SELECTIVE ATTENTION FOR AUDIOVISUAL INTEGRATION OF SPEECH

SYLVAIN J. BOUTROS
Bachelor of Science, University of Lethbridge, 2019

A thesis submitted
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

NEUROSCIENCE

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

© Sylvain J. Boutros, 2022

SELECTIVE ATTENTION FOR AUDIOVISUAL INTEGRATION OF SPEECH

SYLVAIN J. BOUTROS

Date of Defence: December 8, 2022

Dr. M. Tata Thesis Supervisor	Professor	Ph.D.
Dr. D. Euston Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. C. Gonzalez Thesis Examination Committee Member	Professor	Ph.D.
Dr. R. Gibb Thesis Examination Committee Chair	Professor	Ph.D.

Dedication

This paper is dedicated to my past self, under whose constant grit, I have completed this dissertation. To my beloved wife, whose support has helped me tremendously.

Imperare sibi maximum imperium est.

-Seneca

Abstract

The perceptual brain decomposes the audiovisual world into a set of audiovisual features such as color, shape, and pitch. An important question in perception science is whether selective attention is required to bind audiovisual features back into unified perceptual objects. In visual displays, targets defined as conjunctions of bound features typically cannot be searched for in parallel across a complex scene. Instead, attention must be scanned through a visual scene to find such targets. This means that conjunction of features requires selective attention. Only a few prior studies have investigated this process in the crossmodal audiovisual case. These prior studies left some ambiguity as to how temporal dynamics interact with selective attention, and none have used dynamic audiovisual speech as stimuli. In two experiments, this thesis explored whether the crossmodal conjunction of audio and visual speech features requires selective attention. In Chapter 2, we presented observers with displays of multiple visual faces and one audible voice. The task was to determine whether one of the faces was saying a sentence that matched the voice. In Chapter 3, we presented observers with displays of multiple audible voices and one visual face. Similarly, the task was to determine whether one of the voices was saying the sentence that matched the lip movements of the visual face. We compared response times to complete the task as the number of distracting items was increased. We found that it took longer to complete tasks with increasing number of distractors. This means that audiovisual speech targets cannot be registered in parallel across the scene and strongly suggests that selective attention is required to bind speech features across modalities.

Acknowledgments

I would like to express my deepest appreciation to my supervisor Dr. Matthew Tata for his help, guidance and extreme patience throughout my degree.

I am also thankful to my former & current committee members Dr. Chelsea Ekstrand, Dr. Claudia Gonzalez and Dr. David Euston with whom I had the chance to work with, get feedback and guidance during our meetings. I would also like to thank Dr. Robbin Gibb, chair of thesis examination, for accepting the role.

I'd like to acknowledge the members of our lab, Chris Perlette and Lukas Grasse for all their help throughout this journey.

Lastly, I'd like to mention my wife whose support, patience and love was needed throughout these two years.

Contents

List of Figures

Chapter 1

The Problem of Perceiving Features and Objects

1.1 Introduction

We easily perceive objects that differ from the background and from other objects in their environment, but this seemingly effortless process is supported by substantial computational mechanisms in the perceptual systems of the brain. A key step in this process is that the visual and auditory systems seem to decompose sensory input into *features*. Features are the fundamental characteristics of sensory objects. For example, in the visual domain, features include the colour, shape, brightness, etc. of an object. In the auditory domain, features include the pitch, timbre, loudness, chroma, etc. of a sound.

Two major physiological characteristics of the human visual and auditory systems, which helps perceptual system extract and bind features, are *retinotopy* and *tonotopy*, respectively. Retinotopy refers to the mapping of visual input onto the two-dimensional retina at the back of the eye. The afferent visual pathway, ascending through the Lateral Geniculate Nucleus to the primary visual cortex (V1) maintains the spatial configuration of retinotopy such that the cortex has a two-dimensional representation of the image projected onto the retina. Retinotopy exposes spatial relationships to the visual brain, which enables the visual system to organize and bind features using the dimension of space.

Our brains uses mainly binaural cues to locate sounds in complex scenes. Interaural time differences (ITDs) and interaural level differences (ILDs) are the two main types of cues in which our brains uses to help us find the azimuthal arrival angle of sound waves.

The delay for the propagation of the sound between the two ears results in ITD, while the difference between the sound levels due to attenuation by the head results in ILDs. By comparing these two cues the brain can isolate sounds by angle from the midline [3] .

Prior research has demonstrated a biological basis for feature extraction in the ascending sensory pathways. In the visual system the Lateral Geniculate Nucleus (LGN) has cells that are tuned to wavelength and cells that are tuned to the presence of edges, but not orientation. The primary visual cortex (V1) has cells that are tuned to the orientation of local line segments. Area V4 cells are tuned to different colors, and V5 cells are tuned to velocity of motion. In the Inferior Temporal Cortex (ITC) cells are tuned to complex shapes and identity. While in the auditory system the Superior Olivary Nucleus has cells that are tuned to frequency and location. In the primary auditory cortex (A1) cells are tuned to complex movements in the frequency domain (i.e., tone sweeps). In the Inferior Temporal Pole cells are tuned to the identity of sound objects, and in the Temporo-parietal Junction and Planum Temporale (posterior to A1) cells are tuned to spatial locations.

The auditory system creates its own form of mapping by decomposing the component frequencies of sounds along the frequency spectrum. Tonotopy refers to this mapping of auditory input onto the basilar membrane. Different locations along the basilar membrane are sensitive to different frequencies due to its mechanical properties. This tonotopic representation is maintained by the afferent auditory pathway, such as in the Medial Geniculate Nucleus (MGN) and onto the primary auditory cortex (A1). Like retinotopy in the visual system, tonotopy constitutes the primary dimension (frequency) that is used by the auditory system to organize and bind features. However, the notion of what constitutes an auditory object and how its features are bound is less intuitive than in vision, and remains relatively uninvestigated [10, 17].

An important and often overlooked dimension over which both visual and auditory features can be organized is that of *time*. In the visual world, features such as colour or shape change over space and time, often with some regular periodicity. Likewise, auditory objects

comprised of features such as frequency or intonation, change over time. In this thesis we look at the binding of crossmodal features which are temporally dynamic and coherent in time, in other words, the features from the visual and auditory world are joined together in a manner that are changing together in time.

Despite having low-level mechanisms dedicated to feature extraction and registration, we are consciously aware of whole objects, not their individual unbound component features. Thus, the perceptual system must have an additional step for the binding of features into perceptually whole objects. The computational mechanisms by which the brain accomplishes this step are not well understood and are collectively referred to as *The Binding Problem*. Thus, the brain binds features together so that objects are *conjunctions* of their features. Interestingly, we are able not only to bind features within modalities, but also across modalities. For example, we might perceive a red firetruck with a siren as a unified object; moreover, we might perceive a pleasant smell associated with a specific flower. The objective of this thesis was to further explore this phenomenon of *crossmodal* feature conjunction.

Much of what is known about how the *visual* system decomposes and re-binds features has been investigated using the paradigm of *visual search* [2, 26, 27, 32, 33]. We know that both the visual system and auditory system are tuned to primitive features such as orientation in the primary visual cortex and tone sweeps in the primary auditory cortex that are aligned with Treisman's description of features. Visual search is an experimental approach in which an observer tries to find a target among distractors. The target can be defined and discriminated from distractors on the basis of one or more visual features (see Figure 1.1). For example, in Figure 1.1A, the target is a red circle and the distractors are green circles. The observer's task is to indicate whether the target is present (as shown in 1.1A) or absent (not shown). Treisman and others found that when the target is distinct from the distractors by a single feature (i.e., it is a singleton), then the time it takes to perform this task remains relatively constant even as the number of distractors is increased,

as depicted in Figure 1.1B. By contrast, when the target of a visual search could only be identified by the binding of two or more features into a conjunction, it was found that the response time varies linearly with the number of distractors. For example, in Figure 1.1C the target is the only object that is the conjunction of green and square, and the slope of the response times in Figure 1.1D are positive. An important observation was that, for conjunction searches, the slope of the line relating response time to distractor set size tends to be about twice as steep on target-absent trials (i.e., the observer should respond "no") relative to target present trials. The putative explanation being that, when a target is absent, an observer is required to search the entire visual display, item by item, in order to come to a self-terminating searching process. By contrast, when the target is present, the observer needs to search, on average, only half of the items in the display.

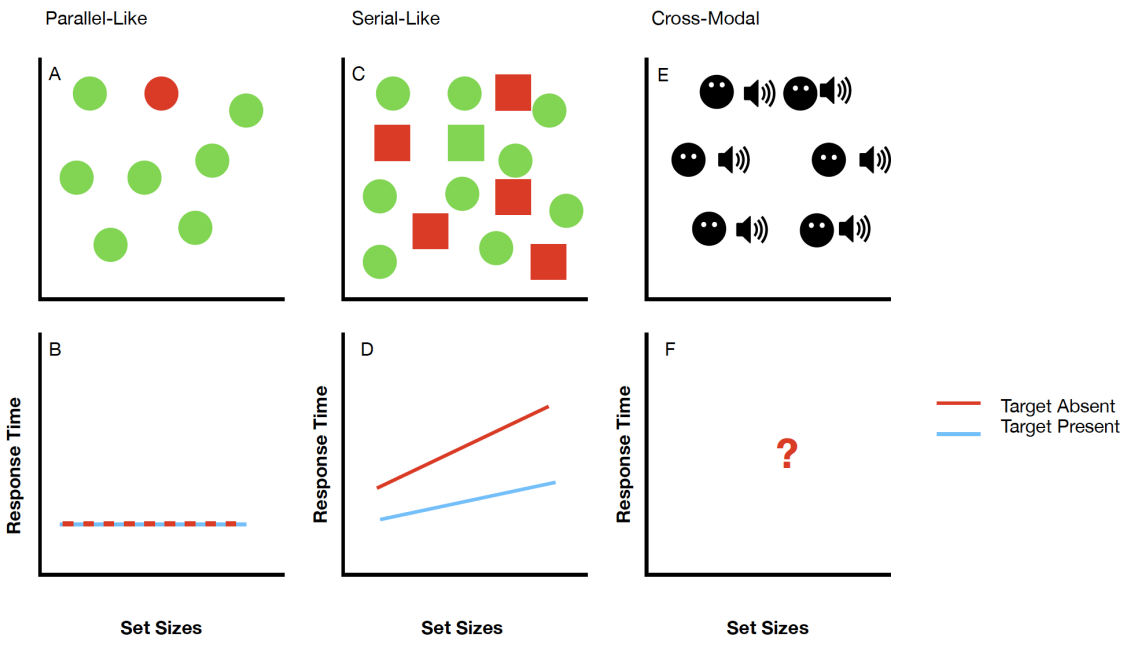

Figure 1.1: **Search Paradigm and Respective Slopes.** Figure 1A and C are examples based on different visual search paradigm and their respective search slopes, below them. Figure 1E and F represent the crossmodal paradigm.

The subjective experience of the observer performing these two kinds of searches is quite different. In the case of search for a singleton, the target seems to "pop-out" of the display, whereas in conjunction searches the observer has to deliberately scan through the

display, focusing attention on each object in a serial fashion. A well-known example of this second serial attentive search process is the popular "Where's Waldo" pictures in which observers search for the character Waldo, who is defined by the conjunction of his blue pants, striped shirt, etc. Treisman and others [2, 26, 27, 32, 33] interpreted the search slope results and the phenomenological impression of pop-out to imply that an "efficient" search mechanism registers features across the retinotopy of the display, and that this feature registration happens in parallel across locations and without the need for focused attention. Indeed, the registration of a singleton feature precedes and guides spatial attention in "pop-out" displays [32, 34]. By contrast, it was proposed that a further binding step is required to solve conjunction searches, which requires the serial allocation of attention to each spatial location in order to bind features into objects. A substantial subsequent body of work has elaborated on the initial ideas of Treisman and Wolfe (for example, [19, 23, 24, 25] (see [15] for a review)). Regardless of the theoretical details, most models invoke a parallel-like preattentive stage and a serial-like selective attention stage. The goal of this thesis was not to differentiate between visual search models, but rather consider where audiovisual conjunctions might fit into the parallel-like or serial-like dichotomy.

Almost as a rule, search for *visual* targets defined by *uni-modal* conjunctions of features (e.g., colour and shape) requires serial attentive search (i.e., they do not "pop-out"). Our aim in this thesis was to consider whether conjunctions of features across modalities follow this same general rule. The most parsimonious hypothesis would suggest that targets defined as *multisensory* conjunctions of features should, as in the uni-modal visual search case, require serial attentive search. There are, however, some observations that suggest there might be an exception in the case of coherently dynamic audiovisual stimuli, particularly audiovisual speech. First, tight temporal binding of audiovisual input in the case of speech perception has been shown in the well-known McGurk Illusion [16]. Second, the perceived spatial location of dynamic auditory events can be captured to the location of temporally coherent visual events – the well-known Ventriloquist Illusion [22]. Third,

providing *video* of talkers helps listeners solve the speech-in-noise problem, as long as the audio and video are closely synchronized in time [21]. These examples of tight, putatively low-level temporal integration across vision and audition suggest that a preattentive cross-modal feature binding mechanism might exist, at least for dynamically coherent stimuli such as audiovisual speech.

More recently, Fujisaki and colleagues (2005) [31] tested whether flashing of rotating visual targets, which were synchronized with either amplitude-modulated tone pips or frequency-modulated sweeps, could be detected efficiently amongst flashing or rotating distractors. The authors did not find any evidence that these simple auditory and visual events integrate in a preattentive step, and thus concluded that the conjunction of audiovisual events requires a serial-like process. Although the results from Fujisaki et al. failed to find any "efficient" searching mechanism for audiovisual search, a pair of studies by Van der Burg [29, 30] claimed the opposite. Van der Burg and colleagues (2008) showed that flickering visual targets, which would normally require attentive search to locate, were made to "pop-out" when coupled with temporally coincident auditory tones. Van der Burg (2010) further showed that this efficient "pop-out" mode of attention orienting only happened under specific conditions of temporal dynamics. That is, the optimal pop-out condition was one in which both visual and auditory events had very sharp temporal transients - their amplitude envelopes were described by square waves. By contrast, when auditory and visual events exhibited smooth (i.e. sinusoidal) amplitude modulation, there was a distinctly serial-like nature to the search task, possibly explaining why Fujisaki et al. failed to find crossmodal "pop-out". Importantly, naturalistic speech has both sharp transient events (i.e., the consonant K and T) and smooth amplitude envelope modulations (i.e., co-articulated vowels), leaving an unanswered question as to whether audiovisual speech might pop out of a cluttered scene. Thus, this thesis explored audiovisual sensory binding of the temporal features of audiovisual speech with respect to preattentive vs. attentive mechanisms. Considering these crossmodal phenomenon, one might expect multisensory search to be

possible using a parallel-like search mechanism.

1.2 Summary

To conclude this introduction, this thesis has the following structure: Chapter one (this chapter) has described the concepts of retinotopy and tonotopy, and the importance of feature decomposition and binding. It then introduced the paradigm of visual search and mental mechanisms that are engaged during serial-like and parallel-like searches. Chapter two and three describe experiments that we used to test the hypothesis that crossmodal conjunction search would behave in similar manner to uni-modal conjunction search (i.e., require serial attentive search). In all experiments, we borrowed the search paradigms used for visual search and created crossmodal versions by adding coherent dynamic speech to investigate and interpret the search slopes and whether we would find any "efficient" search mechanisms registering features across both retinotopy of the display and tonotopy of the sound (see Figure 1.1E and F). Finally, chapter four covers a discussion of our findings and limitations of this work.

Chapter 2

Finding a Face/Voice Match in a Visual Display Requires a Serial-Like Attentive Searching Process

2.1 Introduction

Even if we easily perceive objects that differ from the background, and from other objects in their environment, it still requires a substantial computational mechanism in the perceptual systems of the brain to decompose and re-bind features into perceptual objects. This chapter will explore the cross-modality aspect of this phenomenon by bridging two paradigms; First, we incorporated sound in the search paradigm used successfully in visual search studies [2, 26, 27]. Second, we modified the search paradigm to investigate cross-modal conjunction of audiovisual speech targets. We did this by adding coherent dynamic speech to a visual display of multiple talkers (see Figure 2.1 for the visual layout). Treisman and others contributed substantial prior evidence supporting an ”efficient” searching mechanism that registers features in parallel across the display. This mechanism allows for parallel ”pop-out” searches for certain visual target/distractor relationships. Do we have similarly ”efficient” searching mechanisms that can decompose and then register temporally coherent auditory and visual features across a multisensory display? This chapter describes the experimental paradigms used to investigate whether there is evidence of ”efficient” low-level searching mechanisms in conjoining speech features across auditory and visual modalities.

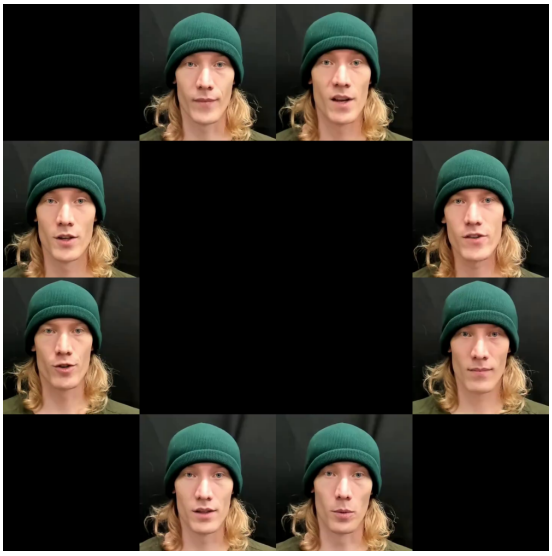

Figure 2.1: **Visual Display.** An example of the visual layout for a set size of 8, whether target was present or absent. If the target was present, the trial had the visual stimulus' lips movements matching one of the auditory stimulus and seven distractors which were saying (unheard) different sentences than the auditory stimulus. In contrast, if the target was absent, there was no match between what was heard and what was seen, and all faces were considered distractors. The visual layout was arranged in a notional circle around a central fixation point which was not visible in the visual display.

Our approach was to present audiovisual displays of talkers, and require observers to indicate whether a face/voice stimulus matched (i.e., target present), despite the presence of distractor faces (see Figure 2.1). In other words, to successfully complete the task, participants needed to search for a face saying the audio that they were hearing and determine if it was present or not. Two contrasting outcomes were envisioned: if an efficient preattentive mechanism for binding auditory and visual speech exists, then search for a single coherent face/voice combination should be independent of the number of talkers; by contrast, if conjoining a face with its matching voice requires attentional selection, then search for a coherent talker/voice combination should take longer with increasing distractor set size. Furthermore, as depicted in Figure 1.1, if an efficient search for coherent audio/visual speech exists, then target-present and target-absent search slopes should be identical and nearly flat, whereas if attentional selection is required, the target-absent slope should be

significantly steeper than target-present slope.

2.2 Methodology

2.2.1 Ethics Statement

Procedures were in accordance with the Declaration of Helsinki and approved by the University of Lethbridge Human Subject Review Board.

2.2.2 Stimuli and Task

Stimulus Recordings

Audiovisual speech stimuli to be used in the experiment were created by recording talkers with a Logitech webcam (Logitech Brio Ultra HD 4K, 60 frames per seconds (FPS), 44100 Hz audio sampling rate (SR)) using Logitech's proprietary software (Logitech Capture). For consistency with other studies, the sentences chosen to record were those of the TIMIT corpus. The TIMIT corpus is a dataset containing various phonetically rich sentences [9]. In total, there was 60 videos recorded and each video segment recorded was six sentences long, varying between 13 and 26 seconds in duration. A custom Python script presented the paragraphs to be read from the TIMIT Corpus by the talkers at the top of the screen and directly underneath the camera. The talkers were instructed to perform as naturally as possible by looking directly into the camera. In exchange of their time, talkers were offered a $30 gift card. In total, we recorded four talkers (two males/females).

Stimulus displays for the subsequent experiment were then constructed using a custom Python script and the MoviePy library [18]. The display for each trial was composed by extracting ten-second segments from the raw recordings mentioned above and arranging those video clips in a ring around a central visual fixation point to generate the visual presentation stimuli for the various set size conditions of 1, 2, 4, 5, 7, and 8. The reasons behind the ring set up was simply to have each of the video clips distance at the same locations from one another, so that each face retained the same retinal eccentricity. Since we had predicted,

based on most prior research, a linear relationship between set size and response time, we decided to keep the set size conditions to six to reduce the overall experimental time. An example of the visual display can be seen in Figure 2.1 depicting a set size of 8 - one target and seven distractors (i.e., on target-present trials) or eight distractors (i.e., on target-absent trials). Each individual ten-second segment was included in a target-present and a target-absent condition. No display contained two copies of the same segment; however, the display always contained the same talker (i.e., the same face saying different sentences).

2.2.3 Participants

Nineteen undergraduate students were recruited from University of Lethbridge courses and gave informed consent before participating. All reported normal hearing, as well as corrected-to-normal or normal vision. Participants received one course credit for one hour of their participation and were informed, both verbally and in writing, of their options to withdraw at any time during the experiment. Of these, the data from five participants were excluded due to failure to complete the experiment. Hence, the data from 14 participants (12 females; all right-handed; $\mu_{age} = 22.57$ years; $\sigma_{age} = 5.8007$; 13 native English speakers) were analyzed.

Stimulus Presentations

For the search experiments, the stimuli were presented on an iMac (mid 2014, macOS Mojave, using the native display resolution of 1920x1080 and an intel HD Graphics 5000 1536 MB card). The refresh rate of the iMac display was 60 Hz. All audiovisual speech stimuli were recorded as described above. Stimulus presentation was controlled by customized MATLAB code using the Psychophysics Toolbox extension [4, 14, 20]. The auditory speech was presented using the native iMac stereo speakers.

Task Requirements

On each trial, participants heard one voice and saw a varying number of faces arranged around a central fixation point (as can be observed in Figure 2.1). On half of the trials, the audio of the voice matched the lip movements of one of the faces (target present). On the other half of trials, the audio of the voice did not match the lip movements of any of the faces (target absent). On each trial, participants were asked to respond by pressing on the ”p” or ”a” key, for target-present and target-absent respectively. Emphasis on both accuracy and speed was made both verbally and in the written instructions.

2.2.4 Experimental Setup

Participants sat approximately 30 cm from the display monitor in an acoustically treated room (RT 60 reverberation time approximately 0.33 ms at 1 kHz; noise-floor approximately 41.6 dBA) under LED lighting. Participants responded via the iMac's keyboard. The experiment started with both verbal and written instructions, followed by four practice blocks consisting each of 12 trials for a total of 48 practice trials. The practice trials were pseudo-randomized such that each set size X target conditions were divided equally between all four talker stimuli. These practice stimuli were then excluded from the subsequent experimental trials, such that the sentences presented in the experimental trials, on either target conditions, were not presented in the practice trials. Experiment trials likewise were counterbalanced so that individual talkers were pseudo-randomized across conditions and the sequence was shuffled so that the pseudo-random order of presentation was not repeated for participants. Each block of the experimental trials consisted of six trials in both target conditions pseudo-randomized such that each set size X target condition was presented once per block. The experiment was divided into three sections, first participants were verbally and visually instructed of the task they needed to accomplish; second the practice blocks; third experimental blocks. Responses were not collected during the practice trials. Feedback (correct/incorrect) was given after each trial, in both the practice and experimen-

tal blocks. Breaks were allocated after each experiment block of 12 trials; participants were instructed to resume whenever they wanted after a break and were encouraged to take longer breaks if needed.

2.2.5 Analyses

To analyze the data, we compiled an average of reaction times (RTs) across all correct-response trials per participants as a function of set size and target presence. To consider whether participants took longer with increasing set sizes, and whether there was an effect of target presence (both classical hallmarks of serial-like search processes) we ran a two-way repeated-measure Analysis of Variance (RM-ANOVA) on the RTs with six levels of the factor Set Size (1, 2, 4, 5, 7, and 8) and two levels of the factor Target Presence (present vs. absent). To further consider whether the slopes of the target-absent trials were steeper than those of target-present trials (a hallmark of serial-like search), we fitted a linear regression to each participant's RTs independently for target-present and target-absent trials. We then compared the slopes of these regression lines with a paired one-tailed t-test, testing the hypothesis that the target-absent slopes were steeper than the target-present slopes against the null hypothesis that the slopes were equal. Additionally, we considered whether set sizes or target presence affected accuracy by computing the proportion of correct responses for each level of each factor. These accuracy data were also compared through a RM-ANOVA with set size and target presence as factors.

2.3 Results

Figure 2.2 shows the average response-time performance for each participant on correct trials as a function of both set size and target presence. Response times generally increased in a linear way as we added distractors to the visual display, and this increase varied depending on target presence. A two-way repeated-measure Analysis of Variance performed on the data supports this observation. There was a significant main effect of target pres-

ence [F(1, 156) = 93.9877, $p < .001$] and a significant main effect of set sizes [F(5, 156) = 86.1189, $p < .001$]. Furthermore, the effect of set size was different (steeper) for target absent relative to target present trials indicated by a significant interaction between the two factors [F(5, 156) = 5.4717, $p < .001$].

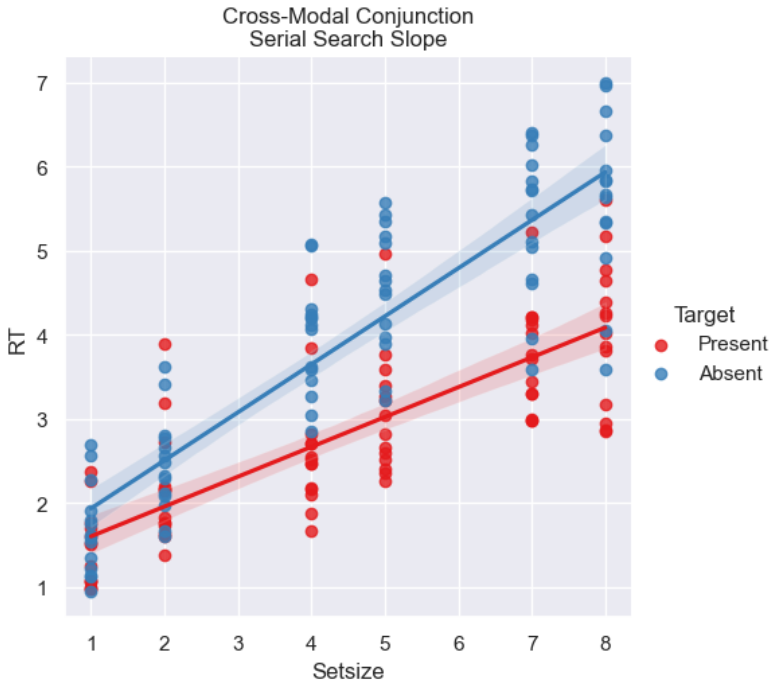

Figure 2.2: **Linear Regression Model.** The average of each participants' response times on each set size for correct trials were used to build a linear regression model with a confidence interval of 95%. The red slope (lower) represents the target-present data, while the blue slope (upper) represents the target-absent data. A two-way repeated-measure ANOVA showed a significant main effect of set size F = **86.1189** and a significant main effect of target presence F = **93.9877**, as well as an interaction between the two factors F = **5.4717**.

Comparing the regression line slopes for each subject, we calculated the target absent and target present slopes across participants (μ target absent slope = 0.5742, σ target absent slopes = 0.1337; μ target present slope = 0.3583, σ target present slopes = 0.0546). We saw a significant difference of the slopes as a function of target presence supported by a significant paired one-tailed t-test [t(13) = 7.5309, $p < .001$].

Figure 2.3 shows the accuracy (correct responses) both as a function of target presence and across all the set sizes. There was a significant main effect of target presence [F(1, 156) = 11.7003, p $<$ .001] and a significant main effect of set size [F(5, 156) = 11.3543, p $<$ .001]. Furthermore, the effect of set size was different (more accurate) for target-absent relative to target-present trials indicated by a significant interaction between the two factors [F(5, 156) = 2.4964, p = .03].

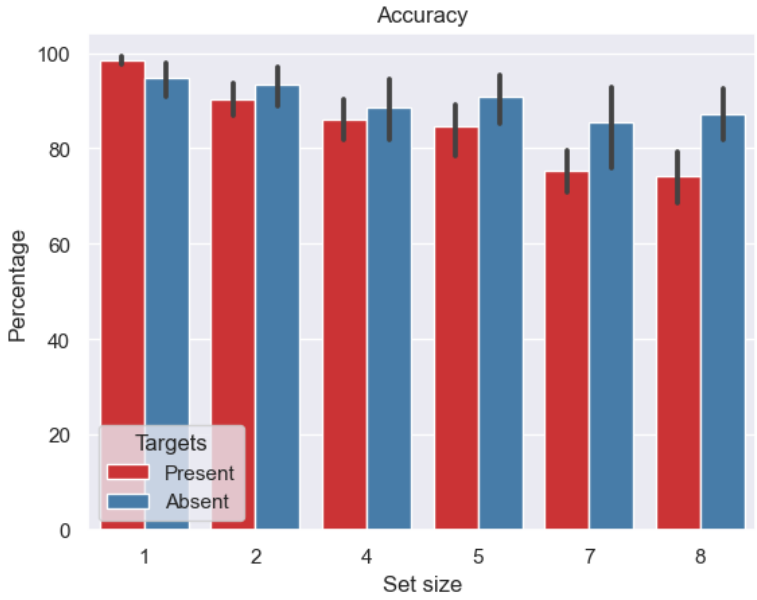

Figure 2.3: **Accuracy.** The average of all the participants' correct responses classified by set sizes and by target presence. The red strips (left) represents the average accuracy for the target-present trials, while the blue strips (right) represent the average accuracy for the target-absent trials with error bars representing a confidence interval of 95%. A two-way repeated-measure ANOVA showed a significant main effect of set size F = **11.3543** and a significant main effect of target presence F = **11.7003**, as well as an interaction between the two factors F = **2.4964**. Participants had an average **87.45%** success rate at discriminating whether a target was present or absent.

2.4 Discussion

In this study, our objective was to determine whether coherent audiovisual speech could be found amongst distraction through an "efficient" low-level mechanism (i.e., "pop-out"),

or instead requires a serial-like searching process. To do so, we modified a proven visual-search paradigm by presenting audiovisual displays of talkers, and requiring participants to indicate whether a face/voice stimulus matched (i.e., target-present) or there was no face/voice stimulus match (i.e., target-absent). Furthermore, we predicted that, if an efficient preattentive mechanism for binding auditory and visual speech exists, then search for a single coherent face/voice combination should be independent of the number of talkers presented in the visual display. Otherwise, if the conjunction of face and voice requires attentional selection, search for a coherent face/voice match should take longer and increase as distractors are added to the visual display. Additionally, an efficient mechanism would yield flat RT slopes similar to Figure 1.1B; otherwise, similar to Figure 1.1D with slopes being steeper for target-absent compared to target-present trials (approximately in a two-to-one ratio).

Analyses of the response times across participants as a function of set sizes showed that an increasing number of distractors resulted in longer time to complete the tasks; in other words, on average it took longer to complete trials as the number of distractors increased. The results of the analysis suggest that auditory and visual feature conjunction search, in this crossmodal paradigm, shares similarities with typical unimodal visual feature conjunction searches. Furthermore, analysis of the response times across participants as a function of target presence showed that there was a difference in response times when they were grouped by target presence. It took longer to complete the tasks when trials were of type ”target-absent”. Additionally, there was an interaction between set-size and target presence such that the increase in RT with set size was greater for target absent than target present trials. This phenomenon is characteristic of unimodal visual feature conjunction searches. The putative explanation for the steeper increase in average searching times on target-absent trials can be explained by having to search the entirety of the visual scene on every trial, item-by-item, in order to come to a self-terminating searching process to complete the trials.

Analyses of the linear regression slopes (see Figure 2.2) as a function of target presence showed that there was a difference between target-absent and target-present trials. The results suggest that response times on target-absent trials increased faster with increasing set size than target present trials. This is also characteristic of the unimodal visual search case when serial visual search is required. Thus, adding distractors into the visual presentation on target-absent trials requires participant to scan through more items on average.

Analyses of the accuracy as a function of set size and target presence (see Figure 2.3) helped in solidifying the results obtained from the slope analyses. For both conditions, an increasing number of distractors caused a lower accuracy; in other words, more mistakes were made when there were more distractors. Furthermore, the significant main effect of target presence on accuracy suggests that accuracy was better on target-absent trials than on target-present trials. We consider each type of error individually as follows: On target-absent trials an incorrect response (i.e., a false alarm) meant that a face/voice pair was matched when none of the faces were matching the voice. On target-present trials an incorrect response (i.e., a miss) meant that no face/voice match was found when one of the faces was in fact matching the voice. In the case of the former (false alarms), it is possible that a crossmodal illusion takes place, making it seem like one of the faces was a match to the voice heard. Such illusory conjunctions have been described in the unimodal visual search case [28]. Indeed, the McGurk Effect could be considered a type of audiovisual illusory conjunction, although in that case the perceived phoneme matches neither the audio nor the visual input. The case of errors on target-present trials (i.e. a miss) is perhaps more interesting: Because of the serial-like positive search slopes we find in this experiment, we consider that selective attention is indeed being allocated sequentially to individual faces. However, we speculate that increased visual clutter in higher set sizes causes a ”competing environment” for adjacent stimuli. In this case the attention system is perhaps not able to exclude the temporal dynamics of nearby visual distractors causing the true dynamics of the visual target to be lost. Another possibility is that participants need to selectively attend to a

face and dwell on it (for an unknown but sufficient time) to accumulate sufficient evidence for matching dynamics; since the slopes are less than 1.0 we know that participants are allocating less dwell time for each item at higher set sizes, perhaps thus failing to achieve enough evidence to score a hit (i.e., find face/voice match).

Many decades of research have supported the notion that there is tight temporal integration between auditory and visual stimuli. For example, Thomas [22] showed that visual stimuli influenced the localization of auditory stimuli. By placing two lights in different spatial locations and sounding a buzzer, the author observed that participants tended to misjudge and report the sound to be coming from nearer the light that was flickering in-rhythm with the buzzer. Later, Sumby and Pollack [21] showed that intelligibility of speech mixed with noise was improved when observers where provided with the face and lip movements of the speakers. Additionally, the now-classic paper by McGurk and MacDonald [16] provided evidence on audiovisual illusions caused when a pair of syllables were presented to the two different sensory systems, seeing the lip movements for ba-ba and hearing the sounds ga-ga which when fused resulted into hearing da-da. All of these examples suggest a tight co-registration of dynamic auditory and visual events in the service of cross-modal perception.

Our results are consistent with the hallmarks of serial search processes. In other words, the conjunction of audiovisual temporally dynamic speech features required a serial attentive mechanism rather than an efficient preattentive mechanism. This is consistent with results described by Fujisaki [31], who also described a serial-like audiovisual search task, albeit with simple non-speech stimuli. Our results are inconsistent, however, with a study conducted by Van der Burg and colleagues (2008) [30], which showed a situation in which auditory tone onsets that were synchronous in time with the onsets of visual search targets led to parallel-like search for those targets. That is, synchronous audio helped visual search. Subsequently, [29] showed that to achieve this efficient pop-out search in such tasks, it was necessary to have auditory and visual stimuli with sharp onset and offset transients. Our

study filled a gap in knowledge because it remained unclear whether natural speech, which includes both smooth and abrupt transients, is a sufficient stimulus to enable pop-out. We speculate that, in fact, the amplitude envelope of speech is not sufficiently abrupt to mediate pop-out search, at least in the situation of a single speech stream and multiple visual distractors. The complimentary situation of a single visual target and multiple auditory distractors is considered in the following chapter.

Chapter 3

Finding a Face/Voice Match in an Auditory Display Requires a Serial-Like Attentive Searching Process

3.1 Introduction

In Chapter 1, we noted that visual objects could be quickly and easily found when they sufficiently differ from their backgrounds and from other objects around them (i.e. target pop-out). However, in some circumstances a serial attentive searching process is required to find target objects amongst distractors. This is particularly common when the target is defined as the conjunction of features (for example, when the target is defined as the only red square). In Chapter 2, we showed that a serial attention searching process is required to find the conjunction of audiovisual speech. In that experiment, targets were defined as a unique conjunction of a single stream of auditory speech and a single visual target out of several visual distractors. In other words, the distractors were visual objects, and those distractors were at unique spatial locations around the visual display. The search task itself was to scan through visual (i.e., spatiotopic or retinotopic) space. However, much of auditory perception involves scanning auditory selective attention through a complex *auditory* scene. As in vision, auditory objects often occupy different spatial locations in natural scenes. Furthermore, it is well-known that separating auditory streams in space makes it much easier to perceptually unmix streams from one another. This phenomenon has been called spatial release from masking [12, 13].

Much of what is known of how we selectively attend to auditory objects has been re-

searched in the context of the *Cocktail Party* problem. The "Cocktail Party" problem, a term coined by Cherry in the early 1950s [5] refers to our ability to segregate an auditory signal amongst many (i.e., listening to one person in a room full of people talking) and lock our attention onto one specific target. It is known that we can solve the cocktail party problem more efficiently when there is congruent visual and auditory information. Visible lip movements help in understanding speech both in noise [21] and when mixed with other speech [6].

Although there has been much research done on scanning *visual* attention through the visual scene, and we know that selectively focusing auditory attention is required in complex *auditory* scenes, there has been little research on actually scanning *auditory* attention through the auditory scene. To be strictly analogous with the notion of scanning visual attention in visual search paradigms, an auditory search task would require the listener to find a target sound from a varying set size of spatially dispersed auditory distractors. Some work relating auditory perception to set size has been published. For example, Fairnie et al. (2016) [7] showed that sensitivity to detect a target sound decreased and response time increased with increasing distractor sounds rendered in virtual auditory space. Freyman and colleagues (2004) performed a study in which they presented sound emanating from two speakers with the target coming from one speaker and up to ten distractors from the other speaker. The authors reported an increase in set size resulted in reduced speech intelligibility for the target sentence [8]. In an electroencephalography (EEG) experiment, Hambrook and Tata (2019) found that the degree to which brain dynamics tracks the envelope of an attended talker is reduced with increasing set size of concurrent distractor talkers presented at up to six spatial location [11]. Addleman and colleagues (2019) [1] presented to their observers a spoken digit (target) along with up to three spoken letters (distractors) from different speakers. Participants were asked whether the number was odd or even, or to report the location of the target. Importantly, in this study the target was also cued with either a valid (better than chance) or invalid (equal to chance) arrow. Listeners responded faster for

validly but not invalidly cued targets. Because this study used a spatial cue it is unlike a typical search task. Although the aforementioned studies varied the set size of distractors at different spatial locations (analogous to visual search), to our knowledge, no previous work has investigated crossmodal conjunction of features in searching through auditory space.

To investigate crossmodal search in the auditory scene, the next experiment used a similar approach to that taken in Chapter 2. Our approach was to present audiovisual displays of talkers, and require observers to indicate whether a face/voice stimulus matched (i.e., target-present), despite the presence of distractor *voices*. Similarly to our prediction in Chapter 2, we conceived two contrasting possible outcomes; First, if there exists an efficient preattentive searching mechanism for binding auditory and visual speech in the cocktail party problem, then search for a single coherent face/voice combination should not take longer with increasing distractor set sizes. Furthermore, as depicted in Figure 1.1, if an efficient auditory search for coherent audio/visual speech exists, then target-present and target-absent search slopes should be identical and nearly flat, whereas if serial attentional scanning is required, the target-absent slope should be significantly steeper than the target-present slope. We know from the results in Chapter 2 that audiovisual conjunction of speech dynamics was not of sufficient saliency to enable efficient visual search through a visual space. However, in this chapter we ask whether the conjunction of audiovisual temporal dynamics is sufficiently salient to allow for an efficient "pop-out" search process through a spatial auditory scene with multiple auditory objects.

3.2 Methodology

3.2.1 Ethics Statement

Procedures were in accordance with the Declaration of Helsinki and approved by the University of Lethbridge Human Subject Review Board.

3.2.2 Stimuli and Task

Stimulus Recordings

The stimulus recordings for the following experiment were those used in Experiment 1 and are described under Stimulus Recording in Chapter 2.

Participants

Thirteen undergraduate students were recruited from University of Lethbridge courses and gave informed consent before participating. All reported normal hearing, as well as corrected-to-normal or normal vision. Participants received one course credit for one hour of their participation and were informed, both verbally and in writing, of their options to withdraw at any time during the experiment. Of these, the data from five students were excluded, two of whom required to withdraw from the study and three who did not meet the cut-off criterion of achieving a 65%+ correct answers on the overall experiment. Hence, the data from eight participants (3 female; all right-handed; μ_{age} = 20.5 years; σ_{age} = 2.4495; five native English speakers) were analyzed.

Stimulus Presentation

The auditory stimuli were presented on up to eight Mackie HR624-MK2 powered studio monitors using a Focusrite Scarlett 18i20 3rd Generation audio interface with balanced analog outputs. The studio monitors were arranged in an arc around an iMac desktop; using the center of the iMac screen as a reference, each audio monitor was spaced by 22.5° and 0.9144 m away from the listener (see Figure 3.1 for reference). However, the centre speaker was not used. Thus, target talker and distractor stimuli could appear at +/- 22.5, 45, 67.5, and 90°

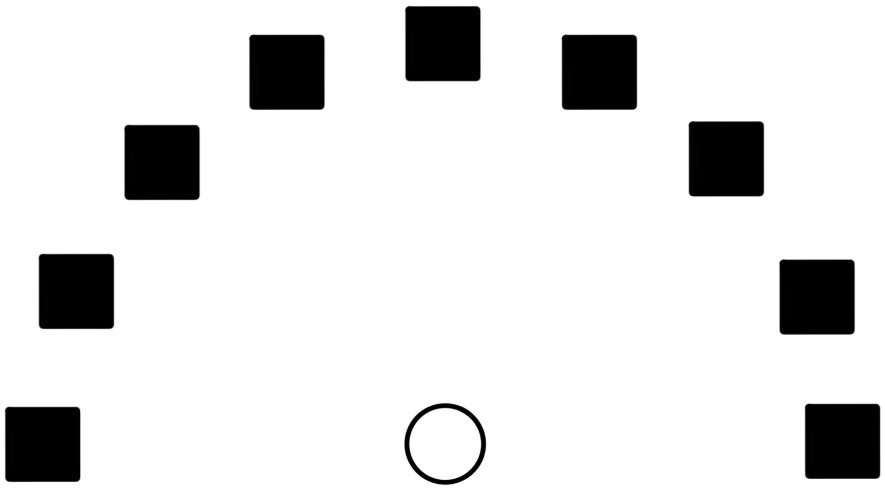

Figure 3.1: **Studio Monitor Arrangement.** The studio monitors were arranged in an arc around an iMac Desktop; using the center of the iMac screen as a reference, each audio monitor was spaced by 22.5° .

Task Requirements

On each trial, participants saw one face displayed on the iMac screen and heard a varying number of voices (the visual display can be observed in Figure 3.2). On half of the trials, the audio of one of the voices matched the lip movements of the face (target present). On the other half of the trials, none of the audio sentences matched the lip movements of the face (target absent). On each trial, participants were asked to respond by pressing on the "p" or "a" key, for target-present and target-absent respectively. Emphasis on both accuracy and speed was made both verbally and in the written instructions.

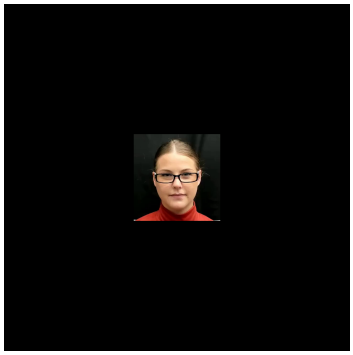

Figure 3.2: **Visual Display.** An example of the visual layout for all set sizes, whether target was present or absent. If the target was present, the trial had the visual stimulus' lips movements matching one of the voices heard. In contrast, if the target was absent, there was no match between what was heard and was was seen, and all were considered distractors. The visual layout was arranged at the center of the screen.

3.2.3 Experimental Setup and Analyses

Experimental setup and analyses were as in Chapter 2 under Experimental Setup and Analyses, respectively.

3.3 Results

Figure 3.3 shows the average response-time performance for each participant on correct trials as a function of both set size and target presence. Response times generally increased in a linear way as we added distractors to the display. A two-way repeated-measure Analysis of Variance performed on the data supports this observation. There was a significant main effect of target presence [$F(1, 84) = 9.96$, $p = .0022$] and a significant main effect of set sizes [$F(5, 84) = 13.2342$, $p < .001$]. However, the results revealed no statistically significant interaction between the two factors [$F(5, 84) = .3515$, $p = .8799$].

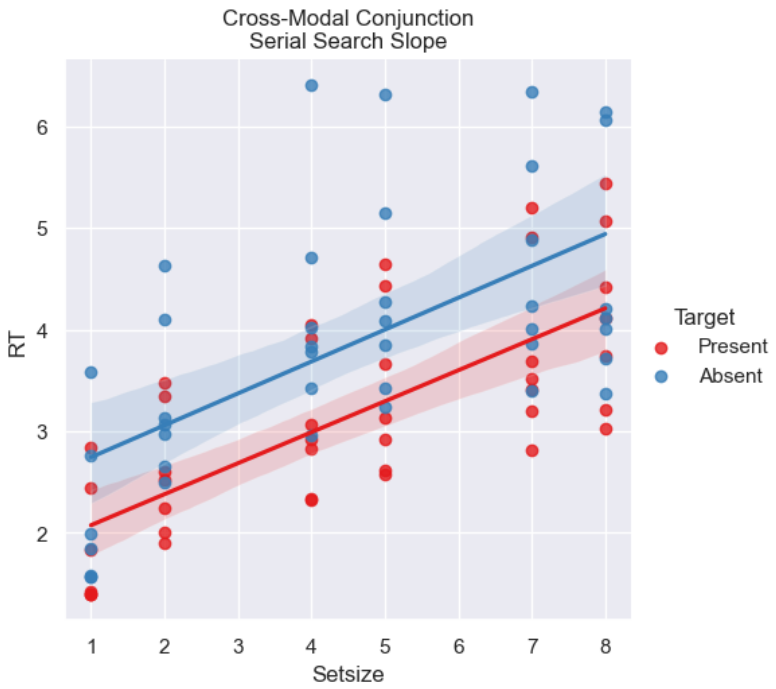

Figure 3.3: **Linear Regression Model.** The average of each participants' response times on each set size for correct trials were used to build a linear regression model with a confidence interval of 95%. The red slope (lower) represents the target-present data, while the blue slope (upper) represents the target-absent data. A two-way repeated-measure ANOVA showed a significant main effect of set size = F = **13.2342** and a significant main effect of target presence F = **9.96**, however, we did not see an interaction between the two factors F = **.3515**.

Comparing the regression line slopes for each subject, we calculated both the target absent and the target present slopes across participants (μ target absent slopes = 0.2973, σ target absent slopes = 0.1066; μ target present slopes = 0.2976, σ target present slopes = 0.0811). We saw no statistically significant difference between the target-present and target-absent slopes (paired one-tailed t-test [t(7) = -0.0192, p = .4924]).

Figure 3.4 shows the accuracy (correct responses) both as a function of target presence and across all set sizes. There was a significant main effect of target presence [F(1, 84) = 20.3237, $p < .001$] and a significant main effect of set size [F(5, 84) = 47.3656, $p < .001$]. Furthermore, the effect of set size was different (more accurate on higher set sizes) for

target-absent relative to target-present trials, indicated by a significant interaction between the two factors [F(5, 84) = 6.7887, $p < .001$].

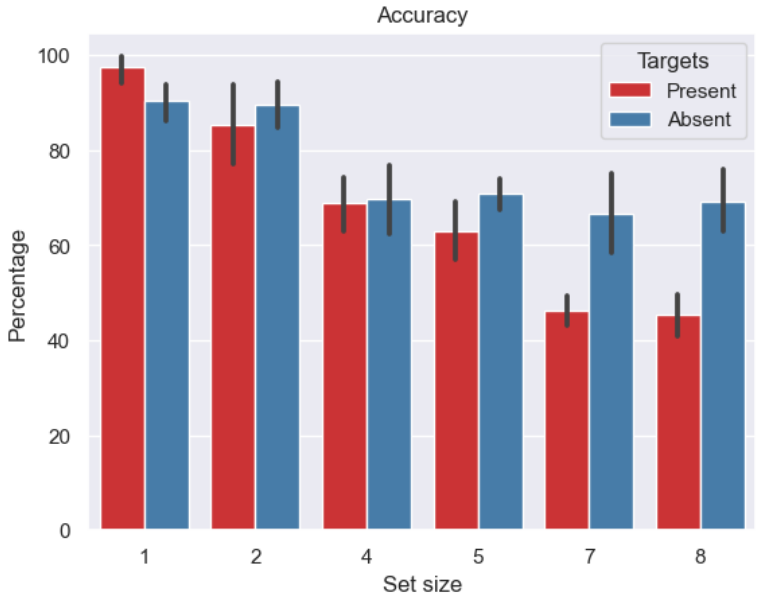

Figure 3.4: **Accuracy** The average of all the participants' correct responses classified by set sizes and by target presence. The red strips (left) represents the average accuracy for the target-present trials, while the blue strips (right) represent the average accuracy for the target-absent trials with error bars representing a confidence interval of 95%. A two-way repeated-measure ANOVA showed a significant main effect of set size F = **47.3656** and a significant main effect of target presence F = **20.3237**, as well as an interaction between the two factors F = **6.7887**. Participants had an average **72.78%** success rate at discriminating whether a target was present or absent.

3.4 Discussion

In this study, our objective was, as per the previous chapter, to determine whether coherent audiovisual speech could be found amongst a set of distractors through an "efficient" low-level mechanism (i.e., "pop-out") or instead required a serial-like searching process. As opposed to our previous chapter, the predominant searching domain we used to conduct the search task was the auditory domain. To do so, we built an experimental paradigm in which we presented audiovisual displays of a speaker and required participants to indicate

whether a face/voice stimulus match existed (i.e., target-present) or there was no face/voice stimulus match (i.e., target-absent).

Analyses of the response times across participants as a function of set sizes showed that an increasing number of distractors resulted in longer time to complete the tasks. Furthermore, it took longer for participants to complete the task on target-absent trials. These results, similar to Chapter 2 results, suggest that search through auditory space for an audiovisual conjunction shares similarities with the typical unimodal visual conjunction searches. Similarly to visual search, audiovisual conjunction search seems to require selectively attending to each item in the display.

Analyses of the linear regression slopes (see Figure 3.3) as a function of target presence showed that there was no difference between target-absent and target-present trials. This result was unexpected. Typical visual conjunction searches, as well the audiovisual conjunction of Experiment 1, result in a stepper target absent slope relative to target present slope. Although there was no statistical difference in slopes, we suggest this null result is due to a lack of statistical power rather than a fundamental difference in how audiovisual conjunctions are found when attention scans through auditory rather than visual space.

Analyses of the accuracy as a function of set size and target presence (see Figure 3.4) suggest an important caveat in understanding the results obtained by the response times. For both conditions, an increasing number of distractors caused lower accuracy. Furthermore, the significant main effect of target presence on accuracy suggests that accuracy was better on target-absent trials than on target-present trials. Importantly, there was a significant interaction between the two factors of target presence and set size: that is, accuracy dropped to near chance performance at higher set size, but only on target-present trials. One possible reason for this interaction is that listeners likely defaulted to a target-absent response on higher set sizes. Unlike complex visual displays which can be parsed effectively using foveal vision, the temporal envelopes of speech distractors are mixed and cannot easily be unmixed when there are many distractors in the scene [11]. Presumably, when a listener

cannot identify any distinct speech envelopes, they tend to respond that the target was absent, which is the correct response on a target-absent trials but not on target-present trials. Hence the interaction between the target presence and set size probably simply reflects the pronounced difference in ability to resolve distinct elements of visual and auditory displays. An alternative speculation relates to overlapping energetic masking. Energetic masking reduces the discriminability of a target among distractors because their energy overlaps on the tonotopy of the basilar membrane. Since the degree of energetic masking depends on the number of sound sources in the auditory scene, this factor was systematically confounded with set size and could explain the increasing error rate as set size increased. This could, in turn, cause a speed-accuracy trade-off resulting in slower responses for higher set sizes.

In the cocktail party problem, described in the introduction of this chapter, we have a clear example of the benefits afforded to perception by selectively attending to a single talker amongst distractors. Those benefits involve the spatial separation of sounds [7, 8, 12, 13] and it is helpful to have visual information to solve this problem (i.e., lip movements) [6, 21]. This work on solving the cocktail party problem specifically addresses the benefits of selective attention but not how the brain manages to move selective attention through auditory space. Our experiment suggest that even in the simple case of matching a single visual target to a set of auditory target requires serial selective attention. In other words, audiovisual targets cannot ”pop-out” of the cocktail party. That fact might impose limits on effective listening on complex scenes. An important caveat is that the target and distractors in our experiment were the same voice and thus shared fundamental frequency and harmonics. In a realistic auditory scene, a target and a distractor would have different voices and thus be separated in the frequency spectrum. It remains uncertain whether audiovisual conjunctions pop-out when sounds can be separated along two dimensions (i.e., space and frequency).

Chapter 4

Discussion and Future Directions

4.1 Discussion

Objects that are sufficiently unique from their background, and from other objects around them, tend to ”pop-out” from the visual scene by a parallel search mechanism. Other objects, which are not entirely unique or are defined by conjunctions of features, require us to serial scan of attention through the visual scene in order to find them. This problem of finding a target amongst distractors is not unique to vision. Auditory objects can also appear mixed amongst other distracting sounds. Thus in realistic scenes, the perceptual brain is constantly trying to decompose and re-bind features together into multisensory objects. Very little is know about how attention is guided through the multisensory world. In this thesis, we investigated whether crossmodal conjunction of audiovisual speech could be found by an efficient ”pop-out” mechanism of search, or whether it requires serial scanning of the multisensory scene.

In Chapter 2, we described a multisensory attention search experiment. In this experiment the target was the conjunction of a voice and a face speaking the same sentence. A varying set size of distractors was also presented. These distractors were visual faces saying other sentences at other locations in the visual display. Thus in this experiment there was one audible voice and a varying number of visual faces. Only one face was coherent and matching the voice on half of target trials (i.e., target-present). Our prediction was as follows: If a matching face/voice pair can be found by an efficient parallel search mechanism, then we would expect the slope of the line relating response time to the set size to be

nearly flat and identical for target present and target absent trials. Otherwise, if a matching face/voice pair required a serial search mechanism, then we would expect this line to have positive slope and be steeper on target-absent relative to target-present trials. We found the second pattern in our results: response times to find the conjunction of audiovisual features increased with distractor set size. This strongly suggests the requirement of a serial search mechanism.

These results fill a gap in the literature. Fujisaki and colleagues (2005) [31] had found a similar result with simple audiovisual stimuli, while Van der Burg and colleagues (2008, 2010) [30, 29] had shown that audiovisual ”pop-out” can occur under specific conditions of temporal dynamics. It was unclear whether audiovisual speech offers the same temporal dynamic that allow for ”pop-out” of targets in an audiovisual scene. Chapter 2 showed that serial search is required to find a face/voice match, at least when the face appears amongst visual distractors but the audio appeared alone.

Our experiment in Chapter 2 did not address the complementary situation in which one face could appear amongst varying set size of voices. Thus in Chapter 3, we described an analogous multisensory attention search experiment in which the audio target appeared amongst varying set size of audio distractors and only one visual face was presented. This paradigm calls to mind the cocktail party problem in which a single target talker is selectively attended in a complex auditory scene. In this case that single talker was defined as having the same dynamics as the visual face. Similarly to Chapter 2 we found evidence that a serial attention scanning was required through the space of auditory distractors to find the crossmodal target.

4.2 Limitations

Although the results presented here were unequivocal in showing that search for audiovisual speech fits a serial-like pattern, there were some factors that limited the interpretability of the research. We had good representation of natural speech in the stimuli used, but

the sentences had a low ecological validity because they were read, rather than spontaneously spoken sentences, and they were presented outside of their context (i.e., ”She had your dark suit in greasy wash water all year”). Furthermore, this thesis was conducted during a pandemic which limited our ability to collect data for a set of experiments that might have explored some interesting caveats about the audiovisual search described in Chapter 3. For example, the auditory objects presented in Chapter 3 were arrayed throughout auditory space but not frequency space (i.e., they were from the same talker saying different sentences, thus had similar frequency characteristics). In a real world cocktail party problem, an observer would not need to unmix identical voices and would instead have access to variations along the frequency dimension as well as the spatial dimension. The auditory system is primarily tonotopic, this would be similar to faces overlapping at the same location in the visual display. Moreover, we suspect that in Chapter 3 we lacked enough statistical power to show an interaction between set size and target presence, and we discovered that our task became nearly impossible to solve at larger set sizes (e.g., participants performed near chance on higher set sizes for target present trials). Therefore, although the pattern of results obtained in Chapter 2 neatly matches the pattern expected in a serial-search process (i.e., an interaction between set size and target presence), we do not obtain that pattern in Chapter 3. This might be due to a simple lack of statistical power. Future experiments are anticipated to extend the results of Chapter 3 and to better explore the various factors that account for search in a complex audiovisual scene.

4.3 Conclusion

As a result of the two experiments presented in this thesis, we conclude that audiovisual conjunction search requires serial scanning of attention. Despite the ubiquity of audiovisual conjunction in the sensory world, and the importance of audiovisual speech for human communication, the brain seems to lack an efficient mechanisms to bind audio and visual temporal dynamics of speech, in these experimental conditions. This represents an impor-

tant computational limitation on the perceptual brain. It remains unclear why dynamic sensory mechanisms should lack an efficient temporal binding mechanism for such important stimuli and why selective attention is required to perform binding of audiovisual speech.

Bibliography

[1] Addleman Douglas A. and Jiang Yuhong V. The influence of selection history on auditory spatial attention. Journal of Experimental Psychology: Human Perception and Performance, 45(4):474–488, 2019.

[2] Treisman A. Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology, 8(2):194–214, 1982.

[3] Jens Blauert. Spatial hearing: the psychophysics of human sound localization. MIT press, 1997.

[4] D. H. Brainard. The psychophysics toolbox. Spatial Vision, 10:433–436, 1997.

[5] Cherry E. Colin. Some experiments on the recognition of speech, with one and with two ears. The Journal of the Acoustical Society of America, 25(5):975–979, 1953.

[6] Zion Golumbic Elana, Cogan Gregory B., Schroeder Charles E., and Poeppel David. Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". Journal of Neuroscience, 33(4):1417–1426, 2013.

[7] Jake Fairnie, Brian C. J. Moore, and Anna Remington. Missing a trick: Auditory load modulates conscious awareness in audition. Journal of Experimental Psychology: Human Perception and Performance, 42(7):930–938, 2016.

[8] Richard L Freyman, Uma Balakrishnan, and Karen S. Helfer. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. The Journal of the Acoustical Society of America, 115 5 Pt 1:2246–56, 2004.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgreen, and V. Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993.

[10] T. D. Griffiths and J. D. Warren. What is an auditory object? Nat Rev Neurosci, 5(11):887–92, 2004.

[11] Dillon A Hambrook and Matthew S Tata. The effects of distractor set-size on neural tracking of attended speech. Brain and language, 190:1–9, 2019.

[12] Gerald Kidd, Christine Mason, Virginia Best, and Nicole Marrone. Stimulus factors influencing spatial release from speech-on-speech masking. The Journal of the Acoustical Society of America, 128:1965–78, 10 2010.

[13] Gerald Kidd, Christine R Mason, Virginia M Richards, Frederick J Gallun, and Nathaniel I Durlach. Informational masking. Auditory perception of sound sources, pages 143–189, 2008.

[14] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. What's new in psychtoolbox-3. Perception, 36(14):1–16, 2007.

[15] H.R. Liesefeld and H.J. Müller. A theoretical attempt to revive the serial/parallel-search dichotomy. Attention, Perception, and Psychophysics, 8:228–245, 2020.

[16] H. McGurk and J. McDonald. Hearing lips and seeing voices. Nature, 264:746–748, 1976.

[17] Amanda R. McMullan, Dillon A. Hambrook, and Matthew S. Tata. Brain dynamics encode the spectrotemporal boundaries of auditory objects. Hearing Research, 304:77–90, 2013.

[18] MoviePy. `https://github.com/Zulko/moviepy`. Accessed: 2022-09-01.

[19] J. Palmer. Attention in visual search: Distinguishing four causes of a set size effect. Current Directions in Psychological Science, 4(4):118–123, 1995.

[20] D. G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10(4):437–442, 1997.

[21] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. Journal of The Acoustical Society of America, 26:212–215, 1954.

[22] G.J. Thomas. Experimental study of the influence of vision on sound localization. Journal of Experimental Psychology, 28(2):163–177, 1941.

[23] J.T. Townsend. A note on the identification of parallel and serial processes. Perception and Psychophysics, 10:161–163, 1971.

[24] J.T Townsend. Serial and parallel processing: Sometimes they look like tweedledum and tweedledee but they can (and should) be distinguished. Psychological Science, 1(1):46–54, 1990.

[25] J.T. Townsend and M.J. Wenger. The serial-parallel dilemma: A case study in a linkage of theory and method. Psychonomic Bulletin and Review, 11:391–418, 2004.

[26] A. Treisman and G. Gelade. A feature integration theory of attention. Cognitive Psychology, 12(1):97–136, 1980.

[27] A. Treisman and S. Sato. Conjunction search revised. Journal of Experimental Psychology. Human Perception and Performance, 16(3):459–478, 1990.

[28] Anne Treisman and Hilary Schmidt. Illusory conjunctions in the perception of objects. Cognitive Psychology, 14(1):107–141, 1982.

[29] E. Van der Burg, J. Cass, C.N. Olivers, J. Theeuwes, and D. Alais. Efficient visual search from synchronized auditory signals requires transient audiovisual events. PLoS One, 5, 2010.

[30] E. Van der Burg, C.N. Olivers, A.W. Bronkhorst, and J. Theeuwes. Pip and pop: Nonspatial auditory signals improve spatial visual search. Journal of Experimental Psychology Human Percept Performance, 34(5):1053–1065, 2008.

[31] Fujisaki W., Koene A., Arnold D., Johnston A., and Nishida S. Visual search for a target changing in synchrony with an auditory signal. Proceedings of the Royal Society B:Biological Sciences, 273:865–874, 2005.

[32] J.M Wolfe. Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review, 1:202–238, 1994.

[33] J.M. Wolfe. Approaches to visual search: Feature integration theory and guided search. The Oxford handbook of attention, 11:35–44, 2014.

[34] J.M. Wolfe. Guided search 6.0: An updated model of visual search. Psychonomic Bulletin and Review, 28:1060–1092, 2021.