

Audiovisual decisions require a shorter period of evidence accumulation than unisensory decisions and rely most on visual information

A. M. van Harmelen

Assessor: dr. F. van Opstal

Supervisor: R. R. M. Tuij, MSc

Cognitive and Systems Neuroscience,
part of the Neurosciences collaboration
of the Swammerdam Institute for Life Sciences
Universiteit van Amsterdam 

Submission date: June 19th, 2020

Word count: 5937

Contact information

To contact the author, please use the following contact information:

Anna M. van Harmelen: anna@vanharmelen.me

Audiovisual decisions require a shorter period of evidence accumulation than unisensory decisions and rely most on visual information

van Harmelen, A. M.

19/06/2020

Abstract

Environmental information is collected and transduced by separate sensory organs, leading to multiple distinct modalities in the brain. Multisensory integration of these separate streams is then required in order to construct a unified percept of the external environment. While much is known about the behavioural advantages and requirements of multisensory decisions, it remains largely unknown how evidence accumulates for these decisions. The present study strives to understand whether multisensory evidence accumulation follows the same (temporal) rules as unisensory evidence accumulation, and how this is reflected in behaviour. Participants performed a two-alternative forced-choice decision task with moment-to-moment fluctuations of visual and auditory stimuli to determine how much each point in time contributes to a multisensory decision. Participants responded correctly to a larger percentage of multisensory audiovisual trials in comparison to either type of unisensory trials, and multisensory audiovisual and unisensory auditory trials were both performed more quickly than unisensory visual trials. A logistic regression indicated that both visual information and auditory information are more important at the start of multisensory evidence accumulation, and that visual information remains important for far longer than auditory information during the same perceptual decisions. This indicates that multisensory audiovisual decisions seem to rely more on visual information. It was therefore concluded that multisensory evidence accumulation does not follow the same temporal rules as unisensory evidence accumulation: multisensory evidence seems to be accumulated more quickly and is less sensitive to errors than evidence within a single modality. This conclusion is a first step towards understanding how a unified percept of the external environment is constructed.

Keywords

Audiovisual integration; Evidence accumulation; Multisensory integration; Unisensory auditory information; Unisensory visual information; Two-alternative forced choice

Introduction

Daily life is filled with perceptual decisions: determining the colour of a traffic light, locating your favourite breakfast cereal on the shelf in the grocery store or deciding whether you know (and therefore have to greet) the person walking on the other side of the street, are only some examples. While perceptual decisions can be made using information from a single sense, it is sometimes necessary to combine information from multiple modalities, a process known

as multisensory integration. Environmental information may contain noise and can even be self-contradicting, therefore behaviourally relevant information from only one sense can be too weak to base decisions on. For instance, when trying to understand someone in a crowded room it may help to not only listen, but also to watch the movement of the speaker's lips. In such situations, all environmental information is collected and transduced by separate sensory organs, leading to multiple distinct modalities in the brain. The integration of these separate streams

is then required in order to acquire a unified percept of the external environment (Tononi, 2008).

The concept of sensory integration is not new: it seems to be Aristotle who first introduced the concept of a common sense, or "*sensus communis*", which would monitor and coordinate the five known senses to form our integrated conscious experience (Guellai et al., 2019). This places the concept of sensory integration back even further than 300 B.C.E., however it was not until the beginning of the 20th century that studies into sensory processing were performed on more than one sense at a time (Fodor, 1983). Still, up until the end of the 20th century most sensory research remained unimodal, with more studies being done on the sense of sight than on all other sensory modalities combined (Hutmacher, 2019).

Multisensory integration is not only needed to create a unified experience of the outside world, but also increases fitness: perceptual decisions based on multisensory information are proven to be taken more quickly than unisensory decisions (Stein, 2012). This fast decision-making is not only beneficial to animals, but may even be necessary for their survival. Even when behaviourally relevant information from the environment is abundantly present within one modality, it may still not be sufficiently conclusive for reliable perceptual decisions: environmental information may be too ambiguous at any one moment or location. For example, when trying to determine if a predator is approaching or retreating, one can listen to see whether the noise it makes is growing louder or fainter, which involves listening for longer. Reliable sensory-motor decisions therefore not only require integration of environmental information between the senses, but also over space and time (Gold & Shadlen, 2007).

Some of the most important principles of the underlying mechanisms for multisensory integration have been established based on early behavioural and electrophysiological data from cats (Meredith, Nemitz, & Stein, 1987). Meredith et al. showed that auditory and visual stimuli, when delivered together in the correct temporal window, can be perceived as belonging to the same event even when temporal differences are present. The correct temporal window

is flexible in the sense that events that are spatially more separated allow for larger temporal disparities between auditory and visual stimuli. Sugita & Suzuki (2003) investigated the flexibility of this temporal window for co-presenting stimuli, and found that the brain reliably compensates for the delay between audio and visual inputs of stimuli that are further away, up to a distance of approximately 40 metres.

The combination of sensory information from different modalities is important for perceptual decision-making on a behavioural level: human studies show multisensory integration leads to improved stimulus detection (Driver & Spence, 1998; Frens, Van Opstal, & Van Der Willigen, 1995; Jaekl & Hris, 2009; McDonald, Teder-Saälejärvi, & Hillyard, 2000; Vroomen & De Gelder, 2000) and shortened reaction times (Gielen, Schmidt, & Van Den Heuvel, 1983; Hershenson, 1962). This multisensory integration between visual and auditory information is not only seen on a behavioural level: it is likewise seen on a physiological level, where neurons in the superior colliculus respond more strongly to a combination of visual and auditory stimuli presented closely together in space and time than to any stimulation of either modality alone (Stein, Huneycutt, & Meredith, 1988). More recent research showed that V1 activity can be modulated when visual and auditory stimulus features are modulated at the same rate (Ibrahim et al., 2016; Meijer, Montijn, Pennartz, & Lansink, 2017).

Originally it was thought that this temporal congruency between the senses was a prerequisite for multisensory integration (Meredith et al., 1987; Stein & Wallace, 1996; Van Atteveldt, Formisano, Blomert, & Goebel, 2007). However, contradictory evidence has been found: Raposo, Sheppard, Schrater, & Churchland (2012) instructed participants and trained rats to report the rate at which brief auditory and/or visual events were presented. In some trials the events in each modality were presented simultaneously, and in other trials both modalities were presented independently from one another (i.e. asynchronously). The experiment revealed that multisensory integration improved judgement in rats and humans in dependent as well as in independent presentation of both modalities, showing that multisensory enhancement

on a behavioural level is also present when auditory and visual information is presented asynchronously. This contrasts the temporal synchrony mechanism of multisensory integration. These findings were further corroborated by Tuip, van der Ham, van Opstal, & Lorteije (n.d.), who showed that integration of auditory and visual information does not depend on the synchronicity of both modalities and that asynchronous modulation of visual and auditory information does not affect the behavioural enhancement caused by multisensory integration.

Even though there has been an increase of knowledge on the behavioural advantages and requirements of multisensory integration, it remains largely unknown how evidence accumulates for these multisensory decisions. Although many models exist which attempt to explain how evidence accumulates in unisensory decisions, it is unclear to which extent these models also describe the way information is collected per modality and subsequently combined in multisensory decisions. As it is unclear how multisensory evidence is accumulated and integrated over time to form perceptual decisions and whether this deviates from decision-making within separate modalities, the following question will be investigated in this psychophysical study: does multisensory evidence accumulation follow the same (temporal) rules as unisensory evidence accumulation? And how is this reflected in behaviour? Based on a single-subject pilot study it is expected that (1) multisensory decisions are made quicker and more accurately than unisensory decisions, (2) and for audiovisual multisensory decisions it is expected that these decisions are mostly taken based on visual information.

To determine the validity of these hypotheses, a two-alternatives forced-choice decision (2AFC) task was created, where either two Gabor gratings of stochastically fluctuating contrast are presented on a screen (the visual condition), two pink-noise stimuli of fluctuating intensity are presented through headphones (the auditory condition), or both are presented simultaneously (the audiovisual condition). Participants were instructed to report on which side of the screen, headphones, or both, the stimulus with the highest intensity (contrast or volume, respectively)

appeared. The moment-to-moment fluctuations of the stimuli allowed us to identify which moments in time are important for the perceptual decision, by using a logistic regression model. If the first hypothesis is correct the data will reflect this by showing that participants have shorter reaction times and a higher percentage of correct decisions on multisensory trials than unisensory trials. If the second hypothesis is correct, the logistic regression will show that more timepoints of the visual information will contribute to the decision than of the auditory information. Additional analyses are performed per modality to evaluate whether there is a difference in evidence accumulation in unisensory versus multisensory trials.

Materials and methods

Participants

In total 8 participants (6 women, 2 men) performed the psychophysics task as described below. All participants were recruited through the personal network of the researcher, due to the recent COVID-19 pandemic. The participants were aged from 20 to 59 years old, with an average of 32 ± 16 (M \pm SD), and all were predominantly right-handed. All but one participant performed the task 3 to 6 times, while the remaining participant performed the task 21 times, adding up to a total of 50 sessions. All participants had no visual or auditory impairments, with the exception of corrected-to-normal vision. The study was approved by The Faculty Ethics Review Board of the Faculty of Social and Behavioural Sciences (ERB) of the University of Amsterdam and all participants provided informed written consent.

Procedure

The experiment consisted of a two-alternatives forced-choice decision task, developed and executed in MATLAB R2019b using the Psychtoolbox library (Brainard, 1997). The task consisted of in total 600 auditory, visual, and audiovisual trials presented in blocks, each block consisting of eight trials in which participants were asked to respond within a certain amount of time. In each trial a target and a distractor stimulus were presented, to which a keyboard response was required (figure 1D). After each trial a feedback signal was produced, which was either the

text ‘correct!’ or ‘incorrect!’, depending on whether the provided keyboard response was correct or not, respectively. Participants started with one visual and auditory practice block.

The auditory blocks consisted of two pink-noise stimuli, presented to each ear separately over headphones of which the volume fluctuated every 50 ms (figure 1A). Participants were instructed to press the ‘f’ key if the stimulus with the highest volume (i.e. the target) was presented to the left ear, and the ‘j’ key if the target was presented to the right ear. It is important to note that the target stimulus did not necessarily always have a higher volume than the distractor, due to the 50 ms fluctuations (figure 2). However, a trend could be observed as to which stimulus has a higher volume on average.

The visual blocks consisted of two black and white Gabor gratings, of which the contrast fluctuated every 50 ms (figure 1B). Participants were instructed to press the ‘f’ key if the grating with the highest contrast (i.e. the target) was on the left-hand side of the screen, and the ‘j’ key if the target was on the right-hand side of the screen. In the audiovisual blocks participants were shown the Gabor gratings in combination with the

pink-noise stimuli (figure 1C), and were asked to indicate (using the same keyboard keys as above) which combination of grating and pink-noise had a higher intensity, i.e. on which side the visual stimulus had a higher contrast and the auditory stimulus had a higher volume.

For the data-analysis, all participants needed to responded incorrectly to a substantial number of trials, therefore the difficulty of the trials had to be centred round the perceptual threshold of each participant. This was done by changing the difficulty of the trials in real-time, using a 1-up-2-down staircasing method (PsychStairCase of the Psychophysics Toolbox for MATLAB; Brainard, 1997). Each block consisted of trials of difficulty 1, 2 and 3: the difficulty of a trial was randomly generated, but approximately equal numbers of each difficulty were present in each block. The target stimuli in trials of difficulty 2 had an average contrast value of 60%, with a maximum variation of 14 percentage points, resulting in a range from 46% to 74%. This led to a response accuracy of $0.61\% \pm 0.03$ percentage points (M±SD). The target stimuli of trials with difficulty 1 always had average contrast values of 1.5 times the target contrast in trials with difficulty 2.

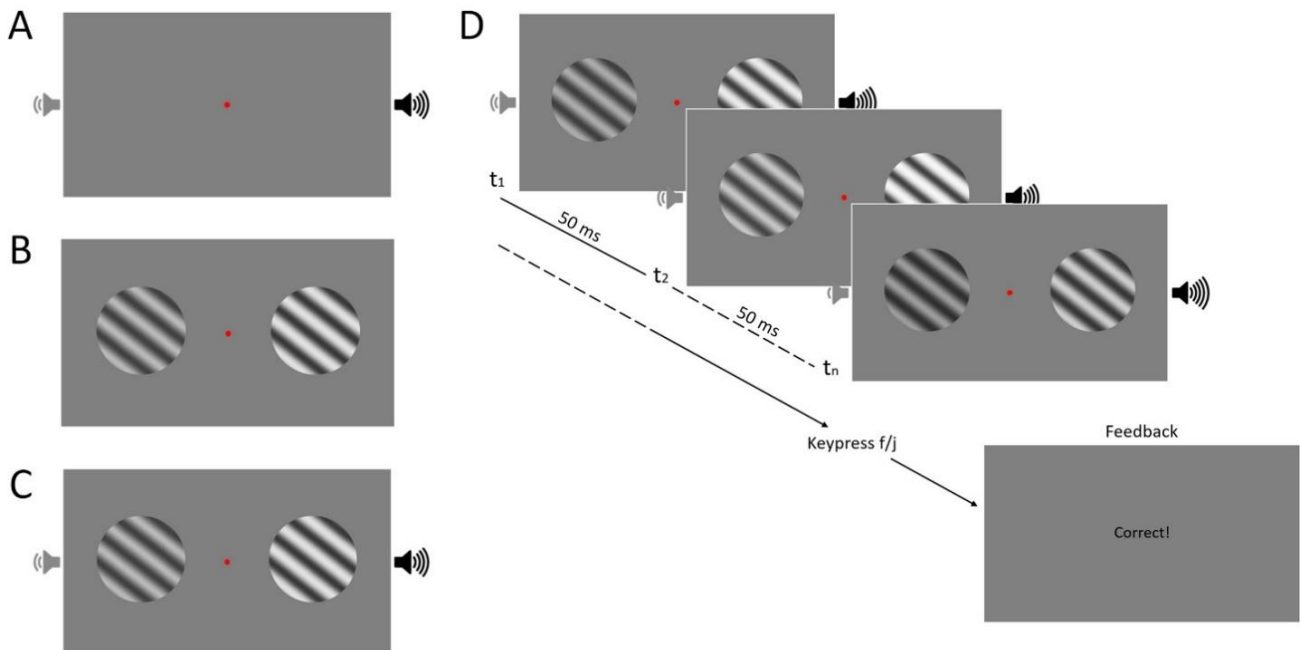


Figure 1: Overview of trial types and flowchart of the 2AFC task. Subfigures A, B and C show one frame of an auditory, visual and audiovisual trial, respectively, in the current experiment. Subfigure D shows the procedure of one trial, where all stimuli (visual and auditory) fluctuate every 50 ms, until a key-press is provided. The key-press is followed by textual feedback.

For trials of difficulty 3 this was 0.5 times that target contrast value.

The distractor stimulus in all trials had an average contrast of 80% of the target contrast value of that trial difficulty, but if a trial was generated of difficulty 2 and all responses since the start of that block were incorrect, the contrast value of the distractor in trial type 2 decreased by two percentage points (e.g. from 60% to 58%). If a trial was generated of difficulty 2 and the two previous responses were correct, the average contrast value of the distractor stimulus increased by two percentage points. As trials of difficulties 1 and 3 are based on the contrast values of difficulty 2, this staircasing procedure resulted in a two-percentage point increase or decrease of the average contrast in all difficulty levels.

Data-analysis

As a first step in the data-analysis, differences in reaction times were investigated between participants, between difficulty levels and between modality of the trials. These same analyses were performed on the

percentage of correctly answered trials. Multiple one-way repeated measures ANOVA's were performed in R studio (version 1.2.5033) to compare each analysis described here. Additionally, a correlation test using the Pearson correlation coefficient was performed on all participants to establish whether any training effect was present, in spite of the staircased difficulty. All assumptions of normality were checked using a Shapiro-Wilk test; equal variances were tested using Levene's test.

A logistic regression analysis was performed on each participant separately, and on the group as a whole, in MATLAB R2019b. The logistic regression estimates the weight of the fluctuations of all stimuli at each time-point, by using the following formula's:

$$Y_A = [1 + \exp(-(\beta_0 + \beta_{1_t}F_{AT,t} + \beta_{2_t}F_{AD,t} + \beta_{3_k}P_k))]^{-1}$$

$$Y_V = [1 + \exp(-(\beta_0 + \beta_{1_t}F_{VT,t} + \beta_{2_t}F_{VD,t} + \beta_{3_k}P_k))]^{-1}$$

$$Y_{AV} = [1 + \exp(-(\beta_0 + \beta_{1_t}F_{VT,t} + \beta_{2_t}F_{VD,t} + \beta_{3_t}F_{AT,t} + \beta_{4_t}F_{AD,t} + \beta_{5_k}P_k))]^{-1}$$

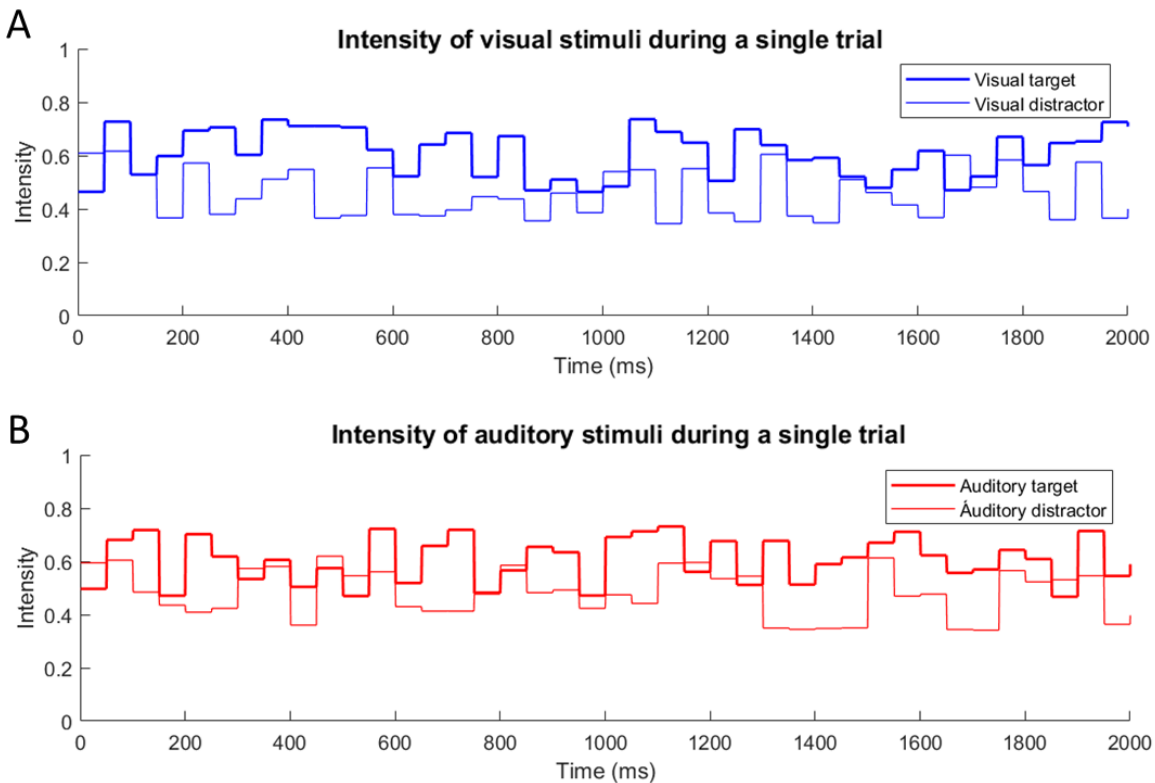


Figure 2: An example of the intensity of the target and distractor stimuli presented in a single trial. Subfigures A and B respectively show the fluctuating intensity of the visual and auditory stimuli during a single trial. The target stimulus is represented by the thicker line. Stimulus intensity fluctuates every 50 ms in both modalities.

Y_A , Y_V and Y_{AV} represent the response accuracies of, respectively, unisensory auditory, unisensory visual and multisensory audiovisual trials (i.e. correct or incorrect). β_0 reflects overall accuracy, all following β -values represent the fitted coefficients for each corresponding predictor. $F_{VT,t}$ represents the normalized fluctuations of the visual target contrast at each timepoint t , $F_{VD,t}$ represents the normalized fluctuations of the visual distractor contrast at each timepoint t , $F_{AT,t}$ represents the normalized fluctuations of the auditory target contrast at each timepoint t , $F_{AD,t}$ represents the normalized fluctuations of the auditory distractor contrast at each timepoint t and P_k is a dummy variable identifying each participant, to investigate whether participant number is also a significant predictor. A visual analysis was performed by viewing the evidence accumulation in different trial types and different participants.

Results

Reaction times

The goal of the current experiment is to determine how multisensory evidence is accumulated and integrated over time to form perceptual decisions and whether this deviates from unisensory decision-making within separate modalities. First, we compared the average reaction times of all participants to determine whether normalization is required for further analysis of the reaction times. Eight participants performed the task multiple times to end up with an average of 3675 ± 3209 ($M \pm SD$) performed trials per person. Figure 3 shows the average reaction times of all participants, which were respectively 0.41 ± 0.09 , 0.95 ± 0.32 , 0.66 ± 0.15 , 1.09 ± 0.33 , 1.11 ± 0.35 , 0.71 ± 0.30 , 0.46 ± 0.14 and

0.44 ± 0.13 . It should be noted that participants 4 and 5, who show the two highest reaction times, are also the oldest. Both participants are 59 years old, while all other participants have an age ranging from 20 to 28. The data did not meet the assumptions of normality and equal variances, hence a Kruskal-Wallis test was performed which showed a significant difference was present between at least two groups ($\chi^2 = 18344$, $df = 7$, $p < 2.2e-16$). All possible participant combinations (28 in total) were tested using a post-hoc Nemenyi test to show that all average reaction times were significantly different, except for the combinations of 4 and 5 ($q = 0.523$, $p = 0.999$) and 3 and 6 ($q = 3.48$, $p = 0.214$) (all other p-values were smaller than $0.05/28 = 0.00179$).

Because all participants performed the current task multiple times, it was deemed necessary to establish whether a training effect is present in the data. For all participants a correlation test was therefore performed, which showed that only participants 1 and 6 showed any significant training effect (respectively: $t = -5.74$, $df = 18$, $p = 1.94e-05$ and $t = -6.12$, $df = 2$, $p = 0.026$). The average reaction times of these participants over the number of performed sessions is shown in subfigures A and B of figure 4, together with a trendline. It is assumed that this learning curve does not need to be a problem for the analysis of the data, as all trials within one session are grouped together and it is expected that within a session differences can still be seen in both the reaction times and performance on trials of either different difficulty or modality. Therefore participants 1 and 6 were not excluded from any further data analysis.

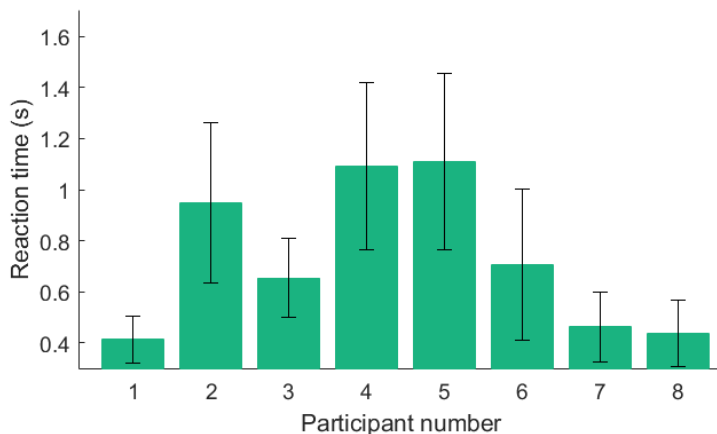


Figure 3: Average reaction times for all participants.

The bars in this figure show the average reaction times per participant, black error bars show one standard deviation from the mean. Average reaction times per participant are, respectively, 0.41 ± 0.09 , 0.95 ± 0.32 , 0.66 ± 0.15 , 1.09 ± 0.33 , 1.11 ± 0.35 , 0.71 ± 0.30 , 0.46 ± 0.14 and 0.44 ± 0.13 ($M \pm SD$). All possible combinations of participants (28 in total) had significantly different reaction times ($p < 0.05/28 = 0.00179$), except for the combinations of 4 and 5 ($p = 0.999$) and 3 and 6 ($p = 0.214$).

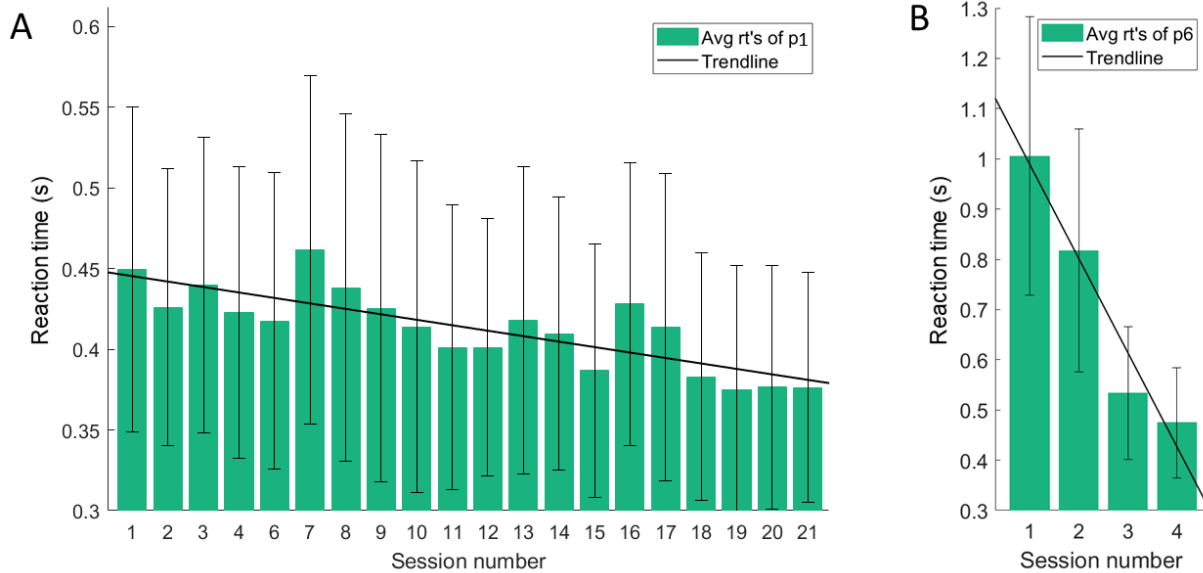


Figure 4: Average reaction times for participants 1 and 6 over all sessions. Subfigures A and B respectively show the average reaction times of participant 1 and 6 per recorded session. The bars show average reaction times, while the black error bars display one standard deviation of the mean. The continuous black lines are linear trendlines. The correlation found between session number and reaction time was significant for both participant 1 and 6 (respectively, $p = 1.94e-05$ and $p = 0.026$).

To determine whether the applied staircasing method led to significantly different reaction times, a following analysis was performed on the reaction times for each difficulty level (figure 5). Because nearly all participants had significantly different reaction times, all reaction times were first normalised per session by dividing all reaction times by the largest reaction time from that session, and averaged for each difficulty level. Note that trials of all modality types were grouped together to analyse the three difficulty levels. An average reaction time for each difficulty level was calculated for each session, which were then again averaged to show the average reaction time for trials of difficulty 1 (0.50 ± 0.07 , $M \pm SD$), 2 (0.51 ± 0.07) and 3 (0.52 ± 0.07). Following three Shapiro-Wilk tests the data met all assumptions of normality (difficulty 1: $W = 0.979$, $p = 0.505$; difficulty 2: $W = 0.971$, $p = 0.264$; difficulty 3: $W = 0.975$, $p = 0.374$) and a Levene's test revealed equal variances were present ($df = 2$, $F = 0.0593$, $p = 0.942$), therefore, a one-way repeated measures ANOVA was performed. The ANOVA revealed that all three group means did not differ significantly from one another ($df = 2$, $SS = 0.0047$, $F = 0.48$, $p = 0.619$).

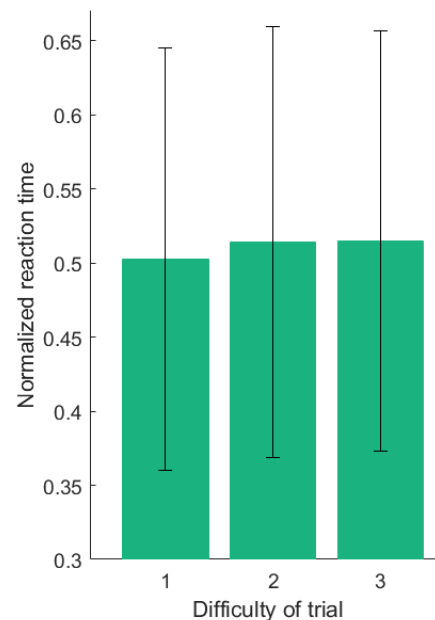


Figure 5: Average reaction times for each trial difficulty. The bars in the figure show the average reaction times per trial difficulty, normalized per session by dividing by the largest reaction time. The average normalized reaction time for difficulty 1 is 0.50 ± 0.07 ($M \pm SD$), for difficulty 2 0.51 ± 0.07 and for difficulty 3 0.52 ± 0.07 . Black error bars show one standard deviation from the mean. No significant differences were present.

To infer whether there is a significant difference in how fast participants respond to trials of different modality, the normalized reaction times were also averaged over all trial types (figure 6). Note that trials of all difficulty levels were grouped together to analyse the three modality types. This showed unisensory auditory trials were responded to fastest (0.49 ± 0.15), followed closely by multisensory audiovisual trials (0.50 ± 0.14). The response to visual trials was the slowest (0.54 ± 0.14). All data met the assumptions of normality (auditory: $W = 0.981$, $p = 0.611$, visual: $W = 0.973$, $p = 0.331$, audiovisual: $W = 0.986$, $p = 0.827$) and equal variances ($df = 2$, $F = 1.60$, $p = 0.205$); hence a one-way repeated measures ANOVA was applied. The ANOVA revealed that at least one of the three group means differed significantly from the others ($df = 2$, $SS = 0.0665$, $F = 13.3$, $p = 7.75e-06$). Tukey's HSD test was applied post-hoc to investigate which groups differ significantly, this revealed that both the unisensory auditory trials and the multisensory audiovisual trials differ from the unisensory visual trials (respectively: $q = 7.12$, $p = 1.65e-05$ and $q = 5.96$, $p = 3.25e-04$). The unisensory auditory and the multisensory audiovisual trials did not differ significantly from one another ($q = 1.15$, $p = 0.716$).

Validity of applied staircasing method

To determine whether the applied staircasing procedure was successful, an analysis was performed to determine whether trials of different difficulty led to a significantly different performance accuracy (figure 7). Here a clear trend can be seen: participants have the highest percentage of correct responses to trials of difficulty 1 (0.70 ± 0.06), followed by trials of difficulty 2 (0.61 ± 0.03) and difficulty 3 (0.58 ± 0.05), in that order. While all three groups had normally distributed data (difficulty 1: $W = 0.988$, $p = 0.906$, difficulty 2: $W = 0.976$, $p = 0.395$, difficulty 3: $W = 0.968$, $p = 0.207$), the assumption of equal variances was violated ($df = 2$, $F = 8.5$, $p = 3.1e-4$), hence the non-parametric Friedman test was applied. The Friedman test showed a significant difference was present ($\chi^2 = 76.41$, $df = 2$, $p < 2.2e-16$). To gain further insight the Nemenyi test was performed post-hoc ($\alpha = 0.05/3 = 0.0167$, following

the Bonferroni correction to compensate for multiple comparisons). This revealed the percentage of correct trials was significantly higher in trials of difficulty 1 when compared to trials of difficulty 2 ($q = 8.57$, $p = 4.1e-09$) and difficulty 3 ($q = 12.0$, $p = 2.9e-14$), but trials of difficulty 2 and 3 did not differ significantly ($q = 3.43$, $p = 0.041$). This finding shows that the staircasing of the trials does produce at least two significantly different difficulty levels, with indication of a third level. These findings imply that the applied staircasing procedure was successful, and it is therefore reasonable to assume all following findings based on the assumption of these three difficulty levels are credible.

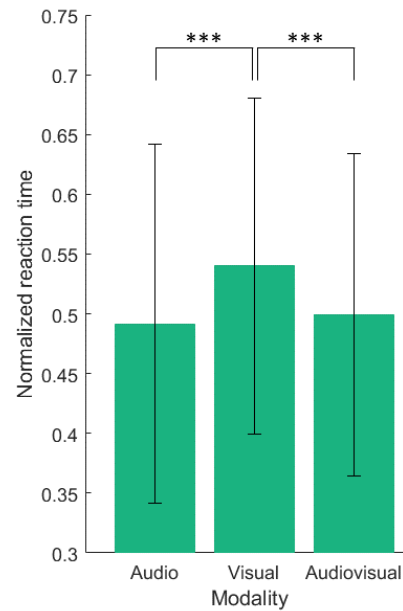


Figure 6: Average reaction times for each trial modality. The bars in the figure show the average reaction times per trial modality, normalized per session by dividing by the largest reaction time. Black error bars show one standard deviation from the mean. The average normalized reaction time for the auditory, visual and audiovisual trials are, respectively, 0.49 ± 0.15 ($M\pm SD$), 0.54 ± 0.14 and 0.50 ± 0.14 . All significant differences are marked with an asterisk; three asterisks indicate a significant difference of $p < 0.001$.

Differences in performance accuracy between modalities

To determine whether there is any behavioural reaction to trials of different modality (respectively: unisensory auditory, unisensory visual and multisensory audiovisual), two more analyses were performed: a comparison of normalized reaction times

for all modalities, and a comparison of percentage of correct trials for all modalities. The percentages of correctly answered trials for all modalities, figure 8, show a clear trend where performance is worst on unisensory auditory trials (0.62 ± 0.04), a slight improvement in performance is seen in unisensory visual trials (0.63 ± 0.05) and performance is best on multisensory audiovisual trials (0.70 ± 0.07). The data of all groups were distributed normally (auditory: $W = 0.981$, $p = 0.609$, visual: $W = 0.974$, $p = 0.351$, audiovisual: $W = 0.976$, $p = 0.423$), but the assumption of equal variances was violated ($df = 2$, $F = 8.3$, $p = 4.0e-4$), hence the analysis continued non-

parametrically. Since the data are paired (i.e. all participants performed trials of all modalities), the Friedman test was applied, which revealed significant differences were present ($\chi^2 = 43.70$, $df = 2$, $p < 3.24e-10$). To investigate these differences further, the Nemenyi test was applied post-hoc. This revealed that the unisensory auditory condition differed significantly from the audiovisual condition ($q = 8.50$, $p = 5.5e-09$), as did the unisensory visual condition ($q = 7.57$, $p = 2.6e-07$). The two unisensory conditions did not differ significantly from one another ($q = 0.929$, $p = 0.79$).

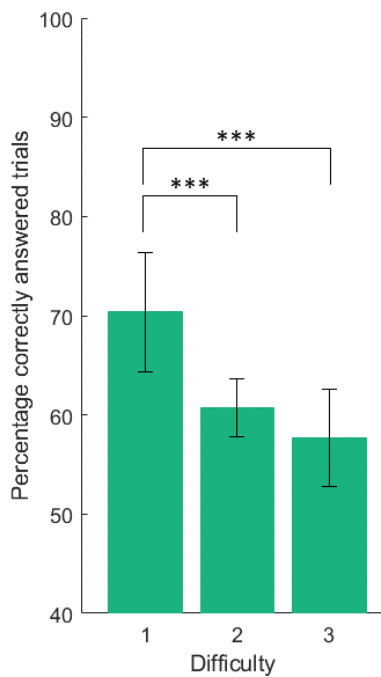


Figure 7: Percentage of correctly answered trials for each trial difficulty. The bars in this figure represent the average percentage of correctly answered trials after averaging over each of the 49 sessions. Black error bars show one standard deviation from the mean. The percentage of correctly answered trials for difficulty 1, 2 and 3 was respectively 0.70 ± 0.06 ($M\pm SD$), 0.61 ± 0.03 and 0.58 ± 0.05 . All significant differences ($p < 0.0167$) are marked with asterisks; three asterisks indicate a significant difference of $p < 0.001$.

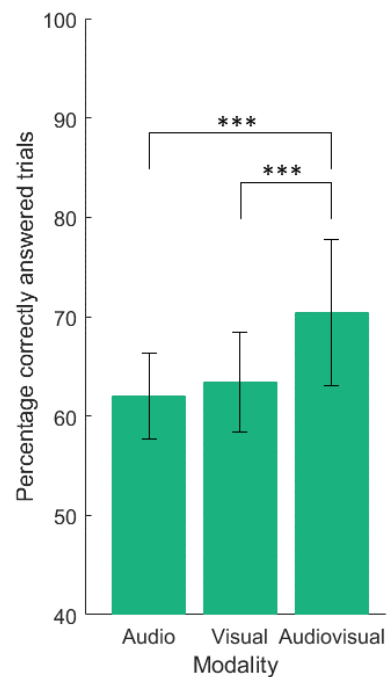


Figure 8: Percentage of correctly answered trials for each trial modality. The bars in the figure represent the percentage of correctly answered trials after averaging over each of the 49 sessions. Black error bars show one standard deviation from the mean. The percentage of correctly answered trials was 0.62 ± 0.04 ($M\pm SD$) for the auditory trials, 0.63 ± 0.05 for the visual trials and 0.70 ± 0.07 for the audiovisual trials. All significant differences are marked with an asterisk; three asterisks indicate a significant difference of $p < 0.001$.

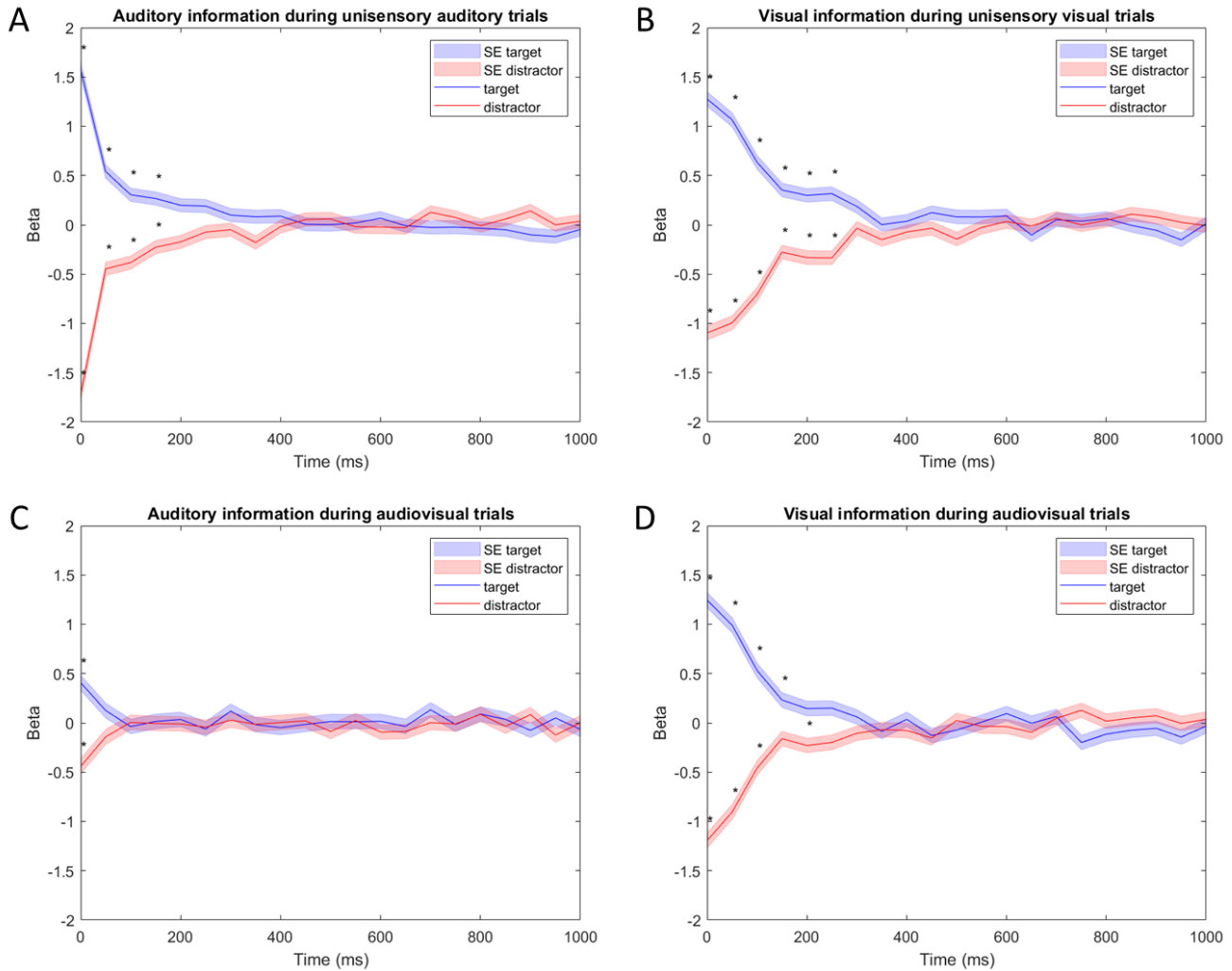


Figure 9: Logistic regression analyses of auditory and visual data in unisensory and audiovisual trials. Subfigures A and B respectively show the relative importance of auditory and visual information during unisensory trials; subfigures C and D show the relative importance of auditory and visual information during audiovisual trials. Solid blue lines show the average beta value of the information provided by the target stimulus for each 50 ms timepoint, with the shaded blue area indicating one standard error of the mean. Solid red lines show the average beta value of the information provided by the distractor stimulus for each 50 ms timepoint, with the shaded red area indicating one standard error of the mean. All timepoints that contribute significantly to a decision are marked with a black asterisk ($p < 0.0024$, due to Bonferroni correction). The beta-values of the first timepoint are 1.58 and 0.42 for the auditory information in, respectively, the unisensory and multisensory conditions. For visual information these beta-values are 1.30 and 1.26.

Unisensory versus multisensory evidence accumulation

As a next step a regression analysis was performed on the data, using the formulas as shown in the methods section. Figure 9 shows the relative importance of each 50 ms fluctuation of stimulus contrast/intensity in each modality for the perceptual decision made in each trial. Subfigures A, B, C and D respectively show the relative importance of all timepoints of auditory information during unisensory trials, visual information during

unisensory trials, auditory information during audiovisual trials and visual information during audiovisual trials. All timepoints that contribute significantly to a decision are marked with a black asterisk ($p < 0.05/21 = 0.0024$, as each trial consists of 21 time samples). A visual inspection of the data reveals that auditory information is significantly important for four times as long in unisensory trials (4 timepoints = 200 ms) than in audiovisual trials (50 ms). For visual information this difference between unisensory trials

and audiovisual trials is absolutely and relatively smaller (300 ms in unisensory trials vs. 200 ms in audiovisual trials).

When comparing auditory and visual information during audiovisual trials (subfigures C and D of figure 9, respectively), it is revealed that visual information significantly contributes to the decision for longer than auditory information (200 ms vs. 50 ms), revealing that it is possible that audiovisual decisions are largely based on visual information. When comparing the beta-values at the first timepoint a similar trend is discovered: in unisensory trials auditory and visual information both have beta-values around 1.5 (for the target stimuli, respectively, $\beta = 1.58$ and $\beta = 1.30$), indicating that the information is of high relative importance. In multisensory trials the beta-value of the visual information remains relatively unchanged ($\beta = 1.26$), however, the beta-value of the auditory information has reduced threefold ($\beta = 0.42$).

However, closer inspection of this theory is necessary, as it is not known whether this is a universally applied strategy or whether different participants employ different perceptual strategies. Therefore, the regression analysis was also performed on data from each single participant. A visual analysis of these figures (appendix A) revealed that all participants who showed significant use of the provided sensory evidence based audiovisual decisions more on visual information than on auditory information, with the exception of participant 8, who showed no significant use of visual information for audiovisual decisions. Participant 8 did show significant use of visual information in unisensory visual trials, implying that the strategy found in audiovisual trials is an actual strategy and not the result of visual impairments or incomplete data. It is therefore important to note that personal differences in audiovisual integration strategies can be present, even though the majority of participants showed a preference for the use of visual information.

The logistic regression formulas also contained a dummy variable identifying each participant, this predictor variable was insignificant for all time-points in the unisensory visual and multisensory audiovisual trials. For the unisensory auditory trials, participant

number was a significant predictor for all timepoints (all β -values > 0.0310 , all p -values < 0.001). This indicates that individual differences between participants are only a significant predictor for the accuracy on unisensory auditory trials, which is in line with a quick visual analysis of the logistic regressions per participant (appendix A): more differences can be seen between participants in the way information is used in auditory trials than in visual or audiovisual trials.

Flexibility of multisensory evidence accumulation

To understand whether this disparity in the importance of the two information streams is universally present or based on the presented information in any given situation, a further analysis was performed. If the multisensory combination strategy employed is flexible one would expect a higher use of information from one modality if the evidence provided in the other modality is inconclusive. Evidence is defined as the absolute contrast/intensity difference between target and distractor. When information from both modalities is relatively conclusive, an equal combination can be made to reach a perceptual decision as quickly as possible. To see whether the multisensory combination strategy is indeed flexible, the same logistic regressions as above were applied to trials where in one modality relatively high or low evidence was present (high and low evidence were defined as trials where the evidence at the first timepoint fell into the top or bottom quartile, respectively).

By comparing subfigures A and B of figure 10, we see auditory information is in fact slightly more important as it has a higher beta-value when the visual information at the start of a trial is inconclusive ($\beta = 0.40$ for high visual evidence, in comparison to $\beta = 0.42$ for low visual evidence). Interestingly, the auditory information is not used for longer. The same trend can be seen in visual information to a lesser extent when comparing subfigures C and D of figure 10. The beta-values of visual information in trials with inconclusive auditory information are slightly higher ($\beta = 1.03$ for high auditory evidence, in comparison to $\beta = 1.24$ for low auditory evidence), and the information of the visual target stimulus is also used for 50 ms longer.

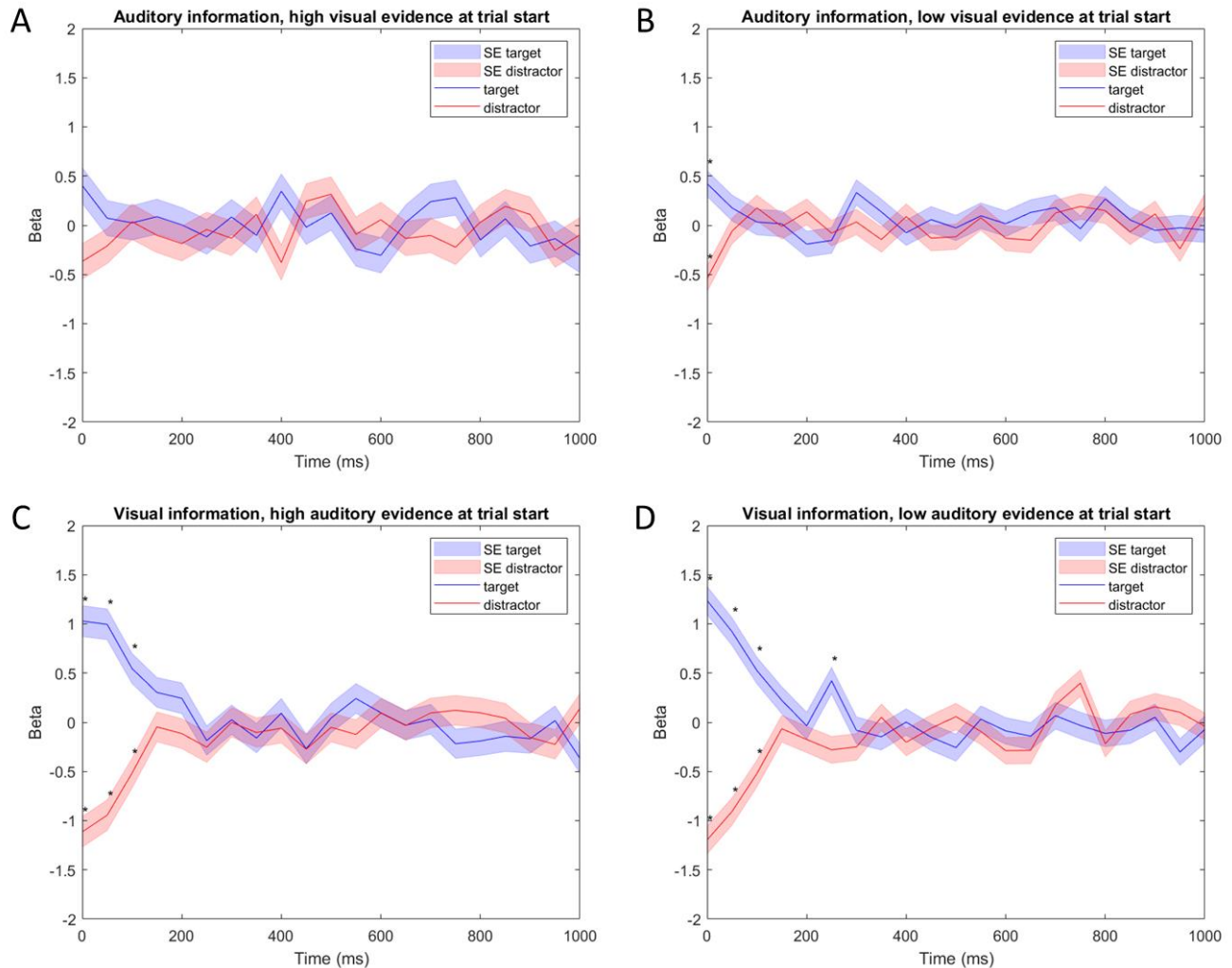


Figure 10: Logistic regression analyses of auditory and visual data in multisensory trials with high or low evidence in the other modality at the trial starts. Subfigures A and B show the relative importance of auditory information during multisensory trials with, respectively, high or low visual evidence at the start of the trial; subfigures C and D show the relative importance of visual information during multisensory trials with, respectively, high or low auditory evidence at the start. Solid blue lines show the average beta value of the information provided by the target stimulus for each 50 ms timepoint, with the shaded blue area indicating one standard error of the mean. Solid red lines show the average beta value of the information provided by the distractor stimulus for each 50 ms timepoint, with the shaded red area indicating one standard error of the mean. All timepoints that contribute significantly to a decision are marked with a black asterisk ($p < 0.0024$, due to Bonferroni correction). The beta-values of the first timepoint are 0.40 and 0.42 for auditory information in trials where the visual evidence is, respectively, high and low. For visual information these values are 1.03 and 1.24.

Discussion

The current study set out to determine how multisensory evidence is accumulated and integrated over time to form perceptual decisions and whether this deviates from unisensory decision-making within separate modalities. To answer these questions, behavioural data was collected and multiple logistic regression analyses were performed. The behavioural

results revealed that participants have a significant increase in performance on multisensory audiovisual trials in comparison to either type of unisensory trials, indicating that it is useful to integrate two streams of information into one perceptual decision as it may lead to better stimulus detection in real-life situations. Concerning the reaction times a different trend was seen: multisensory audiovisual and unisensory auditory trials were both performed much quicker than

unisensory visual trials, suggesting that in humans adding an auditory information stream makes the decision process quicker if the first information stream is visual. Surprisingly, vice versa this is not the case.

The logistic regression analysis revealed that auditory and visual information is collected and used for unisensory perceptual decisions for approximately the same amount of time, with visual evidence being accumulated over a slightly longer period (figure 9). This is in line with the unisensory behavioural results, where the unisensory visual perceptual decisions take slightly longer to be made than the unisensory auditory decisions. However, in multisensory trials a clear disparity is seen: while visual evidence is accumulated for only a slightly shorter time in multisensory trials than in unisensory trials, auditory information contributes only very briefly to the perceptual decision made in multisensory trials. Moreover, the same trend is seen when comparing the beta-values of the first timepoints in each trial type: auditory information is significantly less important in multisensory trials compared to unisensory trials, while the importance of visual information remains unchanged between the two trial types. An additional logistic regression analysis revealed that evidence from a single modality becomes more important for the multisensory decision when evidence from the other modality is inconclusive at the start of the trial (figure 10). This indicates that in multisensory decisions not all modalities are always equally important: it seems that the combination of modalities is flexible in the sense that ambiguity in one modality may be compensated when information is presented more clearly in another modality. This could explain the improved accuracy seen in multisensory decisions in comparison with unisensory decisions.

The behavioural findings lead to the conclusion that multisensory decisions are more accurate than both unisensory decisions, and are made quicker than unisensory visual decisions, but not than unisensory auditory decisions. The logistic regressions showed that visual information is more important than auditory information at the start of multisensory evidence accumulation, and remains important for far longer than auditory information during the same perceptual decisions. These findings explain the shortened

reaction time seen in multisensory decisions, as the logistic regression revealed that evidence from both modalities is accumulated for a shorter period of time in multisensory trials compared to unisensory trials. A possible explanation for this shorter period of evidence accumulation is that evidence from two modalities, when combined, may sooner lead to the same amount of evidence one could get from one modality, since both modalities provide evidence at a certain fixed rate (in this experiment, once every 50 ms). However, this does not explain why multisensory decisions are not also made quicker than unisensory auditory decisions. This could possibly be because auditory information is less complex than visual information (Hutmacher, 2019), since there would be no benefit from adding a more complex information stream to a simple one with regard to response speed. However, accuracy is still increased in multisensory decisions compared to unisensory auditory decisions, so there is a benefit from adding a second (more complex) information stream.

All findings are in line with the expectations given at the beginning of this thesis, specifically that multisensory decisions are made more accurately than unisensory decisions and that these decisions are mostly taken based on visual information. Only the speed at which multisensory decisions are made was unexpected: it was predicted that multisensory audiovisual decisions would be made quicker than either unisensory decisions, but auditory unisensory decisions were performed non-significantly quicker than multisensory decisions. To answer the research questions posed in the introduction one could state that multisensory evidence accumulation does not follow the same temporal rules as unisensory evidence accumulation: multisensory evidence seems to be accumulated quicker and is less sensitive to errors than evidence within a single modality, possibly due to the flexible combination of all available modalities.

Our novel experimental design used moment-to-moment fluctuations of intensity and contrast in, respectively, auditory and visual stimuli to determine which timepoints are important for a perceptual decision. This is necessary as a first step to unravel the temporal dynamics of multisensory decision-making in a controlled manner. Furthermore, participants were

not explicitly instructed to use both sets of stimuli when both modalities are presented, which approximates real world environments where multiple modalities for one perceptual decision are present. Additionally, to obtain conclusions more applicable to the real world, it would be interesting and valuable to investigate multisensory integration with stimuli that better approximate real-life. The low ecological validity of the stimuli in the current study was necessary to be able to fluctuate them in a controlled manner. Nonetheless, by using more ecologically valid stimuli we can further understand whether multisensory evidence accumulation works similarly when it is directly tied to a higher-level goal. For instance, a virtual reality locator task can be used where both sound and images are required to precisely pinpoint the location of a certain object, to further understand whether multisensory evidence accumulation works similarly when it is directly tied to a higher-level goal.

While the current set-up enabled us to evaluate which timepoints are important for a perceptual decision, it would also allow for analysis of which aspects of the stimulus are used in order to determine which stimulus is the target. It is possible that the contrast values of both stimuli are separately averaged over a certain time period, or it could be that the contrast differences between both stimuli indicate which of the two is the target stimulus. The data from the current study is sufficient to at least determine whether either of these strategies is employed, hence further analysis of the data could reveal what type of perceptual strategy participants employ to determine which stimulus has a higher contrast/intensity.

Furthermore, concerning the data-analysis another change could be made in the future. To perform the logistic regressions a dummy variable was added which shows whether participant number is a significant predictor for the accuracy on a certain trial. However, this dummy variable does not compensate for the repeated measures component of the current set-up. Therefore, in future use of this experimental set-up it is advised to use a general linear mixed model

(GLMM) for repeated measures. This had the additional value that the linear predictor also contains random effects, in addition to the usual fixed effects which are present in the logistic regressions.

Further attention could also be focussed on the findings from participant 8, which do not conform to the conclusions based on the data from the entire group, as participant 8 uses auditory information for multisensory decisions instead of visual information. Insight could be gained into whether multisensory integration strategies differ from person to person or maybe even from moment to moment.

More research could also be done into where this multisensory integration observed in the present study takes place in the brain. It is possible that V1 and A1 neurons show early modulation when both stimuli are presented together, resulting in changed evidence accumulation in the very first steps of perception (Watkins, Shams, Tanaka, Haynes, & Rees, 2006). It is also possible that this integration arises later, in higher order areas, as this specific task requires some temporal summation, which might be a challenge for primary sensory areas (Chaplin, Rosa, & Lui, 2018; Li, Xi, Zhang, Liu, & Tang, 2019).

Summarising, this study investigated the temporal aspects of evidence accumulation in audiovisual multisensory trials, and proved consistent with previous research with regard to multisensory perceptual decisions being faster and more accurate than unisensory decisions. We extended this knowledge by showing that multisensory decisions are based more on visual information. Because the stimuli used in the current experimental set-up have a low ecological validity, further investigation is recommended into how multisensory evidence accumulation works when it is directly tied to a higher-level goal. To conclude, this thesis has proven significant differences are present in the way evidence is accumulated in each modality to be used in unisensory versus multisensory decisions, which is the first step into understanding how a unified percept of the external environment is constructed.

Literature

- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, Vol. 10, pp. 433–436. <https://doi.org/10.1163/156856897X00357>
- Chaplin, T. A., Rosa, M. G. P., & Lui, L. L. (2018, October 26). Auditory and visual motion processing and integration in the primate cerebral cortex. *Frontiers in Neural Circuits*, Vol. 12. <https://doi.org/10.3389/fncir.2018.00093>
- Driver, J., & Spence, C. (1998). Cross-modal links in spatial attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373), 1319–1331. <https://doi.org/10.1098/rstb.1998.0286>
- Fodor, J. A. (1983). *The modularity of mind : an essay on faculty psychology*. Cambridge, Mass: MIT Press.
- Frens, M. A., Van Opstal, A. J., & Van Der Willigen, R. F. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Perception & Psychophysics*, 57(6), 802–816. <https://doi.org/10.3758/BF03206796>
- Gielen, S. C. A. M., Schmidt, R. A., & Van Den Heuvel, P. J. M. (1983). On the nature of intersensory facilitation of reaction time. *Perception & Psychophysics*, 34(2), 161–168. <https://doi.org/10.3758/BF03211343>
- Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1), 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
- Guellaï, B., Callin, A., Bevilacqua, F., Schwarz, D., Pitti, A., Boucenna, S., & Gratier, M. (2019). Sensus Communis: Some Perspectives on the Origins of Non-synchronous Cross-Sensory Associations. *Frontiers in Psychology*, 10(MAR), 523. <https://doi.org/10.3389/fpsyg.2019.00523>
- Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology*, 63(3), 289–293. <https://doi.org/10.1037/h0039516>
- Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, 10, 2246. <https://doi.org/10.3389/fpsyg.2019.02246>
- Ibrahim, L. A., Mesik, L., Ji, X., Fang, Q., Li, H., Li, Y., ... Tao, H. W. (2016). Cross-Modality Sharpening of Visual Cortical Processing through Layer-1-Mediated Inhibition and Disinhibition. *Neuron*, 89(5), 1031–1045. <https://doi.org/10.1016/j.neuron.2016.01.027>
- Jaekl, P. M., & Hris, L. R. (2009). Sounds can affect visual perception mediated primily by the pvocellul pathway. *Visual Neuroscience*, 26(5–6), 477–486. <https://doi.org/10.1017/S0952523809990289>
- Li, Q., Xi, Y., Zhang, M., Liu, L., & Tang, X. (2019). Distinct Mechanism of Audiovisual Integration With Informative and Uninformative Sound in a Visual Detection Task: A DCM Study. *Frontiers in Computational Neuroscience*, 13, 59. <https://doi.org/10.3389/fncom.2019.00059>
- McDonald, J. J., Teder-Saälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, 407(6806), 906–908. <https://doi.org/10.1038/35038085>
- Meijer, G. T., Montijn, J. S., Pennartz, C. M. A., & Lansink, C. S. (2017). Audiovisual modulation in mouse primary visual cortex depends on cross-modal stimulus configuration and congruency. *Journal of Neuroscience*, 37(36), 8783–8796. <https://doi.org/10.1523/JNEUROSCI.0468-17.2017>

- Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, 7(10), 3215–3229. <https://doi.org/10.1523/jneurosci.07-10-03215.1987>
- Raposo, D., Sheppard, J. P., Schrater, P. R., & Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *Journal of Neuroscience*, 32(11), 3726–3735. <https://doi.org/10.1523/JNEUROSCI.4998-11.2012>
- Stein, B. E. (2012). *The new handbook of multisensory processes*. Cambridge, Mass: The MIT Press.
- Stein, B. E., Huneycutt, W. S., & Meredith, M. A. (1988). Neurons and behavior: the same rules of multisensory integration apply. *Brain Research*, 448(2), 355–358. [https://doi.org/10.1016/0006-8993\(88\)91276-0](https://doi.org/10.1016/0006-8993(88)91276-0)
- Stein, B. E., & Wallace, M. T. (1996). Comparisons of cross-modality integration in midbrain and cortex. *Progress in Brain Research*, 112, 289–299. [https://doi.org/10.1016/s0079-6123\(08\)63336-1](https://doi.org/10.1016/s0079-6123(08)63336-1)
- Sugita, Y., & Suzuki, Y. (2003). Implicit estimation of sound-arrival time. *Nature*, 421(6926), 911. <https://doi.org/10.1038/421911a>
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, Vol. 215, pp. 216–242. <https://doi.org/10.2307/25470707>
- Tuip, R. M., van der Ham, W., van Opstal, F., & Lorteije, J. A. M. (n.d.). *Synchrony is not required for audiovisual discrimination*.
- Van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex*, 17(4), 962–974. <https://doi.org/10.1093/cercor/bhl007>
- Vroomen, J., & De Gelder, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5), 1583–1590. <https://doi.org/10.1037/0096-1523.26.5.1583>
- Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage*, 31(3), 1247–1256. <https://doi.org/10.1016/j.neuroimage.2006.01.016>

Appendix A

Figures 11-18: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of all participants. Below are the logistic regression analyses of all 8 participants: the top left subfigures show the relative importance of auditory information during unisensory auditory trials; the bottom left subfigures show the relative importance of auditory information during multisensory audiovisual trials; the top right subfigures show the relative importance of visual information during unisensory visual trials and the bottom right subfigures show the relative importance of visual information during multisensory audiovisual trials. Solid blue lines show the average beta value of the information provided by the target stimulus for each 50 ms timepoint, with the shaded blue area indicating one standard error of the mean. Solid red lines show the average beta value of the information provided by the distractor stimulus for each 50 ms timepoint, with the shaded red area indicating one standard error of the mean. All timepoints that contribute significantly to a decision are marked with a black asterisk.

Participant 1

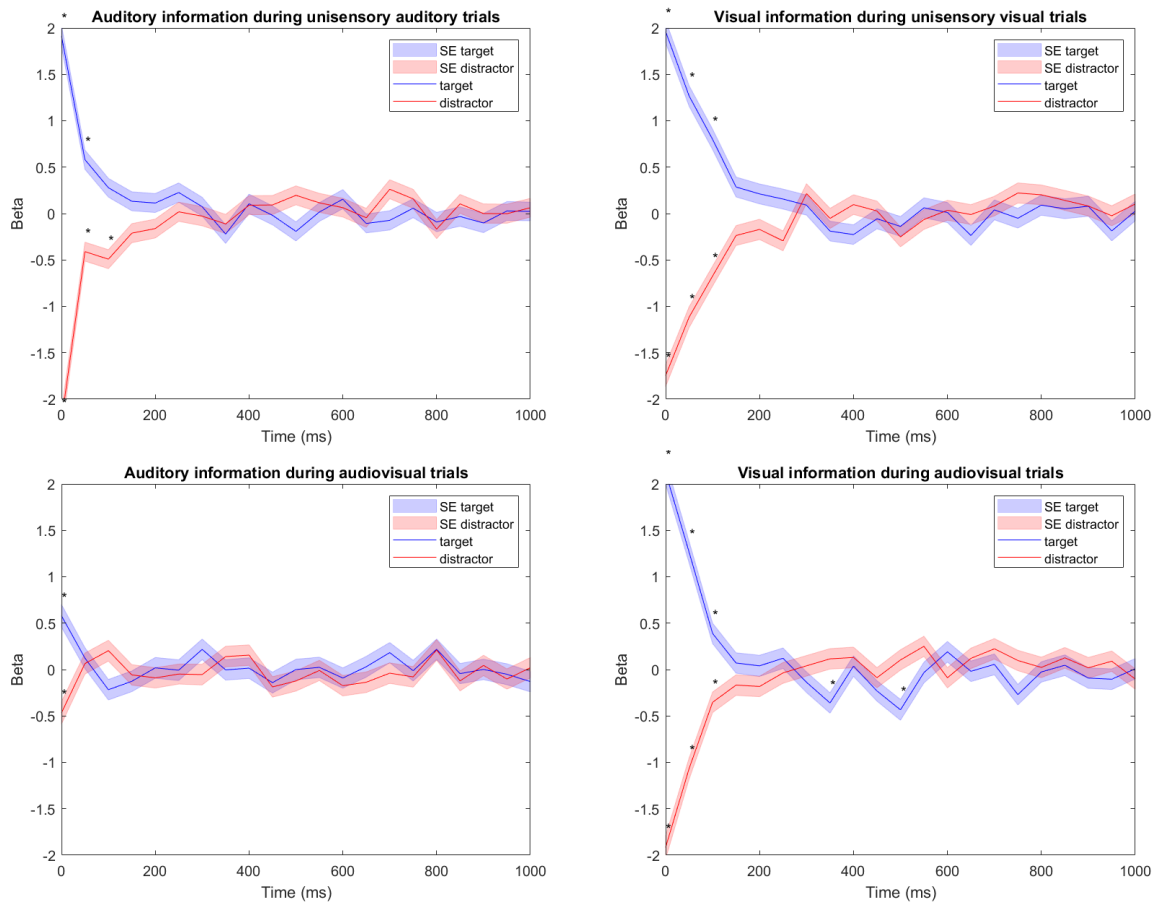


Figure 11: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 1. See above for figure caption.

Participant 2

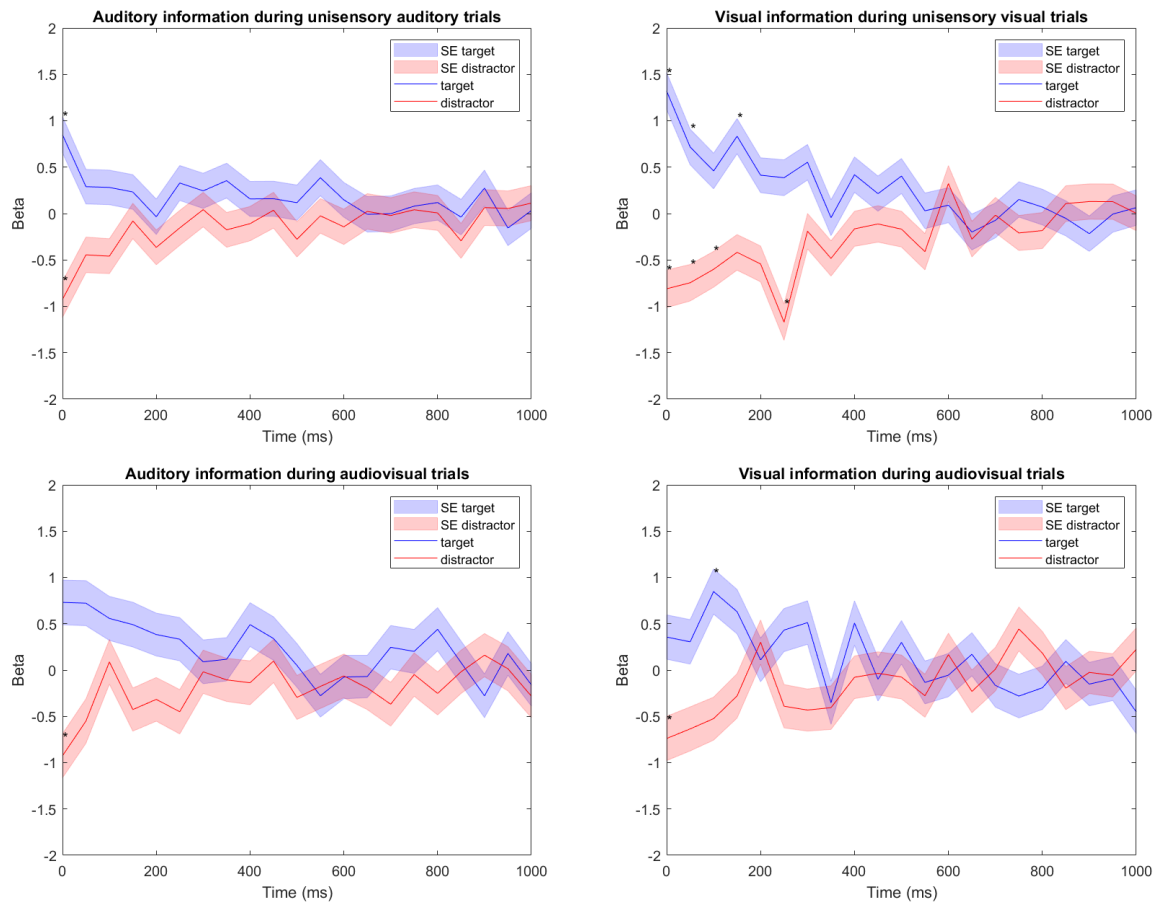
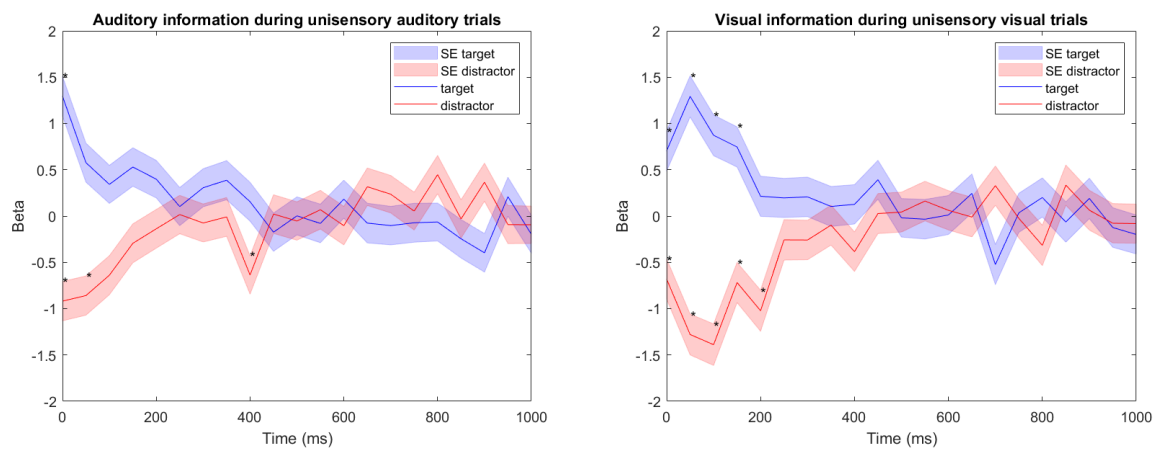


Figure 12: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 2. See above for figure caption.

Participant 3



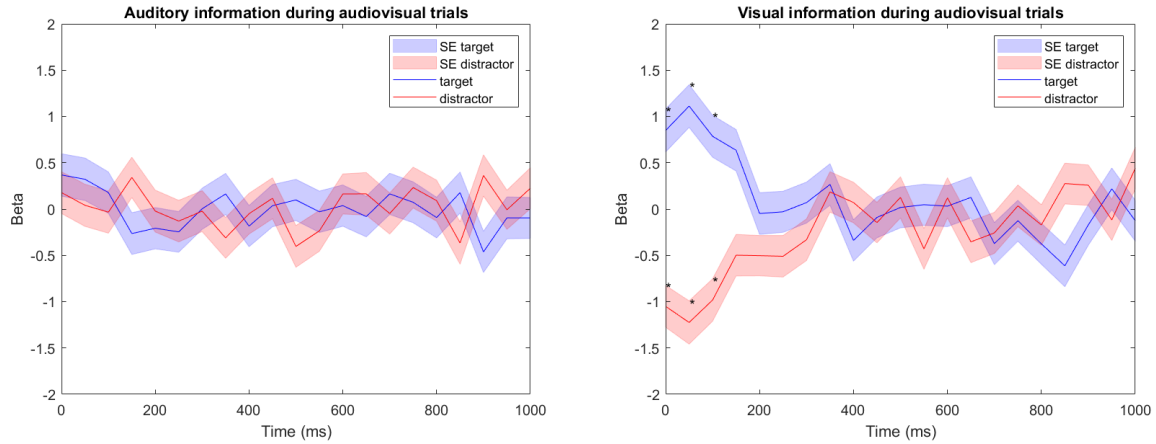


Figure 13: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 3. See above for figure caption.

Participant 4

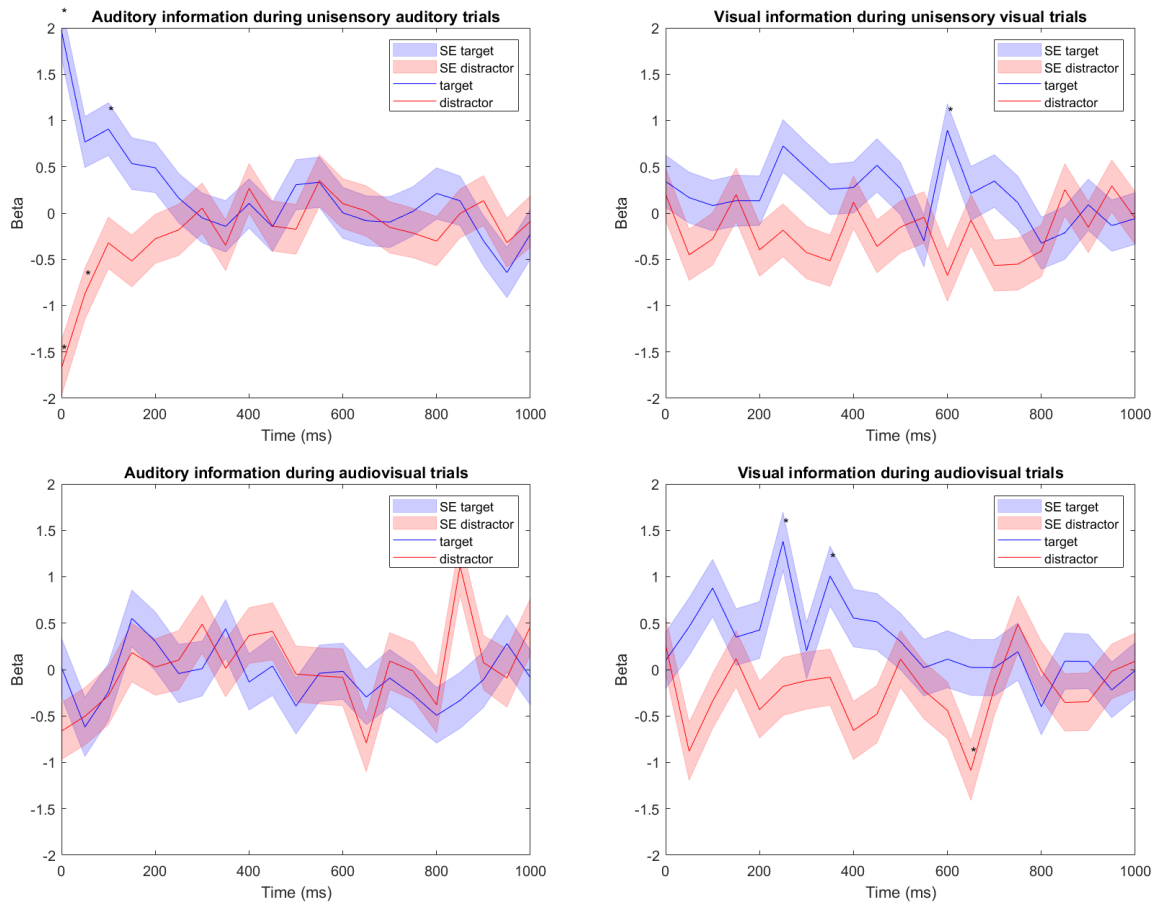


Figure 14: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 4. See above for figure caption.

Participant 5

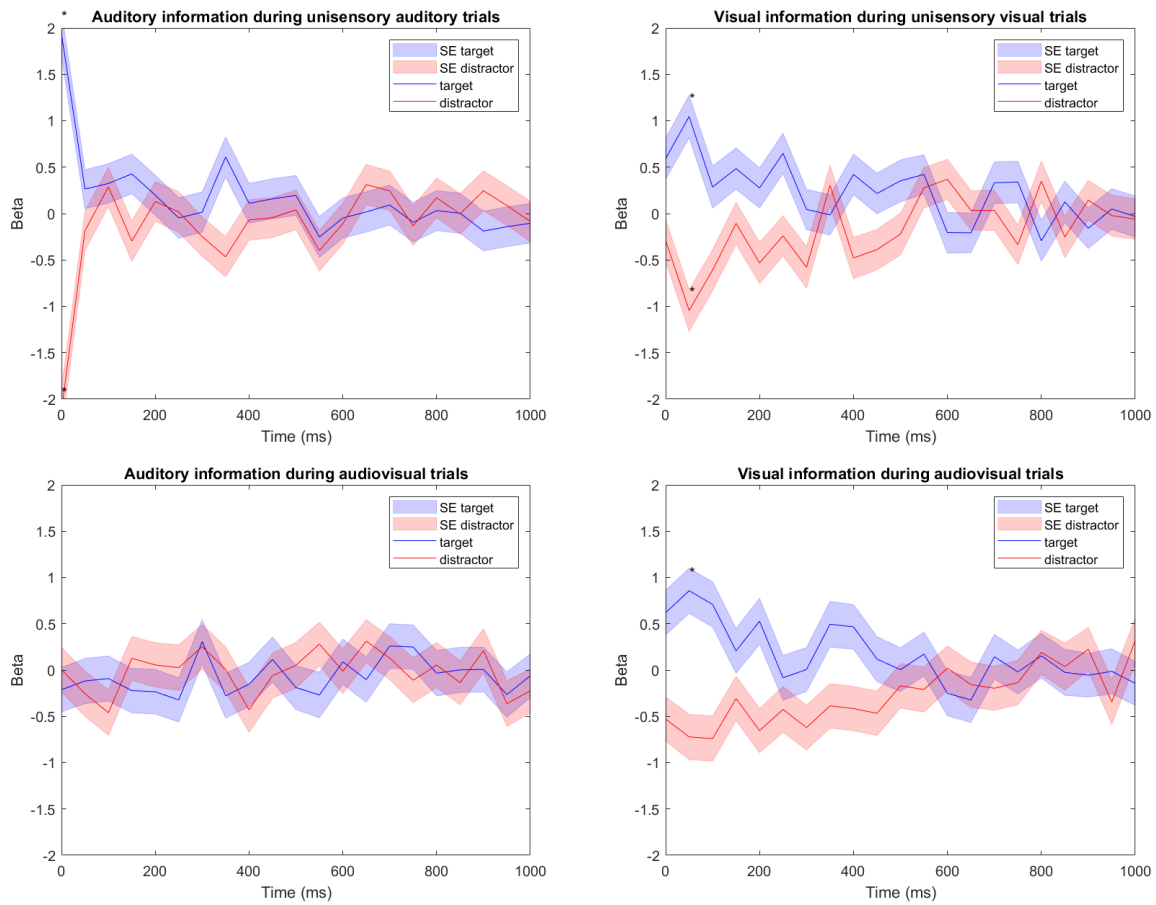
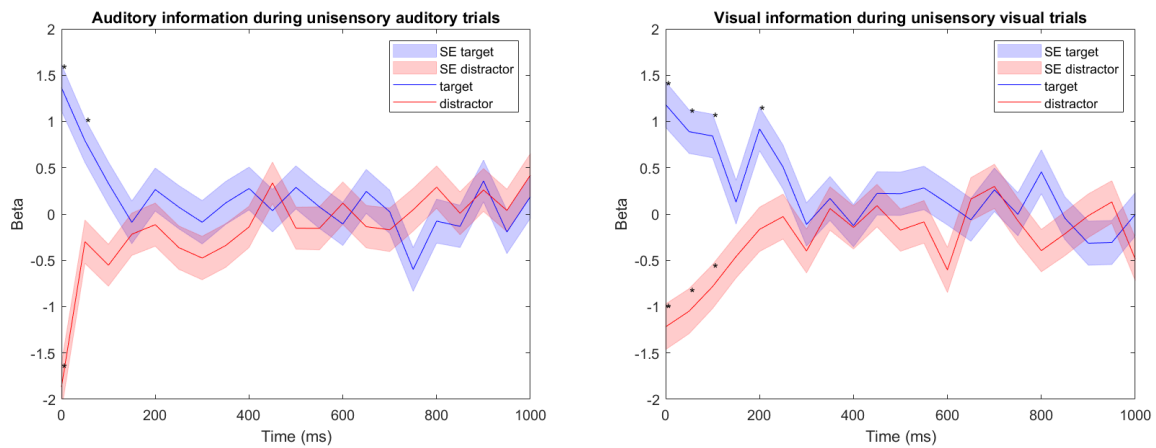


Figure 15: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 5. See above for figure caption.

Participant 6



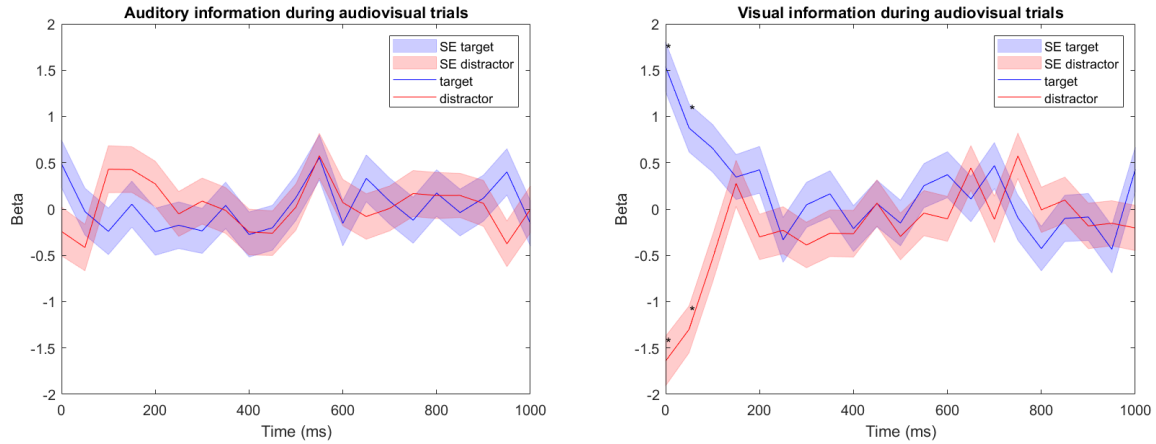


Figure 16: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 6. See above for figure caption.

Participant 7

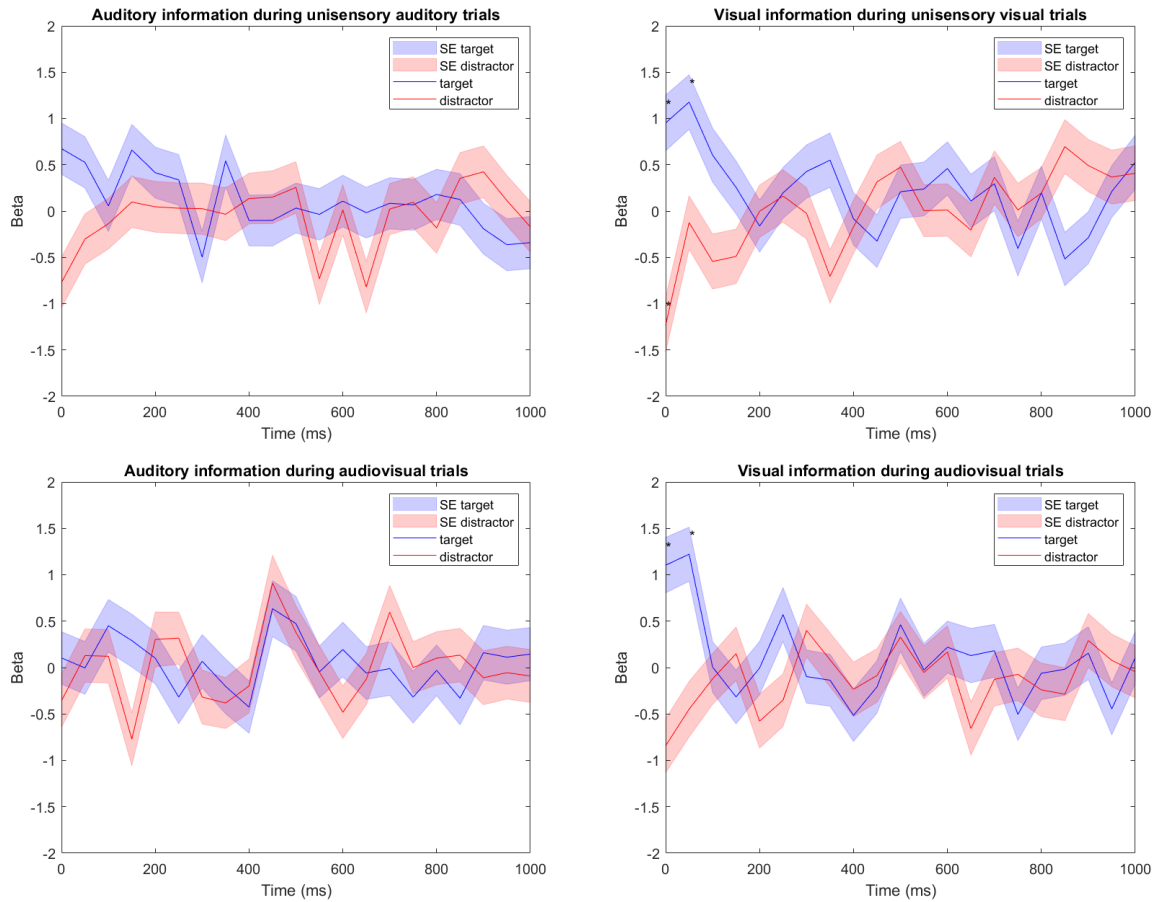


Figure 17: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 7. See above for figure caption.

Participant 8

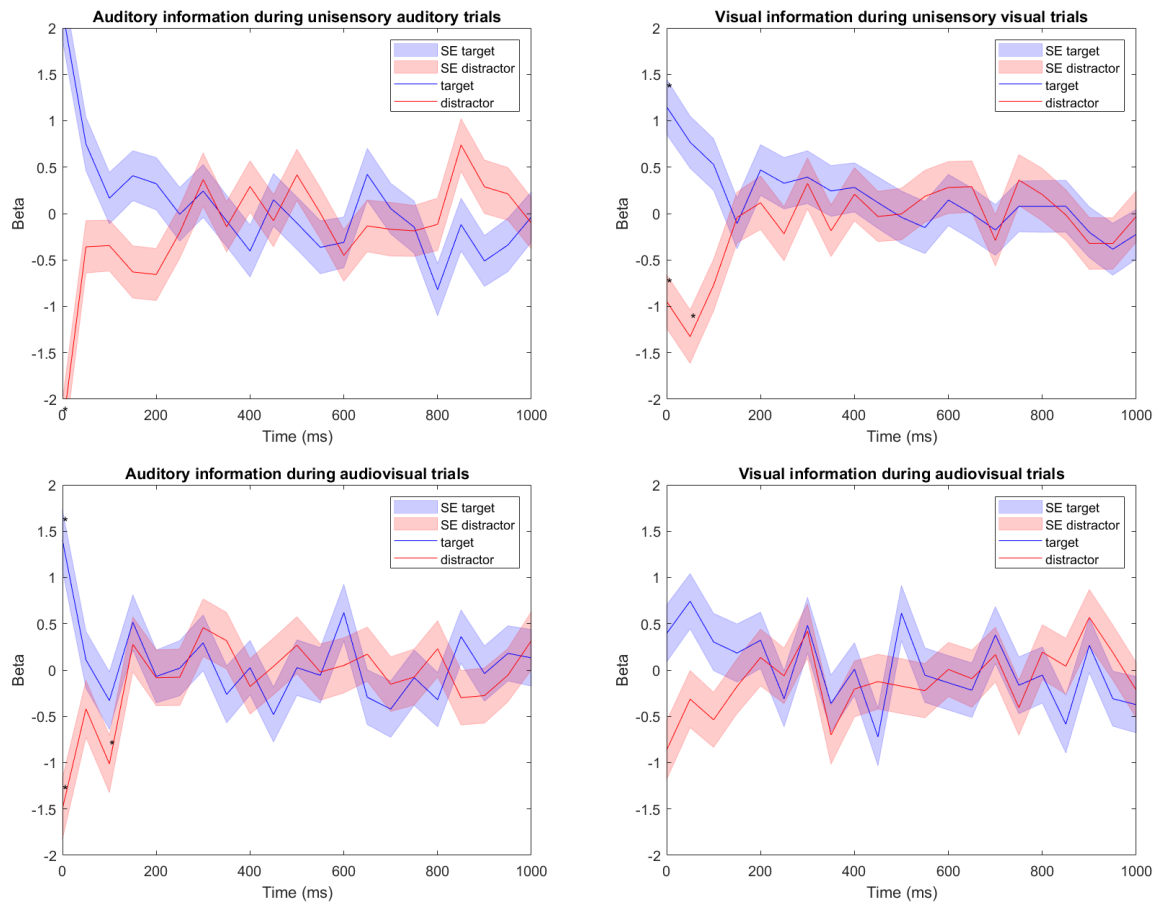


Figure 18: Logistic regression analyses of auditory and visual data in unisensory and multisensory trials of participant 8.
See above for figure caption.