Investigating Compensatory Mechanisms for Sound Localization: Visual Cue Integration and the Precedence Effect

by

Christopher Montagne

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved July 2015 by the
Graduate Supervisory Committee:

Yi Zhou, Co-Chair
Christopher Buneo, Co-Chair
William Yost

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Sound localization can be difficult in a reverberant environment. Fortunately listeners can utilize various perceptual compensatory mechanisms to increase the reliability of sound localization when provided with ambiguous physical evidence. For example, the directional information of echoes can be perceptually suppressed by the direct sound to achieve a single, fused auditory event in a process called the precedence effect (Litovsky et al., 1999). Visual cues also influence sound localization through a phenomenon known as the ventriloquist effect. It is classically demonstrated by a puppeteer who speaks without visible lip movements while moving the mouth of a puppet synchronously with his/her speech (Gelder and Bertelson, 2003). If the ventriloquist is successful, sound will be “captured” by vision and be perceived to be originating at the location of the puppet. This thesis investigates the influence of vision on the spatial localization of audio-visual stimuli. Participants seated in a sound-attenuated room indicated their perceived locations of either ISI or level-difference stimuli in free field conditions. Two types of stereophonic phantom sound sources, created by modulating the inter-stimulus time interval (ISI) or level difference between two loudspeakers, were used as auditory stimuli. The results showed that the light cues influenced auditory spatial perception to a greater extent for the ISI stimuli than the level difference stimuli. A binaural signal analysis further revealed that the greater visual bias for the ISI phantom sound sources was correlated with the increasingly ambiguous binaural cues of the ISI signals. This finding suggests that when sound localization cues are unreliable, perceptual decisions become increasingly biased towards vision for finding a sound source. These results support the cue saliency theory underlying cross-modal bias and extend this theory to include stereophonic phantom sound sources.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Chapter 1 Introduction

Critical for the survival of many species, one fundamental task of the auditory system is to determine the spatial locations of acoustic stimuli. While the activities of sensory receptors in the somatosensory and visual systems are directly correlated with spatial location, those for the auditory system do not. To signal the location of a sound source, the auditory system computes centrally the timing and level disparity information between two ear-canal signals for sensing horizontal direction and monaural spectral cues for sensing vertical direction. Furthermore, these monaural and binaural cues must be integrated across frequency channels in the central auditory stages to realize a space map. This process can become quite noisy, especially in reverberant environments, and result in ambiguously computed binaural cues (Mcfadden et al., 1973; Tollin, 2003; Yin and Chan, 1990; Zurek, 1980). Fortunately, the brain can use various perceptual compensatory mechanisms to increase the reliability of sound localization judgments when provided ambiguous physical evidence. For example, short latency echoes can be perceptually fused into a single perceived direction of sound in a process called the precedence effect (Brown et al., 2015; Litovsky et al., 1999). Also, redundant sensory information is available where the receptive fields of the auditory and visual systems overlap. These redundancies can be utilized in what is referred to as the ventriloquist effect, a phenomenon in which visual cues that are temporally or contextually synchronous with auditory cues ‘capture’ the location of a sound (Gelder and Bertelson, 2003; Jackson, 1953).

While there is an extensive amount of literature on the effects of audio-visual (AV) cue integration on sound localization, most of these studies have assumed that unimodal sound localization is a deterministic process. That is, sound will always be perceived as originating from its source’s physical location. This assumption does not account for the localization of stereophonic phantom sound sources designed to be perceived at locations with varying reliability. With growing amounts of evidence indicating that sound localization depends on a probabilistic inference (Abe et al., 2010; Blauert, 1997; Wendt, 1963; Willert et al., 2006), it is necessary to study AV integration from the perspective of the ‘Bayesian coding hypothesis’, in which the brain represents sensory information in the form of probabilistic distributions rather than deterministic judgments (Knill and Pouget, 2004).

This thesis investigates the influence of sound localization cue saliency on AV integration. The precedence effect and sound localization cues were exploited to create stereophonic phantom sound sources with variable cue saliency. Participants were recruited to complete a sound localization task using these phantom sound sources paired with and without temporally aligned flashing lights. The degree of cross-modal bias was calculated by comparing audio-only to audio-visual localization results. The sound localization cue saliency of the auditory stimuli was assessed through an acoustic analysis and then compared to the degree of cross-modal bias from the localization results. Overall, the results provide strong support for the Bayesian coding hypothesis in AV integration, as visual cues increasingly bias sound localization with decreasing sound localization cue ambiguity.

Chapter 2 Background and Motivation

2.1 Sound Localization Cues for a Single Sound Source

Binaural cues are primarily used for localizing a sound source on the horizontal plane. The two binaural cues for sound localization are interaural level differences (ILDs) and interaural time differences (ITDs). Wavelengths approximately equal to or shorter than the diameter of the head create a shadowing effect by which sound energies received by the ear furthest from a sound source are attenuated, resulting in an ILD. For lower frequency sounds, less than 2000 Hz, the wavelengths become larger than the diameter of the head and create a detectable ITD due to the travel length difference between signals to the left and right ears. ITDs and ILDs are first processed and converted to a spike-rate code in two brainstem structures - the medial superior olive (MSO) and lateral superior olive (LSO), respectively. The ITD discrimination threshold for humans can be as low as 10 μs and 1-2 decibels for ILD thresholds, allowing for sound localization angle discrimination of up to 1-2 degrees (Blauert, 1997; Klumpp, 1956; Von Bekesy, 1930).

In addition to binaural cues, monaural spectral cues are used to determine source location in the vertical plane. These cues arise from frequency-specific modifications in the magnitude and phase of a sound reaching the eardrum caused by interactions with the head and ears. More specifically, the pinna and concha of the external ear filter the spectral content of sound waves before they enter the ear canal in a function referred to as the head-related transfer function (HRTF). These spectral cues for sound localization manifest in a spectrogram as spectral notches created by the HRTF, changing in exact frequency and magnitude in respect to elevation. Spectral cues are also used to resolve front-back confusions in sound localization, a computation that cannot be performed with binaural cues alone (Blauert, 1997).

2.2 The Precedence Effect in Sound Localization with Reflections

In reverberant environments sound localization is more complex as a sound source that reaches the ear will be followed milliseconds later by its reflections from nearby surfaces. Fortunately, the auditory system is capable of localizing sounds in reverberant environments through a group of phenomenon called the precedence effect. Termed by Wallach et al., the precedence effect was demonstrated by seating a subject equidistant from two loudspeakers in a sound-deadened room or anechoic chamber then asking the subject to judge acoustic features of signals projected from the speakers (Wallach et al., 1949). One speaker presented a sound signal and then after a time delay an identical signal was presented from the other. If the time delay is long enough an echo will occur, as the two sound stimuli are individually perceived. The echo threshold depends on the sound stimulus used (e.g., clicks, broadband noise, tones, or speech) and can vary from 2 ms to 100 ms (Litovsky et al., 1999). If the delay is shorter than the echo threshold then listeners will experience fusion, perceiving one signal rather than the two presented. At delays between 1 ms and the echo threshold, the sound source was perceived as originating from the side of the leading sound, while the echo direction was suppressed. This phenomenon is known as localization dominance. At delays under 1 ms a phenomenon known as summing localization occurs, in which a phantom sound source is perceived at a location intermediate of the two speakers rather than at the location of the leading sound source. Overall, the precedence effect demonstrates that auditory spatial perception does not obey a one-to-one mapping to the physical location of sound sources.

2.3 Audio-Visual Cue Integration For Sound Localization With Competing Spatial Cues

Multisensory integration can also influence sound localization. A prime example of this is the ventriloquist effect, in which a speaker speaks without visible lip movements while moving the mouth of a puppet synchronously with his or her speech. If the ventriloquist is successful, the words will be perceived as originating from the location of the puppet and not the speaker.

In the laboratory the ventriloquist effect has been studied by simultaneously presenting audio and visual stimuli in discordant locations while having subjects perform either a discordance detection task or a selective unimodal localization task. A discordance detection task has been used to investigate visual capture by having subjects indicate if the audio-visual stimulus pairs originate from a single location or from separate ones (Jack and Thurlow, 1973; Thurlow and Jack, 1973; Thurlow and Rosenthal, 1976). If subjects reported one location for a spatially discordant AV pair then ‘visual capture’ was said to have occurred. Visual capture has been reported to occur for angles as large as 30º and was found to decrease with increasing AV eccentricity (Jack and Thurlow, 1973). In 1975 Choe et al. questioned whether the auditory signals were completely being captured by the visual cue, or if auditory localization was simply being biased towards the visual cue in a statistical decision process (Choe et al., 1975). To determine if the latter mentioned cross-modal bias was occurring selective unimodal localization tasks were developed.

A selective unimodal localization task has subjects point to auditory stimuli (e.g. clicks, white noise bursts, or tones), while ignoring temporally aligned visual stimuli (Bertelson and Radeau, 1981; Bertelson and Aschersleben, 1998; Charbonneau et al., 2013; Hairston et al., 2003; Weerts and Thurlow, 1971). The sound localization responses typically deviate away from the sound’s actual location and towards the visual cue, in spite of instructions to ignore the visual cue. The degree of cross-modal bias can then be measured by comparing bimodal localization results with a subject’s unimodal results.

A reverse version of the ventriloquist effect, the auditory bias of visual location, can also occur when visual cues become ambiguous. This bias was created by severely blurring visual stimuli, making them more difficult to localize than sounds, and then having subjects localize the visual stimuli among temporally aligned auditory stimuli (Alais and Burr, 2004). This study suggested that AV integration is not simply vision dominating audition, but that AV integration can be more broadly modeled as near-optimal Bayesian integration between the two modalities.

2.4 Motivation

Visual information has been illustrated to readily influence the precedence effect. If the leading or lagging auditory stimulus is visually reinforced, echo suppression has been shown to increase and be inhibited respectively (Bishop et al., 2011). Summing localization and echo thresholds have also shown to increase or decrease significantly depending on the contextual relationship and dynamics of source movement between audio-visual stimuli pairs (Harima et al. 2009). While vision has been shown to influence both the ventriloquist and precedence effects separately, their interactions have yet to be studied together.

All previously cited ventriloquist effect studies used single speakers as sound sources and treated unimodal sound localization as a deterministic process. Alais and Burr 2004 introduced variable visual cue saliency to the ventriloquist paradigm by blurring visual cues, but no study has inspected the effects of variable auditory cue saliency on AV integration. Psychoacousticians however have shown that stereophonic phantom sound sources, created by exploiting the precedence effect and binaural cues, can be designed to have more interaural ambiguity than their single speaker counterparts (Zurek, 1980). By designing stereophonic phantom sound sources, one can simultaneously alter the perceptually computed location of an auditory stimulus as well as the variability of its localization. With the phantom sound source, the effects of auditory cue saliency on AV integration can finally be investigated.

Chapter 3 Experimental Procedures

Chapter 3.1 Subjects

In accordance to procedures approved by Arizona State University, fourteen subjects (4 female, 10 male, mean age 23.2, range 21-26) were recruited to participate in this study. All subjects had self-reported normal hearing and normal or corrected to normal vision. All subjects provided written informed consent and received financial compensation for their participation.

Chapter 3.2 Apparatus and Stimuli

Free-field testing was conducted in a double-walled sound-attenuated chamber (Acoustic Systems RE-243, [2.1 m x 2.1 m x 1.9 m]) lined with 3" acoustic foam. Figure 1A illustrates the spatial arrangement of acoustic and visual stimulation. A listener was seated in front of a black, acoustically transparent curtain, behind which two hidden loudspeakers (Adams F5, frequency response 50-50 kHz) were positioned at lateral angles of +/- 45^o at a distance of 1.1 m relative to the center of the listener's head. Visual stimuli were provided by three high-power LEDs. They were positioned on the acoustic curtain at 45^o to the left, 0^o at the middle, and 45^o to right at eye level of the listener. The LEDs were placed in white ping-pong balls to produce diffuse light flashes. The listener faced forward and put his/her head on a chin rest to minimize head movement.

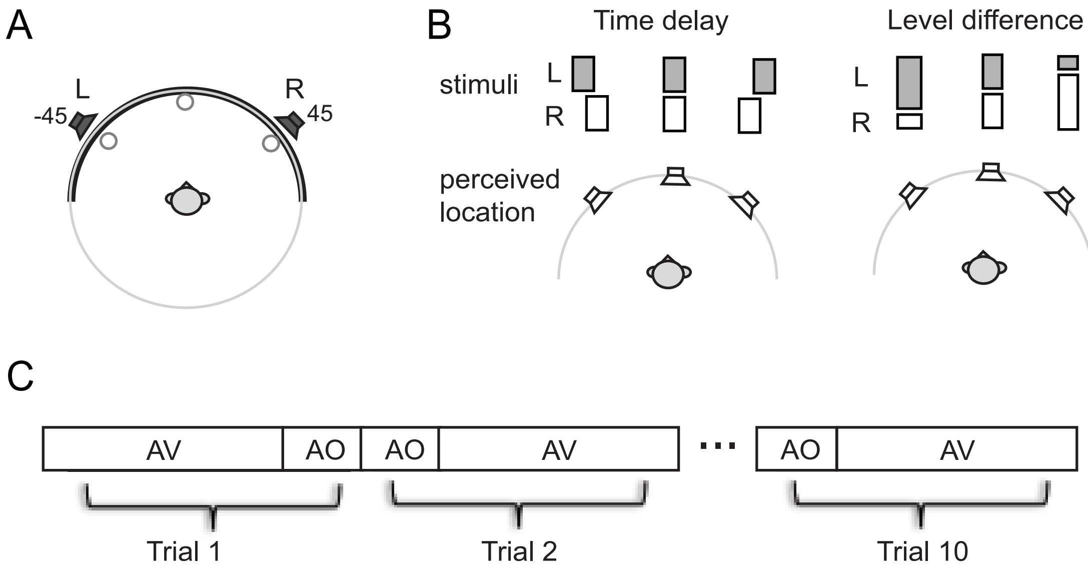

Figure 1: Schematic diagram of the experimental setup for stereophonic and visual stimuli used in experiments.
(A) The spatial arrangement of A and V stimulation. Two loudspeakers were positioned behind an acoustic transparent curtain at ±45^o and in front of the curtain hung were three LEDs at 0^o and ±45^o. (B) Stereophonic setup for auditory fusion. In the time-delay stereophonic method, the onset time difference between the left and right speaker signals was changed from left-lead to right-lead to evoke a perceived sound source location from left to right positions. In the level-difference stereophonic method, the intensity difference between the left and right speaker signals was changed to perceptual move the location of the perceived sound source. (C) Test procedures. AO and AV blocks were presented in random order on each trial. AO block of each trial has 9 stimuli (7 two-speaker fusion and 2 one-speaker controls). For the time-delay condition of each AV block, the light was turned on and off with either the lead or lag speaker, this yields a total of 54 stimuli (3 lights x 2 timing x 9 sounds). For the level-difference condition of each AV block, the light was turned on and off with both speakers, this yields a total of 27 stimuli (3 lights x 1 timing x 9 sounds). The total numbers of stimuli were 630 for the time-delay and 360 for the level-difference stereophonic experiments, respectively.

Stimulus controls were achieved through custom-designed software (written in MATLAB) that generates auditory (A) and visual (V) stimuli and records subject response. Digitized stimuli were sent to an external sound card (RME Multiface II) at a sampling rate of 96 kHz to activate the loudspeakers. The analog outputs of the sound card (via different ports) were also used to activate LEDs, and the analog signals were routed through a simple transistor circuit to yield DC currents for LED activation. Since the same digital-analog device was used for A and V stimulation, their timing error was very small and measured at a sub-microsecond scale by an oscilloscope.

Auditory stimuli were 15-ms, frozen, broadband noise bursts. Identical signals were presented from the two loudspeakers. The noise token was randomly selected for each experimental session. The auditory fusion was introduced by changing the time delay (Fig. 1B) or level difference (Fig. 1C) between left (L) and right (R) loudspeaker signals. Note that the timing and level manipulations referred to here are associated with inter-channel differences between two loudspeakers, not interaural time and level differences commonly studied in the literature. In the time-delay method, 7 inter-stimulus intervals (ISIs, at -1, -0.666, -0.333, 0, 0.333, 0.666, 1 ms) were used to perceptually move a perceived sound source to the left, middle or right positions. In the level-difference method, 7 level differences (-30, -20, -10, 0, 10, 20, 30 dB) were used to achieve this effect. The average intensity of signals from two-speaker stimulation (denoted as the fusion signals) and single-speaker stimulation (denoted as control signals) was adjusted to remain at a constant level of 65 dB SPL in both methods (Brüel & Kjær 2250-L). Visual stimuli were 15-ms light flashes generated by the LEDs. The onset of the light stimulus was synchronized with the sound onset. In the time-delay condition, the light was synchronized with the lead or the lag speaker signals. As a result, there were twice more AV trials in time-delay conditions than the level-difference conditions.

Chapter 3.3 Procedures

Auditory stimuli were presented with or without lights in randomized blocks, which were denoted as audio-visual (AV) and audio-alone (AO) blocks, respectively. Each block contained a set of 9 auditory stimuli, 7 for testing fusion (through time-delay or level-difference methods as mentioned earlier) and 2 for controls (L or R speaker alone). In the level-difference method, there was one AO block and 3 AV blocks for 3 light positions, respectively. In the AV blocks for the time-delay methods, the light flash was in synch with either the leading or lagging speaker stimuli. This resulted in one AO block and 6 AV blocks for 3 light positions.

Sound localization tasks were executed using a graphical user interface (GUI) shown on a touch screen monitor (10”x8”). Before the task participants used a training panel to familiarize themselves with localizing a sound using the GUI. The familiarization mainly involved learning the association between seven response buttons and seven auditory fusion stimuli. Participants were allowed to use the training panel as long as they wanted to until they indicated they were ready to begin the task. Once the task began, a participant would initiate a trial by pushing a “Next” button on the GUI. Upon hearing a sound, the participant indicated the perceived direction of the sound source by pushing one of seven buttons, which were numerically labeled from “1” to “7” and horizontally positioned from left to right on the GUI. After a response there was a 1s delay before the subject could begin the next trial. Breaks were encouraged to be taken every 15 minutes to ensure a participant remained engaged in a task. No feedback was provided during and after the experiments. Participants were not informed about the exact number and spatial locations of the loudspeakers. They often reported, however, there were at least 5 to 7 speakers in the exit survey.

Chapter 3.4 Data analysis

Responses were grouped into AO and AV conditions to be analyzed. The AV conditions are associated with left, middle, and right light positions – AV_L, AV_M, and AV_R, respectively. The lateral angle of a perceived sound source was obtained by mapping the seven choice buttons (from “1” to “7”) to seven angles, -45^o, -30^o, -15^o, 0^o, 15^o, 30^o, 45^o, where the two furthest angles ($+/-45^o$) marked the positions of L and R loudspeakers, respectively. The confusion matrix was constructed based on the stimulus-response data for each subject at AO or AV conditions. The mean and standard deviation (SD) of subject responses were then used to analyze the population response of all participants.

Visual influences on sound source localization were evaluated with respect to: (1) the extent of cross-modal shift and (2) the change in response reliability. Cross-modal shift was defined as the difference in mean response of a subject between AV and AO conditions tested with identical auditory stimuli, $\Delta AV = mean(AV) – mean(AO)$. This yields three sets of shifts, ΔAV_L, ΔAV_M, and ΔAV_R. A positive ΔAV is associated with a rightward shift in response. Response variability was evaluated based on trial-to-trial variability of responses (SD) and the change in response variability was determined by the difference in SDs between AO and AV conditions tested with identical auditory stimuli, $\Delta SD = SD(AV) – SD(AO)$. A positive ΔSD indicates increased response variability by including light stimulation.

Chapter 3.5 Ear-canal signal analysis

Left and right ear-canal signals were collected using a KEMAR binaural head dummy placed in the center of the room. The two ears of KEMAR were fitted with G.R.A.S. 40BP microphones and the microphone output signals were amplified by a G.R.A.S. 26AS preamps and a RME Multiface's +4 dBu analog line. The signals were digitized by the Multiface's 24-bit analog inputs at a sampling rate of 96 kHz. Each stimulus used in time-delay and level-difference methods was recorded 10 times and averaged to minimize the noise fluctuation. The spectral analysis was conducted on monaural signals to investigate the extent of comb filtering as a result of inter-channel interaction between L and R speaker signals.

Binaural spectral differences were then calculated to obtain the interaural level difference as a function of frequency for each stimulus condition. An energy-normalized cross-correlation analysis was further conducted on the narrow-band signals obtained by passing ear-canal signals through an ERB filter-bank that simulates basilar membrane functions (Slaney, 1993). A total of 128 channels were implemented between 50 Hz and 48 kHz and the outputs of the first 51 filters (with characteristic frequencies ranging between 46 and 1852 Hz) were used for the correlation analysis. The equation used is formerly referred to as the Binaural Coherence equation in previous studies is defined as:

$$\rho(\tau) = \frac{\int x_L(t) x_r(t+\tau)dt}{\sqrt{E_L E_R}}$$

(Blauert, 1997; Rankerd and Hartmann, 2010). The cross-correlation function ρ is a function of time lag τ where $x_L(t)$ is the is the left monaural signal, E_L is the energy in the left ear, and E_R is energy for the right ear. The interaural time differences associated with the peaks of the correlation function were transformed into interaural phase differences and plotted as a function of frequency.

Chapter 4 Results

4.1 Visual capture of a perceived sound source location in stereophonic listening

Figure 2 shows the sound localization results of one subject tested with the time delay (Fig. 2A) and level difference (Fig. 2B) methods. On each panel, the top row depicts the stimulus-response relationship for four stimulus conditions (AO, AV_L, AV_M, AV_R) and the bottom row is the corresponding mean and SD of responses. This subject's sound localization performance (AO, leftmost column) demonstrates the classic stereophonic perception (Wendt, 1963; Leakey, 1957) when the timing and level differences were systematically varied between two loudspeaker signals. The perceived sound source location points towards the direction of the loudspeaker emitting earlier or louder signals and when the signals from the two loudspeakers are equal in amplitude and timing, a center-location sound source is reported.

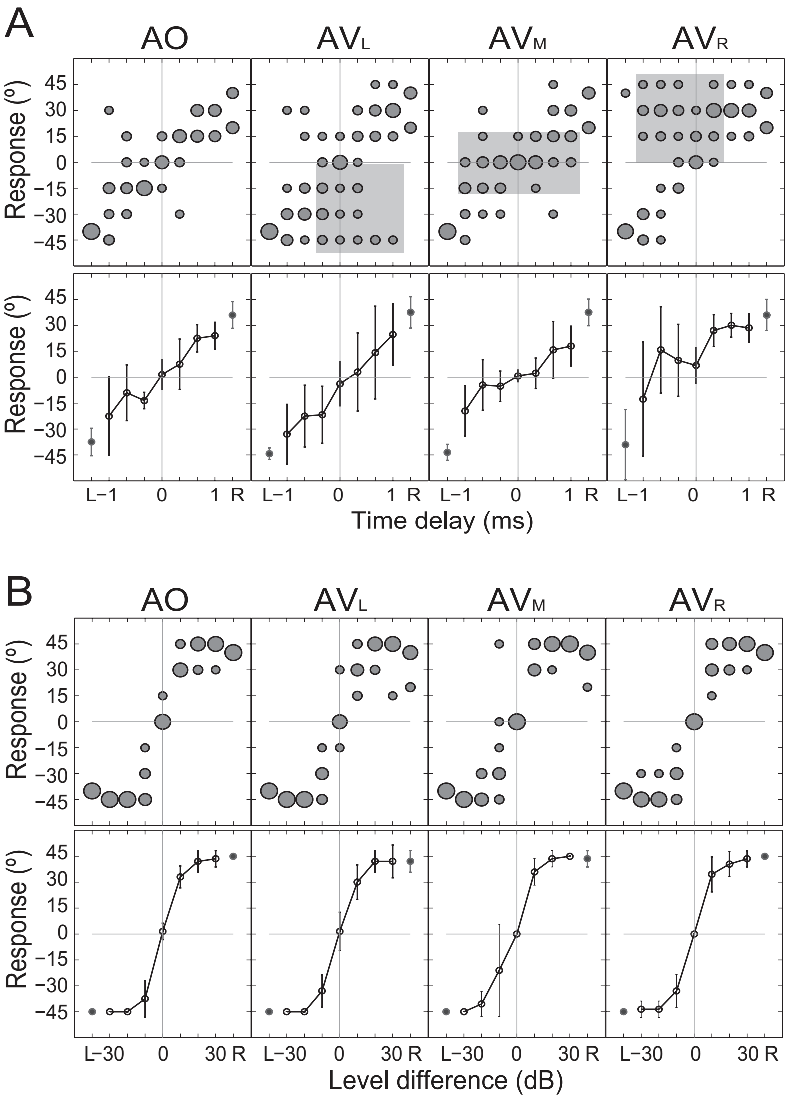

Figure 2: Visual influences on the auditory localization performance of one subject. (A) Auditory localization using time-delay stereophony with and without light stimulation. The top row shows the stimulus-response correlation in bubble plots. The bottom row shows the mean and SD of responses for each stimulus. The three AV conditions (AV_L, AV_M, AV_R) are associated with light flashes in the left, middle, and right directions, respectively. The one-speaker control stimuli are labeled as L and R on each panel. (B) Auditory localization using level-difference stereophony with and without light stimulation. The data were plotted in the same format as those in (A).

The inclusion of spatially discordant visual stimuli appeared to be distractive and caused increasing incidences of responses towards the direction of light when auditory spatial responses were induced by time-delay stereophonic stimuli (Fig. 2A). To help visualize this effect, stimulus-response regions with erroneous localization results were highlighted in light-gray boxes. Compared to the AO responses, light flashes misled auditory localization and caused (1) a shift in response mean towards the light direction and (2) an increase in response variability. In contrast to the time-delay results, auditory localization in response to level stereophonic manipulations was affected to a much less degree independent of the light position. In Fig. 2B, localization performances remain nearly unchanged after light stimulations. This subject's data indicates that although the auditory localization shows similar patterns with the timing and level stereophonic stimuli, visual information may not be equally effective in inducing a cross-modal response bias.

Figure 3 shows in detail the patterns of audio-visual localization of each subject. Each data point linked pair wise the response decision of a subject between two stereophonic stimuli. In the AO condition, the reported sound source direction between the time-delay (abscissa) and level-difference (ordinate) conditions are highly correlated (*p*<0.001; linear regression). Nevertheless results for the level difference method showed a noticeable bias towards more lateral positions when inter-channel level difference exceeded 20 dB. This bias can be visualized by data points below and above the diagonal line on the AO panel. After light stimulation, the vertical spread of data for the level method remains largely unchanged, whereas there is a noticeable shift in the horizontal spread of data towards the light direction for the timing method. This asymmetry suggests that the magnitude of visual bias is not correlated with the lateral extent of perceived sound direction. Otherwise, coordinate shifts along the diagonal line would be expected.

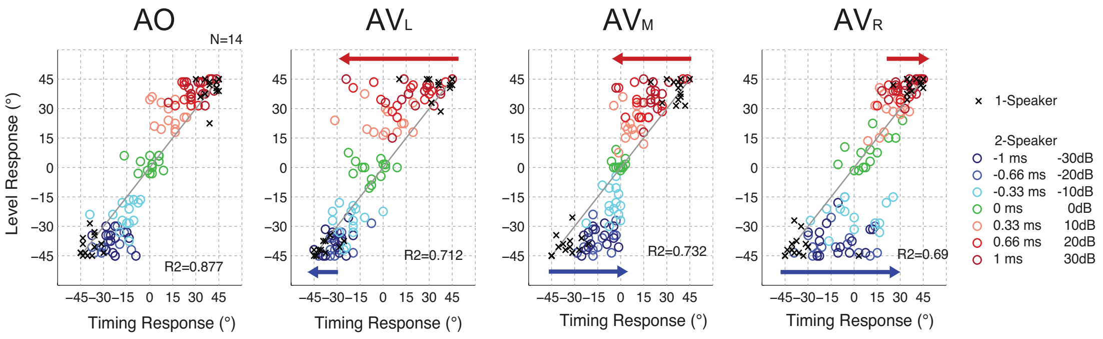

Figure 3: Pair-wise comparison of the auditory localization and visual bias between two stereophonic methods
On all panels each data point is one subject's mean response to a time-delayed stimulus against the mean response to a level modulated stimulus. Time delayed and level modulated stimuli were compared by magnitude (e.g. 0 dB vs. 0 ms, ±10 dB vs. ±0.33 ms, ±20 dB vs. ±0.66 ms, and ±30 dB vs. ±1 ms). Negative values of stimuli elicited leftward AO responses. In experiments, the subject responses were button choices from "1" to "7", which were transformed to -45 to 45 degrees for data analyses. The correlation of determination (R^2) was calculated for the two stereophonic responses in AO and each of the three AV conditions. Red and blue arrows on each AV panel illustrate the direction of visual bias for leftward and rightward AO responses, respectively. Single-speaker control responses are shown by "x" markers.

The data also revealed that for the timing condition the magnitude of visual bias, or the strength of visual capture, increases with the spatial disparity between unitary modality signals. The colored arrows above each AV condition may help illustrate this effect. Light-induced shift becomes greater with increasing distance between the light and the perceived location of a sound from the AO condition, as shown by increased incidences of responses towards left (AV_L), midline (AV_M), and right (AV_R) light directions, respectively. For example, greater shifts were observed when left-side lights were presented simultaneously with right-side sound than those presented from the same side. On the other hand, reliability of responses appears to increase for spatially congruent A and V stimuli. This can be seen by the enhanced density of data clustering (blue in AV_L, green in AV_M and red in AV_M) relative to those in the AO condition.

4.2 Population analysis of the magnitude of visual capture in two stereophonic conditions

Figure 4 shows the averaged mean and SD of AO and AV responses of all 14 participants. In the absence of light (black lines in Panels A and B), the population data show the typical stereophonic perception as reported in the literature (Blauert, 1997; Wendt, 1963). Comparing the results of two stereophonic manipulations, a general shift towards the light direction can be seen for the time-delay (Fig. 4A), but not level-difference (Fig. 4B), methods. A two factor ANOVA [auditory stimulus x visual stimulus] showed a significant interaction between auditory stimulus and visual stimulus for the time delayed stimuli (*p* = 0.0066), while this interaction was not present for the level modulated stimuli (*p* = 0.29). The main effect in level difference localization was the auditory stimulus (*p* = 0.0002). Furthermore, visual cues did not induce a significant response shift for the single-speaker control conditions (L or R). This suggests that although lateral localization was reported in both fusion and control conditions, fusion responses due to the time-delay manipulation appear to be more susceptible to visual influence.

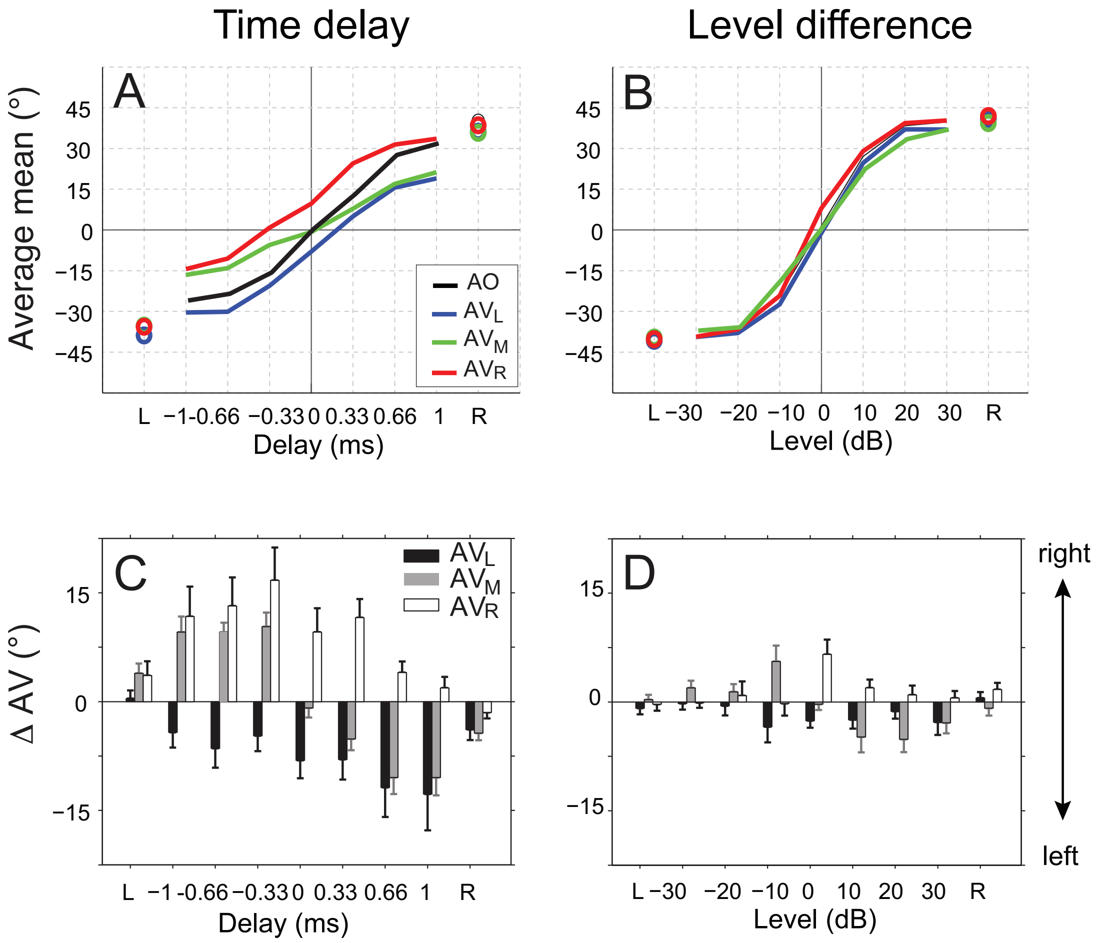

Figure 4: Population analysis of auditory fusion and control responses and the magnitude of visual bias in two stereophonic conditions.
(A)(B) Population average of the mean responses of all subjects in AO and AV conditions for the time-delay and level-difference manipulations. One-speaker control responses are shown in circles and stereophonic responses are shown in color lines. (C)(D) The population average of magnitude of visual bias at the three light positions for the time-delay and level-difference manipulations. Rightward visual bias (positive ΔAV) is strongest for leftward AO responses paired with right light, whereas leftward visual bias is strongest for rightward AO responses paired with left light. The magnitude of visual capture is much stronger for stereophonic responses induced by the time-delay than level-difference method. Error bars show standard error of the mean.

To quantify how localization performances of individual subjects changed with visual stimulation, we further analyzed for each subject the magnitude of visual bias [ΔAV = mean(AV) - mean(AO)] relative to his/her own AO responses as a function of time delay and level difference between two loudspeaker signals. Figures 4C and D show the average visual bias and visual capture (ΔAV) across 14 participants (mean±SEM). For the time delay method, the magnitude of visual bias shows mirror-symmetric patterns between AV_L and AV_R conditions with increasing bias towards the light direction for spatially discordant A and V stimuli. Strong visual capture is correlated with the eccentricity of visual cues. In comparison, this pattern of cross-modal bias was relatively weak and visual bias was much reduced in results from the level difference method. It is also clear that responses to single-speaker control signals are more robust against the influences of visual capture than the two-speaker stereophonic signals (e.g., L vs. -1ms and R vs. 1ms). Overall, the sheer contrast between time-delay results and those of control and level difference indicates that the magnitude of visual capture does not solely depend on the lateral extent of the perceived sound source location in the auditory only condition.

4.3 Population analysis of the response variability in two stereophonic conditions

According to the recent work on multisensory interactions, cross-modal bias increases as the reliability of unitary sensory cues decreases (Alais and Burr, 2004). To investigate potential causes for the greater visual capture observed in the time-delay condition, we compared the response variability for the two stereophonic stimuli. Figure 5 shows the analysis of the response SD of all subjects, which are presented in similar formats as the mean data shown in Fig. 4. In the AO condition (black lines in Panels A and B), the time-delay responses show higher SD values than those of the level-difference responses and of the single-speaker control responses (p<10^{-5}). This observation provides one explanation for the noted larger visual capture shown in Fig. 4. That is, the time-delay fusion responses are less reliable relative to the level-difference and control responses. The data also reveal that response variability is affected by the eccentricity of visual cues. Compared to the AO conditions, visual cues from the peripheral field (AV_L and AV_R) mostly increased the response variability, whereas visual cues from the central field (AV_M) decreased response variability. This trend was more pronounced in results from the time delay method. For both timing and level conditions, response variability was smaller when the perceived sound was spatially concordant with the visual cue than when they were not. The least variability, or greatest response reliability, was achieved when the light was presented from the center and two-loudspeaker signals were perfectly balanced with a fused percept of sound source position in front (the dip on the green curves, AL_M).

We further calculated the changes in response variability [ΔSD= SD(AV) – SD(AO)] of each subject as a function of time delay and level difference between two loudspeaker signals. The patterns of ΔSD shown in Fig. 5C and D provide quantitative estimates of how a visual signal can either reduce or enlarge the response variability in auditory localization performance. For the time-delay methods, similar mirror-symmetric patterns between AV_L and AV_R conditions were observed in ΔSD as shown in the mean shift data. More specifically, vision can either distract the auditory localization by increasing SD when the unitary A and V were on the opposite sides, or attract the auditory localization by decreasing SD or increasing the reliability of auditory localization performance when they were from the same side. The most effective “attractor” is the central location light when it is paired with balanced stereophonic inputs, so that A and V events were both perceived directly in front. Vision’s roles of being either a “distractor” or “attractor” are more pronounced in results from the time delay than level difference conditions.

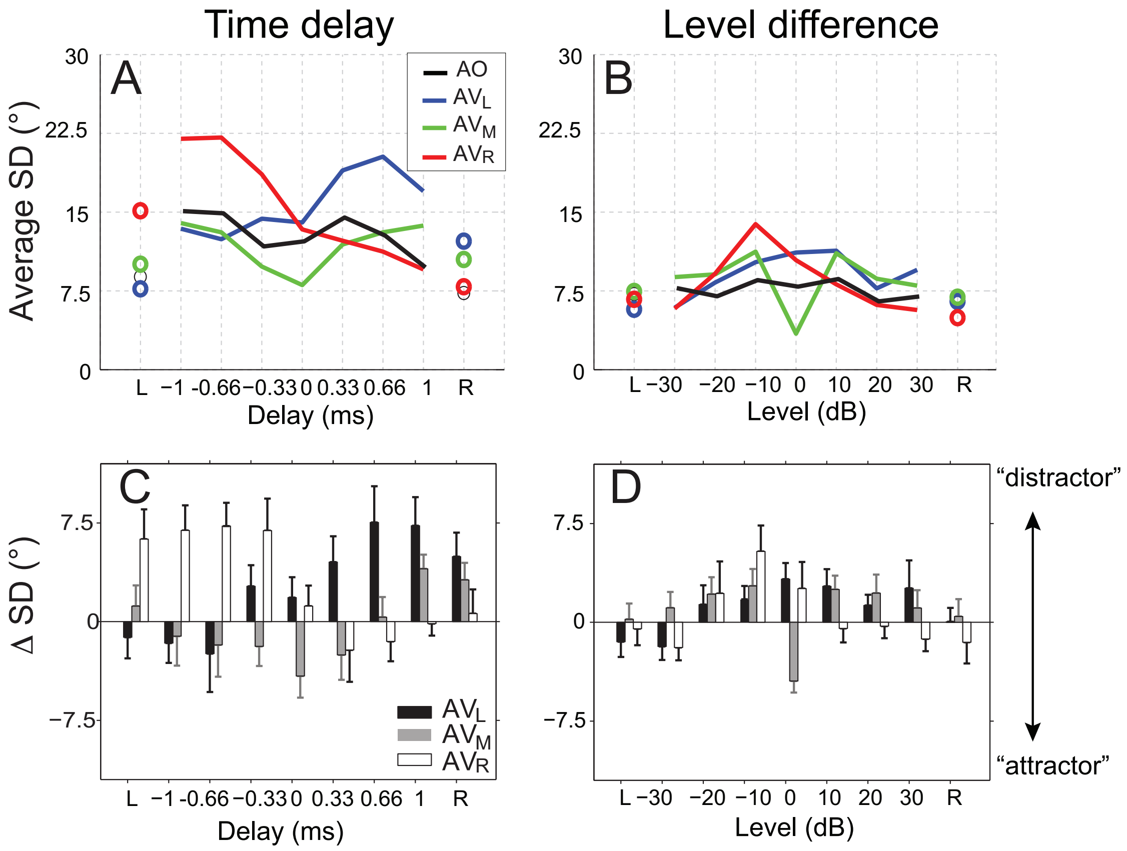

Figure 5: Population analysis of the response variability with and without light in two stereophonic conditions. (A)(B) Population average of the standard deviation (SD) of responses for each stimulus for the time-delay and level-difference manipulations. One-speaker control responses are shown in circles and stereophonic responses are shown in color lines. Response variability was overall higher in time-delay localization (A) than in level-difference localization (B). (C)(D) The population average of changes in SDs of individual subjects' responses at the three light positions. Response variability increased with increasing eccentricity between auditory and visual stimulus locations and decreased for spatially concordant A and V cues. This trend was much greater in time-delay conditions than in level-difference conditions. Error bars show standard error of the mean.

Figure 6 summarizes the main findings in these results shown in Figs. 3-5. Panel A compares the averaged response variability SD in the AO condition (mean of the AO data in Figs. 5A and B). The SDs for the time-delay responses were significantly higher than the level-difference response (p=0.001) and the control responses ($p<1e^{-5}$). This distinction in response variability provides some explanation for the magnitude of visual bias shown in Fig. 6B, which compares the overall visual bias (mean of the absolute visual biases in Figs. 4A and B, averaged across all time-delay or level-difference values). Pairwise comparisons reveal that (1) visual bias is significantly higher in localizing a stereophonic auditory sound source generated by the time-delayed method than level-modulated signals; (2) visual bias is significantly higher for two-speaker stereophonic stimuli than one-speaker stimuli; (3) visual bias of single speaker responses is slightly higher in the time-delay than level-difference conditions, despite that the control signals were identical in these two conditions and non-significant difference in their SDs were found (Fig. 6A). The control results suggest that the contextual factors across trials may influence the subject performance in addition to visual cues.

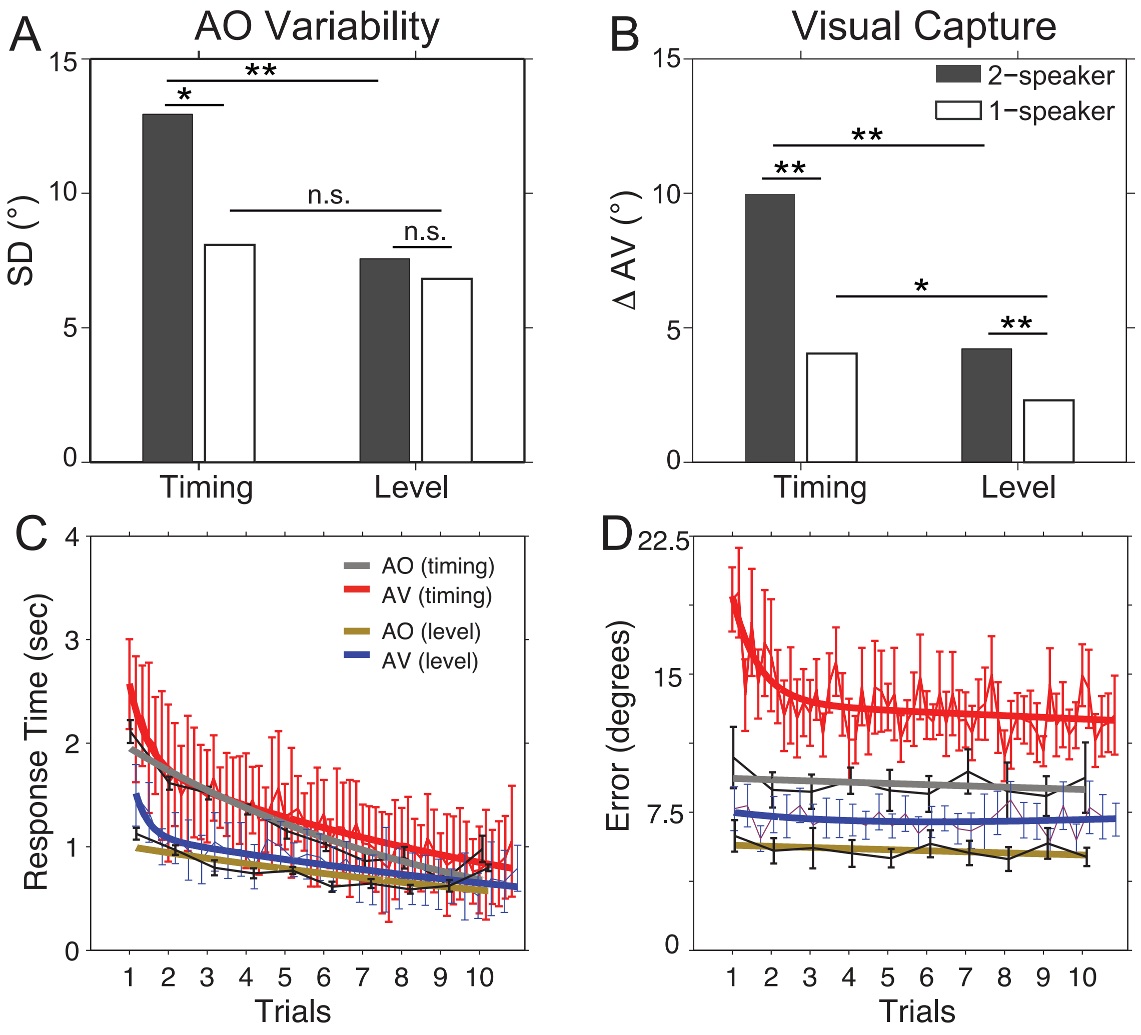

Figure 6: Comparison of overall visual capture, response time and response variability between two stereophonic conditions.
Total average of visual capture (A) and total average of response variability (B) across time delays and level differences for two-speaker and one speaker responses. Response time (mean±sem, C) and error size (mean±sem, D) are plotted as a function of trial number. Thick color lines are exponential fit. For AO results in D, the error is related to the deviation of subject's trial-to-trial response from the overall AO mean shown in Figs. 4A and B. For AV results in D, the error magnitude is a slightly different measure than the response bias shown previously. Here the error is related to the absolute shift (positive or negative) after visual stimulation, which was higher than the directional dependent shift in B. In the latter case the positive and negative shift might cancel out in the mean estimate. Since AV blocks had more conditions than AO blocks, response size and error size were estimated every 9 stimuli similar to AO blocks. [* p=0.001; **p<1e^{-5}; n.s., not significant.]

To discern the potential influences of contextual factors, the response time and error size between the four data sets were compared (AO vs. AV; timing vs. level). Figure 6C shows the averaged results from 14 individual subjects (mean±sem, thin lines) superimposed with exponential fit (thick lines) as a function of the trial number. Table I reports the reaction times for the first and last trials to assist pair-wise comparisons. Overall, reaction time for AO and AV blocks for the same experiment (time or level methods) follows the same patterns of evolution. This is expected as AO and AV blocks were tested back-to-back (Fig. 1C) and therefore may be subject to similar associative or procedural learning rules. However, the inter-subject variability was much smaller in the AO than AV conditions as manifested by the size of the error bars. Comparing results from different stereophonic experiments (time-delayed or level-modulated) reveals that at the beginning (Trial 1), it took a longer time for a subject to report a response to time-delay stimuli than to the level-difference stimuli with or without light (see data in Table I). As trials continued, there was a gradual decline in reaction time and the decay rate was faster for the time-delay stereophonic stimuli. Towards the end of an experiment (Trial 10), the reaction times became remarkably similar for the timing and level methods in both AO and AV blocks. These results suggest that procedural learning is likely involved in each task and learning strategies may speed up the reaction times for unambiguous AV stimuli reaching to a fixed value.

We then analyzed the time course of error size and evaluated the extent of cross-modal learning with increased listening time. Figure 6D shows the absolute magnitude of response error relative to the baseline AO responses (i.e., black lines in Fig. 4A and B); data are plotted in the same format as those in Fig. 6C. A quick reduction in error size after one trial of listening is seen in AV blocks with the time-delay method (red). However, unlike the reaction time, no improvement in error size was observed with repeated listening in the AO blocks and the AV blocks for the level method. It can be seen that both AV and AO error magnitudes are greater for the time-delayed than level-modified responses at all trials. These observations suggest that the stimulus uncertainty casts an immediate and long-lasting effect on auditory localization with and without vision.

Table I: Reaction Time in AO and AV conditions.

	AO		AV	
	Trial 1 (sec)	Trial 10 (sec)	Trial 1 (sec)	Trial 10 (sec)
Time-delay	3.09±0.92	1.16±0.17	2.53±0.45	0.92±0.16
Level-difference	1.34±0.34	1.14±0.21	1.22±0.21	0.73±0.13
ranksum test	*p*=0.011	*p*=0.535	*p*=0.012	*p*=0.323

4.4 Binaural localization cues underlying stimulus uncertainty

Auditory spatial hearing in the horizontal plane relies on binaural disparity information – interaural time difference (ITD) and interaural level difference (ILD). To address the reliability of these binaural cues in relation to the observed visual capture (Figs. 3-6), we measured the ear-canal signals using a KEMAR binaural dummy head for all types of stimuli used in our experiments. The ILDs were measured from the binaural spectral differences between left and right HRTFs, which are simply the power spectrum density functions of the ear canal signals. For ITDs, the auditory system relies on the coincidence-based, cross-correlation comparison of peripherally filtered signals at the brainstem. Here we filtered the ear signals through a 128-channel ERB filter bank (covering frequency range between 50 Hz and 48 kHz) to simulate the auditory peripheral processing. The outputs of the first 47 channels (center frequencies ranging from 46 Hz to 1546 Hz) were then used to extract ITDs through the binaural coherence analysis (See details in Methods). Finally, ITDs were transformed to IPDs ($IPD=2\pi*ITD*frequency$) bounded between $-\pi$ and π. Figure 7 compares the results between time-delay and level-difference methods at selective individual conditions for leftward and central localization (left-speaker control; delays from -1 to 0ms; level from -30 to 0dB). On each panel, the top row is the distribution of IPD/ILD cues across frequencies and the bottom row is the overall density function of IPD/ILD.

We observed that in a stereophonic setup, changing the time-delay and level-difference between two loudspeaker signals alter the ITDs and ILDs in different ways.

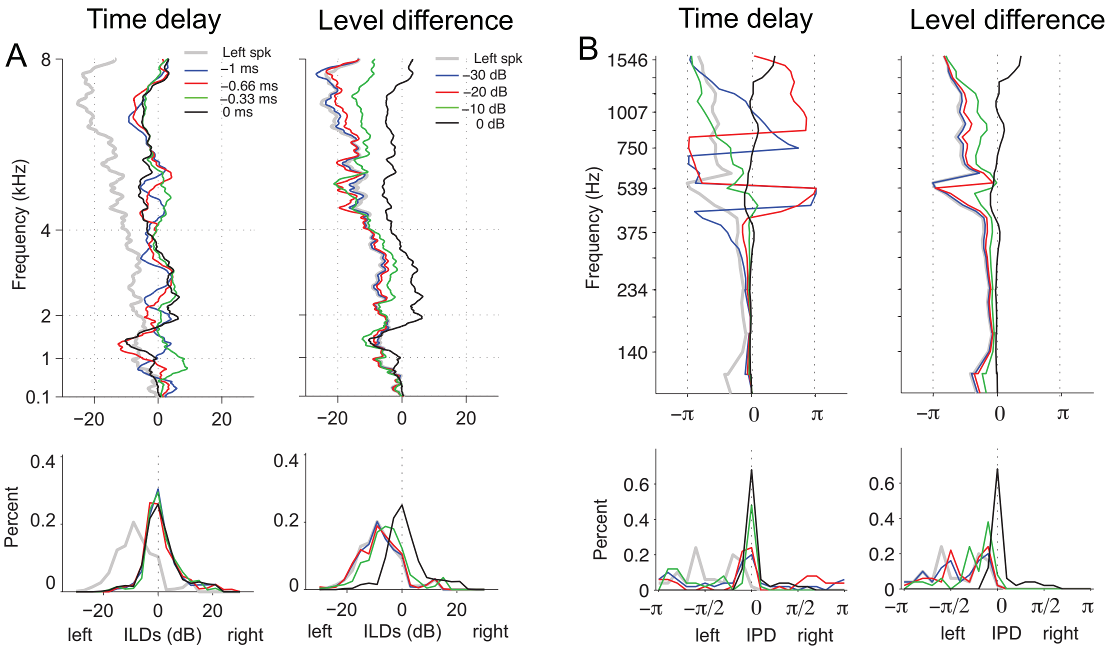

Figure 7: Comparison of binaural cues ITDs and ILDs between two stereophonic conditions.
(A) ILDs at different time delays and level ratios of stereophonic signals. Top panel shows the ILD distribution across frequencies (up to 8 kHz) and the bottom panel shows the overall distribution of ILDs. The ILD for left-speaker control signals was plotted as a reference. (A) IPDs at different time delays and level ratios of stereophonic signals. The data were plotted in the same format as those in A. The binaural coherence analysis was limited to ERB filters with center frequencies between 47 and 1546 Hz. For the overall distribution of binaural cues, bin width is 3dB for ILDs and π/10 for IPDs.

Figure 7A shows that the ILD is a much weaker cue for localizing time-delayed stimuli than the level-modified stereophonic stimuli. More specifically, ILDs appear to be inconsistent in terms of its sideness at different delay values. For example, while ILDs were mostly at zero for delay of 0 ms, they pointed to either left or right side for other nonzero delays, depending on frequency. As a result, unlike the clearly leftward distribution of left-speaker response, the overall distribution of ILDs at three non-zero delays (-0.33, -0.66, and -1 ms) all centered around zero (left, button panel).

The patterns of ILDs for the time-delayed stimuli can be explained by the spectral variations or ripples in monaural HRTFs due to the lead-lag interaction of stereophonic signals (Blauert 1997, Zurek 1980). In simple terms, adding a delayed copy to the original signal, as done in the time-delayed method, lead to comb-filtering in the combined signals, where the spectral ripple frequency is inversely proportional to the delay. Since left and right ear canals contained identical copies of stimuli but with different delays, they exhibit different ripple frequencies. The widely varied ILDs in the time-delayed condition manifest varying spectral peaks between left and right ear signals. In contrast, ILDs for level-modified stimuli consistently pointed to the left side because differential comb filtering does not occur for level-modified stimuli.

As shown in Fig. 7B, ITDs/IPDs for the time-delayed stereophonic stimuli also showed left-right confusions at multiple frequency bands, whereas ITDs for the level-modified stereophonic stimuli consistently pointed to the left side. It can be seen that only for the level method, both ILDs and ITDs show graded changes in accord with the magnitude of level difference. Taken together, for the time-delayed stimuli, the associated binaural ITD and ILD cues are inconsistent with each other and across frequencies. We conclude that the binaural ambiguities in localization cues may underlie the observed visual bias with respect to (1) strong visual capture; (2) large response variability and (3) high error magnitude over time as found in the results from time-delayed stimuli.

Chapter 5 Discussion

Earlier work that has studied the ventriloquist effect generally used single-light sources paired with single speakers and consequentially treated sound localization on its own as a deterministic process (Bertelson and Radeau, 1981; Charbonneau et al., 2013; Hairston et al., 2003). More recently, Bayesian methods that assume the brain represents sensory information probabilistically have successfully been used to model processes of multisensory integration like sensory-motor control and the ventriloquist effect (Knill and Pouget, 2004). The modulation of visual cue saliency in conjunction with the ventriloquist effect has been exploited to show near-optimal Bayesian audio-visual integration (Alais and Burr, 2004), but sound localization cues were a control in that experiment. By using phantom sound sources with variable binaural cue saliency, this thesis was able to explore the effects of auditory cue saliency on the ventriloquist effect.

The analyses shown in Figs. 3 and 4 indicate that on average, cross-modal bias influenced auditory spatial perception to a greater extent for the time-delayed auditory stimuli than level-difference modulated stimuli and single speaker controls. Figs. 4 and 5 showed that both the degree of cross-modal bias and response variability increase with eccentricity. While this phenomenon has been previously reported (Charbonneau et al., 2013; Thurlow and Jack, 1973), these data go on further to demonstrate that both cross-modal bias and response variability are also modulated by auditory cue saliency. Fig. 6 analyzed time course reaction times and error size. Procedural learning was found to decrease reaction times over the time course of the experiment, but there was no improvement in error size with repeated listening. The acoustic analysis in Fig. 7 illustrated that the time-delayed auditory stimuli had more ambiguous ILDs and ITDs/IPDs than the level-modulated stimuli and control speakers. All together, these data demonstrate that unisensory reliability between auditory and visual cues correlates with the degree of visual capture. That is, the more unreliable sound localization becomes, the more vision will dominate and capture the sound.

The methods developed in this thesis highlighted the impact binaural cue saliency can have on the ventriloquist effect. While these data provide evidence of audio-visual cues integrating in a Bayesian manner, future research should test this data within existing Bayesian models for audio-visual integration. Previous models can be vetted and a new model could be proposed. In this experiment’s task, sound localization was limited to the frontal hemifield. As the receptive field of the auditory system extends beyond the field of vision, it would be interesting to see if the results of this research would still occur with visual cues coming from the frontal hemifield and sounds originating from the rear. This thesis also illuminated the utility of the phantom sound source for research in multisensory integration, an avenue left widely unexplored by current psychophysical and neurophysiologic audio-visual integration studies.

References

Abe, K., Sunada, D., Takane, S., and Sato, S. "Relationship between the Summing-localization Behavior and Perceived Width of Sound Image." *Acoustical Science and Technology Acoust. Sci. & Tech.* 31.4 (2010): 260-66.

Bertelson, P. and Aschersleben, G. "Automatic Visual Bias of Perceived Auditory Location." *Psychonomic Bulletin & Review* 5.3 (1998): 482-89. 01 Sept. 1998.

Bertelson, P. and Radeau, M. "Cross-modal Bias and Perceptual Fusion with Auditory-visual Spatial Discordance." *Perception & Psychophysics* 29.6 (1981): 578-84.

Bertelson, P., Vroomen, J., De Gelder, B., and Driver, J. "The Ventriloquist Effect Does Not Depend on the Direction of Deliberate Visual Attention." *Perception & Psychophysics* 62.2 (2000): 321-32.

Bishop, C. W., London, S., and Miller, L. M. "Visual Influences on Echo Suppression." *Current Biology* 21.3 (2011): 221-25.

Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT, 1997.

Brown, A. D., Stecker, G. C., and Tollin, D. J. "The Precedence Effect in Sound Localization." *Journal of the Association for Research in Otolaryngology JARO* 16.1 (2014): 1-28.

Charbonneau, G., Véronneau, M., Boudrias-Fournier, C., Lepore, F.,and Collignon, O. "The Ventriloquist in Periphery: Impact of Eccentricity-related Reliability on Audio-visual Localization." *Journal of Vision* 13.12 (2013): 20.

Choe, C. S., Welch, R. B., Gilford, R. M., and Juola, J. F. "The “Ventriloquist Effect”: Visual Dominance or Response Bias?" *Perception & Psychophysics* 18.1 (1975): 55-60.

Gelder, B., and Bertelson, P. "Multisensory Integration, Perception and Ecological Validity." *Trends in Cognitive Sciences* 7.10 (2003): 460-67.

Hairston, W. D., M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo. "Visual Localization Ability Influences Cross-Modal Bias." *Journal of Cognitive Neuroscience* 15.1 (2003): 20-29.

Harima, T., Abe, K., Takane, S., Sato, S., and Sone, T. "Influence of Visual Stimulus on the Precedence Effect in Sound Localization." *Acoustical Science and Technology Acoust. Sci. & Tech.* 30.4 (2009): 240-48.

Jack, C., and Thurlow, W. "Effects Of Degree Of Visual Association And Angle Of Displacement On The "Ventriloquism" Effect." *Perceptual and Motor Skills* 37.3 (1973): 967-79.

Jackson, C. V. “Visual Factors in Auditory Localization.” *Quarterly Journal of Experimental Psychology*. 28. (1941): 163-175

Klumpp, R. G. "Some Measurements of Interaural Time Difference Thresholds." *J. Acoust. Soc. Am. The Journal of the Acoustical Society of America* 28.5 (1956): 859.

Knill, D. C., and Pouget, A. "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation." *Trends in Neurosciences* 27.12 (2004): 712-19.

Leakey, D. M. “Some Measures On the Effects of Interchannel Intensity and Time Differences In Two Channel Sound Systems.” J. Acoust. Soc. Am. 31 (1957): 997-986

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. "The Precedence Effect." *J. Acoust. Soc. Am. The Journal of the Acoustical Society of America* 106.4 (1999): 1633.

Mcfadden, D., Jeffress, L. A., and Russell, W. E. "Individual Differences In Sensitivity To Interaural Differences In Time And Level." *Perceptual and Motor Skills* 37.3 (1973): 755-61.

Rakerd, B., and Hartmann, W. M. "Localization of Sound in Rooms. V. Binaural Coherence and Human Sensitivity to Interaural Time Differences in Noise." *J. Acoust. Soc. Am. The Journal of the Acoustical Society of America* 128.5 (2010): 3052.

Roach, N. W., Heron, J., Whitaker, D., and Mcgraw, P. V. "Asynchrony Adaptation Reveals Neural Population Code for Audio-visual Timing." *Proceedings of the Royal Society B: Biological Sciences* 278.1710 (2010): 1314-322.

Saberi, K., H. Farahbod, and M. Konishi. "How Do Owls Localize Interaurally Phase-ambiguous Signals?" *Proceedings of the National Academy of Sciences* 95.11 (1998): 6465-468.

Shackleton, T. M., and Palmer, A. R. "Contributions of Intrinsic Neural and Stimulus Variance to Binaural Sensitivity." *Journal of the Association for Research in Otolaryngology JARO* 7.4 (2006): 425-42.

Shelton, B. R., and Searle, C. L. "The Influence of Vision on the Absolute Identification of Sound-source Position." *Perception & Psychophysics* 28.6 (1980): 589-96.

Slaney, M. "Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work." *Apple Technical Report #45* (1993)

Thurlow, W. R., and Jack, C. E. "Certain Determinants Of The ‘Ventriloquism Effect’" *Perceptual and Motor Skills* 36.3c (1973): 1171-184.

Thurlow, W. R., and Rosenthal, T. M. "Further Study Of Existence Regions For The "Ventriloquism Effect"" *Journal of the American Audiology Society* 1.6 (1976): 280-86.

Tollin, D. J. "Spectral Cues Explain Illusory Elevation Effects With Stereo Sounds in Cats." *Journal of Neurophysiology* 90.1 (2003): 525-30.

Tollin, D. J., and Henning, G. B. "Some Aspects of the Lateralization of Echoed Sound in Man. II. The Role of the Stimulus Spectrum." *J. Acoust. Soc. Am. The Journal of the Acoustical Society of America* 105.2 (1999): 838.

Von Bekesy G. Zur Theorie des Hörens: Über das Richtungshören bei einer Zeitdifferenz oder Lautstärkeungleichheit der beidseitigen Schalleinwirkungen. Phys Z 824–838, 1930.

Wallach, H., Newman, E. B., and Rosenzweig, M. R. “The precedence effect in sound localization.” Am. J. Psychol. LXII(3) (1949): 315-336

Weerts, T. C., and Thurlow, W. R. "The Effects of Eye Position and Expectation on Sound Localization." *Perception & Psychophysics* 9.1 (1971): 35-39.

Wendt, K. “Directional Hearing With Two Superimposed Sound Fields In Intensity- and Delay-difference Stereophony. Dissertation, Technische Hochschule, Aachen. (1963)

Willert, V., Eggert, J., Adamy, J., Stahl, R. and Korner, E. "A Probabilistic Model for Binaural Sound Localization." *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics) IEEE Trans. Syst., Man, Cybern. B* 36.5 (2006): 982-94.

Yin, T. C. T., and Chan, J. C. K. "Interaural Time Sensitivity in Medial Superior Olive of Cat." *Journal of Neurophysiology* 64.2 (1990): 465-88.

Zurek, P. M. "The Precedence Effect and Its Possible Role in the Avoidance of Interaural Ambiguities." *J. Acoust. Soc. Am. The Journal of the Acoustical Society of America* 67.3 (1980): 952.