Time-compressed spoken words enhance driving performance in complex visual scenarios

Evidence of crossmodal semantic priming effects in basic cognitive experiments and applied driving simulator studies

Dissertation zur Erlangung des akademischen Grades eines Doktors
der Philosophie der Philosophischen Fakultät III
der Universität des Saarlandes

vorgelegt von
Angela Castronovo (geb. Mahr)
aus Fürstenfeldbruck

Saarbrücken, Januar 2014

Dekan:

Prof. Dr. Roland Brünken, Universität des Saarlandes

Berichterstatter:

Prof. Dr. Dirk Wentura, Universität des Saarlandes

Prof. Dr. Roland Brünken, Universität des Saarlandes

Tag der Disputation: 30.06.2014

Acknowledgements

I want to thank

- my supervisor Dirk Wentura who has sparked my interest in cognitive psychology. I am particularly grateful for his enthusiasm, experience, and advice regarding the transfer of applied ideas to solid experimental grounds.
- the head of our Automotive Group at DFKI, Christian Müller, who opened up the possibility for me to work in an open-minded, interdisciplinary environment, and who has encouraged me to generate and pursue my own ideas.
- Rafael Math and Till Maurer for developing the OpenDS driving simulator, and for granting any conceivable support regarding my specific requirements.
- Michael, Rafael, Sandro, Mehdi, Monica, and all my other colleagues at DFKI for creating a pleasant working atmosphere, and for their fantastic support in little and big matters.
- my friends Birte, Melanie, Andrea, Kalina, and Charlotte from the Department of Cognitive Psychology, for all the enjoyable Mensa lunches, the practical advice, and discussions.
- my Hiwis, Serkan, Frank, and Verena for conducting the experiments.
- Birte and Kirstin for their valuable comments on this work.
- my husband Sandro who kept me motivated, after all.
- Vivian for clearing my mind after sitting at the desk for too long.

Author note

Chapter 3 and Chapter 4 of this thesis are based on a published article (Mahr, A. & Wentura, D. (2014). Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets. *Attention, Perception, & Psychophysics, 76,* 575-590). I am the first author of this article. In order to allow smooth reading, the respective parts are not especially marked. Furthermore, I employ the term “we” instead of “I” throughout the whole thesis.

Abstract

Would speech warnings be a good option to inform drivers about time-critical traffic situations? Even though spoken words take time until they can be understood, listening is well trained from the earliest age and happens quite automatically. Therefore, it is conceivable that spoken words could immediately preactivate semantically identical (but physically diverse) visual information, and thereby enhance respective processing. Interestingly, this implies a crossmodal semantic effect of auditory information on visual performance.

In order to examine this rationale, eight experiments were conducted in this thesis. Starting with powerful interference paradigms from basic cognitive psychology, we approached more realistic scenarios step by step. In Experiments 1A and 1B, we employed a crossmodal Stroop color identification task with auditory color words as primes and visual color patches as targets. Responses were faster for congruent priming in comparison to neutral or incongruent priming. This effect emerged also for auditory primes which were time-compressed to 30 % or 10 % of their original length, and turned out to be even more pronounced under high-perceptual-load conditions. In order to rule out stimulus-response compatibilities as a cause of the congruency effects, we altered the task in Experiment 2: After brief target displays merely target-present or -absent decisions had to be made. Nevertheless, target detection (d') was increased by congruent primes in comparison to incongruent or neutral primes. These results suggest semantic object-based auditory–visual interactions which automatically increase the denoted target object's salience. Importantly, intentional or strategic listening was ruled out since the presentation of primes did not predict the identity of the subsequent targets: The conditional probability of a given prime being followed by a semantically matching target was only at chance level (no contingency). Nevertheless, crossmodal semantic priming effects were efficient and fast: They particularly occurred in complex visual scenes and at an SOA of only 100 ms.

For all following experiments we exchanged colors as the relevant object feature by more specific automotive target icons and their two-syllable denotations. Using these materials and time-compression (to 50 % and 30 % of

the original length) in Experiment 3A, we could replicate our earlier findings from Experiments 1A and 1B. Moreover, since warning systems likely present meaningful information, we slightly increased contingency from chance level (25 %) to 50 % in Experiment 3B. Overall, the pattern of benefits and costs remained similar to Experiment 3A. Interestingly, spoken word primes that were compressed down to 30 % of their original duration effectively led to faster reaction times when they were slightly predictive than when they were completely irrelevant (Exp. 3A vs. 3B). Accordingly, time compression improved responses already when listening became minimally useful.

In another important step towards more realistic scenarios, we essentially transferred our approach to three experiments (4, 5A, and 5B) in a 3D driving simulator. Now gantry road signs showed task-relevant visual information, and, besides, we replaced explicit target identification by simple color classification (red vs. green) of the respective target. Thereby, primes and responses were orthogonal, and stimulus-response compatibility effects could again be eliminated. Pronounced benefits were revealed both in a simple button-press task on the steering wheel (Exp. 4), and also when participants executed complex driving maneuvers (Exps. 5A and 5B). When semantic contingency was increased considerably (to 80 %) in Experiment 5B, crossmodal facilitation and interference were both increased to a similar extent with the typical asymmetric pattern being retained.

In summary, while employing different materials, paradigms, and dependent variables, we repeatedly found crossmodal benefits in visual task performance due to semantically congruent spoken words. These pronounced effects occurred even though words were presented only extremely shortly before object detection or classification, and, most importantly, even if listening to spoken words was not beneficial (or intended). Based on our findings, we assume that spoken words generally have the potential to rapidly preactivate visual information via semantic representations. Moreover, strategic listening effects seem to even add on top of this automatic process. These basic findings are highly encouraging for information transfer via speech warnings in time-critical on-road situations, since faster situation assessment and improved reactions could mitigate or even prevent accidents.

Kurzzusammenfassung

Wäre es in zeitkritischen Straßenverkehrssituationen sinnvoll, dem Fahrer relevante Informationen via Sprachwarnung zu übermitteln? Obwohl Zeit verstreicht bis gesprochene Wörter verstanden werden können, wird das Sprachverständnis von frühester Kindheit an trainiert und läuft weitgehend automatisch ab. Deshalb ist es grundsätzlich denkbar, dass gesprochene Wörter semantisch identische (aber physikalisch verschiedene) visuelle Informationen direkt voraktivieren können und somit deren Verarbeitung verbessern. Interessanterweise impliziert dies einen modalitätsübergreifenden semantischen Effekt von auditiven Informationen auf die visuelle Performanz.

Zur Untersuchung dieser Annahme wurden in der vorliegenden Arbeit acht Experimente durchgeführt. Ausgehend von präzisen, grundlagenpsychologischen Interferenzparadigmen wurden schrittweise realitätsnähere Bedingungen verwendet. In Experiment 1A und 1B wurde eine modalitätsübergreifende Stroop-artige Farbidentifizierungsaufgabe eingesetzt. Hierbei dienten auditive Farbwörter als *Primes*[1] und visuelle Farbkreise als Zielreize (SOA 100ms). In Durchgängen mit kongruenten *Primes* reagierten die Versuchspersonen schneller auf die Zielreize als bei neutralen oder inkongruenten *Primes*. Dieser Effekt trat auch für zeitlich komprimierte Wörter (30 % bzw. 10 % Länge) auf, und verstärkte sich sogar bei visuell belastenden Aufgaben. Um Kompatibilitätseffekte zwischen *Primes* und Reaktionen als Ursache für semantische Kongruenzeffekte auszuschließen, wurde in Experiment 2 die Aufgabe gewechselt: Es musste nach kurzer Darbietungsdauer lediglich entschieden werden, ob ein Zielreiz präsentiert worden war. Selbst unter diesen Umständen wurde die Entdeckungsleistung (*d'*) durch kongruente *Primes* im Vergleich zu neutralen oder inkongruenten *Primes* verbessert. Diese Ergebnisse legen nahe, dass semantische, objektbasierte auditiv-visuelle Interaktionen die Salienz des bezeichneten Objektes weitgehend automatisch erhöhen. Hierbei ist relevant, dass eine intentionale oder strategische Nutzung der Wörter dadurch ausgeschlossen wurde, dass die dargebotenen *Primes* die Identität der nachfolgenden Zielreize in keiner Weise

[1] *Prime*: Ein vorab präsentierter Reiz, der die Verarbeitung eines nachfolgenden Zielreizes (möglicherweise) beeinflusst.

vorhersagten: Es bestand keine Kontingenz, da die bedingte Wahrscheinlichkeit für eine semantische Passung nur auf Zufallsniveau (25%) lag. Dennoch waren die modalitätsübergreifenden semantischen *Priming* Effekte effizient und schnell.

Für alle weiteren Experimente wurden konkrete Icons aus dem Straßenverkehrskontext und deren (zeitkomprimierte) zweisilbige Bezeichnungen eingesetzt. In Experiment 3A konnten damit die bisherigen Befunde repliziert werden. Da Warnsysteme eher inhaltlich passende Informationen darbieten, wurde zudem in Experiment 3B die Kontingenz leicht auf 50 % erhöht. Dennoch zeigte sich hier ein ähnliches Ergebnismuster. Allerdings führten bei stark zeitkomprimiertem Material die leicht informativen *Primes* zu kürzeren Reaktionszeiten (Exp. 3A vs. 3B). Demzufolge verbesserte die zeitliche Kompression die Reaktionen interessanterweise bereits dann, wenn das Zuhören geringfügig nützlich war.

In einem weiteren Schritt in Richtung realitätsnaher Szenarien wurde die bisherige Vorgehensweise auf drei Experimente (4, 5A, und 5B) in einem Fahrsimulator übertragen. Schilderbrücken zeigten die aufgabenrelevanten Informationen an. Die Zielreize mussten nicht explizit identifiziert sondern entsprechend ihrer Farbe klassifiziert werden. Dadurch verhielten sich *Primes* und Zielreize orthogonal und Kompatibilitätseffekte zwischen *Primes* und Reaktionen konnten ausgeschlossen werden. Sowohl bei einer einfachen Tastendruckaufgabe (Exp. 4) als auch bei komplexen Fahrmanövern (Exp. 5A und 5B) konnten erneut starke Reaktionszeitvorteile für kongruente *Primes* gefunden werden. Wenn die Kontingenz deutlich angehoben wurde (80 % in Exp. 5B), erhöhten sich sowohl Reaktionszeitvorteile als auch -kosten in gleichem Ausmaß.

Zusammenfassend zeigt sich, dass unter Verwendung verschiedener Materialien, Paradigmen und abhängiger Variablen wiederholt deutliche, modalitätsübergreifende Vorteile durch semantisch kongruente gesprochene Wörter bei der visuellen Aufgabenperformanz gefunden wurden. Diese starken Effekte traten auf, obwohl die Wörter nur sehr kurz vor dem Zielreiz dargeboten wurden, und bemerkenswerterweise sogar dann wenn deren Beachtung weder vorteilhaft noch beabsichtigt war. Auf Grundlage unserer Ergebnisse nehmen wir an, dass gesprochene Wörter generell das Potenzial

haben, über semantische Repräsentationen sehr schnell visuelle Information zu aktivieren. Darüber hinaus scheint eine strategische Beachtung der Wörter den Effekt noch zu verstärken. Diese grundlegenden Erkenntnisse sind vielversprechend für die sprachliche Übermittlung von Informationen in zeitkritischen Straßenverkehrssituationen, denn ein schnelleres Erfassen der Situation und kürzere Reaktionszeiten können die Schwere von Unfällen reduzieren oder diese bisweilen so gar verhindern.

Table of contents

List of Figures

List of Tables

Preface

Even though driving is a very complex process, it becomes highly automated and less demanding after some practice – above all if it is a routine part of our daily life. However, occasionally and unforeseeably, there occur critical situations drivers cannot easily cope with. Imagine you are driving in an unfamiliar urban environment trying to find your way to your destination. As roads are busy and you are already late for the meeting, your colleague tries to support you from the co-driver's seat. You have just overtaken a cyclist and are approaching a large intersection. While you are just switching to the proper lane for the next turn, your colleague yells all the sudden: "Red light!". You instantly notice the red traffic light and you are able stop the car just in time before hitting a crossing pedestrian. Probably, you will thank your colleague for this support.

Fortunately, a co-driver is not the only potential support in safety-critical situations, since today many cars are already equipped with intelligent assistance systems. However, it is still a pending issue, how these systems should warn and/or inform the driver about the specific relevant event in time-critical situations. With respect to the example given above: Would drivers prefer to hear a warning sound and have a look at a display in order to find out what the sound denotes? Should designers of in-car warnings take into consideration what a vigilant co-driver would (intuitionally) do to assist the driver? All in all, effectiveness and efficiency of warnings need to be closely investigated before making a decision, since these are the most relevant factors in safety-critical driving situations. Even time saving of only a few milliseconds can be crucial with regard to crash avoidance or crash severity (see, e.g., Gray, 2011).

Nowadays, advanced driver assistance systems (ADAS) offer a multitude of information about the road environment, which can be used to support drivers in critical road situations. For example, the driver can be explicitly warned via different modalities, the vehicle could actively intervene, or safety functions are pre-activated. The ADAS systems, which are currently available in series-production cars, are based upon passive sensor technologies.

These sensors are self-contained within the individual vehicle and they collect information about the environment, the car, and the driver. Due to advances in the domain of cooperative vehicles, information from other cars can additionally be taken into account. Especially with regard to time-critical safety-relevant applications, the usage of modern Vehicle2X technologies improves the scope of ADAS, since additional real-time information becomes available.[2] During the last decades, numerous research projects (e.g., DRIVE C2X [3], Stahlmann, Festag, Tomatis, Radusch, & Fischer, 2011, sim^{TD} [4], Assenmacher, 2009) have fostered technological advances in the Vehicle2X domain resulting in a large number of additional use cases (e.g., warnings of congestions, emergency vehicles, bad road weather conditions, obstacles on the road, red traffic lights, cross-traffic). However, the majority of these projects focused on technical and standardization issues, whereas human factors issues received only little attention (for a review, see Strandén, Uhlemann, & Ström, 2008). Accordingly, the design of adequate human-machine interfaces for specific information transfer from technical systems to the driver in time-critical situations is still an open issue. So far, the in-car presentation of either low-priority messages or of only very few time-critical warnings has already been investigated in numerous human factors projects. For example, the assignment of different information types to different codes, modalities, or locations has been identified as an effective approach (we will return to related findings later in Chapter 1.2). However, during the last decade, the increased number of established ADAS functions has almost exhausted the potential of hitherto valuable information transfer approaches. Accordingly, it would be premature and shortsighted to simply apply those existing concepts also to the near future when a considerable multitude of upcoming diverse warnings needs to be presented. Nowadays, for example, using a simple tonal or haptic warning in combination with a display showing textual information is a common practice (see, e.g., Campbell, Richard, Brown, & McCallum, 2007;

[2] Vehicle2X refers to vehicles using wireless communication technologies to enable information exchange. X, in this case, denotes communication partners, e.g., vehicles, road infrastructure, or traffic authorities.

[3] DRIVE C2X – Accelerate cooperative mobility, http://www.drive-c2x.eu/project

[4] Sim^{TD} – Safe Intelligent Mobility Test Area Germany, www.simTD.de

Lee, McGehee, Brown, & Reyes, 2002). Even though this concept is easy to implement and can be applied to all different types of use cases, it is obviously not beneficial to adopt this strategy in complex road scenarios when a driver needs to reliably recognize which out of numerous time-critical warnings is being presented. For disambiguation of an alerting tone and in order to understand the content of the concrete warning, the driver would be forced to check the display with the gaze being drawn off the road. At the same time, external situation assessment and possibly even a response would already become necessary. For example, the driver might need to immediately notice a red traffic light, cross-traffic at an intersection, give way to an emergency vehicle, circumnavigate an obstacle, or initiate emergency braking. Since any of these situations might occur at a given point of time, an unspecific alert will probably not provide sufficient information for the driver to optimally cope with the situation. In contrast, transferring additional (semantic) information about the critical incident would probably be beneficial for instantaneous situation assessment and response selection. The question arises whether, instead of tonal or haptic alerts with visual disambiguation, there are other promising approaches available for more efficient and less distractive transfer of rather complex information. For time-critical situations, we suggest that the system could, analogous to a vigilant co-driver, present time-compressed speech to quickly inform the driver about the current safety issue or a suggested action. However, before evaluating the usability of any new warning approach, general knowledge about the involved modalities needs to be systematically recollected, and, more specifically, it is necessary to review related work on previous findings. Accordingly, we will revisit the usage of speech warnings in the automotive domain. Besides, we definitely argue for the consideration of more general knowledge about influences from spoken words on visual situation awareness and subsequent responses. In our opinion, methodologically sound findings from cognitive psychology can lay valuable and reliable foundations for recommendations on concrete use cases. Consistently, we will consider and interpret our empirical findings under a more general theoretical light of human information processing. Contrary to most experiments on applied research questions, which are tailored to specific use-cases, our principle-based and stepwise approach enables a transfer of

fundamental findings to an even broader range of applications, systems, or domains.

This thesis addresses the high-level research question identified above: How could specific warnings from diverse technical systems be efficiently transferred to the driver in time-critical situations? Structure and contents of this thesis reflect our general approach to pursue specific, applied research questions down to a level of general knowledge regarding human cognitive functioning. Therefore, we begin with a description of the challenges regarding in-car warnings for a multitude of time-critical warnings (Section 1.1). After briefly revisiting literature on warning modalities and codes (Section 1.2), we especially focus on spoken words and other auditory semantics in Chapter 2. In order to derive additional knowledge for speech-warning scenarios, we subsequently aggregate findings from cognitive psychological literature regarding general crossmodal influences from spoken words on the detection of visual targets (Section 3.2). On grounds of several inconsistencies and pending issues in related literature (Section 3.3), we derive respective hypotheses and outline a sequence of tailored experiments (Section 3.4) previous to the subsequent empirical chapters. Chapter 4 comprises three basic computer experiments on crossmodal speech priming in order to analyze relevant preconditions, limiting factors, and underlying mechanisms. This provides a reliable basis for further steps towards more applied and less controlled investigations. With Chapter 5, we turn towards more complex stimuli in terms of visual and auditory demands. Moreover, the crucial issue of contingent crossmodal information presentation and its influence on effects is addressed. In Chapter 6, we take another step towards an applied speech-warning scenario and present the results of three driving simulator experiments. For this purpose, we designed an icon classification task including gantry road signs and complex driving maneuvers. In Chapter 7 we discuss what information can be derived from our experimental series regarding crossmodal priming by spoken words, on the one hand, and warnings in time-critical driving situations, on the other hand. Besides, we mention some issues that are left for future research previous to the final section with our concluding remarks.

1 Informing drivers in safety-critical traffic situations

Before warning and alerting strategies can be integrated into series-production vehicles as part of driver assistance systems, they need to be put to the acid test. The reason is that design decisions have the potential to severely influence drivers' reactions in critical situations and accordingly affect crash rates and accident severity (see, e.g., Dingus et al., 1996; Lee, Hoffman, & Hayes, 2004; ISO/TR 16352:2005(E); Lee et al., 2002). Given this backdrop, consideration of existing theoretic frameworks on driving performance can help to structure design decisions according to human cognitive processes. Driving can be seen as a process comprising three sub-components (Ho & Spence, 2008; Kantowitz, Triggs, & Barnes, 1990): A perceptional sub-component predominantly dealing with visual information (Sivak, 1996), a decisional sub-component which also influences attentional processes, and several response-related sub-components for the selection and execution of motor responses. If knowledge about these basic psychological components is taken into consideration, the choice of information presentation modalities, codes, density, and timing for warnings can be presumably structured and optimized. Accordingly, we will revisit this framework in the corresponding sections.

The present chapter provides insights, how technical advances can increase the safety potential in time-critical driving situations. Besides, efficient information transfer from car to driver is underlined as a necessary precondition in time-critical situations. Currently existing approaches focus on presentation via three modalities (visual, auditory, and tactile), and on combinations thereof. Each of these approaches is briefly revisited with particular emphasis on urgent information transfer in order to underpin which presentation modalities can be recommended. This superordinate choice of modality allows narrowing the subsequent investigation down to the most promising options.

1.1 Technical advances in driver assistance systems enable a multitude of informational warnings

During the last decades, numerous in-vehicle information systems (IVIS) and (advanced) driver assistance systems (ADAS) have been introduced into cars and other vehicles (Stevens, 2009). The development of these systems has primarily been pushed by the assumption that technological progress has a great potential to increase road safety. Information is aggregated and pre-evaluated by a technical system and then presented to the driver. This is supposed to enhance driving performance and to avoid or mitigate accidents. However, the concrete design of respective human machine interfaces has turned out to play a crucial role for information transfer from car to driver (see, e.g., ISO/TR 16352:2005(E); Lee, Gore, & Campbell, 1999) and hence for a system's effectiveness, efficiency, and the user's satisfaction (ISO 9241-11:1998(E)). Therefore, extensive research has been conducted, in order to choose or invent the best possible interface solution for each system.

During the next years, applications based on emerging wireless technologies (e.g., Vehicle2X communication) will provide extensive additional information for the driver. In this regard, the fact that many safety-related applications require instantaneous presentation will become a critical issue. For example, warnings about icy roads, red light violations, emergency vehicles, crossing vehicles, or braking cars are being developed today among a large number of other applications (see, e.g., project sim^{TD}). Accordingly, on the one hand, prioritization of applications will be required (see, e.g., Castronovo, Mahr, & Müller, 2013; Wolf, Zöllner, & Bubb, 2005), and, on the other hand, highly efficient information presentation will be decisive for the enhancement of safety.

Due to an increased number of highly diverse, complex, and time-critical use cases, the traditional solutions for information transfer gradually reach their limits. For example, alerting tonal sounds have been recommended and successfully used as long as only a few in-car warnings had to be presented (see, e.g., Campbell et al., 2007; Lee et al., 2002). Under these circumstances, each sound was directly associated with a situation or meaning. However, with an increasing number of warnings the situation changes fundamentally: Drivers

are no longer able to reliably distinguish and remember a large number of single tones as warning sounds. Even after a learning phase, pilots were not able to remember more than eight to nine out of ten auditory warnings after one week (Patterson & Milroy, 1979). As a result, for warnings in military aircrafts, a maximum of only four tones must not be exceeded. Thereby, absolute discrimination of information can be ensured and confusion should be prevented (MIL-STD-14-72F, 1999). Accordingly, simply assigning new tones to the additional warnings cannot be seriously considered as a meaningful warning approach for untrained drivers.

Besides the multitude of different applications, warnings based on new technologies also offer an increased level of information detail. For example, a warning of an emergency vehicle can also comprise its approaching direction. In this case, a spatial tonal or verbal warning could inform the driver about the respective direction (Ho & Spence, 2005; Selcon, Taylor, & McKenna, 1995; Wang, Pick, Proctor, & Yeh, 2007). However, spatial mapping within a car is not undisputed (Fricke, 2009), and, most important, a spatial tone would not let the driver know which out of many possible objects around might be critical, or which type of situation has emerged (e.g., emergency vehicle, cross-traffic, or red light). Even though modern technology offers this information, a driver could not infer the complete situational information from a simple spatial tone. If sufficient time is available, visual displays are a valuable approach for the disambiguation of tonal alerts (e.g., Campbell et al., 2007). However, it remains a critical issue for further consideration, whether drawing the driver's gaze off the road is advisable in time-critical situations when a hazards needs to be detected (see, e.g., Horrey & Wickens, 2004; Seppelt & Wickens, 2003).

The following chapter briefly revisits existing warning approaches for driver assistance systems, and it evaluates, which modality seems most suitable for the upcoming challenge that we have just introduced: The presentation of multiple different and complex warnings in time-critical situations.

1.2 Driver assistance systems: Existing modalities and codes

Whenever a warning system presents information, it is first of all necessary that drivers perceive it. For information presentation purposes, besides optical

(visual) displays, also acoustic (auditory), and haptic (tactile) displays are commonly used in modern passenger vehicles. The following paragraphs comprise a condensed introduction to the human factor literature on the denoted presentation alternatives. Furthermore, attentional aspects and effects on human performance will be highlighted.

1.2.1 Visual warnings

According to ISO/TR 16352:2005 (E), visual information can be presented via different codes: Analogue display elements (e.g., conventional speedometer), discrete display elements (e.g., low fuel indicator), digital display elements (e.g., mileage reading), alphanumerical display elements (e.g., "Fasten seatbelt!" text), or symbols (e.g., oil pressure indicator) can be used.

For visual information presentation in general, sensorial-related parameters are crucial for signal perception. For example, size, brightness, contrast, or consistency of the presented information contribute to visibility, readability, and efficient search processes (Dingus et al., 1996; ISO/TR 16352:2005(E). Static display of visual information can lead to slower signal detection or even to complete miss of a warning. In contrast, blinking visual information is more conspicuous, however, it might also interfere with ongoing visual or mental processes while driving. Blinking should therefore be exclusively reserved for urgent information presentation and for not more than two pieces of information at a time (ISO/TR 16352:2005 (E)).

In cases when rather complex information needs to be conveyed, textual or symbolic information presentation appears to be suitable. However, complex visual in-car presentation particularly affects gaze behavior and driving performance (Jamson & Merat, 2005; Recarte & Nunes, 2000). Besides, in time-critical situations reading duration for written messages, on the one hand, and interpretation issues of pictorial information, on the other hand, need to be born in mind (Campbell et al., 2007). Besides, any kind of visual information presentation needs to consider the visual sampling process in general and drivers' gaze distribution while driving in specific situations (e.g., ISO/TR 16352:2005 (E); Dingus & Hulse, 1993; Land & Lee, 1994; Tijerina, Kiger, et al., 1996). This implicates, on the one hand, that information has to be

presented with sufficient presentation duration, and, on the other hand, that is presented close to the forward line of sight, in order to allow safe glance distribution between the forward and the in-vehicle view (see, e.g., Dingus et al., 1996; Seppelt & Wickens, 2003). Moreover, visual distraction from the road should generally be kept at a minimum, since lane keeping, headway control, and reactions to unexpected events largely depend upon a driver being visually aware of the current environment.

Modern head-up displays present information projected to the lower front windshield. This can reduce space-based distraction, and facilitate gaze shifts as well as respective lens focusing and lead to faster reaction times (Pierowicz, Jocoy, Lloyd, Bittner, & Pirson, 2000; Srinivasan & Jovanis, 1997). However, object-based attentional engagement must still be considered, since this might even lead to cognitive capture and inattentional blindness: A driver might miss (unexpected) critical events on the road while concentrating on the information presented in a head-up display – even though it is displayed close to the critical object from the driver's line of sight (Logan, 1996; May & Wickens, 1995; Most et al., 2001; Wickens & Long, 1995). Visual clutter might further decrease performance (Martin-Emerson & Wickens, 1997; Pierowicz et al., 2000; Yeh, Merlo, Wickens, & Brandenburg, 2003). Moreover, these displays might elicit a deceptive sense of safety and hence lead to behavioral effects like fewer control shifts to the current road situation. To sum up, head-up displays are a valuable approach, but cannot serve as a universal remedy for all types of urgent warnings situations.

In summary, visual information processing is already loaded by the primary driving task and, moreover, by numerous in-vehicle messages that are not time-critical. In addition, visual presentation interferes with unnecessary but common visually distracting activities (e.g., watching the countryside, looking at passengers, reaching for objects) in which warnings might be especially relevant. Last but not least, eye movements and the observation of visual displays is dependent on tiring (Schleicher, Galley, Briest, & Galley, 2008). Accordingly, it is rather inappropriate to rely solely on visual presentation – above all in cases of urgent, safety-critical warnings on road incidents with immediate perception, processing, and adequate responses being

required. Correspondingly, many authors have suggested the auditory modality as a suitable complement for inconspicuous and infrequent visual information or warnings (e.g., Campbell et al., 2007; Dingus & Hulse, 1993; ISO/TR 16352:2005(E). Thus, the following section will focus on advantages and disadvantages of auditory information presentation with particular emphasis on time-critical warnings.

1.2.2 Auditory warnings

Generally speaking, responses to simple auditory stimuli are faster than responses to simple visual stimuli (e.g., Baxter, 1942), and this holds true above all under loading conditions by another visual task (Johnston, Mayyasi, & Heard, 1971). Besides, auditory stimuli are omnidirectional and inherently more alerting than visual stimuli (Posner, Nissen, & Klein, 1976), allowing for information transfer even when operators do not expect a warning or look elsewhere (Noyes, Hellier, & Edworthy, 2006; Stanton & Edworthy, 1999). Due to the fact that listening cannot be switched off, the auditory modality has even been labeled as a "natural" warning approach (Stanton & Edworthy, 1999). Using the auditory channel allows for in-car alerts and warnings that are both "hands free" and "eyes free". That is, spatial orientation or direct contact is not essential for the perception of auditory signals (Graham, 1999). Furthermore, auditory signals can be perceived simultaneously with other spatial visual or haptic information. Hence, auditory signals – as compared to visual signals – might cause only minor interference with the ongoing driving performance, which is basically a visual feedback loop with motor control (Dingus et al., 1996; Donges, 1978; McRuer, Allen, Weir, & Klein, 1977; Seppelt & Wickens, 2003).

Nowadays, for in-car warnings the auditory channel is rather confined to alerting the driver instead of transferring meaningful information (Nees & Walker, 2011): Simple tonal signals accompany important visual information in order to ensure that the driver immediately takes notice (e.g., a sound accompanies low fuel indicator or a seatbelt indicator). These simple tonal signals can alert a driver, but they only have limited potential to inform about the concrete situation (as was already mentioned earlier). Based on these

considerations, usage of simple tonal alerts cannot accomplish the forthcoming transfer of numerous warning contents in modern cars. Instead, semantically-enriched warnings like natural sounds or speech are a promising approach for information transfer from technical systems to the driver (Gray, 2011; Nees & Walker, 2011). Up to date, auditory semantic warnings have been investigated both for driving scenarios and for other applied domains. However, they have not yet been introduced into serial production cars (for a review of warnings in other domains see, e.g., Dingus et al., 1996). Later in Chapter 2 we will provide further details on auditory warnings that contain semantics.

Despite numerous advantages, auditory information presentation also has some constraints and limitations. For example, auditory information presentation via tones or speech is not recommended for very long and complex messages (e.g., Campbell, Carney, & Kantowitz, 1998; ISO/TR 16352:2005 (E)). It should be noted that especially under noisy conditions, aspects like acoustic quality, distinctiveness, or duration of a sound are crucial for effective use (Hellier & Edworthy, 2000; Stanton & Edworthy, 1999). These sound parameters, as well as the occurrence frequency of auditory presentations, influence the perceived annoyance (Marshall, Lee, & Austria, 2007). Another disadvantage of auditory information presentation is its transient nature. Whenever a recipient misses an auditory warning or does not comprehend a speech warning, information is lost or distracting requests become necessary. As already complementary mentioned in the section about visual warnings, a possible approach for mitigation of this issue is to provide complementary information in another modality as a fallback. For example, redundant textual or pictorial presentation of the relevant information could be used, if proper understanding of auditory signals is uncertain (Dingus & Hulse, 1993; Dingus et al., 1996). Alternatively, additional tactile stimulation could be used to mitigate masking of tonal signals in noisy environments or during an ongoing conversation (Ho, Reed, & Spence, 2007; Mohebbi, Gray, & Tan, 2009). In line with this approach, we briefly turn towards potentials and drawbacks of tactile warnings in the next section.

1.2.3 Tactile warnings

Proprioceptive tactile warnings can be communicated via any body part of a driver that is currently in physical contact with the car. Mostly, tactile warnings are presented via vibrations or counterforces of the control elements (steering wheel, pedals), the seat, or the seatbelt (see Spence & Ho, 2008 for a review). Except for immediate intervention into controls, this warning modality offers similar advantages as auditory warnings: Tactile warnings are also alerting and independent of the current attentional orientation. For example, in forward collision, lane departure, and stop requirement warning scenarios, it has been shown that haptic warnings positively influence driving performance (Campbell et al., 2007; Ho, Reed, & Spence, 2006; LeBlanc et al., 2006; Pierowicz et al., 2000; Tijerina, Jackson, Pomerleau, Romano, & Petersen, 1996). Tactile warnings allow unobtrusive presentation and they are well accepted (Brown, 2005; Hoffman, Lee, & Hayes, 2003; Van Erp & Van Veen, 2001). However, information presentation is limited to, either general alerts, or to basic sensorial information imitating physical information from outside the car (e.g., vibration from driving on a road shoulder). Similar to simple tonal signals, arbitrary combinations of haptic information and semantic content have to be learned and remembered first. In case of a critical warning situation, the meaning of a simple haptic warning needs then to be retrieved in a first step. For complex (semantic) information that needs to be transmitted rapidly tactile-only warnings are hence not sufficient. In such cases, the presentation of complementary information via another modality is inevitable.

1.2.4 Combination of modalities

As previously mentioned, modalities have been combined in numerous warning scenarios. Redundant or complementary message presentation, as well as multimodal feedback, can help to overcome disadvantages of single-modality approaches (Campbell et al., 2007; Edworthy, Stanton, & Hellier, 1995; Ho, Reed, & Spence, 2007; ISO/TR 16352:2005 (E); Pierowicz et al., 2000). For example, transient tonal warnings can be disambiguated by continuous visual presentation, or tonal support is advisable when visual warnings are likely to be overlooked (Dingus et al., 1997). Auditory signals

used in conjunction with visual signals have the potential to decrease reaction time and to improve user acceptance in comparison with purely visual signals (Belz, Robinson, & Casali, 1999; Srinivasan, Yang, Jovanis, Kitamura, & Anwar, 1994; Voss & Bouis, 1979). Referring to ongoing visual-manual tasks in their recent meta-analyses of task performance, Lu et al. (2013) pointed out on a far more general level that the presentation of an additional interrupting task via two differing modalities revealed a speed-accuracy trade-off, with more accurate, but eventually slower response times for redundant as compared to unimodal presentation. When combining multiple modalities, information should be presented in a congruent fashion, and simultaneous presentation of contradictory or competing information pieces via different modalities should be avoided (ISO/TR 16352:2005(E); Reeves et al., 2004). Thereby, misinterpretation or miss of information can be prevented. In some cases, when a graded sequence of warnings from mild to severe is applied, a modality switch might also be beneficial (Horowitz & Dingus, 1992; Lee et al., 2004). If a warning is not imminent, the system can start with an unobtrusive warning in one modality, and, if urgency increases and it becomes necessary, it could subsequently add further codes or modalities. Thus, startling effects of alerting warnings (Fagerlönn, 2010) might be mitigated (Fagerlönn, Lindberg, & Sirkka, 2012). However, this stepwise strategy takes time and is therefore not suitable for imminent and extremely time-critical warnings.

To conclude, the selection of one suitable warning modality for a system or use case is not necessarily the end of the story. Depending on the concrete use case, it is a good idea to consider whether additional modalities can enhance a selected warning approach by redundant or complementary information. Nevertheless, even if a combination of modalities is chosen for the final warning design, this needs to be based upon the most effective components from each modality. Accordingly, in the next section we progress with the selection of the most preferable modality and code for time-critical information transfer.

1.3 Comparison of prevalent warning approaches regarding time-critical information transfer

Our short review on the major three modalities for warnings (and of combinations thereof) in the former sections lays the foundation for an interim decision for the initial question: Which modality should be considered for the presentation of a multitude of warnings in both highly time-critical and, at the same time, rather rare situations? Table 1 provides a summary on advantages and disadvantages of visual warnings, tonal auditory warnings, and haptic warnings. Campbell et al. (2007) recommend the visual modality for the presentation of complex, long, and numerous messages, unless the situation is time-critical. Conversely, tonal and tactile warnings are recommended for the presentation of few, time-critical warnings with low-complexity.

Since visual warnings presented on the dashboard are not recommended for the transfer of time-critical information, and benefits and costs of head-up displays are still disputed, the potential of tactile and auditory warnings should be considered more closely. Due to the fact that both tactile and auditory warnings are omnidirectional and can attract a driver's attention, they are good candidates for the presentation of urgent warnings during an ongoing visual task (Lu et al., 2013; Scott & Gray, 2008). However, tactile warnings as well as simple tonal auditory warnings do not offer the possibility to transfer the more complex informational messages that are enabled by new technologies. We want to point out that auditory semantic warnings, and in particular speech warnings, have the potential to overcome these relevant disadvantages of simple tonal alerts. Even though the majority of the literature on automotive warnings and also serial production implementations give rise to a different picture, the auditory modality naturally offers possibilities beyond the presentation of simple tones: For example, environmental sounds or speech have the potential to convey complex information. Accordingly, the auditory modality turns out to be a promising candidate for time-critical information transfer from car to driver and it is definitely worth a closer inspection. The subsequent chapter provides an overview on spoken words and further auditory semantics in the automotive context, and, moreover, it evaluates which of these should be further pursued with regard to our applied research question.

Table 1. Heuristics for the selection of a warning modality. Relevant criteria for urgent transfer of informational messages are highlighted by means of italics. Table derived from Campbell et al. (2007).

Heuristics to use when selecting a warning modality	
Visual warnings	
Recommended for	Not recommended for
• Unobtrusive warnings / non-urgent information / self-paced presentation • *Complex, long, and numerous messages* • Discrete and continuous information • Spatial information • Temporally and spatially free access	• *Conveying time-critical information /forced-paced presentation* • Poor illumination conditions • Configuration with unrestricted driver viewing angle and position
(Tonal) auditory warnings	
Recommended for	Not recommended for
• *Getting the attention of a driver who is distracted or looking away from a visual warning* • *Time-critical information* • Low-complexity, *high-priority* messages • *Few and short messages* • Discrete, sequential, or spatially-localized information	• Frequent warning messages because they are obtrusive and can be annoying • Continuous information • *High complexity/informational messages* • High-noise environments that can mask auditory warning signals
Tactile warnings	
Recommended for	Not recommended for
• *Obtrusive, omnidirectional attention-getting* • Providing warning information if other modalities are overloaded • Providing simple information if it is given in the appropriate context and if it provides direct intervention in the manual control process (e.g., steering torque naturally advises a driver against further steering against the force).	• *Providing complex or potentially ambiguous information* • Systems that provide limited exposure to warnings because drivers are likely to require some learning to distinguish them from natural driving sensations (e.g. rumble strips)

2 Semantically-enriched in-car warnings: Spoken words and further auditory semantics

When considering auditory semantics for the transfer of various in-car information and warnings, there are several approaches available. As a first possibility for the transfer of semantic information via the auditory modality, we present the usage of arbitrary synthetic tones (*auditory earcons*). Users have to learn meanings of several sounds in advance and can accordingly derive semantic information from the tones and from structured combinations thereof. Secondly, meaningful everyday sounds (*auditory icons*) offer a more direct possibility to transfer semantic contents without presenting lexical information. For example, a car horn sound informs the driver that another car is present, and most probably, it further comprises from which direction the car should be expected. Last but not least, speech is, of course, the most conventional way to convey semantic information to another person. Information transfer can be accomplished by presenting either complete sentences or carefully selected keywords that match the prevailing situation. Since benefits and drawbacks of auditory information presentation in general have already been discussed earlier, the following paragraphs reveal deeper insights about challenges of especially semantic auditory warnings concepts. Differences between the three auditory semantics approaches are highlighted, and results from literature are summarized, in order to support the selection of a promising approach for subsequent work on urgent in-car warnings in diverse situations.

2.1 Auditory earcons

A first possibility to transfer semantic information via auditory warnings is the use of *auditory earcons*. This auditory signal category denotes abstract, synthetic sounds that are arbitrarily associated with a specific piece of information. *Auditory earcons* offer a variety of different auditory signals, each of which is related to a specific meaning. Associations typically base upon (systematic) variations of physical characteristics like for example fundamental frequency, pitch, or the rate of presentation (Blattner, Sumikawa, & Greenberg,

1989; Brewster, Wright, & Edwards, 1992; Haas & Edworthy, 1999). Before *auditory earcons* can be used, recipients have to learn the link between an auditory earcon and its associated meaning. Due to the requirement of a relatively long learning phase, auditory earcons are not recommended for in-car warnings in time-critical situations (Dingler, Lindsay, & Walker, 2008; Ho & Spence, 2008). Even though professional pilots might complete trainings on warning signals, this would not at all be feasible for nonprofessional drivers. Moreover, even if drivers would be willing to learn the relevant meanings, they would most probably forget the meaning of rare *auditory earcon* warnings in the meantime, or they would, at least, need a considerable amount of time for retrieval (McKeown & Isherwood, 2007; Vilimek & Hempel, 2005). Hence, *auditory earcons* should be discarded from the list of valuable approaches in serial-production cars.

2.2 Auditory icons

Auditory icons are sounds, which naturally occur in the real world, containing information analogous to everyday events. These sounds are representatives for their respective objects (Gaver, 1989; for a discussion of the term *"auditory icon"* see also Petocz, Keller, & Stevens, 2008), and there exist three categories of auditory mappings (Gaver, 1986): *Symbolic,* for well-learned arbitrary combinations of sounds and events (e.g. the siren of an ambulance to signalize a hospital), *metaphorical,* using similarities between event and sound (e.g. rising pitch for filling a data carrier), and *nomic,* which is directly based upon physical causations (e.g. a skidding tire indicating a braking car). Other authors (Petocz et al., 2008; Stevens, Brennan, Petocz, & Howell, 2009) have recently pointed out that, contrary to what the term *auditory icon* might suggest, these sounds are rather indexical than iconic. Accordingly, prior learning of observers needs to be considered, thus forming a triadic relation between referent, signal, and observer. This theoretical consideration of former learning also applies to any other type of auditory warning.

While *auditory icons* have initially been designed for desktop applications (Buxton, Gaver, & Bly, 1994; Gaver, 1989), researchers have already transferred this concept to the automotive domain (e.g., Fricke, 2009; Graham, 1999; McKeown & Isherwood, 2007; Stevens, Brennan, & Parker,

2004). Since most of these studies compare *auditory icons* with other relevant auditory approaches, details and relevant results will be presented at a later point in the overview Section 2.4. Particularly for a successful application of *auditory icons* in warning situations, ecological frequency might be crucial as well as the mapping of sound and situation in the eye of the beholder. Unfortunately, the use of *auditory icons* is limited, as soon as relevant information lacks a unique, corresponding sound (e.g., abstract concepts). This is a relevant drawback to keep in mind when turning to speech as the third category of semantic auditory warnings.

2.3 Speech and *spearcons*

From an early age onwards, speech processing is well practiced and rather automatic for everybody in her first language. Hence, no learning is required for users to understand speech utterances containing familiar vocabulary and content. Especially for the presentation of rare, and at the same time highly diverse warnings, this natural understanding of spoken words offers great opportunities. In imminent warning situations, speech seems to be an adequate means to assist the driver, and to inform about the current situation. For decades, speech warnings have already been investigated and used in other domains like for example aircrafts or nuclear power plants (e.g., Dingus et al., 1996; Simpson & Williams, 1980). Accordingly, the transfer of some already existing knowledge appears to be worthwhile. In a general review on speech warnings in several domains, Noyes, Hellier, and Edworthy (2006, p. 565) conclude that speech warnings are particularly appropriate in "high workload and information overload situations where people are working hard and intensely at a number of tasks, [under] poor viewing conditions where visual information is not easy to assimilate, [in] situations where there is a need to convey information rapidly, … [in] safety-critical situations where there is a requirement for people to respond quickly and take appropriate action", and – besides other context conditions – in "situations in which a warning is infrequent but high priority, [since] in these situations a listener would be unlikely to remember the meaning of an infrequently heard non-speech audible warning". The circumstances, under which speech warnings are preferable, seem to almost perfectly match on-road driving conditions. Nevertheless, there

still appears to be a lack of profound investigation considering speech warnings in the driving domain. This deficiency might be due to some disadvantages of speech warnings.

In comparison to simple auditory sounds, a first obvious drawback of spoken words is that each speech warning takes time. Thus, a warning might not be understood until is has been (almost) presented completely (Simpson & Williams, 1980). Speech inherent duration until a recipient comprehends the message is, of course, extremely critical in a warning context. In this regard, increased word rates or elimination of redundant words can serve as approaches to reduce the overall length of speech presentation. In natural situations speakers use increased speech rates and raised pitch to signal emphasis (Simpson & Marchionda-Frost, 1984). For example, instead of using complete sentences in a critical situation a co-driver might quickly yell "red light ahead" in order to warn and assist the driver.

An average word rate for reading out printed (English) text is about 120 words per minute, whereas silent print reading is faster at about 250 words per minute. For visually impaired participants Aldrich and Parkin (1989) and Foulke and Sticht (1969) confirmed listening rates for auditory speech up to 275 words per minute without loss of intelligibility. These rates were determined for male voices, which comprise a naturally lower pitch than female voices. It is not surprising that understandability is further improved, if the original pitch is retained for the compressed speech (Aldrich and Parkin, 1989). In any case, severe truncation of spoken words by acceleration beyond 300 words per minute leads to a decline in understanding (Mangold, 1982). In the human factors context Simpson and Marchionda-Frost (1984) investigated synthesized speech warnings for helicopter pilots varying speech rates and pitch of messages. Response times were faster for higher speech rates (178 wpm) than for lower speech rates (123 wpm), whereas medium speech rates (156 wpm) were rated most alerting, informative, and valuable. Dingus, Hulse and Jahns (1996) recommend a similar word rate of 150 words per minute for advanced traveller information systems. Campbell, Richard, Brown and McCallum (2007) recommended the same rate for cautionary warnings, whereas a word rate even up to 200 words per minute is recommended for imminent collision warnings. However, indicating compression levels using

"words per minute" seems to be a rather imprecise specification that is – if at all – only suitable for longer utterances containing several words, thus compensating for differences in the number of syllables per word. Unfortunately, this measurement unit does not clearly indicate up to which percentage of (an arbitrary) original speech duration a single word has been compressed.

Spearcons is a label introduced for spoken items that are speeded to about 50 % of their original duration. This type of speeded speech has especially been created for interaction purposes with secondary tasks, when users need to quickly navigate in a system's menu structure (Dingler, Lindsay, & Walker, 2008; Jeon, Davison, Nees, Wilson, & Walker, 2009; Walker, Nance, & Lindsay, 2006). A study by Walker, Nance, and Lindsay (2006), in which participants had to find a target within lists of items, revealed that *spearcons* and also normal speech led to faster and more accurate target localization than *auditory icons* or *auditory earcons*. Even though warnings have not yet been investigated, the *spearcon* literature provides another hint that speeded speech might be useful in time-critical situations: Speeded speech has the potential to speed reaction times without drawbacks in accuracy.

In the context of understandability of speech in warning situations, also the choice of words and sentences based on comprehensibility is crucial, since users might not be prepared to listen. Familiar words and longer words containing more syllables lead to better understanding. Besides, natural speech leads to better understanding than synthesized speech (Hirsh, Reynolds, & Joseph, 1954; Van Coile et al., 1997). It is probably valuable to carry over these general conditions to time-compressed speech.

When considering common speech output in the automotive domain, navigation systems spring to mind, since nowadays they are already well-established applications. When using navigation systems, it can be noticed, on the one hand, that speech output is convenient and keeps visual distraction to a minimum (Labiale, 1990). On the other hand, however, too much or redundant speech output tends to be annoying. Accordingly, speech cannot be a panacea for any kind of information presentation, but should be reserved for rather important system output in time-critical situations. Besides, all kinds of urgent auditory warnings potentially lead to acceptance issues, since they are

perceived as relatively unpleasant. There exists a considerably high negative correlation between perceived pleasantness and perceived urgency of auditory presentations (Fagerlönn, 2007; McKeown & Isherwood, 2007). Set against this general backdrop, in a simple computer task McKeown and Isherwood (2007) found speech messages producing medium urgency ratings together with the highest pleasantness ratings as compared to *auditory icons*, abstract sounds, or nonspecific environmental sounds. This constitutes a promising result for the usage of spoken words in urgent situations, since speech transferring high urgency appears to implicate less subjective unpleasantness than urgent tonal warnings.

For auditory warnings in general, there exists quite some knowledge on spatial presentation and respective crossmodal influences on visual perception (McDonald, Teder-Salejarvi, & Hillyard, 2000; Wang et al., 2007). Besides, even the combination of spatial signal presentation and spatial semantics of speech warnings has been successfully investigated (Ho & Spence, 2005). However, beyond those direction specific auditory warnings, considerably less research has dealt with cause-specific auditory spoken warnings. One relevant issue in this regard is that speech has so far rather been investigated as one out of many output alternatives for a single system presenting a single type of information (see, e.g., Baldwin, 2007; Dingler et al., 2008; Lee et al., 2007; McKeown, Isherwood, & Conway, 2010; Tan & Lerner, 1995; Walker et al., 2006). In these cases, speech mostly leads to fast reaction times, but provides unnecessary, dispensable information for the experimental task. Most importantly, we assume that speech did not perform to its full potential, since in these studies only one type of information needed to be encoded. Accordingly, under these circumstances a simple tone would have already been sufficient. Only a few studies comprised a combination of different information contents, traffic situations, or hazards, with one out of multiple contents potentially presented. Under these circumstances, subjects needed to decode the presented information to improve task performance or to successfully fulfill their task at all. Given this background, some authors investigated *auditory icons* as semantically-enriched stimuli without a comparison to speech warnings (Belz et al., 1999; Fricke, 2009), some authors investigated speech warnings as compared to (textual) visual presentation (Lee et al., 1999; Seppelt

& Wickens, 2003), and some authors even contrasted speech warnings with other auditory semantics (McKeown & Isherwood, 2007; Vilimek & Hempel, 2005). Based on this literature, we assume that potential benefits of semantically enriched warning approaches will particularly become observable for task-relevant presentation. In the following overview section, we will, inter alia, provide more detailed information on this issue.

2.4 Comparison of auditory semantics approaches

At a first glance, each of the three approaches, *auditory earcons*, *auditory icons*, and speech warnings bears potential to support drivers' assessment of ambiguous situations, since each of them includes auditory semantics. However, in earlier studies researchers have found interesting differences between the approaches with regard to reaction times and accuracy. In the following, we will review the results from four different studies. Whereas the first two studies included desktop tasks only, and hence were less related with actual driving performance, the latter two were conducted in a driving simulator.

Vilimek and Hempel (2005) assessed to what extent verbal messages (i.e., spoken keywords or longer speech), *auditory earcons*, and *auditory icons* impact short-term memory performance. The dual-task setting was designed according to real-life situations, in which a user has to actively keep relevant situation information in memory, while at the same time receiving auditory system messages (and, besides, selecting a response to them). The best serial recall performance was achieved when the additional messages were presented via *auditory icons* or spoken keywords. Most interestingly, choice reaction times for *auditory icons* and speech (long speech or keywords only) were significantly faster than for *auditory earcons*.

McKeown and Isherwood (2007) used a simple desktop task, in which participants heard a speech message (3-4 words each) or a sound (abstract sound, *auditory icon*, or unrelated environmental sound), and had to select the respective picture with a computer mouse. Even though the setup did not resemble driving, and even though (only) the auditory presentation indicated the participant's task, it is an interesting finding that, again, speech messages and *auditory icons* provided highest accuracy rates along with shortest reaction

times, as compared with arbitrary sounds (abstract sounds or unrelated environmental sounds). Taken together, the findings from the latter two studies implicate that the usage of *auditory earcons* is not advisable for time-critical situations. This is probably related with a non-automatic retrieval process that has already been mentioned earlier: If an *auditory earcon* is presented, tones need to be retranslated into semantic meanings. Due to memory decay of learned associations over time, this additional retrieval process might have even more detrimental effects when *auditory earcons* are applied for rare events.

In a driving simulator experiment, Graham (1999) investigated (solely) two types of urgency situation, in order to compare three types of auditory warnings: Two *auditory icons*, one speech, and one simple tonal warning were applied. He recorded fastest brake reaction times for the two *auditory icons*: a "tyre skid" and a "car horn". For speech, reaction time largely depended on the situation: If the spoken keyword "ahead" matched the situation, reaction time was also short, otherwise, it was rather long. Simple tonal warnings led to intermediate reaction times in both situations. Moreover, Graham found an increased number of inappropriate responses for *auditory icons* as compared to the simplified speech warning and the tonal warning. This was largely due to false positive reactions in uncritical situations. This study indicates that a match between the prevailing situation and the specific semantic information conveyed by speech is crucial. Otherwise, no positive effect can be expected. Beyond, this dependency of the effect implies that a spoken keyword is processed immediately and that this influences a driver's response accordingly. Rather low false positive reactions for keywords point into the same direction. *Auditory icons* seem to speed reaction times as well, but they appear to be less specific, which might also lead to higher false positive rates. Another issue with *auditory icons* is their restriction to situations and events, which naturally include a specific, un-ambiguous, and well-known sound. Accordingly, for example abstract information cannot be conveyed via *auditory icons*, whereas speech has the potential to cover many more diverse situations and information contents for a wider range of applications.

Unfortunately, there exists a backlog in literature for the comparison of different warning contents that are presented via different modalities, including

speech. As opposed to unspecific alerts, in this case participants would need to decode the presented information to fulfill their task successfully. This constitutes a situation, in which semantic information might tap its full potential. Cao, Mahr, Castronovo, Theune, Stahl, and Müller (2010) were the first to compare pure speech warnings, visual warnings enhanced by a blinking element, visual warnings combined with a tone, and visual warnings combined with speech in a driving simulation study on diverse local danger alerts. The four warnings could either be presented in a stand-alone version, or be additionally preceded by a short, spoken action suggestion ("change lane", "brake"). The short auditory action suggestions elicited very fast responses and high preference ratings, indicating that compact spoken information is a promising approach under certain conditions. However, when comparing the (subsequent) four combinations for detailed warning content presentation, a different result was revealed. The relatively complex and long speech warnings led to more unsafe behaviors and longer reaction times than all other presentations, unless they were preceded by a short reliable action suggestion, or unless they were accompanied by graphical display information. To conclude, in this experiment spoken keywords containing action suggestions improved driving performance, whereas long speech utterances containing several pieces of information led to slower responses, and should therefore not be employed in urgent situations. The authors concluded that additional visual support could mitigate detrimental effects of overly long speech information. However, it needs to be kept in mind that presenting redundant information via two modalities does not automatically ensure "the best of both worlds" (Wickens & Gosney, 2003, p. 1591; see also, Lu et al., 2013).

Overall, the potential of brief in-vehicle speech warnings in the context of multiple, diverse systems and contents is still being extremely underestimated. Therefore, this thesis suggests an elaborate reconsideration of speech warnings for urgent situations. In this regard, it is beneficial to initially clarify the more general question, whether spoken words have the potential to immediately influence visual scene assessment. If so, this would clearly support the concrete recommendations on speech warnings. Unfortunately, applied research only provides insufficient indications on this issue, but, conversely, we assume that the consideration of basic research will help to

shed some light on crossmodal effects from auditory semantics on visual perception and on the nature of underlying crossmodal processes. Accordingly, in the next chapter we turn towards these cognitive processes and, besides, we highlight the additional value that basic cognitive experiments can provide with regard to speech warnings.

3 Crossmodal influences of spoken words on processing of visual information

As already introduced in Chapter 1, continuous sampling of visual information is crucial for ongoing driving performance. This holds especially true for situations when safety-critical events need to be detected. When considering speech warnings as an auditory alternative for information transfer in time-critical hazard situations, a more general question should be addressed in a first step: Are there any immediate crossmodal influences of spoken words on the processing of visual information? If so, knowledge of respective cognitive processes and of potential preconditions or moderators would greatly facilitate deriving meaningful and principle-based recommendations for in-car warning scenarios. Accordingly, we start this chapter by underpinning the strong points of basic psychological experiments regarding our applied research question. Subsequently, we review cognitive psychological literature on crossmodal effects of spoken words on the performance in visual tasks. Taking automotive speech warning scenarios into account, we confined our literature survey to studies comprising a semantical match between verbal and visual stimuli. Besides, we focused on denotations of objects or object features, whereas we excluded, for example, the broad field of verbal spatial information[5].

3.1 Why would recommendations for speech warnings benefit from basic cognitive experiments?

As already outlined earlier, we assume that brief in-vehicle speech warnings offer great potential in the context of multiple, diverse systems and contents, and we suggest an elaborate reconsideration of speech warnings for urgent situations. At this point, one might expect a series of driving simulator experiments comprising several hazard situations along with the presentation

[5] Even though informing a driver about a specific situation might indeed include spatial information, auditory spatial directions do not in themselves comprise further information about the type of hazard. Hence, merely considering spatial directions would contradict our approach to convey a variety of relevant details according to a specific critical situation. Nevertheless, in a second step it might be valuable to complement auditory semantics with spatial information.

of (variants of) concrete speech warnings. Importantly, we consider it essential to clarify the underlying processes of effects prior to such ecologically more valid, but rather uncontrolled experiments. For example, in a time-critical context with drivers not expecting warnings, it is highly relevant whether crossmodal effects of speech warnings would occur immediately and effortless, or whether slower and more controlled processes would cause effects. Unfortunately, in applied scenarios exact starting conditions and timing of events are hardly controllable, and besides, from the surrounding traffic conditions each warning would naturally be considered as meaningful and relevant. These factors complicate both the determination of the speed of auditory-visual interactions and the separation of automatic and controlled processing. In contrast, interference paradigms from basic cognitive psychology especially support the examination of such processes and frame conditions.

Stroop and flanker tasks are traditional cognitive psychological interference paradigms (Eriksen & Eriksen, 1974; Eriksen, 1995; MacLeod, 1991; Stroop, 1935).[6] Formulated in an abstract manner, in both of these paradigms, participants have to respond upon a relevant target stimulus or a relevant feature of a target stimulus. Additionally, one or several other irrelevant distractor stimuli (or features within the target stimulus) are presented that could either match the response upon the target stimulus, or conflict with the correct reaction (or be neutral). Typically, responses are faster if distractors imply congruent responses than for those that imply conflicting responses – even if targets and distractors are completely independent (uncorrelated). Since in this case attending to irrelevant, non-informative distractors would not be worthwhile, participants would probably rather try to ignore them than attend to them. Thereby, these paradigms enable the assessment of automatic response activation by distractors. In a congruent case the additional (and as the case may be also earlier) activation of the correct response leads to faster responses. Contrarily, in an incongruent case the automatically activated incorrect response interferes with the correct response

[6] Please note that the effects in the Eriksen flanker task could again be considered as a special case of response priming effects, with the flankers (= distractors) corresponding to primes that are presented simultaneously with the target (SOA = 0 ms) at a nearby location.

and needs to be controlled. This inhibition of an activated incorrect response results in slower reaction times upon the target, and, potentially, also in higher error rates. Moreover, in these interference paradigms exactly controlling the stimulus onset asynchrony (SOA) of target and distractor (or relevant and irrelevant dimension of a single stimulus) allows deriving information on how fast the underlying processes occur.

In a nutshell, even though, at a first glance, invalid and uninformative presentation of spoken words during visual tasks seems paradox for our applied scenarios, exactly such circumstances allow for gaining more insights about the automaticity of crossmodal auditory-visual processes. In this context, a brief excursus on automatic versus non-automatic processes will support further clarification of our point.

> **Excursus: Automatic versus non-automatic processes**
> In the context of constructive appraisal, Moors (2010, p. 141) recently defined automaticity of processes comprising "a number of individual features such as uncontrolled, unintentional, unconscious, efficient, and fast" (for similar approaches see also Bargh, 1992, 1994; Moors & De Houwer, 2006). She claims that the more automatic a process is, the more "it operates under suboptimal conditions, such as minimal time, minimal attentional capacity, no conscious input, and/or no intention to engage in the process". On the contrary, non-automatic processing needs optimal conditions, like "abundant time, abundant attentional capacity, conscious input, and/or the intention to engage in the process". Moreover, Moors delineates a gradual view of automaticity, with these diverse features applying more or less, and their sum contributing to the overall degree of automaticity. Hence, in order to determine whether a given process is rather automatic or not, the individual features should be considered.

In case that crossmodal semantic effects occur even under suboptimal conditions like the ones introduced in the excursus, we would assume it to offer additional value by automatic information transfer. Moreover, if applicable, we would strongly recommend spoken words as a robust alternative for road hazard scenarios. Accordingly, the consideration and execution of basic interference-paradigm experiments can reveal relevant insights for our applied research question – especially with regard to the assessment of automaticity. Therefore, our approach is not meant to be only a simplification of real-world

circumstances, but it rather constitutes an essential prerequisite for meaningful and reliable recommendations. Against this background of automatic processes and the clarification of how findings from basic experiments can support addressing applied research questions, we will, in the following, revisit earlier findings from basic psychological literature regarding the crossmodal influence of verbal information on visual processing.

3.2 Empirical evidence of auditory semantic information affecting performance in visual tasks[7]

For everyday actions like, for example, driving a car, it is essential to integrate information cross-modally. Nevertheless, only in the last decade psychological research has engaged in exploring these multisensory processes: Multisensory interactions and crossmodal processes have gradually received more and more attention, and researchers have set out to test, extend, and complete former unimodal theories and concepts of human attention, perception, memory, and behavior (e.g., Calvert, Spence, & Stein, 2004; Driver & Spence, 1998; Spence, Senkowski, & Röder, 2009).

When considering the ubiquitous co-occurrence of visual and auditory information, crossmodal influences of audition on vision have been found in some research domains: Numerous studies investigated how spatial sounds influence perception of visual objects in space (e.g., Ho & Spence, 2005; Keetels & Vroomen, 2011; Mazza, Turatto, Rossi, & Umiltà, 2007; Spence & Driver, 1997) or how naturalistic sounds influence visual object perception (Chen & Spence, 2010; Iordanescu, Grabowecky, Franconeri, Theeuwes, & Suzuki, 2010; Iordanescu, Guzman-Martinez, Grabowecky, & Suzuki, 2008; Schneider, Engel, & Debener, 2008). However, we still lack detailed knowledge about a highly relevant question, which we have derived in the preceding chapters: Do auditory spoken words have the ability to immediately and crossmodally influence performance in complex visual tasks?

[7] Please note that parts of Section 3.2 have been reported in Mahr, A., & Wentura, D. (2014). Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets. *Attention, Perception, & Psychophysics, 76*, 575-590. Copyright © 2014 by Springer. Adapted with permission. No further reproduction or distribution is permitted without written permission from Springer.

Previous research using semantic auditory context information and visual targets has not yet been able to specify unambiguously the circumstances under which crossmodal influences occur. It is still in dispute whether crossmodal effects occur automatically, whether the underlying processes are facilitative or interfering, or both, and whether effects are solely attributable to response compatibility effects or to encoding facilitation processes. A detailed overview of what we know from the most relevant studies will be presented in the following sections.

3.2.1 Evidence from Stroop literature

A paradigm that has often been used in this field is the Stroop color-naming task. In the color-word Stroop task, participants have to name the font color of written color words (Stroop, 1935). Typically, responses are much slower and more erroneous in naming the font color of incongruent color words (e.g., the written word RED is presented in blue) than in naming the color of a series of Xs in a control condition (i.e., Stroop interference). Furthermore, when font color and color word match, responses are faster than in the control condition (i.e., Stroop facilitation). The performance difference between incongruent and congruent stimuli is known as the Stroop effect.

Early studies exploring crossmodal Stroop effects with auditory distractor words and visual color patches (or other colored stimuli) yielded somewhat contradictory result patterns (Cowan & Barron, 1987; Elliott, Cowan, & Valle-Inclan, 1998; Shimada, 1990). However, besides other limitations these studies differed in their presentation parameters. Roelofs (2005) therefore conducted a comprehensive study about the visual-auditory color-word Stroop asymmetry, varying stimulus onset asynchronies (SOAs) over a wide range (from 300 to -300 milliseconds [ms] in steps of 100 ms).[8] His study showed that non-predictive spoken color words yielded crossmodal Stroop effects in naming color patches. That is, irrelevant spoken color words have the potential to influence responses to visual color stimuli. Consistent with prevalent patterns found with the traditional unimodal Stroop task

[8] Throughout the thesis, a negative SOA indicates post-exposure of the auditory distractor; a positive SOA indicates pre-exposure of the auditory distractor relative to the onset of the visual target stimulus.

(MacLeod, 1991), Roeloefs found mainly interference from incongruent distractors (for all SOAs except the -200 ms condition), but almost no facilitation from congruent distractors (only at an SOA of 300 ms). Though this might be due to usage of a silent control condition (in contrast to neutral auditory distractors; see our discussion below), Roelofs argued that the failure to find a facilitation effect might have been due to the fact that color naming of pure color patches is a very simple task, which might have produced ceiling effects in the (silent) control condition. He recommended using a more difficult task with respect to visual object detection/classification in follow-up research.

3.2.2 Evidence from perceptual-load literature

Indeed, Tellinghuisen and Nowak (2003) employed a more difficult task for the investigation of crossmodal impacts of verbal distractors on visual search. Given the backdrop of Lavie's perceptual-load theory (1995), Tellinghuisen and Nowak presented a visual target stimulus (i.e., letter X or N) among a set of visual task-irrelevant filler stimuli, arranged in a circle. The fillers were either homogeneous and target-dissimilar (five O's; low perceptual load) or heterogeneous and of a similar complexity as the targets (K, H, V, Z, W; high perceptual load). In addition, a distractor (i.e., X or N) was presented, thereby following a task arrangement introduced by Lavie and Cox (1997). However, whereas Lavie and Cox presented a visual distractor next to the circular arrangement, Tellinghuisen and Nowak presented a spoken distractor letter. They found compatibility effects; that is, if the auditory distractor matched the visual target, responses were faster as compared to the nonmatch condition. Two remarkable details emerged in the results: First, Tellinghuisen and Nowak reported compatibility effects that were more pronounced under high perceptual load as compared to low perceptual load. This is in contrast to the results of Lavie and Cox, who predicted that complex visual input (i.e., high perceptual load) would cause more focused processing, thereby leading to early filtering of distractors. Consistent with this, Lavie and Cox found compatibility effects only under low perceptual load. Secondly, by using a neutral-letter condition (inter alia) for comparison, Tellinghuisen and Nowak (2003) separated benefits (i.e., neutral - compatible trials) and costs (i.e., incompatible - neutral trials): For the high-load condition, they found large benefits and

small costs, whereas the low-load condition elicited only small effects and no distinct pattern of benefits and costs.[9]

Similar to Roelofs (2005), Tellinghuisen and Nowak (2003) interpreted their crossmodal effects as response-priming effects. However, Stroop-like designs are characterized by a confound of stimulus-response compatibilities (i.e., the distractor is either compatible or incompatible with the correct target response) and stimulus-stimulus compatibilities (i.e., distractor and target share features on a semantic level; see De Houwer, 2003; see also Wentura & Degner, 2010). Thus, it is possible that crossmodal semantic priming (additionally or exclusively) caused effects: A compatible auditory distractor might facilitate encoding of the respective visual target.[10] That is, benefits might be the consequence of both response priming (i.e., stimulus-response compatibility) and semantic priming (i.e., stimulus-stimulus compatibility). Therefore, benefits might be larger than costs, since costs would merely occur due to response priming. This is precisely what Tellinghuisen and Nowak found. Moreover, the encoding facilitation process might implicitly act as a valuable hint for target detection, especially under high-load conditions and in more complex visual search tasks.

3.2.3 Evidence from semantic-priming literature

In this regard, a study by Chen and Spence (2011) is important; it focused on crossmodal semantic priming via naturalistic sounds – also known as *auditory icons* (e.g., a barking dog) – and spoken words of variable duration. The participants' task was to detect a visual target that was briefly presented in target-present trials and absent in target-absent trials. Congruent sounds led to enhanced object detection as compared with incongruent sounds, but only for a long SOA (346 ms), not for a shorter SOA (173 ms). Spoken words did not cause any facilitation, with one exception (Exp. 4B): In this Experiment a long

[9] We especially refer to the neutral condition with irrelevant letters, as it constitutes the initial neutral condition known from Lavie and Cox (1997).

[10] Please note, that (like Chen & Spence, 2011) we do not use the term „crossmodal semantic priming" equivalent to crossmodal "semantic priming". "Semantic priming" involves semantically (un)related or (un)associated primes and targets (McNamara, 2005), whereas in "crossmodal semantic priming" they are typically semantically identical (refer to the same semantic representation), but physically diverse due to different modalities. Hence, our usage of the term rather denotes crossmodal priming via a shared semantic representation.

SOA (346 ms) was used, and a picture identification task was added after the object detection task. Chen and Spence concluded, firstly, that sufficient time (200–350 ms) is necessary for an auditory stimulus to activate semantic associations (disambiguation duration); secondly, that associations between sounds and pictures are stronger than between spoken words and pictures; and thirdly, that occurrence of crossmodal effects strongly depends on the participants' task goals. In contrast to the third assumption, Lupyan and Ward (2013; Exp. 3) found crossmodal semantic congruency effects on detection sensitivity even without target identification when using circles and squares and a considerably longer SOA (spoken word duration ~ 650 ms plus interstimulus interval (ISI) 450 ms) in a similar experiment, (see also Lupyan & Spivey, 2010; Exp. 2, for another similar experiment with letter targets). Accordingly, so far there exists no consistent view whether target identification is a necessary precondition.

The previously mentioned findings on crossmodal semantic priming with visual objects and their spoken denotations were all derived on the basis of a signal detection approach. However, some additional evidence was revealed in reaction time paradigms. For example, Lupyan and Thompson-Schill (2012; Exps. 3A and 3B) presented two identical pictures on a target screen with either the left or the right one turned upside-down. Classification latencies for the location of the upright picture (left or right) were measured. Spoken object denotations or white noise were presented prior to the target onset with a relatively long ISI of 400 ms.[11] Responses were faster following congruent as compared to incongruent denotations. However, these effects might largely base on strategic effects of participants, since auditory prime and visual target matched in 80 % of the trials that comprised a spoken prime word. Hence, auditory primes became considerably valid and supported participants to prepare for the upcoming pictures (see also Lupyan & Ward, 2013, Exps. 1 and 2 comprising analogical issues).

With regard to this contingency issue, it is worth taking a look at an article by Salverda and Altmann (2011; Exp. 3), which disentangled contingency and congruency. A pair of objects was presented on a screen for a

[11] Due to the differing stimulus durations no consistent SOA was employed.

period of 3,000 ms before the onset of a spoken object denotation. 400 ms after auditory stimulus onset, one of the two pictures moved slightly upwards or downwards, and participants should classify the respective movement via key press (up or down). Any semantics were completely irrelevant for the actual task, since the response was orthogonal to both the spoken denotation and to any evaluation of the semantic content of the respective picture. Nevertheless, when auditory prime and moving target were semantically congruent, responses were faster than when the auditory prime matched the static non-target object. By comparison of congruent and incongruent semantics with neutral denotations of objects that were not included in the current display, the authors additionally identified pronounced interference and minor facilitation. Accordingly, by applying a quite different experimental procedure with ample time for the initial visual investigation of present objects, Salverda and Altmann could identify crossmodal congruency effects implicating attentional capture – even without any contingency or task relevance of spoken words.

3.3 Conclusions from former findings and open issues

Summing up the findings from several different studies introduced in the preceding section, there is converging evidence for considerable effects of spoken words on processing of semantically related visual targets. However, authors assume somewhat discordant preconditions, for example regarding minimum SOA, target identification, or task-relevance of auditory stimuli. Most importantly, findings diverged with regard to the facilitative or detrimental nature of effects. Some findings indicated pronounced facilitation and no or only minor interference (Chen & Spence, 2011; Lupyan & Spivey, 2010; Tellinghuisen & Nowak, 2003), some findings point towards equal facilitation and interference (Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013), and some revealed major or exclusive interference (Roelofs, 2005; Salverda & Altmann, 2011). However, direct comparisons between these studies are complicated, since differences in paradigms, materials, and design might already cause inconsistencies between findings. For example, from the unimodal Stroop literature, it is known that the ratio between facilitation and interference is affected considerably by the choice of the neutral baseline condition (Duncan-Johnson & Kopell, 1981; Kahneman & Chajczyk, 1983;

MacLeod, 1991). Accordingly, it is crucial which baseline is chosen. Presenting silence or white noise might lead to relatively fast responses upon visual targets and hence an underestimation of facilitation as compared to a single unvarying neutral word, or even several neutral words (see also Tellinghuisen & Nowak, 2003, for a comparison of different neutral conditions). In this light, inconsistent findings between former experiments regarding facilitation and interference become more comprehensible on the one hand, and underline the need for further investigation on the other hand.

Moreover, another issue needs close consideration. Above, we have already introduced the issue of response priming (i.e., stimulus-response compatibilities) as opposed to semantic priming (i.e., stimulus-stimulus compatibilities). Whenever responses upon targets are completely independent of those responses that are implicated by primes, response-compatibility effects can be excluded, since congruent and incongruent prime-target pairs only differ regarding their semantic concordance.[12] This allows for a separation of semantic priming effects from response-priming effects. However, both Roelofs (2005) and Tellinghuisen and Nowak (2003) did not separate these accounts, leaving doubt about the underlying process(es) for their results.

Another critical issue that has also not been sufficiently considered in the hitherto studies, is the role of contingency, which resembles the "relatedness proportion" in semantic-priming contexts (Bodner & Masson, 2003; Logan & Zbrodoff, 1979; Logan, 1980; Melara & Algom, 2003; Perea & Rosa, 2002). Contingency can be derived from the experiment-wide or block-wide proportion of experimental trials regarding various conditions. When, for example, a given prime is followed by a (congruent) target stimulus with a probability above chance, this prime reveals contingent information on target identity. If this correlation between prime and target is mentioned in the instructions, or if it is rather obvious, participants most likely (strategically) use this knowledge for generation of expectancies regarding the target, thus

[12] The term "distractor" is used in the Stroop literature as well as in the perceptual-load theory for an additional stimulus that is either compatible or incompatible (or neutral) with regard to the required target response, whereas the term "prime" is not only used in research on semantic priming (where a prime can be semantically related or unrelated to a target), but also in response priming tasks that are structurally equivalent to Stroop tasks. In the following, we will consistently use the term "prime" for the additionally presented stimulus that does not require a response itself.

improving overall task performance. However, even if contingency is not explicitly disclosed, participants might (implicitly) learn this coherence during the experiment and adapt their behavior according to previous episodes and retrospective memory retrieval (Bodner & Masson, 2003). Therefore, primes revealing contingent information potentially affect the interpretability of crossmodal effects. Unfortunately, most authors of literature on crossmodal semantic effects have paid no or only little attention to this issue (Chen & Spence, 2011; Lupyan & Spivey, 2010; Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013, Exps. 1 and 2; Salverda & Altmann, 2011, Exp. 1). Consequently, we especially considered this issue, and, moreover, we even explicitly investigated effects of increased contingency on facilitation and interference in our experiments.

Besides, we target to clarify the issue of processing speed of spoken words in the context of semantically related visual targets. However, examination of the effect by varying SOAs is more complicated for crossmodal than for unimodal presentations, since the latencies needed to access the semantic representations are much shorter for visual stimuli (e.g., Potter, 1975; Thorpe, Fize, & Marlot, 1996) than for auditory (semantically-enriched) stimuli (e.g., Murray & Spierer, 2009). Up to date, Tellinguisen & Nowak (2003) and Roelofs (2005) were the only authors who found a crossmodal effect of verbal utterances at an SOA of zero ms, albeit only with single letters or monosyllabic color names, and, besides, without controlling for response-priming effects. In contrast, Chen and Spence (2011) stated on the basis of their findings with naturalistic sounds that the presentation of auditory semantics must start at least 300 ms before the visual stimulus onset because auditory signals only evolve over time, whereas the meaning of a picture can be accessed rapidly. Likewise, in all further articles on crossmodal semantic effects, considerably longer SOAs than 300 ms were applied (Salverda & Altmann, 2011), mostly even above one second (Lupyan & Spivey, 2010; Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013). However, the minimum SOA to cause crossmodal semantic effects is highly interesting, since it allows drawing conclusions about speed and efficacy, respectively. Crossmodal effects by irrelevant verbal stimuli with extremely short SOAs would point towards fast and rather automatic semantic evaluation and

processing (see also our consideration of automatic processes in Section 3.1). However, this would be quite astonishing since the understanding of speech is assumed to involve higher-level cognitive processes (Eysenck & Keane, 2000; Goldstein, 2002), and therefore implies to be being rather slow. Regarding immediacy of crossmodal effects, the inherent duration of spoken words seems to be a confounding factor. Referring to this, we suggest the usage of time-compressed spoken words for closer investigation of the time factor. Unfortunately, there exist no related findings in literature on crossmodal effects so far, thus leaving space for an initial investigation.

Recalling our applied research question on warnings in diverse, time-critical driving situations the basic findings from cognitive psychological literature reveal quite promising insights: While using diverse approaches, many authors confirmed crossmodal effects of spoken words on visual target detection and classification. However, before more concrete recommendations can be derived, answers to numerous open questions are still required. For example, with regard to the delivery of time-critical information, it is so far unclear whether these crossmodal effects will occur fast enough to immediately affect visual performance. Accordingly, we put emphasis on incremental investigation of crossmodal effects reaching from controlled personal computer setups in order to clarify basic preconditions and processes to the point of driving simulator experiments and the investigation of rather complex driving maneuvers. Moreover, this approach does not only support clarification of a concrete, applied research question (i.e., delivery of diverse, time-critical in-car warnings for drivers), but our basic findings can also be applied to other domains (e.g., power plants, aviation, gaming).

3.4 Hypotheses and overview of experiments

The overall goal of this thesis was to contribute valuable information to the design of a warning type that enhances a driver's situation assessment and response in highly diverse, time-critical, and hazardous traffic situations. From former studies in the automotive domain, we have derived that short speech has the potential to convey diverse, urgent information within a short period of time, thus initially constituting a promising approach. In this regard, however, even cognitive psychological literature has not satisfactorily answered one

important underlying research question so far: Does speech have the potential to crossmodally enhance visual target identification and classification?

Clarification of this basic issue could, in turn, importantly affect the choice of semantic warning contents. Hence, our major hypothesis was that congruent spoken words would directly enhance visual performance as compared to a neutral condition. This relevant issue was repeatedly addressed in each of our eight experiments (1A, 1B, 2, 3A, 3B, 4, 5A, and 5B). Therefore, we initially designed a visual task and consistently applied it to all our experiments, albeit slightly adapting it where necessary. Participants always learned a set of four target items prior to the main phase of the experiment. This should resemble drivers having a general knowledge of various (potentially) critical objects as against irrelevant or uncritical objects in their surroundings. In all experiments, participants should either indicate which of the targets was present on a given target screen (Exps. 1A-1B, 3A-B), whether any of the targets had been present (Exp. 2), or indicate a feature of the present target (Exps. 4-5B). Since we were interested in immediate effects of spoken words on the performance in the visual task, we consistently started auditory presentation only 100 ms before the onset of the visual task. This allows roughly imitating a driving situation with an assistance system providing time-critical information just before a driver visually notices a hazard. Exemplary reasons for such a time lag might be: Driver distraction or fatigue, failure in critical situation appraisal, superiority of technical sensors over human beings regarding information extraction, processing speed, or prioritization. We were especially interested whether congruent spoken words, representing adequate warnings, and incongruent words, representing inappropriate information, would differently affect visual processing. In order to separate benefits and costs, we additionally applied a neutral condition in all experiments. For this purpose, we used merely one (or several) equally alerting (non)word(s) that lacked any semantic overlap with the targets (Exps. 1A-5B). Transferred to the driving context a neutral word would correspond to a single, unspecific word being presented like a “master alarm”. In some experiments, we additionally employed a silence condition (Exps. 2, 4-5B) in order to test whether any kind of speech would already interfere with visual performance. An affirmation of

our first hypothesis regarding crossmodal performance enhancement would lay important foundations for the general usage of speech warnings.

Our second hypothesis was that effects would, at least partly, base on crossmodal stimulus-stimulus compatibilities. Accordingly, strategic listening or stimulus-response compatibilities would not constitute a necessary precondition for the occurrence of effects. For on-road usage of speech warnings, this is a critical issue, since very rare warnings should not require careful listening or compliance based on prior experience. Furthermore, for mere hazard denotations no response compatibility with driving would be given. Therefore, in driving scenarios crossmodal effects due to stimulus-response compatibilities would be less decisive than those based on stimulus-stimulus compatibility. We excluded stimulus-response compatibilities in Experiment 2 by using a signal detection approach, and, moreover, in Experiment 4 by using responses orthogonal to the target identity. In case that crossmodal semantic priming effects were replicated under these conditions, we could conclude that stimulus-stimulus compatibilities on a semantic level have indeed the potential to influence visual processing.

Thirdly, we hypothesized that crossmodal effects would become more pronounced under higher perceptual-load conditions in more difficult visual tasks: A less efficient search process would especially profit from crossmodal congruency. This factor was addressed in Experiments 1A and 1B by the usage of two different perceptual-load levels, since it denotes an important point regarding speech warnings for highly complex traffic situations.

Fourthly, we assumed the duration of spoken words to significantly influence effects: Under time-critical circumstances, time-compressed, but still understandable, words would cause more pronounced effects due to earlier disambiguation and termination. Initially, this hypothesis was inspired by the observation that people commonly tend to increase their word rate in situations requiring urgent information transfer. Four different levels of time compression were applied in our experiments while keeping the pitch constant. Stimuli could be presented in their original duration (Exps. 1A, 1B, 3A, and 3B), or compressed down to 50 % (Exps. 3A-5B), to 30 % (Exps. 1A-3B), or even to 10 % (Exps. 1A, and 1B) of their original duration, with Experiments 1A, 1B, 3A, and 3B particularly addressing the issue of time compression. Findings

regarding this hypothesis would help to decide whether time compression is advisable for speech warnings, and, if applicable, which compression level(s) might be suitable.

With regard to our fifth hypothesis that strategic use of contingent spoken word presentation would increase both facilitation and interference (Exps. 3A-B, 5A-B), we sixthly hypothesized that time-compressed spoken words would moderate this effect: The effect of time compression (see Hypothesis 4) would become especially obvious under contingent presentation of spoken word and visual target. In Experiments 3A, 3B 5A, and 5B, we compared non-contingent presentation (Exps. 3A and 5A) with both a contingency level only slightly above chance (Exp. 3B), and a high contingency level (Exp. 5B). We assumed that in case of contingent presentation observers were able to recognize meaningful time-compressed spoken words more rapidly, thereby mitigating perceptual degradation.

Last but not least, our seventh hypothesis was that crossmodal effects would be robust enough to be replicated even under highly dynamic driving conditions, when performance of complex driving maneuvers is measured instead of simple key-press responses. Only if practical relevance can be confirmed under more realistic conditions, speech warnings for hazardous driving situations should be further pursued.

In the following three chapters (Chapter 4 – Chapter 6) we present eight experiments in order to investigate the seven hypotheses introduced above. With each chapter, we take another step from highly controlled stimuli in computer experiments towards more applied scenarios. This is supposed to reflect our approach to initially gather basic insights followed by incremental transfer of these findings to real-world scenarios.

4 Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets[13]

This chapter comprises three experiments (Exps. 1A, 1B, and 2), which were supposed to reveal basic insights into crossmodal semantic priming effects. In order to especially address our applied research question on warnings in time-critical driving situations, which we have introduced in the preceding chapters, we focused on spoken prime words and visual targets. Moreover, we designed Experiments 1A, 1B, and 2 closely in line with the most relevant literature on this topic, thus allowing for comparisons regarding our key assumptions on: Crossmodal semantic facilitation and interference, differences in perceptual load and word duration affecting effects, and stimulus-stimulus compatibilities on a semantic level. The following paragraphs will revisit these points and explain the respective choice of our experimental design and materials.

As already described in more detail in Chapter 3, a study by Tellinghuisen and Nowak (2003), shows findings of pronounced facilitation and a non-typical moderation of their Stroop effects by perceptual load. Moreover, the results of Chen and Spence (2011; Exp. 1) indicate that congruent auditory stimuli (naturalistic sounds) enhance processing of visual targets, but incongruent fail to impair. In contrast, Lupyan and Ward (2013, Exp. 3) found both interference and facilitation with a silence baseline condition. However, the differences in procedure and materials between the studies make a comparison somewhat difficult. Therefore, we aimed to integrate the aforementioned crossmodal approaches for our Experiments 1A, 1B, and 2. First, following the Stroop tradition, we chose simple color stimuli

[13] Please note that Experiments 1A, 1B, and 2 have been reported in Mahr, A., & Wentura, D. (2014). Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets. *Attention, Perception, & Psychophysics, 76,* 575-590.

and used them as visual targets, visual filler items, and auditory spoken word primes.[14]

Secondly, we manipulated perceptual load, by using a simple and a more difficult visual search task (cf. Lavie & Cox, 1997; Lavie, 1995). In contrast to Tellinghuisen and Nowak, who merely used two single letters as visual targets and auditory primes, we selected four different visual color patches and, as primes, their spoken labels. An extremely short SOA was required to maintain consistency with Lavie and Cox's and Tellinghuisen and Nowak's experiments on perceptual load. Stroop effects being most pronounced at an SOA around zero (Elliott et al., 1998; Roelofs, 2005; Shimada, 1990) further supported this decision. Besides, we found it intriguing, whether crossmodal effects could be replicated even with a considerably shorter SOA than applied by Chen and Spence (2011) or Luypan and Ward (2013), while still presenting complex verbal stimuli. Analogous to both Roelofs (2005) and Tellinghuisen and Nowak, we varied prime and target orthogonally (i.e., identical trial numbers in the cells of the 4 [prime color] × 4 [target color] matrix). Thus, the conditional probability of a target given a prime is the same for all targets, such that the primes did not reveal any predictive information about the subsequent targets (see Melara & Algom, 2003).

Third, because a speech signal unfolds over time, processing of auditory spoken color words might inherently be slowed, and might therefore not lead to effects in a relatively fast visual search task. Remember that Chen & Spence (2011) found effects only with rather long SOAs. However, they used sounds and words that might reveal their meaning comparably late. Accordingly, we wanted to investigate whether color word duration would moderate crossmodal effects (Exps. 1A and 1B). For this purpose, we introduced the factor time compression, by using three different word lengths: The original duration of the spoken color words (i.e., 400 ms), compression up to a level that was fast but still understandable (compression to 30 % of original duration, i.e., 120 ms), and compression up to a small fraction of the original duration (10 % - i.e., 40 ms), which made it hard to even differentiate

[14] As already explained above (p. 35), we decided to use the term "prime" instead of the term "distractor" for the additionally presented stimulus that does not require a response itself.

between the four target color words. Accordingly, the pure perception time for auditory stimuli could be reduced considerably and was brought more in line with perception of visual stimuli (Potter, 1975; Schneider et al., 2008; Thorpe et al., 1996).

We decided to use manual responses (instead of vocal responses which are more typical for Stroop experiments) although this might yield less pronounced Stroop effects (e.g., Brown, Joneleit, Robinson, & Brown, 2002; MacLeod, 1991). The reason for this is, on the one hand, that we considered it critical that participants might start naming the target while auditory prime presentation was not yet completed. On the other hand, manual responses are more closely related with those expected in the driving context. We chose an SOA of 100 ms for all experiments, thus resembling a situation when a warning is presented right before the occurrence of a hazardous situation (see also Chapter 3). The effects of time compression were evaluated in Experiments 1A and 1B. In accordance with the respective results, we then chose to apply merely a medium compression rate for Experiment 2. Note that for this experiment, we also switched to a signal detection theory approach analogous to Chen and Spence (2011), and Lupyan and Ward (2013). This allows for investigation of whether crossmodal stimulus-stimulus compatibilities lead to enhanced perception of visual colors when searching them in a rather complex scene.

4.1 Experiment 1A: Spoken words crossmodally accelerate reaction times for semantically congruent simple color targets: The role of time compression and cognitive load

Experiment 1A was designed to establish crossmodal Stroop effects with auditory primes. We wanted to explore whether effects depend on (a) visual search task difficulty and (b) the presentation duration of spoken word primes. In addition, we were especially interested in analyses of costs (of incongruent primes) and benefits (of congruent primes). In accordance with Tellinghuisen and Nowak (2003), we expected to find pronounced effects for tasks with higher visual perceptional load.

4.1.1 Method

Participants. A total of 36 students (22 females, 14 males) from Saarland University took part in Experiment 1A. Participants received course credit for taking part. All participants had normal or corrected-to-normal vision, and none of them reported any color-blindness or hearing problems.

Design. A 3 (semantic congruency of auditory prime and target color: congruent, incongruent, neutral) × 3 (duration of the auditory prime: 100 % [i.e., 400 ms], 30 % [i.e., 120 ms], 10 % [i.e., 40 ms]) × 2 (perceptual load of the target screen: low, high) design was employed with all factors manipulated within participants. Technically, the congruency factor was realized by a 5 (auditory prime type: red, green, blue, yellow, and neutral) × 4 (visual target type: red, green, blue, yellow) design. The prime and target were uncorrelated (i.e., each combination of prime color and target color was presented equally often) resulting in auditory priming without contingency. Thus, participants had no benefit from listening to the words. Congruency was manipulated on a trial-by-trial basis, whereas compression level and perceptual load were manipulated blockwise, resulting in a total of six blocks. The order of compression level presentation was counterbalanced between participants. Within each compression level, we presented the two high and low-perceptual-load conditions in consecutive blocks, again counterbalancing order between participants.

Materials. Each target display contained eight round, visual stimuli presented in a ring-like arrangement on a black background (see Figure 1). The ring was presented with a visual angle of 18.9° (diameter 400 pixels ≙ 20 cm), with each circle in the ring spanning 3.4° (diameter 73 pixels ≙ 3.6 cm). One of these visual stimuli was the target and therefore appeared in one of the four target colors. Seven filler items were either presented in different shades of gray (low-load condition) or seven other non-target colors (high-load condition). These seven filler colors (pink, orange, turquoise, light green, purple, and two shades of brown), as well as the seven gray circles, were clearly distinguishable from the target colors, and randomly located at the seven remaining non-target positions in each trial (see also Appendix A).

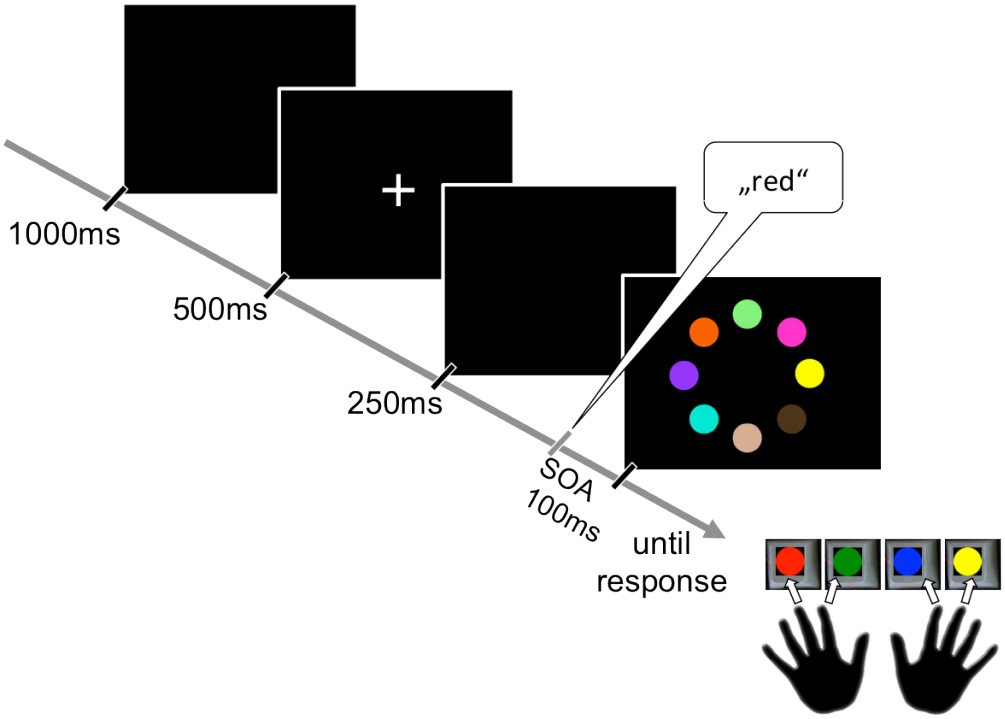


*Figure 1. Trial sequence of Experiment 1A or 1B (incongruent trial, high perceptual load). The prime word (red) is presented via headphones 100 ms prior to the visual target (yellow).*We generated auditory word stimuli by having a male speak the words several times and recording them with a Sennheiser condenser microphone (ME104; 16 bit mono; 44100 Hz digitization). *Audacity* (for *Windows*) served as recording software and a *MOTU* (*8pre*) sound card as hardware. We selected sound files that lasted 400 ms, which seemed to be the average length of the one-syllable words red (“rot”, [ro:t]), green (“grün”, [gry:n]), blue (“blau”, [blaʊ]), and yellow (“gelb”, [gɛlp]). The German word for ‘there’ (“hin” ([hɪn]) served as the neutral prime.

We created two different time-compressed versions of each of the five sound files: One with a duration of 120 ms (i.e., compression to 30 % of their original length), and another one with a duration of 40 ms (i.e., compression to 10 % length of the original length). Time compression of the stimuli was performed with the linguistic Praat freeware program (Boersma & Weenink, 2011) using the PSOLA (Pitch-Synchronous-Over-Lap-Add) Algorithm, which ensures a phonologically adequate time compression with an unchanged pitch. The sounds were presented over closed-ear headphones (TerraTec headset master 5.1 USB) and ranged in loudness from 68 to 72 dB SPL. The 30 %-

compression files were still understandable, whereas the 10 %-compression files were hardly discriminable without the context of the given task: That is, arbitrary 10 %-compression words without any constraining expectations are not identifiable; however, in the present context, with the usage of only five known words the categorization was feasible after hearing samples. This could be confirmed in a separate study (N = 13), in which we tested recognition of the five prime words for normal duration and both levels of time compression. One word was presented at a time and participants had to recognize which of the five words was presented. Duration was blocked, and words were presented five times each (per duration) in a random order, resulting in a total of 75 trials. Detailed accuracy rates are presented in Appendix B.

Procedure. Participants were individually tested in a sound proof chamber. They were seated in front of a 15-inch monitor (60 Hz refresh rate, resolution 640 × 480 pixels) controlled by a personal computer in an experimental cabin with dimmed light. The participants' viewing distance was about 60 cm and they wore closed-ear headphones. The experiment was conducted using E-prime software (E-prime 1.1). In each trial of the experiment, participants had to categorize the target color within the ring-like arrangement of eight items by pressing a corresponding key (keys d, f, j, and k on a standard keyboard). Participants were informed that they would not have any benefit from listening to the color words, since these were not predictive of the subsequent target color.

To start each block participants pressed the space bar. In each trial, following a 1,000-ms blank (black) screen, a white fixation cross appeared for 500 ms, followed by a blank screen for another 250 ms. Subsequently, the target screen was presented until a response was given (see Figure 1). The auditory prime presentation started with an SOA of 100 ms. Four warm-up trials (not included in the analyses) preceded each block. As an introduction to the experimental task, participants completed three practice phases. In Phase one, they were simply presented with one of the four target colors in a given trial (four times each) and had to learn the response key assignment. To support this process, they got stickers with the four target colors affixed to their left and right index and middle fingers. Feedback on task accuracy was provided on each trial. Phase two had 22 trials, and participants had to categorize color

circles (again presented as stand-alone stimuli) as either one of the four target colors (using the four response keys) or as one of the filler item colors (by pressing the space bar). Feedback was given on each trial. This phase should make participants familiar with the filler colors and their discriminability from target colors. The final practice phase (Phase three) comprised 30 trials identical to the trials of the main experiment (i.e., each trial contained one target color that had to be categorized by pressing the assigned key). The only difference with regard to the main experimental trials was that feedback was provided on each trial to ensure participants had understood their task and responded to visual targets only.

Each of the six experimental blocks consisted of 100 trials composed of 20 neutral, 20 congruent, and 60 incongruent trials, randomly intermixed. The trial list was completely balanced with regard to the four colors (i.e., each of the 16 possible target-prime combinations was presented five times, and each target color was used five times in the neutral condition). In order to control for response repetition effects, target colors were not repeated from one trial to another. Between blocks, participants took a self-timed break. The experiment lasted for approximately 45 minutes. Mean reaction times (RTs) constituted the dependent variable in this experiment.

4.1.2 Results

Error trials (6.5 %) and outliers (4.8 %; i.e., RTs that were 1.5 interquartile ranges above the third quartile or below the first quartile, respectively, with respect to the individual distribution; Tukey, 1977) were discarded. Table 2 shows the mean RTs for the conditions of our design.

RTs were analyzed in a 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) × 3 (semantic congruency: neutral vs. congruent vs. incongruent) MANOVA (i.e., we used the multivariate approach to the repeated measures analysis; see, e.g., Dien & Santuzzi, 2005; O'Brien & Kaiser, 1985). Besides the omnibus tests for the semantic congruency factor (with $df = 2$), we additionally report the results for the Stroop contrast between

congruent and incongruent conditions (the contrast of greatest interest in the present context).[15]

Table 2. Mean reaction times (RTs in ms; error rates in parentheses) in Experiment 1A as a function of semantic congruency conditions, search difficulty levels, and compression levels.

		Semantic congruency condition		
Perceptual load	Duration	Congruent	Neutral	Incongruent
Low	100 %	683 (7.2)	721 (7.8)	726 (6.8)
	30 %	689 (5.7)	693 (6.5)	712 (7.8)
	10 %	699 (4.7)	698 (8.1)	698 (6.8)
High	100 %	870 (7.2)	928 (5.7)	941 (6.4)
	30 %	858 (6.4)	922 (6.0)	937 (6.2)
	10 %	890 (6.9)	903 (7.2)	929 (5.1)

First of all, the analysis revealed a significant main effect of perceptual load, $F(1, 35) = 236.90$, $p < .001$, $\eta_p^2 = .87$. As expected, mean RTs were slower in the high than in the low-perceptual-load condition, revealing a difference in visual search difficulty. The main effect of duration and the interaction of perceptual load and duration were not significant, with both Fs < 1. With regard to the semantic congruency manipulation, the analysis revealed a significant main effect, $F(2, 34) = 31.71$, $p < .001$, $\eta_p^2 = .65$ [$F(1, 35) = 56.24$, $p < .001$, $\eta_p^2 = .62$, for the contrast incongruent vs. congruent] that was moderated by perceptual load, $F(2, 34) = 22.81$, $p < .001$, $\eta_p^2 = .57$ [$F(1, 35) = 45.03$, $p < .001$, $\eta_p^2 = .56$, for the contrast incongruent vs. congruent]. Besides,

[15] Since we used the multivariate approach to the repeated measures analysis, the tripartite factor of the semantic congruency condition is – as part of the procedure – transformed into a vector of two orthogonal contrast variables (see, e.g., Dien & Santuzzi, 2005). We chose the first contrast as the contrast between congruent and incongruent trials a priori. For the second contrast, scores were averaged across congruent and incongruent trials and contrasted with the neutral stimuli. This contrast is of minor interest in the present context.

the main effect of the semantic congruency manipulation was moderated by duration, when considering the contrast of main interest, $F(2, 34) = 3.28$, $p = .05$, $\eta_p^2 = .16$ [$F(4, 32) = 2.17$, $p = .10$, $\eta_p^2 = .21$, for the omnibus test]. The three-way interaction was not significant, $F(4, 32) = 1.05$, $p = .40$, $\eta_p^2 = .12$. As can be seen in Figure 2a, Stroop effects decreased from high load to low load and from 100 % to 10 % duration. Nevertheless, all effects were significantly different from zero, ts(35) > 2.92, ps < .006, except the one for the shortest duration in the low-perceptual-load condition, $t(35) < 1$.

In addition, difference scores for facilitation (neutral - congruent) and interference (incongruent - neutral) were calculated for all six experimental conditions (see Figure 2b/c). Both indicators were submitted to 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) MANOVAs. For facilitation, the main effect of duration was significant, $F(2, 34) = 3.87$, $p = .03$, $\eta_p^2 = .19$, as well as the main effect of perceptual load, $F(1, 35) = 13.03$, $p < .001$, $\eta_p^2 = .27$ but not the interaction, $F(2, 34) = 2.12$, $p = .14$, $\eta_p^2 = .11$. Figure 2 shows that the pattern of facilitation scores closely mimics the pattern for the overall Stroop effects (i.e., the comparison of congruent and incongruent trials), with higher facilitation effects in the high than in the low-perceptual-load conditions. The analysis of interference revealed neither significant main effects nor a significant interaction, all Fs < 1.33, ps > .26, η_p^2s < .06. However, averaging the interference effects across the six conditions yielded a mean significantly above zero, $F(1, 35) = 9.71$, $p = .004$, $\eta_p^2 = .22$ [$F(1, 35) = 17.88$, $p < .001$, $\eta_p^2 = .34$ for the corresponding averaged facilitation score].

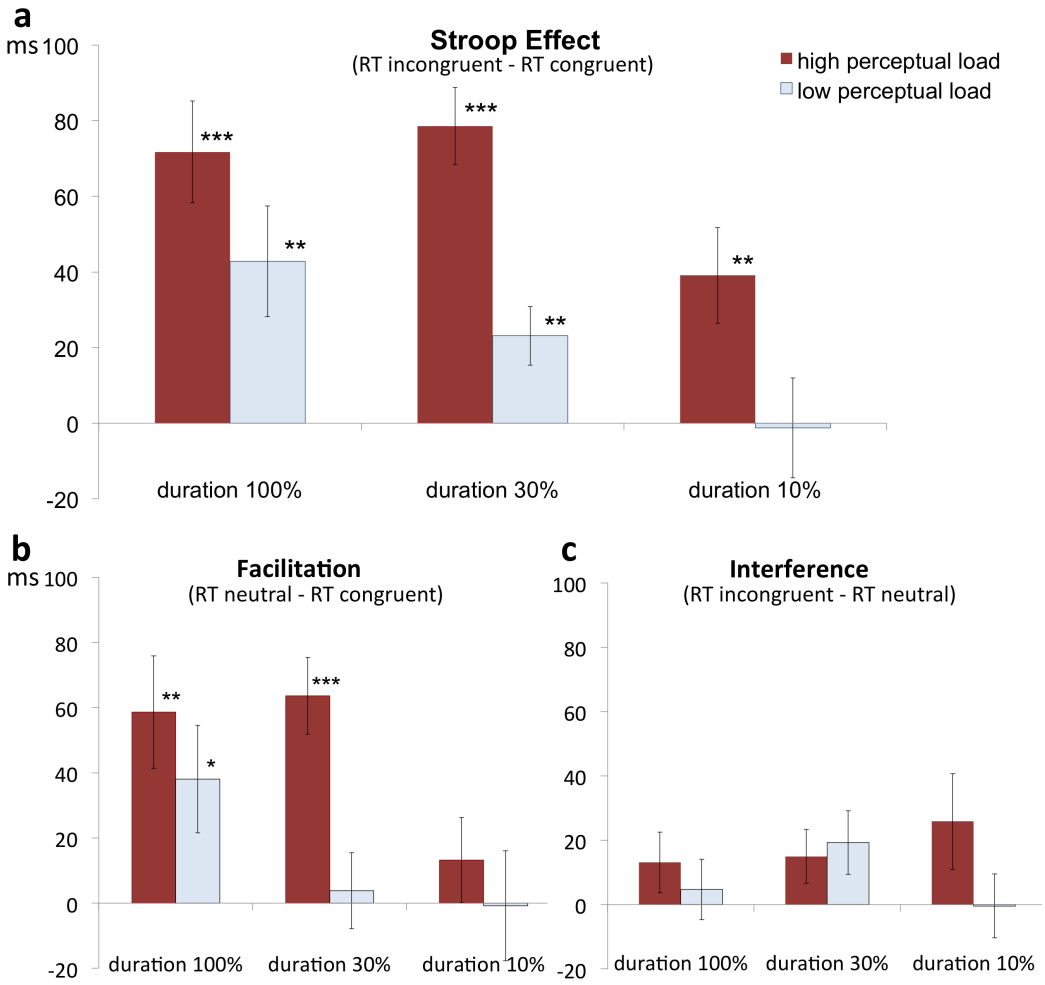


*Figure 2. Reaction time (RT) differences (Stroop: incongruent trials – congruent trials; facilitation: neutral – congruent; interference: incongruent – neutral) for the three compression rates (duration) and the two perceptual-load levels in Experiment 1A. The error bars indicate ± 1 SEM (see Franz & Loftus, 2012). * $p < .05$, ** $p < .01$. *** $p < .001$.*

For the error rates (see Table 2), we conducted the same analysis as for the RTs. The 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) × 3 (semantic congruency: neutral vs. congruent vs. incongruent) MANOVA did not yield any results contradicting the RT analysis. Neither the main effects (all Fs < 1.49, ps > .23, η_p^2s < .22), nor the two-way interactions (all Fs < 2.76, ps > .06, η_p^2s < .24), nor the three-way interaction ($F < 1$) were significant. Despite the non-significance of the MANOVA results, we tested the 3 (duration) × 2 (perceptual load) simple Stroop effects (i.e., the comparisons of the incongruent and congruent conditions) to provide full

transparency with regard to potential speed-accuracy trade-offs. We used the Bonferroni-Holm method of protection against α-error accumulation (Holm, 1979). None of the Stroop effects was significant ($|t|$s < 1, except for $t(35) = 2.36$, $p = .024$, for 30 %/low, $t(35) = 2.25$, $p = .031$, for 10 %/low; $t(35) = -1.96$, $p = .058$, for 10 %/high): There were two numerically positive effects and three clear null effects. Only one effect was numerically negative, but even it would be non-significant if we used nonadjusted p values.

4.1.3 Discussion

We investigated whether spoken color words – even without revealing any contingent information – influence the categorization of visually presented color patches in two visual search contexts. Our results clearly support this assumption by the significant and remarkably large difference in RTs between congruent and incongruent trials in five out of six experimental conditions. The only condition that did not reveal an effect contained highly compressed color words, lasting for merely 10 % of the original length, in combination with a low search task difficulty. Not having found effects in this quite extreme condition might be due to a combination of overall weaker effects for time-compressed stimuli and a quite easy search task that generally benefits less from unpredictable support.

We also investigated whether the visual perceptual-load level modifies the RT difference between congruent and incongruent crossmodal stimulus presentation. This hypothesis was supported, since the high-perceptual-load condition revealed larger effects at all duration levels. While performing a visually demanding task, participants seem to be more susceptible to crossmodal auditory influences than during performance of a rather simple task. Thereby, we conceptually replicated the finding by Telllinghuisen and Nowak (2003), which was in contrast to the original findings by Lavie and Cox (1997) for visual distractors. Note that, not unexpectedly and in line with Tellinghuisen and Nowak's results, the mean RTs were significantly faster for the low-perceptual-load conditions than for the high-perceptual-load conditions (see Table 2). Thus, prima facie, one might argue (for both studies) that this could have simply left more time for the auditory stimuli to be entirely processed in the high-perceptual-load condition, resulting in an increased

influence on visual task performance. According to this rationale, time-compressed stimuli with an earlier ending of spoken word presentation should also lead to increased effects – above all in the faster low-load condition. By contrast, time-compressed stimuli do not increase Stroop effects when comparing, for example, 30 % duration with original duration in the low-load condition. The main effect of duration implies that time-compressed, shorter speech even leads to declined crossmodal Stroop effects. Even though recognition accuracy rates remain high even for highly compressed color words (see Appendix B), the declined Stroop effects in the highly compressed word condition might be due to the unusual presentation of the word primes. The primes do not comprise contingent information, and additionally, they are explicitly known to be unhelpful. Therefore, participants could conclude that they were better off trying to ignore them. Apparently, this was easier for highly and unnaturally compressed spoken words.

Another remarkable point is the cost-benefit partitioning of the Stroop effect. Stroop-like effects are mostly caused more by interference than by facilitation, (MacLeod, 1991; Roelofs, 2005). In our study, however, for understandable word durations, facilitation seems to be the major component (see Figure 2). This asymmetrical pattern for crossmodal presentation is consistent with the findings of Tellinghuisen and Nowak (2003). For effects originating from stimulus-response compatibility (De Houwer, 2003), both benefits in congruent cases and costs in incongruent cases would be expected. Yet, our experiment only revealed modest costs (that were significant only if all conditions were collapsed). With regard to this issue, it is important to note that we used only a single, unvarying word for the neutral condition. On a critical note, we have to admit that the interpretation of crossmodal benefits and costs might be complicated if it is based on only one simple type of neutral condition. We will soon return to this issue. However, the “best guess” for a critique of our neutral condition would be that it might have artificially improved performance, since the neutral word was easily distinguishable from the four target-relevant words, which varied across trials. If this were the case, having found remarkable facilitation effects even with possibly overestimated performance in the neutral condition is all the more surprising.

To sum up, we found clear crossmodal effects of spoken words on object classification at an SOA of only 100 ms. Furthermore, the problem of the duration of spoken words as a fundamental problem of crossmodal research was addressed by applying two time compression levels. Even with a duration of only 40 ms we found a Stroop effect (albeit only for high perceptual load). In contrast to Lavie's theory (1995) – which, however, merely targets unimodal stimulus presentation – but in concordance to Tellinghuisen and Nowak (2003), our results for crossmodal presentation show increased Stroop effects in a high-perceptional-load (as compared to low-load) condition when using crossmodal presentation.

From the unimodal Stroop literature it is known that presence of facilitation effects depends on the choice of the neutral condition (Duncan-Johnson & Kopell, 1981; Kahneman & Chajczyk, 1983; MacLeod, 1991). The neutral condition of our Experiment 1A comprises merely a single unvarying word. We have already stated that this choice might have led to relatively fast responses in the neutral condition, and therefore to an underestimation of (our already large) facilitation effects. However, since we emphasized the relative size of facilitation and interference in crossmodal semantic priming, we decided to scrutinize crossmodal Stroop effects by means of another neutral baseline condition: A set of four neutral stimuli was chosen for Experiment 1B. This conceptual replication would make our interpretation more credible, if results are highly comparable to those of Experiment 1A.

4.2 Experiment 1B: Is the choice of the baseline condition essential? A replication of crossmodal effects

Experiment 1B was an almost exact replication of Experiment 1A, except for the choice of the neutral condition. We generated a set of four pronounceable neutral nonwords, which contained different phonemes and sounded different from the color words. Hearing nonwords in the neutral condition would elicit a phonological impact comparable to hearing the color words, but no semantic effects were expected.

4.2.1 Method

Participants. A total of 37 students (26 females, 11 males) from Saarland University took part in Experiment 2. Participants received 8 Euro or course credit for taking part. All participants had normal or corrected-to-normal vision, and none of them reported any color-blindness or hearing problems.

Design, materials, and procedure. These were identical to Experiment 1A, except for the fact that the neutral condition of the congruency factor was altered: The single neutral word was replaced by four neutral nonwords. Instead of the German word for 'there' ("hin" ([hɪn]), four one-syllable nonwords were used: "liez", [liːʦ], "tän", [tɛːn], "nux", [nʊks], and "töff", [tœf]. Duration and the time compression process regarding the nonwords were also equivalent to the treatment of the color words. For this set of eight words, we conducted an equivalent study as for the stimulus set in Experiment 1A. Accordingly, each of the 13 (new) participants executed 120 trials. Detailed accuracy rates can be taken from the Appendix B. Overall recognition was again almost perfect. Besides, misclassifications of neutral stimuli on the 10 %-compression level as color words decreased from 10.8 % for the set used in Experiment 1A to only 1.7 % for the set designed for Experiment 1B, allowing for clearer interpretation of the effects.

4.2.2 Results and Discussion

Error trials (4.8 %) and outliers (2.5 %; i.e., RTs that were 1.5 interquartile ranges above the third quartile or below the first quartile, respectively, with respect to the individual distribution; Tukey, 1977) were discarded. Table 3 shows the mean RTs for the conditions of our design.

RTs were again analyzed in a 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) × 3 (semantic congruency: neutral vs. congruent vs. incongruent) MANOVA. Besides the omnibus tests for the semantic congruency factor (with $df = 2$), we additionally report the results for the Stroop contrast between congruent and incongruent conditions (the contrast of greatest interest in the present context).

Table 3. Mean reaction times (RTs in ms; error rates in parentheses) in Experiment 1B as a function of semantic congruency conditions, search difficulty levels, and compression levels.

		Semantic congruency condition		
		Congruent	Neutral	Incongruent
Perceptual load	Duration			
Low	100 %	716 (5.4)	748 (6.5)	744 (4.8)
	30 %	709 (5.9)	736 (4.7)	733 (5.0)
	10 %	709 (4.6)	720 (4.7)	722 (5.0)
High	100 %	948 (4.5)	1027 (4.1)	1080 (5.1)
	30 %	945 (5.1)	1042 (5.0)	1074 (4.0)
	10 %	946 (4.9)	1017 (4.6)	1017 (4.1)

First of all, the analysis revealed a significant main effect of perceptual load, $F(1, 36) = 190.41$, $p < .001$, $\eta_p^2 = .84$. As in Experiment 1, mean RTs were slower in the high-perceptual-load condition, revealing a difference in visual search difficulty. The main effect of duration and the interaction of perceptual load and duration were not significant, with both Fs < 1.18. With regard to the semantic congruency manipulation, the analysis revealed a significant main effect, $F(2, 35) = 38.79$, $p < .001$, $\eta_p^2 = .69$ [$F(1, 36) = 75.60$, $p < .001$, $\eta_p^2 = .68$, for the contrast incongruent vs. congruent] that was moderated by perceptual load, $F(2, 35) = 31.47$, $p < .001$, $\eta_p^2 = .64$ [$F(1, 36) = 64.51$, $p < .001$, $\eta_p^2 = .64$, for the contrast incongruent vs. congruent]. The effect of semantic congruency manipulation was furthermore moderated by duration, $F(4, 33) = 3.88$, $p = .01$, $\eta_p^2 = .32$ [$F(2, 35) = 7.63$, $p = .002$, $\eta_p^2 = .30$, for the contrast incongruent vs. congruent]. The three-way interaction was not significant, $F(4, 33) = 1.86$, $p = .14$, $\eta_p^2 = .18$. As can be seen in Figure 3a, Stroop effects decreased from high load to low load and from 100 % to 10 % duration. Nevertheless, all effects were significantly different from zero, ts(36)

> 2.88, *p*s < .007, except the one for the shortest duration in the low-perceptual-load condition, $t(36) = 1.36$, p = .18.

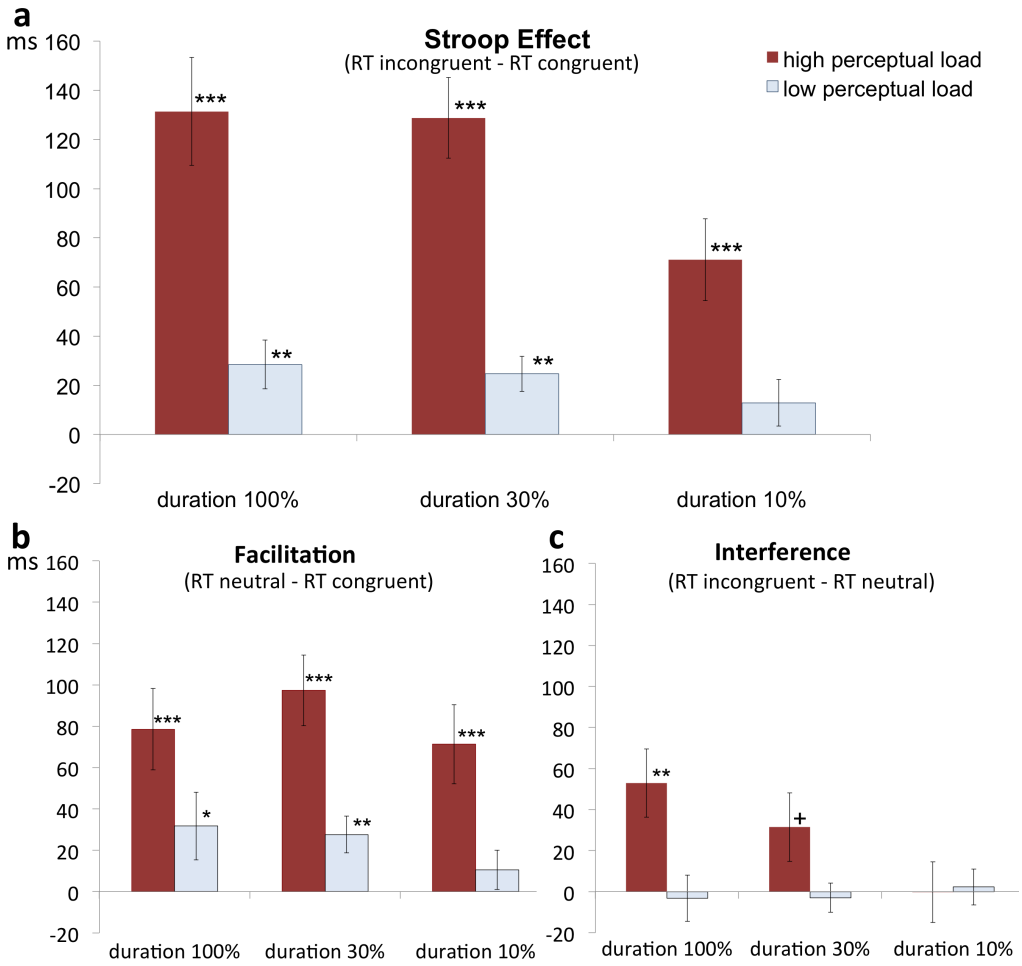


*Figure 3. Reaction time (RT) differences (Stroop: incongruent trials – congruent trials; facilitation: neutral – congruent; interference: incongruent – neutral) for the three compression rates (duration) and the two perceptual-load levels in Experiment 1B. The error bars indicate ± 1 SEM (see Franz & Loftus, 2012). + p < .05, one-tailed, * p < .05, ** p < .01. *** p < .001.*

In addition, difference scores for facilitation (neutral - congruent) and interference (incongruent - neutral) were calculated for all six experimental conditions (see). Both indicators were submitted to 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) MANOVAs. For facilitation, the main effect of perceptual load was significant, $F(1, 36) = 28.90$, $p < .001$, $\eta_p^2 = .45$ but neither the main effect of duration, $F(2, 35) = 1.14$, $p = .33$, $\eta_p^2 = .06$, nor the interaction, $F < 1$, was. Figure 3 shows that the pattern of

facilitation scores closely mimics the pattern for the overall Stroop effects (i.e., the comparison of congruent and incongruent trials), with higher facilitation effects in the high-perceptual-load conditions compared with the low-perceptual-load conditions. The analysis of interference revealed a significant main effect of perceptual load, $F(1, 36) = 9.43$, $p = .004$, $\eta_p^2 = .21$. Interference effects were clearly present in the uncompressed high-perceptual-load condition, but no interference was found for any of the low-perceptual-load conditions. Neither the main effect of duration, $F(2, 35) = 1.16$, $p = .33$, $\eta_p^2 = .06$, nor the interaction, $F(2, 35) = 2.75$, $p = .08$, $\eta_p^2 = .14$, were significant. Averaging the interference effects across the six conditions yielded a mean significantly above zero, $F(1, 36) = 6.31$, $p = .02$, $\eta_p^2 = .15$ [$F(1, 36) = 60.35$, $p < .001$, $\eta_p^2 = .62$ for the corresponding averaged facilitation score].

For the error rates (see Table 3), we conducted the same analysis as for the RTs. The 3 (duration: 100 % vs. 30 % vs. 10 %) × 2 (perceptual load: low vs. high) × 3 (semantic congruency: neutral vs. congruent vs. incongruent) MANOVA did not yield any results contradicting the RT analysis. Neither the main effects of duration and semantic congruency condition (both Fs < 1, ps > .55, η_p^2s < .15), nor the effect of perceptual load [$F(1, 36) = 1.90$, $p = .18$, $\eta_p^2 = .27$], nor the two-way interactions (all Fs < 1, ps > .50, η_p^2s < .16), nor the three-way interaction [$F(4, 33) = 1.16$, $p = .35$, $\eta_p^2 = .32$] were significant. Despite the non-significance of the MANOVA results, we tested the 3 (duration) × 2 (perceptual load) simple Stroop effects (i.e., the comparisons of the incongruent and congruent conditions) to provide full transparency with regard to potential speed-accuracy trade-offs. We used the Bonferroni-Holm method of protection against α-error accumulation (Holm, 1979). None was significant (all $|t|$s < 1, except for $t(36) = -1.80$, $p = .08$, for 30 %/high): There were five clear null effects and one even non-significant numerically negative effect.

Overall, the results from Experiment 1B confirm our findings from Experiment 1A. Even though we used four nonwords as stimulus material for the neutral condition, the crossmodal effects of spoken words on visual target classification remained comparably clear. Furthermore, the dominance of facilitation was replicated.

As already mentioned in the introduction, one important question remains unanswered by the preceding experiment: Should crossmodal semantic Stroop effects be attributed solely to semantic stimulus-response compatibility (De Houwer, 2003) or could our effects be (additionally) based on perceptual priming effects of semantics? In order to control for the effects of response priming, and thereby separate the aforementioned two accounts, we conducted Experiment 2.

4.3 Experiment 2: Spoken words crossmodally enhance the detection sensitivity for semantically congruent visual targets

As we outlined above, the crossmodal effects of the first two experiments could either be attributed to crossmodal response priming or to crossmodal semantic facilitation. The response-priming account suggests that the prime involuntarily triggers its corresponding response. This preactivation of responses could influence reactions to targets, even if these were presented in a different modality: It could accelerate or slow down responses, depending on the response-related congruence or incongruence of prime and target. On the other hand, the crossmodal semantic-priming account implies that the semantic content of an irrelevant auditory prime involuntarily enhanced visual sensitivity regarding semantically related objects, even if these were presented in another modality, and hence were physically diverse (crossmodal stimulus-stimulus compatibilities). With Experiment 2, we wanted to specify the processes underlying our previous results. Therefore, we separated stimulus-response compatibility effects from stimulus-stimulus compatibility effects using a signal detection approach. In Experiment 2, each target display was presented only for a short duration, and participants had to decide whether it had (target-present key) or had not (target-absent key) contained one out of the four target colors. For target-present responses, no indication of the concrete color had to be given. Since the same key was pressed for all target-present responses, stimulus-response compatibility effects could have an effect on the respective discrimination. However, if the spoken words elicited crossmodal stimulus-stimulus compatibilities by increasing visual sensitivity, Stroop effects would still be expected for target discrimination in this signal detection approach.

4.3.1 Method

Participants. The participants of Experiment 2 were a new group of 38 undergraduates (29 females, 9 males) from Saarland University, who participated to fulfill course requirements or received a chocolate bar for voluntary participation. All had normal or corrected-to-normal vision and none reported color-blindness or hearing problems.

Design. Two factors, semantic congruency (congruent vs. incongruent vs. neutral [featuring either a control prime word or silence], and target (present vs. absent), were manipulated.

Materials. In this experiment, each target display contained eight visual stimuli comparable to the displays in Experiments 1A and 1B. The target stimulus was either one of these stimuli or was absent. Only the high-perceptual-load condition of Experiments 1A and 1B was employed. In each trial, seven (target-present condition) or eight patches (target-absent condition) were chosen from a set of 16 distracting filler colors and were randomly placed at the available positions. The filler colors were generated by mixing each original target color with black, or white, or both, creating four different filler colors from each target color. Each of the 16 filler colors was clearly distinguishable from the four target colors. Auditory prime stimuli were identical to those used in the 30 %-duration condition in Experiments 1A and 1B.

Procedure. The laboratory, equipment, and software were identical to Experiment 1. On each trial of the experiment, participants had to decide whether or not one of the four target colors was present. No categorization of the perceived target color was required. Participants were again informed that the auditory words were uncorrelated with the subsequent target display.

To start each trial, participants pressed the space bar. After a 1,000-ms blank (black) screen, a central white fixation-cross appeared for 500 ms, followed by a blank screen for another 250 ms. Subsequently, the target screen was presented for only 150 ms followed by a blank screen for 2,000 ms. The SOA was 100 ms, and prime presentation time was for 120 ms (i.e., 30 % compression). The maximum RT was 2,150 ms; if participants did not respond in the given time window, they were instructed to speed up their responses.

As an introduction to the experimental task, participants completed two practice phases. In Phase 1, participants were initially given the opportunity to familiarize themselves with the four target color circles. They were then given 26 sample trials in which single color patches were presented, and participants decided whether the patches matched one of the four target colors or one of the filler colors (by pressing the "target present" ["j"] key with their right index finger or the "target absent" ["f"] key with their left index finger). Feedback was provided on each trial. Phase two had 24 practice trials and was identical to the main experiment, except for the provision of feedback. Each of the two main experimental blocks consisted of 120 randomly intermixed trials (both preceded by four warm-up trials which were not included in the analyses). Half the trials contained a target color. The target-present trials were composed of 10 congruent, 30 incongruent, and 20 neutral trials. The neutral trials comprised 10 trials with a control word ("hin" as in Exp. 1A) and 10 trials with silence. For the nontarget trials, the same proportion of auditory color words, control words, and silence was presented as in the target-present trials. The number of target colors was counterbalanced for the different semantic congruency conditions. Participants had the opportunity to take a break in-between the two blocks. The experiment lasted for approximately 30 min. Accuracy of target detection constituted the dependent variable in this experiment.

4.3.2 Results

The participants' target-present responses in the target-present condition indicated their overall hit rate, whereas their error rate in the target-absent condition constituted their overall false alarm rate. As a first step, hit and false alarm rates were calculated separately for the three different prime types (see Table 4): A target-relevant color word, a neutral word, or no prime word could be presented preceding the target-present or target-absent screen. Since these three prime types were presented in an intermixed block, three false alarm rates in contrast to four hit rates could be determined separately. The false alarm rates were derived for each of the aforementioned three prime types in target-absent trials. For target-present trials, the hit rates for target-relevant color words could be determined separately for congruent and incongruent trials,

resulting in four different hit rates. Hit rates of 100 % (applied to 3.3 % of all hit rates) or false alarm rates of 0 % (applied to 5.3 % of all false alarm rates) were adjusted as recommended by (Macmillan & Creelman, 2005).[16]

Table 4. Parameter overview for prime conditions in Experiment 2 (standard errors in parentheses)

Prime	Hit Rate	False Alarm Rate	*d'*	*c*	RT[ms]
Task-relevant	.77 (0.02)	.22 (0.03)		0.05 (0.05)	
Congruent	.85 (0.02)		2.10 (0.14)		629 (12.1)
Incongruent	.75 (0.02)		1.64 (0.12)		667 (15.1)
Neutral control	.77 (0.02)	.23 (0.02)	1.65 (0.17)	0.03 (0.05)	669 (18.0)
Neutral silence	.77 (0.03)	.25 (0.03)	1.61 (0.14)	0.01 (0.06)	668 (19.4)

Note. The hit rate for task-relevant primes is composed of 20 congruent and 60 incongruent trials, and is therefore not identical with the mean of congruent and incongruent hit rates. The false alarm rate for task-relevant color was used to calculate *d'* of both the congruent and incongruent conditions (see text for further explanation). The mean reaction times (RTs) of correct target-present decisions are specified.

As a first step, we wanted to find out whether – despite intermixed presentation – the three different prime types (task-relevant, neutral control, and neutral silence) led to different response criteria.[17] For this purpose, *c* indices for the three prime types (see Table 4) were analyzed in a repeated measures MANOVA. We obtained neither a significant main effect of prime type, $F(2, 36) = 1.17$, $p = .32$, $\eta_p^2 = .06$, nor a significant result in any Helmert contrast (task-relevant vs. neutral [pooled] and neutral [control] vs. neutral [silence]), both Fs$(1, 37) < 1.53$, ps $> .22$, η_p^2s $< .04$. These results indicate that task-relevant primes did not influence participants' response criterion.

For the *d'* data, we split the hit rates for the color word prime condition according to congruent and incongruent trials, in order to analyze whether

[16] A hit rate of 100 % in any condition was adjusted as 1-(1/ 2N), where N is the number of trials in that condition. Furthermore, false alarm rates of 0 % were adjusted as 1/ 2N.

[17] At this stage of analysis, congruent and incongruent prime conditions cannot be distinguished because *a priori* there are no independent target-absent trials for congruent versus incongruent conditions due to congruency being defined by the relationship between the prime and a (present) target.

word prime compatibility would influence detection sensitivity. Analogous to Chen & Spence (2011; see also Lupyan & Spivey, 2010), the false alarm rate for target-relevant color word primes was used for the determination of *d'* for both congruent and incongruent trials. The *d'* data (see Table 4) for all four semantic congruency conditions was analyzed in a repeated measures MANOVA. We found a significant main effect of semantic congruency, $F(3, 35) = 10.16$, $p < .001$, $\eta_p^2 = .47$, indicating an influence on detection sensitivity. We found a significant difference between congruent trials and all other conditions, $F(1, 37) = 27.22$, $p < .001$, $\eta_p^2 = .42$, indicating that in congruent trials detection sensitivity was highly increased. For all further contrasts (incongruent vs. neutral [pooled]; neutral [control] vs. neutral [silence]) we did not obtain any significant differences between conditions, all Fs < 1.

Even though participants were not forced to complete the detection task quickly, we additionally analyzed the RTs of correct target-present decisions. We discarded outliers (2.6 % of all correct trials) – that is, RTs that were 1.5 interquartile ranges above the third quartile or below the first quartile, respectively, with respect to the individual distribution (Tukey, 1977). Mean RTs are shown at the right of Table 4. In a repeated measures MANOVA, we found a significant main effect of semantic congruency, $F(3, 35) = 7.13$, $p = .001$, $\eta_p^2 = .38$. Analogous to the detection sensitivity results, we found a significant difference between congruent trials and all other conditions, $F(1, 37) = 15.48$, $p < .001$, $\eta_p^2 = .30$. RTs for the incongruent condition were not significantly different from those in the neutral conditions, $F < 1$, and the comparison between the neutral conditions (neutral [control] vs. neutral [silence]) was not significant, as well, $F < 1$.

4.3.3 Discussion

When presenting time-compressed spoken color words with an SOA as short as 100 ms, we found higher detection sensitivity for targets in trials with a congruent crossmodal semantic presentation than for neutral (control word or silence) or incongruent trials. Since we controlled for response-priming effects in this experiment, our results imply crossmodal semantic priming on visual perception. The priming effect was found in the sensitivity parameter *d'* as well

as in response latencies. This semantic priming effect extends beyond Tellinghuisen and Nowak's (2003) interpretation that their crossmodal effects in a visual search task occurred solely due to response priming. Experiment 2 thus demonstrates that the crossmodal effects of Experiments 1A and 1B are (at least partially) based on enhanced (and occasionally deteriorated) target perception. Moreover, this is the first report of a crossmodal semantic priming effect for time-compressed verbal primes with an SOA considerably less than 300 ms, within a target detection task not requiring target identification. With an SOA of 346 ms, Chen and Spence (2011) did find effects for naturalistic sounds on picture detection without identification, but they did not find the same effects for spoken words. They argued that for spoken words, reporting the identity of a target would be necessary to facilitate participants' visual picture identification performance in congruent trials. In contrast to this auxiliary assumption, we showed that a mere target detection task was sufficient to elicit crossmodal benefits. Accordingly, crossmodal semantic priming effects can occur even without explicit target identification as a precondition. Chen and Spence's (2011) assumption that semantic representations of spoken words are generally too weak or their activation is too slow was overcome by our study. At this point, we want to highlight the fact that our participants had to keep the four relevant target colors active in mind in order to perform the task. This circumstance implies a strong mental set for four color words, which might have boosted the impact of those spoken words matching the set. By contrast, in Chen and Spence's (2011) task pictures of 30 items were used. The pictures were introduced to the participants once before the main experiment, but they did not have to be memorized. Participants probably did not keep these items active in a mental set, since their task was solely to indicate in each trial whether any picture or nothing had been presented. The situation might have changed if the task were to additionally identify the picture, and this might be a reason for Chen and Spence (2011) finally finding crossmodal effects.

Our findings support the more general assumption about involuntary crossmodal activation of (semantic) associations without identification as a necessary precondition. However, we assume that each task elicits a mental set whose intensity might modulate activation magnitude of respective items. Our

task comprised a small and concise mental set that might be responsible for semantic speech stimuli accessing semantically associated representations especially quickly (in only 100 ms, as opposed to Chen and Spence's 2011, suggestion of 200-350 ms).

Our consideration of crossmodal activation of semantic associations is further supported by the "unity assumption" (Chen & Spence, 2011; Spence, 2007), which claims that the binding of information from different sensory modalities is facilitated by semantic congruency. Binding of visual and auditory stimuli occurs with larger spatiotemporal disparities when they are physically plausible (Jackson, 1953) or semantically congruent (Driver & Spence, 1998) than when they are physically unrelated or semantically incongruent. The associative strength between semantically congruent visual and auditory stimuli might be a modulating factor for crossmodal effects. Our results indicate that strong associations of colors (as part of an activated attentional set) lead to remarkable crossmodal effects.

Another issue remains to be addressed: In summary, irrelevant spoken words did not hamper detection performance, as compared with the silence condition. Accordingly, spoken words did not in and of themselves present a hindrance to visual performance in our task. Moreover, our results show that (even irrelevant) congruent primes do actually enhance visual detection compared with all other conditions, and that this cannot merely be interpreted as arising from less interference being provoked by congruent primes than by neutral or incongruent primes. Unlike Chen and Spence (2011), we could distinguish between facilitation and interference effects in crossmodal semantic priming by introducing a clearly defined neutral condition. Besides, by additionally comparing a neutral spoken word condition to a silence condition in Experiment 2, our findings provide information beyond Lupyan and Ward (2013), who found intermediate performance for trials without auditory stimuli but did not present neutral words. Most importantly, our findings demonstrate significant crossmodal facilitation, as opposed to merely the absence of interference. Regarding our applied research question on speech warnings in time-critical situations, this finding is highly interesting: Warnings that match the critical situation might enhance performance not only above an unspecific "master alert" level, but also above cases without warnings.

4.4 Discussion of crossmodal effects: Spoken word primes enhance the processing of visual targets

The results of our first three experiments clearly show that even if spoken words are semantically uncorrelated (i.e., non-predictive) with subsequent visual targets, they can influence target processing and the subsequent response selection process significantly. RTs for congruent trials were remarkably faster than for both neutral and incongruent trials (Exps. 1A and 1B). Furthermore, significant facilitation was found for visual color detection: Following congruent spoken word primes, participants detected visual target colors much more easily than after a neutral word prime, neutral silence, or incongruent primes (Exp. 2). We would like to especially emphasize Experiment 2, where the results indicate that stimulus-stimulus compatibilities, and hence semantic priming, made an important contribution to our crossmodal effects. Accordingly, stimulus-response compatibilities cannot be the sole cause for crossmodal enhancement. This result fits with findings from Chen and Spence (2011), who found crossmodal semantic priming effects of spoken words on picture detection with a signal detection approach. However, Chen and Spence found crossmodal effects only with a (comparably long) SOA of 346 ms, whereas we employed an SOA of only 100 ms.

Crossmodal effects of spoken words were even larger in high-perceptual-load conditions than in low-perceptual-load conditions, and facilitation was pronounced. While performing a visually demanding task, participants seem to be more susceptible to crossmodal auditory influences. This resembles Tellinghuisen and Nowak's (2003) results but stands in contrast to Lavie's theory for unimodal stimuli (Lavie & Cox, 1997; Lavie, 1995). If we assume that the effects in Experiments 1A and 1B occurred (at least partly) due to stimulus-stimulus compatibilities, Lavie's theory would need to be extended and adjusted regarding crossmodal stimulus presentation: Distractors presented in a modality different from the target modality can apparently access semantic representations independently of perceptual load in the target modality. In case of semantic congruency, the distractor can enhance target object detection even if prime and target are uncorrelated. This effect emerges even more in tasks with higher perceptual load, since the overall potential of support is higher.

As we said, we found asymmetrical result patterns in our experiments, with facilitation as the major component. This held true for the different neutral conditions that we applied in Experiments 1A and 1B. Pronounced facilitation is opposed to Roelofs' (2005) findings with a silence control condition, but consistent with the conclusion of Tellinghuisen and Nowak (2003) that was based on several control conditions. As was already mentioned earlier, the choice of the baseline condition is known to be a crucial factor for finding reliable facilitation. If Roelofs had used neutral words instead of silence for his neutral condition, he might have found facilitation as well. Likewise, facilitation might have been more pronounced in the findings by Lupyan and Ward (2013). However, when we employed silence as a neutral condition in Experiment 2, we did not find different effects as compared to a neutral word condition. In addition, we found clear facilitation in color detection relative to both neutral conditions. For effects originating from stimulus-response compatibility (De Houwer, 2003), we would expect both benefits in congruent cases and costs in incongruent cases. Nevertheless, Experiments 1A and 1B only revealed minor costs, as compared to relatively strong facilitation. This result already points toward semantic stimulus-stimulus compatibility being substantially involved in crossmodal semantic effects. However, in Experiments 1A and 1B the process components (i.e., response-related components and semantic-priming components) could not be separated, and hence this assumption needed consideration of Experiment 2.

The factor time compression was introduced as a novel approach to the problem of a possible confound by "disambiguation duration." For the first time we were able to show crossmodal Stroop and semantic priming effects with time-compressed stimuli. If time compression was not too intense and the spoken word remained understandable, crossmodal effects of spoken words on a visual search task were reliably found in all our experiments. However, we found that an earlier end of the spoken word presentation did not increase effects (which might be expected on the basis of the disambiguation duration argument). At an SOA of 100 ms, we found decreased crossmodal effects of highly time-compressed primes for both perceptual-load conditions. It might be the case that the effect of compression level on Stroop effects would be decreased, or even reversed (i.e., larger Stroop effects with higher

compression), for negative SOAs that leave an even smaller time window for prime-induced facilitation until the response is given. This might hold especially for predictive primes, since participants would try to decode auditory prime information as soon as possible in order to support reactions to visual targets. We will return to this assumption later in Chapter 5.

To sum up, spoken word primes could significantly influence target processing in complex visual scenes – even if they were time-compressed and no benefit could be drawn from listening. These crossmodal effects were primarily facilitative in nature: Congruent spoken words led to benefits in the processing of visual targets, whereas incongruent spoken words only led to minor costs (as compared to neutral spoken words or silence). Moreover, these effects were at least partly due to crossmodal stimulus-stimulus compatibilities in semantic priming, and could not merely be explained by stimulus-response compatibilities.

With regard to the driving context, it is highly interesting that spoken words have the potential to crossmodally enhance the performance in a visual task – even if they do not predict the subsequent target and listening is completely task-irrelevant. Our findings suggest a very fast and efficient transfer of semantic information that does not require intentional processing. Such a rather automatic process would be extremely valuable for warnings in time-critical situations. Therefore, we were optimistic about the usage of speech warnings in general and decided to further scrutinize the frame conditions of crossmodal semantic priming effects.

Numerous modifications of our experiments were worth further consideration, in order to receive a more precise picture of the nature of crossmodal Stroop and semantic priming effects. More details about necessary preconditions needed to be revealed in order to facilitate a transfer of our basic findings to more applied scenarios. For example, when using higher levels of contingency with a large proportion of congruent trials in semantic priming, we would expect increased benefits, but also increased interference (see also Lupyan & Thompson-Schill, 2012, Exps. 3A and 3B). With regard to in-car warning systems for time-critical situations, it is absolutely reasonable to assume contingency levels above chance. Accordingly, a close investigation how contingency affects crossmodal semantic effects was a relevant next step

towards more applied scenarios. Moreover, we assumed that time compression would become more effective with predictive primes. Last but not least, it remained to be investigated whether our crossmodal semantic effects are limited to color perception, or whether they are more general in nature and would extend to other object features or object perception in general. Of course, this would constitute a necessary precondition for the usage of spoken warnings in driving scenarios. Correspondingly, we will address the aforementioned open issues in the next chapters.

5 Crossmodal influences of time-compressed spoken denotations on automotive icon classification and the role of contingency

The preceding experiments revealed detailed insights into the underlying processes of crossmodal presentation. However, an important question still remained unanswered: Do findings of crossmodal Stroop effects that were achieved with simple color stimuli equally apply to more complex objects? For the automotive context it might not only be interesting, whether object features as colors can be primed by spoken words, but also whether entire objects can be primed by their denotations. For this purpose, we switched to icons from the automotive context and their spoken two-syllable denotations in Experiments 3A and 3B. We were interested whether we would achieve equivalent crossmodal effects as in Experiment 1A, even though presenting longer words and more complex visual stimuli.

Furthermore, we reconsidered time compression for the two-syllable words, since words with a longer duration might benefit more from an earlier ending – and hence disambiguation – than short words. If this were the case, we would expect to find duration moderating crossmodal differences between congruent and incongruent trials, with larger differences for time-compressed auditory stimuli. As in Experiments 1A and 1B, we applied three different levels of spoken word duration in Experiments 3A and 3B. Once again, we used an uncompressed version of each word (100 % of the original duration), and words compressed down to 30 % of their original duration. When compressing the two-syllable words down to 10 % of their original duration, it was actually impossible to recognize which denotation they comprised. Therefore, we decided this time to replace the most extreme level of time compression with an intermediate one (50 % of the original duration).

Due to the fact that we found stronger crossmodal semantic enhancement for high-perceptual-load conditions than for low-perceptual-load conditions in Experiments 1A and 1B, we additionally confined the design of Experiments 3A and 3B to solely a high-perceptual-load condition. Also in the light of our applied research question on time-critical warnings, we

analogously assumed that speech warnings would particularly support drivers in highly complex and rather ambiguous situations than in unambiguous cases. Accordingly, our focus on more difficult visual search tasks did not only match previous findings, but also coincided with practical relevance/considerations.

Last but not least, we addressed one major issue in this chapter: How would increased contingency between prime and target influence crossmodal semantic effects? Former studies on crossmodal semantic effects have – more or less deliberately – applied different levels of conditional probability (or validity) between auditory and visual semantics. For crossmodal semantic priming this issue applies to prime and target, whereas for crossmodal Stroop effects it applies to task-irrelevant and task-relevant stimulus dimensions. In some experiments authors have explicitly excluded contingent presentation, and thus applied semantically congruent trials only at chance level in their experiments (Lupyan & Ward, 2013, Exp. 3; Roelofs, 2005; Salverda & Altmann, 2011, Exp. 3; Tellinghuisen & Nowak, 2003). In contrast, in other experiments levels of (considerably) increased semantic contingency between prime and target were used (Chen & Spence, 2011; Lupyan & Spivey, 2010; Lupyan & Thompson-Schill, 2012, Exps. 1 and 2; Lupyan & Ward, 2013; Salverda & Altmann, 2011, Exp. 1). We consider contingency as a highly relevant factor regarding crossmodal semantic effects, since an increased level of contingency would (implicitly) induce participants to pay attention to (otherwise task-irrelevant) semantics, thus leading to improved task performance in congruent and deteriorated task performance in incongruent trials (see, e.g., Logan & Zbrodoff, 1979; Logan, 1980). However, none of the hitherto studies has systematically addressed the influence of contingency on crossmodal semantic priming. For a start, we decided to investigate this issue by slightly increasing the correlation between prime and target from no contingency in Experiment 3A (at chance level: a given prime word is followed by its semantically congruent visual target in 25 % of the cases; alike in Exps. 1A, 1B, and 2), to slight contingency in Experiment 3B (doubled probability: a given prime word is followed by its semantically congruent visual target in 50 % of the cases). In this regard, it was furthermore interesting, whether effects of time compression would additionally be affected by increased contingency between prime and target. Eventually, a main effect of duration on RTs might

be decreased, or might even be reversed (i.e., larger Stroop effects with higher compression) for predictive primes, since participants would try to decode auditory prime information as soon as possible in order to support reactions to visual targets.

5.1 Experiment 3A: Towards more realistic stimuli – Automotive icon targets and two-syllable denotations

Experiment 3A was designed to replicate our crossmodal priming effects from Experiment 1A with more complex auditory primes and visual target objects. Now applying considerably longer, two-syllable spoken prime words, we addressed once more the issue, whether crossmodal semantic effects would depend on the auditory presentation duration. In addition, we were interested whether, in accordance with our previous findings, benefits of congruent primes would outweigh the costs of incongruent primes. The auditory primes did not denote any valid information about subsequent visual targets in Experiment 3A.

5.1.1 Method

Participants. A total of 19 students (16 females, 3 males) from Saarland University took part in Experiment 1. Participants received course credit for taking part. All participants had normal or corrected-to-normal vision, and none of them reported any color-blindness or hearing problems.

Design. A 3 (semantic congruency of auditory prime and target icon: congruent, incongruent, neutral) × 3 (duration of the auditory prime: 100 % [i.e., 700 ms], 50 % [i.e., 350 ms], 30 % [i.e., 210 ms]) design was employed with all factors manipulated within participants. Technically, the congruency factor was realized by a 5 (auditory prime type: traffic light, tractor, ambulance, children, and object) × 4 (visual target type: traffic light, tractor, ambulance, children) design resulting in non-contingent auditory priming. That is, participants had no benefit from listening to the words, since they did not contain any information about the subsequent target. Congruency was manipulated on a trial-by-trial basis, whereas compression level was manipulated blockwise, resulting in a total of three blocks (each of which comprised 100 trials: 20 congruent, 60 incongruent, and 20 neutral trials). The

order of compression level presentation was counterbalanced between participants.

Materials. Each target display contained eight visual stimuli presented in a ring-like arrangement on a white background (see Figure 4). The ring was presented with a visual angle of 18.9° (diameter 400 pixels ≙ 20 cm), with each icon in the ring (40-72 pixels width × 50-68 pixels height) centered on a white squared patch (size 73 × 73 pixels ≙ 3.6 × 3.6 cm) spanning 3.4°. One of these visual stimuli was the target and therefore appeared as one of the four target objects (traffic light, tractor, ambulance, or children). Seven filler icons (motorcycle, stop sign, truck, bicycle, cow, fallen tree, and roadworks; see also Appendix D) were clearly distinguishable from the target icons, and randomly located at the remaining seven non-target positions in each trial.

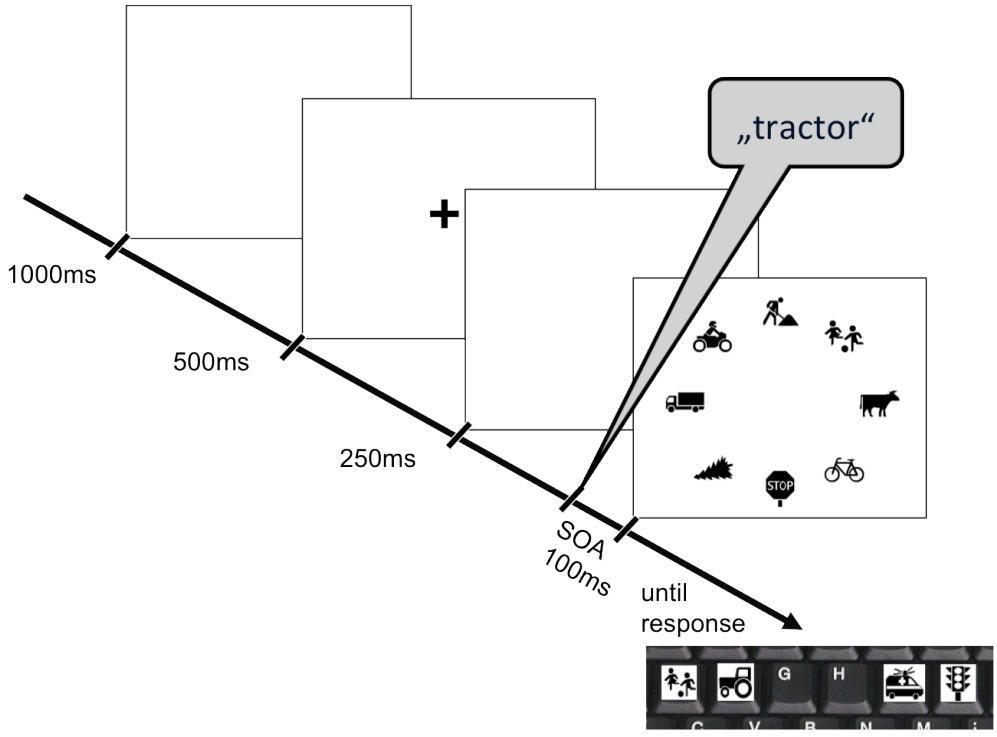


Figure 4. Trial sequence (incongruent trial) from Experiment 3A or 3B. The prime word ("tractor") is presented via headphones 100 ms prior to the visual target (children).

We generated auditory word stimuli by having the same adult male (as for color words) speak the icon denotations several times. Utterances were recorded with the identical equipment as for the preceding experiments. We selected sound files that lasted exactly 700 ms, which seemed to be the average

length of the two-syllable words traffic light (“Ampel”, [ˈampl̩]), tractor (“Traktor”, [ˈtʁaktoːɐ̯]), ambulance (“Notarzt”, [ˈnoːtʔaʁtst]), and children (“Kinder”, [ˈkɪndɐ]). The German word for ‘object’ (“Objekt” ([ɔpˈjɛkt]) served as the neutral prime.

We created two different time-compressed versions of each of the five sound files: One with a length of 350 ms (i.e., compression to 50 % of their original duration), and another one with a length of 210 ms (i.e., compression to 30 % of the original duration). The 50 %-compression files were still understandable, whereas the 30 %-compression files were hardly discriminable in the context of the given task: That is, arbitrary 30 %-compression two-syllable words without any constraining expectations are not identifiable; however, in the present context with the usage of only five known words the categorization is rather easy after hearing samples. Initially, we tried to use compression up to 10 % of the original duration analogous to the preceding experiments. Unfortunately, recognition of these spoken two-syllable denotations was considerably harder than for one-syllable (color) words. Accordingly, we switched to less severe time compression for the *auditory icon* denotations. Time compression was achieved analogous to the color stimuli. Also the auditory presentation of stimuli remained identical, except for a range in loudness from 67 to 72 dB SPL.

Procedure. Lab equipment, viewing distance, et cetera were identical to the preceding experiments. Besides, the whole procedure was identical to Experiment 1A, except for the fact that colors were exchanged with automotive icons (see Figure 4) and only one (high) level of perceptual load was implemented. Participants were informed that the auditory words preceding the target display did not contain any information about the subsequent target. The experiment lasted for approximately 45 minutes. Mean RTs constituted the dependent variable.

5.1.2 Results

Error trials (4.3 %) and outliers (3.4 %; i.e., RTs in correct trials that were 1.5 interquartile ranges above the third quartile or below the first quartile, respectively, with respect to the individual distribution; Tukey, 1977) were discarded. Table 5 shows the mean RTs for the conditions of our design.

Table 5. Mean reaction times (RTs in ms; error rates in parentheses) in Experiment 3A as a function of compression levels and semantic congruency conditions.

		Semantic congruency condition		
		Congruent	Neutral	Incongruent
Contingency	Duration			
No	100 %	1001 (5.0)	1095 (2.1)	1155 (4.4)
	50 %	1041 (5.5)	1140 (5.0)	1168 (3.7)
	30 %	1108 (5.5)	1163 (4.0)	1205 (4.4)

RTs were analyzed in a 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) MANOVA. The analysis revealed a significant main effect for the semantic congruency manipulation, $F(2, 17) = 38.05$, $p < .001$, $\eta_p^2 = .82$. Helmert contrast further showed that neutral trials did not lead to different RTs than congruent and incongruent trials (pooled), $F(1, 18) = 2.29$, $p = .15$, $\eta_p^2 = .11$, and that congruent trials led to considerably faster RTs than incongruent trials, $F(1, 18) = 80.01$, $p < .001$, $\eta_p^2 = .82$ (see also Figure 5a). All three Stroop effects were significantly above zero, ts(18) > 4.62, ps < .001. The interaction of semantic congruency condition and duration was not significant, $F < 1$, as well as the main effect of duration, $F(2, 17) = 1.88$, $p = .18$, $\eta_p^2 = .18$.

In addition, difference scores for facilitation (neutral - congruent) and interference (incongruent - neutral) were calculated for all three duration conditions (see Figure 5b/c). We analyzed both indicators in a repeated measures MANOVA with regard to the duration factor (100 % vs. 30 % vs. 10 %). Overall, striking facilitation significantly above zero was revealed, $F(1, 18) = 26.12$, $p < .001$, $\eta_p^2 = .59$, but we did not find a significant main effect of duration, $F < 1$. Similarly, overall interference was significantly above zero, $F(1, 18) = 11.17$, $p = .004$, $\eta_p^2 = .38$, without revealing a significant main effect of duration, $F < 1$. According to these results and consistent with our findings in the preceding experiments, facilitation appears to be more

pronounced than interference (see also Figure 5b/c), and thereby, in particular, contributing to the Stroop effects.

*Figure 5. Reaction time (RT) differences (Stroop: incongruent trials – congruent trials; facilitation: neutral - congruent; interference: incongruent – neutral) for the three compression rates (duration) and the two contingency levels in Experiments 3A and 3B. The error bars indicate ± 1 SEM (see Franz & Loftus, 2012). + $p < .05$, one-tailed, * $p < .05$, ** $p < .01$, *** $p < .001$.*

For the error rates (see Table 5), we conducted the same analysis as for the RTs. The 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) MANOVA did not show any contradictory results with regard to RTs. Neither the main effect for duration, $F = 2.47$, $p = .12$, $\eta_p^2 = .23$, nor the main effect for semantic congruency, $F = 1.03$, $p = .38$, $\eta_p^2 = .11$, nor the two-way interaction, $F = 1.06$, $p = .41$, $\eta_p^2 = .22$, were significant.

Despite the non-significance of the MANOVA results, we tested the 3 simple Stroop effects (i.e., the comparisons of the incongruent and congruent conditions) for the three duration levels, in order to provide full transparency with regard to potential speed-accuracy trade-offs. None was significant (all $|t|$s(18) < 1.25, ps > .23): Even though all three differences were numerically negative, only null effects were found.

5.1.3 Discussion

We investigated whether spoken object denotations – even without revealing any contingent information – influence categorization of automotive target icons. Our results clearly support this assumption by the significant and remarkably large difference in RTs between congruent and incongruent trials in all experimental conditions. Another remarkable point is the repeated pattern of cost-benefit partitioning of the Stroop effect. In accordance with our former findings, large benefits and only minor interference were revealed. Furthermore, analogous to Experiments 1A and 1B no main effect of duration on RTs was identified. With regard to time-compression, it is interesting that we found clear crossmodal effects of semantic congruency independent of the duration of spoken word primes. Thereby, our assumption of increased effects for highly time-compressed spoken words was not confirmed. Accordingly, for non-contingent presentation of spoken word stimuli time-compression seems to be neither a beneficial, nor a detrimental factor.

To sum up, we replicated crossmodal semantic Stroop/priming effects even with more complex materials, namely two-syllable spoken words as auditory primes and automotive icons as visual targets. Initially, we assumed the potential confound of spoken word presentation by its inherent duration to especially affect longer utterances. We addressed this issue by applying two reasonable time compression levels to the irrelevant primes, but no difference in effects was revealed. Overall, our results are promising since they indicate that crossmodal semantic priming is effective even for more complex objects and their denotations. This constitutes a major step towards critical real-world road scenarios.

The preceding experiments, in which primes did not contain any valid information about subsequent targets, were extremely relevant, since this

allowed for a highly controlled identification of the basic, underlying crossmodal effects. By keeping contingency at chance level, any (attentional) strategies or implicit learning could be prevented. Thus, we were able to demonstrate instantaneous and irrepressive crossmodal effects of semantics. However, another important aspect regarding our applied research question remains to be addressed: Warning systems would not present non-informative semantics that only match the relevant object or situation incidentally. Instead, since drivers would only accept meaningful warnings, respective systems would only be introduced, if they were able to present information matching the actual hazard. Accordingly, we were also interested, how higher levels of semantic contingency would additionally affect these effects. Therefore, we increased contingency between prime and target in the next Experiment 3B.

5.2 Experiment 3B: Contingent presentation of two-syllable denotations of automotive icon targets

In case of meaningful information presentation, drivers would most probably try to make use of semantics or even rely on a warning system. Thus, implicit learning or strategic effects would come into play. However, it still remained to be investigated whether and how an increased rate of semantically congruent trials would affect crossmodal semantic priming. Strategic usage of contingent semantic information might lead to pronounced benefits of congruent primes on the one hand, and to increased costs of incongruent primes on the other hand. Participants might no longer try to ignore the auditory primes as in our hitherto experiments, but (implicitly) try to benefit from the spoken words. The other side of the coin, however, might be that in incongruent trials auditory primes might interfere with visual task performance. Besides, time compression of spoken words might affect response times. So far, moderate time-compression has not turned out to be a relevant factor regarding crossmodal effects. However, for increased levels of contingency the situation might change: Since the meaning of time-compressed auditory stimuli is disclosed earlier this might additionally speed up responses – above all in congruent cases.

As a first approach to the issue of semantic contingency, we adapted Experiment 3A by switching from semantically uncorrelated primes and targets

to contingent crossmodal presentation in Experiment 3B. For a start, we decided to just slightly increase the probability for a congruent target following a given prime from 25 % (chance level) to 50 %. Accordingly, after a given auditory prime word a congruent target icon was presented in half of the cases. In the remaining cases one out of the remaining three target icons appeared. In the following, we initially present Experiment 3B, before turning to the comparison of Experiment 3A with 3B regarding the contingency manipulation.

5.2.1 Method

Participants. A new group of 19 undergraduates (13 females, 6 males) from Saarland University took part to fulfill course requirements. All had normal or corrected-to-normal vision and none reported any known color-blindness or hearing problems.

Design. A 3 (semantic congruency of auditory prime and target color: congruent, incongruent, neutral) × 3 (duration of the auditory prime: 100 % [i.e., 700 ms], 50 % [i.e., 350 ms], 30 % [i.e., 210 ms]) design was again employed with all factors manipulated within participants. Presentation of trials and blocks was equivalent to Experiment 3A, except that we changed the semantic contingency between prime and target. Each of the three experimental blocks consisted of 100 trials composed of 20 neutral, 40 congruent and 40 incongruent trials, randomly intermixed. For the 40 congruent trials, the list was completely balanced (each target object was presented ten times with its respective prime). Also for the 40 incongruent trials, each target object was presented ten times. Here, we presented each of the three incongruent prime words three times plus in one trial a designated incongruent prime word was chosen. Altogether, each incongruent prime word was presented ten times. In the neutral condition each target object was used five times.

Materials. Materials were identical to Experiment 3A.

Procedure. The procedure and the participants' task were identical to Experiment 3A.

5.2.2 Results

One participant was discarded from analysis because of an error rate of 51 %. Prior to aggregation, we discarded error trials (5.5 %) and outliers (4.7 %; see Exp. 3A). Table 6 shows the mean RTs and error rates for all conditions.

Table 6. Mean reaction times (RTs in ms; error rates in parentheses) in Experiment 3B as a function of compression levels and semantic congruency conditions.

		Semantic congruency condition		
		Congruent	Neutral	Incongruent
Contingency	Duration			
	100 %	936 (5.1)	1093 (3.3)	1109 (6.0)
Yes/Slight	50 %	943 (5.8)	1056 (7.5)	1102 (4.9)
	30 %	902 (5.8)	998 (5.6)	1070 (5.1)

RTs were analyzed in a 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) MANOVA. The analysis revealed a significant main effect for the semantic congruency manipulation, $F(2, 16) = 35.63$, $p < .001$, $\eta_p^2 = .82$. Helmert contrasts further showed that neutral trials differed from congruent and incongruent trials (pooled), $F(1, 17) = 25.22$, $p < .001$, $\eta_p^2 = .60$. Most importantly, congruent trials again led to faster RTs than incongruent trials, $F(1, 17) = 64.88$, $p < .001$, $\eta_p^2 = .79$. The main effect of duration was not significant, $F(2, 16) = 2.09$, $p = .16$, $\eta_p^2 = .21$, as well as the interaction of semantic congruency and duration, $F < 1$.

In addition, difference scores for facilitation (neutral - congruent) and interference (incongruent - neutral) were calculated for all three duration conditions (see Figure 5). A MANOVA with respect to the duration factor (100 % vs. 30 % vs. 10 %) was conducted for both indicators. Overall, significant facilitation above zero was revealed, $F(1, 17) = 73.50$, $p < .001$, $\eta_p^2 = .82$, but no significant main effect of duration was found $F(2, 16) = 1.47$, $p = .26$, $\eta_p^2 =$

.16. Significant interference above zero was revealed overall, as well, $F(1, 17) = 14.91$, $p = .001$, $\eta_p^2 = .47$, but again no significant main effect of duration was found, $F(2, 16) = 1.36$, $p = .29$, $\eta_p^2 = .15$. These results correspond to the findings in Experiment 3A: Effects of facilitation appear to be more pronounced than those of interference (see also Figure 5b/c).

For the error rates (see Table 6), we conducted the same analysis as for the RTs. The 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) MANOVA did not show any contradictory results with regard to RTs. Neither the main effects, all $Fs < 1.56$, $ps > .24$, η_p^2s $< .16$, nor the two-way interaction, $F = 1.44$, $p = .27$, $\eta_p^2 = .29$, were significant.

Despite the non-significance of the MANOVA results, we tested the 3 simple Stroop effects (i.e., the comparisons of the incongruent and congruent conditions) for the three duration levels, in order to provide full transparency with regard to potential speed-accuracy trade-offs. None was significant (all $|t|$s$(17) < 1$): Two differences were numerically negative, but only clear null effects were revealed.

5.2.3 Discussion

With a slightly increased contingency between speech primes and visual targets, we again found significant and remarkably high difference in RTs between congruent, neutral and incongruent trials in all experimental conditions of Experiment 3B. Duration of the primes did, once again, not significantly influence RTs. Even though, at a first glance, the pattern of semantic congruency conditions and time compression seemed to be quite similar to Experiment 3A, it was worthwhile to directly compare results from both experiments regarding the contingency manipulation.

5.3 Highly compressed auditory primes: Contingency matters

Besides the manipulation of contingency, Experiment 3A and Experiment 3B were designed equivalently with participants randomly assigned to the two contingency conditions. Hence, we were able to additionally analyze effects between participants.

Figure 6 shows a combination of the mean RTs for all conditions in Experiments 3A and 3B. RTs were analyzed in a 2 (contingency: No/Exp. 3A vs. yes/Exp. 3B) × 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) mixed model MANOVA. In order to avoid redundancy with regard to the preceding results on this data, and in order to keep the analyses tight, we only report the effects involving the contingency factor in the following paragraph.

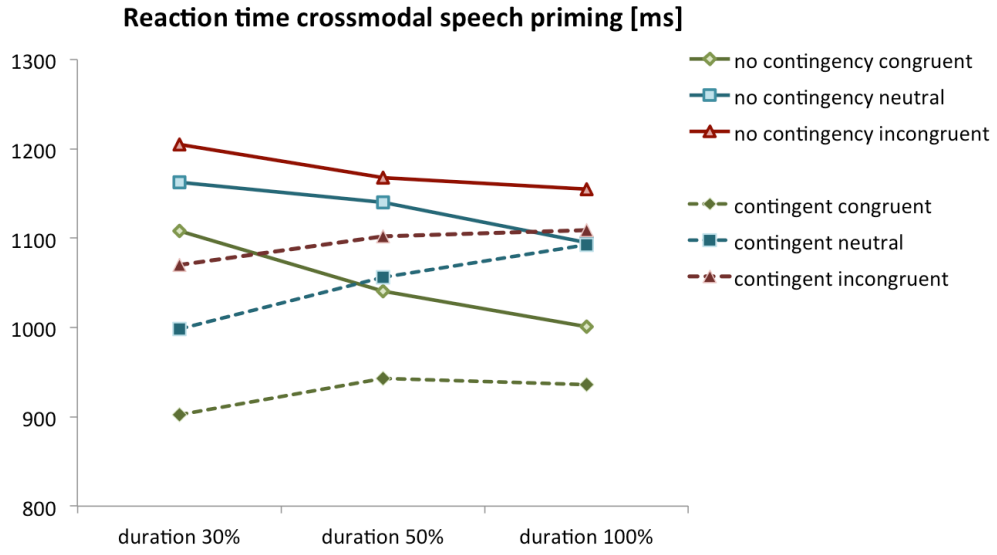


Figure 6. Reaction times (RTs) for the three compression rates (duration) and the two contingency levels (no contingency vs. slight contingency) in Experiments 3A and 3B.

The analysis revealed neither a significant main effect for contingency, $F(1, 35) = 2.34$, $p = .14$, $\eta_p^2 = .14$, nor a significant three-way interaction, $F < 1$. The interaction between contingency and semantic congruency was not significant, as well, $F(2, 34) = 1.72$, $p = .20$, $\eta_p^2 = .09$. Accordingly, also reporting further analyses on contingency affecting facilitation or interference were dispensable. By contrast, a significant interaction between duration and contingency was revealed, $F(2, 34) = 3.72$, $p = .03$, $\eta_p^2 = .18$. Therefore, we had a closer look at the Helmert contrasts. No difference was found for this interaction for the comparison of medium compression (50 % duration) and both other duration conditions (pooled), $F < 1$. However, for the contrast of the most extreme conditions – the contrast between the uncompressed (100 % duration) and the highly compressed (30 % duration) primes – the interaction

revealed a significant interaction of duration and contingency, $F(1, 35) = 6.52$, $p = .02$, $\eta_p^2 = .16$. Accordingly, contingency has a different effect on RTs for uncompressed versus highly compressed speech primes. For uncompressed word primes (100 % duration), RTs between contingent and non-contingent presentation did not differ significantly (mean: 1046 ms vs. 1084 ms), $t(35) < 1$, whereas for highly compressed word primes (30 % duration), RTs were significantly faster for contingent (mean: 990 ms) than for non-contingent presentation (mean: 1159 ms), $t(35) = 2.39$, $p = .02$.

For the error rates (see Table 5 and Table 6), we conducted the same analysis as for the RTs. The 2 (contingency: No/Exp. 3A vs. yes/Exp. 3B) × 3 (duration: 100 % vs. 50 % vs. 30 %) × 3 (Semantic congruency: neutral vs. congruent vs. incongruent) mixed model MANOVA did not show any contradictory results with regard to RTs. Neither the three-way interaction, $F < 1$, nor the two-way interactions of semantic congruency and contingency and of duration and contingency, both $Fs < 1$, nor the main effect of contingency, $F < 1$, were significant.

In sum, we achieved similar crossmodal effects for non-contingent and slightly contingent primes. Slightly informative primes did neither increase interference, nor lead to a reversed pattern with minor facilitation and larger interference. Interestingly, highly time-compressed speech prime trials led to shorter RTs when they were slightly contingent than when prime and target were uncorrelated. For uncompressed spoken word primes, in contrast, we did not find an effect of contingency on RTs. Accordingly, for highly time-compressed auditory primes it makes a difference whether they comprise the potential to support target detection: If this is the case, RTs are faster than for completely irrelevant and unhelpful spoken words. The reason might be that in the former case participants (implicitly) tried to make use even of hardly understandable primes – since slight overall benefits could be expected. Under these circumstances the earlier disclosure of a time-compressed word's meaning might become effective and speed average response times.

When summing up the results for Experiments 3A and 3B, it must not be overlooked that even under most adverse or unexpectable conditions – with no potential support for target detection by spoken words (no contingency) and highly compressed primes – we found clear congruency effects on RTs.

Transferred to time-critical situations, our findings from the preceding experiments suggest that there is no need for time-compressed presentation of spoken words. However, as soon as (even minor) contingency between warnings and the subsequent situations comes into play, time-compression should be reconsidered as a promising solution. We assume that drivers would especially benefit from (time-compressed) speech warnings if they were obviously suitable and useful. Both technological advances and customer acceptance (as a prerequisite for the admittance of new warnings) will certainly lead to particularly high contingency levels for warnings in applied driving scenarios. Therefore it remains to be investigated how considerably higher contingency levels than the one employed in Experiment 3B would affect crossmodal semantic priming effects, and whether speech duration would play an important role under such more extreme circumstances. Accordingly, we will return to the contingency issue later in Chapter 6.

5.4 Interim summary of experiments

At this point, a brief summary regarding the preceding five experiments seems to be valuable. Clear crossmodal congruency effects of spoken words on RTs were repeatedly found – even if stimulus-response compatibilities were ruled out. So far, we consistently found pronounced facilitation and only minor interference. Besides, result patterns are comparable for simple color stimuli and for more complex automotive icons. A reasonable interpretation of our findings would be that spoken semantics were able to pre-activate associated visual features via semantic representations. As long as recognition of prime words is not affected, time-compressed words seem to cause comparable effects as words with normal duration. Besides, for highly time-compressed stimuli already slightly increased contingency speeds RTs as compared with non-contingent semantics. This becomes especially obvious in congruent trials. We expect that this interaction of duration and contingency would become more obvious if contingency was further increased, if speech messages were longer, or if there were more diverse speech warnings. In these cases participants might especially benefit from early disambiguation of semantics.

So far, many relevant issues have already been addressed in the preceding experiments. By approaching the relevant applied question in a

stepwise and very controlled way, we could not only reveal effects of spoken words on a visual task, but we were also able to identify underlying processes accounting for those crossmodal effects. Accordingly, we are in the position to claim that crossmodal semantic effects can occur extremely fast, without any involvement of implicit learning of correlations, and without strategic attentional processes. If one considers that speech comprehension is generally characterized as a complex process involving a mixture of top-down and bottom-up processes (Eysenck & Keane, 2000), it is highly astonishing to find those fast and irrepressive effects of spoken words on visual target identification/classification. This forms a promising basis for the effective application of speech warnings. In case of increased contingency, and hence usefulness of spoken word primes, strategic effects might even add on top of rather automatic semantic effects.

When transferring our findings from basic computer experiments to real-world scenarios, ecological validity needs to be considered as another major challenge: Are crossmodal semantic effects restricted to a highly controlled single-task lab environment, or are they pronounced and robust enough to affect performance under driving conditions? In the next chapter we approach this issue by using a driving stimulation environment and even complex driving maneuvers indicating visual target classification. In the light of driving safety, this allowed for the investigation of crossmodal semantic priming effects under dual task and optical flow conditions.

6 Time-compressed spoken word primes crossmodally improve the classification of semantically congruent road signs and associated driving performance

In this chapter, we address several further issues regarding the transfer of our previous findings to more dynamic real-world scenarios. Nevertheless, we decided to retain our stepwise and deliberate strategy regarding changes in materials or procedure in order to have maximum control over respective effects. Hence, we did not directly investigate effects of speech warnings in completely realistic road hazard situations. In such a case, we would suppose that numerous factors like, for example, interindividual differences in vigilance and situation awareness, differences in the foreseeability and urgency of critical incidents, as well as several different adequate driving responses, would severely affect our effects and their measurability. Considerable uncontrolled noise might conceal basic effectiveness of semantic priming and other independent variables. Therefore, we specifically designed and implemented a driving task that maintained the typical trial-based structure of our former computer experiments. This enabled us to leave many experimental parameters unchanged, while at the same time investigating whether crossmodal effects of spoken words could be replicated during a dynamic driving task comprising continuous visual and motor demands. Moreover, in Experiments 4, 5A, and 5B we returned to the separation of crossmodal stimulus-stimulus and stimulus-response compatibilities: This time, merely the color of an automotive target icon should be classified, thus keeping responses orthogonal to the semantic identity of primes and targets. Similar to Experiment 2, this excluded response-priming effects and we were able to derive more precise insights about underlying processes. Besides, we altered the response type from simple key press in Experiment 4 to demanding driving maneuvers in Experiments 5A and 5B: Depending on the target color, participants had to either brake or switch to a designated lane. The introduction of such complex motor responses constituted another controlled step towards in-car warning scenarios, in which

drivers might have to respond appropriately to imminent road hazards. In some cases such a response might be (emergency) braking, whereas in other cases it might rather be swerving. Besides, sometimes it might of course be sufficient to observe a potentially critical item for a while, thus improving anticipation of sudden, critical events. In any case, it was highly interesting whether crossmodal semantic benefits and/or costs could still be replicated when assessing complex driving maneuvers. Only if driving performance enhancement by congruent speech warnings could be confirmed, the claim that our previous findings are relevant for critical road incidents could be maintained.

Last but not least, we addressed the issue of contingent information presentation once more, since we wanted to investigate whether and how strategic effects based on semantic relations might additionally affect semantic priming effects. However, this time we even compared highly contingent with non-contingent auditory primes (Exps. 5A and 5B).

6.1 Experiment 4: Spoken word primes improve the classification of semantically congruent visual targets in an on-road scenario

For the new driving scenario, we transferred the target detection task to gantry road signs that repeatedly occurred while driving. In a first step, we recorded responses upon targets simply via steering wheel buttons. This allowed for retaining our hitherto well-examined procedure and materials, while at the same time switching to dual task and optical flow conditions. In order to separate stimulus-response compatibility effects and stimulus-stimulus compatibility effects (see also Exp. 2), we switched from target identification to target feature identification: Instead of the icon identification response (one out of four) which might be directly influenced by the corresponding spoken word, participants had to indicate the color of the present target icon (one out of two). By using a response independent of the target denotation, we were able to explore effects of spoken word stimuli on visual perception within a RT paradigm, while controlling for influences from stimulus-response compatibilites. When considering this simple target feature classification under an applied perspective, parallels to assessing the criticality of an object or situation might be drawn. For Experiment 4, we decided to use a time

compression level of 50 % of the original duration for the auditory primes. This corresponds to the medium duration of auditory primes in Experiments 3A and 3B, with primes remaining well intelligible.

6.1.1 Method

Participants. The participants of Experiment 4 were a new group of 25 students (13 females, 12 males) from Saarland University, who were payed 8 euros. All had normal or corrected-to-normal vision and none reported color-blindness or hearing problems. All participants possessed a valid driver's license for at least two years.

Design. Experiment 4 employed a within-subjects design. The independent variable semantic congruency (congruent vs. incongruent vs. neutral [featuring either a control prime word or silence]) was technically realized by a 5 (auditory prime type: traffic light, tractor, ambulance, children, and a control stimulus [prime word object or silence]) × 4 (visual target type: traffic light, tractor, ambulance, children) design. The conditional probability for a target to appear after a given prime was at chance level, resulting in non-contingent auditory priming. That is, participants had again no benefit from listening to the words. Congruency was manipulated on a trial-by-trial basis. Each experimental track comprised 160 trials with a target icon present: 32 congruent trials, 96 incongruent trials, and 32 neutral trials. The neutral trials consisted of 16 trials with a control word ("Objekt" [object] as in Exps. 3A and 3B), and of 16 trials with silence instead of a prime word. In twelve additional catch trials (7.0 % of all trials) no target icon was present and each prime type (four target icon denotations, neutral word, silence) was presented twice. Overall, each track comprised 172 randomly intermixed trials (see also Table 7) – except for target icons not being repeated from one trial to the next. RT of target color classification (red or green) constituted the dependent variable in this experiment.

Table 7. Overview of the 172 trials presented in Experiment 4 and 5A. An auditory prime and a target icon specified a trial. The color of all target icons was counterbalanced, and auditory primes did not reveal any information about subsequent target (no contingency).

	Target icon									
	Ampel (traffic light)		Kinder (children)		Notarzt (ambulance)		Traktor (tractor)		No target	Sum
Auditory prime	red	green	red	green	red	green	red	green		
Ampel	4	4	4	4	4	4	4	4	2	34
Kinder	4	4	4	4	4	4	4	4	2	34
Notarzt	4	4	4	4	4	4	4	4	2	34
Traktor	4	4	4	4	4	4	4	4	2	34
Neutral control	2	2	2	2	2	2	2	2	2	18
Neutral silence	2	2	2	2	2	2	2	2	2	18
Sum	20	20	20	20	20	20	20	20	12	172

Materials. The experimental track comprised a straight road (width 18.5 m) with five lanes next to each other (width 3.7 meters [m] each). With a distance of 210 m gantry road signs (height 7.3 m) were positioned (see Figure 7 and Figure 8). According to the number of trials (172) the track was about 37 km long. A start and a goal symbol signalized beginning and end of data recording. Each gantry road sign (width above street 18.5 m) contained square displays (2.5 m × 2.5 m) referring to one lane each with horizontal spacing of 1.2 m in between. Except for the middle lane, one display was presented for each lane resulting in a total of four displays.

Each of the displays showed one luminous automotive icon from the same set of targets and distractors as in Experiments 3A and 3B (see also Appendix E). Each icon could be presented in red or green, with always two icons being red and two being green, in order to prevent participants from using strategies for target color identification. Either one of the four target icons was presented at one of the four displays of the gantry road sign, or no target stimulus was presented at all (catch trial). In each trial, the remaining three (target-present condition) or four (target absent condition) icons were chosen from the set of seven filler items, which were also the same as in Experiments 3A and 3B (see also Appendix E). Filler items were randomly placed on the spare positions.

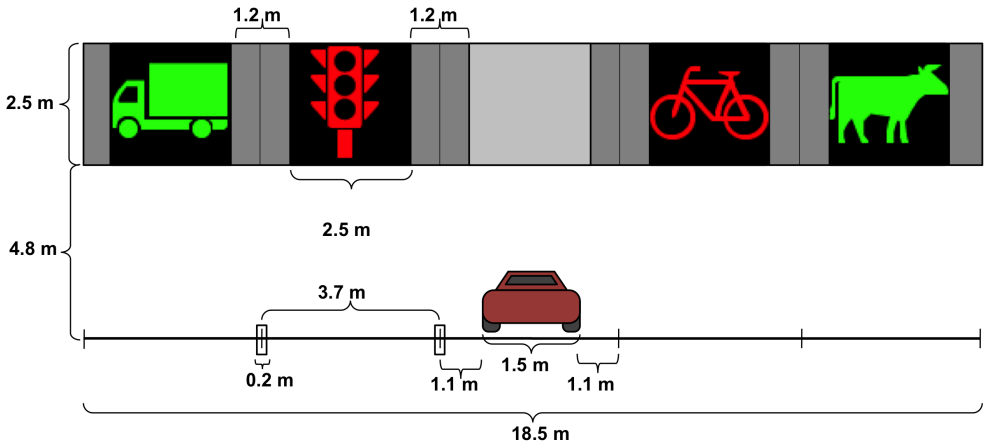


Figure 7. Schematic cross-section of the three-dimensional simulation environment in Experiments 4, 5A and 5B including measurements of the car, the traffic way, and the gantry road signs.

Auditory prime stimuli were identical to the medium compression rate (50 %) used in Experiments 3A and 3B. In addition, a neutral silence condition was included in order to assess whether any auditory presentation impairs performance. In congruent trials, auditory primes matched the target stimulus, in incongruent trials they differed.

The experiment was conducted and data was collected using the open source driving simulation software OpenDS 1.0 (Math, Mahr, Moniri, & Müller, 2012).[18] In order to generate the different tracks for each participant according to our requirements, we used the adjunct tool GenXDS, before feeding the respective files into OpenDS.

Procedure. Participants were tested individually in an experiment room with modest daylight. They were seated in front of a 19-inch monitor (60 Hz refresh rate, resolution 1280 × 1024 pixels) controlled by a personal computer. The participants' viewing distance was about 70 cm and they wore closed-ear Sennheiser PC 151 headphones. The auditory stimuli ranged in loudness from 70 to 72 dB SPL. A MiMo game steering wheel and pedals were used to control the vehicle within the simulation.

Throughout the experimental tracks participants were constantly driving at a speed of 100 kilometers per hour (km/h) leading to time intervals of about 7.6 seconds between two gantry road signs. Their task was to continuously

[18] Available on the website http://www.opends.eu/

keep the center of the middle lane and to react upon target icons appearing on the gantry road signs (see Figure 8 for a screenshot). 60 m before each gantry road sign the icons were displayed and remained visible until participants passed under the sign (about 2.2 s). For each encounter of a gantry road sign, participants had to decide by key press whether a red target object ("target red"-button on the steering wheel with right thumb) or a green target object ("target green"-button on the steering wheel with left thumb) was presented at one of the four locations of the gantry road sign. No semantic categorization of the perceived target icon was required. In target-absent trials no answer should be given.

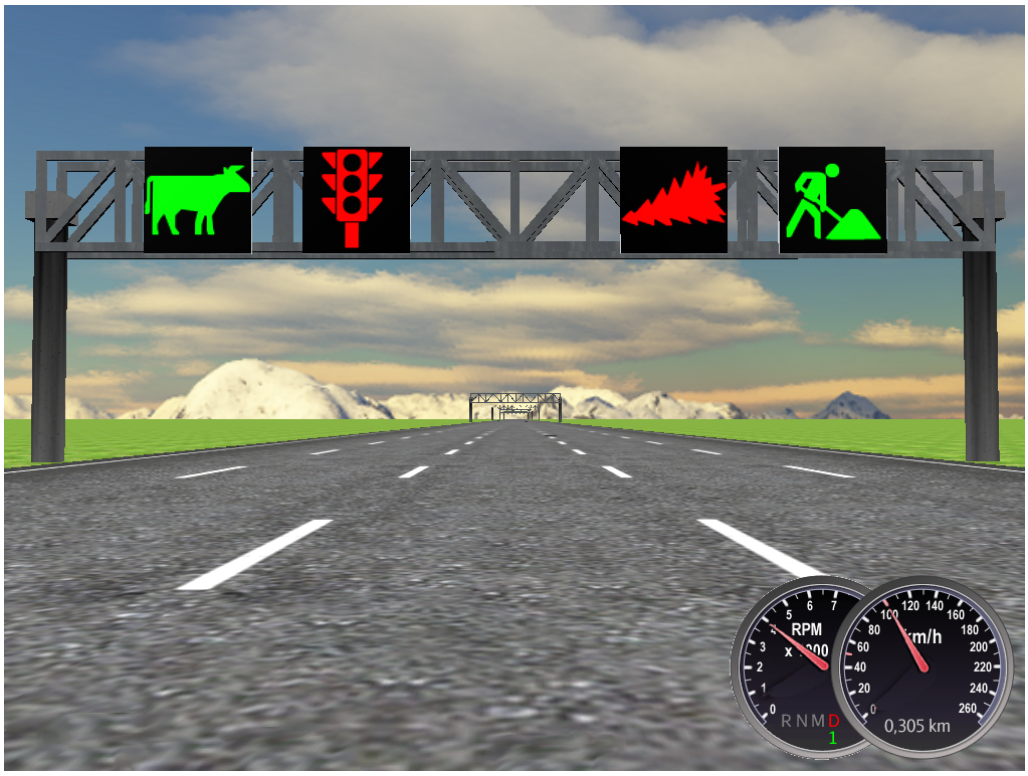


Figure 8. Screenshot of the track and an overhead sign gantry recorded during a simulation run (Exp. 4).

Participants could respond from the onset of the target display onwards until the next trial started (maximum response time: 7.5 s). Auditory primes were presented just before the visual target onset with an SOA of 100 ms (see Figure 9 for a schematic overview of trials aligned within a track). The presentation duration for the auditory primes was 350 ms (i.e., 50 % compression). Participants were informed that the auditory words preceding the target display were non-informative.

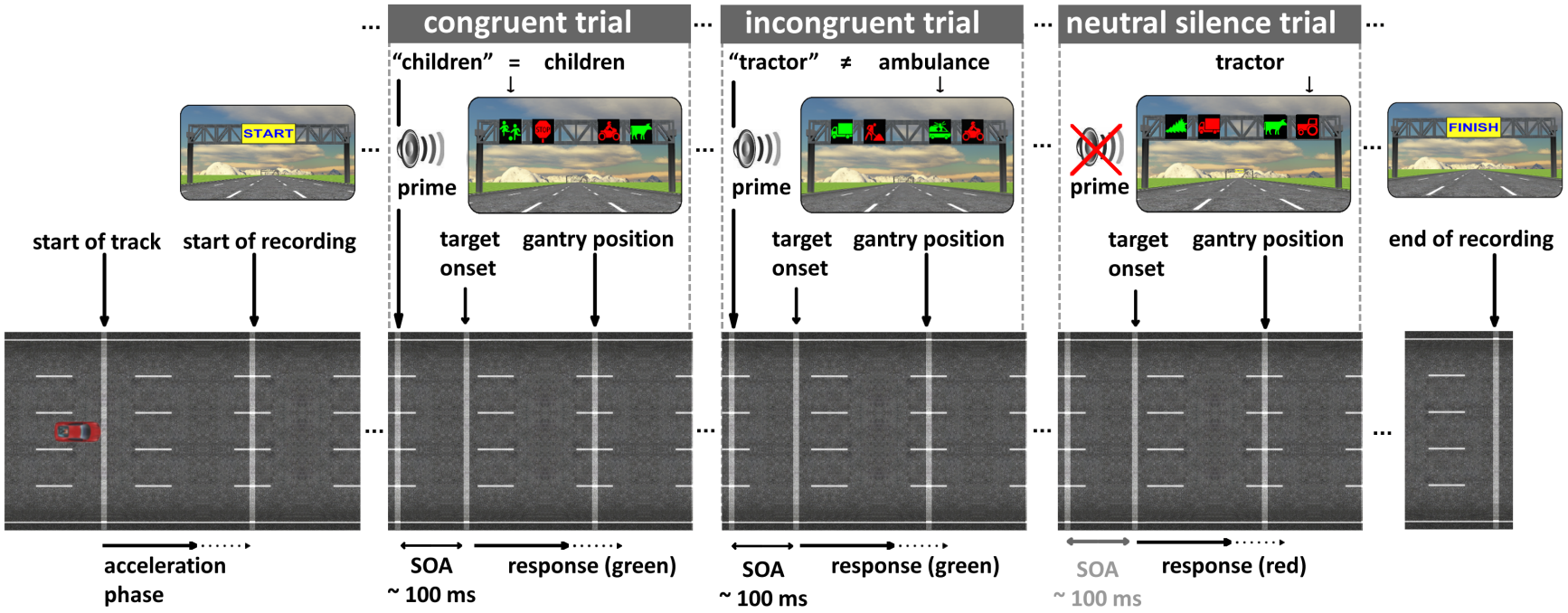


Figure 9. Schematic overview of a track built for Experiment 4 including examples for a congruent, an incongruent, and a neutral silence trial. Each trial sequence comprised an overhead gantry road sign showing four automotive icons. Depending on the color of the target icon, the participants' task was to press either the left (red) or the right (green) key on the steering wheel, or not to press any key (none of the four target icons present). An auditory prime was presented via headphones 100 ms prior to the onset of the automotive icons on the overhead gantry road signs, and could be semantically congruent, incongruent, or neutral (control word or silence), with regard to the target icon.

As an introduction to the experimental task, participants completed an initial practice phase to familiarize with the simulator. They drove on a straight road without road signs present. Afterwards, participants were given the opportunity to familiarize with the four target icons (each in red and green) presented on a static screen. Participants could watch and memorize the icons as long as they preferred and they were told to carefully keep the icons in mind. Subsequently, a short practice track containing 20 sample trials had to be performed. This short practice track was composed equivalently to the subsequent main experimental track and accordingly all auditory prime types were included. Besides, it also did not comprise any contingency.

The main experimental track comprised 172 and was designed not to reveal any contingency between auditory prime and visual target. Besides, each track started with four additional warm-up trials which were not included in the analyses. The experiment lasted for approximately 40 minutes.

6.1.2 Results

For the following evaluation catch trials without target were not included. Prior to aggregation, we discarded error trials (3.5 % incorrect reaction, 0.4 % no reaction) and individual outliers (2.7 %; Tukey, 1977). Table 8 shows mean RTs and error rates for all conditions.

Table 8. Mean reaction times (in ms; error rates in parentheses) in Experiment 4 as a function of semantic congruency conditions.

	Semantic congruency condition			
	Congruent	Neutral control	Neutral silence	Incongruent
Contingency				
No	1015 (3.0)	1074 (2.8)	1079 (4.3)	1078 (4.3)

The RTs for the four semantic congruency conditions (Semantic congruency of auditory prime and target object: congruent vs. neutral control vs. neutral silence vs. incongruent) were analyzed in a repeated measures MANOVA. The analysis revealed a significant main effect of the semantic congruency

manipulation, $F(3, 22) = 19.25$, $p < .001$, $\eta_p^2 = .72$. Consistent with Experiment 2, Helmert contrasts further showed that congruent trials led to shorter response times than all other trial types (see also Table 8), $F(1, 24) = 49.93$, $p < .001$, $\eta_p^2 = .68$. Incongruent trials did not lead to longer RTs than both neutral trial types, and neutral trials also did not differ from each other, both Fs < 1.

In addition, difference scores for facilitation (RT neutral control - RT congruent) and interference (RT incongruent – RT neutral control) were calculated (see also Figure 10). This allows separating the factors of semantic priming effects and helps to understand how the neutral control condition differs from the two relevant prime word conditions. Facilitation significantly differed from zero, $t(24) = 4.69$, $p < .001$, whereas interference did not, $t(24) < 1$.

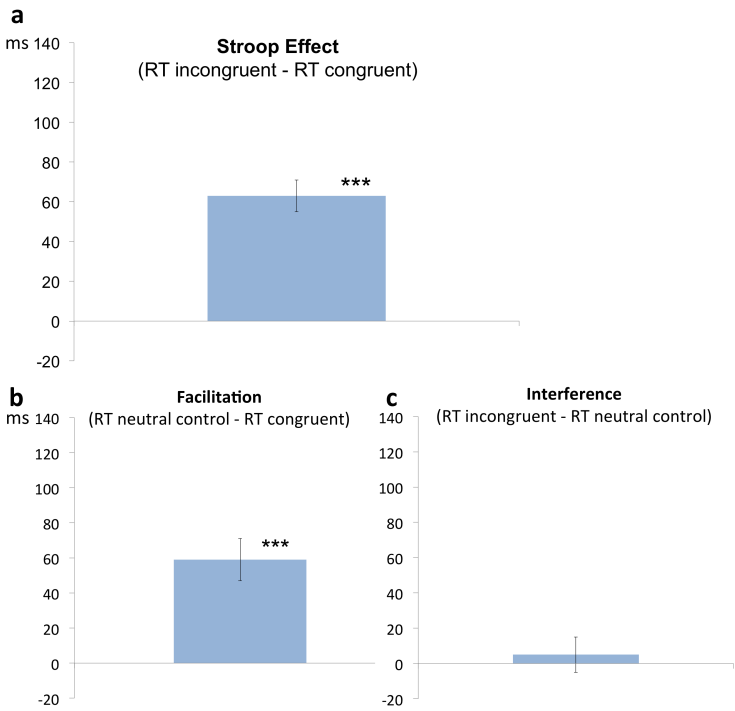


*Figure 10. Reaction time (RT) differences (Stroop: incongruent trials – congruent trials; facilitation: neutral - congruent; interference: incongruent – neutral) for Experiment 4. The error bars indicate ± 1 SEM (see Franz & Loftus, 2012). * p < .05, ** p < .01, *** p < .001.*

For the error rates (see Table 8, incorrect reaction or missing reaction), we conducted the same analysis as for the RTs. A repeated measures MANOVA (Semantic congruency: congruent vs. neutral control vs. neutral silence vs.

incongruent) revealed no significant overall effect of semantic congruency, $F(3, 22) = 1.99$, $p = .15$, $\eta_p^2 = .21$. Besides, the positive difference between the error rates in incongruent and congruent trials contradicts a potential speed-accuracy trade-off.

6.1.3 Discussion

When presenting time-compressed spoken denotations of objects just before their appearance, we could find faster responses in trials with a congruent crossmodal semantic presentation than for neutral (control word or silence) or incongruent trials. This replication of our crossmodal semantic priming effects is highly interesting. First of all, because the effects seem to be robust and pronounced enough to occur even under dynamic driving conditions. Secondly, since we controlled for stimulus-response compatibility effects in this experiment, our results imply that semantic priming and not response priming seems to be the major cause for our crossmodal effects. In Experiments 1A, 1B, 3A, and 3B, we did not control for response priming effects, thus not being able to separate whether the finings were mainly based on semantic priming (stimulus-stimulus compatibilities) or rather on response priming (stimulus-response compatibilities). Even though we already found out in Experiment 2 that semantic priming enhanced detection sensitivity for semantically related targets, with Experiment 4 we could confirm clear crossmodal semantic facilitation of visual processing in a RT paradigm, as well. Interestingly, corresponding result patterns were revealed for Experiments 2 and 4 – even though completely different paradigms and materials were applied.

As a next step towards ecological validity, we decided to switch to more complex responses in Experiments 5A and 5B: Instead of simply pressing one out of two keys for target classification we introduced more complex driving maneuvers. Participants had to either brake hard or change to a designated lane immediately. For both driving maneuvers required in Experiments 5A and 5B, we decided to log several performance indicators. The onset of the automotive icon stimuli on each gantry road sign designated the start point for each response time measure. We were primarily interested in the earliest assessable performance measure for both maneuver types, since during later in the process random noise might increase considerably and mask effects.

For braking the first measure was the time until the gas pedal was released, whereas for lane changes it was the time until the steering wheel was turned three degrees from the straight position. For further braking and lane change metrics that we have logged additionally, please see Appendix F. Overall, we were interested to find out, whether crossmodal semantic benefits could still be identified when evaluating performance in driving maneuvers with increased motor demand.

6.2 Experiment 5A: Spoken word primes crossmodally enhance driving performance

Basically, we reused the materials from Experiment 4 in Experiment 5A. However, since conducting the driving maneuvers took considerably more time than simple key presses, some adaptations regarding the task and the track were necessary: We chose to slightly reduce the instructed velocity, and to increase the distances between the gantry road signs, thereby leaving enough time for drivers to complete the maneuvers and to speed up or to return to the middle lane before the next trial started.

6.2.1 Method

Participants. The participants of Experiment 5A were a new group of 24 students (16 females, 7 males) from Saarland University, who were payed 8 euros. All had normal or corrected-to-normal vision and none reported color-blindness or hearing problems. All participants possessed a valid driver's license for at least two years.

Design. In Experiment 5A, semantic congruency (congruent vs. incongruent vs. neutral [featuring either a control prime word or silence] was manipulated exactly like in Experiment 4. Target color classification (red or green) enabled drivers to choose the respective maneuver: Changing to the target icon lane or braking. RTs based on driving performance in lane change (until a steering wheel turning angle of 3 degrees from the straight position was exceeded) or braking (Time until gas pedal was released) maneuvers constituted the dependent variable in this experiment.

Materials. The track used for Experiment 5A was built analogous to the experimental track in the preceding Experiment 4. Hence, it did not comprise

any contingency between auditory prime word and the target icon. The duration of the auditory primes (350 ms, i.e., 50 % compression) and the SOA of auditory prime and visual target (100 m) were kept constant. The only slight adaptation for the tracks was the distance between gantry road signs, which was slightly reduced to 200 m. Accordingly, the entire track with 172 trials was about 35 km long.

Procedure. Testing facility and hardware were identical to Experiment 4, except for a Logitech Driving Force GT steering wheel and the respective pedals for throttle and brake. This allowed for a better haptic interaction during the driving maneuvers. The general driving speed was reduced to a maximum of 60 km/h, in order to allow more time for driving maneuvers between gantry road signs. Throughout the experimental tracks participants should constantly drive at the maximum speed by keeping the gas pedal pressed, resulting in a time window of about twelve seconds between two gantry road signs. The driving task was to continuously keep the center of the middle lane and to react upon target icons appearing on the gantry road signs with designated driving maneuvers. About 48.3 m before each gantry road sign the icons were displayed and remained visible until participants passed under the sign (minimum 2.9 s). For each gantry road sign encounter, participants had to decide whether a red target icon or a green target icon was presented at one of the four locations of the gantry road sign. As in Experiment 4, no semantic categorization of the target icon was required. In cases where the target icon was red, participants should immediately brake and decelerate sharply to a target speed of 20 km/h or lower. They were told not to leave the middle lane in this case. Here, we were interested in the RTs from target onset until gas pedal release. In case of a green target icon, they should keep their speed and switch to the respective lane (over which the icon was presented). Accordingly, lane changes were one or two lanes, and to the left or to the right. For the lane change maneuvers, we were interested in the RTs from target onset until the steering wheel was turned three degrees from the straight position. After reaching the designated speed or lane for 2,000 ms, a signal tone was displayed and drivers should speed up to 60 km/h again, or change back to the middle lane. In target-absent catch trials no maneuver should be performed. For a correct response, participants could accomplish the required maneuver within

5,000 ms after target display onset. If they reacted incorrectly, too late, or missed a reaction, the signal tone was anyway presented 7,000 ms after target onset.[19] This ensured that participants returned to the middle lane or speeded up in time before the next trial started. Participants were again informed that the auditory words preceding the target display were non-informative.

As an introduction to the experimental task, participants completed an initial practice phase to familiarize with the simulator. They drove on a straight road for 4.3 km with a single icon presented on each of the 20 gantry road signs (see Appendix E). In case of a red symbol with the "x" participants should immediately brake and reach a target speed of 20 km/h. In case of a green symbol with the arrow, they should change to the respective lane as fast as possible. After hearing the signal tone they should return to the initial speed or the middle lane. Afterwards, participants were given the opportunity to familiarize with the four automotive target icons as in Experiment 4. Subsequently, they performed another practice track containing 20 sample trials (4.3 km), which were equivalent to the subsequent main experimental track. All auditory prime types were included without contingency between primes and targets. The experiment lasted for approximately 45 minutes.

6.2.2 Results

Catch trials without target were not included in the following evaluation, since no driving maneuver should be performed. Data from one participant had to be excluded due to a remarkably lower accuracy rate (55 %) than average for the remaining participants. Prior to aggregation, we discarded error trials (1.5 % false reaction or no reaction, 0.7 % for braking, 2.4 % for steering) and individual outliers (braking: 2.7 %; lane changes: 4.1 %; Tukey, 1977). Table 9 shows mean RTs and error rates for all conditions, after averaging braking and steering performance (detailed information on both maneuvers is included in Appendix G).

[19] An incorrect reaction was recorded for the braking task, if participants slowed down to a speed above 20 km/h, if they turned the steering wheel beyond three degrees, or if they exceeded the lane markings of the middle lane, whereas changing to an incorrect lane or taking the foot off the gas pedal was recorded as an incorrect response for the steering task. "No response" was recorded in target-present trials, if participants did neither respond correctly nor incorrectly within the given time window.

Table 9. Mean reaction times (RTs in ms; error rates in parentheses) averaged for both maneuver types in Experiment 5A as a function of semantic congruency conditions.

	Semantic congruency condition			
	Congruent	Neutral control	Neutral silence	Incongruent
Contingency				
No	980 (1.1)	1071 (1.9)	1037 (1.4)	1063 (1.6)

The RTs for the four semantic congruency conditions (Semantic congruency of auditory prime and target object: congruent vs. neutral control vs. neutral silence vs. incongruent) were analyzed in a repeated measures MANOVA. The analysis revealed a significant main effect of semantic congruency, $F(3, 20) = 14.53$, $p < .001$, $\eta_p^2 = .69$. Helmert contrasts further showed that trials without prime word presentation (neutral silence) did not differ from the trials with prime word (all other trial types), $F < 1$, and that the neutral prime word trials did significantly differ from trials with relevant prime words (congruent and incongruent [pooled]), $F(1, 22) = 13.97$, $p = .001$, $\eta_p^2 = .39$.[20] Moreover, the contrast of main interest (congruent vs. incongruent trials) revealed a significant difference, $F(1, 22) = 42.71$, $p < .001$, $\eta_p^2 = .66$.

In addition, difference scores for facilitation (neutral control - congruent) and interference (incongruent - neutral control) were calculated. This helps to understand how the neutral control condition differs from the two relevant prime word conditions, and gives insights into potential costs and benefits. Facilitation (Mean: 91 ms) significantly differed from zero, $t(22) = 5.24$, $p < .001$, whereas interference (Mean: -7 ms) did not, $t(22) < 1$.

For the error rates (see Table 9), we conducted the same analysis as for the RTs. A repeated measures MANOVA (Semantic congruency: congruent vs. neutral control vs. neutral silence vs. incongruent) revealed no significant overall effect of semantic congruency, $F < 1$. Besides, the difference between

[20] We will additionally present difference scores for facilitation and interference in the following paragraph, thereby providing details regarding this significant contrast.

incongruent and congruent trials was again numerically positive, contradicting a potential speed-accuracy trade-off.

6.2.3 Discussion

In brief summary, with Experiment 5A we were able to replicate the result pattern from Experiment 4. Even though the driving maneuvers were much more complex than simply pressing a key, we again found a clear Stroop effect between congruent and incongruent trials. Once more, the effect occurred due to considerable facilitation with no significant interference involved. Since the average RTs in Experiment 5A were similarly long as in Experiment 4, it was furthermore interesting to compare the extent of facilitation. Under the restrictions, that there were some differences between the tracks (driving speed and spacing distance between two gantry road signs), and that participants were not randomly assigned to these two experiments, a glance at differences between the two experiments might still be worthwhile. Surprisingly, facilitation seemed to be even slightly more pronounced for the more demanding driving maneuvers (91 ms) than for the key presses (59 ms).

As a next step, we investigated whether our result pattern would change after switching from non-contingent prime presentation conditions to highly increased contingency. As already introduced earlier, for in-car warning systems rather valid warnings could be expected that mostly match the upcoming road incident. Accordingly, it was important for us to go even beyond fast and irrepressive crossmodal semantic priming effects with our investigation. We tried to find out whether increased contingency, and hence for example strategic attentional effects, additionally affected effects. These insights would help to assess how drivers' former experience with a quite reliable speech warning system might influence basic effects in (semantically) compatible cases, and which detrimental effects could be expected in incompatible cases.

6.3 Experiment 5B: Highly contingent spoken word primes considerably affect driving performance

Since impact of a considerably increased contingency between semantics of spoken word primes and visual targets is highly relevant with regard to real-

world in-car warnings, we addressed this aspect in Experiment 5B. In Experiment 3B, we had already raised contingency slightly above chance level. However, we consider in-car warning systems to provide conspicuously higher levels of contingency. Accordingly, we adapted our track from Experiment 5A: In Experiment 5B a congruent target ensued a given prime word in 80 % of the cases. This enabled us to contrast rather automatic crossmodal semantic effects with the influence of rather strategic usage of time-compressed primes.

6.3.1 Method

Participants. The participants of Experiment 5B were a new group of 24 students (18 females, 5 males) from Saarland University, who were payed 8 euros. All had normal or corrected-to-normal vision and none reported color-blindness or hearing problems. All participants possessed a valid driver's license for at least two years.

Design. Experiment 5B employed a within-subjects design. The independent variable semantic congruency (congruent vs. incongruent vs. neutral [featuring either a control prime word or silence] was manipulated like in Experiments 4 and 5A. This time, however, we increased the conditional probability for a congruent target to appear after a given prime word considerably above chance level (80 %), thus resulting in highly contingent auditory priming. That is, from listening to the primes participants could infer about the upcoming (mostly congruent) targets. Congruency was manipulated on a trial-by-trial basis. Each experimental track comprised 152 trials with a target icon present: 96 congruent trials, 24 incongruent trials, and 32 neutral trials. The neutral trials (16 control and 16 silence) and catch trials (12 ≙ 7.3 % of all trials) were identical to Experiments 4 and 5A. Overall, each track comprised 164 randomly intermixed trials (see also Table 10) – except for target icons were not being repeated from one trial to the next.

Table 10. Overview of the 164 trials presented in Experiment 5B. An auditory prime and a target icon specified a trial. The color of all target icons was counterbalanced, and auditory primes comprising a target-relevant denotation semantically matched the subsequent target icon in 80 % of the cases (24 out of 30 trials; high contingency).

	Target icon									
	Ampel (traffic light)		Kinder (children)		Notarzt (ambulance)		Traktor (tractor)		No target	Sum
Auditory prime	red	green	red	green	red	green	red	green		
Ampel	12	12	1	1	1	1	1	1	2	32
Kinder	1	1	12	12	1	1	1	1	2	32
Notarzt	1	1	1	1	12	12	1	1	2	32
Traktor	1	1	1	1	1	1	12	12	2	32
Neutral control	2	2	2	2	2	2	2	2	2	18
Neutral silence	2	2	2	2	2	2	2	2	2	18
Sum	19	19	19	19	19	19	19	19	12	164

Like in Experiment 5A, the classification of the target color (red or green) enabled drivers to choose the respective driving maneuver (braking or changing to the respective lane). The driving performance RTs constituting the dependent variable were identical to Experiment 5A.

Materials. The identical auditory stimuli and SOA as in Experiment 5A were used. This held true also for the distance between gantry road signs, Accordingly, the entire track with 164 trials was about 33 km long.

Procedure. The procedure was analogous to Experiments 4 and 5A. The second practice drive was correspondingly adjusted to contain the same contingency as the main experimental track. Participants were informed that the auditory words predicted the target object in most cases.

6.3.2 Results

Catch trials without target were not included in the following evaluation, since no driving maneuver should be performed. Data from one participant had to be excluded due to a remarkably lower accuracy rate (34 %) than average for the remaining participants. Prior to aggregation, we discarded error trials (1.1 % false reaction or no reaction, 0.5 % for braking, 1.8 % for steering) and individual outliers (braking: 3.9 %; lane changes: 4.1 %; Tukey, 1977). Table 11 shows mean RTs and error rates for all semantic congruency conditions

after braking and steering performance were averaged (detailed information on both maneuvers is included in Appendix G).

Table 11. Mean reaction times (RTs in ms; error rates in parentheses) averaged for both maneuver types in Experiment 5B as a function of semantic congruency conditions.

	Semantic congruency condition			
	Congruent	Neutral control	Neutral silence	Incongruent
Contingency				
High	927 (0.6)	1070 (0.8)	1051 (1.6)	1110 (3.1)

The RTs for the four semantic congruency conditions (Semantic congruency of auditory prime and target object: congruent vs. neutral control vs. neutral silence vs. incongruent) were analyzed in a repeated measures MANOVA. The analysis revealed a significant main effect of semantic congruency, $F(3, 20) = 36.30$, $p < .001$, $\eta_p^2 = .85$. Helmert contrasts further showed that trials without prime word presentation (neutral silence) did not differ from the trials with prime word, $F(1, 22) < 1$, and that the neutral prime word trials significantly differed from trials with relevant prime words (congruent and incongruent [pooled]), $F(1, 22) = 11.43$, $p = .003$, $\eta_p^2 = .34$. [21] Moreover, the contrast of main interest (congruent vs. incongruent trials) revealed a significant difference, $F(1, 22) = 98.63$, $p < .001$, $\eta_p^2 = .82$.

In addition, difference scores for facilitation (neutral control - congruent) and interference (incongruent - neutral control) were calculated. Both, facilitation (Mean: 143 ms), $t(22) = 8.49$, $p < .001$, as well as interference (Mean: 40 ms), $t(22) = 2.14$, $p = .04$, significantly differed from zero.

For the error rates (see Table 11), we conducted the same analysis as for the RTs. A repeated measures MANOVA (Semantic congruency: congruent vs. neutral control vs. neutral silence vs. incongruent) revealed a significant

[21] We will additionally present difference scores for facilitation and interference in the following paragraph, thereby providing details on the nature of this significant contrast.

effect of semantic congruency, $F(3, 20) = 3.97$, $p = .02$, $\eta_p^2 = .37$. Helmert contrasts further revealed that neutral silence trials did not differ from all other trial types, $F(1, 22) < 1$. Besides, neutral control word trials did not differ from trials with relevant prime words (congruent and incongruent pooled), $F(1, 22) = 2.06$, $p = .17$, $\eta_p^2 = .09$. For the contrast of main interest (congruent vs. incongruent), a significantly higher error rate was found for incongruent than for congruent trials, $F(1, 22) = 7.41$, $p = .01$, $\eta_p^2 = .25$. This finding is in line with the pattern of RTs, and contradicts a potential speed-accuracy trade-off.

6.3.3 Discussion

In a nutshell, we found a clear Stroop effect comprising both facilitation as well as interference. Accordingly, after we increased the contingency between prime and target interference came into play regarding both RTs and error rates. In cases when the number of congruent trials obviously exceeds the number incongruent trials, performance in the latter trials seems to be decreased by incongruent semantics between prime and target. However, interference was still considerably less pronounced than facilitation. This points towards contingency effects to constitute an additive component on top of the rather automatic Stroop effects (Logan, 1980). In the following section, we directly compare Experiments 5A and 5B in order to investigate effects of contingency on crossmodal facilitation and interference.

6.4 Highly contingent presentation boosts both crossmodal facilitation and interference

Besides the manipulation of contingency, Experiments 5A and 5B were designed in an equivalent fashion. Moreover, participants were randomly assigned to the conditions. Hence, we were able to additionally analyze the effect of contingency as a between-participants factor.

Table 9 and Table 11 show the mean RTs for all conditions in Experiments 5A and 5B. RTs were analyzed in a 2 (contingency: No [Exp. 5A] vs. High [Exp. 5B]) × 4 (semantic congruency: congruent vs. neutral control vs. neutral silence vs. incongruent) mixed model MANOVA. In order to avoid redundancy, we only report effects involving the contingency factor.

The analysis revealed that there was no significant main effect of contingency, $F < 1$, but a significant interaction of contingency and semantic congruency, $F(3, 42) = 3.47$, $p = .02$, $\eta_p^2 = .20$, For this interaction, Helmert contrasts revealed that neutral silence trials did not differ from all other trial types, and neutral control word trials also did not differ significantly from relevant prime word trials (congruent and incongruent [pooled]), both Fs < 1. However, for the comparison of congruent and incongruent trials, a significant interaction of contingency and semantic congruency was revealed, $F(1, 44) = 19.50$, $p < .001$, $\eta_p^2 = .31$. When having a look at the RTs in Table 9 and Table 11, this finding implicated that for congruent trials, highly contingent prime word presentation led to faster RTs than non-contingent presentation. In contrast, for incongruent trials the pattern seemed to be reversed. In accordance with this consideration, also the Stroop difference (RTs incongruent - RTs congruent) was higher for contingent than for non-contingent prime word presentation (see Figure 11a), $t(44) = 4.42$, $p < .001$. Nevertheless, both differences are positive and hence point towards basically similar crossmodal effects of semantic congruency.

In order to get a more detailed view on this issue, also facilitation and interference scores were analyzed with respect to contingency. For facilitation, we found a significantly higher effect for highly contingent presentation than for presentation without contingency (see Figure 11b), $t(44) = 2.14$, $p = .04$. As we already pointed out in the former results sections, both facilitation effects were significantly above zero. For interference, we also found a significantly higher effect for highly contingent presentation than for presentation without contingency (see Figure 11c), $t(44) = 2.17$, $p = .04$. Resuming from the respective results sections, interference significantly differed from zero for highly contingent presentation, but not for non-contingent presentation.

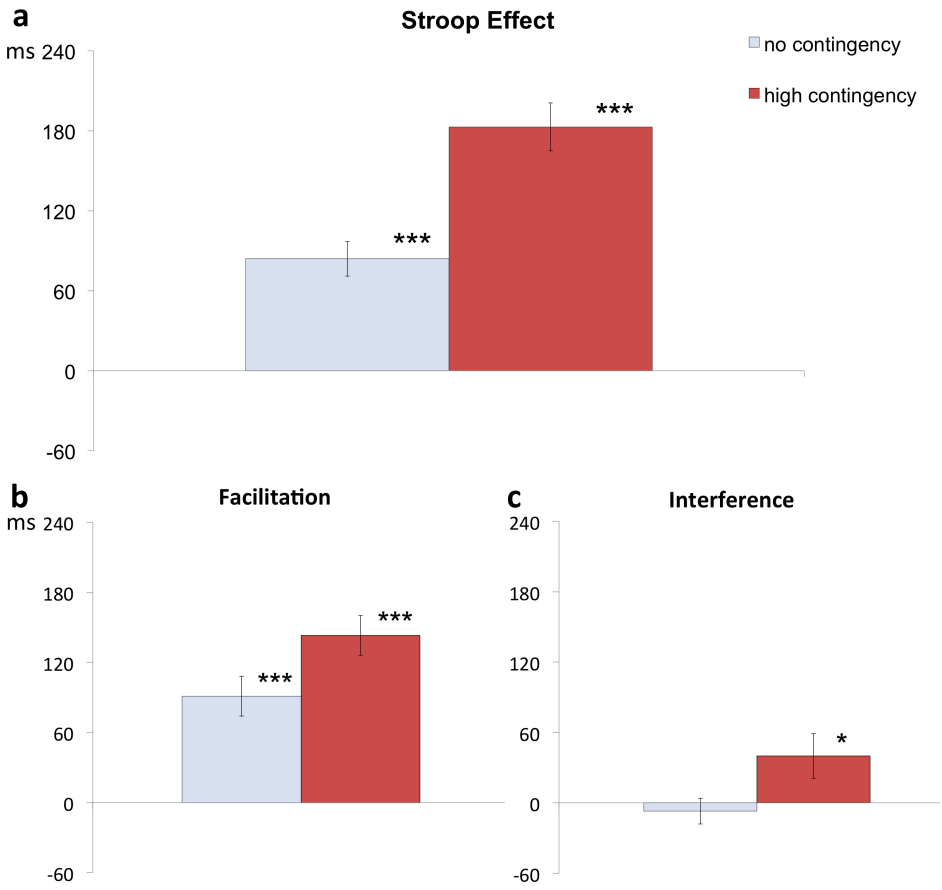


*Figure 11. Reaction time (RT) differences (Stroop: incongruent trials – congruent trials; facilitation: neutral - congruent; interference: incongruent – neutral) for the two contingency levels in Experiments 5A and 5B. The error bars indicate ± 1 SEM (see Franz & Loftus, 2012). * $p < .05$, ** $p < .01$, *** $p < .001$.*

For the error rates (see Table 9 and Table 11), we conducted the same analysis as for the RTs. We merely add reports on the effects involving contingency. The 2 (contingency: No [Exp. 5A] vs. High [Exp. 5B]) × 4 (Semantic congruency: neutral silence vs. neutral control vs. congruent vs. incongruent) mixed model MANOVA did not show any contradictory results with regard to RTs. Neither the main effect of contingency, $F(1, 44) < 1$, nor the interaction of contingency and semantic congruency, $F(3, 42) = 1.38$, $p = .26$, $\eta_p^2 = .09$, were significant.

In the preceding section, we investigated how a considerably increased level of contingency between spoken prime words and visual targets would affect the pattern of crossmodal semantic effects. In Experiment 5A the prime words did not reveal any contingent information about the subsequent targets,

whereas in Experiment 5B a given prime word was most probably followed by its respective (congruent) target icon. Our findings can be briefly summarized by contingency increasing effects. In fact, both facilitation and interference were significantly higher for highly contingent presentation. Therefore, we suggest that if contingency is included in a crossmodal semantic priming design, (strategic, attentional, controlled) components come into play, thus leading to additive effects on top of the fast and irrepressive crossmodal semantic priming effects (Logan, 1980). These insights imply that former experience with a quite reliable speech warning system might additionally influence fundamental effects: If a system frequently presents meaningful warnings with spoken words matching the hazardous situations, even higher benefits could be expected after a while. However, the flip side of the coin might become obvious in cases of system failure when speech is not in accordance with the upcoming hazard. In such cases, drivers might respond to critical situations slightly slower than without a warning (silence condition) or than with a neutral “master alert”. Because of those potentially detrimental effects of speech warnings, it would be important to carefully consider the accuracy of a warning system. Hit and false alarm rates might thus become relevant for the recommendation, whether only master alerts should be used, or whether the presentation of more specific speech warnings would be advisable. Of course, with respect to different reliability levels of warning systems further experiments would become necessary in order to derive definite conclusions about respective benefits and costs of speech warnings.

7 Reflection

In this chapter, we first briefly summarize how our applied research question has successively led us to concrete hypotheses in the domain of cognitive psychology. In a next step, we recapitulate our experimental findings along the lines of these hypotheses (Section 7.1.1) and discuss them in a more general light of basic cognitive psychology regarding crossmodal auditory-visual effects of semantics (Section 7.1.2). Later on, we return to our applied research question. Here, we examine our findings in light of speech warnings in time-critical, hazardous road situations (Section 7.1.4). Besides, for each of the two aforementioned perspectives we additionally point out open issues and future directions (Sections 7.1.3 and 7.1.5). Finally, we conclude the thesis with a brief summary of the most relevant findings and insights, accompanied by an evaluation regarding our extraordinary approach: Bringing together the "two worlds" of applied research questions and basic cognitive psychology (Section 7.2).

7.1 General discussion

With this thesis, we investigated how a wide variety of different information could be suitably presented to drivers in time-critical road hazard situations. Relevant literature on this issue revealed that the usage of the auditory modality, and especially the presentation of speech warnings, is a promising approach for urgent and specific information transfer. Even though there exist surprisingly few publications regarding speech warnings for drivers, and none could provide sufficient details about fast, semantic information transfer, the synopsis of those earlier findings (e.g., Graham, 1999; McKeown & Isherwood, 2007; Seppelt & Wickens, 2003; Vilimek & Hempel, 2005) resulted in a rather positive summary. Unfortunately, the key issue whether congruent speech actually has the potential to crossmodally support processing of visual scenes still remained unanswered. Hoping to find out more about this assumption, we turned to basic findings from cognitive psychology in a next step. Also in this research domain, results were promising on the whole, since they pointed towards verbal stimuli and spoken words crossmodally affecting visual tasks (e.g., Chen & Spence, 2011; Lupyan & Thompson-Schill, 2012;

Roelofs, 2005; Salverda & Altmann, 2011; Tellinghuisen & Nowak, 2003). Nevertheless, considerable differences in materials, paradigms, timing, and design (potentially confounding conditions) complicated a consistent conclusion about the nature of effects. Since some of these issues were highly relevant for the transfer of those basic findings to an applied warning scenario, we decided to address them by conducting a series of experiments. For example, we considered it highly relevant whether effects actually occur due to a semantic match between prime and target, whether strategic listening is a necessary precondition, whether effects emerge immediately, or whether auditory presentation duration influences effects. Based on our findings in Experiments 1A – 5B, we carefully revisit our initial hypotheses about crossmodal semantic effects in the next section.

7.1.1 Summary of the results

Our key hypothesis throughout this thesis was that spoken words have the potential to crossmodally affect – or more precisely, enhance – target identification and classification of semantically related, visual targets. Therefore, we addressed this issue in all experiments, receiving a consistent approval: In each of our eight experiments we found that semantic congruency of spoken words and visual targets clearly affected performance. More precisely, when separating benefits and costs by using a neutral word condition, we consistently found considerable facilitation in congruent trials and only minor interference by incongruent trials. Besides, we also found better performance in congruent trials than in neutral silence trials (e.g., Exp. 2).[22] Importantly, despite an extremely short SOA of only 100 ms crossmodal priming effects via semantics occurred reliably. Accounting for substantial differences in materials, tasks, and further conditions between the eight experiments, we conclude that our results consistently support our key hypothesis.

Another crucial issue was whether crossmodal effects would occur due to stimulus-response compatibilities, or rather based on stimulus-stimulus

[22] Please note that also in Experiments 4, 5A, and 5B, a neutral silence condition was applied and responses in congruent trials were always faster than in silence trials. However, we did not explicitly test significance within our a priori defined Helmert contrasts.

compatibilities (De Houwer, 2003). This question was highly interesting, since the former account would indicate that auditory primes matching motor responses would cause effects, whereas the latter account denotes that a match between semantically identical objects (or object features) can facilitate performance – even though their presentation is physically diverse. In this regard, our second hypothesis could also be affirmed: In several experiments (2, 4, 5A, and 5B) we excluded response-priming effects by requesting responses that were orthogonal to spoken words. Nevertheless, we still found considerable crossmodal effects. Accordingly, a spoken word in the auditory modality facilitated the encoding of a semantically congruent item in the visual modality based on crossmodal stimulus-stimulus compatibility. It seems obvious to assume a shared semantic representation as cause for this compatibility, and we will soon return to this issue in more detail. Besides, it is important to note that our clear effects occurred unintentionally: Even if the spoken words did not reveal any information about the following target, and thus were completely irrelevant for participants, clear effects were revealed (Exps. 2 and 4). This showed that the influence of spoken words on visual performance does not depend on conscious listening and is not necessarily based on a strategic process.

Furthermore, we assumed perceptual load of the visual task to be a critical factor for crossmodal semantic effects. When addressing this issue in Experiments 1A and 1B, we found that effects were actually more pronounced under high than under low perceptual load – even though any benefits from listening were excluded. Above all for more demanding visual search conditions when there is less spare attentional capacity, spoken words seem to affect target detection, thus confirming our third hypothesis about perceptual load.

When turning to our fourth hypothesis that under time-critical circumstances crossmodal effects are significantly increased for time-compressed spoken words, a closer look needs to be taken. Even though pitch-synchronous time-compressed spoken words (50 %, 30 %, or 10 % of the original duration) allowed for earlier disambiguation and ending, we could neither affirm this hypothesis with one-syllable words (Exps. 1A and 1B), nor with two-syllable words (Exps. 3A and 3B). As compared to spoken words

with normal duration, time-compressed stimuli led either to similar crossmodal priming effects (Exps. 1A and 1B: 30 % duration, Exps. 3A and 3B: 50 % and 30 % duration), to slightly decreased effects (Exps. 1A and 1B: 10 % duration/ high perceptual load), or effects were even eliminated completely (Exps. 1A and 1B: 10 % duration/ low perceptual load). Accordingly, at least if spoken words are not useful or informative for the visual task, time-compression would not improve performance. On the contrary: Extremely high levels of time compression decreased crossmodal semantic priming effects, or even eliminated them completely.

Next, we summarize our findings regarding contingent presentation of spoken words and visual targets. We were especially interested, whether strategic (or implicit) use of the auditory primes would increase our crossmodal priming effects. While for most experiments we applied a non-contingent presentation with primes and targets only matching at chance level (1A, 1B, 2, 3A, 4, and 5A), we explicitly included contingency in two experiments. In Experiment 3B, primes were slightly predictive about the subsequent targets (conditional probability of a congruent target after a given prime word: 50 %), whereas in Experiment 5B, the auditory primes were even highly predictive of the visual targets (conditional probability: 80 %). Since those two experiments were in all other respects identical with Experiments 3A and 5A, respectively, we could directly investigate effects of contingency. When contingency was only slightly increased above chance level (Exp. 3A vs. 3B), we did not find significant changes in crossmodal semantic priming effects. However, with a considerably higher level of contingency (Exp. 5A vs. 5B), crossmodal semantic facilitation from congruent trials increased significantly. Besides, significant interference was identified in incongruent trials. This suggests that if contingency was rather obvious, participants tried to improve their performance by listening to the spoken words, thus deriving mostly valid information about the subsequent targets. However, taking spoken words into consideration also increased costs in incongruent trials when auditory information was misleading. Our findings point towards strategic effects arising from highly contingent presentation (Logan, 1980). We will return to this issue later in Section 7.1.2.

The sixth hypothesis comprised that time-compression levels of spoken words would moderate effects of contingency. This issue was addressed in Experiments 3A and 3B. In fact, a significant interaction of word duration and contingency was revealed. For spoken words that were compressed up to 30 % of their original duration, slightly contingent spoken words led to significantly faster RTs than spoken words without any predictive character. On the contrary, for spoken words with longer duration, no significant difference in RTs was found. Thus, even though we could not directly confirm strategic usage of auditory semantic information and increased crossmodal priming effects for a relatively low contingency level earlier (Exp. 3B), the findings regarding the interaction between contingency and time compression might indicate that participants have, nevertheless, (implicitly) started considering the primes. We assume that fast disambiguation of highly time-compressed stimuli would only affect RTs differently for non-contingent versus slightly contingent presentation, if contingency changed how participants listened to the primes. The diverse effects in the highly time-compressed condition reveal that already slightly increased contingency can actually become relevant and make listening worthwhile, albeit eventually not by a conscious and strategic process but rather by an implicit decoding of semantics.

Finally, our last hypothesis could also be confirmed, since we could clearly replicate significant crossmodal semantic priming effects under dynamic driving conditions (Exps. 4, 5A, and 5B), and even when complex driving maneuvers had to be performed (braking and lane changes; Exps. 5A and 5B). This constitutes a first step towards the investigation of speech warnings in real-world driving scenarios.

In brief summary, the results obtained with our sequence of experiments revealed that spoken words have the potential to influence the processing of visual scenes by immediately and efficiently activating respective semantic stimulus-stimulus compatibilities. Most importantly, these effects seem to be much more facilitative in congruent cases than detrimental in incongruent cases. We will discuss both the basic and the rather practical implications of our findings more precisely in the following sections.

7.1.2 Discussion of crossmodal auditory-visual semantic effects[23]

In this section we start with the discussion of our findings in light of existing literature on crossmodal effects of auditory semantic primes (see Chapter 3), and highlight some insights that were achieved by our series of experiments.

High-perceptual-load visual search conditions: Decreased inhibition of auditory distractors, or increased effects of stimulus-stimulus compatibilities?

For a start, we consider it worthwhile to revise Tellinghuisen and Nowak's (2003) assumption that their crossmodal distractor effect patterns were merely due to response-priming processes (and hence above all stimulus-response compatibilities) combined with impaired inhibition of distractors. This rationale cannot be entirely upheld, given our data. First of all, and as already mentioned earlier, our effects still occurred when response priming was ruled out (Exps. 2, 4, 5A, and 5B). Secondly, we replicated their asymmetric pattern of pronounced crossmodal facilitation and only minor interference in all of our experiments. A pure response-priming account does not provide a rationale for extremely asymmetric facilitation and interference effects. Thirdly, their interpretation of larger crossmodal effects in a high-load visual search task, as compared with a low-load visual search task, rests on the intricate assumption of lower spare capacity to inhibit distractors under high-load conditions (see also Lavie & Cox, 1997). Tellinghuisen and Nowak claimed that increased cognitive resources were necessary for auditory distractor inhibition as compared to visual unimodal distractor inhibition, leading to more pronounced distractor effects in the crossmodal condition. However, according to our data it is more plausible and generally more parsimonious to assume that enhanced target processing in the case of congruent priming considerably contributes to the effects found with crossmodal Stroop-like designs.

Unlike Roelofs (2005) and Tellinghuisen and Nowak (2003), we controlled for stimulus-response compatibilities as a source of effects in several

[23] Please note that parts of this section have been reported in Mahr, A., & Wentura, D. (2014). Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets. *Attention, Perception, & Psychophysics, 76,* 575-590. Copyright © 2014 by Springer. Adapted with permission. No further reproduction or distribution is permitted without written permission from Springer.

experiments (2, 4, 5A, and 5B). Compatibly, we suggest that our effects are mostly based on semantic stimulus-stimulus compatibilities, and hence crossmodal semantic priming. In this regard, it is highly interesting to consider the underlying process and to discuss whether spoken word primes could activate modality-independent semantic representations.

Modality-independent semantic representations? Referring to the reading-naming interference model (Glaser & Glaser, 1989)

From an extension of Glaser and Glaser's reading-naming interference model (1989) by Chen and Spence (2011), it can be inferred that the auditory features of a spoken word automatically activate semantic information via lexical systems. This activated (modality-independent) semantic information potentially enhances detection or identification of corresponding pictures. Accordingly, when presenting a congruent auditory word prime during processing of a rather complex visual scene, facilitated object detection or classification would be expected as compared with a control condition.

In two aspects, however, the results of Chen and Spence (2011) were not completely in line with this theory. Firstly, they did find crossmodal effects of spoken words only if the target had to be identified (and not only be detected). Secondly, they found their effect only with a rather long SOA (346 ms). Given these constraints, they argued that semantic representations of spoken words might generally be too weak or their activation might be too slow to find effects with (a) a mere detection task and (b) short SOAs. Consonant with the latter precondition, also Lupyan and Ward (2013) applied an extremely long SOA above one second, when finding that language-based activation of two visual shape representations improved respective detection sensitivity. Similarly, Salverda & Altmann (2011) obtained attentional capture without target identification using task-irrelevant spoken language cues with a rather long SOA (400 ms). However, both studies did not entirely confirm Chen and Spence's assumptions since they did not comprise target identification. Even beyond that, we repeatedly found crossmodal semantic priming effects by spoken words without target identification and at an SOA of

only 100 ms (Exps. 2, 4, 5A, and 5B).[24] Therefore, at least given specific conditions (see below), short SOA-priming with a mere detection task is possible. Chen and Spence's assumption that semantic representations of spoken words are generally too weak or their activation is too slow was finally overcome by our experiments. In conclusion, crossmodal semantic priming effects are in line with the "unity assumption" (Spence, 2007; Chen & Spence, 2011), which claims that binding of information from different sensory modalities is facilitated by semantic congruency.

Four target objects are constantly kept in working memory: How could attentional sets influence crossmodal effects?

Of course, the differences of our results to those of Chen and Spence (2011) might be due to the fact that, contrary to Chen and Spence (2011), our task implies forming a strong attentional set: Four relevant target items had to be actively kept in working memory throughout the experiment. Up to around four items can typically be attended to, enumerated, and retained simultaneously in attention and working memory tasks (Alvarez & Cavanagh, 2004; Cowan, 2000; Luck & Vogel, 1997; Phillips, 1974) and attentional sets are powerful determinants of performance (e.g., Olivers, Meijer, & Theeuwes, 2006). If an auditory prime is related to the content of a currently active attentional set, it might influence perception of the respective item. In our experiments, spoken word primes might accordingly accentuate processing of one of the four visual (or even amodal) working memory items, since visual working memory can prioritize one feature over the other (Olivers et al., 2006). Independent of the target screen the semantic representation of the boosted item will be especially (pre)activated (Desimone & Duncan, 1995), leading to increased visual sensitivity.

Given this backdrop, we can easily integrate our finding that crossmodal semantic effects became most obvious in the high-perceptual-load condition. We designed the perceptual-load manipulation in Experiments 1A and 1B analogous to Lavie and Cox (1997) and Tellinghuisen and Nowak

[24] Besides, in another (so far unpublished) experiment, we were even able to replicate our findings on crossmodal semantic priming effects with a simultaneous onset of prime and target (SOA 0 ms).

(2003), except, of course, for increasing the number of potential targets from two to four. However, the question might arise whether the manipulation of load is exclusively a manipulation of perceptual complexity or whether – at least indirectly – the contribution of working memory processes is manipulated as well.[25] As was said in the preceding paragraph, the four different target colors had to be held active in working memory. This maintenance process would cause slightly varying activations for the different colors across the time span. Metaphorically speaking, at a given point in time, a color might be more in the "foreground" or "background" of the working memory. Besides active rehearsal, priming a color will put this color into the "foreground". In the low-perceptual-load condition, the highly salient target color strongly resonates with the corresponding working memory content (Olivers et al., 2006), boosting it into the "foreground". Thus, variations in "foregrounding" caused by the auditory prime would not make a big difference here. In the high-perceptual-load condition, with a strongly reduced salience of the target, however, nonefficient top-down guided search processes would become more dominant. Plausibly, whether a target color was "foregrounded" in working memory or not would now make a difference, because the "foregrounded" item would dominate the top-down guided search process. These varying activations have just been explicated for the case that primes did not reveal any information about the subsequent targets. Nevertheless, the same basic mechanism might also be applicable in case of contingent presentation of prime and target. We will shortly return to this assumption after discussing the effects of contingent presentation on crossmodal semantic effects.

Crossmodal semantic priming effects: Disentangling semantic stimulus-stimulus compatibilities and contingency

Most related experiments on crossmodal semantic effects controlled only for stimulus-response compatibilities by responses to targets being orthogonal to prime semantics. However, the conditional probability that a semantically congruent target followed a given prime was above chance level (Lupyan & Spivey, 2010; Lupyan & Thompson-Schill, 2012, Exps. 3a and b; Lupyan &

[25] We thank Tracy Brown for pointing out this ambiguity.

Ward, 2013, Exps. 1 and 2; Salverda & Altmann, 2011, Exp. 1). Unfortunately, this left unclear whether the crossmodal findings based on stimulus-stimulus compatibilities or on contingent presentation of primes and targets. In contrast, we additionally applied non-contingent presentation in several experiments (1A, 1B, 2, 3A, 4, and 5A). Thereby, we could exclude that any strategic listening would contribute to crossmodal effects and confirm the efficacy of crossmodal, semantic stimulus-stimulus compatibilities. Subsequently, we explicitly compared non-contingent and contingent presentation of primes and targets (Exp. 3A vs. 3B: slight contingency, Exp. 5A vs. 5B: high contingency). At least for highly contingent presentation, a significant interaction of contingency and congruency was revealed: If the conditional probability of a semantic match between target and a given prime was obviously increased, both facilitation and interference were clearly increased, as well. These findings point towards strategic effects arising from highly contingent presentation. In this regard two processes seem plausible: Either strategic effects could quantitatively increase the rather automatic crossmodal priming effects (with their major facilitation and minor interference), or they could equally add on top of both of these crossmodal priming components as an even qualitatively different component (for related considerations in the context of a Stroop-like task see also Logan, 1980). Further clarification of these two accounts and the exact mechanisms would be worthwhile in the future. In any case, these strategic effects support our skepticism towards many earlier studies that confounded the investigation of semantic congruency by contingent presentation of semantics. Moreover, attentional effects and controlled processes might become all the more relevant and pronounced for longer SOAs, since during a longer period of time a more conscious and intentional preparation for semantically congruent targets would be enabled. Therefore, as long as semantics of primes and targets are somehow correlated (within sequences, blocks, or experiments), effects of semantic congruency have to be interpreted with caution. Accordingly, many earlier findings on crossmodal semantic effects left rather unclear whether effects were mainly unintentional effects of semantic congruency, or rather occurred due to primes revealing information about subsequent targets. In contrast, we can claim credit for shedding light on this issue by a deliberate separation of the two accounts.

As already mentioned previously, the assumption regarding activation of working memory items by spoken words could even be extended to contingent presentation of primes and targets. Here, the "foregrounding" of a working memory item by a spoken word might become even more pronounced over time, since on average this "foregrounded" representation matches the subsequent target above chance. Below the bottom line, this would "reward" activation: On the one hand, increased benefits would be caused in numerous congruent trials, and, on the other hand, increased costs occurred in incongruent trials. In addition to such a (gradual) reinforcement, which probably already occurs for less apparent contingency, also strategic listening to spoken words would probably come into play – especially for explicit or obvious prime validity. Overall, this working memory activation account seems to provide plausible reasoning for our findings, but it is certainly left for future research to test it more generally. Besides, we suggest scrutinizing speed and strength of working memory item activation for low versus high levels of contingency.

Crossmodal semantic priming effects: Faster disambiguation of time-compressed spoken words?

Contingency between semantic representations of primes and targets leads us to another remarkable interaction that should be discussed: For words that were presented with original duration, contingency did not lead to a difference in RTs. However, for highly compressed spoken words (30 % of their original duration), slight contingency already led to significantly faster RTs than no contingency (i.e., conditional probability at chance level). Besides, this difference seemed to become especially obvious for congruent trials. Thus, even though we could not confirm a moderation of semantic congruency effects by contingency between prime and target, the interaction between contingency and time compression indicated that participants might still (implicitly) listened to the primes and decoded semantics when the conditional probability of a target matching a given prime was slightly increased above chance level (Exp. 3A vs. 3B). This might have led to reinforced activation of the respective semantic representations. In this situation, faster disambiguation of highly time-compressed stimuli could become effective and mitigate

perceptual degradation. Against this background, we anticipate the effect to become considerably more pronounced for higher contingency levels and for longer, polysyllabic speech.

Is crossmodal semantic priming an automatic process? A consideration of Moors' (2010) features on automaticity

Next, we address the interesting question whether crossmodal priming via semantics occurs rather automatically, or is rather based on strategic processes. In order to gain insights about the level of automaticity of our crossmodal priming effects, we revisit the individual features summarized by Moors (2010) which we have already introduced earlier in our excursus on automatic processes in Section 3.1. Firstly, it is interesting, whether processes occur unintentionally – and are hence uncontrolled in a promoting sense. Suitably, in many of our experiments we found effects even though spoken words were completely irrelevant for participants and this circumstance was also clearly announced (Exps. 1A, 1B, 2, 3A, 4, and 5A). Therefore, still finding clear effects in these experiments revealed that the goal to engage in the crossmodal priming process was not a prerequisite, thus indicating an unintentional process. Secondly, crossmodal semantic priming might also be uncontrolled in a counteracting sense: On inquiry, some participants indicated after Experiment 1A that they had actively tried to ignore the auditory information, but, nevertheless, no difference in priming effects was found for this group of participants (see Appendix C). This points towards rather robust effects and an uncontrolled process that is not stopped by the goal to avoid it. The third feature for automatic processes denotes whether it operates even if one is not aware of the relevant input or process. Since, we presented stimuli that were clearly perceivable and the meaning of the spoken words remained recognizable even for highly time-compressed spoken words, participants were clearly aware of the input. With regard to the crossmodal process, participants were explicitly asked after Experiment 1A, whether they thought that the irrelevant primes speeded their reactions in the visual task in case of a match, and slowed them in case of a non-match. Consent was revealed for both of these claims (see Appendix C), and hence, also this second aspect regarding unconscious processes was not confirmed for our experiments. Accordingly,

spoken word primes seem to affect visual performance rather consciously. With regard to "efficiency" as the fourth feature for automatic processes, we assume that crossmodal semantic priming occurs rather effortless and only with minimal use of attentional capacity. Otherwise, we would not have found even more pronounced effects under high than low-perceptual-load conditions in Experiments 1A and 1B. Last but not least, we could confirm that crossmodal semantic priming operates within a considerably short amount of time, since we repeatedly found effects with an SOA of only 100 ms. To summarize, four out of five features for automatic processes proposed by Moors could be confirmed in our experiments. Hence, we would argue that crossmodal priming via semantics occurs rather automatically under the given circumstances. On the one hand, this seems to be plausible due to the fact that the recognition of speech is constantly being well trained since our very early days. On the other hand, our findings are quite astonishing, since spoken words need some time until they can be disambiguated and comparatively slow higher-level processing is generally deemed necessary for understanding of speech (see, e.g., Goldstein, 2002). Against this background, our evaluation on automaticity of effects is a valuable contribution to the field of (crossmodal effects of) speech perception.

What does the cost-benefit partitioning reveal about crossmodal semantic priming effects and the underlying mechanisms?

While we generally found pronounced semantic effects, it is moreover worth discussing underlying costs and benefits. Therefore, we initially reconsider striking differences in earlier findings regarding benefits or costs causing crossmodal effects of semantics. In some cases pronounced facilitation and no (or only minor) interference were revealed (Chen & Spence, 2011; Lupyan & Spivey, 2010; Tellinghuisen & Nowak, 2003), in some cases findings pointed towards equal facilitation and interference (Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013), and in some cases major or exclusive interference occurred (Roelofs, 2005; Salverda & Altmann, 2011). Throughout our eight experiments, the results consistently show an asymmetrical pattern with pronounced facilitation and only minor interference. However, direct comparisons between different studies are complicated, since it is unclear

whether differences in paradigms, materials, or design caused inconsistencies between findings. [26]

Besides basic mechanisms that lead to facilitation and interference, it should initially be recalled that the choice of a baseline condition could considerably affect the result pattern (Tellinghuisen & Nowak, 2003, for findings based on unimodal Stroop literature see Duncan-Johnson & Kopell, 1981; Kahneman & Chajczyk, 1983; MacLeod, 1991). Since both facilitation and interference are calculated based on their absolute difference to a neutral reference category, a slightly better overall performance in this baseline condition would, for example, apparently decrease facilitation and increase interference. Accordingly, it is important to take into account whether a neutral condition comprised one or several noise signal(s), neutral sound(s), non-word(s), neutral word(s), word(s) that are phonetically similar to targets, or merely silence. Similar to findings from the irrelevant speech paradigm, where acoustic characteristics and similarities are crucial for performance decrements in visual working memory tasks (e.g., Jones & Macken, 1993; Larsen, Baddeley, & Andrade, 2000; Tremblay, Nicholls, Alford, & Jones, 2000), we assume that using different auditory stimuli as baselines probably leads to different RTs. For example, RTs in a baseline condition would probably be faster for a neutral silence condition than for several neutral words, since for the silence condition interference by a prime would simply be excluded. In extreme cases, this difference between neutral conditions might even completely shift the result pattern from sole interference to sole facilitation, even though RTs for compatible and incompatible trials would remain absolutely constant. Due to the fact that in many of the earlier experiments on crossmodal effects of verbal stimuli only silence or white noise was applied as a baseline, this might have led to a considerable underestimation of facilitation and overestimation of interference (e.g., Lupyan & Ward, 2013; Roelofs, 2005).[27] Hence, it cannot be emphasized strongly enough that the baseline condition must be chosen deliberately in order to enable meaningful inferences

[26] Of course, also general problems like floor or ceiling effects might complicate the interpretation symmetrical or asymmetrical costs and benefits.

[27] Besides, in some experiments there was not even a baseline condition included, thus not allowing for a separation of benefits and costs at all (e.g., Chen & Spence, 2011, Exps. 3 and 4; Salverda & Altmann, 2011, Exp. 1).

about the cost-benefit partitioning of effects and to allow for comparisons across experiments. With Experiments 1A and 1B, we have partly addressed this issue by presenting either a single neutral word or several neutral nonwords. Pronounced facilitation and minor interference were consistently revealed for both experiments, thus underlining the reliability of our findings. If anything, facilitation seemed to appear even more clearly for a baseline condition with several non-words (Exp. 1B) than for a baseline condition with a single neutral word (Exp. 1A).

In contrast to the choice of a baseline condition, other factors actually influence absolute performance in congruent and/or incongruent trials and hence affect crossmodal facilitation and interference magnitude. With reference to our results, we will in the following identify important factors that contribute differently to crossmodal semantic facilitation and interference.

We designed our initial experiments (1A and 1B) closely according to earlier crossmodal experiments which were either based on the Stroop (Roelofs, 2005) or the flanker (Tellinghuisen & Nowak, 2003) paradigm. These are quite powerful interference paradigms which typically include remarkable stimulus-response compatibilities causing both benefits and interference. Also for crossmodal auditory-visual tasks, these compatibilities would be expected to affect performance. In case the semantics of a task-irrelevant spoken word are compatible with the semantics of a response, verbal stimuli might directly preactivate the respective response. In the experiments of Roelofs (2005) and Tellinghuisen and Nowak (2003), those responses again corresponded (semantically) with visual targets. Accordingly, in congruent trials performance would be increased and facilitation might occur – even without any direct involvement of the (congruent) semantic representations of visual target objects. The same consideration applies in reverse, if spoken words are incompatible with responses. Since in this instance an incorrect response would be preactivated analogously, inhibition of this response would become necessary. This would lead to decreased task performance relative to a neutral condition. In summary, for effects that occur due to stimulus-response compatibilities we would generally expect a symmetrical pattern of facilitation and interference. Against this background, our findings with considerably pronounced facilitation and only minor interference were all the more

surprising. This raised doubts about stimulus-response compatibilities being the sole cause for crossmodal semantic effects, and resulted in particular consideration of stimulus-stimulus compatibilities as a relevant factor. However, if stimulus-response and stimulus-stimulus compatibilities coincide like in many Stroop-like designs, separation of the two components and interpretation of relative strength is complicated (De Houwer, 2003). Consistently, when interpreting the findings of our Experiments 1A, 1B, 3A, and 3B, we could not separate the influences of each process. Therefore, we excluded stimulus-response compatibilities in Experiments 2, 4, 5A, and 5B. Interestingly, we could still reconfirm a similar and particularly asymmetrical pattern when only stimulus-stimulus compatibilities were included. This indicates, that our crossmodal priming effects largely base on semantic compatibility between prime and target, which facilitates performance in congruent trials, whereas in incongruent trials a non-match of representations across the two different modalities would only marginally interfere with visual processing and thereby not result in considerable performance detriments. Above all this finding of pronounced facilitation and (almost) no interference at SOAs below 200 ms resembles the result pattern found in semantic priming studies (McNamara & Holbrook, 2012; McNamara, 2005) rather than that in interference paradigms (with symmetric facilitation and interference, or even pronounced interference, e.g., Eriksen & Eriksen, 1974; MacLeod, 1991). Besides, semantic priming effects are typically based on semantic associations between prime and target, and hence also stimulus-stimulus compatibilities (rather than on stimulus-response compatibilities). However, despite these marked similarities, our effects did not occur due to associations between related semantic concepts like in semantic priming. While our stimuli differed in physical terms, semantic representations between prime and target were even identical in congruent trials. Moreover, in contrast to semantic priming which usually employs rather large stimulus sets (McNamara, 2005), we used a substantially smaller set with only four target icons. This, again, conforms to the stimulus set size in interference paradigms (e.g., Stroop, flanker, response priming), and might be responsible for automatic processes leading to benefits even at a way shorter SOA (100 ms) than in the semantic priming literature (400 ms, McNamara, 2005). As already mentioned above, such a small

stimulus set might generally reduce latencies of processes because all target stimuli are already actively kept in working memory.

Before closing our discussion of the cost-benefit partitioning, we would like to finally draw attention to the issue of strategic effects and contingency in a crossmodal context. Our findings revealed that the degree of contingency between auditory primes and visual targets is also a relevant factor for crossmodal facilitation and interference. As already described earlier, highly contingent presentation might lead to strategic listening to prime words. Thereby, semantic representations might be activated far beyond non-contingent presentation and they might also be processed more thoroughly. Most importantly, participants might also actively start looking for the respective visual target, since overall this would improve performance throughout the experiment. In numerous congruent trials, benefits would be expected for this strategy, whereas in relatively few incongruent trials, consideration of and searching for a visual target that will not appear would plausibly cause costs. However, when separately determining average benefits and costs on a trial basis, we assume that facilitation and interference are about equally high. Suitably, our findings in Experiments 5A and 5B support the assumption that increased contingency led to a symmetrical increment of crossmodal semantic priming effects. [28] This is a similar pattern like the findings Logan (1980) achieved when manipulating contingency in a Stroop-like task with an SOA of 0 ms, but it is somehow opposed to findings from semantic priming, where strategic effects are typically only revealed for longer SOAs above 400 ms (McNamara, 2005). In any case, it is still unclear whether contingency and stimulus-stimulus compatibilities contribute to facilitation and interference independently of each other with contingency causing qualitatively different effects, or whether, eventually, only a quantitative change in effects might occur by a correlation between semantics of primes and targets. Even without contingency we overall found small interference, and this might

[28] Overall, we assume that the level of contingency would be positively correlated with the amount of benefits and costs. Please note, however, that in extremely rare cases of invalid semantics the pattern would become asymmetrical: People would most probably rely almost completely on spoken words. At a certain point, this would probably lead to ceiling effects in congruent trials. In rare and unexpected incongruent trials, responses would be slowed remarkably and also error rates would increase considerably.

already constitute a sufficient basis for high contingency to increase interference to a similar extent as facilitation. Accordingly, our findings do not definitely specify, whether contingency affects crossmodal benefits and costs quantitatively or qualitatively.

Summing up the previous considerations, we applied a somehow hybrid approach in between interference paradigms and semantic priming. This led to pronounced facilitation at extremely short SOAs, and enabled us to identify stimulus-stimulus compatibilities as an underlying process for crossmodal semantic priming. Noticeably, increased contingency also revealed exceptionally fast effects on task performance. It remains for future research to investigate the exact frame conditions under which fast, crossmodal facilitation via semantic representations occurs. In this regard, gradually shifting towards a crossmodal semantic priming paradigm by increasing the target icon set might, for instance, constitute a promising approach. Besides, this would shed additional light on Chen and Spence's (2011) failure to find crossmodal semantic priming effects without target identification.

After having closely discussed some core issues regarding crossmodal semantic priming in the preceding sections, we pinpoint several open issues that would be worth closer investigation in the next section.

7.1.3 Open questions and future directions regarding crossmodal semantic priming

Besides some aspects that have already been addressed in the respective text passages, some further experimental modifications are worth consideration in the future. For example, in order to receive a more precise picture of the time course of our crossmodal effects, varying SOAs could be applied. Especially when contingency between primes and targets is also varied, longer SOAs might help to separate automatic semantic effects and conscious, strategic exploitation of spoken words. Negative SOAs might be especially interesting with regard to another open issue: Could spoken words directly affect visual search processes that are already in progress? Suitably, and with regard to attentional capture by spoken denotations, it might be worth to slightly adapt our design to that of Salverda and Altmann (2011). We would assume that even in well-examined visual scenes spoken words would instantaneously enhance

change detection regarding respective visual objects. Applying far lower SOAs than Salverda and Altmann would allow for testing this assumption.

So far, we have only tested effects of time-compression for verbal stimuli that were anyway quite short and that sounded differently. Under these special circumstances, already for words of normal duration disambiguation might have occurred extremely early, thereby concealing beneficial effects of time-compression. We would assume that time-compression would become more effective for rather similar-sounding utterances. The same would probably hold true, if a larger set of spoken words was applied since in this situation the first syllable would no longer be sufficient for disambiguation and exclusive activation of one semantic representation. In this regard, it might also be worth to only repeat each target item once per experiment, since denotations and objects would then be completely unexpected and the meaning of spoken words would not already be revealed by the first syllable. This could, for example, be achieved by using a fixed set of filler items and having participants classify the only new item which is not a well-known filler item. Moreover, such an experiment with unrepeated and unexpected targets would allow for further inferences about the involvement of active working memory items. But even when keeping the current experiment design with four target items, changing from a repeated presentation of a small set of familiar and visually unvarying target stimuli to visual stimulus materials with more variation might reveal insights whether the crossmodal effects we found are confined to rather artificial setups, or whether they are more universally valid. For example, presenting photographs of objects that were taken from various perspectives or presenting different instances of one category would be reasonable steps into this direction.

Furthermore, it remains to be investigated, whether crossmodal semantic effects are limited to a perfect match of semantic representations, or whether effects are more general in nature. Effects might even extend to semantic associations between spoken words and visual targets (equivalent to semantic priming). For example, it might be interesting, whether the prime word ambulance would also improve the detection of a police car, or whether the prime word animal would increase the detection of a deer and a dog.

Finally, we suggest the investigation whether spoken words also have the potential to directly affect motor responses (in visual tasks). For example, it might be interesting whether compatible and incompatible action instructions like, for example, “push” or “pull” would enhance and deteriorate motor responses like pushing and pulling. We assume that also in such a case spoken words have the potential to prime motor responses.

After having discussed our results and future directions of crossmodal effects of spoken words on visual perception in general, we return to our applied research question. In the following, we will elaborate on insights that can additionally be derived from a speech warning perspective.

7.1.4 Discussion of the results from an in-car speech warning perspective

In this chapter, we discuss insights that could be derived from our findings of crossmodal effects from spoken words on visual perception for recommendations on speech warnings. In this regard, our literature overview in Chapters 1 and 2 resulted in the principal hypothesis that spoken words constitute a highly promising alternative for time-critical road hazard situations. However, there was a lack of detailed knowledge about the cause for considerably good performance when speech warnings were presented. Therefore, our series of experiments was supposed to support clarification of the hitherto open issue whether speech could immediately and crossmodally influence visual scene assessment via semantics, and thus support the transfer of diverse urgent information from car to driver in time-critical situations. Compatibly, our hypotheses, which were presented in Section 3.4 were also developed against this background. Hence, it is worthwhile to discuss the relevance of our findings also from a perspective of applied research.

First of all, we found evidence that congruent spoken words can directly and crossmodally enhance visual performance as compared to neutral words or silence. Clarification of this basic issue could importantly affect warning design in general, since our findings indicate that suitable speech warnings might improve visual hazard detection beyond no warning or an unspecific master alert (sound or word).

Besides, timing constitutes another important aspect which has to be considered in the context of warnings. In this regard we found reliable effects of spoken words on the performance in the visual task – even though the auditory information was only presented 100 ms prior to the visual target. This was supposed to roughly imitate a driving situation in which an assistance system provides time-critical information just before a driver visually notices a hazard. For example, if a driver was distracted, drowsy, or inattentive, or if technical systems had more rapidly detected and appraised a critical incident than the driver, the presentation of a speech warning might start a few milliseconds before the driver starts visual situation assessment. In a first step, we showed that crossmodal semantic priming could even be effective under such extreme conditions, thereby initially supporting usage of speech warning for information transfer in time-critical situations. Nevertheless, in real-world scenarios this offset of warning and event might vary considerably, and hence further investigation on this influence of timing is needed here.

Regarding our intention to investigate speech warnings especially for highly complex traffic situations, it was remarkable that higher visual perceptual load actually increased the crossmodal effects of spoken words (Exps. 1A and 1B). Even though this effect might not have been caused exclusively by a manipulation of perceptual complexity, but also by an altered contribution of working memory load (see Section 7.1.2), our findings point towards speech warnings being a promising alternative – particularly under more demanding conditions with only minimal spare attentional capacity. According to these findings, we have limited the investigations in our following experiments to a rather demanding visual task, and we were able to reconfirm clear crossmodal effects. This gives grounds for optimism that even in complex and adverse road situations task performance could be similarly improved by speech warnings.

In light of speech warnings, also the duration of spoken words and respective influences on effects denotes an important point. Based on the observation that people commonly tend to increase their word rate in situations requiring urgent information transfer, we initially hypothesized that time-compressed, but still understandable, words would cause more pronounced effects due to earlier disambiguation and termination. Since time compression

did not increase our crossmodal semantic priming effects, this hypothesis was not supported. Hence, we would not generally recommend applying time compression, so far. However, throughout our experiments we have presented only short words that were well known, differed in pronunciation, and only comprised a single piece of information. Under these special and somehow artificial circumstances, disambiguation might have occurred extremely early anyways, thus masking beneficial effects of time-compression. If the spoken words comprised more similar-sounding phonemes, more syllables, a variety of different or completely unexpected contents, there might, nevertheless, be a plausible chance to confirm our hypothesis. Accordingly, a final evaluation of time compression for more realistic contents of speech warnings remains for future work.

Having identified stimulus-stimulus compatibilities to cause considerable crossmodal effects of spoken words on visual objects (e.g., Exps. 2 and 4), we could exclude that a match of semantic representations with a potential response was a necessary precondition. Accordingly, mere hazard denotations might effectively warn of objects or situations by directly enhancing respective visual processing. Furthermore, the red-green-classification of target icons that we have applied from Experiment 4 onwards could resemble a driver categorizing a denoted on-road situation as critical or uncritical. Even though we chose this artificial target classification task and orthogonal responses for methodological and theoretical reasons, it happened to correspond to real-world conditions in some relevant aspects. In an on-road scenario, the semantic content of warnings might – equivalently to our task – be uncorrelated with the driver's evaluation of the situation and also with the subsequent response. Finding effect under these conditions is thus promising for specific road scenarios.

For on-road usage of speech warnings, it is especially relevant that we have repeatedly identified clear benefits and only minor costs for completely irrelevant spoken words that did not contain any contingent information about subsequent visual targets. With strategic listening or implicit learning being excluded as a necessary precondition, we are optimistic that speech warnings would have the potential to automatically support drivers even without active listening or compliance, or already at the first occurrence. Finding robust

crossmodal priming effects via semantics even under adverse conditions denotes one of our core findings for the applied warning context. In this regard, we have successfully demonstrated how basic psychological paradigms can contribute highly relevant information to applied research questions – even though such an approach might seem counterintuitive at a superficial glance. Furthermore, we point out based on our findings that semantics in speech warnings would apparently have the potential to outperform auditory master alerts or no warnings in case of a match, whereas almost no costs would be expected for a non-match.

Aside from basic, rather automatic effects, drivers would most probably start to expect suitable information after several valid warnings and try to utilize the semantic information. Analogous to our increased effects for obvious contingency (Exp. 5A vs. 5B), we would assume even higher crossmodal effects for this instance. Noticeably, reliance on speech warnings would probably increase both the major semantic facilitation and the minor semantic costs by the same extent. Besides the asymmetric and rather beneficial crossmodal semantic effects, system validity would affect the ratio of accumulated benefits and costs over time: A higher rate of suitable speech warnings would result in an increased sum of benefits and in a lower sum of costs, thereby increasing performance at the bottom line.[29] However, it is left for further investigation to clarify whether costs of incompatible warnings would still be tolerable for specific safety-critical scenarios.

After a period of system usage, drivers will hopefully have experienced that speech warnings match respective situations quite well, and start listening intentionally. Under such circumstances, we suggest that speech warnings could also be presented in a time-compressed way. Referring to our findings, semantic congruency effects would probably remain comparably high and overall RTs would not be slowed for these stimuli as compared to spoken words with normal duration. As already reasoned earlier, effects of time

[29] For extremely high rates of contingency drivers might start to completely rely on auditory information thus resulting in a disproportionate performance decrease for incompatible warnings. This could manifest, for example, in extremely retarded responses to an actual hazard, if it has not been denoted in the warning. We do not further address this issue here, since it constitutes a very general research question in applied human factors literature and it is not essential to specifically revisit it for crossmodal effects of speech warnings.

compression might even become more pronounced for a larger set of (longer) speech warnings, thus creating room for further investigations with a special focus on the influence of contingency.

In order to address the transferability of our results to more realistic driving conditions, we investigated driving maneuvers instead of key-press reactions in a first step. Spoken words crossmodally affected visual processing even when measuring complex motor performance like swerving and braking. Hence, we are optimistic that our general findings have practical relevance on actual driving performance in critical on-road situations. Eventually, accident severity would be reduced, or accidents might even be avoided completely. Potential limitations in our experiments might be that it was clearly defined in advance which out of two maneuvers should be performed in which situation, and, besides, maneuvers were quite well practiced. Under natural driving conditions adequate responses might be less obvious and habitual, thereby leaving room for further investigations under more ambiguous conditions.

At this point, it is furthermore worth to reconsider the framework proposed by Ho and Spence (2008) about driving as a special behavioral phenomenon, which comprises perceptual, decisional (attentional), and response-related components (see Chapter 1). Spence and Ho (2008) claimed that "the future design of multisensory warning signals should … be optimized … by combining the knowledge regarding each of the sub-processes involved in multisensory attention, event perception, response selection and execution" (p. 533-534). By our series of experiments, we have addressed each of these three components: Firstly, in all our tasks we have investigated influences on visual perception, which is dominant in the driving context. Above all the signal detection approach applied in Experiment 2 provided valuable findings on crossmodally enhanced visual detection sensitivity. Moreover, the application of two levels of perceptual load in the visual task could be assigned to the perceptional sub-component, but eventually also to the attentional processes, since visual search difficulty might also affect working memory processes (see our reasoning in Section 7.1.2). Besides, we altered the information that auditory primes revealed about subsequent targets (contingency) from absolutely non-informative to rather predictable. Thereby, we could infer how attention and strategic listening affected crossmodal

semantic priming. Last but not least, we have investigated, whether more complex response selection and response execution concealed or even eliminated perceptional effects. In this regard, we could confirm the robustness of our effects even for braking or lane change responses. Even though these sub-processes and their frame conditions need to be further scrutinized in the future, we claim that we have, in summary, conducted comprehensive research by accumulating knowledge about each of the central psychological components regarding driving performance. By combining our literature review and our own findings, we are able to considerably support the choice of information presentation modality, code, density, and timing of in-car warnings.

In summary, speech warnings can be recommended due to their general potential to automatically affect visual processing, especially under high perceptual load. If semantic representations of spoken words and visual objects match, considerable and immediate benefits can be expected, whereas in case of a mismatch interference seems to be less pronounced. The robust performance increase for compatible speech warnings as compared to non-specific or no warnings points towards an important safety benefit, and hence a viable proposition for a successful real-world application. Moreover, time-compressed speech warnings might improve responses for warnings that are reliably compatible with the actual hazard. Eventually, application of time compression would especially improve longer, polysyllabic warnings.

7.1.5 Open questions and future directions for speech warnings

Even though we made initial, important steps towards real-world in-car speech warnings by deliberately investigating crossmodal priming effects via semantic representations, there remain a few open issues regarding direct transferability of our results to more realistic driving conditions.

The first aspect addressed here is our small set of only four target icons which had to be learned and kept active in working memory. On the one hand, one could argue that drivers have a general latent knowledge about critical road incidents as against other irrelevant or uncritical objects. Compatibly, our target set could be regarded as a subset of those potentially critical and plausible objects with the distractors standing for irrelevant objects in the

surroundings. On the other hand, the limited number of target items is quite artificial and differs from real on-road situations in which one out of numerous potential events might occur without a driver even considering it beforehand. For this matter, it remains unclear how well the controlled implementation of our experimental task would actually reflect real-world circumstances. Driving simulator experiments with more realistic situations would, for instance, provide further insights here.

Besides, drivers would experience speech warnings for critical situations only once in a while. On the contrary, in all our experiments participants experienced spoken words every few seconds in (almost) each trial. For words and their representations that have currently been activated their repeated crossmodal activation might be facilitated. It is left for future work to investigate whether such an "artificial" pre-activation of semantics constitutes a necessary precondition for the extremely rapid crossmodal effects that we have found. With regard to low hazard and hence warning frequency, it would, moreover, be highly interesting to assess whether additional phenomena (e.g., startling effects) might become relevant for unexpected and abrupt presentation of speech.

Furthermore, our findings open up some more distantly related possibilities for future work on crossmodal semantic priming effects. For example, intentional generation of driver distraction could constitute a useful component in the assessment of the alerting character of spoken words in comparison with visual warning presentation. In this regard it should be noticed that we found remarkable benefits even though participants expected the visual task and were well prepared to perform it, and even though they tried to ignore the irrelevant, congruent spoken words. If the conditions for speech warnings were more realistic and less adverse (e.g., alerting and concurrently informative characteristics of speech warnings), we would overall expect even more pronounced performance benefits. In this regard, distraction by secondary tasks and the usage of eye-tracking measures might especially help to uncover whether a denoted, critical object would be focused and attended sooner, or whether it would be focused and attended for a different period of time, after a semantically compatible prime word. Besides, it might be interesting whether

spatially presented verbal warnings about critical objects would lead to additive combined effects from both sources of information.

Moreover, it would be worth investigating how semantic priming of responses via action suggestions would influence drivers' behavior. For example, it is left for future research to investigate whether the warning "swerve left" might be more effective than "deer (coming) from right". In this regard, actions might also be tonally coded since they are typically only a few response alternatives while driving. If this was effective, such a tone could then be combined with a speech warning and would eventually lead to a fast transfer of information about the situation as well as the suggested action.

One issue that has been addressed in Section 1.2.4, is the combination of modalities. With regard to speech warnings, probably the visual modality would suitably complement time-compressed speech warnings, since it is not transient and could be presented for a longer period of time in case that an auditory warning would be missed. Nevertheless, it remains unclear whether such a combination would be beneficial under extremely time-critical circumstances and, hence, further investigation on this issue would be required.

In this chapter, we have introduced a few ideas for promising future work based on our findings on crossmodal semantic priming. In summary, concerning the transferability of our results we would recommend two major working points: A larger set of target objects and more (similar) spoken denotations would be beneficial, as well as considerably reducing the warning frequency and expectancy for at least some objects. Thereby, pre-activation of semantic representations right before the presentation of respective warnings could be avoided. This should, however, neither distract from the concrete relevance of our findings, nor from the reliable foundations that we laid by our experiments. At this point it is worth mentioning, that beyond the domain of in-car warnings even a much broader range of applications (e.g., aviation, nuclear power plants, smart objects) could benefit from the consideration of our findings regarding automatic crossmodal links between verbal presentation and visual performance.

7.2 Conclusion

With this thesis, we contributed to the issue how a wide variety of different information could be suitably presented to drivers in time-critical road hazard situations. From reviewing relevant literature on this applied issue we could derive that the usage of the auditory modality, and especially the presentation of speech warnings is a generally promising approach for urgent and specific information transfer. However, neither in applied, nor in basic cognitive psychology literature reliable information was provided whether concordant speech essentially has the potential to instantaneously support processing of visual scenes. By contrast, the results of the experiments conducted for this thesis provide evidence that spoken words can automatically enhance the detection (Exp. 2), identification (Exps. 1A, 1B, 3A, and 3B), or classification (Exps. 4, 5A, and 5B) of semantically congruent visual targets. We derived this interpretation since effects occurred even if semantics were completely task irrelevant and did not contain any contingent information about subsequent targets (Exps. 1A, 1B, 2, 3A, 3B, 4, and 5A), and even if stimulus-response compatibilities were ruled out by usage of orthogonal responses (Exps. 2, 4, 5A, and 5B). Importantly, crossmodal semantic priming occurred at an extremely short SOA of only 100 ms even though different paradigms, different dependent variables, and different materials were employed. Furthermore, we found evidence that higher perceptual load in the visual task (Exps. 1A and 1B) increased crossmodal priming benefits, whereas higher contingency between prime and target increased both benefits and interference (Exp. 5A vs. 5B). With regard to time-compression, spoken words with a considerably shorter duration (30 % of the normal duration) effectively led to shorter RTs for slightly predictive than for completely irrelevant spoken word primes (Exp. 3A vs. 3B), thereby opening up positive future perspectives for the applied context. Last but not least, we could replicate considerable crossmodal semantic priming effects even under dynamic driving simulation conditions when measuring performance in complex driving maneuvers.

In a nutshell, our findings revealed that spoken words could automatically enhance processing of visual targets via crossmodal stimulus-stimulus compatibilities. Even though primes and targets were physically

diverse, a direct crossmodal link for matching semantic representations seemed to improve performance. These results strongly support the application of speech warnings in time-critical road situations when information needs to be immediately transferred to the driver. Seen from a bird's eye view, we demonstrated a successful linkage between basic and applied research questions and methods with mutual benefits and synergies for both fields.

References

Aldrich, F. K., & Parkin, A. J. (1989). Listening at speed. *British Journal of Visual Impairment*, *7*, 16–18.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*, 106–111.

Assenmacher, S. (2009). simTD: Field operational test for determining the effectiveness of cooperative systems. In *Proceedings of the 16th World Congress on Intelligent Transport Systems*. Stockholm, Sweden.

Baldwin, C. L. (2007). Acoustic and semantic warning parameters impact vehicle crash rates. In G. P. Scavone (Ed.), *Proceedings of the 13th International Conference on Auditory Display* (pp. 143–145). Montréal, Canada: Schulich School of Music, McGill University.

Bargh, J. A. (1992). The ecology of automaticity: Toward establishing the conditions needed to produce automatic processing effects. *The American Journal of Psychology*, *105*, 181–199.

Bargh, J. A. (1994). The Four Horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In J. R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1–40). Hillsdale, NJ: Erlbaum.

Baxter, B. (1942). A study of reaction time using factorial design. *Journal of Experimental Psychology*, *31*, 430–437.

Belz, S. M., Robinson, G. S., & Casali, J. G. (1999). A new class of auditory warning signals for complex systems: Auditory icons. *Human Factors*, *41*, 608–618.

Blattner, M. M., Sumikawa, D. A., & Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *Human-Computer Interaction*, *4*, 11–44.

Bodner, G. E., & Masson, M. E. J. (2003). Beyond spreading activation: An influence of relatedness proportion on masked semantic priming. *Psychonomic Bulletin & Review*, *10*, 645–652.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer. http://www.praat.org/.

Brewster, S. A., Wright, P. C., & Edwards, A. D. N. (1992). A detailed investigation into the effectiveness of earcons. In G. Kramer (Ed.), *Proceedings of the First International Conference on Auditory Display* (pp. 471–498). Santa Fe, NM: Addison-Wesley.

Brown, S. B. (2005). *Effects of haptic and auditory warnings on driver intersection behavior and perception*. Unpublished master's thesis, Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

Brown, T. L., Joneleit, K., Robinson, C. S., & Brown, C. R. (2002). Automaticity in reading and the Stroop task: Testing the limits of involuntary word processing. *The American Journal of Psychology*, *115*, 515–543.

Buxton, W., Gaver, W. W., & Bly, S. (1994). Auditory interfaces: The use of non-speech audio at the interface. Unfinished book retrieved from: http://www.billbuxton.com/Audio.TOC.html (March 2010).

Calvert, G., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

Campbell, J. L., Carney, C., & Kantowitz, B. H. (1998). *Human factors design guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO)*. (Federal Highway Administration Technical Report No. FHWA-RD-98-057). Retrieved from http://www.fhwa.dot.gov/publications/research/safety/98057/toc.cfm.

Campbell, J. L., Richard, C. M., Brown, J. L., & McCallum, M. (2007). *Crash warning system interfaces: Human factors insights and lessons learned*. (National Highway Traffic Safety Administration Technical Report No. DOT HS 810 697). Washington, DC: U.S. Department of Transportation.

Cao, Y., Mahr, A., Castronovo, S., Theune, M., Stahl, C., & Müller, C. (2010). Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction. In *Proceeding of the 15th International Conference on Intelligent User Interfaces* (pp. 239–248). New York, NY, USA: ACM.

Castronovo, S., Mahr, A., & Müller, C. (2013). What, where, and when? Intelligent presentation management for automotive human machine interfaces and its application. In S. Yamamoto (Ed.), *Human Interface and the Management of Information, Part II, HCII 2013, LNCS 8017* (pp. 460–469). Berlin / Heidelberg: Springer.

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*, 389–404.

Chen, Y.-C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance*, 1–15.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N., & Barron, A. (1987). Cross-modal, auditory-visual Stroop interference and possible implications for speech memory. *Perception & Psychophysics*, *41*, 393–401.

De Houwer, J. (2003). On the role of stimulus-response and stimulus-stimulus compatibility in the Stroop effect. *Memory & Cognition*, *31*, 353–359.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Dien, J., & Santuzzi, A. M. (2005). Application of repeated-measures ANOVA to high-density ERP datasets: A review and tutorial. In T. Handy (Ed.), . Cambridge: MIT Press.

Dingler, T., Lindsay, J., & Walker, B. N. (2008). Learnabilitiy of sound cues for environmental features: Auditory icons, earcons, spearcons, and speech. In *Proceedings of the 14th International Conference on Auditory Display* (pp. 1–6). Paris, France.

Dingus, T. A., & Hulse, M. C. (1993). Some human factors design issues and recommendations for automobile navigation information systems. *Transportation Research Part C: Emerging Technologies*, *1*, 119–131.

Dingus, T. A., Hulse, M. C., Jahns, S. K., Alves-Foss, J., Confer, S., Rice, A., … Sorenson, D. (1996). *Development of human factors guidelines for advanced traveler information systems and commercial vehicle*

operations: Literature review. (National Highway Traffic Safety Administration Technical Report No. FHWA-RD-95-153). McLean, VA: U.S. Department of Transportation.

Dingus, T. A., McGehee, D. V, Manakkal, N., Jahns, S. K., Carney, C., & Hankey, J. M. (1997). Human Factors Field Evaluation of Automotive Headway Maintenance/Collision Warning Devices. *Human Factors*, *39*, 216–229.

Donges, E. (1978). A Two-Level Model of Driver Steering Behavior. *Human Factors*, *20*, 691–707.

Driver, J., & Spence, C. (1998). Crossmodal attention. *Current Opinion in Neurobiology*, *8*, 245–253.

Duncan-Johnson, C., & Kopell, B. (1981). The Stroop effect: brain potentials localize the source of interference. *Science*, *214*, 938–940.

Edworthy, J., Stanton, N. A., & Hellier, E. (1995). Warnings in research and practice. *Ergonomics*, *38*, 2145–2154.

Elliott, E. M., Cowan, N., & Valle-Inclan, F. (1998). The nature of cross-modal color-word interference effects. *Perception & Psychophysics*, *60*, 761–767.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*, 101–118.

Eysenck, M. W., & Keane, M. T. (2000). *Cognitive psychology: A student's handbook* (4th ed.). New York, NY, USA: Psychology Press.

Fagerlönn, J. (2007). Expressive musical warning signals. In Gary P. Scavone (Ed.), (pp. 430–436). Montréal, Canada: Schulich School of Music, McGill University.

Fagerlönn, J. (2010). Distracting effects of auditory warnings on experienced drivers. In *Proceedings of the 16th International Conference on Auditory Displays*. Washington, DC, USA.

Fagerlönn, J., Lindberg, S., & Sirkka, A. (2012). Graded auditory warnings during in-vehicle use: using sound to guide drivers without additional

noise. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 85–91). New York, NY, USA: ACM.

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Psychological Bulletin*, *72*.

Franz, V. H., & Loftus, G. R. (2012). Standard errors and confidence intervals in within-subjects designs: generalizing Loftus and Masson (1994) and avoiding the biases of alternative accounts. *Psychonomic bulletin & review*, *19*, 395–404.

Fricke, N. (2009). *Gestaltung zeit- und sicherheitskritischer Warnungen im Fahrzeug*. (Doctoral dissertation). Technische Universität Berlin.

Gaver, W. W. (1986). Auditory icons: using sound in computer interfaces. *Human-Computer Interaction*, *2*, 167–177.

Gaver, W. W. (1989). The SonicFinder: An interface that uses auditory icons. *Human-Computer Interaction*, *4*, 67–94.

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General*, *118*, 13–42.

Goldstein, E. B. (2002). Auditive Sprachwahrnehmung. In M. Ritter (Ed.), *Wahrnehmungspsychologie* (2nd ed., pp. 465–497). Heidelberg: Spektrum.

Graham, R. (1999). Use of auditory icons as emergency warnings: evaluation within a vehicle collision avoidance application. *Ergonomics*, *42*, 1233–1248.

Gray, R. (2011). Looming auditory collision warnings for driving. *Human Factors*, *53*, 63–74.

Haas, E. C., & Edworthy, J. (1999). The perceived urgency and detection time of multitone auditory signals. In N. A. Stanton & J. Edworthy (Eds.), *Human factors in auditory warnings* (pp. 129–149). Aldershot: Ashgate.

Hellier, E., & Edworthy, J. (2000). Auditory warnings in noisy environments. *Noise and Health*, *2*, 27–39.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America*, *26*, 530–537.

Ho, C., Reed, N., & Spence, C. (2006). Assessing the effectiveness of "intuitive" vibrotactile warning signals in preventing front-to-rear-end collisions in a driving simulator. *Accident Analysis & Prevention*, *38*, 988–996.

Ho, C., Reed, N., & Spence, C. (2007). Multisensory in-car warning signals for collision avoidance. *Human Factors*, *49*, 1107–1114.

Ho, C., & Spence, C. (2005). Assessing the effectiveness of various auditory cues in capturing a driver's visual attention. *Journal of Experimental Psychology: Applied*, *11*, 157–174.

Ho, C., & Spence, C. (2008). *The multisensory driver: Implications for ergonomic car interface design*. Aldershot: Ashgate.

Hoffman, J. D., Lee, J. D., & Hayes, E. M. (2003). Driver preference of collision warning strategy and modality. In *Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*. Park City, Utah.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Horowitz, A. D., & Dingus, T. A. (1992). Warning signal design: A key human factors issue in an in-vehicle front-to-rear-end collision warning system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 36, pp. 1011–1013). Atlanta, Georgia.

Horrey, W. J., & Wickens, C. D. (2004). Driving and Side Task Performance: The Effects of Display Clutter, Separation, and Modality. *Human Factors: The Journal of the Human Factors and Ergonomics Society* , *46* , 611–624.

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics*, *72*, 1736–1741.

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review*, *15*, 548–554.

ISO 9241-11:1998(E), Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability. (1998). Geneva, Switzerland: International Organization of Standards.

ISO/TR 16352:2005(E), Road vehicles - Ergonomic aspects of in-vehicle presentation for transport information and control systems - Warning systems. (2005). Geneva, Switzerland: International Organization of Standards.

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, *5*, 52–65.

Jamson, A. H., & Merat, N. (2005). Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*, 79–96.

Jeon, M., Davison, B. K., Nees, M. A., Wilson, J., & Walker, B. N. (2009). Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle technologies. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 91–98). New York, NY, USA: ACM.

Johnston, W. L., Mayyasi, A. M., & Heard, M. F. (1971). The effectiveness of a vibrotactile device under conditions of auditory and visual loading. In *Proceedings of the 9th Annual Symposium of the Survival and Flight Equipement Association* (pp. 36–41). Las Vegas, NV.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 369–381.

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 497 – 509.

Kantowitz, B. H., Triggs, T. J., & Barnes, V. E. (1990). Stimulus-Response Compatibility and Human Factors. In R. W. Proctor & T. G. Reeve (Eds.), *Advances in Psychology* (Vol. 65, pp. 365–388). North-Holland.

Keetels, M., & Vroomen, J. (2011). Sound affects the speed of visual processing. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 699–708.

Labiale, G. (1990). In-car road information: Comparisons of auditory and visual presentations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 623–627). Orlando, FL.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*, 742–744.

Larsen, J. D., Baddeley, A., & Andrade, J. (2000). Phonological similarity and the irrelevant speech effect: implications for models of short-term verbal memory. *Memory*, *8*, 145–57.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 451–468.

Lavie, N., & Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. *Psychological Science*, *8*, 395–398.

LeBlanc, D., Sayer, J., Winkler, C., Ervin, R., Bogard, S., Devonshire, J., … Gordon, T. (2006). *Road departure crash warning field operational test*. Washington, DC.

Lee, J. D., Gore, B. F., & Campbell, J. L. (1999). Display alternatives for in-vehicle warning and sign information: message style, location, and modality. *Transportation Human Factors*, *1*, 347–375.

Lee, J. D., Hoffman, J. D., & Hayes, E. M. (2004). Collision warning design to mitigate driver distraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 65–72). New York, NY, USA: ACM.

Lee, J. D., McGehee, D. V, Brown, T. L., & Reyes, M. L. (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Human Factors*, *44*, 314–334.

Lee, S. E., Perez, M. A., Doerzaph, Z. R., Stone, S. R., Neale, V. L., Brown, S. B., … Dingus, T. A. (2007). *Intersection collision avoidance - violation project: Final project report*. (National Highway Traffic Safety Administration Technical Report No. DOT HS 810 749). Washington, DC: U.S. Department of Transportation.

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology*, *12*, 523–553.

Logan, G. D. (1996). The CODE Theory of visual attention: An integration of space-based and object-based attention. *Psychological Review*, *103*, 603–649.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*, 166–174.

Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S. D., Sarter, N. B., & Sebok, A. (2013). Supporting interruption management and multimodal interface design: Three meta-analyses of task performance as a function of interrupting task modality. *Human Factors*, *55*, 697–724.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: Verbal but not visual cues enhance visual detection. *PLoS ONE*, *5*, e11452.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*, 170–186.

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 14196–14201.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*, 163–203.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Mahr, A., & Wentura, D. (2014). Time-compressed spoken word primes crossmodally enhance processing of semantically congruent visual targets. *Attention, Perception, & Psychophysics*, *76*, 575-590.

Mangold, S. S. (1982). Nurturing high self-esteem in visually handicapped children. In S. S. Mangold (Ed.), *A teacher's guide to special needs of blind and visually handicapped children.* (p. 153). New York: American Foundation for the Blind.

Marshall, D. C., Lee, J. D., & Austria, P. A. (2007). Alerts for in-vehicle information systems: Annoyance, urgency, and appropriateness. *Human Factors*, *49*, 145–157.

Martin-Emerson, R., & Wickens, C. D. (1997). Superimposition, symbology, visual attention, and the head-up display. *Human factors*, *39*, 581–601.

Math, R., Mahr, A., Moniri, M. M., & Müller, C. (2012). OpenDS: A new open-source driving simulator for research. In *Adjunct Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Appilcations* (pp. 7–8). Portsmouth, NH, USA.

May, P. A., & Wickens, C. D. (1995). The role of visual attention in head-up displays: Design implications for varying symbology intensity. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 50–54).

Mazza, V., Turatto, M., Rossi, M., & Umiltà, C. (2007). How automatic are audiovisual links in exogenous spatial attention? *Neuropsychologia*, *45*, 514–522.

McDonald, J. J., Teder-Salejarvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. *Nature*, *407*, 906–908.

McKeown, D., & Isherwood, S. (2007). Mapping candidate within-vehicle auditory displays to their referents. *Human Factors*, *49*, 417–428.

McKeown, D., Isherwood, S., & Conway, G. (2010). Auditory displays as occasion setters. *Human Factors*, *52*, 54–62.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition.* New York: Psychology Press.

McNamara, T. P., & Holbrook, J. B. (2012). Semantic memory and priming. In I. B. Weiner (Ed.), *Handbook of psychology, Vol. 4, Experimental psychology* (2nd ed., pp. 449–471). New York: Wiley.

McRuer, D. T., Allen, R. W., Weir, D. H., & Klein, R. H. (1977). New results in driver steering control models. *Human Factors*, *19*, 381–397.

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review*, *110*, 422–471.

MIL-STD-14-72F. (1999). United States of America. Retrieved from http://www.public.navy.mil/navsafecen/Documents/acquisition/MILSTD1472F.pdf.

Mohebbi, R., Gray, R., & Tan, H. Z. (2009). Driver reaction time to tactile and auditory rear-end collision warnings while talking on a cell phone. *Human Factors*, *51*, 102–110.

Moors, A. (2010). Automatic constructive appraisal as a candidate cause of emotion. *Emotion Review*, *2*, 139–156.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, *132*, 297–326.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: the contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological science*, *12*, 9–17.

Murray, M. M., & Spierer, L. (2009). Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans. *Hearing Research*, *258*, 121–133.

Nees, M. A., & Walker, B. N. (2011). Auditory displays for in-vehicle technologies. *Reviews of Human Factors and Ergonomics*, *7*, 58–99.

Noyes, J. M., Hellier, E., & Edworthy, J. (2006). Speech warnings: a review. *Theoretical Issues in Ergonomics Science*, *7*, 551–571.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*, 316–333.

Olivers, C. N. L., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: visual working memory content affects visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1243–1265.

Patterson, R., & Milroy, R. (1979). *Existing and recommended levels for the auditory warnings on civil aircraft.*

Perea, M., & Rosa, E. (2002). Does the proportion of associatively related pairs modulate the associative priming effect at very brief stimulus-onset asynchronies? *Acta Psychologica*, *110*, 103–124.

Petocz, A., Keller, P. E., & Stevens, C. J. (2008). Auditory warnings, signal-referent relations, and natural indicators: Re-thinking theory and application. *Journal of Experimental Psychology: Applied*, *14*, 165–178.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception and Psychophysics*, *16*, 283–290.

Pierowicz, J., Jocoy, E., Lloyd, M., Bittner, A., & Pirson, B. (2000). *Intersection Collision Avoidance Using ITS Countermeasures*. (National Highway Traffic Safety Administration Technical Report No. DOT HS 809 171). Washington, DC: U.S. Department of Transportation.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*, 157–171.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*, 965–966.

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied*, *6*, 31–43.

Reeves, L. M., Lai, J., Larson, J. A., Oviatt, S., Balaji, T. S., Buisine, S., … Wang, Q. Y. (2004). Guidelines for multimodal user interface design. *Communications of the ACM*, *47*, 57–59.

Roelofs, A. (2005). The visual-auditory color-word Stroop asymmetry and its time course. *Memory & Cognition*, *33*, 1325–1336.

Salverda, A. P., & Altmann, G. T. M. (2011). Attentional capture of objects referred to by spoken language. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1122–1133.

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, *51*, 982–1010.

Schneider, T. R., Engel, A. K., & Debener, S. (2008). Multisensory identification of natural objects in a two-way crossmodal priming paradigm. *Experimental Psychology*, *55*, 121–132.

Scott, J. J., & Gray, R. (2008). A comparison of tactile, visual, and auditory warnings for rear-end collision prevention in simulated driving. *Human Factors*, *50*, 264–275.

Selcon, S. J., Taylor, R. M., & McKenna, F. P. (1995). Integrating multiple information sources: using redundancy in the design of warnings. *Ergonomics*, *38*, 2362–2370.

Seppelt, B., & Wickens, C. D. (2003). *In-vehicle tasks: Effects of modality, driving relevance, and redundancy*. (Aviation Human Factors Division Institute of Aviation Technical Report No. AHFD-03-16/GM-03-2). Seattle, IL: University of Illinois.

Shimada, H. (1990). Effect of auditory presentation of words on color naming: The intermodal Stroop effect. *Perceptual and Motor Skills*, *70*, 1155–1161.

Simpson, C. A., & Marchionda-Frost, K. (1984). Synthesized speech rate and pitch effects on intelligibility of warning messages for pilots. *Human Factors*, *26*, 509–517.

Simpson, C. A., & Williams, D. H. (1980). Response Time Effects of Alerting Tone and Semantic Context for Synthesized Voice Cockpit Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society* , *22* , 319–330.

Sivak, M. (1996). The information that drivers use: is it indeed 90% visual? *Perception*, *25*, 1081–1089.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical Science and Technology*, *28*, 61–70.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, *59*, 1–22.

Spence, C., & Ho, C. (2008). Multisensory warning signals for event perception and safe driving. *Theoretical Issues in Ergonomics Science*, *9*, 523–554.

Spence, C., Senkowski, D., & Röder, B. (2009). Crossmodal processing. *Experimental Brain Research*, *198*, 107–111.

Srinivasan, R., & Jovanis, P. P. (1997). Effect of Selected In-Vehicle Route Guidance Systems on Driver Reaction Times. *Human Factors*, *39*, 200–215.

Srinivasan, R., Yang, C.-Z., Jovanis, P. P., Kitamura, R., & Anwar, M. (1994). Simulation study of driving performance with selected route guidance systems. *Transportation Research C*, *2*, 73–90.

Stahlmann, R., Festag, A., Tomatis, A., Radusch, I., & Fischer, F. (2011). Starting European Field Tests for CAR-2-X Communication: The DRIVE

C2X Framework. In *Proceedings for the 18th ITS World Congress Exhibition*. Orlando, FL.

Stanton, N. A., & Edworthy, J. (1999). *Human factors in auditory warnings*. Aldershot: Ashgate.

Stevens, A. (2009). European approaches to principles, codes, guidelines, and checklists for in-vehicle HMI. In M. A. Regan, J. D. Lee, & K. L. Young (Eds.), *Driver distraction: Theory, effects, and mitigation.* (pp. 395–410). Boca Raton, FL: CRC Press.

Stevens, C. J., Brennan, D., & Parker, S. (2004). Simultaneous manipulation of parameters of auditory icons to convey direction, size, and distance: Effects on recognition and interpretation. In S. Barrass & P. Vickers (Eds.), *Proceedings of the 10th International Conference on Auditory Display*. Sydney, Australia: International Community for Auditory Display (ICAD).

Stevens, C. J., Brennan, D., Petocz, A., & Howell, C. (2009). Designing informative warning signals: Effects of indicator type, modality, and task demand on recognition speed and accuracy. *Advances in Cognitive Psychology*, *5*, 84–90.

Strandén, L., Uhlemann, E., & Ström, E. G. (2008). State of the art survey of wireless vehicular communication projects. In *15th World Congress on Intelligent Transport Systems*. New York, NY, USA.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Tan, A. K., & Lerner, N. D. (1995). *Multiple attribute evaluation of auditory warning signals for in-vehicle crash avoidance warning systems*. (National Highway Traffic Safety Administration Technical Report No. DOT HS 808 535). Washington, DC: U.S. Department of Transportation.

Tellinghuisen, D. J., & Nowak, E. J. (2003). The inability to ignore auditory distractors as a function of visual task perceptual load. *Perception & Psychophysics*, *65*, 817–828.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Tijerina, L., Jackson, J. L., Pomerleau, D. A., Romano, R. A., & Petersen, A. D. (1996). Driving simulator tests of lane departure collision avoidance systems. In *IProceedings of the ITS America Conference*. Houston, Texas.

Tijerina, L., Kiger, S., Rockwell, T., Tomow, C., Kinateder, J., & Kokkotos, F. (1996). *Heavy vehicle driver workload assessment. Task 6: Baseline data study* (p. 163). (National Highway Traffic Safety Administration Technical Report No. DOT HS 808 467). Washington, DC: U.S. Department of Transportation.

Tremblay, S., Nicholls, A. P., Alford, D., & Jones, D. M. (2000). The irrelevant sound effect: does speech play a special role? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 1750–1754.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.

Van Coile, B., Rühl, H.-W., Vogten, L., Thoone, M., Goß, S., Delaey, D., … Willems, S. (1997). Speech synthesis for the new Pan-European traffic message control system RDS-TMC. *Speech Communication*, *23*, 307–317.

Van Erp, J. B. F., & Van Veen, H. A. H. C. (2001). Vibro-tactile information presentation in automobiles. In *Proceedings of Eurohaptics 2001* (pp. 99–104).

Vilimek, R., & Hempel, T. (2005). Effects of speech and non-speech sounds on short-term memory and possible implications for in-vehicle use. In E. Brazil (Ed.), *Proceedings of the 11th International Conference on Auditory Display* (pp. 344–350). Limerick, Ireland.

Voss, M., & Bouis, D. (1979). *Der Mensch als Fahrzeugführer: Bewertungskriterien der Informationsbelastung, visuelle und auditive Informationsübertragung im Vergleich*. *FAT Schriftenreihe* (Vol. 12). Frankfurt a. M.: Forschungsvereinigung Automobiltechnik.

Walker, B. N., Nance, A., & Lindsay, J. (2006). Spearcons: speech-based earcons improve navigation performance in auditory menus. In *Proceedings of the 12th International Conference on Auditory Display* (pp. 63–68). London, UK: Department of Computer Science, Queen Mary University of London.

Wang, D.-Y. D., Pick, D. F., Proctor, R. W., & Yeh, Y. (2007). Effect of a side collision avoidance signal on simulated driving with a navigation system. In *Proceedings of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 206–211). Washington, USA.

Wentura, D., & Degner, J. (2010). Practical guide to sequential priming and related tasks. In B. Gawronski & B. K. Payne (Eds.), (pp. 95–116). New York: Guilford.

Wickens, C. D., & Gosney, J. L. (2003). Redundancy, modality, and priority in dual task interference. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 1590–1594).

Wickens, C. D., & Long, J. (1995). Object versus space-based models of visual attention: Implications for the design of head-up displays. *Journal of Experimental Psychology: Applied*, *1*, 179–193.

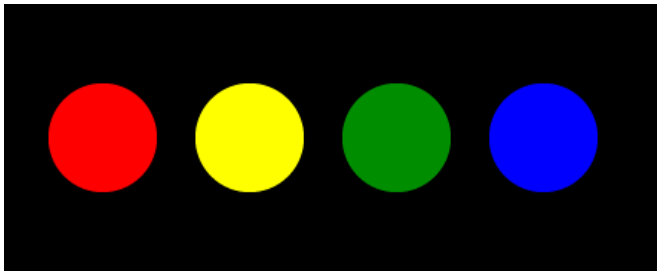
Wolf, H., Zöllner, R., & Bubb, H. (2005). Ergonomische Aspekte der Mensch-Maschine-Interaktion bei gleichzeitig agierenden Fahrerassistenzsystemen. *Zeitschrift für Verkehrssicherheit*, *3*, 119–124.

Yeh, M., Merlo, J. L., Wickens, C. D., & Brandenburg, D. L. (2003). Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue E effectiveness for display signaling. *Human Factors*, *45*, 390–407.

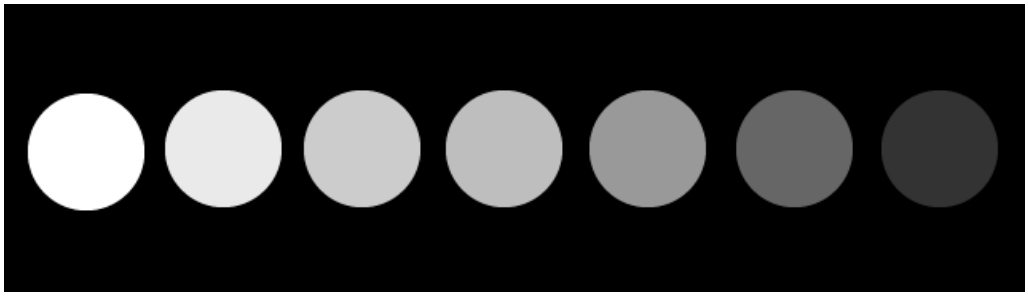
Appendices

Appendix A Visual stimuli employed in Experiments 1A and 1B

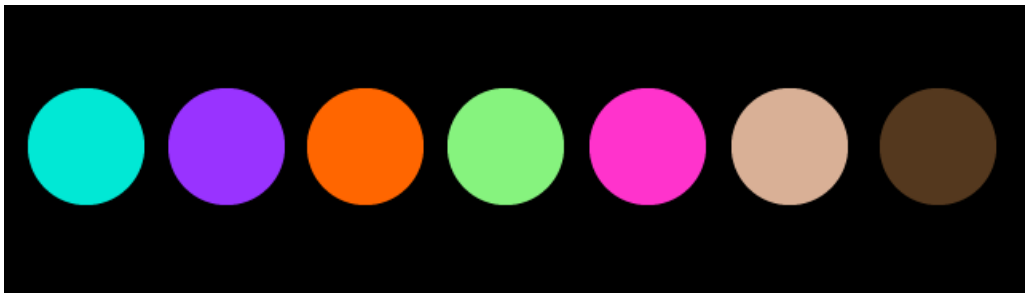
Target stimuli



Distractor stimuli low-perceptual-load condition



Distractor stimuli high-perceptual-load condition



Appendix B Recognition accuracy for spoken color words and nonwords

Identification accuracy (%) regarding the stimulus sets used in Experiments 1A and 1B (see Materials of Exps. 1 A and 1B).

		Duration		
Spoken Word		100 %	30 %	10 %
Set Experiment 1A				
Red	(rot)	100	100	100
Green	(grün)	100	100	100
Blue	(blau)	100	100	100
Yellow	(gelb)	98.5	100	100
Neutral Word	(hin)	100	100	89.2[a]
Set Experiment 1B				
Red	(rot)	100	100	100
Green	(grün)	100	100	100
Blue	(blau)	100	98.5	100
Yellow	(gelb)	100	100	100
Neutral	(*Combined*)	100	100	98.1[b]
Nonword 1	(liez)	100	100	100
Nonword 2	(nux)	100	100	81.5[c]
Nonword 3	(tän)	100	100	90.8[d]
Nonword 4	(töff)	100	100	100

Note. The sample size for assessing the identification accuracies of Set Experiment 1A and Set Experiment 1B were N=13 in both cases.

[a] response 'green' in all cases of incorrect categorizations

[b] percentage of nonword responses (incl. within nonword confusions)

[c] other nonword response in all cases of incorrect categorizations

[d] response 'green' in five cases of incorrect categorizations, other nonword in one case.

Appendix C Questionnaire after Experiment 1A

Question regarding conscious facilitation.

Wenn das gehörte Wort identisch war mit der Zielfarbe habe ich schneller reagiert als beim neutralen Wort.

1 = stimme nicht zu

2 = stimme eher nicht zu

3 = teils, teils

4 = stimme eher zu

5 = stimme zu

(Mean = 3.75, SD = 1.00, significantly above scale average (3) with $t(35) = 4.52$, $p < .001$; no significant correlation with facilitation effects)

Question regarding conscious interference.

Wenn das gehörte Wort abweichend war von der Zielfarbe habe ich langsamer reagiert als beim neutralen Wort.

1 = stimme nicht zu

2 = stimme eher nicht zu

3 = teils, teils

4 = stimme eher zu

5 = stimme zu

(Mean = 3.66, SD = 0.87, significantly above scale average (3) with $t(35) = 4.46$, $p < .001$, no significant correlation with interference effects)

Question regarding strategies.

Hast du irgendwelche Strategien verwendet?

Ja/ Nein (Wenn ja, welche?)

(Eight participants explicitly indicated that they had tried to ignore the spoken words. However, these participants did not show reduced crossmodal priming effects compared to the remaining participants, $F(1, 34) < 1$, n.s.)

Further comments

Sonstige Anmerkungen?

Appendix D Visual stimuli employed in Experiments 3A and 3B

Target stimuli

Distractor stimuli

Appendix E Visual stimuli employed in Experiments 4, 5A, and 5B

Target stimuli

Distractor stimuli

Stimuli for the practice phase

Appendix F Performance metrics recorded for steering and braking maneuvers in Experiments 5A and 5B

Due to the fact that we were mostly interested in the time until responses were initiated for both maneuver types, we restricted our analyses of Experiments 5A and 5B to the initial responses that we have logged. Besides these initial response indicators, we recorded three additional subsequent brake responses and two additional lance change responses, which were of minor interest.

Whenever braking was required, we recorded the period of time until the foot released the gas pedal (braking 1). In order to control whether potential effects continue during later aspects of the response, we additionally recorded the time until the brake pedal was initially pressed (braking 2), the time until the brake pedal was pushed down to 80% of the maximum pedal deflection (braking 3), and the time until the braking maneuver was completed by reaching the target speed of 20 km/h (braking 4).

Whenever changing to a designated lane (below a green icon) was required, we logged the time until the steering wheel was turned to 3 degrees from the center (steering 1). Besides, we also recorded the time until a car's front wheel crossed the pavement marking between the middle and the adjacent left or right lane (steering 2), and the time until the car completely reached the designated lane (steering 3).

In the interest of full transparency, we provide all RTs that we have logged for braking and steering maneuvers here in Table 12.

Table 12. Mean RTs (in ms; Standard deviation in parentheses) for all braking and steering maneuver responses that were logged in Experiments 5A and 5B (see text for details) as a function of semantic congruency conditions.

		Semantic congruency condition			
Contingency	Maneuver type	Congruent	Neutral control	Neutral silence	Incongruent
No – 5A	Braking 1	992 (179)	1089 (217)	1072 (199)	1112 (197)
	Braking 2	1184 (186)	1274 (212)	1268 (198)	1309 (195)
	Braking 3	1306 (200)	1400 (229)	1380 (218)	1424 (212)
	Braking 4	1857 (189)	1936 (233)	1930 (2016)	1976 (204)
	Steering 1	968 (181)	1053 (239)	1002 (190)	1015 (190)
	Steering 2	1975 (309)	2059 (355)	2022 (359)	2006 (284)
	Steering 3	2432 (394)	2515 (427)	2462 (400)	2468 (364)
Yes – 5B	Braking 1	914 (126)	1067 (210)	1058 (232)	1101 (174)
	Braking 2	1123 (143)	1274 (240)	1319 (230)	1342 (183)
	Braking 3	1239 (167)	1402 (287)	1431 (254)	1450 (205)
	Braking 4	1795 (151)	1953 (256)	1987 (244)	2001 (200)
	Steering 1	940 (113)	1073 (160)	1043 (163)	1119 (191)
	Steering 2	1936 (209)	2052 (242)	2092 (295)	2174 (292)
	Steering 3	2392 (276)	2541 (318)	2561 (388)	2621 (323)

Appendix G The role of maneuver type in Experiments 5A and 5B: Steering versus braking

Despite quite similar RTs for both (initial) maneuver types in Experiments 5A and 5B, overall error rates seemed to be obviously higher for the changing lanes (2.4 % and 1.8 %) than for braking maneuvers (0.7 % and 0.5 %). This might indicate differences in task difficulty or strategies. In order to provide full transparency regarding our results, we additionally present the analyses of Experiments 5A and 5B split by maneuver type. For sake of clarity and in order to avoid redundancy with Sections 6.2.2 and 6.3.2, only results involving the maneuver type factor are presented here.

Experiment 5A

Prior to aggregation, error trials and individual outliers were discarded (separately for both maneuver types) as was already denoted in Section 6.2.2. Table 13 shows mean RTs and error rates for all conditions.

Table 13. Mean RTs (in ms; error rates in parentheses) for both maneuver types in Experiment 5A as a function of semantic congruency conditions.

		Semantic congruency condition			
		Congruent	Neutral control	Neutral silence	Incongruent
Contingency	Maneuver type				
No	Braking	992 (0.3)	1089 (1.1)	1072 (1.1)	1112 (0.6)
	Steering	968 (1.9)	1053 (2.7)	1002 (1.7)	1015 (2.6)

RTs were analyzed in a 4 (Semantic congruency of auditory prime and target object: congruent vs. neutral control vs. neutral silence vs. incongruent) × 2 (maneuver type: braking vs. steering) MANOVA. The analysis revealed no significant main effect for maneuver type, $F(1, 22) = 2.75$, $p = .11$, $\eta_p^2 = .11$, but a significant interaction of semantic congruency and maneuver type, $F(3,$

20) = 9.67, $p < .001$, $\eta_p^2 = .59$. Helmert contrasts revealed that this interaction was neither based on the comparison of neutral silence trials versus all other trials types, nor on the comparison of neutral silence trials versus congruent and incongruent trials (pooled), $Fs(1, 22) < 1.20$, $ps > .29$, η_p^2s $< .05$. Instead, we found a significant interaction of maneuver types and the most important contrast of congruent versus incongruent trials. $F(1, 22) = 26.67$, $p < .001$, $\eta_p^2 = .55$. In this regard, braking and steering maneuvers did not lead to significantly different RTs in congruent trials, $t(22) < 1$, but to significantly slower RTs for braking maneuvers than for steering maneuvers in incongruent trials, $t(22) = 2.80$, $p = .01$. According to our findings, there exists an ordinal interaction between maneuver type and congruent versus incongruent semantics.

In addition, difference scores for facilitation (neutral control - congruent) and interference (incongruent - neutral control) were calculated for braking and steering performance. This denoted how the neutral control condition differed from the two relevant prime word conditions, and provided insights into potential benefits and costs. Facilitation for braking maneuvers (Mean: 97 ms) did not differ significantly from facilitation for steering maneuvers (Mean: 85 ms), $t(22) < 1$, and both indicators were significantly above zero, $ts(22) > 3.62$, $ps < .01$. In contrast, interference for braking maneuvers (Mean: 23 ms) was significantly larger than interference for steering maneuvers (Mean: - 38 ms), $t(22) = 2.76$, $p = .01$. Interference was not significantly above zero for braking maneuvers, $t(22) = 1.52$, $p = .14$, and even significantly below zero for steering maneuvers, $t(22) = -2.33$, $p = .03$.

For the error rates (see Table 13), we conducted the same analysis as for the RTs. A 4 (Semantic congruency: congruent vs. neutral control vs. neutral silence vs. incongruent) × 2 (maneuver type: braking vs. steering) MANOVA revealed no significant interaction of semantic congruency with maneuver type, $F < 1$. The main effect of maneuver type missed the level of significance, $F(1, 22) = 3.86$, $p = .06$, $\eta_p^2 = .15$. However, this slightly points

towards the steering task actually resulting in more errors (2.2 %) than the braking task (0.8 %).[30]

Experiment 5B

Analogous to Experiment 5A, error trials and individual outliers were discarded prior to aggregation as already described in Section 6.3.2. Table 14 shows mean RTs and error rates for all conditions.

Table 14. Mean RTs (in ms; error rates in parentheses) for both maneuver types in Experiment 5B as a function of semantic congruency conditions.

		Semantic congruency condition			
		Congruent	Neutral control	Neutral silence	Incongruent
Contingency	Maneuver type				
High	Braking	914 (0.3)	1067 (0.0)	1058 (0.5)	1101 (1.8)
	Steering	940 (1.0)	1073 (1.6)	1043 (2.7)	1119 (4.4)

RTs were analyzed in a 4 (Semantic congruency of auditory prime and target object: congruent vs. neutral control vs. neutral silence vs. incongruent) × 2 (maneuver type: braking vs. steering) MANOVA. We neither found a significant main effect for maneuver type, nor a significant interaction of semantic congruency and maneuver type, both Fs < 1.

In addition, difference scores for facilitation (neutral control - congruent) and interference (incongruent - neutral control) were calculated for braking and steering performance. Facilitation for braking maneuvers (Mean: 153 ms) did not differ significantly from facilitation for steering maneuvers (Mean: 133 ms), $t(22) < 1$, and both indicators were significantly above zero, $ts(22) > 5.79$, $ps < .001$. Also interference for braking maneuvers (Mean: 34 ms) did not significantly from interference for steering maneuvers (Mean: 46

[30] Please note that these error rates slightly differed from the overall braking and steering error rates. This is due to the fact that, within this analysis, they are aggregated for the semantic congruency conditions which differ in the number of trials.

ms), $|t\ (22)| < 1$. Interference was neither significantly above zero for braking, $t(22) = 1.50$, $p = .15$, nor for steering maneuvers, $t(22) = 1.71$, $p = .10$.

For the error rates (see Table 14), we conducted the same analysis as for the RTs. A 4 (Semantic congruency: congruent vs. neutral control vs. neutral silence vs. incongruent) × 2 (maneuver type: braking vs. steering) MANOVA revealed a significant main effect of maneuver type, $F(1, 22) = 10.00$, $p = .005$, $\eta_p^2 = .31$, with the steering task resulting in more errors (2.4 %) than the braking task (0.7 %).[31] The interaction between semantic congruency and maneuver type missed the level of significance, $F(3, 20) = 2.73$, $p = .07$, $\eta_p^2 = .15$. A glance at Table 13 provides a hint at a potential reason for this interaction tendency: Congruent trials possibly only led to decreased error rates as compared to neutral conditions for the overall more difficult, and hence error-prone, steering task. The respective Helmert contrast for the interaction did not contradict this assumption, but it slightly missed significance, as well, $F(1, 22) = 3.66$, $p = .07$, $\eta_p^2 = .14$.

Discussion

Overall, for braking and steering maneuvers similar result patterns with major facilitation and no or only minor interference were revealed.

For non-contingent presentation of prime words (Exp. 5A), RTs only differed for incongruent trials (but not for congruent trials) with braking leading to slower responses than steering. In this regard, eventually a speed-accuracy trade-off has surfaced: Lane changes might have been initiated faster, but might also result in more errors – especially when the irrelevant words conflicted with the visual task. However, it remains somewhat unclear why steering was even performed slower in the neutral control condition than the incongruent condition. Despite this interaction and in order to keep the analyses for Experiment 5A as straightforward as possible, we decided to average performance over both maneuver types in Section 6.2. Nevertheless, it should be kept in mind that interference was by trend positive for braking but negative for swerving maneuvers.

[31] Please note that, for the same reason as in Experiment 5A, these error rates slightly differed from the overall braking and steering error rates.

For the highly contingent presentation of prime words (Exp. 5B), only error rates revealed a significant difference between the two maneuver types. Consistent with the tendency revealed in Experiment 5A, steering involved higher error rates than braking. Maybe, this is caused by the braking task being less confusing than the steering task, where participants had to initially identify the respective lane and swerving direction. In any case, since there was no significant interaction between maneuver type and semantic congruency, we decided to simply average over both response types for our analyses in Section 6.3.