USING AUDITORY AUGMENTED REALITY TO UNDERSTAND VISUAL SCENES

SCOTT STONE
Bachelor of Science, University of Lethbridge, 2015

A Thesis
Submitted to the School of Graduate Studies
of the University of Lethbridge
in Partial Fulfilment of the
Requirements for the Degree

MASTER OF SCIENCE

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

USING AUDITORY AUGMENTED REALITY TO UNDERSTAND VISUAL SCENES

SCOTT STONE

Date of Defence: August 22, 2017

Dr. Matthew Tata Co-supervisor	Associate Professor	Ph.D.
Dr. Claudia Gonzalez Co-supervisor	Associate Professor	Ph.D.
Dr. Jon Doan Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Howard Cheng Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. Robbin Gibb Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Dedication

To my parents, and everyone who has supported me along the way.

Abstract

Locating objects in space is typically thought of as a visual task. However, not everyone has access to visual information, such as the blind. The purpose of this thesis was to investigate whether it was possible to convert visual events into spatial auditory cues. A neuromorphic retina was used to collect visual events and custom software was written to augment auditory localization cues into the scene. The neuromorphic retina is engineered to encode data similar to how the dorsal visual pathway does. The dorsal visual pathway is associated with fast nonredundant information encoding and is thought to drive attentional shifting, especially in the presence of visual transients. The intent was to create a device capable of using these visual onsets and transients to generate spatial auditory cues. To achieve this, the device uses the core principles driving auditory localization, with a focus on the interaural time and level difference cues. These cues are thought to be responsible for encoding azimuthal location in space. Results demonstrate the usefulness of such a device, but personalization will probably improve the effectiveness of the cues generated.

In summary, I have created a device that converts purely visual events into useful auditory cues for localization, thereby granting perception of stimuli that may have been inaccessible to the user.

Acknowledgements

First and foremost, I would like to thank my supervisors Dr. Matthew Tata and Dr. Claudia Gonzalez. Thank you for believing in me as a student and giving me the opportunity to explore my passion. Your advice and guidance has been invaluable and has helped shape me as a person. I would also like to thank Drs. Jon Doan and Howard Cheng for agreeing to be on my committee and humoring my (usually insane) ideas for projects.

Thank you to everyone in the Brain in Action and Tata laboratories for putting up with my inane chatter. I'd like to thank Jason, Lara, and Nicole for being a second set of ears to bounce my ideas off as well as being good friends. Most of my computer science and coding knowledge I owe to Marko and Lukas.

Obviously, I couldn't have done this without my family. My parents, Dawn and Brian, to whom I have dedicated this thesis, have been the inspiration that has driven me to be curious and think critically about everything. You are the example of the kind of person I'd like to be. Kayla, thank you for being a kind person to me (even more so when you moved to the Netherlands) and giving me valuable life advice.

And finally, thank you, whoever you are, for taking the time to read this thesis.

Table of contents

List of Figures

CHAPTER 1. How visual and auditory systems engage attention

Preface

Sensory systems in humans are capable of processing vast amounts of complex information in real-time. Sensory events typically come in the form of transients; changes in brightness, loudness, or sudden motion. While some events are multi-modal, such as a cell phone ringing and flashing, others may be monomodal. That is, some events have exclusively auditory or visual features. Not everyone has access to both modalities. If someone is blind, they cannot use visual information to locate the phone and must instead rely solely on auditory cues. One possible way to solve this problem is by converting visual events into auditory events. Effectively, this would remove the perceptual barriers blind individuals face with the loss of vision. Currently, there are no widely-available prostheses for the blind that do this, much less any evidence that such a device is even possible to create. The goal of this thesis was to develop a method of augmenting spatial cues into the auditory scene using purely visual information, with the end goal being to implement it into a device to be used as an attentional prosthesis. This device would allow the blind to glean crucial visual information through auditory cues.

This thesis will be organized as follows: chapter 1 will focus on the basic principles that drive visual and auditory perception that are critical to understanding how we might create valid auditory localization cues from visual events. Chapter 2 will describe how these localization principles are implemented computationally using neuromorphic retina hardware with a focus on the interaural time and level difference cues to calculate azimuthal location. Chapter 3 demonstrates the efficacy of the device and algorithms in two key populations: the regularly sighted (i.e. controls) and congenitally blind individuals. Chapter 4 discusses the overall performance and design of the device when compared to similar technologies as well as the possible improvements and future directions of the project.

Auditory and visual system function

Localizing objects in space can be achieved using auditory information, visual information, or a combination of both. For example, when searching for a ringing phone, visual and auditory information can be used independently or integrated to increase the chances of finding the phone. Greater accuracy can be achieved when integrating information from the two modalities, such as when the phone is vibrating and moving slightly while ringing. However, not everyone has access to both modalities. One population is known as the blind, those who have reduced or no access to visual information. Therefore, it is important for the blind to have adequate auditory localization skills to make up for the loss of vision. Given that not all stimuli contain both visual and auditory components, the blind are simply unable to give attention to purely visual stimuli. In the example of a completely blind user interacting with a computer, the information normally visible on the screen must be translated into audio descriptions of what is happening on the screen. An unfortunate side effect of this is all spatial information is lost in the translation from vision to audition. To maintain visuospatial characteristics about the scene, the location of the visual stimulus on the retina must be known. Therefore, a system designed to improve the ability of a blind person to locate objects in space should rely on spatial information. We developed a system to inform the user of visual events via auditory cues, allowing someone with no access to visual information to still be able to perceive these events. To extract only the important salient cues from the scene, basic principles of visual system function were used so as not to include redundant or irrelevant information in the auditory scene.

Visual system function

Generating auditory cues based on visual events requires an understanding of how the visual system collects and processes information. Much of the visual system is organized retinotopically. Essentially, visual brain areas are organized similarly to the retina. The retina receives inputs via photoreceptors, which are connected to bipolar cells that provide excitatory signals to ganglion cells (Kolb, 1991; Kolb, Linberg, & Fisher, 1992). Ganglion cells are spread densely around the retina, with each individual cell encoding a particular part of the retina. In the geniculostriate pathway, retinal ganglion cells pass information into the lateral geniculate nucleus (LGN) of the thalamus via the optic nerve (Meissirel, Wikler, Chalupa, & Rakic, 1997). Magnocellular (M) and parvocellular (P) projections originate in the retina and project into separate M and P layers of the LGN (Meissirel et al., 1997; L. G. Ungerleider & Haxby, 1994). These layers provide efferent signals to the different areas of the cerebral cortex which are thought to carry out roles such as discriminating motion (M; (Meissirel et al., 1997; W. Merigan, Byrne, & Maunsell, 1991; W. H. Merigan & Eskin, 1986; W. H. Merigan & Maunsell, 1990)) and colour (P; (Gegenfurtner & Kiper, 2003; W. Merigan, Katz, & Maunsell, 1991; Snowden, 2002; L. G. Ungerleider & Haxby, 1994)). This functional segregation was initially characterized by L. Ungerleider and Mishkin (1982) , where it was suggested that there are two visual streams that play complementary roles to one another. These are known as the dorsal (where or vision for action) and ventral (what or vision for perception) visual streams. Because the focus of this thesis is on generating auditory cues from salient visual events, the ventral stream will not be discussed.

The dorsal visual stream is of great interest because it is thought to be responsible for processing spatial information such as the location and velocity of objects, as well as visual transients (Milner & Goodale, 2006). For a prosthetic attention system to provide useful information to the user in real-time, key operating principles need to be borrowed from dorsal visual stream function. Deriving this information is important if the goal is to locate an object of interest. The dorsal stream is thought to be important in attention switching. In fact, lesions to posterior parietal cortex (i.e. along the dorsal pathway) are associated with an inability to disengage from a stimulus or to shift attention to a new source (Posner & Petersen, 1990; Posner, Walker, Friedrich, & Rafal, 1984). The parietal lobes are central to the dorsal visual stream, which Posner argues are critical to shifting attention behaviours.

Because the primary input fibres to the dorsal stream are magnocellular (and thus myelinated), two operating characteristics are high temporal and low spatial resolution. These characteristics allow the dorsal stream to be sensitive to visual transients. This means representations in the dorsal stream will be passed to cortical areas and processed very quickly, but low amounts of detail are available. These characteristics are key for quickly noticing sudden visual changes or onsets. Important areas along projections of the dorsal stream include visual brain areas such as V1, V2, and the middle temporal (MT) area. The dorsal stream processes information very quickly, and as such, is thought to drive the attention orienting response to transient stimuli. Using the information processed by the dorsal visual stream to augment the auditory scene in real-time should promote fast attentional shifting, where a sensor approximating dorsal stream function to detect transients would be ideal for driving attention orienting responses.

Areas found along the dorsal visual stream perform distinct roles; area MT is of particular importance because it is thought to process visual motion (Born & Bradley, 2005; Felleman & Van Essen, 1991; Mikami, Newsome, & Wurtz, 1986; Newsome & Pare, 1988;

Nowlan & Sejnowski, 1995; Simoncelli & Heeger, 1998). Area MT receives inputs from many cortical areas including V1, V2, V3, V3A as well as subcortical areas such as the pulvinar and LGN (primarily koniocellular projections) (Figure 1 adapted from Born and Bradley (2005)). Born and Bradley (2005) argue that the most important input to MT originates from a magnocellular projection in layer 4B of V1. Given that inputs to MT can be as low as five synapses away from the photoreceptors (Born & Bradley, 2005), it is not surprising to find a retinotopic organization. Spatial information processed in area MT would be useful to generate auditory events because it already encodes visuospatial location. The proposed auditory augmented reality system would process motion information similar to area MT to generate auditory cues.

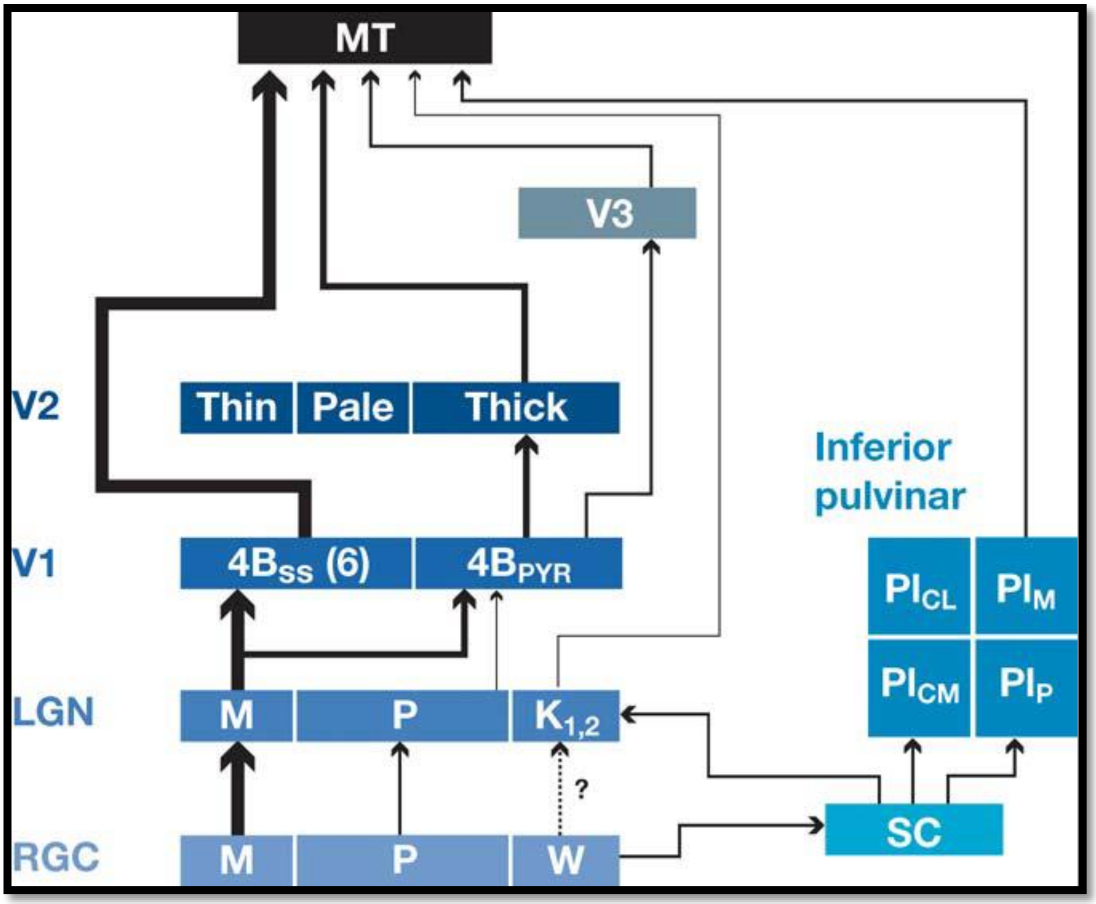

Figure 1. Adapted from Born and Bradley (2005). Inputs to area MT are shown. V1 (from layer 4B) are shown as thicker because it is thought that inputs from this area are denser in information than other inputs and is thus very important. Other cortical inputs include V2, V3 and V3A (not shown). Subcortical inputs include the pulvinar and koniocellular inputs originating from the LGN.

Overall, dorsal stream function is key to understanding how we might generate spatial auditory cues to drive attention. Such a system would be based on the information the dorsal stream produces. While spatial information tends to be sparse, temporal information is encoded much faster and less redundantly. These principles are what allow the dorsal stream to be sensitive to visual transients. The spatial information encoded in the visual events should be able to be converted into valid auditory cues to direct the user’s attention.

Horizontal and vertical auditory localization mechanisms

Because we are interested in converting visual events into spatial auditory cues, it is important to understand how the brain processes cues encoding position. Valid auditory cues are imperative to generate convincing attentional percepts. Typically, the eyes are thought to be more useful for locating objects in space. For example, searching for your cell phone requires visual scanning of the room. However, a ringing cell phone can be easier to locate because we now have access to an extra modality: audition. Auditory localization is possible in both the vertical and horizontal dimensions, though the mechanisms for encoding each are distinct. Horizontal spatial auditory signals encode location through a couple of cues: interaural time and level differences. Vertical localization is achieved through the pinnae transfer function: spectral cues in the form of a notch filter performed at the level of the pinna. Together, these cues create a percept of location in two dimensions. The mechanisms that encode both horizontal and vertical localization are thought to be performed in the dorsal brainstem, though not in the same areas. It is imperative we understand how these cues work on a computational level, as replicating these cues in software gives us the ability to render objects in virtual space to provide spatial cues for the listener.

Horizontal localization

The foundation of most horizontal localization research is largely based on work by Lord Rayleigh (1907). In this work, Rayleigh attempted to account for localization along the horizon using the interaural time (ITD) and level (ILD) difference cues at the ears. When a sound is presented from the side, the signal will enter one ear slightly before the other (ITD), as well as be shadowed by the head and thus quieter in the ear furthest from the sound (ILD). These cues, he argued, can be used to calculate the azimuth of a sound in space. Rayleigh argued that either cue is sufficient to localize in the horizontal plane, almost as if one of the cues was not necessary. To better understand why both cues are helpful for localization, a model of optimal frequencies was created for each cue. The amount of shadowing is dependent on the frequency of sound. Thus, these cues should be optimal at different frequencies. Sinusoidal signals with frequencies below 1000 hertz are physically several times the size of the head, and therefore useless for the ILD cue. Differences in phase (i.e. ITD), would therefore be more important at these frequencies. At higher frequencies, the level difference becomes more important because differences in amplitude can be detected. In an early attempt to demonstrate these cues, a pair of tuning forks tuned at slightly different frequencies was presented to participants (Rayleigh, 1907). Because the forks are slightly out of tune with each other, a steadily varying phase difference is produced. Participants reported the percept of a sound source moving back and forth in front of them. This was taken as good evidence that the ITD cue is capable of being used to compute a position in space.

It is not possible to completely dissociate interaural delay cues from each other. For example, it is not possible to have a signal that only has ITD information encoded, because it must have an amplitude. It is also not possible to have a signal with only ILD information encoded, because it must exist in time. Therefore, it is difficult to determine if these cues are calculated in different parts of the brain. Early anatomical studies of auditory pathways showed an early site of convergence from both signals in the superior olivary complex (SOC), which is comprised of the medial superior olive (MSO) and lateral superior olive (LSO). Studies involving lesions in the SOC in ferrets and cats have shown a massive decrease in ability to localize sounds in the horizontal plane (Kavanagh & Kelly, 1992; Masterton, Diamond, Harrison, & Beecher, 1967; Moore, Casseday, & Neff, 1974), suggesting that this area of the brain is playing a role in the calculation. The LSO is the most likely area where ILD information is processed.

Jeffress (1948) proposed a mechanism capable of calculating position based on an ITD signal. At its core was the idea of a coincidence detector; a neuron that would only fire if two signals activate it simultaneously (i.e. coincidentally; Figure 2.). Jeffress postulated that primary auditory fibres (i.e. shortly after the tympanic membrane) give rise to secondary fibres which project to ipsilateral and contralateral areas of the brain. Secondary fibres produce ladder-like tertiary fibres that act like a delay line. Delay lines are critical to Jeffress' model, as this is the primary way that phase differences can be calculated. As an example, if a sound source is one inch to the left of the midline, the sound will enter the left ear about a tenth of a millisecond earlier than the right ear. Thus, the primary, secondary, and tertiary fibres emanating from the left ear have a time advantage and can travel further down the ladder delay line. The neuron the signal settles on (i.e. the coincidence detector) depends on when the signal arrives from the right ear. In this example, the activated neuron would probably be closer to the beginning of the tertiary fibres from the contralateral projection (i.e. the right side) and therefore encoding the sound as being closer to the left side. While convincing evidence of delay lines has not been shown in humans, some work has shown that animals such as cats (Smith, Joris, & Yin, 1993) and barn owls (Carr & Konishi, 1990; Andrew Moiseff & Konishi, 1981) have physiological delay-line-like systems. Smith et al. (1993) injected horseradish peroxidase into the trapezoid body, which is thought to bilaterally innervate the MSO with inhibitory inputs. They found that the projections from the contralateral, but not the ipsilateral side show a delay line configuration which approaches rostrally, turns caudally and branches off much like Jeffress (1948) predicted. One caveat is that the delay line configuration was only found in the contralateral MSO. It is worth noting, however, that this configuration will still work because the side with the delay line on it will be the side ipsilateral to the sound source.

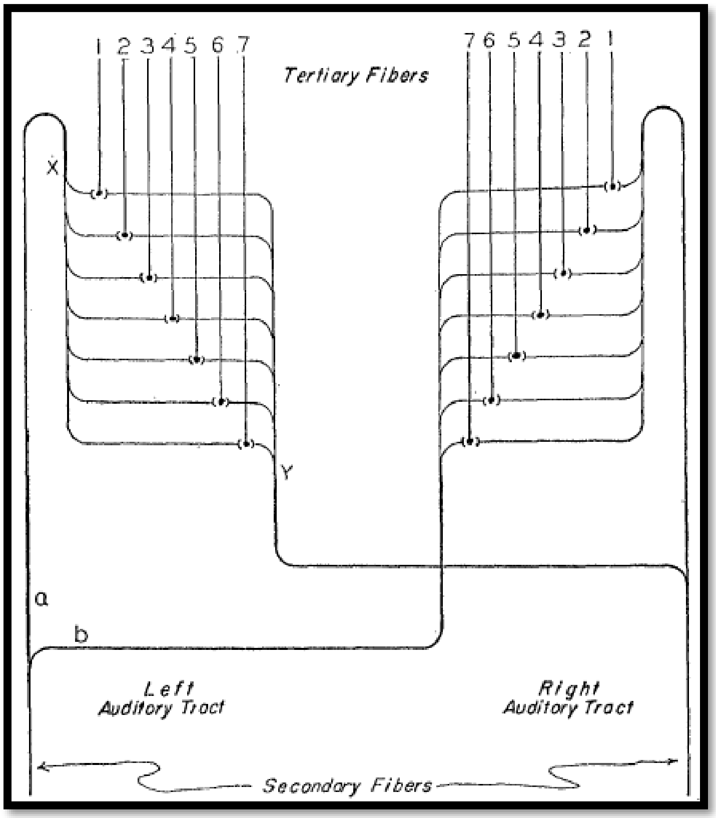

Figure 2. Adapted from Jeffress (1948). Jeffress' model demonstrating how secondary and tertiary fibers can excite coincidence detectors. It should be noted that this diagram is meant as an illustration only, and such a system would need a much larger number of secondary/tertiary fibres to properly encode space at a reasonable resolution.

The ILD cue is not processed in the same way as ITD. The LSO is thought to be the area where azimuth is calculated from the ILD cue (Masterton et al., 1967; Moore et al., 1974; Pecka, Brand, Behrend, & Grothe, 2008; Tollin, 2003). The LSO is thought to process information by passing information from the cochlea to ipsilateral nuclear angularis (NA) neurons in the brainnstem to provide excitatory efferents to the contralateral posterior ventral nucleus of the lateral lemniscus (VLVp) within the LSO (Mogdans & Knudsen, 1994; A Moiseff & Konishi, 1983). Each

VLVp, while simultaneous receiving excitatory input from the contralateral NA neurons, send inhibitory signals to each other. Because cochlear input drives NA activity, this mechanism allows for the calculation of the ILD signal. If one side is louder than the other, the VLVp activity contralateral to the sound source will reflect that. Put simply, this neural circuit is calculating proportionally how much louder one signal is than another. This type of cue should be relatively easy to model computationally when compared to the ITD cue. Each cue provides helpful for localization, so an auditory augmented reality system should include both for optimal performance. However, it should be noted that, like the ITD cue, much of this work was performed on the avian brain and may not exist in its current form in the mammalian brain.

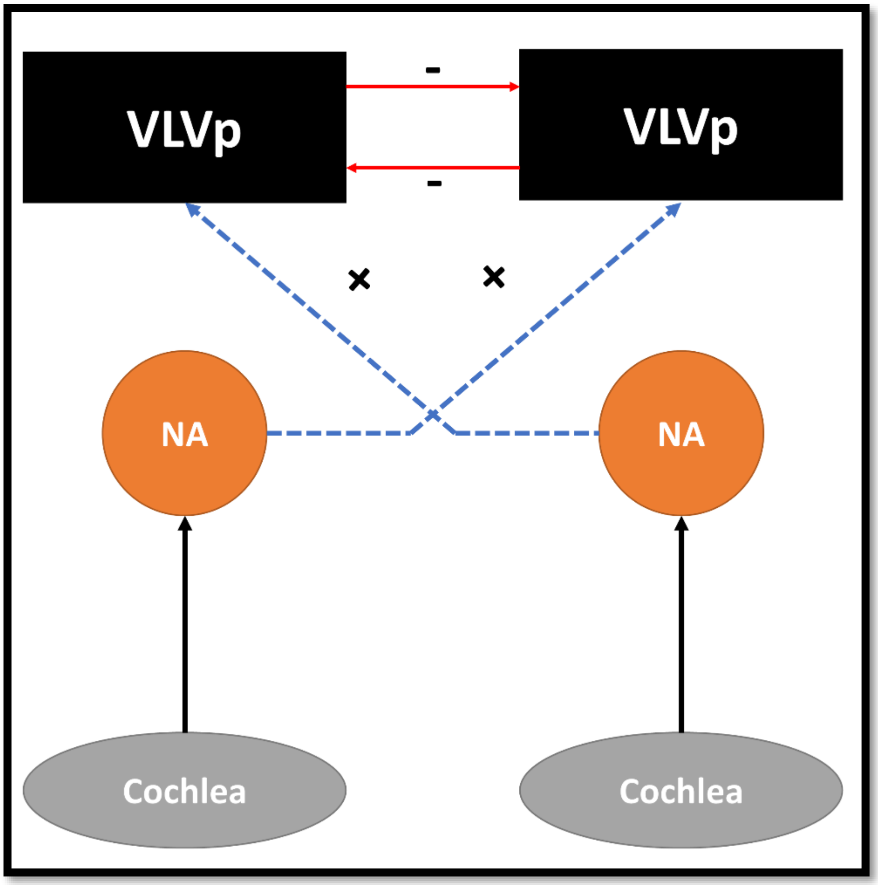

Figure 3. Visual description of the circuit thought to calculate ILD. Cochlear input drives activity of the ipsilateral nuclear angularis neurons (NA, in orange). The NA in turn provided excitatory input to the contralateral ventral lemnisces (VLVp, in black). Here, the lemnisces directly inhibit one another. It is thought that this competition processes the ILD cue.

There is evidence that sounds can be localized well in low- and high-frequency tones, but mid-frequency tones pose a problem (Tollin, 2003). As discussed earlier, the low-frequency sounds are best localized using the ITD cue, while the high-frequency sounds are localized easier using the ILD cue. Together, the relationship between frequency and processing ITD and ILD is known as duplex theory. Performance at mid-frequencies (i.e. tones that fall between the optimal ranges of ITD and ILD) is poor, suggesting that ITD and ILD are not calculated in the same brain area. One possible explanation for this poor performance is that a transition is occurring between the brain areas that calculate ITD and ILD, producing poor localization cues.

This is another prediction that falls out of duplex theory. Broadband signals, such as a short click, should be easy to localize because they contain information along all frequency bands. Within the ascending auditory pathway, one of the earliest areas to receive converging bilateral inputs are the primary nuclei of the SOC: the MSO and the LSO. Cells in MSO have shown to be strongly modulated by ITD with a periodicity matching the stimulus (Joris, Smith, & Yin, 1998; Yin & Chan, 1990). In fact, Carr and Konishi (1990) investigated the brain stem of the barn owl and found a circuit very similar to Jeffress' model. However, it is unknown if cells in the human MSO have an arrangement similar to what Jeffress predicted. There is evidence that low-threshold potassium channels limit the ability of bilateral inputs to be integrated in time (Grothe & Sanes, 1994; Smith, 1995). This suggests that the MSO is processing information similarly to how Jeffress predicted. It should be noted the exact mechanisms are not currently known and may not be similar to how the avian system processes ITD cues.

As much of the early work in auditory localization was performed in free field environments (e.g. (Blauert, 1969; Burger, 1958; R. Butler, 1969)) , it was not known if headphones were able to accurately represent auditory cues in space through manipulation of the audio to mimic the interaural and spectral cues the brain uses to locate sound sources. Wightman and Kistler (1989) presented wideband noise bursts to participants in either free field or by headphones. They found that participants wearing the headphones localized the sound source to virtually the same location as the free field participants. This is good evidence that headphones are a valid tool to emulate spatial audio. In the context of auditory augmented reality, this means auditory cues can be presented to the user through a pair of headphones or earbuds.

Taken together, the ITD and ILD cues give rise to the duplex theory of sound localization. Further work in the localization of sound refined the frequency ranges that ITD and ILD work best at. The

ITD cue is thought to encode space at lower frequencies (i.e. below ~1500 hertz) and the ILD cue is thought to encode space at higher frequencies (i.e. above ~1500 hertz). Early work by Stevens and Newman (1936) demonstrated that participants could localize (via pointing) a sound source on a loudspeaker emitting a broadband noise on a boom twelve feet high, where performance was highly dependent on the tone of the sound used. Further, they reported that clicks were localized faster than any tone, suggesting that the qualities in a click allow for faster and easier localization. Sandel, Teas, Feddersen, and Jeffress (1955) had participants adjust the position of a broadband noise source to correspond with the position of a sinusoidal source. As reported by Stevens and Newman (1936), sinusoidal sound sources can be difficult to localize because they contain a single frequency band, which may fall outside of the usable range for the duplex cues. They found that performance tended to be worst for sinusoids between 1500 – 3000 hertz. Essentially, the 1500 – 3000 hertz range is too high for usable ITD cues as well as too long for adequate calculation using ILD. This was taken as evidence that differential frequencies work for each of the cues. Overall, duplex theory creates a framework that explains the mechanisms for the localization sound sources in the horizontal dimension. To summarize, the ITD and ILD cues can be used to calculate horizontal position, though each works best at different frequency ranges. Clicks (i.e. short bursts of broadband noise) may be the most useful signal to localize because it spans frequency, allowing the auditory system to process the ITD and ILD cues at maximum efficiency. While clicks are not necessarily pleasant to listen to, they may be the best choice for an auditory augmented reality device.

Horizontal cues are especially useful in an auditory augmented reality system because they encode the azimuthal position of the object. The ITD cue is the difference of arrival times at either ear, which can be represented in software through the shifting of one audio signal relative to another. The ILD cue is the difference in amplitude at either ear, which can be created through attenuating one audio signal relative to another. Essentially, the fundamental principles driving horizontal localization can be emulated to provide additional (i.e. augmented) information to the user.

Vertical localization

Overall, ITD and ILD cues are sufficient to render azimuth, but not elevation. The brain uses different cues to compute elevation. We only intended to render azimuth, so vertical localization cues will not be discussed in detail. However, a complete auditory augmented reality system should include the ability to calculate both elevation and azimuth.

The head and ears are largely symmetrical, so most sounds occurring anywhere along the median plane (i.e. vertically in front or behind) do not produce any interaural difference cues. However, many early studies have shown humans can discriminate different points in vertical space accurately (Batteau, 1967, 1968; Woodworth, 1937). This suggests another mechanism is used to localize sounds in the median plane. Batteau (1967, 1968) suggested that the outer pinnae of the ears are providing a source of spatial cues that encode location. Essentially, the pinnae are producing changes on the spectrum which can be used to focus elevation. These changes in spectrum are known as spectral shape cues and are thought to be performed by the cavum concha leading into the ear canal. It is worth noting that these cues are monaural in nature, meaning only one ear is required to effectively determine elevation. This is a key difference from the duplex theory of sound localization (which is limited to azimuth).

Spectral shape cues are thought to be major contributors for vertical localization through changing the spectral information available to the tympanic membrane rather than interaural time or amplitude differences (R. A. Butler & Helwig, 1983; Gardner & Gardner, 1973; Oldfield & Parker, 1984). R. A. Butler and Helwig (1983) used a stimulus with a 1 kHz wide noise band with center frequencies ranging from 4 to 14 kHz and reported that sounds in the 4 – 12 kHz range appeared to move from the front to the rear along the median sagittal plane (i.e. the plane bilaterally bisecting the ears). At frequencies above 13 kHz, the sounds began to come from the front again. The authors suggest that this is due to the sound containing spatial referents that are used for encoding specific locations along the median sagittal plane. Further, they found that when increasing the bandwidth to 4 kHz (from 1 kHz), performance significantly increased. It is likely that the convolutions of the pinnae are performing this function. In fact, when bypassing the folds of the pinnae, vertical localization is no longer possible (Gardner & Gardner, 1973; Oldfield & Parker, 1984; Roffler & Butler, 1968).

Motion detection in auditory space

Much like the visual system, there are thought to be two auditory pathways the are functionally and anatomically segregated. One of the pathways is thought to be analogous to the dorsal visual pathway, where the planum temporale is thought to process motion (Rauschecker & Tian, 2000; Warren, Zielinski, Green, Rauschecker, & Griffiths, 2002). Motion is easier to understand in the context of vision. As such, the mechanisms that encode motion on the retina are much more intuitive than the auditory system. The eye is organized such that space is encoded on the retina, so it is easier to understand how motion could then be processed. The auditory system must complete several steps prior to getting a percept of space; the primary sensory organ, the basilar membrane, does not perform this function. Movement on the retina indicates either object movement or eye movement (or both), whereas movement on the basilar membrane indicates a pitch sweep. Combining both modalities allows for better localization of targets, with the auditory percept of motion biasing the visual percept (Meyer & Wuerger, 2001). Specifically, Meyer and Wuerger (2001) tested whether supra-threshold motion in the auditory system can bias the visual motion system. They found that the auditory direction of motion biases the visual direction of motion, with the bias being stronger if the motion is consistent (i.e. constant in velocity and the target is co-localized), suggesting that the auditory cue is powerful enough to modify the visual perception of motion. For horizontal motion, it is thought that the changing ITDs and ILDs are used to calculate a location over time. In an auditory augmented reality system, this information can be integrated in the brain to keep track of the object's location.

The ITD and ILD of a sound-emitting object will both change as it circles a listener. For example, a listener is positioned forward (0°) with a sound source to the right (90°) moving to the left (270°). Initially, the sound will enter the right ear first. As it moves in space, the ITD between the ears will shorten and eventually disappear as the object is located directly ahead. Eventually, as the sound continues moving left, the sound will enter the left ear first. The ILD will shift similarly; louder in the right ear followed by equal at both ears then louder in the left ear. Taken together over time, the percept is that of an object in motion. Computationally, the auditory augmented reality device can do this in real-time by shifting the audio channels and attenuating one signal relative to the other, while using very little resources. The cues are simple to program, making such a system viable for use given intuitive inputs.

The exact mechanisms for processing auditory motion are not completely understood, but there are two prevailing theories: *snapshot theory* and the existence of *specialized motion systems*. Snapshot theory states that the percept of motion in auditory space emerges the comparison of multiple static "snapshots" of an auditory source in time and space (Grantham, 1985, 1989; Neuhoff, 2004). Grantham (1985) showed that the minimum audible movement angle (i.e. the minimum angle a horizontally moving sound source must move to be known as a source in motion) of a source in motion increased significantly when the inter-stimulus interval (ISI) decreased from 500 to 50 milliseconds. Interestingly, this was not found to be true of a stationary sound source condition. Grantham suggests that this result is due to the dynamic nature of the stimulus and not the brief duration. This means that the same systems responsible for localizing auditory sources are also used for auditory motion processing. As location is already known, calculating features such as velocity is simply by integrating the change in position over time.

However, there is evidence that the auditory system will respond directly to objects in motion. Perrott, Costantino, and Ball (1993) showed that participants could discriminate between objects that were equally displaced in position and time, but the acceleration and deceleration phases required to get there were different. Because snapshot theory predicts that the velocity is calculated from the change in distance over time, the velocity percept should be identical in both cases. However, the results showed that the minimum duration needed to discriminate between the two conditions were different and had differing displacements. These results suggest that the velocity percept may not be singly dependent on simply integrating change in distance over time. Fundamentally, there must be more to determining motion than position displacement and movement time.

It is possible that both theories are at least partially correct. Perhaps the bulk of the velocity calculation is done via snapshot theory whilst changes in acceleration are detectable by some area within the auditory system. Grantham (1997) investigated this by creating two conditions to test auditory motion detection: a) a dynamic signal where a single wideband noise was presented to the listener and asked to say whether it was moving or stationary and b) a static signal presentation where two stationary noise bursts were played in close succession and the listener was asked to determine whether they came from different or the same spatial location. The results showed that at a speed of 20°/sec, more information was used during the time the target was in motion when compared to onset and offset information. However, at 60°/s onset and offset information led to equal performance to the target being present during the entire trajectory. This suggests that the auditory system may contain a specialized motion perception system that appears to only be active at lower speeds (i.e. probably below 60°/s). This relationship is like duplex theory, but in speed instead of frequency; the specialized motion system processes lower speeds whereas the snapshot mechanism processes higher speeds.

Summary

Overall, there are many different principles in auditory and visual localization that engage attention. For a proof-of-concept auditory augmented reality system designed to provide additional localization cues to the user, only the most useful of these principles should be implemented, at least initially. ITD and ILD cues are extremely important for horizontal localization, and thus should be amongst the first of the features implemented. Other cues, such as vertical object position, information specific to object motion, and object distance are also important, and should be added to the system in a future iteration for completeness. Because the purpose of this thesis was to determine if computer-generated auditory cues are sufficient to engage attention in a horizontal localization task, only the generation of ITD and ILD cues will be discussed in detail.

CHAPTER 2. Interaural time and level difference cue algorithms and generation

Choice of camera

Frame-based cameras are a conventional method of capturing visual information in a scene. A conventional frame-based camera will capture frames at some frame-rate (typically measured in hertz or reported as frames per second), which are stitched together to generate a video. This is suitable for most applications. However, if measuring changes in the scene are key to the project's operation, a frame-based camera becomes less suitable. Each recorded frame will likely contain a large amount of redundant information; parts of the scene that have not changed from the last frame are re-recorded. This is unlike the function of the retina, which is sensitive to transients or sudden changes in the scene. Because we are interested in calculating auditory cues from just the salient objects in the scene, a camera that functions similarly to the retina would be better suited for this job. One such sensor, the DAVIS 240B, uses neuromorphic technology to mimic human retina function. The neuromorphic sensor is much better suited for this kind of project because it is capable of capturing just salient events in a scene while ignoring stationary (or unimportant) objects. The DAVIS 240B is much more efficient than a frame-based camera; minimal processing is required to process the output data, and very little current is needed to power the device. See Figure 4 for a visual representation of the differences between a conventional and neuromorphic sensor.

Neuromorphic engineering is inspired by the natural algorithms that arise in biological systems. Devices such as the neuromorphic retina are created to imitate these biological system algorithms using silicon chips. The DAVIS 240B, a neuromorphic retina, operates by collecting salient visual information in the form of logarithmic luminance changes. These changes coincide with motion stimuli, which tend to be attention-grabbing. Using the data collected by the neuromorphic sensor, I have created a device that can convert the purely visual stimuli collected by the sensor into localizable auditory events. The camera operates similarly to the magnocellular visual stream through its natural fast computation of object motion. Using key principles of duplex theory, these visual events are converted into spatial audio to provide a user with additional localization cues. In some special populations, such as the blind, visual-to-auditory conversion gives access to a previously inaccessible modality. Using only horizontal auditory localization cues, the user is provided with the auditory spatial information necessary to deduce the azimuthal position of a purely visual stimulus.

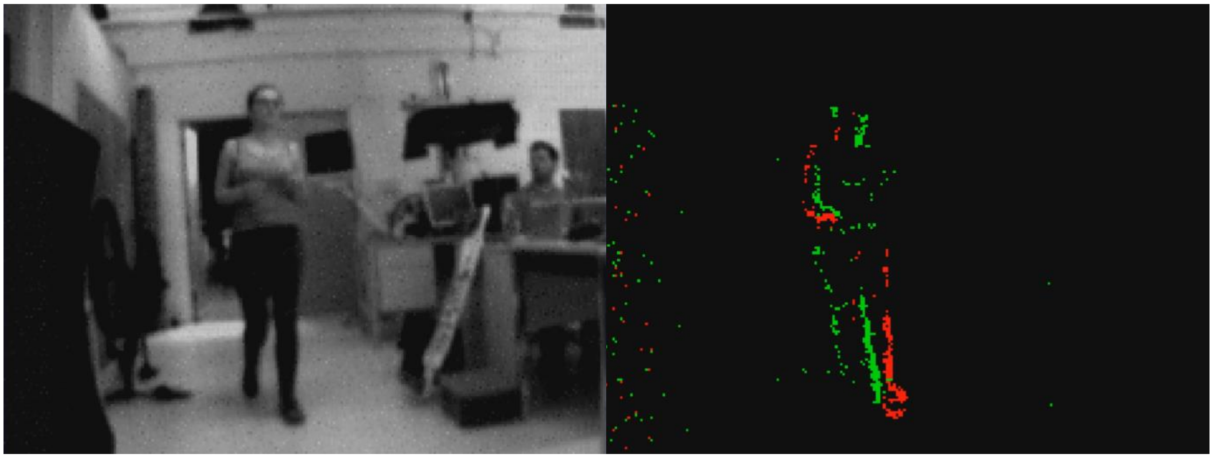

Figure 4. Comparison of the conventional camera (left) and the neuromorphic retina (right). Both scenes are identical, but recorded using either a frame-based capture (conventional camera) or through detecting luminance changes (neuromorphic retina). The conventional camera records frames of data, which are represented in black and white here. Each frame does not account for any changes that may have occurred since the last frame was recorded, resulting in a high amount of redundancy in the data. The neuromorphic sensor uses address event representation (AER) to collect changes in the scene and is much more sparse with data collection (though at a much higher collection rate). The green pixels represent when there is an increase in luminance and red when there is a decrease.

The neuromorphic engineering philosophy

Neuromorphic engineering is a relatively young field which involves the design of silicon chips that mimic biological sensory system function (Indiveri & Horiuchi, 2011; Liu & Delbruck, 2010). At its core, computational principles used by the brain are emulated by low-powered silicon chips. For example, two of the most prominent inventions in the field are the neuromorphic cochlea and retina. A neuromorphic cochlea works by approximating the function of the cochlea while a neuromorphic retina works by approximating the function of the human retina; light enters the sensor (where each pixel can be thought of as a ganglion cell) which detect logarithmic changes in luminance. When the sensor detects a change, it sends a *spike* to the connected computer in the form of an address event representation (AER;(Mahowald, 1994)). AER is a digital communication protocol that sends the address of the event with a matching timestamp. These events mark where, in pixel coordinates, the stimulus occurred as well as a microsecond-scale timestamp. AER events are asynchronous, meaning each event occurs independently of every other event. A conventional frame-based camera will capture frames at some set interval, regardless of any change in the scene which leads to redundant capture. At high computational cost, each frame is individually processed for salient changes in brightness or motion to capture the true changes in the scene. In contrast, the neuromorphic retina naturally does this, much faster and with lower power consumption. This is analogous to retinal function, where each ganglion cell can fire in synchrony or on its own into the optic nerve. The sensor on the neuromorphic chip is only sensitive to luminance changes and does not encode colour information. Thus, the output of the sensor closely approximates the magnocellular visual pathway in that it has very high temporal but low spatial resolution.

One such neuromorphic sensor, the DAVIS 240B (iniLabs Ltd., Zurich, Switzerland), is capable of encoding motion using minimal amounts of power (i.e. USB power from a laptop). The sensor records events that are similar to what the magnocellular output of the LGN would project into the dorsal visual stream. Events are encoded on the DAVIS using the AER communication protocol. In addition to location and timestamp, the polarity of the event is also captured. That is, the sensor can determine the direction of the luminance change (i.e. brighter to darker or darker to brighter). This can be used, for example, to detect objects approaching or receding from the sensor. Because the visual events are marked in space and time, it is therefore possible to use these events as the basis for equivalent auditory events. Essentially, if a purely visual event occurs somewhere in the scene, it should be possible to use the spatial information of that event to create an auditory event that occurs in that same space. One advantage of using a neuromorphic sensor for this type of processing is that the events of interest are already extracted from the scene with minimal processing. In fact, the motion detection is done via the hardware of the DAVIS sensor. This is unlike a conventional frame-based camera, where motion detection must be done on multiple still frames using computationally expensive image processing algorithms (for a review see Radke, Andra, Al-Kofahi, and Roysam (2005)). The goal of this thesis is to create valid spatial auditory cues using the visual input events from the DAVIS 240B neuromorphic sensor. Auditory cues generated allow the user to give attention to events that were previously exclusively visual. In this way, visual events can essentially be converted into auditory events.

Why generate spatial auditory events at all?

An auditory stimulus, even in the absence of a correlated visual stimulus, can be a useful signal to determine location. Visual localization is conceptually easier to understand when compared to auditory localization, but visual events are not useful to all populations. For example, purely visual events do not have matching auditory signals and are therefore inaccessible to the blind. Augmenting auditory cues into the auditory scene gives the blind user access to information previously impossible to perceive. Visual motion events tend to be important, especially in busy urban environments. Crossing the street without any visual information can be extremely challenging because sounds, unlike vision, do not have spatial boundaries. For example, listening carefully for nearby passing cars at a crosswalk, only to be drowned out by a loud pickup truck significantly increases the difficulty of crossing the street. When presented with many different sounds at differing amplitudes and frequency bandwidths, it becomes difficult to ascribe a sound to a source with great certainty. A fully-developed auditory augmented reality device might operate as follows: a user points the device at an area of interest, such as the road immediate to the crosswalk to determine if the road is safe to cross, and valid auditory localization cues marking the location of any approaching vehicles are generated. The auditory events generated have spatial properties, allowing the user to glean crucial, previously exclusively visual, information. Any events generated by the camera will be relevant to the user, because the sensor naturally generates non-redundant temporal information.

Generating auditory cues from neuromorphic visual events

Auditory interaural cues can be used to encode azimuthal space. Duplex theory describes two cues used to calculate azimuthal position: interaural time (ITD) and level (ILD) differences. The ITD cue is calculated from the difference in time of two signals entering the ears, whereas the ILD cue is calculated from the level (or amplitude) difference between the two signals. For example, a human is standing with an active loudspeaker at 45° to the left. The signal will enter the left ear shortly before the right (i.e. the ITD cue) and will be louder in the left ear relative to the right (i.e. the ILD cue). These two cues are sufficient information to locate the position of the sound source. Even without vision, the spatial percept will be very accurate. Using *PortAudio*, an open-source audio input/output library for C++, these cues can be accurately emulated using a low-powered laptop and a pair of headphones.

To access events from the camera, an application programming interface (API) must be used. The API is a set of functions and routines necessary to get data from the camera. An API, known as *libcaer*, was built by *iniLabs* for the DAVIS line of cameras to allow for access to data generated by the sensor. The camera uses the AER data communication protocol, meaning that data is sent in the form of timestamped pixel coordinates. Pixel coordinates can be used as a spatial marker of where the stimuli are occurring in the scene. Because the camera is only collecting salient motion events, minimal filtering is needed. When data is collected, it is sent to the computer in *packets*. Each packet can contain up to 4096 events, with more events being put into a packet when the scene is visually busy (i.e. generating a lot of events in a short period of time). To maintain a high signal-to-noise ratio, packets that contain a small number of events can be discarded as they likely contain noise. An example of visual noise is the flickering of fluorescent lights. This simple method of filtering allows for a robust collection of salient, yet relevant, events.

Calculating and creating the ITD cue

Sound travels at a finite speed of approximately 340 meters/second. If a human is listening to a sound located directly in front of them, the sound will reach the ears at precisely the same time because the distance to each ear is identical. If, however, the sound is shifted slightly to the right, the distance to reach the right ear is slightly shorter than the left ear (see

Figure 5). This results in a small delay between the time the signal enters the right ear and when it enters the left ear. Because the time difference is so small, it is best expressed in number of microseconds. In audio processing, audio is typically represented as samples instead of time; one sample represents the audio captured at some moment in time. Audio samples are collected at a frequency known as the sample rate. While there are many common sampling rates, 44.1 kHz is commonly used. In other words, 44,100 samples of audio are collected every second. One sample at this rate is equal to 22.7 microseconds. This means one sample of audio is collected approximately every 23 us. Using PortAudio, a two-channel stream of audio can be manipulated by shifting the audio of one channel relative to the other. In a stereo signal, each channel will be sent independently to each earphone. This approximates the basic principle of ITD: an arrival time difference at the level of the ear.

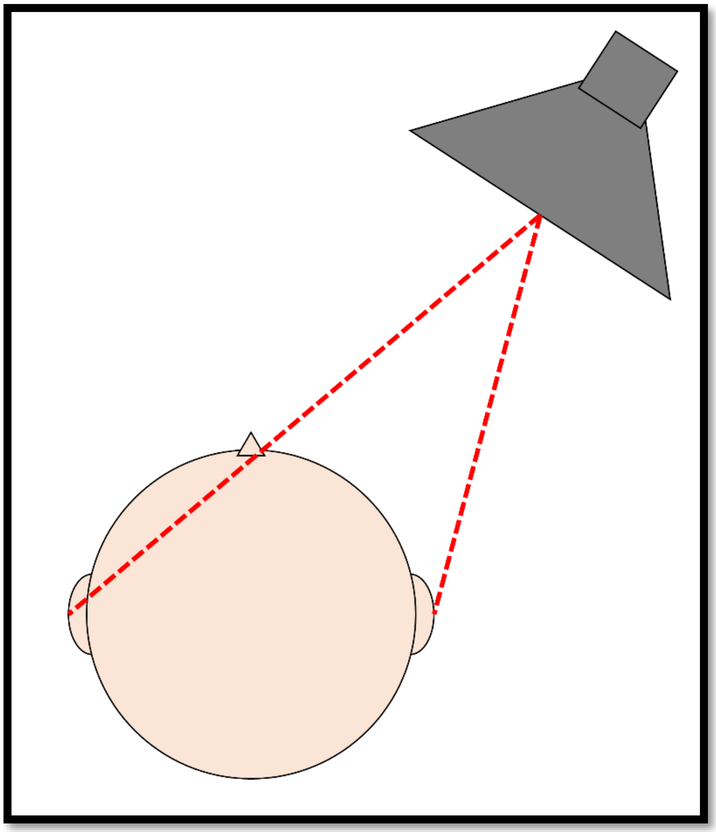

Figure 5. A human listening to a loudspeaker that is shifted to the right. The red lines denote the direct path from the loudspeaker to the ears of the listener. The sound would enter the right ear prior to the left ear. The further shifted the sound source is, the larger the time difference between the ears. This is known as the ITD cue. The signal would be louder in the right ear relative to the left. The further shifted the sound source is, the more attenuated the signals becomes in the second ear because it must travel through the cranium. This is known as the ILD cue.

The algorithm used to approximate ITD works as follows (Figure 5):

1. events are collected from the DAVIS sensor using the AER protocol

2. the horizontal coordinate (i.e. the ‘X’ coordinate) is used to calculate the distance in pixels from the center of the scene

3. the distance from the center of the scene is mapped to an integer value between -18 and +18 (e.g. a distance of 50 pixels to the right: 50/95 * 18 = ~9 samples

4. this value is then used as the number of samples to shift the channel further from the sound source

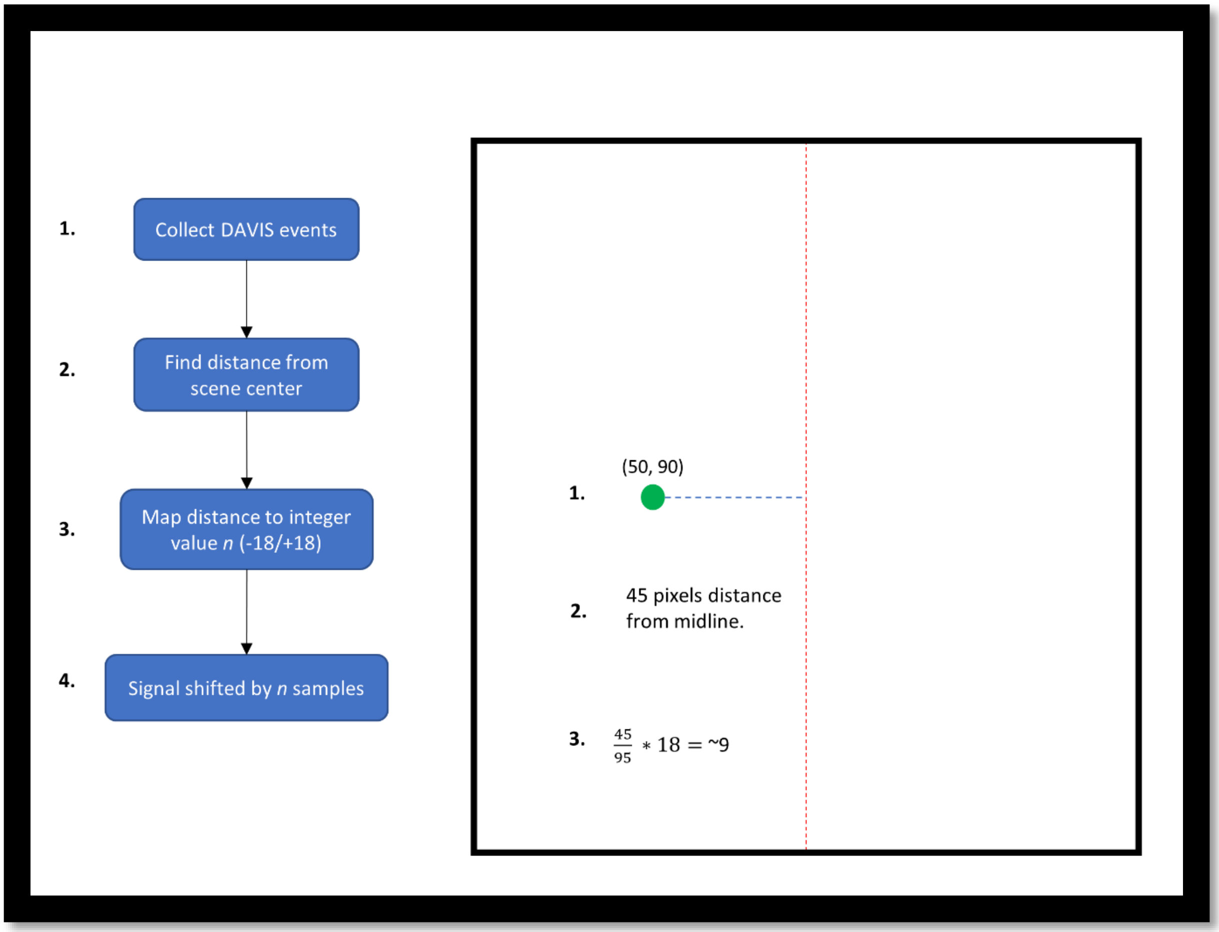
1.
Collect DAVIS events
2.
Find distance from scene center
3.
Map distance to integer value *n* (-18/+18)
4.
Signal shifted by *n* samples
(50, 90)
1.
2.
45 pixels distance from midline.
3.
$\frac{45}{95} * 18 = \text{~}9$

Figure 6. Visual description of how the ITD cue is calculated using the neuromorphic camera for input. The algorithm collects visual events (1; e.g. the enlarged green pixel at coordinates (50, 90) is active) and finds the distance from the center of the scene (2; the red line denotes the horizontal center of the scene, which happens to be 95 pixels). This value is then mapped to an integer between -18 and +18 (3; this value, *n*, is used as the number of samples to shift the audio sample by) then finally the audio is shifted by that amount (4; not shown in visualization).

Calculating and creating the ILD cue

The other major cue of duplex theory, ILD, is thought to be driven by the cranium absorbing the sound as it passes through the head. This is known as the head shadow effect (Schleich, Nopp, & D'haese, 2004). If the head and ears are pointed at the sound source, the ILD effect is minimal because the sound does not have to pass through the head. However, if one ear is pointed towards the source, the sound necessarily must pass through the head to reach the other ear (see Figure 5). In digital audio, this effect can best be emulated by attenuating one channel and leaving the other. This is accomplished by multiplying the channel-to-be-attenuated by a number less than 1. When presented simultaneously with an unattenuated signal, the percept is a valid ILD cue because one signal is simply louder than the other.

The algorithm to approximate the ILD cue works as follows (Figure 6):

1. events are collected from the DAVIS sensor using the AER protocol

2. the horizontal coordinate (i.e. the ‘X’ coordinate) is used to calculate the distance in pixels from the center of the scene

3. the distance from the center is used to calculate an attenuation factor by dividing the distance by the maximum possible distance from the center of the scene (e.g. distance of 50 pixels: 50/95 = ~0.5)

4. to assert that one channel is not reduced by 100%, the attenuation factor is multiplied by a *maximum attenuation factor* (default: 95%). the signal on the adjusted channel is multiplied by the attenuation factor

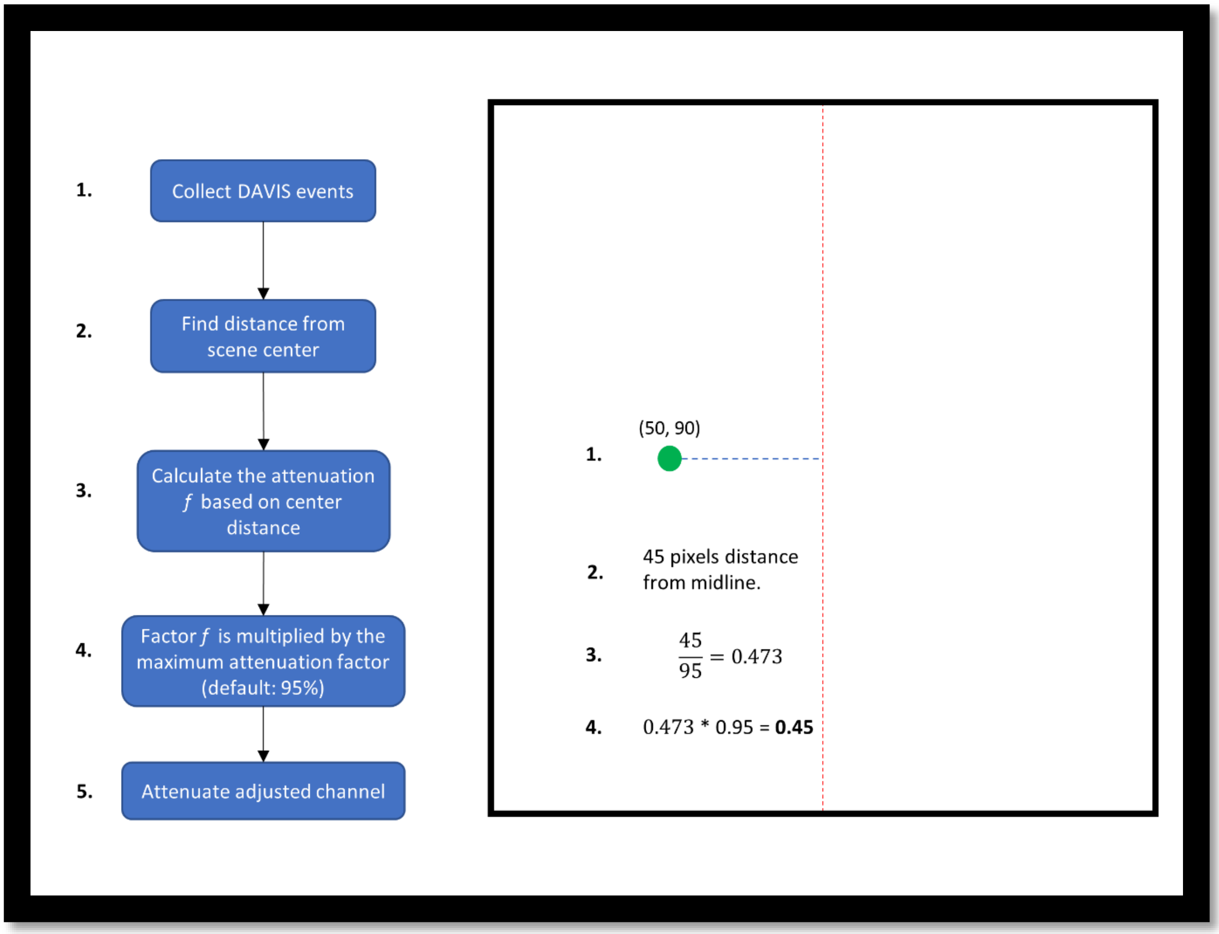
1.
Collect DAVIS events
2.
Find distance from scene center
3.
Calculate the attenuation f based on center distance
4.
Factor f is multiplied by the maximum attenuation factor (default: 95%)
5.
Attenuate adjusted channel
(50, 90)
1.
2.
45 pixels distance from midline.
3.
$\frac{45}{95} = 0.473$
4.
0.473 * 0.95 = **0.45**

Figure 7. Visual description of how the ILD cue is calculated using the neuromorphic camera for input. The first two steps are identical to how the ITD cue is processed and calculated: visual events are collected (1; e.g. the enlarged green pixel at coordinates (50, 90) is active) and finds the distance from the center of the scene (2; the red line denotes the horizontal center of the scene, which happens to be 95 pixels). A value is calculated by dividing the distance by the horizontal center's pixel value to get a ratio, known as *f* (3). To ensure one channel cannot be attenuated to 100%, *f* is multiplied by some maximum attenuation factor (default: 95%; 4). The channel contralateral to the stimulus location is multiplied by the factor *f* to approximate the ILD cue (5; not shown in visualization).

Given that the calculations performed by these algorithms are not computationally intensive, one of the largest limiting factors for performance is auditory signal length. A longer signal will take more time to shift because each sample must be moved individually. However, it is possible to use an extremely short signal, such as a single sample. In the current implementation, one sample is generated in each channel, resulting in only a single value requiring shifting and attenuation. The low workload allows for extremely quick operation.

CHAPTER 3. Assessing efficacy of the device in control and blind individuals

Abstract

Many salient visual events tend to coincide with auditory events, such as seeing and hearing a car pass by. Information from the visual and auditory senses can be used to create a stable percept of the stimulus. Having access to related coincident visual and auditory information can help for spatial tasks such as localization. However not all visual information has analogous auditory percepts, such as viewing a computer monitor. Here, we describe a system capable of detecting and augmenting visual salient events into localizable auditory events. The system uses a neuromorphic camera (DAVIS 240B) to detect logarithmic changes of brightness intensity in the scene, which can be interpreted as salient visual events. Control and congenitally blind individuals participated in this study. Participants asked to use the device to detect new objects in the scene, as well as determine direction of motion for a moving visual object. Results suggest the system is robust enough to allow for the simple detection of new salient stimuli. Control users appeared to perform better than blind users at the motion discrimination task. Future successes are probable as neuromorphic devices are likely to become faster, smaller, and more affordable, allowing for easier individualization and calibration and making this system much more feasible.

Introduction

Attentional orienting mechanisms allow us to notice important sensory events and reallocate perceptual resources to deal with them.   A variety of visual events are known to trigger reorienting - particularly motion stimuli and abrupt onsets of new objects (Franconeri & Simons, 2003; Yantis & Jonides, 1984) and the ability to detect these attentional cues is critical to safely interacting with the world. Indeed, visual deficits or impairment of the attention orienting system due to stroke can be debilitating. Fortunately, visual events often coincide with auditory events, thus providing a multimodal cue.   For someone with a visual deficit, this coupling of auditory and visual information is critical because it affords the only indication of a potentially important change in the sensory world. However, not all important sensory events are multimodal. For example, the driver of a car might try to alert pedestrians with a horn (purely auditory) or might instead flash the headlights (purely visual). Likewise, objects that start moving in a cluttered auditory scene might not be heard above the background noise floor. Failure to notice such events constitutes an important safety hazard for people with visual impairments. Here we report preliminary success in developing an auditory augmented reality system that renders visually salient events (onsets and motion) onto the spatial auditory scene to provide auditory cues about the visual world.

The perception of auditory motion is largely dependent on the ability to detect the speed and direction of the event. Speed and direction can be perceived through use of the interaural time difference (ITD) and interaural level difference (ILD) cues (Lewis, Beauchamp, & DeYoe, 2000). It is believed that spectral notches performed by the pinnae can provide an additional cue about the elevation of the sound. The ITD cue is encoded by the difference of arrival times of sound at each ear (Brand, Behrend, Marquardt, McAlpine, & Grothe, 2002). That is, a sound source closer to the right ear will enter the right ear up to around 700us before entering the left ear. This small amount of time is dependent on the distance between the ears and the azimuthal arrival angle. The ITD cue can be used by the auditory system to calculate the angle of the sound source. When one ear is closer to the sound source, the sound will have a higher intensity relative to the other ear. This difference in sound intensity is known as ILD. For a review of sound localization techniques, see (Tollin, 2003). When these cues are implemented using software and headphones, the percept of a localizable sound source is apparent. In the same way that someone may localize the auditory motion percept of a car as it passes by, a software-generated sound source can also be panned through virtual space. A key difference is that listening in the free acoustic field (i.e. not through headphones) provides extra cues as to the source elevation and distance. It is worth noting that these cues typically need to be interaural in nature, as monaural cues do not sufficiently code elevation (Jin, Corderoy, Carlile, & van Schaik, 2004). When using only the ITD and ILD cues through headphones, high-resolution localization is possible, but only in along the azimuthal dimension.

It has been demonstrated that non-individualized head related transfer functions (HRTFs) can result in accurate localization (Wenzel, Arruda, Kistler, & Wightman, 1993). Wenzel et al. (1993) showed that participants could localize the direction of narrowband noise when using a representative subject’s HRTF. More recent works have focused on measuring individualized head related transfer functions for auditory scene synthesis (S. Xu, Li, & Salvendy, 2007; Zotkin, Hwang, Duraiswaini, & Davis, 2003), suggesting that individualized HRTFs maintain the spectral cues important for resolving location. These methods carry the benefit of higher accuracy, but are much more difficult and time-consuming to implement. ITD and ILD cues can be sufficient for azimuthal localization.

An important computational challenge arises when attempting to render visual events into auditory events: a typical frame-based video camera provides highly detailed raw information about color and luminance at frame rates of around 30 frames-per-second. Extracting important events from the dynamics across such frames is computationally intensive, and the Nyquist limit of 15 Hz imposes an upper limit on the temporal resolution. Biologically, conventional frame-based cameras are analogous to the parvocellular and the ventral visual pathway, which conveys high spatial resolution, texture, and colour information, but is slow (Felleman & Van Essen, 1991; Maunsell et al., 1999). By contrast, the visual pathway thought to drive the posterior parietal attention orienting system, the magnocellular dorsal pathway, is fast, low-resolution, and insensitive to colour. The ideal camera system for an attentional augmented reality would forgo the high computational demands of a parvocellular-like frame-based camera, and instead emulate the fast, low-resolution dynamics of the magnocellular system. For this reason we designed our augmented reality system not around a frame-based camera, but around a neuromorphic Dynamic Vision Sensor (DVS) (Lichtsteiner, Posch, & Delbruck, 2008; Liu, Delbruck, Indiveri, Douglas, & Whatley, 2015).

Neuromorphic sensors are based on the design of silicon chips that mimic the underlying function of biological sensory systems (such as the retina and cochlea). A dynamic vision sensor silicon retina approximates the basic information processing pipeline of the human retina: the sensor sends spikes to a computer. The spikes represent log intensity (brightness) changes. The output of the camera signifies the relative changes in scene reflectance. Moving edges and sudden luminance changes are salient to the DVS and generate bursts of spikes, while slow changes and isoluminant edges do not. The neuromorphic retina uses an address-event-representation (AER) system, which allows for the timestamped ordering of temporal contrast events tagged with their spatial coordinates. Due to the way the AER system works, no temporally redundant data is captured; if there are no triggered events on the sensor, no information is sent on to a processing device. By contrast, a conventional frame-based camera sends frames continuously. Since the only information forwarded by the neuromorphic sensor is related to attentionally-relevant events in the scene, it is possible with minimal computation to render visual events into to an augmented auditory space while still maintaining the spatial characteristics of the scene and with a reduced demand on power and computational resources. Early work in auditory augmented reality systems resulted in the SeeHear system: an aid device for the blind (Cao, Mattisson, & Bjork, 1992; Nielsen, Mahowald, & Mead, 1987). A lens projected light onto a 15x11 matrix of photoreceptors which calculated the light intensity at each point. These intensities were then propagated along a delay line to simulate the time delay of the sound in air. The chip would then output a stereo signal mimicking the spatial properties of the sound in a pair of headphones. Essentially, the device could convert a visual stimulus in motion into an auditory stimulus. The psychophysical efficacy was never evaluated in human studies, so it is unknown how useful such a device would be for users. Other auditory augmentation hardware/software, including the VIS2SOUND (Morillas, Cobos, Pelayo, Prieto, & Romero, 2008) and TESSA (Martinez & Hwang, 2015) have also demonstrated that it is possible to convert visual events into spatial auditory events but no human trials have been conducted.

We created an augmented visual-to-auditory system which takes a neuromorphic visual input and augments it to auditory space while still preserving the spatial characteristics of the scene using the ITD and ILD cues. A user of the device experiences visual onsets as bursts of auditory clicks at some azimuthal angle related to the position of the AER visual event. Likewise, a continuously moving visual object is heard as a train of clicks that pans through auditory space with speed and direction related to the visual stimulus. This visual stimulus used was a large white dot on a LCD computer screen either stationary or moving. The algorithm detects the centroid of the stimulus to use as the location of the dot. In this study, we performed two experiments: 1) testing as a proof-of-concept that the device works for regularly sighted blindfolded controls and 2) testing using a congenitally blind population to assess the efficacy of the device. We show here that a control blindfolded listener with almost no training (experiment 1) can detect visual onsets and can determine the direction of a visual stimulus in motion at varying speeds and displacements. Congenitally blind individuals were able to complete the onset detection as well as blindfolded controls, but all found the motion discrimination task confusing, leading to two individuals dropping out of the task.

Methods

Participants (controls)

Nineteen normally sighted individuals from the University of Lethbridge participated in the present study. The study was approved by the University of Lethbridge Human Subject Research Committee (protocol #2013-037), and all participants gave written informed consent prior to participating.

Participants (blind)

Five congenitally blind individuals recruited from the Canadian National Institute for the Blind participated in this study. Two individuals dropped out of the study during the motion discrimination portion, so only data from the three remaining participants will be discussed.

Apparatus

The neuromorphic camera (DAVIS 240B, inilabs.com) was placed in front of a laptop computer (15.6” Lenovo Y510P, Intel i7-4700MQ) screen (1920x1080 60Hz) with custom MATLAB Psychophysics toolbox code (Brainard, 1997) running to generate still and moving stimuli (Figure 7). The camera was placed 42 centimeters from the screen to ensure the entirety of the screen was captured. The camera was connected to a second laptop computer (Acer Aspire One D255E-1638, Intel Atom N570 1.66 GHz) to generate the spatial audio events of the scene. Clicks (i.e. a ‘1’ in the audio buffer) were generated with ITD and ILD modulation to create the percept of spatial azimuth. For each visual event generated by the neuromorphic camera, a matching auditory click was generated. Participants, wearing a pair of headphones (Sennheiser HD280, Sennheiser, USA) and a blindfold (unless they were blind), were seated next to the apparatus separated using a large wooden barrier to prevent events generated by the participant’s actions to be reflect on the stimulus laptop’s screen (and thus the DVS). The participant did not glean any visual information during the experiment.

Algorithm of ITD and ILD

Custom software developed using C++ allowed for visual events generated by the neuromorphic sensor to be perceived in the auditory domain with the azimuthal spatial quality preserved. The interaural time difference is the difference of arrival time between the ears. To simulate this in a pair of headphones, one of the two channels of identical audio need to be shifted. The common audio sampling rate of 44.1 kHz was used, where shifting by 18 samples creates a 400 us shift in the sound (i.e. $\frac{18\ samples}{44100\ samples/second} = 400\ us$). The ITD algorithm worked as follows: as a visual event occurs, the horizontal distance from the midline of the scene (95 pixels is the center, as there are 190 pixels across) was simply mapped to a value between -18 and +18. This value was used as a relative offset from the other channel to simulate the ITD cue. For example, if a visual event occurs in the far left of the scene (e.g. a midline offset distance of -95 pixels), the audio signal would be shifted by -18 samples in the audio channel's buffer.

The ILD cue was generated by taking each event's absolute horizontal distance from the scene's midline, and attenuating the intensity of the audio. Each sample could be attenuated from a range of 0 to 95% (i.e. midline to far-right or far-left). For example, if a sound occurs in the extreme-far-right portion of the scene, the left channel's audio would be attenuated by 95%.

Onset detection and motion discrimination tasks

The study had two tasks: onset detection and motion discrimination.

For the onset detection phase, 100 trials were generated. There were an equal number of trials showing a stimulus (e.g. a white dot on a black screen) or a blank screen. Participants responded using an external keyboard to indicate whether a stimulus had been perceived in auditory space.

For the motion discrimination task, 500 trials were generated. Five different total displacements were used, ranging equally from 100 pixels to 400 pixels on the computer monitor (i.e. 100, 175, 250, 325, and 400 pixels or in visual degrees: 2.31, 4.05, 5.78, 7.52, and 9.25.). The displacement referred to the number of pixels the stimulus would move from the original starting point. The task was counterbalanced to include an equal number of displacements, equaling 100 total trials for each displacement, making up the total of 500 trials. Similarly, five different stimulus speeds were used, ranging from 5 to 20 pixels/frame (i.e. 5, 8.75, 12.5, 16.25, and 20 pixels/frame or in visual degrees/second: 7.20, 12.3, 17.4, 22.5, 27.6.). Direction of motion was equally counterbalanced for a total of 250 trials moving to the left or right. The task was run at a framerate of 60 Hz. Again, the task was counterbalanced to include an equal number of speeds, equally 100 total trials for each speed. These values were used to determine a threshold, should any exist, in either displacement or speed. Like the onset detection task, participants responded using an external keyboard to indicate whether a stimulus had moved either left or right.

Stimuli were pseudo randomly generated prior to beginning each task. Starting position was pseudo randomly chosen, and counterbalanced on the left and right. All participants had pseudo randomized orders of stimulus presentation. Reaction time and hit rate was recorded for each trial. Reaction time was determined as the amount of time to respond after the stimulus first appeared on the screen.

Timing of the algorithm execution time

An apparatus was built to calculate the time difference between the onset of the visual and auditory events. The time difference was consistently found to be in the range of 6 – 8 milliseconds, demonstrating that the algorithm was working in near-real-time. This test was performed to ensure there was not a large delay between visual and auditory onsets, meaning it can be used to react to stimuli in very quickly.

General analysis

All analysis was performed offline using SPSS Statistics 22.0 for Windows (SPSS Inc., Chicago, IL, USA).

Hit Rate (HR)

To determine performance on the task, each trial was evaluated and marked as either correct or incorrect when compared to the actual known movement of the stimulus for the trial. Hit rate was calculated as the number of correct responses divided by the total number of responses.

Reaction times (RT)

Reaction time analysis was performed to get a sense of the cognitive load required to complete the task. The more difficult the task, the more time it should take to complete. The reaction time was calculated as the amount of time between the visual onset and the key press.

Experiment 1 results

Onset detection HR

Hit rate was calculated as the average number of correct trials. Participants detected onsets with a 95.6% (sd: 9.1%) success rate.

Motion discrimination HR

When collapsing across displacement and speed, participants discriminated the direction of motion with a 62.7% (sd: 3.81%) success rate.

Displacement HR

Percent correct ranged from 55.247% (D100) to 70.699% (D400), with hit rate increasing with each increase in displacement (Figure 8a). Pairwise comparisons revealed significant differences between all measures of displacement, except for D175 & D250 and D250 & D325 (Table 1). A repeated-measures ANOVA revealed a significant effect of displacement ($F(4, 72) = 16.945$, $p < 0.001$).

Speed HR

Most of the hit rates tended to be like one another (Figure 8b). Pairwise t-tests confirmed no significant difference in hit rate between any of the speed measures. A repeated-measures ANOVA revealed no significant effect of speed on hit rate.

Onset detection RT

The average response time for the signal detection task was 2.02 seconds (sd: 2.00 seconds). Further investigation revealed two participants whose data influenced the mean with significantly longer reaction times (z-score > 2.2). Upon removal, the average response time was 1.38 seconds (sd: 0.71 seconds). When removing individual trials that exceeded a z-score of +-3 within each participant's block, the average response time was found to be 0.94 seconds (sd: 0.43 seconds).

Motion discrimination RT

When collapsing across displacement and speed, control participants reacted an average of 1.34 seconds (sd: 0.70 seconds) after stimulus onset.

Displacement RT

A repeated-measures ANOVA was performed, which revealed no significant effect of displacement (F(4,72) = 1.018, p = 0.361)) (Figure 9a), suggesting that reaction time does not change with displacement. Variance was not found to be equal within subjects for this analysis, and as such the Greenhouse-Geisser bounds were used to provide a more accurate F value.

Speed RT

For speed, significant differences between S5 and S8.75 ($t(18) = 9.356$, $p < 0.01$), S5 and S12.5 ($t(18) = 8.379$, $p < 0.01$), S5 and S16.25 ($t(18) = 6.047$, $p < 0.01$), S5 and S20 ($t(18) = 4.107$, $p < 0.01$), S8.75 and S12.5 ($t(18) = 2.988$, $p < 0.01$) and S8.75 and S16.25 ($t(18) = 2.206$, $p < 0.05$) were revealed. A repeated-measures ANOVA revealed a significant effect of speed ($F(4, 72) = 19.485$, $p < 0.001$), with reaction times in S5 being significantly slower than the rest of the measures (Figure 9b).

Guessing

Binary forced-choice tasks can produce results consistent with guessing. As such, it is necessary to perform a one-sample t-test to rule out the possibility that the participants are simply guessing. A one-sample t-test with a mean of 50 (percent) was performed on all data. All p-values were significant (all ps < 0.001), suggesting that any effect generated from the data was not due to guessing.

Experiment 2 results (congenitally blind)

Onset detection HR

Hit rate was calculated as the average number of correct trials. Participants detected onsets with a 96.4% (sd: 2.5%) success rate.

Motion discrimination HR

When collapsing across displacement and speed, participants discriminated the direction of motion with a 54.3% (sd: 1.31%) success rate.

Displacement HR

Percent correct ranged from 48.9% (D100) to 58.4% (D400) (Figure 10a). Because we were interested to see if a particular displacement resulted in better or worse performance relative to others, pairwise comparisons were used. Pairwise comparisons revealed significant differences between D100 and D250 as well as D100 and D400. A repeated-measures ANOVA showed no effect of displacement (F(4, 8) = 2.168, p = 0.168).

Speed HR

Most of the hit rates tended to be like one another (Fig 10b). Pairwise t-tests confirmed no significant difference in hit rate between any of the speed measures. A repeated-measures ANOVA revealed no significant effect of speed on hit rate (F(4, 8) = 0.591, p = 0.679).

Onset detection RT

The average response time for the signal detection task was 0.89 seconds (sd 0.46 seconds). When removing individual trials that exceeded a z-score of +-3 within each participant's block, the average response time was found to be 0.76 seconds (sd: 0.31 seconds).

Motion discrimination RT

When collapsing across displacement and speed, blind participants reacted an average of 1.10 seconds (sd: 0.53 seconds) after stimulus onset.

Displacement RT

A repeated-measures ANOVA was performed, which revealed no significant effect of displacement (F(4,8) = 1.164, p = 0.395)) (Figure 11a), suggesting that reaction time does not change with displacement.

Speed RT

A repeated-measures ANOVA was performed, which revealed no significant effect of speed (F(4,8) = 2.295, p = 0.147)) (Figure 11b), suggesting that reaction time does not change with displacement.

Guessing

All congenitally blind participants complained that the motion discrimination task was not intuitive, and as such it is possible that the results obtained are due to guessing the direction of motion. To determine if the blind users were gleaning useful information from the experiment, a one-sample t-test was used. All measures, except for D250, were not significantly different from a mean of 50 (percent). This suggests it is possible the blind participants were simply guessing for many of the trials.

Experiment 1 Discussion

Experiment 1 investigated whether an auditory augmented-reality system could render visual events as salient auditory events and convey spatial information about the visual scene for regularly sighted university-aged controls. The system converts purely visual stimuli into auditory stimuli with preserved spatial characteristics. Control participants were able, with no training and very little practice, to successfully detect a visual stimulus as well as determine the directionality of the stimulus in motion at a rate significantly better than chance. A neuromorphic camera allowed for the detection of visual stimuli with minimal power and processing requirements. Custom software to approximate the ITD and ILD cues allowed for the visual events encoded by the neuromorphic sensor to be perceived in the auditory domain.

In both tasks, reaction time (RT) was much slower than would be normally expected, where approximately 200 milliseconds would be the expected value (Niemi & Näätänen, 1981; Pain & Hibbs, 2007). An average reaction time across valid trials was 0.94 seconds (sd: 0.43 seconds) in the onset detection task and 1.34 seconds (sd: 0.70 seconds) in the motion discrimination task. As visual-to-audio latency was on the scale of a few milliseconds, these prolonged RTs were due to sensory and perceptual effects. One possible reason for this lengthy RT was the presence of occasional spurious noise events in the camera output, which appeared to users as occasional auditory events with random locations and brief duration. In general, a high signal-to-noise ratio was achieved, with only occasional noise. However, some participants reported that it was difficult to filter out the extraneous noise events. We speculate that the prolonged RTs may have been due to the uncertainty involved with ignoring the noise stimuli, with additional time required to build evidence for the true onset of a visual event. An additional possible reason for the prolonged RT in the motion discrimination task is that the total stimulus presentation times varied randomly, making the participant unsure of how long the stimulus would last.

We found a main effect of displacement for motion discrimination. Essentially, larger displacement meant more time to build evidence for the direction of the motion for the trial, so this was expected. Speed, however, was not found to affect accuracy in any significant way. When collapsing across speed groups, a mean accuracy of 62.7% (sd: 0.69%) was found. This suggests that participants formed their judgments of motion direction using spatial displacement rather than velocity. One strategy could be to determine the start and end points of the task and simply figure out what direction was necessary to get from the start to the end.

These findings suggest that the device creates convincing auditory percepts of objects in space and will therefore be useful for special populations (e.g. the blind). The main effect of displacement was not surprising to find because a larger displacement allows for more information to be gathered. The further the object moves, the more information useful information gained, leading to better performance. It is probable that accuracy would continue to increase for larger displacements. In the example of a stimulus beginning in the extreme left of the scene and ended up on the extreme right (a total of 1080 pixels travelled in total), the auditory percept would move from one headphone to the other.

Experiment 2 Discussion

Experiment 2 assessed the efficacy of the device in a blind population. The device was not necessarily designed for use by the regularly sighted. With approximately 39 million blind individuals in the world (Pascolini & Mariotti, 2012) and limited options for attentional prostheses, the blind population is an excellent target consumer of such a device. We recruited a total of five blind participants for this experiment, of which two dropped out due to feeling confused by the motion discrimination task. Overall, it appears the blind participants performed worse than the controls, though it is difficult to accurately compare the two groups due to the difference in sample size.

Blind participants could successfully detect a visual onset with a 96.4% accuracy. This result suggests a capability of understanding when a stimulus is present, much like the controls. When given the motion discrimination task, blind participants achieved a 55.3% success rate. However, blind participants expressed feeling confused by the motion discrimination task's goal. Unlike the control sighted individuals, it was not possible to explain what the stimulus looked like since the participants had never had a visual experience. When prompted to imagine an indistinct object moving through space, most blind participants tended to perform better. It was not unusual to pause the experiment several times to reassure the participant that they had the correct understanding of the goal.

In the onset detection task, the blind participants reacted on average within 0.89 seconds (sd: 0.46 seconds) during valid trials. During the motion discrimination task, blind participants reacted on average within 1.10 seconds (sd: 0.53 seconds) of the stimulus onset. With a larger sample size, this difference would likely be significant because direction of motion is more difficult to determine than onset detection. While this reaction time is still higher than would be expected from a simple onset detection task, it is reasonable to expect because the task is unusual in that it requires participants to a) detect a valid auditory stimulus from noise and b) determine a direction of motion. Occasional uncorrelated noise events are generated on the sensor, mostly due to fluorescent light flickering. These events are largely filtered out through a scene activity threshold, but noise events slip through infrequently. Many of the blind participants reported being unable to differentiate between noise and legitimate spatial events. In contrast, controls reported no problems. It is possible that the noise events were very salient to the blind users, lowering performance.

The results suggest that blind individuals perform differently than regularly sighted individuals. In fact, studies involving the visual cortex of the congenitally blind have revealed increased activation during sensorimotor tasks (Uhl, Franzen, Lindinger, Lang, & Deecke, 1991), language tasks (Bedny, Pascual-Leone, Dodell-Feder, Fedorenko, & Saxe, 2011), and auditory localization tasks in both congenitally blind (Collignon et al., 2011) and early-to-late blind individuals (Arnott, Thaler, Milne, Kish, & Goodale, 2013; Thaler, Arnott, & Goodale, 2011). A possible explanation is the areas typically thought to process visual information is repurposed due to experience-dependent plasticity in the brain. It has been shown that a blind auditory map (i.e. an auditory map not integrated with visual information) develops differently than a fully integrated visual-auditory map (King, Schnupp, & Thompson, 1998; Thinus-Blanc & Gaunet, 1997; Vercillo, Milne, Gori, & Goodale, 2015).The device uses an auditory map created by the regularly sighted experimenter. It is reasonable to expect, given a differently organized auditory map than the blind user, that the device's map is incompatible.

General Discussion

A key functional advantage of the neuromorphic camera is its ability to extract the luminance dynamics of a scene without any further image processing. This means that algorithms dependent on brightness changes and moving edges can run on less powerful hardware than might otherwise be required with a conventional frame-based camera. Indeed, although the computer used to render the spatial audio signal was a low-power netbook, rendering visual AER events into spatial auditory events took between 6 - 8 milliseconds. In fact, the visual-to-audio rendering ran well on a Raspberry Pi 3 single-board computer, although limitations in that board's native audio output made it unsuitable for the experiments described here.

It is likely that practice will improve performance, just like a new user of a prosthetic device (such as an arm or leg) may perform poorly at first, but will usually become proficient enough to replace the functionality of the lost limb. It is likely that even at very low displacements, a moving auditory stimulus would be unambiguous enough for the attention system to extract a cue about visual events.

The prototype device described here shows proof-of-concept and points toward a prosthetic device for patients with visual or attention orienting deficits. Currently, the system requires the user to wear a pair of headphones, which can interfere with the auditory system. As blind individuals are dependent on accurate auditory cues, this can pose a major problem. Passing through audio of the real world into the headphones would alleviate part of this problem, at the added expense of increased bulkiness of an added microphone. The most critical constraint relates to the rendering of real-world visual scenes with complex dynamics. When many visual events occur across the scene, the auditory augmented reality device becomes largely unusable. The auditory rendering fails to provide the cues needed to un-mix the complex auditory scene - a situation known as the Cocktail Party Problem. This might arise because several visual objects move simultaneously, but also occurs whenever the camera is panned across a stationary background. To handle complexity due to multiple visual events, future implementations will need to attach distinct auditory tags, possibly of varying pitch, as cues for the user to parse the acoustic scene. To handle egocentric motion of the camera, newer iterations of the system will require implementation of an event filter to remove background motion, probably at the cost of computational demands.

The DVS does include an inertial-measurement unit for this purpose and differentiating self-motion from moving targets is an active area of development (Delbruck, Villanueva, & Longinotti, 2014; Rueckauer & Delbruck, 2016). While the current implementation of the augmented auditory software does not include inputs from the inertial-measurement unit, it is a logical next step in development. When combined with visual grouping, the concept that the visual system tends to group related stimuli together (i.e. as uniform objects) (Vidal, Chaumon, O'Regan, & Tallon-Baudry, 2006; Y. Xu & Chun, 2007), the system described could potentially be used to solve extremely complex visual scenes and augment them to auditory space. The size of the camera proves to be a limiting factor, which can be attached to the body but is too large to be unnoticeable to the wearer. As neuromorphic technologies are currently in their infancy, newer, smaller, and faster technologies are likely to make neuromorphic vision sensors viable as wearable visual prosthetic devices.

Individualized auditory maps are probably necessary for blind users. Few problems occurred for the control group, whereas the stimulus noticeably confused blind users. Two blind participants found the stimulus too confusing and asked to withdraw from the study. Because the auditory map was calibrated using a visually and auditorily intact human, it may be necessary to perform similar calibration for users who do not have visual representations of space. Although future implementations would benefit from this, it will necessitate a large portion of the codebase be rewritten to dynamically change the mapping of the ITD and ILD cues as well as require a significantly longer calibration period per user.

Table 1. Paired-samples t-tests across all measures of displacement hit rate.

The t statistic and corresponding p value are shown. All measures were found to be significant, except for D175 & D250 and D250 & D325.

Pair	t statistic	p value
D100 & D175	-3.101	0.006164
D100 & D250	-3.806	0.001295
D100 & D325	-4.069	0.000721
D100 & D400	-5.754	0.000019
D175 & D250	-1.398	0.179058
D175 & D325	-2.282	0.034841
D175 & D400	-4.990	0.000095
D250 & D325	-.777	0.447102
D250 & D400	-4.251	0.000481
D325 & D400	-4.062	0.000732

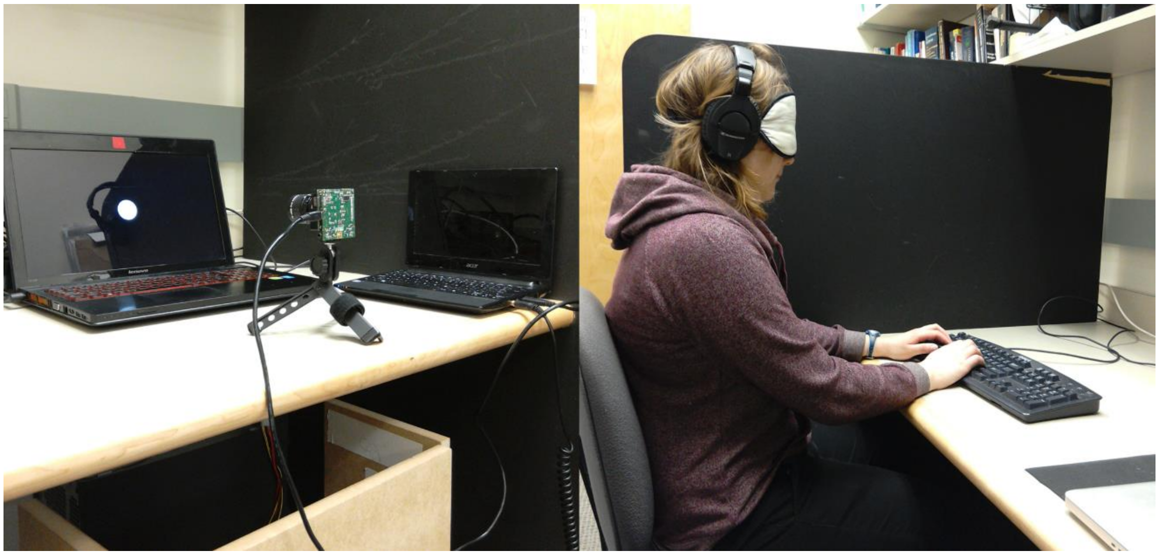

Figure 8. Experimental setup.

Participants were blindfolded and positioned to the right of the black wooden separator to prevent any visual information from the laptop computer. The netbook was running the auditory augmented reality software and providing the headphone audio signal. The camera was positioned 42 cm away from the screen of the laptop to capture the entirety of the screen. The participant wore a blindfold to reduce the likelihood of visual distraction. The individual in this manuscript has given written informed consent (as outlined in PLOS consent form) to publish these case details.

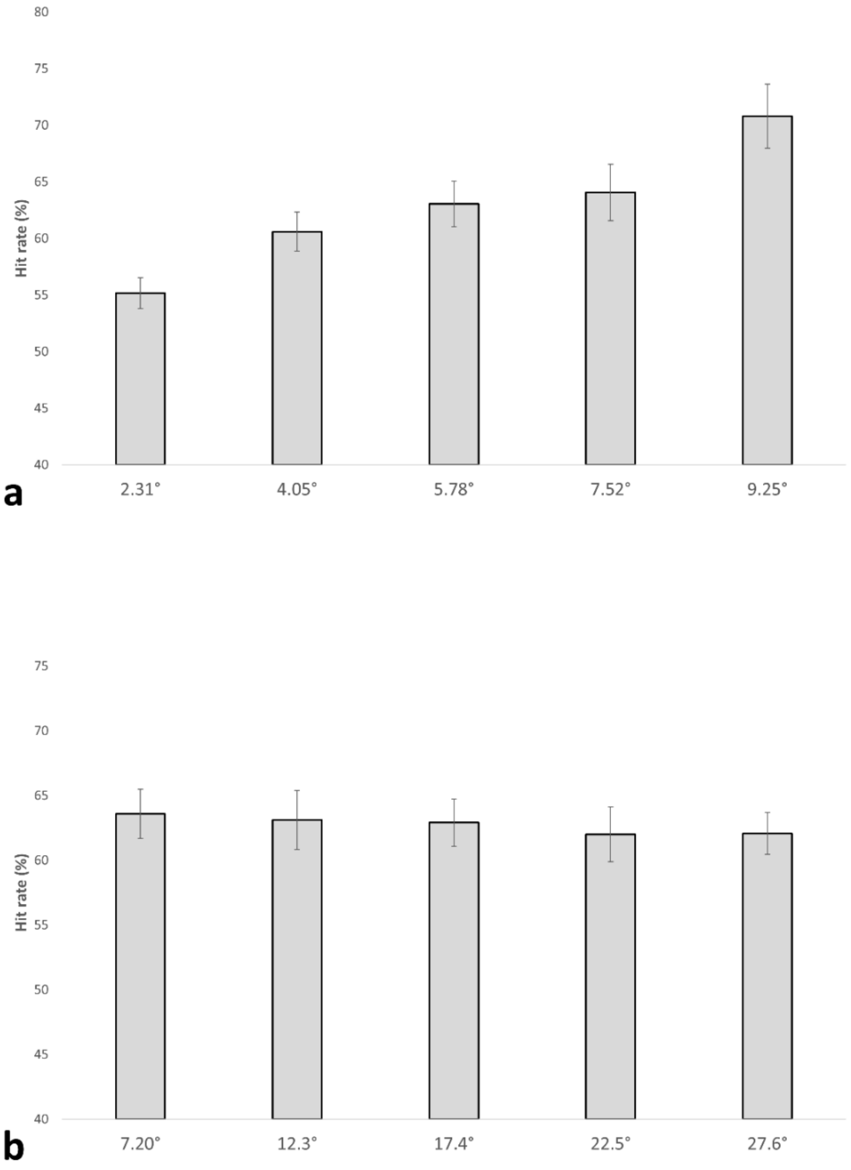

Figure 9. Hit rate during motion discrimination.

a. Bar graph showing the hit rate of each of the displacement trials. The horizontal axis contains the different displacements (in visual degrees) used. The respective matching displacements in pixels are 100, 175, 250, 325, and 400. The vertical axis is the average hit rate in percent. A significant effect of displacement was revealed, where more displacement was found to produce a higher hit rate. Standard error bars are shown.

b. Bar graph showing the hit rate for each of the displacement trials. The horizontal axis contains the different speeds (in visual degrees/second) used. The respective matching speeds in pixels/frame are 5, 8.75, 12.5, 16.25, and 20. The vertical axis is the average hit rate in percent. Standard error bars are shown.

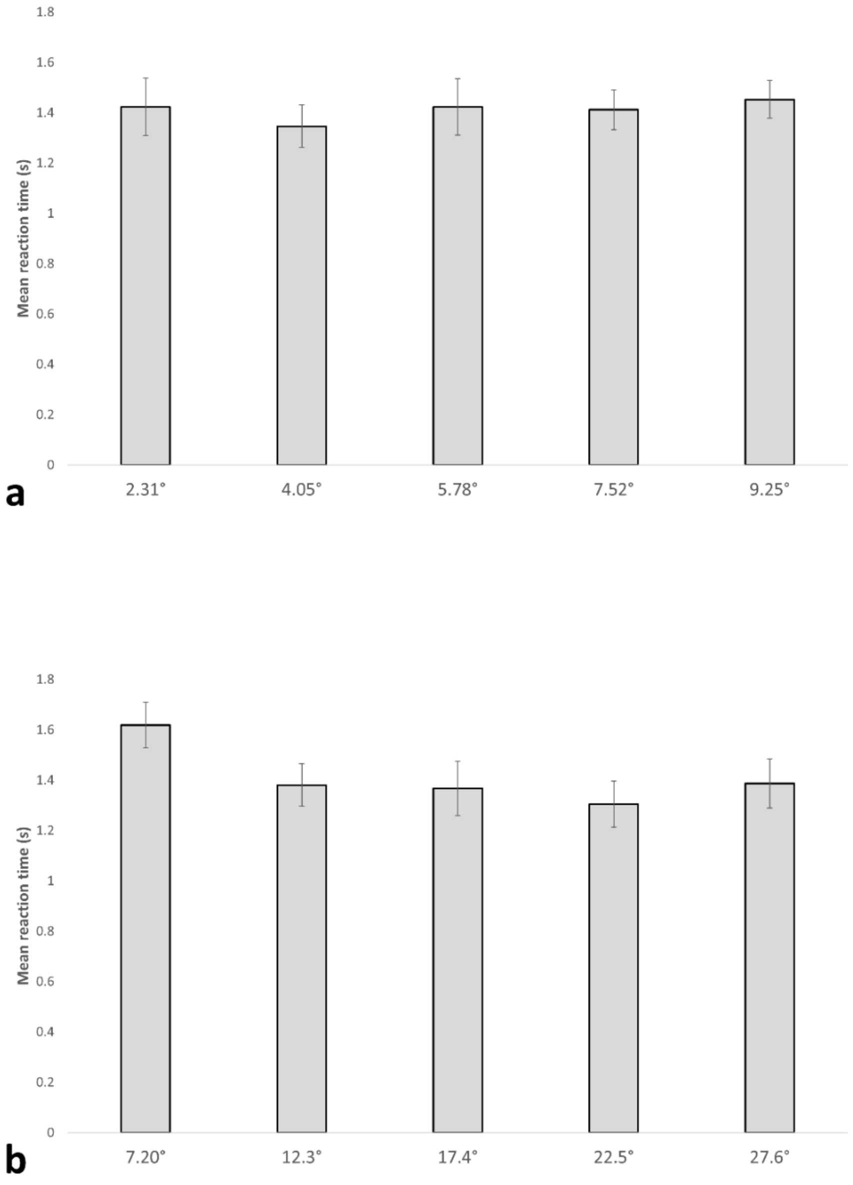

Figure 10. Mean reaction times during motion discrimination.

a. Bar graph showing the mean reaction time of each of the displacement trials. The horizontal axis contains the different displacements (in visual degrees) used. The respective matching displacements in pixels are 100, 175, 250, 325, and 400. The vertical axis contains the average reaction time in seconds. Standard error bars are shown.

b. Bar graph showing the reaction time for each of the speed trials. The horizontal axis contains the different speeds (in visual degrees/second) used. The respective matching speeds in pixels/frame are 5, 8.75, 12.5, 16.25, and 20. The vertical axis contains the average reaction time in seconds. Standard error bars are shown.

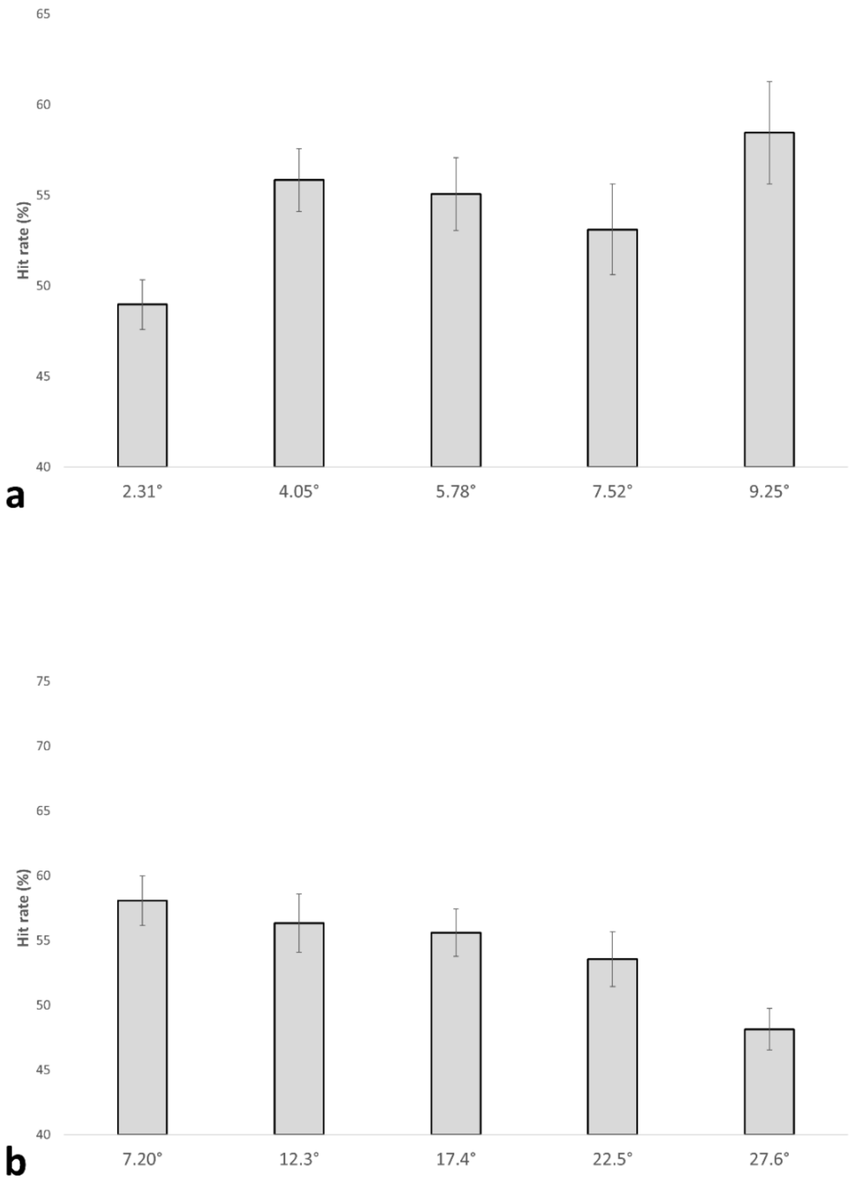

Figure 11. Hit rate during motion discrimination.

a. Bar graph showing the hit rate of each of the displacement trials for the blind participants. The horizontal axis contains the different displacements (in visual degrees) used. The respective matching displacements in pixels are 100, 175, 250, 325, and 400. The vertical axis is the average hit rate in percent. A significant effect of displacement was revealed, where more displacement was found to produce a higher hit rate. Standard error bars are shown.

b. Bar graph showing the hit rate for each of the displacement trials for the blind participants. The horizontal axis contains the different speeds (in visual degrees/second) used. The respective matching speeds in pixels/frame are 5, 8.75, 12.5, 16.25, and 20. The vertical axis is the average hit rate in percent. Standard error bars are shown.

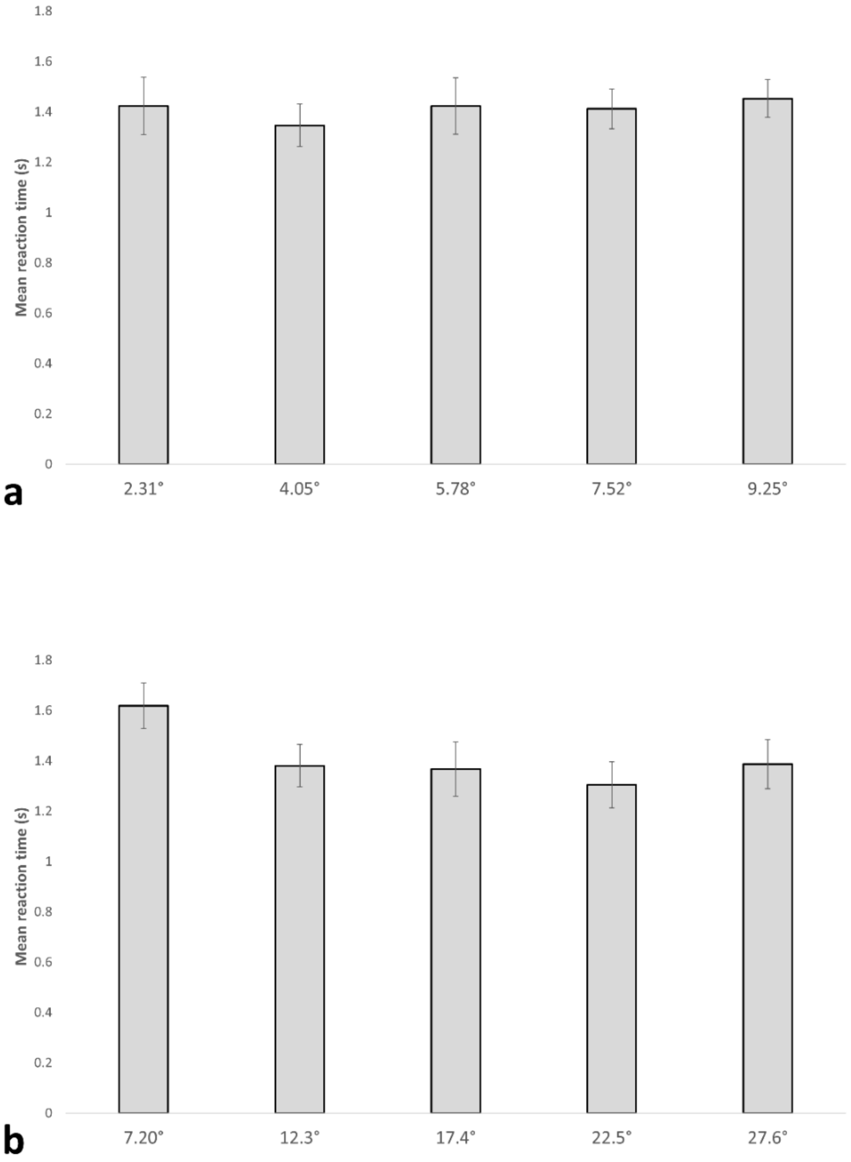

Figure 12. Mean reaction times during motion discrimination.

a. Bar graph showing the mean reaction time of each of the displacement trials for the blind participants. The horizontal axis contains the different displacements (in visual degrees) used. The respective matching displacements in pixels are 100, 175, 250, 325, and 400. The vertical axis contains the average reaction time in seconds. Standard error bars are shown.

b. Bar graph showing the reaction time for each of the speed trials for the blind participants. The horizontal axis contains the different speeds (in visual degrees/second) used. The respective matching speeds in pixels/frame are 5, 8.75, 12.5, 16.25, and 20. The vertical axis contains the average reaction time in seconds. Standard error bars are shown.

CHAPTER 4. Comparable devices, overall discussion, and future directions

Comparable auditory augmented reality systems

Auditory augmented reality could dramatically improve the safety and lifestyle of visually impaired people, yet surprisingly few attempts have been made to create auditory augmented reality systems in the past. For example, the SeeHear system (Cao et al., 1992; Nielsen et al., 1987) was created to convert visual events into spatial auditory events. The system was a more traditional effort to emulate the delay lines thought to be present in the auditory system. The system acquired visual input through a matrix of photoreceptors, where the intensity at each receptor was calculated. Intensities were propagated along a delay line to approximate the ITD cue, and then played through a pair of headphones. The resulting percept could map a moving visual stimulus to an auditory object moving through space. While the system apparently worked, efficacy was never assessed via human trials. It is possible the small array of photoreceptors (15x11) used to mimic vision may not have been large enough to produce convincing stimuli. As a proof-of-concept system, the SeeHear made great strides in field of auditory augmented reality and gave evidence that it is possible to convert visual events to auditory events with preserved spatial characteristics.

The VIS2SOUND system (Morillas et al., 2008) worked using a similar philosophy to the SeeHear system. The VIS2SOUND system was designed as an aid for the visually-impaired to enhance visual information in a scene. Some key differences from the SeeHear and the device described in this thesis are the use of field-programmable gate arrays (FPGAs) and conventional video camera inputs. The FPGAs, which process in parallel and are thus extremely efficient, filter the video stream spatially and temporally to extract features. The features are then fed into a localized sound generator which uses HRTF modeling to create biologically plausible sounds. Notably, customization of the system allows for filtered outputs to generate pulse codes for use with a neuro-stimulation device though this was never assessed. With the possibility to use two cameras simultaneously, depth estimation could also be achieved using this system. Like the SeeHear, the system was never assessed on a human population, so the psychophysical usefulness of such a system is not known. While the system described is technologically impressive, it was not designed for portable use as evidenced by its bulky size. Rather, it was designed to demonstrate the advances in technology that allow the use of conventional hardware to create convincing spatial audio in near real-time.

The vOICe system (Meijer, 2017) uses regular frame-based cameras and edge-detection filtering to extract salient or potentially dangerous features of a space (e.g. the corner of the table) to augment into auditory space. The filtered image file is sequentially scanned from left-to-right and each pixel is encoded as a sound. The resulting auditory scene is very complex and as such requires intensive training to use. From the bottom-up, the system was designed for use by the severely visually-impaired and is very portable. In fact, the system can be used as an app on any modern smartphone with a pair of headphones. Use of the system is supposed to provide additional cues to the user about their environment, which are largely derived from the edges around objects. For example, an exit door will have a sharp edge around it because it is discontinuous with the rest of the wall. Because the system takes in the order of a few seconds to process each frame, fine motion sensitivity is not possible using vOICe. It is not especially useful for detecting scene changes, but rather locating specific features within a scene such as the edge of a table or locating some specific object in which the general shape is known.

Overall, very few bona fide attempts have been made at creating auditory augmented reality systems. Of the systems described, none are identical in function to the system described in this thesis. The vOICe system is designed to be portable, but given the amount of training necessary for use, is not an intuitive system. The VIS2SOUND system demonstrates the processing power available using FPGA technology, where video streams can be processed and spatial audio generated in real-time at up to 60 frames per second. A major limitation of this system is the lack of portability, although a dedicated device could be envisioned that might someday incorporate cameras and FPGA hardware into a relatively small printed circuit board.. The SeeHear system mimics auditory delay lines, which are thought to calculate the ITD cue used by the brain. Limited conduction velocities will naturally produce delay lines, since longer lines will take longer for a signal to travel. If the goal of the system is to imitate the biological system it derived from, this is an excellent choice. If the goal is to be a useful prosthetic device, it is valid to simply generate the output of the delay lines; either system creates a valid ITD cue. Unlike the system described in this thesis, the SeeHear uses a silicon chip that contains both the photoreceptors and delay lines. It should be noted the SeeHear has much lower spatial resolution than a neuromorphic retina. My system uses a DAVIS 240B to measure visual activity and a low-cost laptop to encode these visual events as spatial audio in near real-time. For reference, this is the first time a neuromorphic sensor has been used to augment information into an auditory scene. It should be noted that none of the other systems have been formally tested on a human population to determine efficacy. A neuromorphic sensor based auditory augmented reality system has many advantages, namely the ability to generate convincing localization cues without large amounts of engineering (e.g. SeeHear, vOICe, VIS2SOUND) or computational power (e.g. VIS2SOUND, vOICe).

When compared to other fields of engineering, neuromorphic sensors are still in their infancy. In the early 1990’s, the first chips mimicking biological function using very-large-scale integration circuits were demonstrated (Mead, 1990). Mead argued that the efficiency achieved by biological systems was not matched by the current microprocessors of the time. The difference, he argued, is that the brain operates on different principles than what digital systems used. Further, he argued that adaptive analog technology (i.e. neuromorphic sensors) can represent information by relative analog signals rather than discrete digital values, resulting in increased efficiency. A consequence of developing new technologies is the increased cost of adoption. A cheap computer webcam can be purchased for around $50, whereas the cost of a neuromorphic retina is around $8000, making the retina 160 times the price of a webcam.

Discussion limitations and future development

The auditory augmented reality device described here shows good promise as a useful tool for localizing purely visual events using generated auditory cues, at least in regularly sighted individuals. As a purely prototype device, a number of straightforward limitations exist because of first approximations made in designing the system. When designing the auditory space bounds for the system, my own perception of auditory space was used. It is possible that my auditory map is misaligned to the congenitally blind auditory map. The lateral intraparietal area (LIP) is thought to integrate multimodal information such as auditory, visual and somatosensory signals (Andersen, 1997). As congenitally blind individuals do not have visual input, they probably have a different map organization. This suggests that the regularly sighted auditory map may be different than a blind auditory map. Visual and auditory acuity were never properly assessed in the regularly sighted group, making it impossible to directly equate their spatial maps to my own. However, it appeared that the system worked as expected for the controls; performance improved with more displacement.

Currently, the system exists as several bulky parts: the DAVIS 240B camera, a small laptop, a pair of headphones, and a USB cable for the camera. The system as a research and testing platform works well, but as a practical device to be used by a consumer, it is not ideal. Ideally, the entire system would be contained to a single unit; a neuromorphic retina that can be attached to the body connected to a battery-powered computing board. An embedded version of the DAVIS camera exists and is currently under active development (Conradt, Berner, Cook, & Delbruck, 2009), which will allow for fully modular versions of the current system to exist as a single system-on-chip board. Portability is very important to consider, especially for the blind. Anecdotally, the blind participants in the study were very independent and appreciated any attempt to improve their quality of life through increased freedom of movement. As neuromorphic technology develops,

A fundamental limitation to the technology as currently implemented is that the neuromorphic sensor is extremely sensitive to self-motion. Self-motion, or egocentric motion, generates events on the sensor that are indistinguishable from allocentric, or object motion events. This is a problem because it is not currently possible to distinguish self-motion from important visual events. The DAVIS 240B has a built-in inertial measurement unit (IMU), which is used to measure linear and gyroscopic acceleration in three axes. Using data generated by the IMU, it is possible to filter events generated by self-motion and be left with valid allocentric motion events. This approach involves additional filtering to reduce the number of self-motion events, necessitating additional computing power. The current implementation of the system did not account for self-motion because it was assumed that the neuromorphic retina would be stationary for the duration of the experiments. If the camera is moved while the auditory cue generation algorithm is active, the resulting auditory scene is very complex with many spatially incoherent clicks, and nearly impossible to resolve. Self-motion filters exist for the java-based AER program, *jAER*, but they are computationally expensive to run and require powerful hardware. Because the device is best suited as a portable prosthetic attentional device, future implementations would benefit immensely from a self-motion filter probably at the cost of the added computing cycles needed to resolve the scene. However, if one imagines that the camera would be head-mounted, then it would be usable provided that the listener holds the head still with respect to the allocentric world.

To make the device as simple and accessible as possible, the auditory stimulus generated is identical regardless of the size or shape of the object used. An auditory click is generated for each visual event in the scene. No information about the size or shape of the visual object is preserved. Size could be encoded by using the number of active pixels in the scene to calculate the amplitude of the signal. Shape is a much more difficult problem to solve and is outside the scope of this thesis. It is likely that including information about shape or size would necessitate intensive training to use the device, much like the vOICe system. Because the goal of this thesis is not to create a production-ready device, these changes would probably impose restrictions on the design of the auditory stimulus. A simple click was used because it is a broadband noise signal, which tend to be easier to localize than pure tones (Blauert, 1997; Middlebrooks & Green, 1991). While simple to generate, a common complaint amongst participants was the unpleasantness of the sound itself over the course of the experiments. Different tones could be used to denote the size or shape of the object, for example.

Depth is not currently calculated or accounted for in the auditory signal. A sense of depth can be useful, especially in situations where knowing distance is critical. For example, when crossing the street, it is useful to know how far away vehicles are. Filters calculating depth have been developed for the neuromorphic sensors, so implementation is not a barrier.

Generating depth cues necessitates at least two cameras, but because the cameras are so expensive, this is not a priority for implementation.

Wearing headphones is not an ideal way of interacting with the world. Headphones tend to passively block out most of the ambient noise of an environment, and can be hazardous when navigating areas that require acute audiovisual attention such as a street crosswalk. Headphones were used to create the auditory stimulus because they are an easy way to pass auditory information to the user. Unlike free field solutions, virtual objects can be created using headphones. To include the auditory information in the environment, a microphone could be used to mix the audio from the device and the environment. The mixed audio would allow for a truly augmented auditory experience; auditory signals from the device or the environment could be used to orient attention. State-of-the-art hearing aid technology might be useful in this field of research, as recent devices are nearly “acoustically transparent”.

Personalization of the device, such as individually calibrating the auditory map, might substantially improve the quality of the information given to the user. The device currently uses an auditory map derived from a sighted individual, a factor probably impacting the performance of blind users lower than if the map was individualized. The mapping of ITD and ILD cues onto perceived acoustic space probably approximates the average sighted individual better than the average blind individual. The results of Experiment 1 supported this speculation; sighted participants could use the device as evidenced by the main effect of displacement. These results suggest the current implementation of the device might not be best suited for the visually impaired population, not because of a flaw in the design but rather due to how it is calibrated.

Conclusion

For the visually-impaired, there are very few options to interpret visual information. Existing technologies, such as the white cane, work well for detecting rigid stationary objects but do not allow for perception of non-physical objects such as those on a computer screen. Auditory cues are extremely important to the blind, and as traditionally noisy technologies such as automobiles become quieter (e.g. electric cars), real world auditory might become dangerously sparse. A prosthetic attention system that adds useful localization and motion cues to the auditory scene is a beneficial invention particularly for the blind or visually impaired community.

In this thesis, an auditory augmented reality device using neuromorphic retina technology to generate valid interaural difference cues has been described and assessed. Using purely visual information, auditory cues were generated to encode object position and motion. Thus, stimulus events that were visually salient but acoustically silent could be detected by rendering them into the auditory sense. Software was written to emulate the cues that the brain uses to calculate azimuthal position in space. These cues are known as the interaural time and level difference cues. Psychophysical efficacy was assessed in two key populations: the regularly sighted and the congenitally blind. Both populations could use the device to detect with high accuracy the onset of a visual object using auditory cues. However, regularly sighted individuals could also use the device to glean information about object position and motion in space, whereas the blind participants were not able to use the device in that way. Changes to the way the device operates, such calibrating ITD and ILD cues to better reflect the auditory map of the user, or changes to the sound encoding object position are improvements that can be made.

Future implementations will likely benefit from the active development of smaller, faster, lighter, and cheaper neuromorphic sensors.

Bibliography

Andersen, R. A. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 352*(1360), 1421-1428.

Arnott, S. R., Thaler, L., Milne, J. L., Kish, D., & Goodale, M. A. (2013). Shape-specific activation of occipital cortex in an early blind echolocation expert. *Neuropsychologia, 51*(5), 938-949.

Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London B: Biological Sciences, 168*(1011), 158-180.

Batteau, D. W. (1968). Listening with the naked ear. *Neuropsychology of Spatially Oriented Behavior. Dorsey Press, Homewood, IL. USA*.

Bedny, M., Pascual-Leone, A., Dodell-Feder, D., Fedorenko, E., & Saxe, R. (2011). Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences, 108*(11), 4429-4434.

Blauert, J. (1969). Sound localization in the median plane. *Acta Acustica united with Acustica, 22*(4), 205-213.

Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*: MIT press.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annu. Rev. Neurosci., 28*, 157-189.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision, 10*, 433-436.

Brand, A., Behrend, O., Marquardt, T., McAlpine, D., & Grothe, B. (2002). Precise inhibition is essential for microsecond interaural time difference coding. *Nature, 417*(6888), 543-547.

Burger, J. (1958). Front-back discrimination of the hearing systems. *Acta Acustica united with Acustica, 8*(5), 301-302.

Butler, R. (1969). Monaural and binaural localization of noise bursts vertically in median sagittal plane. *Journal of Auditory research, 9*(3), 230-235.

Butler, R. A., & Helwig, C. C. (1983). The spatial attributes of stimulus frequency in the median sagittal plane and their role in sound localization. *American journal of otolaryngology, 4*(3), 165-173.

Cao, Y., Mattisson, S., & Bjork, C. (1992, 21-23 Sept. 1992). *SeeHear System: A New Implementation.* Paper presented at the Solid-State Circuits Conference, 1992. ESSCIRC '92. Eighteenth European.

Carr, C., & Konishi, M. (1990). A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience, 10*(10), 3227-3246.

Collignon, O., Vandewalle, G., Voss, P., Albouy, G., Charbonneau, G., Lassonde, M., & Lepore, F. (2011). Functional specialization for auditory–spatial processing in the occipital cortex of congenitally blind humans. *Proceedings of the National Academy of Sciences, 108*(11), 4435-4440.

Conradt, J., Berner, R., Cook, M., & Delbruck, T. (2009). *An embedded aer dynamic vision sensor for low-latency pole balancing.* Paper presented at the Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on.

Delbruck, T., Villanueva, V., & Longinotti, L. (2014). *Integration of dynamic vision sensor with inertial measurement unit for electronically stabilized event-based vision.* Paper presented at the 2014 IEEE International Symposium on Circuits and Systems (ISCAS).

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1-47.

Franconeri, S. L., & Simons, D. J. (2003). Moving and looming stimuli capture attention. *Perception & psychophysics, 65*(7), 999-1010.

Gardner, M. B., & Gardner, R. S. (1973). Problem of localization in the median plane: effect of pinnae cavity occlusion. *The Journal of the Acoustical Society of America, 53*(2), 400-408.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual review of neuroscience, 26*(1), 181-206.

Grantham, D. W. (1985). Auditory spatial resolution under static and dynamic conditions. *The Journal of the Acoustical Society of America, 77*(S1), S50-S50.

Grantham, D. W. (1989). Motion aftereffects with horizontally moving sound sources in the free field. *Attention, Perception, & Psychophysics, 45*(2), 129-136.

Grantham, D. W. (1997). Auditory motion perception: Snapshots revisited. *Binaural and spatial hearing in real and virtual environments*, 295-313.

Grothe, B., & Sanes, D. H. (1994). Synaptic inhibition influences the temporal coding properties of medial superior olivary neurons: an in vitro study. *Journal of Neuroscience, 14*(3), 1701-1709.

Indiveri, G., & Horiuchi, T. K. (2011). Frontiers in neuromorphic engineering. *Frontiers in Neuroscience, 5*.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of comparative and physiological psychology, 41*(1), 35.

Jin, C., Corderoy, A., Carlile, S., & van Schaik, A. (2004). Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America, 115*(6), 3124-3141.

Joris, P. X., Smith, P. H., & Yin, T. C. (1998). Coincidence detection in the auditory system: 50 years after Jeffress. *Neuron, 21*(6), 1235-1238.

Kavanagh, G. L., & Kelly, J. B. (1992). Midline and lateral field sound localization in the ferret (Mustela putorius): contribution of the superior olivary complex. *J Neurophysiol, 67*(6), 1643-1658.

King, A. J., Schnupp, J. W., & Thompson, I. D. (1998). Signals from the superficial layers of the superior colliculus enable the development of the auditory space map in the deeper layers. *Journal of Neuroscience, 18*(22), 9394-9408.

Kolb, H. (1991). The neural organization of the human retina. *Principles and practices of clinical electrophysiology of vision*, 25-52.

Kolb, H., Linberg, K. A., & Fisher, S. K. (1992). Neurons of the human retina: a Golgi study. *Journal of Comparative Neurology, 318*(2), 147-187.

Lewis, J. W., Beauchamp, M. S., & DeYoe, E. A. (2000). A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex, 10*(9), 873-888.

Lichtsteiner, P., Posch, C., & Delbruck, T. (2008). A 128$\ times $128 120 dB 15$\ mu $ s Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE journal of solid-state circuits, 43*(2), 566-576.

Liu, S.-C., & Delbruck, T. (2010). Neuromorphic sensory systems. *Current Opinion in Neurobiology, 20*(3), 288-295. doi:http://dx.doi.org/10.1016/j.conb.2010.03.007

Liu, S.-C., Delbruck, T., Indiveri, G., Douglas, R., & Whatley, A. (2015). *Event-based neuromorphic systems*: John Wiley & Sons.

Mahowald, M. (1994). *An analog VLSI system for stereoscopic vision* (Vol. 265): Springer Science & Business Media.

Martinez, C. S., & Hwang, F. (2015). *TESSA: Toolkit for Experimentation with Multimodal Sensory Substitution and Augmentation*. Paper presented at the Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, Republic of Korea.

Masterton, B., Diamond, I. T., Harrison, J., & Beecher, M. D. (1967). Medial superior olive and sound localization. *Science, 155*(3770), 1696-1697.

Maunsell, J. H., Ghose, G. M., Assad, J. A., McADAMS, C. J., Boudreau, C. E., & Noerager, B. D. (1999). Visual response latencies of magnocellular and parvocellular LGN neurons in macaque monkeys. *Visual neuroscience, 16*(01), 1-14.

Mead, C. (1990). Neuromorphic electronic systems. *Proceedings of the IEEE, 78*(10), 1629-1636.

Meijer, P. B. L. (2017). Augmented Reality for the Totally Blind.

Meissirel, C., Wikler, K. C., Chalupa, L. M., & Rakic, P. (1997). Early divergence of magnocellular and parvocellular functional subsystems in the embryonic primate visual system. *Proceedings of the National Academy of Sciences, 94*(11), 5900-5905.

Merigan, W., Byrne, C., & Maunsell, J. (1991). Does primate motion perception depend on the magnocellular pathway? *Journal of Neuroscience, 11*(11), 3422-3429.

Merigan, W., Katz, L. M., & Maunsell, J. (1991). The effects of parvocellular lateral geniculate lesions on the acuity and contrast sensitivity of macaque monkeys. *Journal of Neuroscience, 11*(4), 994-1001.

Merigan, W. H., & Eskin, T. A. (1986). Spatio-temporal vision of macaques with severe loss of P β retinal ganglion cells. *Vision Research, 26*(11), 1751-1761.

Merigan, W. H., & Maunsell, J. H. (1990). Macaque vision after magnocellular lateral geniculate lesions. *Visual neuroscience, 5*(04), 347-352.

Meyer, G., & Wuerger, S. (2001). Cross-modal integration of auditory and visual motion signals. *NeuroReport, 12*(11), 2557-2560.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual review of psychology, 42*(1), 135-159.

Mikami, A., Newsome, W. T., & Wurtz, R. H. (1986). Motion selectivity in macaque visual cortex. I. Mechanisms of direction and speed selectivity in extrastriate area MT. *J Neurophysiol, 55*(6), 1308-1327.

Milner, D., & Goodale, M. (2006). The visual brain in action. In: Oxford University Press.

Mogdans, J., & Knudsen, E. I. (1994). Representation of interaural level difference in the VLVp, the first site of binaural comparison in the barn owl's auditory system. *Hearing research, 74*(1), 148-164.

Moiseff, A., & Konishi, M. (1981). Neuronal and behavioral sensitivity to binaural time differences in the owl. *Journal of Neuroscience, 1*(1), 40-48.

Moiseff, A., & Konishi, M. (1983). Binaural characteristics of units in the owl's brainstem auditory pathway: precursors of restricted spatial receptive fields. *Journal of Neuroscience, 3*(12), 2553-2562.

Moore, C., Casseday, J., & Neff, W. (1974). Sound localization: the role of the commissural pathways of the auditory system of the cat. *Brain Research, 82*(1), 13-26.

Morillas, C., Cobos, J. P., Pelayo, F. J., Prieto, A., & Romero, S. (2008). *VIS2SOUND on reconfigurable hardware.* Paper presented at the Reconfigurable Computing and FPGAs, 2008. ReConFig'08. International Conference on.

Neuhoff, J. (2004). Auditory motion and localization. *Ecological psychoacoustics*, 87-111.

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience, 8*(6), 2201-2211.

Nielsen, L., Mahowald, M., & Mead, C. (1987). SeeHear.

Niemi, P., & Näätänen, R. (1981). Foreperiod and simple reaction time. *Psychological bulletin, 89*(1), 133.

Nowlan, S. J., & Sejnowski, T. J. (1995). A selection model for motion processing in area MT of primates. *Journal of Neuroscience, 15*(2), 1195-1214.

Oldfield, S. R., & Parker, S. P. (1984). Acuity of sound localisation: a topography of auditory space. II. Pinna cues absent. *Perception, 13*(5), 601-617.

Pain, M. T., & Hibbs, A. (2007). Sprint starts and the minimum auditory reaction time. *J Sports Sci, 25*(1), 79-86. doi:10.1080/02640410600718004

Pascolini, D., & Mariotti, S. P. (2012). Global estimates of visual impairment: 2010. *British Journal of Ophthalmology, 96*(5), 614-618.

Pecka, M., Brand, A., Behrend, O., & Grothe, B. (2008). Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic inhibition. *Journal of Neuroscience, 28*(27), 6914-6925.

Perrott, D. R., Costantino, B., & Ball, J. (1993). Discrimination of moving events which accelerate or decelerate over the listening interval. *The Journal of the Acoustical Society of America, 93*(2), 1053-1057.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual review of neuroscience, 13*(1), 25-42.

Posner, M. I., Walker, J. A., Friedrich, F. J., & Rafal, R. D. (1984). Effects of parietal injury on covert orienting of attention. *Journal of Neuroscience, 4*(7), 1863-1874.

Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: a systematic survey. *IEEE transactions on image processing, 14*(3), 294-307.

Rauschecker, J. P., & Tian, B. (2000). Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences, 97*(22), 11800-11806.

Rayleigh, L. (1907). XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 13*(74), 214-232.

Roffler, S. K., & Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane. *The Journal of the Acoustical Society of America, 43*(6), 1255-1259.

Rueckauer, B., & Delbruck, T. (2016). Evaluation of Event-Based Algorithms for Optical Flow with Ground-Truth from Inertial Measurement Sensor. *Frontiers in Neuroscience, 10*, 176. doi:10.3389/fnins.2016.00176

Sandel, T., Teas, D., Feddersen, W., & Jeffress, L. (1955). Localization of sound from single and paired sources. *The Journal of the Acoustical Society of America, 27*(5), 842-852.

Schleich, P., Nopp, P., & D'haese, P. (2004). Head shadow, squelch, and summation effects in bilateral users of the MED-EL COMBI 40/40+ cochlear implant. *Ear and hearing, 25*(3), 197-204.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*(5), 743-761.

Smith, P. H. (1995). Structural and functional differences distinguish principal from nonprincipal cells in the guinea pig MSO slice. *J Neurophysiol, 73*(4), 1653-1667.

Smith, P. H., Joris, P. X., & Yin, T. C. (1993). Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat: evidence for delay lines to the medial superior olive. *Journal of Comparative Neurology, 331*(2), 245-260.

Snowden, R. J. (2002). Visual attention to color: Parvocellular guidance of attentional resources? *Psychological Science, 13*(2), 180-184.

Stevens, S. S., & Newman, E. B. (1936). The localization of actual sources of sound. *The American Journal of Psychology, 48*(2), 297-306.

Thaler, L., Arnott, S. R., & Goodale, M. A. (2011). Neural correlates of natural human echolocation in early and late blind echolocation experts. *PLoS ONE, 6*(5), e20162.

Thinus-Blanc, C., & Gaunet, F. (1997). Representation of space in blind persons: vision as a spatial sense? *Psychological bulletin, 121*(1), 20.

Tollin, D. J. (2003). The lateral superior olive: a functional role in sound source localization. *The neuroscientist, 9*(2), 127-143.

Uhl, F., Franzen, P., Lindinger, G., Lang, W., & Deecke, L. (1991). On the functionality of the visually deprived occipital cortex in early blind persons. *Neurosci Lett, 124*(2), 256-259. doi:http://dx.doi.org/10.1016/0304-3940(91)90107-5

Ungerleider, L., & Mishkin, M. (1982). Analysis of Visual Behavior, eds Ingle DJ, Goodale MA, Mansfield RJW.

Ungerleider, L. G., & Haxby, J. V. (1994). 'What'and 'where'in the human brain. *Current Opinion in Neurobiology, 4*(2), 157-165.

Vercillo, T., Milne, J. L., Gori, M., & Goodale, M. A. (2015). Enhanced auditory spatial localization in blind echolocators. *Neuropsychologia, 67*, 35-40.

Vidal, J. R., Chaumon, M., O'Regan, J. K., & Tallon-Baudry, C. (2006). Visual grouping and the focusing of attention induce gamma-band oscillations at different frequencies in human magnetoencephalogram signals. *Journal of Cognitive Neuroscience, 18*(11), 1850-1862.

Warren, J. D., Zielinski, B. A., Green, G. G. R., Rauschecker, J. P., & Griffiths, T. D. (2002). Perception of Sound-Source Motion by the Human Brain. *Neuron, 34*(1), 139-148. doi:http://dx.doi.org/10.1016/S0896-6273(02)00637-2

Wenzel, E. M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America, 94*(1), 111-123.

Wightman, F. L., & Kistler, D. J. (1989). Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America, 85*(2), 868-878.

Woodworth, R. (1937). Experimental Psychology. New York: Holt, 1938. *Department of Psychology Dartmouth College Hanover, New Hampshire*.

Xu, S., Li, Z., & Salvendy, G. (2007). *Individualization of head-related transfer function for three-dimensional virtual auditory display: a review.* Paper presented at the International Conference on Virtual Reality.

Xu, Y., & Chun, M. M. (2007). Visual grouping in human parietal cortex. *Proceedings of the National Academy of Sciences, 104*(47), 18766-18771.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and sleective attention: Evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 601 - 620.

Yin, T., & Chan, J. (1990). Interaural time sensitivity in medial superior olive of cat. *J Neurophysiol, 64*(2), 465-488.

Zotkin, D., Hwang, J., Duraiswaini, R., & Davis, L. S. (2003). *HRTF personalization using anthropometric measurements.* Paper presented at the Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.