Sensory amplification through crossmodal stimulation

Tonja Machulla
University of Stuttgart
Stuttgart, Germany
tonja.machulla@vis.uni-stuttgart.de

Marc O. Ernst
Ulm University
Ulm, Germany
marc.ernst@uni-ulm.de

Lewis Chuang
MPI for Biological Cybernetics
Tuebingen, Germany
lewis.chuang@kyb.mpg.de

Albrecht Schmidt
University of Stuttgart
Stuttgart, Germany
albrecht.schmidt@vis.uni-stuttgart.de

Francisco Kiss
University of Stuttgart
Stuttgart, Germany
francisco.kiss@vis.uni-stuttgart.de

Proceedings of the CHI 2017 Workshop on Amplification and Augmentation of Human Perception, May 07, 2017, Denver, CO, USA. Copyright is held by the owner/author(s).

Abstract

This paper discusses how stimuli processing in one sensory modality (e.g., vision) can be amplified by co-stimulation of other modalities (e.g., auditory/tactile). The focus lies on explanatory accounts and established findings from experimental psychology and neuroscience. We review models of sensory amplification as well as the physical constraints on amplification. From this, we derive guiding principles for the design innovation of multi-sensory displays for crossmodal amplification.

Author Keywords

multisensory displays; sensory augmentation; Race Model; multisensory integration; human information processing

ACM Classification Keywords

H.1.2 [Models and principles]: User/Machine Systems

Introduction

There are various approaches to amplify human perception. We are most familiar with the use of technology aids, such as a megaphone, that amplify a weak physical signal before it reaches the sensory organ. Amplifying the physical signal allows it to exceed the human sensory threshold, making the signal more easily perceived. Nonetheless, the strength of a sensory signal can continue to be modulated even after it reaches the sensory organ. Physical events that produce

crossmodal stimulation results from environmental events that are perceived through more than one sensory modality

multiple sensory signals (e.g., audio and visual speech) are typically perceived more vividly. In other words, one sensory signal can be crossmodally amplified or modulated by another sensory signal.

To exploit the benefits of crossmodal amplifications, this contribution reviews established computational, behavioral, and neuronal principles from the fields of psychology and neuroscience. In these fields, the common terms for crossmodal sensory amplification are "multisensory interaction" and "multisensory integration". In the following sections, we first describe the effects of crossmodal amplification, introduce two explanatory frameworks for this, list the physical constraints to crossmodal amplification, and account for how the brain resolves physical discrepancies between crossmodal signals.

Amplification through redundant stimulation

Humans, like most higher animals, perceive the world via multiple sensory systems. This has many obvious advantages. First, it increases the range of information available to us. Our senses are complements and compensate for the "blind spots" of each other; some aspects of the environment are uniquely coded through one sense only [6]. For instance, the perception of color is unique to the visual sense.

In addition, multiple sensory systems allow for redundant coding—many environmental properties can be perceived by more than one sense. Examples include the location, shape, and texture of a spatial object or the duration, intensity, rate, or rhythm of a temporal event.

Redundant coding results in many behavioral benefits(for an overview see [4]). For example, displays for parking present both auditory and visual pulsed warnings to communicate proximity to a collision target. In the laboratory, such redundant signals have been shown to result in faster reaction times, higher perceptual sensitivity, as well as faster and more accurate spatial localization. These performance improvements are accompanied by neurophysiological changes such as larger peak responses to multisensory stimuli in EEG and single-cell recordings.

Models of crossmodal amplification

How does crossmodal amplification occur? Two established models can account for this behavioral phenomenon.

The Race Model predicts the gain in reaction time to a stimulus that can be achieved by adding a secondary redundant stimulus [14]. The gain is explained in terms of statistical probability summation: this means that the likelihood of detecting a bimodal event is larger because more data about the presence of the event is available. For instance, in the case of an audiovisual stimulus the auditory and the visual signal "race" against one another for access to the motor response system; the winner triggers the motor response. On average, the runtimes of the winning signals will always be shorter than the runtime of either racer. Interestingly, behavioral studies have demonstrated that the upper limit of predicted runtime improvement is sometimes exceeded, i.e., humans react even faster than would be expected from statistical probability summation alone. This indicates that concurrent signals can interact to elicit "amplified" joint behavior that is superior to the better response of two independent signals.

More evidence for crossmodal amplification of sensory signals comes from the observation of the response behavior of single multisensory neurons [21]. These often show a multisensory response enhancement in the form of an increased spike-count rate, i.e., the neuron responds more vigorously to crossmodal stimulation than to stimulation of

Multisensory neurons respond to sensory input from more than one modality. They are found in various cortical areas as well as the superior colliculus, a mid-brain structure that controls orientation of attention and gaze to a location in space.

Spikes are electric pulses generated by a neuron. Information about external events can be encoded as the temporal pattern of successive spikes.

Superadditivity is a nonlinear neural computation, where the output after crossmodal stimulation exceeds the sum of the outputs after stimulation of the component modalities. [Stein & Stanford]

either sense alone. This neuronal response can be *superadditive*, i.e., the neuron combines sensory information nonlinearly. This form of sensory amplification is proportionally strongest for very weak stimuli—as stimuli become easier to perceive, the amplification gain decreases. This principle of i*nverse effectiveness* is ecologically plausible. Enhancing the sensory signals elicited by an event that is very loud or bright and, therefore, easily detected brings little additional advantage.

Applicational relevance. If the detection, identification, or localization of an environmental event is to be improved in terms of accuracy and speed, the sensory signal can be amplified by adding concurrent presentation of redundant information over different sensory channels, e.g. via head-worn cognition-aware computing devices. This might be particularly useful in situations where a) the to-be-detected signals are weak, such as an approaching car in a foggy environment, and b) it is either technically more economical or less distracting to co-stimulate via another modality rather than to amplify the physical signal. Quantitative predictions of response time improvements from the Race Model can be useful to the system designer in situations where the benefit of faster response has to be balanced against possible costs of the additional stimulation, such as increases in perceptual and mental workload.

Constraints on multimodal interaction

Several physical factors constrain whether and how the information presented to different senses will interact [1]. The most prominent of these is temporal co-occurrence or simultaneity. Redundant signals have to reach the sensory organs within a certain time of each other. Otherwise, they will not be perceived as belonging to the same event (e.g., thunder and lightning of a distant storm) and render no processing advantage for event perception. The time window within which asynchrony between signals is tolerated approximates 200 ms for signals originating from short unitary events like single light flashes and short sound bursts [20]. This time window can be extended for signals that demonstrate noticeable cross-correlation over time, such as is the case of a video stream of a news speaker and a lagging audio stream [16].

Some crossmodal interaction can occur, even if the stimuli are presented one after another with an interval just larger than the time window of asynchrony tolerance, albeit with a negative gain. This should pose a concern for multimodal displays with variable temporal synchrony. There are several well-documented processing deficits that result from asynchronous presentation. The *psychological refractory period* is a phenomenon where the processing of the first stimulus consume so much perceptual resources that the response speed and accuracy to the second stimulus is significantly decreased [17]. An *attentional blink* occurs when the processing of the second stimulus is degraded to such a degree that the stimulus fails to be consciously perceived [18]. In addition, the first stimulus can act as a cue to a specific region in space and initiate an attentional shift to that region—processing of any further stimulus within this region is enhanced for up to 300 ms but afterwards decreases below the level of processing of regions that were not cued. This *inhibition-of-return* of attention to previously attended regions is explained to be an unconscious neural strategy that maximizes the overall area that is explored [19].

The co-location of signals is another constraint on multimodal interaction [1]. Generally, performance improves if two signals originate from the same location in space. However, there will often be an enhancement even if signals are spatially separated. For instance, detection performance for visual warning improves if it is accompanied by a sound presented over headphones. Unless binaural recording is used, the sound source is perceptually located inside the head, while the source of the visual signal is located outside of the body.

Applicational relevance. If a signal is to be amplified by adding crossmodal redundant information, the timing of the additional signal is the most crucial factor. Thus, the temporal calibration of the system providing the additional stimulation should be optimized. System lag can lead to perceptual and behavioral consequences opposite to the desired effect, namely a reduction in performance below the standard level.

Discrepancy between the senses

A topic that is related to the optimal window of crossmodal interaction is how the human cognitive system deals with discrepancies between redundant signals. Such discrepancies are very common and result from two main sources [5]. The first is the random noise which is added to the signal during physical and physiological transmission. If you were to repeatedly point at a sound source with closed eyes, each instance would result in a slightly different location. These transient discrepancies are corrected for on-the-fly, often by adjusting one sensory estimate to the other. This results in a large number of crossmodal illusions, the most famous being the ventriloquist effect [3]. Here, the voice of the puppeteer seems to emanate from the mouth of the dummy, or in the terminology of experimental psychologists, the perceived location of the auditory stimulus is captured by the location of the visual stimulus. We experience this when we watch TV—the location of sounds appear to be correctly matched to events on the screen even if it is produced by a single neighboring speaker.

Capture describes the phenomenon when the multisensory percept is dominated by one sense, e.g., in visual capture of audition, sound is perceived as coming from the same location as the visual stimulus even if the two are spatially separated.

Other capture-based illusions include the temporal ventriloquism effect, where a short visual event is shifted in time towards an auditory sound (auditory capture of vision, e.g., [15]); the McGurk Effect, where visual lip movements influence the perceived speech sound (visual capture of sound [13]); or partial out-of-body illusions such as the Rubber Hand illusion, where the felt position of a hand is shifted towards the seen position of a (fake or Avatar) hand (visual capture of touch and proprioception, e.g., [8]).

Which factors determine whether one sense will dominate the other or vice versa? Under natural, non-degraded environmental conditions, we can apply a rule-of-thumb: the visual sense is better—i.e., less noisy—at resolving spatial detail. Therefore, it can be expected to dominate the auditory and the tactile senses in localization tasks such as pointing and grasping. In contrast, the auditory sense is better at resolving temporal detail and will dominate judgments of rhythmicity, judgments of when something occurred in time and of event duration.

Maximum likelihood estimation models can be applied to arrive at more precise descriptions of how discrepancies between signals are resolved. A large number of behavioral studies have shown that two discrepant sensory estimates (e.g., of the physical location of an event) are adjusted towards each other, with the amount of adjustment determined by the noise associated with each signal (for an overview see [1]). More specifically, the combined estimate can be modelled as the weighted sum of the individual estimates, with weights proportional to the signals' variances, the variance being a measure of noise ([6]; see Figure 1). This model also predicts that if signals in one sense are substantially less noisy, the combined estimate will be dominated by this sense.

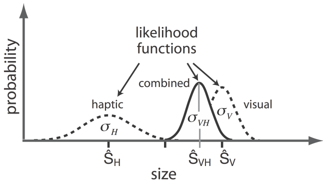

Figure 1: Exemplary likelihood functions of visually and haptically estimated size of an object. Signal noise is represented by the width of the distributions. The combined visual-haptic estimate is closer to the estimate with lower associated noise.

A second source of discrepancy between the senses results from systematic changes in the relationship between sensory signals. For example, when we assume new corrective lenses, we may temporarily experience distortions in the peripheral visual field such that the seen and the felt location of objects no longer coincide perfectly. Nonetheless, we are able to correct for these non-transient discrepancies between the senses over time by sensory recalibration—the signals from the altered sensory environment are realigned with regards to the other senses. This process is fairly fast. Corrections in grasping and throwing behavior in response to prismatically shifted vision set in within a few minutes and are almost perfect within a few days [9], even in extreme cases such as up-down inversions.

Similarly, temporal recalibration sets in if redundant signals arrive with a fix lag, such as in a video of a talking person with a lagging audio stream [7, 12]. For small discrepancies of up to 200–300 ms, the perceptual temporal discrepancy is reduced by 10–30% within one minute—that is, the video and the audio stream appear more and more in sync. Whether such discrepancies can be fully compensated, remains unclear since conducting experiments where one sense is delayed over longer periods of time (e.g., days) has yet to be carried out in the laboratory. Wearable systems present the research potential to systematically evaluate this in everyday life, outside the laboratory.

Applicational relevance. Presenting a secondary, redundant stimulus at a discrepancy—e.g., over head-worn computing devices—may be used to induce immediate biases in perception and performance. However, substantial shifts in perception can only be expected if the signal associated with the secondary stimulus is sufficiently reliable (i.e., not noisy) compared to the signal that is to be influenced. A good example of this has been recently provided by haptic retargeting, where the perceived proprioceptive location of a physical object is shifted in space by providing visually distorted feedback in VR [2]. Similarly, VR users in tight constrained spaces can be led to believe that they are walking on endless straight paths, by manipulating the visual feedback to bias them into walking in curves [10].

However, care is to be taken with exposure to prolonged, unidirectional incongruencies in the mapping between visual cues in VR environments and proprioceptive/vestibular perception. Incongruency can occur even when isometric mappings are applied between physical and VR space [22, 11]. Such discrepancies may cause spatial recalibration between seen and felt locations, which might temporarily carry over into behavior outside of the VR environment and thus pose a safety hazard.

Summary and implications for systems design

The present contribution provides an overview of perceptual and neurophysiological principles of how sensory perception can be modulated crossmodally. From this, we can derive several implications for the design of novel systems dedicated to the amplification of human sensory perception, as summarized in the following: 1. Crossmodal amplification shows the largest gain for weak signals. 2. For two signals to interact, they have to be presented in close temporal proximity and should be strongly correlated. 3. Spatial congruence of redundant signals is helpful but not of absolute necessity to obtain performance gains. 4. Often, in the case of discrepancy between redundant signals, one sense dominates the perceptual interpretation. Depending on the use case, conflicting inputs to the senses should either be avoided or can be exploited.

Acknowledgements

T.M., F.K., and A.S. were supported by the European Research Council, under grant number 683008 (Amplify). A.S., M.O.E., and L.C. were supported by the German Research Foundation (DFG) within the SFB/Transregio 161.

References

[1] David Alais, Fiona N Newell, and Pascal Mamassian. 2010. Multisensory processing in review: from physiology to behaviour. *Seeing and perceiving* 23, 1 (2010), 3–38.

[2] Mahdi Azmandian, Mark Hancock, Hrvoje Benko, Eyal Ofek, and Andrew D Wilson. 2016. Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 1968–1979.

[3] Bjoern Bonath, Toemme Noesselt, Antigona Martinez, Jyoti Mishra, Kati Schwiecker, Hans-Jochen Heinze, and Steven A Hillyard. 2007. Neural basis of the ventriloquist illusion. *Current Biology* 17, 19 (2007), 1697–1703.

[4] Gemma Calvert, Charles Spence, and Barry E Stein. 2004. *The handbook of multisensory processes*. MIT press.

[5] Beatrice De Gelder and Paul Bertelson. 2003. Multisensory integration, perception and ecological validity. *Trends in cognitive sciences* 7, 10 (2003), 460–467.

[6] Marc O Ernst and Heinrich H Bülthoff. 2004. Merging the senses into a robust percept. *Trends in cognitive sciences* 8, 4 (2004), 162–169.

[7] Waka Fujisaki, Shinsuke Shimojo, Makio Kashino, and Shin'ya Nishida. 2004. Recalibration of audiovisual simultaneity. *Nature neuroscience* 7, 7 (2004), 773–778.

[8] Wijnand A IJsselsteijn, Yvonne A W de Kort, and Antal Haans. 2006. Is this my hand I see before me? The rubber hand illusion in reality, virtual reality, and mixed reality. *Presence: Teleoperators and Virtual Environments* 15, 4 (2006), 455–464.

[9] Alan S Kornheiser. 1976. Adaptation to laterally displaced vision: A review. *Psychological bulletin* 83, 5 (1976), 783.

[10] E Langbehn, P Lubos, G Bruder, and F Steinicke. 2017. Bending the Curve: Sensitivity to Bending of Curved Paths. *IEEE transactions on visualization and computer graphics* (2017).

[11] Jack M Loomis and Joshua M Knapp. 2003. Visual perception of egocentric distance in real and virtual environments. *Virtual and adaptive environments* 11 (2003), 21–46.

[12] Tonja-Katrin Machulla, Massimiliano Di Luca, Eva Froehlich, and Marc O Ernst. 2012. Multisensory simultaneity recalibration: storage of the aftereffect in the absence of counterevidence. *Experimental brain research* 217, 1 (2012), 89–97.

[13] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. (1976).

[14] Jeff Miller. 1982. Divided attention: Evidence for coactivation with redundant signals. *Cognitive psychology* 14, 2 (1982), 247–279.

[15] Sharon Morein-Zamir, Salvador Soto-Faraco, and Alan Kingstone. 2003. Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research* 17, 1 (2003), 154–163.

[16] Cesare V Parise, Charles Spence, and Marc O Ernst. 2012. When correlation implies causation in multisensory integration. *Current Biology* 22, 1 (2012), 46–49.

[17] Harold Pashler. 1994. Dual-task interference in simple tasks: data and theory. *Psychological bulletin* 116, 2 (1994), 220.

[18] Salvador Soto-Faraco, Charles Spence, Katherine Fairbank, Alan Kingstone, Anne P Hillstrom, and Kimron Shapiro. 2002. A crossmodal attentional blink between vision and touch. *Psychonomic Bulletin & Review* 9, 4 (2002), 731–738.

[19] Charles Spence and Jon Driver. 1998. Auditory and audiovisual inhibition of return. *Attention, Perception, & Psychophysics* 60, 1 (1998), 125–139.

[20] Charles Spence and Sarah Squire. 2003. Multisensory integration: maintaining the perception of synchrony. *Current Biology* 13, 13 (2003), R519–R521.

[21] Barry E Stein and Terrence R Stanford. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* 9, 4 (2008), 255–266.

[22] Frank Steinicke, Gerd Bruder, Jason Jerald, Harald Frenz, and Markus Lappe. 2010. Estimation of detection thresholds for redirected walking techniques. *IEEE Transactions on Visualization and Computer Graphics* 16, 1 (2010), 17–27.