Multimodal interactions: visual-auditory

Imagine that you are watching a game of tennis on television and someone accidentally mutes the sound. You will probably notice that following the game becomes harder, not just because the narration is no longer available, but also because perceiving the timing of the impact of the ball on the ground and racket will be harder. Now, imagine that instead of watching a game, you are playing tennis yourself. Plugging the ears would strongly interfere with your ability to play because now not only you cannot perceive the timing, speed, and location of the ball as accurately, but also you cannot coordinate your actions accordingly either. Humans are often exposed to visual and auditory information that arises from objects and events in their environment, and the nervous system has evolved and acquires ways of utilizing these two correlated sources of information for achieving a more accurate and reliable perception and action in the environment.

AUDITORY-VISUAL INTERACTIONS IN PERCEPTION

Although we are almost never consciously aware of it, the interplay between auditory and visual modalities is always operating in daily life. The reason we are not consciously aware of the interactions between vision and hearing is that under normal

circumstances, the sights and sounds that correspond to the same object convey consistent information, for example, the sound of the tennis ball hitting the ground and the image of it both agree in the time and location of impact (as well as other attributes, such as speed, weight etc.). Therefore, our overall estimate of the time and location of impact is a unified and coherent one. One may ask, how could the two estimates be anything but consistent if they arise from the same object. The estimates in each sensory modality are always corrupted by noise, noise in the environment (e.g., fog affecting the rays of light, or clutter affecting the sound waves) and noise in the brain (the firing of the neurons is noisy). Therefore, even the same exact stimulus can elicit different neural responses at different times. Therefore, the same tennis ball hitting the same location x on the ground time after time may be heard at location x in one time and at location x+5 cm at another time, and likewise for visual perception. Therefore, even stimuli that arise from the same event can be slightly inconsistent in the sensory estimates they invoke in the nervous system. The reason we are not aware of such inconsistencies is that the nervous system fuses the signals into one unified estimate by combining the estimates according to their respective reliability. For example, if the tennis court is well-lit and the observer has normal visual acuity, then visual estimate of location is likely more reliable than the auditory estimate as the auditory spatial resolution is generally not as good as vision, and

therefore the overall estimate of location will be largely biased towards the visual estimate while also influenced by the auditory estimate. On the other hand, if it is night time, and the court is not well-lit or the observer has bad eye sight, then the visual estimate may be less reliable than the auditory estimate and then overall estimate of location may be determined primarily by the auditory information. The same principle applies to the timing of the bounce. Under normal circumstances the auditory estimates of time are generally more reliable than visual estimates, and therefore dominate the overall estimate of time, but if there is a lot of noise in the background and the auditory estimate is not reliable, then the perception of time may be primarily determined by vision.

AUDITORY-VISUAL ILLUSIONS

Much of the knowledge of crossmodal interactions has been obtained through experiments that induce a conflict between two modalities and probe observer's perception. For example, it has been found that if sound is presented at a location that is moderately different from the location of a visual stimulus, it is often perceived to be originating from the same location as the visual stimulus. This effect is known as the ventriloquism effect, and is the same effect that ventriloquists have exploited for centuries for their puppet shows. It is also the same effect that we all experience every time we watch TV or a movie at the movie theater, where the voice of the actors is

perceived to originate from the same location as the image of the actors on the screen, as opposed to the fixed location of the speakers. This occurs because the visual estimates of location are typically more accurate than the auditory estimates of location, and therefore the overall percept of location is largely determined by vision. Conversely, perception of time, wherein auditory estimates are typically more accurate, is dominated by hearing. One example of this is perception of number of pulsations (which largely involves temporal processing). If a single flash of light is accompanied by two or more beeps, observers often perceive multiple flashes as opposed to a single flash. In this case, sound dominates the perception as it is generally more accurate in this task. This effect is known as sound-induced flash illusion. Another intriguing demonstration of auditory-visual interactions is in the domain of speech perception, and it is known as McGurk effect. If the video of an individual articulating the syllable /ga/ is played synchronously with the sound of an individual saying the syllable /ba/, the syllable /ba/ is often perceived as /da/. This reveals the strong auditory-visual interactions that take place during speech perception, which appears as a purely auditory task.

AUDITORY-VISUAL INTERACTIONS IN MEMORY AND LEARNING

If auditory-visual interactions are so ubiquitous in perception, can they also play a role in learning and memory? Indeed, recent studies have shown that auditory-visual

interactions can facilitate memory and learning. For example, it was found that observers were able to recognize the image of objects that were previously presented accompanied by their corresponding sound better than the image of objects that were initially shown only visually. Even more surprisingly, learning of a visual task was recently shown to be facilitated by congruent sounds during training. Observers learned to detect visual motion much faster and much better when during training the visual motion was accompanied by auditory motion, despite the fact that sound was absent during testing. Therefore, sound appears to help with the encoding and/or retrieval of the visual information. Conversely, visual information has been shown to facilitate auditory learning. Training with voices that are paired with video clips of talking faces is more effective in inducing learning of voices than training with voices alone, even when videos are absent during testin.

Ladan Shams

Cross-References: Bayesian approach to perception, Binding problem, Inference and perception, Multimodal interactions, Perceptual learning, Sensory coding.

Further Readings

Welch, R. B., Warren, D. H. 1980. Immediate perceptual response to intersensory discrepancy. Psychol Bull. 88, 638-667. A early article advancing the idea of that the modality with higher reliability dominates in any given task.

Stein, B. E., Meredith, M. A. 1993. The merging of the senses. Cambridge, Mass: MIT Press. An early book on phenomenology, brain mechanisms, and principles of multisensory integration.

Knill, D. C., Pouget, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in Neurosciences. 27, 712-719. A review of probabilistic coding in the sensory systems, and its implications for optimal sensory processing and cue combination, and the possible underlying neural mechanisms.

Ernst, M. O., Bülthoff, H. H.. 2004. Merging the senses into a robust percept. Trends in Cognitive Sciences. 8, 162-169. A review of the crossmodal cue combination and the computational principles governing it.

Shimojo, S., Shams, L., 2001. Sensory modalities are not separate modalities: plasticity and interactions. Current Opinion in Neurobiology. 11, 505-509. A review of crossmodal interactions with a particular focus on auditory-visual interactions, challenging the traditional modular view of perception.