Chapter 1
Visual Influence on Auditory Perception

Adrian K. C. Lee and Mark T. Wallace

Abstract Auditory behavior, perception, and cognition are all shaped by information from other sensory systems. The research examining this multisensory view of auditory function is rapidly expanding and has demonstrated numerous interactions between hearing and the other senses at levels of analysis ranging from the single neuron to neuroimaging in human clinical populations. A great deal of evidence now suggests that areas of the brain traditionally considered auditory can be strongly influenced by information from other senses. This chapter highlights the multisensory world from an auditory perspective, in particular, focusing on the intersection of auditory and visual processing that has a profound impact on communication in everyday social settings. It is followed by an introduction of the chapters that make up this volume, which provide contemporary and comprehensive discussions on an array of concepts related to the behavioral, perceptual, and physiological aspects of audiovisual processing.

Keywords Attention · Auditory cortex · Autism spectrum disorder · Cross-modal · Multisensory · Neural typical · Oscillation · Plasticity · Prefrontal cortex · Reference frame · Scene analysis · Sensory integration · Speechreading

A. K. C. Lee (✉)
Department of Speech and Hearing Sciences, University of Washington,
Seattle, WA, USA

Institute for Learning and Brain Sciences (I-LABS), University of Washington,
Seattle, WA, USA
e-mail: akclee@uw.edu

M. T. Wallace
Department of Hearing and Speech Sciences, Vanderbilt Brain Institute,
Vanderbilt University, Nashville, TN, USA

Department of Psychiatry and Behavioral Sciences, Vanderbilt Brain Institute,
Vanderbilt University, Nashville, TN, USA

Department of Psychology, Vanderbilt Brain Institute, Vanderbilt University,
Nashville, TN, USA
e-mail: mark.wallace@vanderbilt.edu

A. K. C. Lee et al. (eds.), *Multisensory Processes*, Springer Handbook
of Auditory Research 68, https://doi.org/10.1007/978-3-030-10461-0_1

1.1 Introduction

From the galloping sound of a horse stampede to racing cars zooming past the finish line at a grand prix, the environment is full of events that emit information that is carried as energy and propagated in a variety of forms, such as light and sound. Although individual sensory systems have evolved to transduce this energy into meaningful signals in the nervous system, these individual systems must also work in concert to generate a unified and coherent view of the perceptual world. Additionally, the ability to leverage information across multiple senses can often increase performance, and this construct of multisensory integration undoubtedly impacts survival by conferring a selective advantage. Thus, a gain in the signal attributable to the presence of information in two or more senses can help predators better locate food sources and, conversely, allow potential prey to better avoid or escape their predators.

In humans, face-to-face conversation is a particularly relevant and commonplace multisensory experience. In addition to the meaning derived from the auditory speech signal, visual information from the facial expressions of the interlocutors is also informative for both the content and emotional context of the conversation. If this face-to-face conversation takes place in a crowded restaurant, looking at the talker’s lip movement can substantially improve speech intelligibility. Despite the ubiquity of the multisensory experience and the powerful associated behavioral and perceptual benefits, experiments in the laboratory have mostly focused on understanding how unisensory information is used to perform a task. Take, for example, experiments structured to examine the impact of spatial cues in auditory scene analysis as opposed to examining how visual information is combined with this auditory information in the context of the same task. A second related example are experiments evaluating the impact of available visual information alone in the naturalistic environment to help separate talkers in a crowded acoustic scene.

The aim of this volume is to provide a foundation of knowledge about the current state of understanding in regard to multisensory influences on auditory processes, with the goal of inspiring more rapid growth in scientific studies surrounding this topic.

1.1.1 Basic Concepts and Historical Perspectives

From a computational perspective, integrating information across the different senses is nontrivial. Take the face-to-face conversation as an example. Acoustic energy associated with the talker’s speech is transmitted as a time-varying pressure-wave signal, whereas visual information is transmitted as electromagnetic radiation (i.e., light waves). Exquisite and dedicated structures perform mechanotranduction in the cochlea and phototransduction in the visual system to turn these different forms of energy into electrical signals. These modality-specific or unisensory signals are first processed by dedicated neural pathways, which are assumed to be largely independent and hierarchically organized (see King, Hammond-Kenny, and Nodal, Chap. 6). In such a traditional and serially oriented view of sensory organization and function, it has been believed that only after these modality-specific computations have been performed can information from the different senses be combined and integrated to carry out multisensory computations.

Seminal work in multisensory systems strived to quantify the product of these multisensory computations and began with a framework in which the effectiveness of multisensory integration was operationally defined as the difference in response evoked by a combination of stimuli from two different modalities when compared with the response evoked by the most effective of its component stimuli (Stein and Meredith 1993). From the perspective of the single neuron, in which many of these multisensory operations were first characterized, the quantification of multisensory integration was summarized as the difference in neuronal firing rates and illustrated that multisensory convergence could give rise to either significant increases in firing (response enhancement) or significant decreases in firing (response depression). Furthermore, this work went on to show that these enhancements and depressions of response were often strongly dependent on the physical features of the stimuli that were combined. Thus, multisensory stimuli that were in close spatial and temporal correspondence generally resulted in response enhancements. Such an organization makes a great deal of sense relative to the physical world because stimuli that are in close proximity in space and time are highly likely to have originated from the same source. Hence, the nervous system can make inferences in common origin by evaluating the spatial and temporal statistics of a given stimulus pairing.

As for other fields of inquiry, the perspective of multisensory processing is often influenced by the field in which each scientist was originally trained. Thus, those coming to the multisensory field from a single-neuron neurophysiological background will focus on changes in neuronal encoding of individual neurons associated with having information present from multiple modalities. Conversely, those coming to the field with a neuroimaging perspective are much more interested in providing a more network-based view into multisensory function. Similarly, those trained in different sensory systems bring a different lens to their study of multisensory function. Whereas vision scientists often focus on spatially based tasks and the impact of adding sounds to these tasks, auditory scientists are generally more interested in questions of how auditory discrimination, such as speech comprehension, is impacted by the presence of visual cues (either concordant or discordant). In this volume, a flavor for these differing approaches and perspectives can be gleaned, but all unified from the viewpoint of better understanding how vision can shape auditory function.

1.2 Volume Roadmap

Auditory and visual information are seamlessly combined to form better perceptual estimates of the multisensory world. Alais and Burr (Chap. 2) begin this journey by describing how redundant cues from the different senses are statistically combined in a so-called "optimal" manner. For example and as alluded to in Sect. 1.1.1, a multisensory event or object is typically one in which the sight and sound energies come from the same place at around the same time. However, these energies are likely not of equal value for the nervous system because the spatial resolution of the visual system is better than that of the auditory system and the temporal resolution of the auditory system is better than that of the visual system. To consider this differential weighting, Alais and Burr make use of a Bayesian statistical framework, here known as maximum likelihood estimation, that combines information based on the uncertainty of the individual cues and thus generates an optimal model of behavioral performance that fits the empirical data exceptionally well. They also trace how this statistical cue-weighting model evolves over development as well as how it is impacted in circumstances of sensory loss.

The benefits of combining auditory and visual information are immediately evident when one tries to communicate in noisy and reverberant environments. Grant and Bernstein (Chap. 3) examine the auditory, visual, and audiovisual factors that influence speech intelligibility, such as which spectral regions of the speech signal are most important for audiovisual speech recognition and what information is complementary or redundant across auditory and visual speech cues. Audiovisual speech intelligibility research can be traced back to the seminal research conducted at Bell Laboratories (a part of the earlier iteration of AT&T) in the last century that addressed how different communication channel qualities can affect speech intelligibility. This line of research has expanded to include different metrics to predict speech intelligibility performance in noisy environments and now includes the significant work focused on listeners with hearing impairment. However, an understudied area is how speech intelligibility can be modeled in active speech communication settings involving face-to-face audiovisual input (i.e., beyond the well-studied auditory target-in-noise scenarios). Grant and Bernstein provide a unique perspective on audiovisual integration, with a distinctive focus coming from the design and evaluation of audio systems and hearing-rehabilitation devices.

In addition to boosting speech intelligibility in multitalker environments, visual information can also help listeners attend to the talker of interest, helping to solve the classic cocktail party problem (Cherry 1953). Lee, Maddox, and Bizley (Chap. 4) examine whether and how auditory and visual information can be grouped perceptually. They also argue that this perceptual binding could help select and focus on the source of interest in the presence of competing sounds. They go on to argue that multisensory grouping phenomena should be strictly delineated from multisensory integration (any process in which information across sensory modalities is combined to make a judgment) to facilitate a deeper understanding of how information is combined across senses. In this chapter, many classic multisensory illusions are revisited, and the authors ask whether there is evidence to unambiguously support the often-presumed perceptual binding cited in the literature. Coming from an auditory-centered perspective, the authors also focus on how visual information can help resolve auditory competition and suggest how future studies can focus on this understudied area of research.

Spatial cues are potent features for visual and auditory scene analyses, and combining spatial information from these two senses could potentially help an observer to better perceive surrounding events, especially in a crowded environment. Computationally, however, there is a significant operational challenge: auditory spatial cues are encoded in a head-centered framework (i.e., based on timing and intensity cues to the two ears), whereas visual spatial cues are initially encoded in an eye-centered framework (i.e., based on the eye-centered location of the visual stimuli available from the retina). Willet, Groh, and Maddox (Chap. 5) address this coordinate reference frame problem with a focus on how spatial information is coded in the superior colliculus, a major midbrain hub for multisensory convergence. In addition to focusing on the physiological properties that help solve these coordinate issues, Willet, Groh, and Maddox also provide evidence from behavioral investigations on how eye movements can affect auditory spatial tasks.

King, Hammond-Kenny, and Nodal (Chap. 6) take readers to a deeper exploration of the multisensory neural circuitry along the auditory pathway. In addition to focusing on visual influences on the auditory cortex and their implications for hearing, they also highlight somatosensory inputs along the auditory pathway because of the tight coupling between the motor aspects of speech production and the associated visual articulation cues. The authors argue that such a multisensory-based perspective will not only improve our understanding of the computational mechanisms of auditory cortical neurons but will also illuminate how perception and behavior can be influenced by multisensory interactions.

Multisensory interactions with an auditory component extend far beyond the auditory cortex. In Chap. 7, Plakke and Romanski look at how the frontal lobes support the processing of communication signals via the convergence of sensory inputs from many brain regions. They focus specifically on the ventrolateral prefrontal cortex, a region known to integrate face and vocal stimuli in nonhuman primates. These authors examine how factors such as the timing and congruence of the auditory and visual information shape how this information is integrated by these prefrontal neurons. Furthermore, the authors go on to review the deactivation studies of the ventrolateral prefrontal cortex that show the central role of this area in the integration of socially relevant face and vocalization information.

Where are the neural substrates for audiovisual speech processing in the human cortex? Beauchamp (Chap. 8) elaborates on several of the different neuroimaging approaches that have been used to address this question. As detailed, somewhat surprisingly, the anatomical and functional mapping studies of the early stages of auditory processing in the temporal cortex reveal this question to be one of ongoing and active debate. Based on evidence from postmortem studies as well as structural and functional magnetic resonance imaging studies, including data from the Human Connectome Project (Van Essen et al. 2013), subdivisions of the human auditory cortex are described. In effect, Chap. 8 highlights the challenges of delimiting functional borders given the individual differences across subjects and the limitations of a method that indirectly indexes neural activity. Beauchamp argues that multisensory processing may be a unifying principle that can further aid the functional parcellation of the human auditory cortex and its surrounding regions, particularly in the context of speech processing.

Despite these difficulties in accurately parcellating the human temporal cortex, there is abundant evidence from both human and nonhuman primate studies that the temporal lobe is an important site for multisensory processing. Perrodin and Petkov (Chap. 9) provide an overview of the cortical representations of voice and face content in the temporal lobe. Based on the results from studies that combine microstimulation and functional magnetic resonance imaging in monkeys, the authors provide insights on effective connectivity between the temporal lobe and the prefrontal cortices and suggest that these sites within the temporal lobe are critical convergence sites for auditory and visual information positioned between sensory-specific cortices and the executive control circuits of the frontal cortex.

The strong connectivity between temporal and prefrontal cortices raises the important question of how information across these brain regions is shared and coordinated. To address this question, Keil and Senkowski (Chap. 10) introduce the concept of neural network dynamics, as reflected in neural oscillations, to describe information processing across different cell assemblies. They argue that such analysis of oscillatory cortical activity provides valuable insight on the network interactions that underlie multisensory processing and, more broadly, any perceptual and cognitive tasks. Based on converging empirical observations, the authors conclude that it is likely that different oscillatory frequencies, reflective of different spatial scales of network assembly, index different facets of multisensory processing.

Can auditory perception be changed as different cross-modal experiences are acquired over time? It is well-known that neuroplasticity is pronounced during development, but there is now a great deal of evidence suggesting significant plastic capacity for the mature brain. Bruns and Röder (Chap. 11) review evidence that spatial, temporal, and speech identification tasks carried out by the auditory modality can all be influenced by cross-modal learning. Both brief as well as longer-term cross-modal exposure can trigger sensory recalibration, but the mechanisms underlying short-term and long-term recalibration appear to be distinct. The authors conclude by reviewing the evidence for the neural mechanisms of such cross-modal learning, which suggest that this learning takes place through the modification of both the cortical and subcortical pathways.

The final chapter provides a clinical perspective on multisensory influences on auditory processing. In Chap. 12, Baum Miller and Wallace review how fundamental changes in both auditory and multisensory processing impact perception in autism spectrum disorder. The authors dive into the behavioral and neural correlates of altered sensory processing and examine instances of both enhanced and diminished sensory function and perception in autism spectrum disorder compared with typical development. Furthermore, they propose that differences in the ability to integrate information across the different senses may link sensory abnormalities with the more canonical autism symptoms (e.g., impairment in social communication). If sensory and multisensory functions form the scaffold on which higher order abilities are built, the authors argue that treatment strategies that target strengthening sensory representations may prove useful in improving social communication.

1.3 Outlook

The perceptual world is not constructed on a strict sense-by-sense basis but rather is experienced as a coherent and integrated multisensory gestalt. Despite the self-evident perspective of the multisensory world, the neuroscience community has been slow to acknowledge the importance of multisensory processing and, consequently, delve into its neural bases. An example of this somewhat biased view comes directly from auditory studies. Given that humans can rely on auditory features (e.g., spatial cues, pitch) to help segregate sounds in a complex acoustical scene, is there a need to study the visual impact on solving this cocktail party problem? It is true that the brain can derive an amazing wealth of information about the perceptual environment using only a single sense. However, integrating information across the different senses often leads to striking improvements in human performance and perception (Calvert et al. 2004; Murray and Wallace 2011). More importantly, sensory integration is the natural *modus operandi* in the everyday environment in which humans live and operate.

With the growing interest and emphasis in multisensory systems, many neurophysiological studies have now sought to describe the brain circuits and encoding features associated with multisensory processing. Much of this work has relied on using animal models, focusing on describing the anatomical convergence that provides the neural substrate for multisensory interactions and then on detailing the neuronal operations carried out by neurons and circuits on multisensory stimulation. However, many of the multisensory encoding principles derived to date have come from work carried out in anesthetized animals, using paradigms in which stimulus characteristics are highly constrained (a necessary prerequisite for beginning to better understand multisensory processes; for reviews, see Stein and Meredith 1993; Stein 2012). However, the field needs to transition to studies in awake and behaving animals, with an emphasis on more naturalistic paradigms. Complementing these animal model studies should be human-imaging studies using similar, if not identical, paradigms, with the goal of bridging across levels of analysis. These human studies should take advantage of the host of approaches currently available, including magnetic resonance imaging and electrocorticography, as well as electro- and magnetoencephalography. Furthermore, building off of the wealth of correlative data that have been gathered, studies need to move more in the direction of causation and employ approaches such as transcranial magnetic stimulation and transcranial direct/alternate current stimulation to activate/deactivate brain regions during task performance. Again, these human studies can and should be complemented by animal model studies that make use of new technologies such as chemo- and optogenetics that represent powerful tools to dissect functional circuits.

As with other sensory systems, a new frontier in multisensory research needs to be in the context of active sensing where there is an acknowledgement that sensation and perception in naturalistic settings are a product of an active interplay between the sensory and motor systems. Take, for example, our eye and head movements, which represent powerful filters that decide what aspects of the multisensory world are sampled from at any given moment. Innovative experimental paradigms need to be established so that the question can be answered: how do observers actively sample the environment through movements of the eyes, head, and body to optimize the information they gather from their multisensory surroundings?

Finally, there is another area of research that has not yet been addressed adequately in the field of multisensory research: How does one build a multisensory environment to optimize human performance? The multisensory scene can be constructed de novo (in a virtual reality setting) or realized by injecting additional sensory information to the natural surrounding (in an augmented reality setting). Consumer electronics have progressed to a point that the differentiating factor for the ultimate user's experience might rest on a better multisensory experience. The ergonomics associated with audiovisual (or other multisensory combinations) experiences to improve human-computer interaction capabilities will be fueled by the needs of the consumers and may represent the next frontier of multisensory behavioral research.

The chapters in this volume focus on the neural circuits related to multisensory integration along the auditory pathway, from the brainstem to the prefrontal cortex as well as the perceptual benefits of leveraging other senses for communication in the complex auditory environment of everyday life. It is hoped that the readers will find this overview of multisensory influences an important contribution to the overall understanding of hearing science and, perhaps more importantly, as an inspiration for new research directions that will continue to improve the understanding of how the behavioral and perceptual representations of the multisensory world within which humans live are assembled.

Compliance with Ethics Requirements Adrian K. C. Lee declares that he has no conflict of interest.

Mark T. Wallace declares that he has no conflict of interest.

References

Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processes*. Cambridge: The MIT Press.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975–979.

Murray, M. M., & Wallace, M. T. (Eds.). (2011). *The neural bases of multisensory processes*. Boca Raton: CRC Press.

Stein, B. E. (Ed.). (2012). *The new handbook of multisensory processing*. Cambridge: The MIT Press.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge: The MIT Press.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., & WU-Minn HCP Consortium. (2013). The WU-minn human connectome project: An overview. *NeuroImage, 80*, 62–79.