Chapter 1

Background

Several fundamental aspects of hearing come in pairs: place coding and temporal coding of pitch, localization using interaural time difference cues or interaural level difference cues, signal identification using temporal-envelope cues or temporal-fine-structure cues, and discrete signal sampling in time or continuous temporal integration. When one mechanism is unavailable, the other seems to cover for it, within certain ranges of signal parameters that are particular for these cues. Present auditory models largely consider a single signal path that functionally combines a predetermined weighting of one or two types of signal processing. This is the case even though the auditory signal splits into three parallel pathways (two in non-mammalian vertebrates) in the auditory brainstem that converge only at the inferior colliculus. Can it be that our sense of hearing is set to do everything twice, so that two signal processing outputs are combined to obtain a superior output to either of the two on their own?

The theory put forth in this treatise portrays hearing as both a communication system and an imaging system. Both imaging and communication theories provide a built-in duality that distinguishes between coherent and incoherent (or noncoherent) imaging or detection. While most engineered systems tend to stick to either coherence or incoherence, we shall argue that hearing makes the most of both, which also matches the acoustic environment best, as it tends to be partially coherent. Therefore, this work will employ the notion of degree of coherence in a manner that is germane to hearing with emphasis on analogies to vision, since they are numerous and they can provide novel intuition for the development of alternative concepts in hearing. Namely, the theory highlights some parallels with vision that are simpler to grasp if the spatial and temporal envelope dimensions are swapped. However, unlike vision, whose image can be visually seen on the retina before being demodulated by the photoreceptors[1], the auditory image resides deep inside the brain and cannot be listened to—at least not without sophisticated instrumentation and signal processing. Moreover, auditory processing includes various phase and across-channel interactions that give rise to auditory phenomena, such as harmony, that do not have analogs in vision.

This introductory chapter surveys some of the overarching historical themes in hearing theory. Then, it examines more closely the various comparisons that have been made between hearing and vision, in attempt to garner insight about hearing that is more readily found in vision. This background is used to motivate the main themes that have driven this work—temporal imaging being the foremost one.

A historical review of hearing theory is found in Boring (1942, pp. 399–436) including some interesting notes about discoveries and ideas about the physiology of hearing, which anticipated Helmholtz's work and provided the context for its development. An even more comprehensive

[1]Christoph Scheiner was the first to directly observe the retinal image in 1619 (Le Grand and El Hage, 1980, p. 57).

review of historical theories was given by Wever (1949, pp. 3–94), where he delineated the so-called "place theories" and "frequency theories", which are now referred to as temporal theories. A vivid account of several hearing theories has been presented in Lyon (2018, Chapter 2 and throughout). A recent report on how the understanding of tonotopy has developed is found in Ruben (2020), where the origin of the frequency analysis and tonotopy of the cochlea is traced back to Guichard Joseph Du Verney in 1683, who mistakenly switched the frequency mapping between the cochlear base and apex. A more detailed (but still brief) overview of the anatomy and physiology of the ear is provided in §2.

1.1 The scope and state of hearing theory

Although most scholars did not attempt to formally define the scope of hearing theory, it seems to have meant different things for those who did. For example, Fletcher (1930) delineated several aspects that such a theory should be able to account for: the auditory bandwidth and dynamic range, the just-noticeable differences in pitch and intensity, distortion products, masking phenomena, loudness, binaural effects, frequency selectivity of complex tones, effects of pitch and loudness and their relations to the physical signal. Two decades later, Licklider (1951a, p. 1034) suggested a more general scope: "*The principal tasks of auditory theory are (1) to explain the psychophysics of hearing in terms of aural mechanics and neurophysiology, (2) to give audition its proper setting in a general theory of communication, and (3) to provide a calculus of response to auditory stimulation.*" Although the role of the brain in hearing was recognized from anatomical data and was being crudely incorporated in early theories (e.g, Fletcher, 1930; Wever and Bray, 1930b), hearing theory was regularly considered about equivalent to a mechanical explanation of how the cochlea transduces the acoustical waves to neural spiking patterns as late as 1975 (Békésy, 1956; Schroeder, 1975). More recently, Cariani and Micheyl (2012) proposed a more cautious scope: "*A full theory of audition thus should explain the relations between sounds, neuronal responses, auditory functions, and auditory experience.*" This definition seems to be general enough to be inclusive of just about anything auditory.

Ideally, a complete theory of hearing should be able to explain what hearing does and how it is facilitated by the ears, for a given listener, acoustic environment, stimulus, and situation. Furthermore, it should enable derivation of particular auditory effects, motivate observed behaviors and responses, and help resolve ostensible contradictions in the empirical science. In other words, the ideal hearing theory should be able to reduce the complexity of observations that had seemed disparate before the theory was introduced. At present, no such theory exists[2]. Instead, various

[2]Acknowledgment that hearing theory does not exist has been seldom made. See for example, Nordmark (1970, p. 57) and Bialek and Schweitzer (1985). A passage that may be interpreted as saying essentially that was provided by Reinier Plomp—one of the most prominent psychoacousticians of the second half of the 20th century. In his book, "The Intelligent Ear", he candidly admitted (Plomp, 2002, p. 9): "*Many investigators have assumed that full knowledge of how sinusoidal tones are heard will be sufficient to explain the perception of everyday sounds. A long time ago, a well-known Dutch composer visited me, presuming that current laboratory knowledge of tone perception could help him to a better understanding of how musical sounds are perceived. I had to disappoint him.*" Plomp then went on to suggest that the reason for this lack of knowledge is the discrepancy between the perfect laboratory-based stimuli and sounds encountered in the real world. In the context of comparing visual and auditory processing models, Schönwiesner and Zatorre (2009) stated: "*In auditory neuroscience there is no consensus yet about the suitable set of low-level features.*" In the narrow context of masking phenomena, Durlach (2006) pointed at a "*conceptual chaos*" and that: "*Not only is there no overarching conceptual structure available to organize the area and provide it with scientific elegance, but there are few definitions that evidence even a modest degree of scientific stability (varying across both individuals and time).*" As a final example, in the context of a theory of the auditory thalamus and cortex, Winer (2011b, p. 679) stated: "*There is no global theory of auditory forebrain function since the facts available cannot support such an edifice.*"

"part-theories" (an expression coined by Licklider, 1959) are available, which attempt to account for local phenomena within hearing, such as the cochlear function, pitch perception, sound localization, auditory scene analysis, auditory attention, speech perception, music perception, etc.

All the part-theories notwithstanding, it is not a given that once they reach maturity, they will coalesce into a grand unified theory of hearing. Hearing is a complex biological system, and biology is presently best understood through evolution theory (Dobzhansky, 1973), as Lettvin et al. (1959) noted with regards to vision: "*...since the purpose of a frog's vision is to get him food and allow him to evade predators no matter how bright or dim it is about him, it is not enough to know the reaction of his visual system to points of light.*" Therefore, any account of hearing phenomena beyond evolution that results in a more compact and less complex theory than the present state of knowledge may be seen as a boon.

1.2 Elements of hearing theory

Hearing theory has made a slow progress over about two and a half millennia from the external to the internal hearing organs—from the outer ear to the brain (Boring, 1942, pp. 399–400). With the monumental work of Hermann von Helmholtz, it became abundantly clear that the most critical sensory transformation of sound occurs in the cochlea, where it is transduced to neural information (Helmholtz, 1948, first published in 1863). Helmholtz hypothesized that the capability of the ear to analyze complex tones is due to the radial fibers that make up the basilar membrane of the cochlea, which locally resonate in sympathetic vibration with incoming tones, just like tuning forks. Almost a century later, the collected works of Georg von Békésy (1960) were published, where what may have been the most influential model of cochlear mechanics since Helmholtz's was laid out. Békésy showed that the unique spectral analytical property of the cochlea corresponds to a **traveling wave** that propagates along the basilar membrane. The traveling wave peaks at a place along the cochlea that is mapped to a specific frequency, as a result of the elastic properties and geometry of the basilar membrane.

By the mid-20th century, the maturation of electronic engineering and signal processing precipitated the transformation of empirical science as a whole, including acoustics and neurophysiology. Using a mixture of synthesized pure tones and white noise maskers, it was found that the frequency range of the ear is internally covered by a bank of bandpass filters (also referred to as channels) with overlapping flanks. The response of these filters provides a robust explanation for simultaneous masking phenomena (Fletcher, 1940), loudness summation (Zwicker et al., 1957) and other important effects.

The significance of the sound information transfer to neural signals was hypothesized by Rutherford (1886) in his **telephone theory** where he noted: "*that simple and complex vibrations of nerve energy arrive in the sensory cells of the brain, and there produce, not sound again of course, but the sensations of sound...*" Direct physiological recordings of auditory nerve fibers of cats showed that the spiking patterns are spectrally tuned (Galambos and Davis, 1943; Kiang et al., 1965), in a way that corresponds to the incoming sound, if amplified and played back (Wever and Bray, 1930a). The spiking pattern was hypothesized to follow the **volley principle**. This principle predicts that the auditory nerve fiber bundle, which innervates a given hair cell in the cochlea, tracks the incoming stimulus together, so that all fibers spike in tandem at a certain phase of the stimulus (Wever and Bray, 1930b). This principle can account for how low-intensity stimuli can be heard even though each individual fiber experiences refractory periods and fires stochastically (see also Wever, 1949, pp. 166–441). Indeed, once the neural and mechanical recording techniques became sufficiently precise, it became possible to establish a close correspondence between the cochlear mechanical response of the traveling wave and the auditory nerve fibers (Sellick et al., 1982; Khanna and Leonard, 1982).

Things became more complicated as it had gradually become evident that the cochlea is unmistakably nonlinear, since the dynamic range of the mechanical traveling wave response is compressed relatively to the acoustic input (Rhode, 1971). Dispelling any doubt that the cochlea must contain an active mechanism needed the discovery of **otoacoustic emissions** from the ear by Kemp (1978), which had soon after received the name **the cochlear amplifier** by Davis (1983). This idea resurrected a much earlier model of active hearing that was proposed by Gold (1948) but was prematurely rejected at the time. The discovery of electromotility of the outer hair cells by Brownell et al. (1985), along with converging physiological evidence using various methods and models, has led to the conclusion that the outer hair cells are the main cause of nonlinearity in the cochlea. However, as the complexity of the organ of Corti is very high, and since handling it requires delicate methods and tools if it is to remain intact during experimentation, much work has been carried out through modeling and indirect measurements of cochlear mechanics. Thus, the exact mechanism of amplification (as well as other cochlear effects and the very function of various cochlear features) are still being debated (e.g., Ashmore et al., 2010). Importantly, the same organ appears to be the main cause for a host of other nonlinear phenomena, including the generation of audible distortion products in the cochlea (Avan et al., 2013).

The above observations, which follow the auditory signal in the auditory nerve channels, can partially account for most phenomena on Fletcher's list (§1.1) that underscore the significance of the cochlea in hearing sensation. Binaural effects such as localization, however, are strictly dependent on neural structures in the auditory brain to work, as was hypothesized by Lord Rayleigh (1907b) and has largely been confirmed since (Grothe et al., 2010). But binaural effects are not the only phenomena that extend beyond the auditory nerve. Another important case is the detection of periodicity across the cochlear bandpass channels, which gives rise to the effects of the missing fundamental and periodicity pitch (Schouten, 1940). The highly influential **duplex pitch model** by Licklider (1951b) maintained that the auditory system carries out an autocorrelation operation, centrally, using delay lines and coincidence detectors on top of the cochlear bandpass filtering. The latter can give rise to the standard tonal pitch, whereas the former can give rise to periodicity pitch. Other notable phenomena that require central processing are the processing of amplitude and frequency modulation, broadband sound processing, various adaptations that improve detection of specific sounds in noise and reverberation, forward masking, and also disorders such as tinnitus. Therefore, it is clear that peripheral theories of sound only cover a restricted range of auditory phenomena.

Probably the most influential higher-level theory of hearing in the last years is that of **auditory scene analysis**, which was synthesized by Albert S. Bregman (1990). The main idea behind it is that the auditory system utilizes different acoustic cues in a given stimulus, which enables the mental organization of different sound elements into auditory streams—the auditory counterpart to visual objects—that are mentally localized in space. Different cues may be available to the listener that enable auditory streaming, such as grouping based on common onset time of sounds, common fundamental frequency (harmonic content), cues based on spectral range, location in space, and others. Importantly, this theory provides a framework for an intermediate stepping stone in the auditory information processing of meaningful sounds such as speech, vocalizations, or music. Scene analysis is also related to other cognitive factors such as memory, attention, and decision-making, and it generally opens a much broader perspective for hearing theory. However, while auditory scene analysis provides a powerful framework for understanding sound processing and demystify several auditory illusions, it was originally framed without a clear physiological substrate that can realize the necessary signal processing. Thus, scene analysis does not readily connect with the peripheral auditory output mentioned above, which was traditionally considered the essence of hearing theory. Auditory scene analysis is thought to take place at the level of the cortex (Christison-Lagay et al.,

2015), although there are indications that basic processing of relevant cues begins in the brainstem (Masterton, 1992; Pressnitzer et al., 2008; Felix II et al., 2018).

Harnessing auditory scene analysis to its core, Cariani and Micheyl (2012) outlined the scope for an auditory theory that should take basic auditory (and cognitive) attributes as multidimensional variables of auditory information. The basic auditory attributes are loudness, duration, location, pitch, and timbre, which are processed in the auditory cortex, go through scene analysis, and culminate in conscious perception. Such a theory should draw on the three levels for understanding information processing in the visual cortex that were laid out by David Marr and could be readily generalized to other sensory modalities: 1. "*Computational theory—What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out.*" 2. "*Representation and algorithm—How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?*" 3. "*Hardware implementation—How can the representation and algorithm be realized physically?*" (Marr, 2010, p. 10). In the theoretical framework outlined by Cariani and Micheyl (2012), there is emphasis on elucidating the different neural codes that represent the auditory stimulus, which can then be processed in a broad range of perceptual operations using different neurocomputational architectures that lead toward decision making, which sometimes leads also to action.

The above high-level theoretical frameworks leave a substantial gap in the understanding of the subcortical processing between the auditory nerve and the cortex. It is known that important processing takes place, probably gradually, in the brainstem and midbrain, which reflects the animal's specific needs and ecology (e.g., Casseday and Covey, 1996; Felix II et al., 2018; see §2.4). Such processing is likely directed by the auditory cortex via the descending efferent pathways (Cariani and Micheyl, 2012). But relevant theories tend to be too general, probably due to the complexity of the involved circuitry (a complexity that relates both to their apparent algorithms and hardware, re Marr), as well as the large number and variety of processes (apparent computational goals) implicated by them. This was implied by Winer (2011a, p. 67), who stressed that the auditory system is an artificial construct made up of streams, although in reality it is integrated with many other sensory modalities and processes. Moreover, he underscored that even with the auditory scene analysis theoretical framework, when it is applied to the physiological circuitry, a significant number of auditory nuclei in the midbrain and forebrain are excluded from the processing chain, and their functions within the complete auditory process are poorly understood.

At present, it remains unknown whether the different auditory brain functions can be explained using a compact theory from which lower-level effects may be unambiguously derived, or whether each auditory phenomenon has to be accounted for using its own part-theory. In this sense, the combined part-theories of hearing—covering the entire cascade of periphery, brainstem, and perception—are incomplete.

1.3 Hearing theoretical development and vision

One approach that has been repeatedly invoked to produce insights about hearing is comparison to vision, which is by far the most studied sense in humans. Vision is also touted as the most dominant of all our senses and humans are often said to be "visual creatures". The analogy between the two senses is natural, given that hearing and vision are intuitively juxtaposed—both organs are placed on the same level of the face, both come in pairs, both are based on wave physics, both sensory experiences are ubiquitous and elemental in the human experience, and both can give rise to rich aesthetics and culture.

The cross-inspiration between hearing and vision can be traced back at least to ancient Greece, when various analogies between hearing and vision were conceived (Darrigol, 2010a,b). At that time,

light was thought to emanate from the eyes (originally due to Empedocles) and there was generally no distinction between sensation and perception (Hamlyn, 1961). A more fertile correspondence between the nature of light and sound began only around 1000 A.D., when it was proposed by Ibn al-Haitham—widely considered as the father of modern optics—that light is an entity that is independent of the beholder. Many of the most prominent scholars that studied acoustics and hearing from the Renaissance until the end of the 19th century contributed significantly to both sound and light theories. Notable figures such as Isaac Newton, Thomas Young, Hermann von Helmholtz, and Lord Rayleigh critically advanced both fields. Interestingly, Helmholtz, who is best known in hearing for his influential place-theory of pitch perception (Helmholtz, 1948), is also credited with laying the foundations to the optics and the physiology of the eye (Helmholtz, 1909)—still relevant today as well (e.g., Le Grand and El Hage, 1980; Charman, 2008; see also Wade, 2021 for a historical review).

Perhaps the most epigrammatic of all the hearing-vision analogies is best captured by the aphorism "*architecture is frozen music*[3]." It relates to the fact that visual perception unfolds in space, whereas hearing unfolds in time (e.g., Hirsh, 1952; Massaro, 1972; Jones, 1976; Welch and Warren, 1980; Kubovy, 1988; Näätänen and Winkler, 1999). Hirsh (1952) maintained that the perceived visual dimensions are determined by objects, which are not meaningful auditory entities, since hearing is concerned with events. Using these dimensions for comparison, visual acuity—"*a measure of the interval of space between two visual stimuli that are perceived as two*"—becomes the fundamental descriptor of visual capacity, whereas temporal acuity is the analogous one for hearing. This point was elaborated in Julesz and Hirsh (1972), where it was stressed that visual events do exist—making the comparison asymmetrical between the two senses. They further emphasized the transient nature of the auditory environment compared to the stable one of vision, but they implied that the two senses can be ultimately seen as complementary, as they both detect both space and time of the animal's environment. Handel (1988) pushed back against the adage that visual space and auditory time are analogous, as neither sensory object can exist in space or in time only. Nevertheless, the observation that vision is largely spatial and hearing is largely temporal has been robust. For reasons that will become clearer later, we shall borrow from optics and refer to it as the **space-time duality**.

Other comparisons between hearing and vision were usually made by emphasizing specific dimensions. For example, a popular analogy that appeared early is between pitch and color. Marin Mersenne wrote in 1632 that the lowest notes are akin to black, the highest notes to white, and the colors are anything in between (Darrigol, 2010b). An influential analogy between color and pitch was drawn circa 1665 by Isaac Newton, who tried to impose the natural musical intervals of the seven diatonic notes on the seven primary colors that constitute white light (Pesic, 2006) (Figure 1.1). A similar analogy was independently proposed by Robert Hooke in 1672. Leonhard Euler was the first to relate colors to the frequency of light waves—drawing from sound wave theory instead of the other way round (Pedersen, 2008). Helmholtz too continued this line of thinking, having recognized that both pitch and color relate to the wave frequency, but he eventually abandoned the analogy with musical intervals, seeing that the spectra of light and sound cannot be made to match (Pesic, 2013).

Helmholtz made additional comparisons between the ear and the eye throughout his book from which two points stand out. First, the eye is much slower than the ear in its ability to resolve changes (Helmholtz, 1948, p. 173). Second, the ear has an analytic ability to decompose complex tones to their harmonics, which the eye does not possess. He additionally compared vibrations in the spatial-frequency domain (he used water waves as an object) to the audio frequency domain (Helmholtz, 1948, pp. 29 and 128). As will be seen later in this work, this is confusing the carrier

[3] Attributed to both Johann Wolfgang von Goethe and Friedrich Wilhelm Joseph von Schelling.

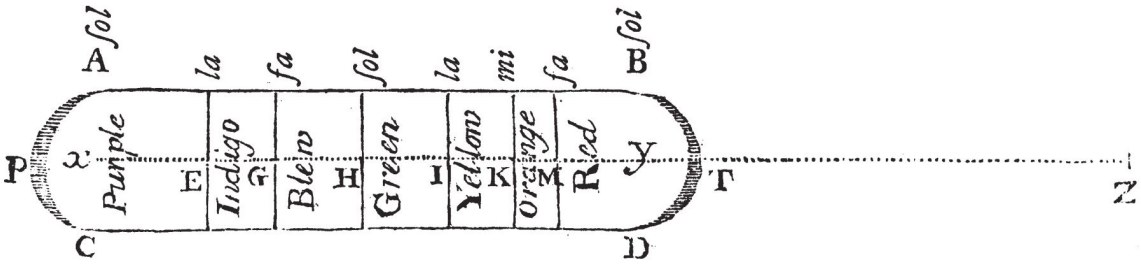

Figure 1.1: Isaac Newton's color-pitch map from his second paper on color and light, read at the Royal Society in 1675.

and modulation domains of the two systems—not an uncommon mistake in hearing and vision that is still occasionally encountered in literature.

Several commonalities between the two senses were contrasted by Harris (1948), who compared certain aspects in the psychophysical and neural coding in both hearing and vision. He found some that were directly comparable and differ in value (e.g., sensitivity, internal noise, neural adaptation, integration time, dynamic range, lateral inhibition), but others that were not directly comparable (e.g., frequency required for tonal modulation to sound continuous, binaural/binocular summation, quantum-effect threshold). While most comparisons are dated, they are valuable in showing the general properties that both the ear and the eye have as physical detectors, whose outputs manifest neurally. This is also the underlying thinking behind the work of Jacobson (1950, 1951a), who was the first to attempt to quantify the informational channel capacity (in bits per second) of the eye and the ear. More narrowly, Stevens (1957) focused his comparison on psychophysical characteristics of the two senses and on the similarity of the power laws that describe them. He contrasted loudness in hearing and brightness in vision as the psychophysical counterparts of intensity, and are comparable with respect to their growth-rate dependence on frequency (wavelength), the effect of masking (glare), the degree of adaptation to baseline signal level, and their pathologies—loudness recruitment in hearing and a rare genetic disorder of congenital stationary night blindness in vision. Stevens also mentioned loudness dependence on bandwidth as an auditory phenomenon with no analog in vision.

The scope of comparison was further broadened by Julesz and Hirsh (1972), who noted the omnidirectionality of hearing, its high alertness, and the fact that it does not cease (as vision does when the eyes are closed). Their discussion is driven by perception-general logic, where hearing is dominated by events and less so by objects, which are not as spatially well-defined as in vision. They began developing the figure-ground concept from Gestalt psychology relevant to hearing, which Bregman greatly elaborated and brought to maturity in his theory of auditory scene analysis that in itself was largely inspired by the analogy to vision. Bregman (1990, p. 6) wrote: *"In vision, you can describe the problem of scene analysis in terms of the correct grouping of regions. Most people know that the retina of the eye acts something like a sensitive photographic film and that it records, in the form of neural impulses, the 'image' that has been written onto it by the light. This image has regions. Therefore, it is possible to imagine some process that groups them. But what about the sense of hearing? What are the basic parts that must be grouped to make a sound?"* Bregman preferred the term **auditory stream**—"*the perceptual unit that represents a single happening*"—to auditory event (Ibid., pp. 10–11), which is the physical occurrence that can be composed of smaller subunits (e.g., footsteps, notes in a melody).

Auditory objects have entered the jargon of auditory scene analysis notwithstanding, as subunits of the auditory stream. For example, Shinn-Cunningham (2008) defined the auditory object to be *"a perceptual entity that, correctly or not, is perceived as coming from one physical source."* As such, it

is taken to be the basic unit of auditory attention. Just as in vision, the listener can attend to only one auditory object at a time, and the different objects effectively compete for attention in a process that combines bottom-up feature extraction with top-down enhancement and control. These high-level similarities probably have physiological correlates, and significant similarities indeed exist between the visual and auditory cortices, which suggest that on an advanced processing level the sensory information may become modality-invariant (Rauschecker, 2015). Projections from both the visual and auditory cortices split into the "what" and the "where" processing paths, which relate to the perceived object attributes that are processed in each path (see §2.3). In a similar vein, it was shown that the human auditory cortex strongly responds to modulated stimuli, which are mathematically analogous to patterns that evoke similar response in vision (Schönwiesner and Zatorre, 2009). Such high-level processing similarities have been revisited several times (e.g., Massaro, 1972; Shamma, 2001; Kubovy and Van Valkenburg, 2001), as, for example, in the case of scene analysis processing similarities (Handel, 2006), or in the context of units that are tuned to spatiotemporal modulation in vision and spectrotemporal modulation in hearing (Shamma, 2001; Schönwiesner and Zatorre, 2009).

It can be seen that several authors have put vision and hearing on equal footing once perceptual analysis begins. However, vision (or visual perception and the analysis thereof) enjoys the existence of a well-defined optical image on the retina that is amenable for processing in the central nervous system, whereas hearing does not. Hearing has its cochlear frequency map (tonotopy) represented throughout the auditory brain (§2.3), but it constitutes a spectral representation and not an image in any intuitive way, as is the retinotopic map between the eye and the visual cortex. Therefore, several attempts have been made to model the **auditory image**, which can feed into higher-order processing such as required by scene analysis.

1.4 Objects and images in hearing research

The roles of the object and the image are well-defined in optical imaging—the underlying physical basis of visual sensation and perception. They can be expressed mathematically in unambiguous terms using the laws of optics and projective geometry, which enable complete prediction of the real optical image on the retina, prior to transduction (e.g., Le Grand and El Hage, 1980). However, a precise definition of a visual object that applies to the perceptual experience that corresponds to the optical object is not as straightforward (Feldman, 2003). Similarly, the concept of an object that applies equally intuitively to hearing has been elusive and controversial. Most hearing theories overlook this discussion, or rather dodge it by referring to the acoustical source as the de-facto object of hearing. Some hearing models relate to the auditory image without accounting for the object that produced it in the first place. Yet other higher-level hearing models posit the existence of a perceptual auditory object that may or may not follow an auditory image and may not have a definitive physical object it corresponds to. Different models of these three imaging stages (object, image, mental object) are reviewed below. It will be seen that they are generally inconsistent among one another, and, when contrasted, they portray rather confused conceptualizations of both the object and the image of hearing. The main features of all auditory image and object models, including the one presented in this work, are summarized in Table 1.1.

1.4.1 The acoustic object

A few accounts of the acoustical object—distinct from the auditory object—are described below. Unfortunately, these two concepts are often conflated in a way that is counterproductive for the understanding of hearing as a physical process. The main complication in arriving at a definition of

an acoustic object is that, unlike optical objects, acoustic objects continuously change in time. Also, one-dimensional time signals arriving at the ears do not map to two-dimensional spatial images as in vision. Auditory objects are coupled to auditory events, which modulate the active duration of the acoustic sources. Does a spoken word constitute an object? Or should the mouth or the person that utters the word should be considered the object? What happens to the object when the source ceases to emit sound? Possible answers are inconsistent between the visual and auditory domains: in optics and vision we will never refer to the reflected light pattern from the object as the object itself, whereas in hearing the object is sometimes taken to be the actual transient sound—e.g., the word—and not the vocal cords, mouth, or person that produces it.

Acoustic sources are routinely presented in a naive manner within physical acoustics, which sidesteps any philosophical challenges regarding objecthood. An acoustic source is anything that creates pressure vibrations, which can acoustically radiate through the medium that surrounds it. For example, it can be a rigid body like a string, a plate, a larynx, a loudspeaker, or a locus in a fluid that undergoes disturbance. Vibration and radiation require an investment of energy, so the source has to be forced in order to vibrate. The acoustic source vibrations have the potential to become the object of sensation if they vibrate at frequencies that are within the animal's hearing range, and arrive to the hearing organ at a level that is above their hearing threshold. A passive acoustic object may be subjected to echolocation—targeted radiation by echolocating animals—whose reflection from the object contains information about its geometry and position.

Another practical way to sidestep the problem of objecthood is to consider auditory stimuli, acoustic signals, or simply "sounds", as the entities that are to be sensed by the ear. This terminology dissociates the radiated vibrations from their particular source, assuming that an arbitrary method can be devised to produce the same vibrational pattern, such as a mechanical instrument, a person, a loudspeaker, an array of loudspeakers, or even a direct electrical stimulation of the cochlea. This useful approach runs the risk of losing touch with the geometrical and mechanical acoustics of real sources, which the animal may have adapted to associate with the sounds, possibly using various imperfections in the signals and multimodal cues.

1.4.2 The auditory image

Despite the substantial commonalities between hearing and vision, there is an ostensible asymmetry between them at the level of the cortex, due to the different peripheral extent associated with each sense. While the visual periphery produces an optical image, the cochlea does not produce an obvious image, but rather a multiband filtered version of the sound stimulus—often compared to a spectrogram or to the output of a Fourier analysis. Several models have attempted to make up for the missing link by hypothesizing an auditory image, which may have some analogous properties to the visual image and may therefore provide a gateway for further processing. All the models skip the acoustical transformations that take place in the outer and middle ear and rather consider the image to emerge either somewhere at the cochlea or on the various auditory nuclei in the brain, where it necessarily manifests in the neural domain.

The simplest conceptualization of an auditory image is due to Kemp (2002), who suggested that the traveling wave in the cochlea is itself an image of the acoustic object. This image represents a size mapping of larger objects—dominated by low frequencies that are mapped to the apical cochlear region, and smaller objects to the basal region of the cochlea. This is the only auditory image model that relates to the concept of image sharpness. According to Kemp, cochlear filters that are actively sharpened by the outer hair cells produce sharper images of the intensity envelopes. Therefore, it is also one of few auditory-image models that relates the image to the physical properties of the

object[4].

There have been two prominent attempts to define auditory images that are not strictly spatial as in vision, but are instead composed of some combination of the temporal and spectral dimensions of the acoustic stimulus. Early incarnations of the model can be found in Lyon (1984), who highlighted the parallel two-dimensional laminae that are found along the auditory neural pathways as a likely target area for an image. The laminae are particularly attractive for imaging because they are tonotopically organized along one dimension found throughout the auditory brain, including the primary auditory cortex (§2.4). Related to Lyon's is the **auditory image model** by Patterson and colleagues (Patterson et al., 1992, 1995; Patterson and Holdsworth, 1996; Patterson, 2001), who proposed that sustained auditory images exist following cochlear and auditory brainstem processing, which includes multichannel compression, half-wave rectification, suppression, phase-alignment, and adaptation. Lyon (2018) further elaborated these ideas and called his model the **stabilized auditory image**. These two models consider the second laminar dimension to correspond to periodicity. They consider sound information to be coded in the temporal patterns of the neural spiking, through synchronization. Such processing readily produces correlational responses that are in line with Licklider's autocorrelation pitch and Jeffress's binaural localization models (Licklider, 1951b; Jeffress, 1948). The stabilization of the image is related to its "movie-like" property of being anchored with zero lag-time, which can be readily obtained using autocorrelation of the neural activity pattern of each auditory channel.

The resultant image from Lyon's and Patterson's models can be visualized by plotting the time series of the neural activity on the x-axis with all the parallel auditory channels on the y-axis. Pure tones and other periodic stimuli tend to appear as stable patterns on these plots (using a long integration time constant of 200 ms), whereas transient sounds and random noise decay much more rapidly and do not form stable images. Fine details are tracked through faster temporal integration (time constant in the order of 10 ms). This may be achieved by generating short pulses (pulselets) from the stimulus through "strobing"[5]. These models were criticized by Carlyon and Shamma (2003) as they fail to account for across-channel information (produced by the relative delay between channels) that is sometimes used by listeners and has to be extracted using a summary measure of the spectrogram (or image).

A similar idea to the image was presented by Carney (2018), who referred to auditory "**fluctuation profiles**" that appear in the inferior colliculus, which correspond to the dynamic changes in the low-frequency temporal envelopes of the coded signals across all channels. Here, the system optimizes the cochlear gain through its efferent system, in order to maintain an adequate level for coding in the auditory nerve that would otherwise be limited in dynamic range. While ocular focus, blur, and accommodation were invoked as motivation, none of these terms was employed in a more rigorous analogy, where acoustical and optical factors were compared.

A more ecologically motivated auditory imaging model was proposed by Simmons and colleagues, specifically for echolocating bats (Simmons and Stein, 1980; Simmons, 1989; Saillant et al., 1993; Simmons et al., 1996, 2014). These bat species produce periodic frequency-modulated vocalizations during flight, which are reflected from objects in their environments and are neurally processed to obtain information about the distance, shape, and movement of a remote target. This intricate

[4]This model may be a distant relative of a historical theory by J. R. Ewald, who posited that the basilar membrane vibrates with standing waves (rather than traveling waves), which give rise to an "acoustic image" that faithfully represents the sound and some of its characteristic effects (Wever, 1949, pp. 45–52).

[5]Patterson and Holdsworth (1996) also used the term "quantization", which in information theory is reserved for amplitude steps in the dynamic range, whereas "discretization" is used for generating samples, or symbols, from a continuous sequence. (The symbols are anyway quantized, assuming a finite dynamic range). The concept of strobing was used with no mechanism to explain it (Patterson et al., 1995; Patterson, 2001), in what can be thought of as a rough "sample-and-hold" processing (§14.4.3) through the various temporal integration stages.

biosonar process is usually compared to man-made sonar and radar systems[6], but here the returning echoes are used to construct an image of the remote target. In contrast to the above-mentioned general-purpose auditory imaging models, bats (and probably other echolocating animals) appear to be able to use the stored knowledge of the probing signal in order to obtain an image of the reflecting object and perform exceptionally fast and precise information processing on it. The bat echolocation system extracts information in at least two time scales: individual echoes merge if they are received with less than 0.3–0.5 ms separation. Thus, target (object) features are detected through direct comparison between transmitted and received reflected frequency chirps, which are on the order of 10 ns (!) or longer. The target and image in this case are much more similar to those that are familiar from visual imaging, as reflected waves are used to produce a spatial image of the target, which endows the animal with a superior three-dimensional model of the remote object compared to that achievable with passive hearing.

It is common in the audio jargon to talk about a **sound image**, especially in the spatial sense, but without defining it precisely. Typically, it relates to where vibrating objects are localized within the listener's mental geometrical space (including inside the listener's head) and how large they are (e.g., Sayers, 1964; Heyser, 1974; Altman and Viskov, 1977; Toole, 2009; Moore, 2013, pp. 245–282). This is the context in which stereo or phantom images are discussed in audio engineering. However, this terminology is sometimes used loosely with respect to the object-image pair. For example, Moore (2013, p. 2) implies that the image is formed in space before it arrives to the ear: "*The sound wave generally weakens as it moves away from the source, and also may be subject to reflections and refractions caused by walls or objects in its path. Thus, the sound "image" reaching the ear differs somewhat from that initially generated.*" This careful wording is not unique, as another classic example can illustrate: "*The sonic image, if one could speak of such, is smeared in space behind the physical loudspeaker*" (Heyser, 1971). Heyser (1974) explained later: "*The subjective sound image, or illusion of sonic presence, is the final form of this figure when we want to study subjective properties.*" These examples show a conflation between objects and images that hint that all sound images are purely mental or subjective (e.g., Whitworth and Jeffress, 1961)—something that is not the case in vision, where a real optical image appears on the retina.

An early theory considered the **preperceptual auditory image**—a sustained version of the sensory information about the stimulus that can serve as the input for perception, in tasks such as pattern recognition, detection, and short-term memorization (Massaro, 1970, 1972). Accessing the preperceptual image is likened to a sequential readout process, which is dependent on the complexity of the stimulus. Conceptually proximate, for McAdams (1984), the auditory image is a metaphor of the internal form of sound objects that are automatically perceived as though they belong together and can be taken as fundamental units in music, for example. He defined (Ibid., p. 11): "*the auditory image is a psychological representation of a sound entity exhibiting an internal coherence in its acoustic behavior.*" He added: "*...there is an identity relation between a percept and its image, where the image notion serves as a kind of bridge between the percept and its interpretation (the concept or schema).*"

McAdams (1984) retained a relatively loose usage of the concept of coherence (a coherent stimulus can be a complex tone whose harmonics are modulated in phase, for example), which nevertheless affords the auditory system with the ability to stabilize and organize its image using a variety of cues and principles, in the spirit of auditory scene analysis[7]. McAdams (1982) noted

[6] In radar (RAdio Detection And Ranging), electromagnetic radiation is mainly used to detect and evaluate the distance and sometimes shape of targets of interest (Levanon and Mozeson, 2004). Sonar (SOund NAvigation Ranging) systems achieve the same using sound waves in underwater environments.

[7] Note that McAdams sometimes used the term "behavioral coherence" instead. See §7 for a breakdown of the different definitions of coherence found in literature relating to hearing.

different mechanisms for separation and grouping of sounds, which result in contiguous images that are also invariant to transformations (such as retaining a sung voice identity, despite vibrato). Later, Yost (1991) drew a stronger parallel between the auditory image and the acoustic source. Following various segregation and grouping processes, listeners can easily identify different simultaneous sources in a mix of sounds, which may seem inseparable just by looking at their neural patterns. It should be noted that Bregman (1990) himself generally reserved the word "image" for vision and only rarely did he use it to designate sound percepts that are distinctly separate, when they do not fuse with other sounds in the same stream.

1.4.3 The perceived auditory object

In the hearing research literature, the basic perceptual unit that is extracted from the auditory scene is often taken to be the "auditory object" (e.g., Kubovy and Van Valkenburg, 2001; Shinn-Cunningham, 2008; Bizley and Cohen, 2013; Nelken et al., 2014). As such, it is dependent on attention, context, familiarity, and other higher-level cognitive factors, which may help to segregate it from other objects and from its background. Griffiths and Warren (2004) suggested that the time scale over which the auditory object is formed need not be fixed and it can be drawn from patterns in the frequency-time plane at different time scales, which can also facilitate the separation of the object from its background.

Discussions about the auditory object do not usually state whether it arises after an auditory imaging step, or what the relation is between the putative image and object. For example, in Yost et al. (1989) the authors suggested that the auditory system uses binding of comodulated bandpass-filtered stimuli to form auditory objects, with no mention of intermediate images. In contrast, Griffiths and Warren (2004) drew on the auditory image model of Patterson et al. (1995) and employed a building block of a "*two-dimensional frequency-time object image in the auditory nerve*" that "*...might correspond to a sound source or an event.*". Other auditory object references are almost indistinguishable from those of auditory images. So, according to Bizley and Cohen (2013), auditory objects are fundamental, stable, perceptual units of hearing, which have neural correlates in the auditory pathways, starting from the cochlear nucleus but more prominent in the cortex—very similar properties to the auditory image models by Patterson and Lyon. In another parallel definition, Griffiths and Warren (2004) emphasized the role of memory and familiarity with the objects that is likely required in forming them—elements that overlap with the image models of Massaro and McAdams.

Treating the acoustic source as an object creates several difficulties, especially if compared to the more familiar visual object. Hirsh (1952) and Julesz and Hirsh (1972) downplayed the importance of auditory objects and preferred to relate to events instead. Bregman, who preferred to use the term "auditory stream" instead of "auditory object", observed that auditory objects are transparent and add up in loudness, whereas visual objects can block each other and generally have little effect on each other's brightness (Bregman, 1990, p. 121).

Another difficulty in the comparison between auditory and visual objects is that vision is mostly concerned with reflected light from surfaces, whereas hearing is concerned with the source itself—not with its reflections (Kubovy and Van Valkenburg, 2001). This neat distinction has been disrupted due to technological shifts over the last century that allowed for light sources to convey information directly (e.g., using traffic lights, or electronic displays) and for sounds to be reproduced using loudspeakers without a clear relationship to natural acoustic sources that are being reproduced. For some authors, these advances weakened the obvious differences between optical and acoustical objects, as there was little left in the acoustic source dimensions, shape, or other visible properties that could be described by listening, which undermines the very notion of acoustic object. Without

a doubt, these difficulties have created some confusion in the field, which tends to focus on mental objects that are elusive and do not exactly correspond to the acoustic sources.

1.4.4 Discussion

From the above overviews about the acoustic object, its auditory image, and the resultant auditory object, it can be seen that definitions are both inconsistent among one another and are in many cases rather vague and even confusing.

Perhaps the most problematic aspect of these concepts is that, much like the word "sound" itself, the term "auditory object" is used to refer both to the acoustical entity that evokes the auditory sensation, as well as its mental representation at the level of perception. The following quotation from Griffiths and Warren (2004, p. 891) illustrates the confusion between image and object aptly: "*Operationally, an auditory object might be defined as an acoustic experience that produces a two-dimensional image with frequency and time dimensions.*" According to this definition, the "acoustic experience" assumes the role of the reflecting surface (the object) in vision. Is the auditory object equivalent to the auditory image? Moreover, is the resultant auditory object external or internal to the perceiver? According to many perception models it is both, giving a hint of the dreaded **homunculus fallacy**[8,9]. But this is clearly an unhelpful conclusion if we aim at constructing a physical understanding of hearing with tractable cause and effect.

The problem of associating the acoustic source with objecthood is that an object is defined from the answer to the question—What do we perceive as an auditory whole (i.e., a single and coherent percept) (e.g., Nudds, 2010)? This is akin to reverse-engineering auditory perception in an all-inclusive manner, where phantom perception from within the system (e.g., tinnitus) produces messy answers. These approaches prescribe that perception becomes intermingled with physics and that objects cannot exist independently of the perceiver's brain. It is not the same as asking—What kind of acoustic radiation can be sensed by the auditory system?—which was answered in § 1.4.1 based on the physical system only. The difference between these two questions may have been a critical motivation for studying auditory scene analysis, wherein a scene is a collection of sound sources that are nevertheless perceived as separate entities by the listener. The inability to disentangle the two—the external sounds from the perceptual experience that they evoke—is a common thread in the early development of acoustics as a science, whose only method of observing (and thus measuring) sound was through listening (Hunt, 1992). It is also a source for a centuries-long debate in the philosophy of perception between the so-called **naive realistic view**, which states that we directly perceive sensory inputs, and **indirect realism**, which states that perception is forever indirect and therefore we can never directly access the world through our senses (e.g. Searle, 2015).

The auditory image is also riddled with problems. First, all auditory image models appear to have been developed without a direct anatomical analogy to the eye's image or its mathematical and physical imaging principles as are known in optics (see §4). This also means that the acoustical object that is imaged in each model is not always well-defined (with the exception of the echolocation imaging model; § 1.4.2). Second, and arguably a consequence of the first, some of the most critical concepts in optical imaging—e.g., focus, sharpness, blur, depth of field, aberrations—do not have

[8]The homunculus ("a little man", in Latin) fallacy suggests that a recursive chain of images is formed in the brain—each one is the object of a successive observer, ad absurdum (Attneave, 1960; Dennet, 1980; Nizami, 2017).

[9]A remnant of the homunculus fallacy may be inferred from the very choice of terminology that associates a mentally perceived entity—the visual or auditory object—with the word "object". We generally assign objecthood and objectivity to elements in the reality that is external to us and does not depend on our knowledge, whereas "subjective" are exactly these mental entities that are perceived or thought within the mind. Evidently, this oxymoronic reversal of the meanings of objective/subjective has already happened around the mid-18th century, whereupon these words had received the meanings that are in modern use (Daston, 1994).

Model	Object	Image	Physiology	Main features
Standard vision and optics	External spatial objects, reflected light, spatial modulation envelopes	Scaled spatial modulation envelope, overlapping color channels	Image on retina	Accommodated focus according to distance; pupil control; binocular integration in cortex; reconstructed 3D shapes
"Preperceptual image" (Massaro, 1970, 1972); Auditory image (McAdams, 1984); Yost (1991); Auditory object in Griffiths and Warren (2004)	Any sound stimulus, acoustic source	Undefined	Psychological; unspecified	Coherent input to perception prior to scene analysis
Simmons and Stein (1980)	Spatial objects under echolocation	Reconstructed objects	Neural; unspecified	Specific to bats; requires comparison between emitted and received sounds
"Auditory Image Model" (Patterson et al., 1992); "Stabilized Auditory Image" (Lyon, 2018); Auditory object in Bizley and Cohen (2013)	Temporal, periodic, broadband	Filtered, processed, autocorrelated	Neural, somewhere past the brainstem	Discrete; stabilized; dual time constant
Kemp (2002)	Spatial acoustic	Traveling wave	In cochlea	Image sharpness stems from filter sharpness; size is mapped to frequency
Carney (2018)	Low-frequency envelopes	"Fluctuation profile"	Inferior colliculus	Dynamic range optimization using the efferents
"Sound image" (audio)	Spatial distribution of sounds	The perceived spatial distribution of sounds	Unspecified	Binaural
"Temporal auditory image" (This work)	Narrowband temporal envelopes and the entirety thereof that relate to the acoustic source	Scaled temporal envelopes in parallel frequency channels and the entirety thereof	Inferior colliculus	Cochlear dispersion; cochlear time lens; neural dispersion; neural aperture; blur and focus; aberrations; depth of field; accommodation; coherent and incoherent dual processing in the brainstem

Table 1.1: Comparison of auditory imaging models and some of their key features. Conceptually similar models are grouped together.

meaningful correlates with these putative auditory images. These fundamental and intuitive concepts in optics, which have major implications in vision (e.g., focus accommodation, refractive errors of vision), are completely foreign to hearing science and remain inaccessible even with the available auditory image models.

While it is arguable whether vision theory is doing significantly better than auditory theory, it is undeniably better with respect to the visual periphery—the eye—its function, its optics, and its mechanics. Roughly, the eye produces an optical image of an object at a distance. The image that appears on the retina can automatically be made sharp by focusing the lens using accommodation, as well as by appropriately controlling the level of light by closing and opening the pupil. Hearing models that drew parallels to vision sometimes hypothesized the existence of an auditory image and/or an auditory object, but ignored all the other elements of visual imaging: the lens, its focus, the degree of image sharpness, accommodation, and the pupil—none of which have obvious analogs in hearing.

The common invocation in hearing of Marr's highly influential three levels of analysis (§1.2) is telling, because his theory takes for granted the sharp optical image that is formed on the retina, which serves as an input to post-retinal central information processing in vision. While nowhere stated explicitly, it is implied that the optics of the eye does not process information and does not execute any algorithm. This mistake is relatively inconsequential in vision[10], but it can be misleading in hearing, because if an auditory image exists anywhere, then it is most likely concealed within the auditory pathways in the brain and not in the periphery (§1.5). Alternatively, the eye can be recast as an analog computer, whose goal is to create an image, which can be understood as a solution to the imaging equations for light waves, using the mechano-optical periphery of the eye[11].

In vision, "object" is an overloaded term. It refers to the optical object that is positioned in front of the lens and is projected as an upside-down image on the retina. It also refers to the visual object that is experienced in perception as a result of the image sensation. According to the philosopher Thomas Reid (quoted in Duggan, 1960): "*Sensation is a name given by philosophers to an act of the mind which may be distinguished from all others by this, that it hath no object distinct from the act itself.*" For Reid, it is necessary to attend to the sensation in order for its output to turn into an object[12]. But optics does not really care about attention or intent—the image exists by virtue of the illuminated optical object, the lens, and the screen. Thus, the image of the eye is primarily a

[10] Marr's theory can be quite easily extended to include analog computation performed by the periphery. However, this would demonopolize the brain (and neurons in general) from being the sole information processor of the animal—something that is not currently discussed in biology and neuroscience, to the best knowledge of the author.

[11] The following passage about analog computation captures this idea almost fully (MacLennan, 2007, p. 27): "*...we may define computation as a physical process the purpose of which is the abstract manipulation of abstract objects (i.e., information processing); this definition applies to analog, digital, and hybrid computation... Therefore, to determine if a natural system is computational we need to look to its purpose or function within the context of the living system of which it is a part. One test of whether its function is the abstract manipulation of abstract objects is to ask whether it could still fulfill its function if realized by different physical processes...*"

[12] Other common definitions for sensation and perception tend to be somewhat circular. For example, in Goldstein (2014, p. 415), sensations are defined as: "*Elementary elements that, according to the structuralists, combine to create perceptions,*" whereas perception is "conscious sensory experience" (p. 412). Definitions in Mather (2011, pp. 140–141) are slightly more helpful—sensation is "*An elementary experience evoked by stimulation of a sense organ, such as brightness, loudness, or saltiness,*", whereas perception is "*A complex, meaningful experience of an external event or object, created from a combination of many different sensations.*" According to Merriam-Webster dictionary, sensation is "*a mental process (such as seeing, hearing, or smelling) resulting from the immediate external stimulation of a sense organ often as distinguished from a conscious awareness of the sensory process.*" And perception is "*awareness of the elements of environment through physical sensation.*" Perhaps it is not a coincidence that the two terms are often used interchangeably in literature.

product of sensation. Why should the auditory image be any different from vision? Why should the auditory object be dependent on the intent of the listener? We would like to break the circular logic of indirect perception theory by physically relating to the acoustic object and the auditory image as components of sensation, even if they are formed neurally—well within the brainstem or midbrain. This is the key for the present analysis.

1.4.5 A note about subsequent auditory imaging terminology

Given the above review and discussion, a note about the terminology that is going to be used in this work would be in place. In the literature, different concepts are encountered that variably relate to objects and images such as acoustic source, acoustic object, sound source, sound object, auditory object, auditory image, and sound image. Sometimes the object is material, whereas in other cases it is strictly visual and does not apply to hearing, and in modern usage it is often perceptual. It is a similar case with the image, which is sometimes taken to exist outside of the listener, or inside as a neural or mental representation. In this work, unless referring to specific jargon of another work, the adjective **auditory** is reserved for signals, stimuli, or entities that are observable within the animal's hearing system, especially within its neural pathways. External sources and objects are invariably considered **acoustic**. Similarly, in reference to vision, external objects and the retinal image are **optical** and all internal representations are **visual**. Although the word "physical" is often useful in this context too—to describe the acoustic or optical objects—perceptual images or objects are physical in the sense that all information has to be manifested physically, even if in neurally encoded form. Thus, the term **material object** is preferred to describe the physical object that is sensed by the animal.

1.5 Rigorous analogies between hearing and vision

In the previous sections, two weak aspects of auditory theory were highlighted. One aspect was the relative opaque role of some of the main auditory areas in the brain—mainly the brainstem. While it is known to be critical in extracting all sorts of low-level auditory cues such as used for localization, it does not have a well-defined function that can be communicated in simple terms, as part of a modular system. The second aspect was the unsatisfactory importation of the object and image concepts into hearing. The main concern in this work is the latter aspect, but through the development of the idea of hearing as an imaging system, several ideas will be explored concerning new hypothesized roles of the auditory brainstem and midbrain as well.

There are at least five different perspectives that motivate the temporal imaging theory that is at the heart of this work: the prominence of direct versus reflected radiation, anatomy and physiology, imaging mathematics, information and communication, and coherence. These perspectives overlap and complement one another and should not be taken as independent. Except for the anatomical perspective that is argued for in more depth, the other perspectives are presented qualitatively and briefly, and more rigorous derivations will be left for their respective chapters.

1.5.1 The prominence of direct versus reflected radiation

Among the senses, vision, hearing, and touch are specifically geared to deal with wave stimuli. In the case of touch, low-frequency vibrations (< 500 Hz) from an object require direct contact with the skin (at least at typical amplitudes) (Bolanowski Jr et al., 1988), which suggests conduction rather than radiation. Hearing and vision are unique in that the radiated waves propagate in the three-dimensional space around the animal, where reflections tend to be rampant—mainly from

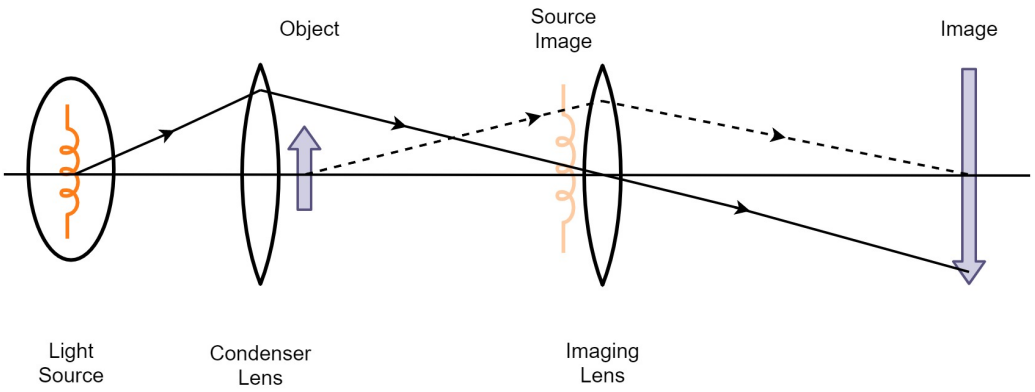

Figure 1.2: A simplified diagram of Köhler illumination. The light source is placed behind a condenser lens that distributes the light uniformly on the object. The object is in sharp focus of the imaging lens on the image plane, whereas the light source is completely defocused at this point. The illustration is based on Figure 32 in Goodman (1988, p. 153).

large objects and boundaries (which are also objects). Perceptually, however, the two senses treat the direct and reflected energy of the radiating source in completely different manners. The visual image is largely based on reflections of the light source, which is not particularly interesting as an object in its own right. In contradistinction, hearing is primarily interested in the source itself and much less in the reflections. And yet, just as the visual image reveals something about the source of light, which was not imaged directly (e.g., the location of the source, its color temperature and identity, its uniformity, its power), so does the auditory image contains some information about the reflecting enclosure that is not directly imaged (e.g., indoor or outdoor environment, room volume, the boundary materials, proximity to walls). In both hearing and vision, the supplementary information might be useful for the animal, but is considerably less accessible in perception than the primary object of interest. Why is this the case?

To illustrate the difference between vision and hearing with respect to their preferential treatment to reflected or emitted radiation, it can be telling to compare these situations to the Köhler illumination system, which is the most prevalent illumination technique used in microscopy and other optical systems (Köhler, 1893; Goodman, 1988). The problem that this technique solves is how to use an external source of light to illuminate an object and produce its image on a screen or in an eyepiece, but without getting the features of the lamp geometry (the shadow of its filament) and light distribution mixed with the features of the object itself (see Figure 1.2). The way in which Köhler solved it was to defocus the lamp's filament, so that the light that illuminates the object arrives to it as a uniform beam. Effectively, this amounts to producing a very blurry image of the lamp, which does not reveal any of its undesirable features[13]. Then, the uniform light is used to illuminate the object that is imaged in sharp focus by a second lens, which in turn projects the image on a screen (or inside the eyepiece). Now, if we compare this system to daylight vision, we see that the normally diffuse sunlight anyway does not disclose any easily observable features of the surface of the sun, nor its exact shape. Therefore, in daylight vision, the first lens of Köhler illumination is superfluous, whereas the lens of the eye assumes the role of the second lens in this system, which produces the sharp image of the object, but not of the sun.

Contrast this with hearing an object inside a room. For the sake of this presentation, we can assume that a source that is clearly heard may be considered to be in sharp focus. But the reflections from the surrounding surfaces, while they go on simultaneously with the source radiation,

[13] In modern implementations, the first lens is usually supplemented with a diffuser that further decoheres the beam.

are generally blurry. Even if we want to discern them, we generally cannot, as numerous reflections mix together and lose their individual character. Not being able to hear these reflections means that we do not directly "hear the walls", although information about them is contained in the acoustic field. Why is it so different from the visual objects? How does hearing achieve this?

Part of the explanation lies in the different wavelength and frequency ranges that are associated with light and sound. Audible frequencies are low enough to be amenable to direct neural processing that has small time constants, which are suitable for analysis of the acoustic signal phase. As it turns out, the phase functions that are associated with the acoustic source and reflections are qualitatively different, in a way that will be quantified later using coherence theory (§8). The auditory system can take advantage of this difference and accentuate it using transformations that are analogous to the optical ones from imaging theory. Effectively, the ear further defocuses partially coherent reflections so they do not come at the expense of the coherent object itself (§15). The combined partially coherent image contains information about both the source and its environment. Once again, this is in contradistinction to vision, which is based exclusively on incoherent imaging, as natural light sources and most artificial lighting are incoherent sources too.

1.5.2 Anatomy and physiology

A fair comparison between hearing and vision should be anchored to anatomically or physiologically homologous structures of both sense organs. Let us try to identify what these structures are.

In the cochlea, each inner hair cell (IHC) is innervated by about 10 nerve fibers (in humans), which then project to the cochlear nucleus (CN) and on to the superior olivary complex (SOC), lateral lemniscus (LL), inferior colliculus (IC), medial geniculate body (MGB), and the primary auditory cortex (A1). Different signal pathways downstream from the CN exist, including via subnuclei that are not mentioned here, although nearly all of them synapse at the IC (see Figure 2.4).

In contrast to the ear, the photoreceptors of the retina are innervated by a network of neurons both vertically (leading to the retinal ganglion cells) and horizontally with interneurons (Sterling, 2003). Depending on lighting conditions, photoreceptor type (cones or rods), and place on the retina, there are usually 1–3 neurons in the direct path between the photoreceptor and the ganglion cells (a combination of horizontal interneuron cells, bipolar cells, and amacrine interneuron cells; see Figure 1.3). The ganglion cells project primarily to the lateral geniculate body (LGB), and from there to the primary visual cortex (V1). In low-light conditions, rod photodetection is dominant, where many rods (120 on average) converge to a single ganglion cell. Considerably less cones (6 on average) converge to each ganglion cell in daylight conditions, with the least convergence taking place around the fovea at the center of the retinal field.

Four possible comparisons between the eye and the ear are considered, which are illustrated in Figure 1.4 and are explained below.

Equal-level receptors

The naive way to compare the two systems is by setting the sensory receptors of the ear—the IHCs—and the photoreceptors of the eye—cones or rods—on comparable levels (Alignment 1A in Figure 1.4).

According to this comparison, the auditory brainstem should be at a comparable processing level to the retinal neurons (e.g., Sitko and Goodrich, 2021). However, the brainstem contains more synapses and has a completely different architecture than the different neural networks possible in the retina. One interpretation for this disparity is that the auditory signal has gone through more processing than the visual one by the time they reach their respective cortices (Nelken et al., 2003; King and Nelken, 2009). This account is unattractive from a system-design perspective, because

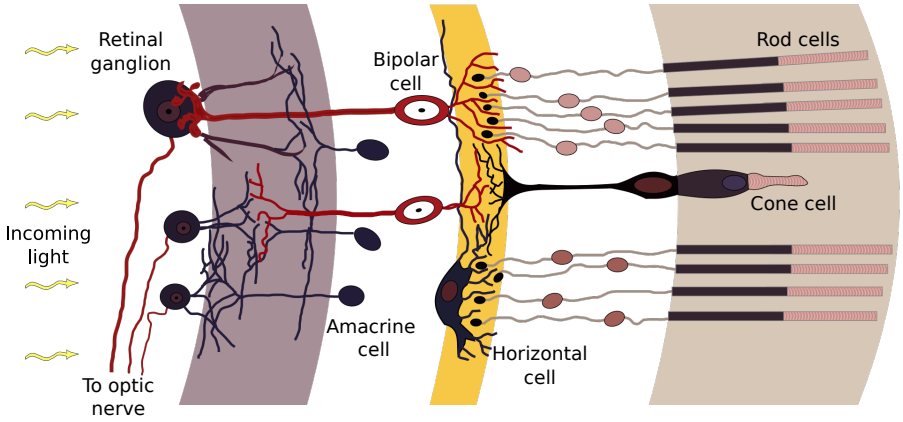

Figure 1.3: A section of the convergent network of the human retina. Light arrives from the left and is detected by the photoreceptor layer on the right after traversing through the intermediate transparent cell layers. Additional processing then goes from right to left. Labels added to illustration by Anka Friedrich and Chris, whose original illustration was derived from Ramón y Cajal (1911), https://en.wikipedia.org/wiki/Retina#/media/File:Retina-diagram.svg.

cortical theory would be more parsimonious if the functions of the cortex and thalamus are relatively consistent regardless of modality, unlike what is suggested by this comparison. For example, it has become increasingly clear that the cortex is highly plastic and areas that were once thought to be dedicated to one function can be repurposed by another, given the right circumstances. So, the areas associated with the auditory cortex in deaf people (and animals) may be used for visual processing (Kral, 2007). Therefore, on some level of processing abstraction, different modalities may be processed as equals, either due to the stimuli or due to the circuitries that process them (e.g., Handel, 2006; Sievers et al., 2021), which makes this comparison unconvincing.

A variation of this comparison (Alignment 1B in Figure 1.4) is that the retina and auditory brainstem contain a comparable number of synapses, which is just enough to have A1 and V1 on analogous processing levels (Rauschecker, 2015). However, the architectural differences between the retina and the brainstem are vast and their functions and complexity do not obviously overlap. Probably the most damning difference between the two networks is that the photoreceptors **converge** to fewer ganglion cells, whereas the IHCs **diverge** to many more auditory nerve fibers than there are IHCs. It suggests that the retina is constructed so to reduce the amount of information that reaches the eye before relaying it to the brain[14], while the auditory periphery is designed to conserve as much information as possible before central processing commences.

Therefore, comparing the hearing and vision by setting their receptors at equal levels seems misguided.

Equal-level ganglion cells

Another variation of the previous comparison is that the peripheral nerve cells of the ear (the spiral ganglion cells) and of the eye (the retinal ganglion cells) are set on comparable levels (Alignment 2 in Figure 1.4). This alignment suffers from unequal processing levels of A1 and V1 and it cannot deal with the convergence/divergence asymmetry, as it places the auditory and optic nerves on equal levels. The former is still composed of many more fibers than IHCs that feed into the auditory brainstem. The optic nerve contains less fibers than photoreceptors and projects primarily to the

[14]In low-light conditions, the convergence of rod inputs optimizes for low signal-to-noise ratio conditions, so that photon-activated rods are enhanced, while the noise from adjacent rods is inhibited. In daylight conditions, the cones are configured to avoid saturation, which also entails information overload (Sterling, 2003).

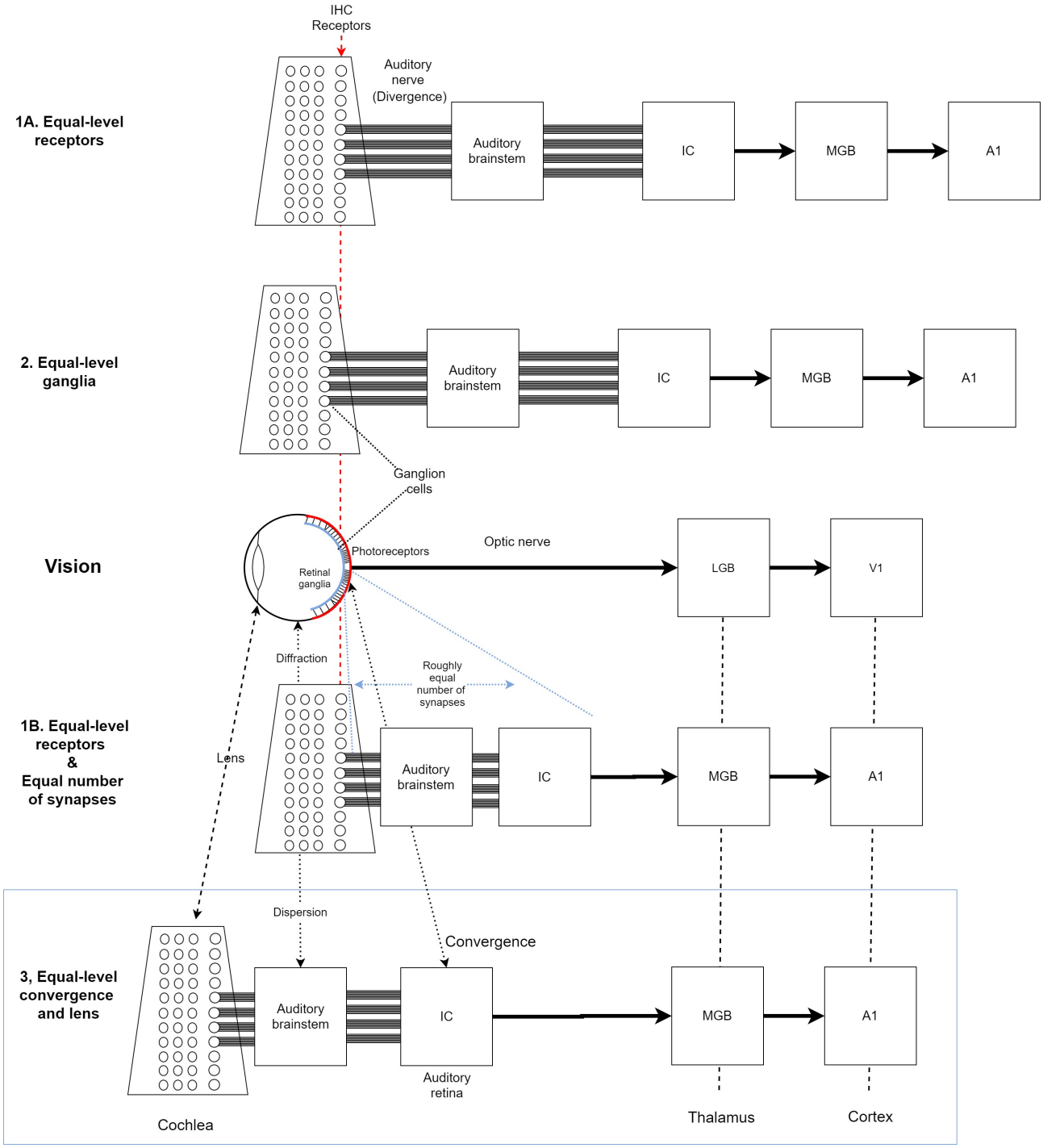

Figure 1.4: Illustration of different anatomical and physiological analogies that are possible between the eye and ear. Four alignments of the auditory system are considered relative to the visual system, which is sketched in the middle row of the plot. If brain areas of the two systems are vertically aligned, it indicates that they are on equal-processing level. In areas that are too crowded in the image, dashed lines indicate alignment. Note that the photoreceptors are located on the right of the retina (drawn in red), while the retinal ganglions are on the left (in blue). See text for further details.

LGB of the thalamus. It is usually noted that although both are counted as cranial nerves, the optic nerve is, in fact, part of the central nervous system, whereas the auditory nerve is part of the peripheral nervous system. Thus, this comparison suggests that by the time that the two signals reach the cortex, considerably more neural processing of the auditory signal has taken place than of the visual one, which once again implies a processing disparity. Therefore, this comparison seems misguided as well.

Equal-level cortex and thalamus

The last possible comparison aligns A1 and V1 of the cortex and the LGB and MGB of the thalamus (Alignment 3 in Figure 1.4). Aligning A1 and V1 makes sense due to the many perceptual analogies that were found between hearing and vision. This, in turn, automatically aligns the auditory and visual thalamic levels as well. The photoreceptors and the retinal neurons should be then aligned against the IHCs, the auditory nerve, and the auditory brainstem and midbrain nuclei. The major point of connection to the auditory thalamus in the MGB is the IC, which should then be comparable to the retinal ganglion cells[15]. However, to avoid the convergence/divergence asymmetry, rather than comparing individual neuron layers, it is more sensible to place the retina as a whole against the IC, which is the only auditory nucleus with considerable and unmistakable convergence (Figure 2.4).

This solution sets the most peripheral parts of the eye—the lens and the vitreous humor (between the lens and the retina; Figure 4.7)—on an equal level to the auditory periphery and brainstem. It suggests that part of the early hearing function is better achieved in the neural domain than in the analog domain. But once processing moves to the neural domain, it is subjected to the same neural reality that other parts of the brain are subjected to, as various auditory circuits can be excited, inhibited, or neuromodulated by other circuits. These diverse functions may be contrasted with visual accommodation, which provides a neuro-mechanical modulation of the peripheral visual function—a kind of preprocessing that is applied before the optical signal reaches the retina.

This anatomical comparison between the eye and the ear blurs the traditional distinction between peripheral and central processing. Or rather, if hearing requires neural processing to achieve a function that is achieved in the "analog" visual domain, then the border between auditory sensation and perception arguably becomes less obvious.

A cartoon comparison of the anatomical levels of the auditory and visual systems up to the image level is given in Figure 1.5. A more thorough overview of the auditory anatomy is given in §2 and arguments for why the IC is the most adequate organ for the auditory image are provided in §11.

1.5.3 Imaging theory

The eye is an optical imaging system. As imaging theory in optics has been thoroughly studied at different levels of abstraction and given that it has several parallels in hearing, it can aid us in refining the concept of auditory imaging.

A perfect optical image is a linearly scaled light pattern of (the projection of) an object (§4.2). When the object is illuminated and placed in front of a lens and a screen is placed behind it, its image is obtained from a simple geometrical law that relates the curvature of the lens to the distances between the lens and the object and the screen. A critical condition for the image to remain linearly scaled (i.e., uniformly magnified or demagnified) is that the object should subtend only small angles from the imaginary axis that connects the object and the lens centers.

[15] An anatomical equivalence between the IC and the retina was noted by Carney (2018). A functional equivalence was indirectly implied as well, although instead of an optical image, the IC deals with "fluctuation profiles".

Applying more rigorous wave physics, spatial imaging can be also shown to be a combination of three processes—a diffraction, a lens curvature operation, and another diffraction. Mathematically, the three can be expressed as quadratic phase transformations, which cancel out when the imaging system is in sharp focus. The image itself is then understood as an intensity pattern of a spatially modulated light source of a much higher carrier frequency. Typically in the optical analysis, light is taken to be monochromatic—a constant frequency, effectively like a pure tone in acoustics, or a fixed carrier in communication—so the changes in intensity relate only to brightness, but not to color, perceptually. In human color vision, three narrowband channels of the cone photoreceptors are normally available, which can be mapped to three monochromatic images. Each image is a demodulated version of the light intensity pattern, in which the high-frequency carrier is discarded. A polychromatic (color) image may then be expressed as the combination (i.e., an incoherent sum) of different monochromatic images within its frequency range.

The mathematical analogy here requires us to use the space-time duality, which was the perceptual analogy found between the spatial dimensions of vision and the temporal dimension of hearing (§1.3). Originally, the mathematical space-time duality was discovered in nonlinear optics by Akhmanov (Akhmanov et al., 1968, 1969), who obtained a formulation to the wave equation that is in every way analogous to the wave equation used in imaging optics, only with interchanged time and space coordinates. To use this analogy, we retain the monochromatic carrier, but apply the modulations in the time domain as instantaneous frequency variations around the carrier, instead of spatial frequency modulations as in standard optics. Mathematically, we lose the two spatial dimensions of the image, which were previously applied to the spatial frequency, leaving intact only the propagation coordinate of the plane wave. Thus, the quadratic phase transformations no longer describe changes in the spatial object, but rather temporal changes to a pulse, which correspond to amplitude and frequency modulation and to the effect of group-velocity dispersion[16], such as exists in the cochlea itself.

In temporal imaging we also resort to narrowband channels that are associated with a monochromatic carrier, which in itself is not used directly in the imaging calculation. The temporal object is the envelope of a pulse input, which relates to the finite range of spectral and temporal variations that can be captured by each narrowband filter. Using the auditory filters of the cochlea as narrowband channels, each one can linearly handle small temporal modulations relatively to the pulse center, which is analogous to the small-angle condition in spatial imaging. The existence of a lens in the auditory system will be explored in depth in §11.6. The image of the pulse is obtained after additional processing beyond the lens, where demodulation may take place as well. It should be mentioned that unlike the eye, which produces a demagnified image, we know of no scaling that takes place in the auditory system. It suggests that the system magnification may be very close to unity. An illustration of the auditory and the visual system analogous parts and functions is given in Figure 1.5.

Dealing with sound pulses as images means that the input is composed of short samples that have to be integrated to be perceived continuously. This is in analogy to the photoreceptors that spatially sample the entire area of the retina, whose exact distribution sets the limits on the maximum spatial frequency that can be detected by the eye.

The analogy between spatial modulation in optics and temporal modulation in psychological acoustics motivated Houtgast and Steeneken (1973) to adopt the modulation transfer function in acoustics, which directly impacts speech reception in reverberation (Houtgast and Steeneken, 1985). More directly related to hearing, Joris et al. (2004, pp. 544 and 565) suggested, in passing, that the

[16]When the velocity of the wave depends on frequency, then the medium is considered dispersive (§3.2). If the group velocity is also dependent on frequency, then the effect is of group-velocity dispersion (§10.2). It is equivalent to talk about group-delay dispersion instead, which relates to the same physics.

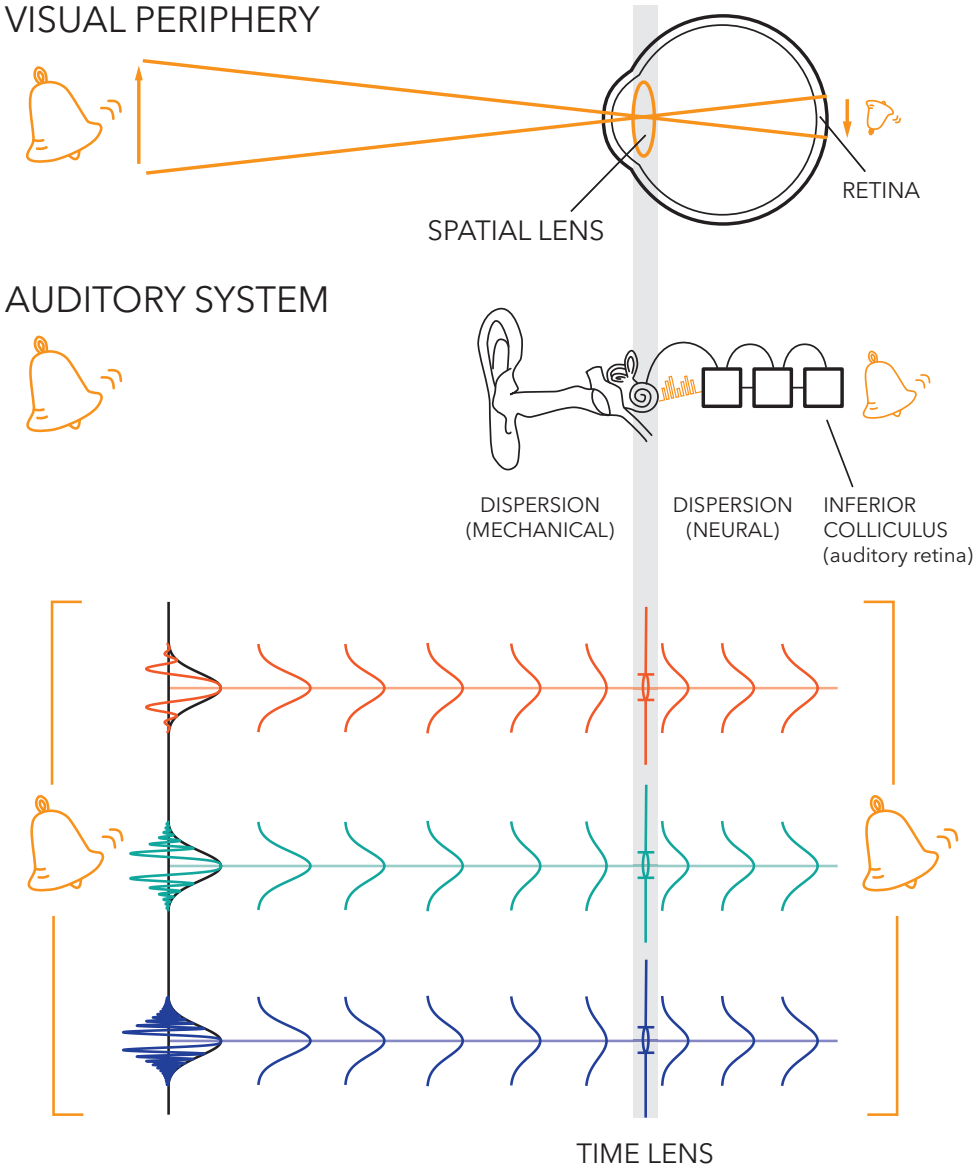

Figure 1.5: Cartoon comparison between auditory temporal and visual spatial imaging. A time lens resides in the cochlea and is analogous in processing level to the crystalline lens of the eye. The auditory image appears in the inferior colliculus, which is analogous in level to the retina. Unlike vision, information from the carrier phase may be conserved in addition to the envelope information that is used in both senses. In both cases, the image is a combination of the monochromatic images from different frequency channels—either tone or color. Original illustration by Jody Ghani (2020).

spectral contents of the temporal envelope of sound may be analogous to the spatial frequencies in vision.

While the mathematical space-time duality is very robust, fully motivating it may seem far-fetched without additional theory that establishes the concepts of group-delay dispersion, time lens, aperture, and coherence, in a way that is applicable to acoustics and hearing. The relevant theory will be developed over the next chapters and the components of the auditory imaging system itself will be explored from §10 onwards.

1.5.4 Information and communication

It is common to ascribe information processing to the brain function, especially with respect to perception of information-bearing signals and stimuli that are detected through sensory channels (e.g., Sterling and Laughlin, 2015). Such analyses tend to examine the receiver's side of the communication chain (comprising a source, a channel, and a receiver), which neglects the physical transfer of information from the source, through the environment, and into the sensory receptors of the animal. However, actual transmission of information depends on engineering a communication system that is shared between the source and receiver over a physical channel[17].

Any kind of communication depends on the ability to physically transmit and receive waves of an arbitrary type that are manipulated to carry messages. Communication signals can be formulated, without loss of generality, as a product of a high-frequency carrier and a slow-varying complex envelope, which modulates the carrier. The message is generally taken to be contained in the complex envelope of the signal rather than in the carrier itself, and is then referred to as baseband signal. In communication, the goal of the receiver is to recover the message from the received signal with minimum loss of information or distortion. This process is done by demodulation, which entails the separation of the envelope and discarding of the carrier. All long-range communication is physically realized by modulating well-defined channels with known frequency bandwidth. If the channels are narrowband (their bandwidth is much smaller than the carrier frequency), then it is possible to unambiguously recover the message[18].

There are cases in which the amount of information in the message requires a large bandwidth that cannot be fitted in the narrowband channel without breaching it, which would result in distortion and loss of information. The simplest solution is to employ a higher-frequency carrier and use a channel that has a relatively narrow but absolutely wide bandwidth. In practice, this is not always possible, if such high frequencies are not available for different reasons, such as prohibitive energetic cost of generating the carrier, interference from other communication or background radiation that occupies the same bandwidth, high absorption by the medium, strong interaction with objects in the environment, etc. If this solution is ruled out, communication cannot be considered narrowband. The main difficulty arises because wideband signals cannot be expressed in a mathematically unique form, so the concept of slowly-varying envelope cannot be well-defined for them. Therefore, receiving

[17] This communication framework is found in Claude Shannon's seminal work that was originally named "A mathematical theory of communication" (Shannon, 1948), but presents a system that is completely abstracted from the nitty-gritty mechanics of generating the communication signals in reality. The universal principles of his theory, which is now referred to as **information theory**, are integrated into the somewhat more mundane communication engineering, which deals with the mathematical principles that are embedded in the electronic system designs that actually deliver the information. Both sciences are tightly related and both contain universal principles that are not limited to a specific implementation in hardware, software, or the choice of physical medium and transmission energy. See §5 for further details.

[18] It should be understood that modulation and demodulation happen below the coding level that is familiar from information theory, even though some complex digital modulation techniques appear like codes in and of themselves. Therefore, possible message encoding can only take place before modulation, while decoding must take place after demodulation (see §5).

wideband signals may give rise to ambiguity in the demodulation and message recovery, unless a system is devised to disambiguate the received signals and, hence, the messages. Technically, it means that the received message may suffer some distortion. Hearing overcomes this problem in part by dividing its bandwidth to narrowband channels, which overlap to such an extent that there can easily be some redundancy in their detection that may be used to minimize ambiguity. While the broadband signal that is recovered from such a configuration may not be a faithful reconstruction of the original input (should this be its goal), it is possible that its informational content may remain largely intact.

There are two general categories of signal modulation detection. In noncoherent detection, only amplitude information is demodulated and the phase is ignored. In coherent detection, both the phase and the amplitude of the complex envelope are demodulated, which is essential in frequency- and phase-modulation techniques. In order to coherently detect the modulation phase, it is necessary to track the signal phase, since its carrier frequency tends to drift in transmission. The most common electronic circuit used to facilitate coherent detection in various engineering applications is the phase-locked loop (PLL), whose output is literally locked on to the carrier phase, so that no information is lost. In general, modulation detection is a well-studied feature of hearing, which behaves as if both coherent and noncoherent detections may be employed in different situations, as will be discussed throughout this work. Interestingly, phase locking is a robust feature of the auditory system, which is nevertheless limited to low-frequency carriers.

Evolutionarily, the hearing system long preceded auditory communication, at least in its verbal, human form. How did the hearing and communication functions converge? Acoustic objects and the cavities they are often coupled to are characterized by resonance frequencies that depend on their geometrical and physical properties. When objects oscillate, it is a result of them being forced into motion, using internal or external sources of energy. The forcing pattern itself can be thought of as modulation that shapes the vibration patterns of the object. For example, listening to a plucked guitar string, the pitch of the string is determined by its resonant frequencies, whereas the plucking is determined by the modulation, which has onset and offset temporal patterns, as well as input power. If, in addition to plucking, the guitar player also bends the string, it results in frequency modulation—a time-dependent change in pitch around its natural tuning. Both plucking and bending may be expressed as slow changes to the complex envelope of the string resonant frequencies. The constant-level frequency resonance is mathematically no different from a carrier. Each frequency may relate to its own channel, as long as it is analyzed independently of other resonant frequencies. Therefore, the mathematical basis for communication and acoustic source hearing has natural commonalities. Treating the guitar sound as a communication signal gives the listener the access to information both about the string itself, as well as to how it was forced into vibration. The two constitute two separate dimensions of information, which nevertheless tend to overlap in frequency and to interact in complex ways.

While the present work has been written with optical imaging theory as its main source of inspiration, communication theory became an indispensable element in many of its chapters. The most pertinent aspects to hearing of communication, information, and imaging theories are presented and contrasted in §5. The analytic signal and the envelope domain in the context of hearing are reviewed in §6. The auditory phase locked loop is in §9. Realistic acoustic sources are reviewed in §3. Aspects relating to the balance between noncoherent and coherent detection are discussed throughout Chapters §§ 16 to 18.

1.5.5 Coherence

The concept of coherence was used above in three different contexts: as a feature of communication detection that preserves the signal phase, as a necessary quality of the complex auditory stimulus that endows it with objecthood, and as a property of the wave field that determines how certain information about the source propagates. These definitions and additional ones that are relevant to hearing are reviewed in depth in §7, where a unifying definition is sought, and in §8 where the relevant coherence theory from physical optics is introduced in consistent manner. Source coherence and its acoustic and then physiological propagation are shown to be key to understanding auditory imaging. However, coherence goes beyond that, because the key to conserve coherence is in phase locking (§9), which is a feature of synchronization in the brain that seems to have correlates that are more universally important than just in hearing. Brain synchronization to a stimulus is an indicator for attention and thus for the engagement of the animal with certain input channels and actions. While it is perhaps a speculation of a sort, on a high level of abstraction, coherence seems to be the currency of important signals in the brain, regardless of their modality. This is speculated to be the key to auditory accommodation in §16 and an overarching processing design motivation in the auditory system as a whole (§18).

1.6 Conclusion

We began the chapter with an overview of some of the milestones of hearing science that have attracted the most attention in research, but also emphasized some gaps in what appears to be a rather loose and fragmented theory. Then, we dwelt on how some commonalities and differences of hearing and vision have repeatedly attracted scholars to formulate hearing models using visual concepts. These ideas have also been used to elucidate how each modality is specialized within perception as a whole, with the recurrent observation that hearing is predominantly temporal whereas vision is predominantly spatial. Perhaps the most common of all visual concepts that have found their way to hearing are that of the object and the image. However, these terms have been inconsistently applied in hearing and have failed to retain the insight that they carry for vision. Subsequently, we presented five different perspectives that can be developed to obtain more insight into hearing theory, using concepts from vision, optics, communication, and wave physics.

Table 1.2 compares some basic parameters of hearing and vision in humans, which also relate to conclusions from the five specific comparisons above. Some parameters are well-known from literature, whereas others are based on the present work and are explained throughout. A similar (and shorter) parametric comparison is found in (Handel, 2006, p. 24, Table 1.1).

The complete auditory system that will be explored in this work in parts is displayed in Figure 1.6 for reference, but will be explained in detail only in §18.2.

	Hearing	Vision
	Carrier	
Energy	Acoustic (pressure)	Light (electromagnetic)
Typical speed	343 m/s (in air at 20°C)	299,792 km/s (vacuum)
Frequency range	20–20000 Hz	400–750 THz (nominal); 360–830 THz (maximum range)
Wavelengths in air	17.15 m–1.71 cm	400–700 nm (nominal); 360–830 nm (maximum range)
Period	50 ms–50 μs	$\approx 10^{-15}$ s
System bandwidth	10 octaves	0.9 octave
Channel bandwidth[1]	5–30%	15–20%
Modulation bandwidth	$\leq$ 2000 Hz (broadband noise)	$\leq$ 90 cycles / degree
Dynamic range	$>$ 120 dB	$10^{-6} - 10^{8}$ cd/m^2 (140 dB)
	Physical image	
Type	Temporal	Spatial
Governing equation	Paratonal (dispersion)	Paraxial (diffraction)
Image field	Complex temporal amplitude envelope, $a(t)$	Spatial intensity envelope, $I(x, y)$
Inverse domain	Complex spectral modulation envelope, $A(\omega)$	Spatial frequency intensity envelope, $I(k_x, k_y)$
Main assumption	Paratonal—slowly varying envelope, narrowband, short aperture (time window)	Paraxial—small angles, monochromatic
Spatial assumption	Plane waves	Paraxial / Gaussian beams
Typical invariance (still image)	Spatial	Temporal
Linearity	Complex amplitude (within channel), Intensity (across channel)	Intensity
Typical mode of imaging	Partially coherent	Incoherent
Primary objects imaged	Direct acoustic sources	Light reflecting objects
Sensory sampling rate	50–2500 Hz	25–75 Hz
	Detector	
Number of detectors	2	2
Channel frequencies	$\approx$ 3000–3500	3 colors (cones) + 1 intensity (rods)
Ability to move	No	Yes
Lensing	Cochlear time lens	Crystalline lens and cornea of the eye
Type of aperture stop	Temporal (neural, cochlear at low frequencies)	Spatial (pupil)
Image information medium	Neural	Optical
Image locus	Inferior colliculus	Retina
Sensitivity	quantum-limited (basilar membrane motion of the order of 10^{-12}–10^{-11} m)	quantum-limited $<$ 1–7 photons
Near point	N/A	25 cm
Far point	N/A	6 m
Means of accommodation	Lens curvature; phase locked loop gain; coherent to incoherent weighting (likely); level gain	Lens curvature (and pupil size and vergence)
Accommodation time	?	Minimum reaction time 0.4 s and response time 0.6 s (Charman, 2010)

Table 1.2: Comparison of properties and attributes of human vision and hearing. [1]Equivalent rectangular bandwidth of the auditory depends on the method used to measure frequency selectivity. The range in the table is based on minimum and maximum values from Glasberg and Moore (1990); Shera et al. (2002). A rough estimate of the full-width half maximum (FWHM) of the photoreceptor sensitivity curves was obtained from Hunt (2004, Figure 2.2a). Relative channel bandwidth in both vision and hearing decreases with frequency.

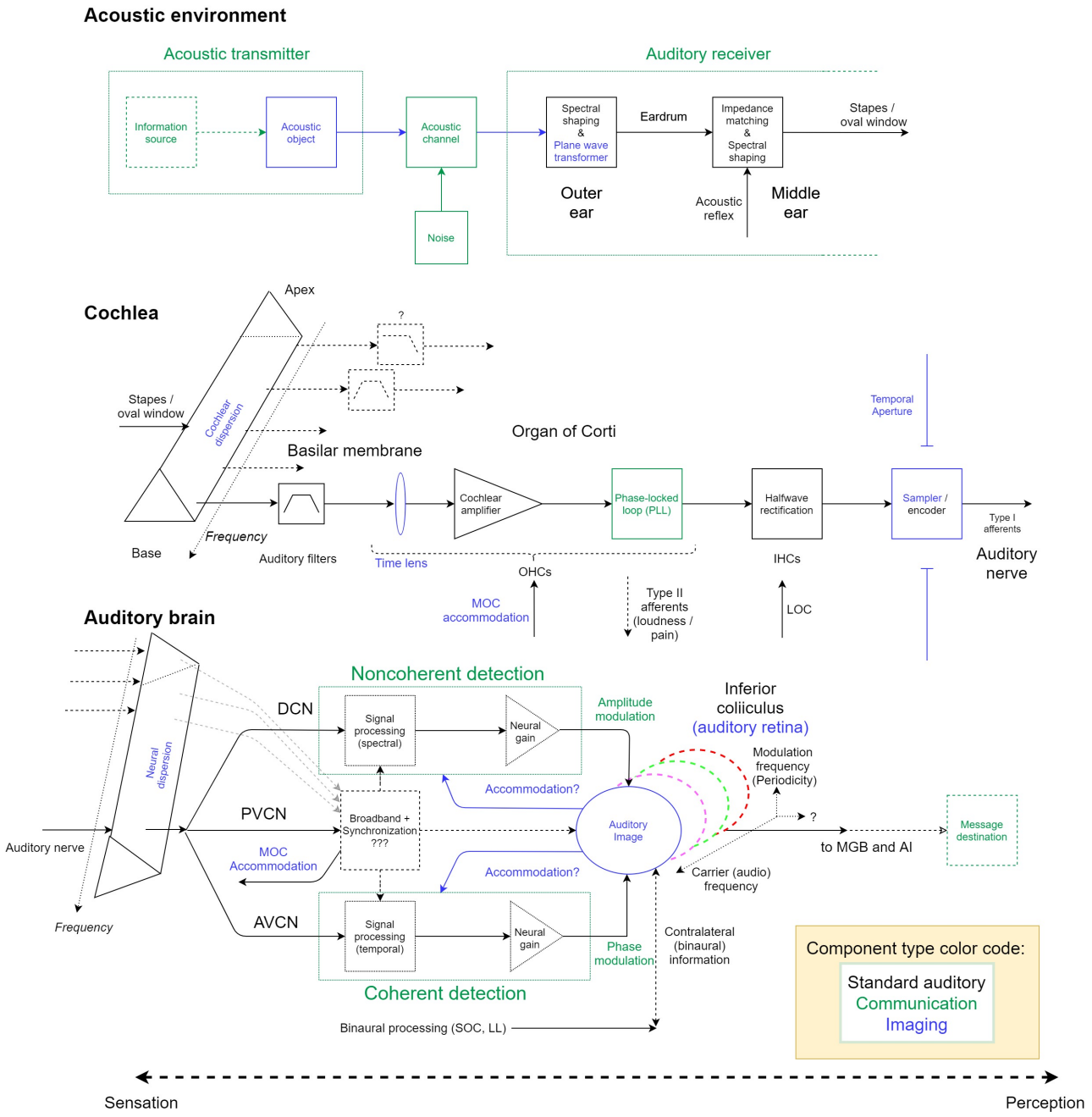

Figure 1.6: A functional diagram of the monaural auditory system as theorized in this work. The model contains standard auditory elements (in black). Novel elements to the standard auditory system are shown in green (inspired by communication) or in blue (imaging). The new components will be considered throughout the text and the full model is revisited in §18.