systems such as video analysis, or surveillance, where both vision and hearing can be applied together, we have opportunities to *fuse* information from the different senses, on the way to the extraction of meaning. Even the simple concatenation of sound features onto image features has already been shown to improve the performance of video classification systems (Gargi and Yagnik, 2008); they may still be half blind and "hard of hearing," but they're no longer completely deaf.

1.9 Machine Hearing Systems "by the Book"

After we survey a range of conventional and novel sound analysis and representation techniques in Part I, we review in Part II the linear system theory that explains why the idea of analyzing sounds into frequencies, or overlapping frequency bands, makes sense, and how important nonlinear concepts such as compression need to be integrated into that view.

In Part III, we go on to apply that concept at other levels of description, culminating in a model of the cochlea that runs as an efficient machine algorithm for processing sounds into a representation that respects what we know about signals on the auditory nerve.

Part IV of the book attempts to do the same for the next levels of processing, in the lower parts of the auditory nervous system: to provide a functional concept, and an efficient process and mechanism that will extract the "auditory image" sound representations needed by the higher levels of hearing, to connect to the information that applications need to understand sound.

In Part V, we get into applications, which we can think of as paralleling the uses to which humans apply the information they extract from sound. We may not yet know enough about the function of neocortex to really leverage that knowledge for building intelligent machines, so at the application level we turn mostly to techniques we understand better, from the field of machine learning. We use various methods to convert the sound representations from the subsystems in Parts I and III into the kinds of features that machine learning systems can easily use, and from there we train transformations that extract the information we want. None of this has much to do with frequency analysis, so we should be careful to not let that concept dominate our thinking about the ear.

Our book develops the idea of a machine hearing system made of four modules, or layers; from the bottom up, as illustrated in Figure 1.5 and detailed here:

1. A model of the cochlea, or auditory periphery, built as a cascade of nonlinear filters, as developed in Part III;

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

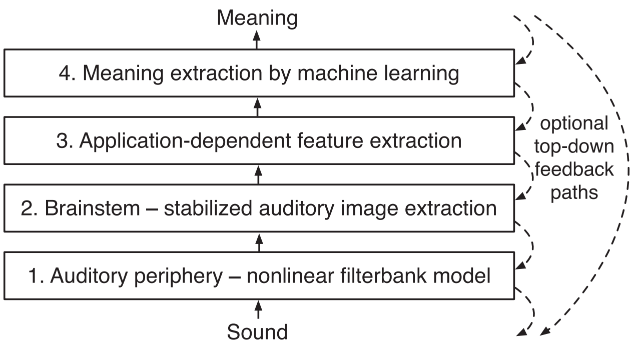

Figure 1.5 The *four-layer model* of machine hearing systems developed in this book—from sound to meaning, and sometimes back the other way. The big feedback loop from meaning to sound is for a system that can make sound and hear itself, for example, a speech conversation system.

2. A model of the auditory brainstem, extracting one or more auditory images appropriate to the range of sounds and tasks to be addressed as developed in Part IV;
3. A feature extraction layer to convert auditory images to a form more suited to the particular application and tailored to the machine learning system chosen, as developed in Part V;
4. A machine learning system that is trained to extract the kind of decisions or *meaning* needed for the target application, as addressed also in Part V.

This layering will focus us on a known-working and factored structure, based closely on human hearing where possible, not specific to the higher-level properties of speech and music, that is open-ended enough to allow expansion into arbitrary applications. From the point of view of many applications, such as speech recognition, most of the action is at the top, in level 4, and the lower three levels just make a black-box front end. The challenge there will be to make sure that the features that come out of level 3 are what the recognizer needs.

Our machine hearing systems are characterized by several special features, in the first two modular layers: the cascade filterbank structure with nonlinearities, and the auditory image approach. Hence, much of our emphasis is on developing an understanding of these hearing-based ideas and their historical precedents, in the corresponding book parts.

These special features are not new or radical, but are not yet widely enough appreciated and used in hearing systems. Both were discussed in the middle of the twentieth century. The notion of a cascade as an alternative to the more common parallel-resonator filterbank was presented by Licklider (1956) as a model of cochlear filtering. He also adopted what we now call auditory images in his "duplex theory of pitch perception" and combined this approach with Jeffress's "place theory of sound localization," to form his "triplex theory" of pitch perception (Licklider, 1956):

> . . . It outlines a mechanism that accounts for the three ways in which acoustic stimulation can give rise to subjective pitch and, at the same time, brings into mutual relation a number of facts from other parts of auditory experience. . . . If the aim is to understand the process of perception,

https://doi.org/10.1017/9781139051699.004 Published online by Cambridge University Press

the inquiry must extend into the higher centres of the brain. At the present time, this is sure to lead one into speculation. However, if there is a lack of anatomical and physiological facts, there is an abundance of psychophysical ones.

It took a few more decades for the understanding of the auditory nervous system, and of the cochlear nonlinearity, to evolve. Auditory image maps in the auditory nervous system are now known and being actively investigated (Knudsen, 1982; Sullivan and Konishi, 1986; Schreiner, 1991; Langner et al., 1997; Velenovsky et al., 2003), as discussed in Part IV. Incorporating appropriate nonlinearities into the cascade filterbank, with results reflected in the auditory image, is straightforward, once the different types of nonlinearity are understood, as emphasized in Part III.

Pierce and David (1958) commented on the fact that different types of meaning extraction involve different types of processing:

> Undoubtedly the nervous system uses a multiplicity of methods in dealing with the range of auditory stimuli presented to it. We don't "perceive" vowels in the same way as gunshots. A machine to emulate the nervous system in these functions would be an intelligent machine indeed. Can we ever understand enough to make such a machine? Before science can answer unequivocally it must look farther, directly or indirectly, into both the problems involved in such recognition and the way in which human beings manage to solve them.

These differences, which we now know more about from studies of psychoacoustics and of the nervous system, would be reflected in the application-dependent feature extraction layer, where we extract different features to localize a gunshot than to classify a vowel, and in the trainable decision system in the final layer.

On the prospects of building machines to do it, Pierce and David (1958) knew it would be a long hard road:

> We have already taken the first few faltering steps toward building machines which will respond to and correctly interpret the sounds of speech. Through further diligent work it may indeed become possible to construct devices which will respond to and reply in human speech, perhaps even to make useful voice typewriters, and maybe, later on, to build that staple of science fiction, a device which translates spoken words of one language into spoken words of another. Whether such machines are ever actually built will depend upon how complex they need be; and, in essence, how much human time, effort, and money man is willing to expend in simulating human functions.

But this was more than a half century ago; the value of building such systems is now generally known to exceed the costs in many application areas. It would be good to reflect on the various dimensions of progress since then, as well as on the remaining difficulties, as we set out to build more such machines.

In proceeding to build such machines, we will learn much about hearing. I hope the first lesson has sunk in already: sounds are not just sums of tones of different frequencies, and the ear is not a frequency analyzer.