Do what I hear see?

The influence of auditory icons and earcons

on multimodal integration

in visual categorization

Dissertation in Social Sciences
at the Catholic University of Nijmegen, the Netherlands

June 2001

Myra P. van Esch-Bussemakers

Cover Design: Maarten Slooves, Grave
Cover Illustration: Joop Russon, Nijmegen
Printed by: PrintPartners Ipskamp B.V., Enschede

The three monkeys

It is said that knowledge, science and wisdom are founded on questions rather than on answers: to be or not to be?

To me the question: *Do I hear what I see?* captures the essence of this dissertation. It brings man back to his senses in our computerized world. It also reminded me of the phrase: *see, hear and speak no evil*, attributed to the famous ornamental carving of the three monkeys: Mizaru (blind), Kikazaru (deaf) and Iwazaru (dumb) that can be found on the Sacred Stable (Shinkyusha) at the Nikko Toshogu Shrine in Japan. In the Shinto religion, a monkey is believed to be the messenger of the shrines.

Inspired by my question title and the three monkeys of the Toshogu Shrine, the Nijmegen artist Joop Russon made a lino-cut illustration for the cover of this dissertation.

NICI Technical Report 21-01

ISBN: 90-9014793-4

Hoor ik ook wat ik zie?

Bij computerprogramma's is het steeds gebruikelijker aan allerlei functies een geluidje te koppelen. Bekende voorbeelden zijn het 'dingdong'-geluid bij een binnenkomend e-mailbericht of het geluid van een papierversnipperaar bij het leegmaken van de prullenbak.
Maar wat is de bedoeling of het effect van deze geluiden? Zijn ze alleen maar leuk of grappig of hebben de geluiden een positieve of negatieve invloed op het uitvoeren van een bepaalde taak? Maakt het wat uit of ik hoor wat ik zie?

Met behulp van proefpersonen zijn twee soorten geluiden onderzocht: concrete geluiden, zoals het blaffen van een hond en abstracte muziekakkoorden in majeur en mineur. Muziek in majeur wordt vaak met iets positiefs geassocieerd en in mineur met iets negatiefs. In de experimenten werd onderzocht wat het effect is van deze geluiden op de uitvoering van een visuele taak.
Uit dit wetenschappelijk onderzoek komt naar voren dat concrete, levensechte geluiden de taak versnellen, terwijl abstracte geluiden de uitvoering vertragen. Als het geluid niet overeenkomt met de visuele taak (bijvoorbeeld geblaf bij de afbeelding van een kat), wordt de uitvoering van de taak extra vertraagd.
Het lijkt niet gewenst zomaar geluid toe te voegen aan een computeromgeving.
Anders gezegd: door gelijktijdige toevoeging van de juiste geluiden, zodat ik hoor wat ik zie, kan het uitvoeren van computertaken versneld worden.

Stellingen

1. Abstracte geluiden werken vertragend op een visuele categorisatietaak, terwijl concrete geluiden versnellend werken (dit proefschrift)

2. De integratie van beeld en geluid is optimaal als beide tegelijk worden aangeboden (dit proefschrift)

3. Als in een visuele categorisatietaak een abstract geluid wordt aangeboden met taakinformatie die tegengesteld is aan de response-informatie, dan leidt dit tot een negatief Simon-effect (dit proefschrift)

4. There is no such thing as silence (Bill Buxton)

5. Geluid wordt vaak aan computeromgevingen toegevoegd zonder onderzoek naar de consequenties ervan

6. Luchtziekte heeft meer met de kwaliteit van de lucht in een vliegtuig te maken dan met het vliegen zelf

7. Een eigen Europese munteenheid is een illusie zolang niet-Europeanen over een Euro-**dollar** blijven spreken

Nijmegen, 19 juni 2001 *Myra van Esch-Bussemakers*

Do I Hear What I See?
The influence of auditory icons and earcons on multimodal integration in visual categorization

Een wetenschappelijke proeve
op het gebied van de Sociale Wetenschappen

Proefschrift

ter verkrijging van de graad van doctor
aan de Katholieke Universiteit Nijmegen,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op dinsdag 19 juni 2001,
des namiddags om 1.30 uur precies

door
Myra Petrina van Esch-Bussemakers
geboren op 27 februari 1973 te Nijmegen

NICI
Nijmeegs Instituut voor Cognitie en Informatie

Promotor:	Prof. dr. G.P. van Galen
Co-promotor:	Dr. A. de Haan
Manuscriptcommissie:	Prof. dr. C.M.M. de Weert Prof. dr. A. Kohlrausch (IPO, TU Eindhoven) Dr. D.J. Hermes (IPO, TU Eindhoven)

Contents

Acknowledgements

In December 1995, when I received a last minute request from Ab de Haan to have a meeting with Mark Hoffman from NCR Human Factors Engineering, I could never have imagined that this interesting PhD-project would be the result. Ab, first of all I would like to thank you for your enthusiastic guidance, your pragmatic solutions and most of all your ability to integrate experimental results at a higher abstraction level into a theory.

Mainly in the last year of the project I have received a lot of support from Gerard van Galen, whose diverse theoretical knowledge and enormous efforts I greatly admire.

For all the work-related discussions, but most of all for the personal and encouraging talks I would like to thank Josine van de Ven. It is wonderful to have you around, first at the NICI and after that also at home.

Within the NICI I would like to thank all my colleagues, especially: Heike Martensen, for her statistical knowledge; Walter van Heuven, for his experimental knowledge; Herbert Schriefers, for a good start; Beppie van Dijk, for taking care of all the extra administrative 'stuff' and her ability to always cheer me up; Ellen de Bruijn for the enjoyable lunches; Rasmus de Gruil, my NCR-partner; and Paul Lemmens. Paul, it is so nice to have shared an office with someone, who is just as crazy about multimodal interaction as I am.

I am also grateful to my colleagues at NCR, especially Julie Huffman, Jackie Huffman, Chris Sutarno, Mike Inderrieden, Sally Cohen and Mimi Ryan. The consultancy projects that we did together probably kept me 'sane' while working on my PhD. Katie Lehman and Jennie Psihogios, you not only introduced me to the world of biomechanics (the dark side), but you also guided me through the world of retail. Thank you for helping me to grow as a person.
Mark Hoffman and Loren Ulrich, thank you for your faith in me, and for making me feel part of the team.

My family and friends I would like to thank for their support and for always asking how my dissertation was coming along. I especially would like to thank my parents and sister, for always being there when I need them.

Finally I am grateful to Erik, for always having faith in my abilities and for his unconditional love: 'It's so much friendlier with two'.

Chapter 1

Sound, Vision and their Multimodal Integration in Categorization

Introduction

When we perceive the world around us, we automatically try to integrate the possible different information streams (e.g. Bussemakers & de Haan, 1998), whether those are the sounds we hear, the pictures we see or the objects we feel. In our interaction with machines, from household appliances to complex computer systems, modalities are used implicitly. Every modality influences our behavior, even if we are not consciously aware of it. The mind is an open system, meaning that it cannot be seen as consisting of isolated parts (e.g. Mesulam, 1998). Vision and audition for example are integrated when information is processed.

Essentially, the mind can be thought of as consisting of interacting areas with (individual, but related) functions: visual, auditory, lexical, semantical, pictorial, spatial etc. In this dissertation we try to shed a new light on the connections or interactions *between* some of these functions related to the activation *within* functions. In what types of situations, i.e. with what types of stimuli, are these connections used; when does activation occur within a function and, more importantly, how does all this influence our responses?

Previously, modalities were mostly studied in isolation. Researchers investigated the visual perceptual system for example, to understand its elements. Although this is valuable knowledge, it does not seem enough. In any task setting it is important to look beyond the individual senses and use a more integrated approach. How, for instance, does sound influence how we see things? What would happen if we would come across a cat in a street somewhere that, instead of meowing, would bark like a dog, or even more interesting, would sound like a piano playing every time it opens its mouth? How does our mind integrate the perhaps semantically unrelated information? Temporally the sound and the visual information co-occur. So, do we conclude that the cat made the sound?[1]

A type of controlled environment where multiple information streams can be perceived is human-computer interaction. Computers have become vital for many operations in today's world; therefore a lot of time and effort should be spent on optimizing this complex environment. By investigating how our mind processes and integrates information, possible cognitive bottlenecks in interfaces within a certain environment can be spotted, repaired or even removed.

[1] In real life this example may seem unrealistic, in cinema, however, it is a technique that is often used to create an effect (see for example the movie 'The Exorcist'), suggesting that the information is integrated into, in this case, something scary.

There is a variety of modalities that can be applied when attempting to optimize the interaction between a user and his/her computer. Information can be presented through vision, touch, sound, in some cases even scent, or any combination of these sensory modalities. An interface designer has the difficult responsibility of choosing the more optimal combination of information streams for a particular task. Apart from finding the best solution from a performance standpoint, the designer also may have to take aesthetics into account. In video games for example users report that they can perform better with the audio on, although this impression has never been empirically validated (Edworthy, 1998). The effect may be caused by mere aesthetics, but it is also possible that there simply is more information available.

In this dissertation an experimental setup is used that is known to provide a way of investigating facilitation and interference between the visual and auditory modalities. Within the setup both the Stroop paradigm and the Simon effect can be manipulated. The Stroop paradigm (e.g. Stroop, 1935; MacLeod, 1991) looks at the integration of semantic information within a modality at the level where feature information is coded and integrated. Traditionally, in one of the experiments, subjects saw words that describe a certain color (red, blue, green, brown and purple) and these color names were sometimes presented in different colors. Results indicate that it takes longer to name the colors if the color is incongruent with the written word. In the categorization investigated in this dissertation, subjects have to categorize pictures, while hearing concrete sounds, i.e. auditory icons, like the picture of a cat and the sound of a barking dog. The Simon effect (e.g. Simon, 1990), another extensively studied paradigm, is caused by more or less congruency between the 'semantic area' of the mind and the spatial features of the response that needs to be made. In the traditional experiment, subjects would perform a categorization task, where the stimulus would be presented on either the left side or the right side of the screen, related to the location of the button that needed to be pressed as a response. Results showed that subjects are able to respond faster if the location on the screen is congruent with the location of the button (for instance both on the left side). In our project, subjects, again categorizing pictures, hear abstract, musical sounds, i.e. earcons, which are related to the response. For instance a subject sees the picture of a cat and hears a major chord played on the piano that with its positive connotation could suggest pressing the 'yes' button.

Generally, the flow of information in the experimental task used in this dissertation can be reconstructed as follows. A subject perceives visual information on a feature level. This information is passed on to the pictorial area of the mind and together with perceptual, categorical or conceptual information, the picture of an animal is recognized. In the case of the Stroop paradigm, a sound is played with the appearance of the picture. The auditory features of that sound are passed on to the auditory area of the mind and the sound is recognized. Having the categorical information in two modalities in general leads to a facilitation, i.e. shorter response times. However, if the information is incongruent (suggesting two different responses) a conflict seems to arise, which seems to lead to longer reaction times than when the multimodal information is congruent.

This seems especially true if the to be negated features (for instance like in the traditional task: reading color names) are more quickly processed than the instructed stimulus dimension (e.g. naming colors).

In the case of the Simon paradigm, the auditory information is again presented together with the visual information. The sound is perceived and its features are processed mostly in terms of concepts. If the abstract information is irrelevant to the task at hand, but is related to the response that needs to be given to the visual stimulus, there is a conflict, but a different level than in the Stroop paradigm. It is the communication between areas of the mind, instead of the communication within an area that is influenced.

Before going into more detail on the experimental setup, the implications and the results from the experiments, the different senses, modalities and how they influence human-computer interaction are briefly discussed.

Sound and Hearing

When an object moves or vibrates, it can cause a pattern of changes in pressure in the surrounding medium (often air), which results in sound. An example of a type of sound is the sine wave. It is a pure sound, like the tone of a tuning fork. A sinusoid has three attributes that can be used to describe it. First of all there is the frequency or the number of times the waveform repeats itself per second (in Hertz, 1 Hz = 1 cycle per second). Secondly, there is the amplitude or the pressure deviation from the mean. Thirdly, there is the phase that defines which part within a cycle is reached at a specific time instant. A complete cycle of the sound is called a period (e.g. Rossing, 1989).

Related to these attributes, sound has four basic perceptual attributes: pitch, loudness, duration and timbre (e.g. ASA, 1960). Pitch can be defined as the attribute of auditory sensation in terms of which sounds may be ordered on a musical scale (ASA, 1960). It is related to the frequency of a sound. If the frequency of a tone is higher, the pitch of that tone will be higher as well. Loudness is the perceived intensity of the sound. The amplitude of the sound wave defines this. Higher amplitude – keeping other parameters like intensity constant - leads to the perception of a louder tone. Timbre is the perceived 'quality' of the sound, which allows us to distinguish between a piano and a violin playing the same note. It is related to the spectral energy of a tone (Hereford & Winn, 1994).

When a sound travels through the air towards us it first reaches the outer ear, or pinna. From there it travels through our ear canal (meatus) to our eardrum, the tympanic membrane. Past the eardrum the middle ear starts. Three little bones, the first of which is attached to the eardrum, pass vibrations onto the cochlea, which is filled with liquid. In the cochlea the vibrations are converted to nerve pulses that travel to the brain via the auditory nerve. The conversion of the vibrations takes place through hair cells. Different hair cells are excited by different frequencies. This frequency-place transformation contributes to our ability to distinguish between sounds with different frequencies (e.g. Cook, 1999).

In contrast to the visual situation where a specific location on the retina corresponds to a specific spatial orientation, the auditory system has to deduce the direction of a sound source from a comparison of the signals arriving at the left and right ear. We are able to hear where a direct sound is coming from in the horizontal plane because of two other major sources of information. First of all if a sound reaches us from the left, it will be louder to our left ear than to our right ear. This is called the Interaural Intensity Difference (IID) (e.g. Brewster, 1994) or Binaural Intensity Difference (e.g. Murch, 1973). For high frequencies this effect is greater than for low frequencies. At low frequencies the sound waves bend around the head. At high frequencies however the sound waves do not bend and an auditory shadow occurs that causes a greater IID.

Similar to an intensity difference between both ears, there is also a time difference when a sound travels to the ears from a location to the left or right of the observer. This is known as the Interaural Time Difference (ITD) or Binaural Time Difference. Since one of the ears is closer to the sound source and the other is further away, it takes longer for the sound to reach the ear farthest from the source. When the sound is opposite one ear, for example, the delay between the two ears is greater than 0.6 ms (Brewster, 1994). At high frequencies the wavelength becomes shorter than the distance between the ears and the ITD becomes unreliable for sinusoids. Therefore the effect can be observed best at low frequencies.

Other aspects that can influence the localization of sound, especially in the vertical plane, are the shape of the pinna and the duration of the sound (e.g. Hofman et al, 1998). When a sound continues for a longer period in time, it is possible to orient the head to maximize the sensitivity to changes in IID and ITD (e.g. Murch, 1973). Finally, the more frequencies are present in a sound, the easier the sound can be localized.

Sound Representation

The end result of the nerve pulses that travel to the brain is a representation of what is going on in the external world. After we have perceived a sound, that auditory stream is broken into components (pitch, loudness, and timbre), which will enable us to make sense of it. How we are able to perceive these components is explained by Gestalt principles (e.g. Moore, 1989; Bregman, 1990; Williams, 1992; Cook, 1999) that describe the factors that are of influence on perceptual organization. Although these rules were originally defined for the visual modality, they also apply to the auditory field. The Gestalt principles are: Similarity, proximity, good continuation, habit or familiarity, belongingness, common fate, closure and stability. It is important to note that these principles are not only based on the acoustical features of the sounds. Often the observed phenomena are also the result of context, attention or prior knowledge.

The first principle, *similarity*, states that components are perceived as related if they share the same attributes. Examples of auditory grouping concepts, which demonstrate the principle of similarity are: common onset, common offset, common frequency modulation and common amplitude modulation.

If for instance a group of tones is played which are all of the same frequency and duration, but of different intensities, a distinct galloping rhythm is heard, because the intense tones are grouped separately from the softer tones to form a beat.

The second principle is called *proximity*. Components that are close to each other are more likely to be grouped together. Examples are: temporal proximity and frequency proximity. Sounds that are close together in frequency or presented together in time are perceived as a group.

The third principle is *good continuation*. Components that display smooth transitions from one state to another are perceived as related. If there is a regular or predictable transition, our perceptual system assumes that the sound is coming from the same source. However if there is an abrupt transition, this indicates that the sound is coming from a different source. For example, if two tones are separated by a noise burst, they are more likely to be heard as continuous, if the start frequency of the second tone matches the end frequency of the first tone.

Habit or *familiarity*, the fourth Gestalt principle, refers to prior expectations about sound groupings, which have been acquired through previous experience. If a certain meaning has been attached to a relationship between sounds, it is likely that the same meaning will be attributed if the same sounds are heard again. However it depends on the level of familiarity or how well the new sounds match the previously heard sounds. This principle for instance helps us to recognize a familiar voice or a piece of music.

Disjoint allocation, or '*belongingness*', means that a component can only be part of one disjunctive object at a time and its percept is relative to the rest of the figure to which it belongs. With some types of stimuli it is possible to have multiple ways of interpreting the input. The perceptual organization can be ambiguous. However when a component of the input is used to form a particular interpretation, it cannot be used to construct another interpretation.

Common fate, the sixth Gestalt principle, is related to the fact that components that undergo the same kind of changes at the same time are perceived as a group. Different auditory signals, coming from the same source, coherently change in for instance frequency and intensity. An example of that can be observed when two sounds that have the same temporal onset, i.e. they begin at the same time, are considered to be coming from the same source (e.g. Williams; 1989).

In some situations, a sound is obscured by other sounds and it is hard to verify whether the sound continued or not. Our perceptual system perceives such a sound as continuous, which is known as the principle of *closure*. This principle is analogous to the phenomenon of visual occlusion, where fragments may be perceived as components of a single object that have been partially obscured. An auditory example of closure can be heard under certain circumstances when a tone is alternated with a noise burst and it is perceived as continuous (e.g. Bregman, 1989).

Finally there is the principle of *stability*. Having achieved an interpretation of an acoustic signal, that interpretation will remain the same through slowly changing parameters, until it is no longer appropriate. The sound is interpreted in the context of what preceded it, not just its current form. This principle is often used in music to direct attention to certain aspects of the musical piece.

It will be interesting to see whether these Gestalt principles that seem to be true for both visual and auditory perception, also can be found in more complex environments, involving multimodal perception.

Audio in Human-Computer Interaction

We are surrounded by sounds all the time. Some of the sounds help us in accomplishing tasks and therefore are considered to be information. Other sounds hinder us and are considered to be noise. There is no such thing as silence (Buxton, 1989). It is the goal of good sound design to increase the amount of informational sounds and to limit the noise.

The effect of using sound in an interface situation has been under investigation for a little over a decade. Especially for situations where the eyes are otherwise occupied, for instance when you are away from your computer, the benefit of having additional auditory information is clearly demonstrated (e.g. Gaver, 1989; Brewster, 1994; Rauterberg, 1998).

One of the properties of sound that make it especially suitable, besides vision, for the use in interfaces, is that sound is in time over space, whereas visual information is in space over time. Thus sound and vision are complementary (Gaver, 1989). This means that sound can be heard anywhere within reach; the user of a computer does not actually have to 'look' at the source of the sound, a computer for example, to perceive it. However it is in time, meaning that if a user does not perceive the sound when it is presented, for instance because he is not present within the range of the sound, the information is lost. Visual information, on the other hand, can only be perceived if the observer is oriented to it, like for instance with a computer screen, but has the advantage that it can remain visible for a longer period of time.

In addition to or instead of visual information, sound can be used to give the user information on the state of the system, it can alarm the user when something is wrong or it can give output from programs (Hereford & Winn, 1994). Continuous tones with repeating patterns are used to provide information on the status of the system. Because of their monotonous repetition, status sounds will perceptually fade into the background. It is only when the status of the system changes, and as a result of this the sound changes, that the user will notice the sound again. Alarms and warning signals are presented when something happens that requires the immediate attention of the user. They are designed to interrupt the ongoing task.

Lastly, it is possible to use sound to present the output of programs by creating melodies or soundscapes (e.g. Buxton, 1989). Music as an additional source of information is often used, when it is not possible to visually represent all the dimensions of the data.

Although sound offers a powerful addition to the interface, it is generally not recommended for sighted users to use sound as the only means of conveying information (Microsoft, *Windows guidelines*), because of the different types of (noisy) environments users work in. Furthermore sound is transient in nature, so if the receiver does not pick up the message (completely), it sometimes cannot easily be retrieved. Instead sound is best used to supplement visual information that is incomplete or unavailable.

Although there is no doubt as to the advantages users have in such situations, it is questionable whether this is solely due to the sound itself. The fact that there is complementary information available, which is not present in the visual information stream, regardless of the auditory nature of that information, could account for some of the effects reported (Edworthy, 1998). It seems interesting to see what results are found when the auditory information is completely redundant with the visual information, i.e. also available in the other modality. When information is presented both visually and auditorily, i.e. through multiple sources, users seem to integrate these informational elements into a single experienced stream (Bussemakers & de Haan, 1998).

For conveying auditory information from the system to the user, two types of sounds can be used: speech or non-speech sounds. Within the non-speech sounds auditory icons and earcons can be distinguished, both of which are studied here.

Auditory Icons and Earcons

In multimedia applications vision can be combined with different types of sound, like real-life sounds (auditory icons, e.g. Gaver, 1989) or abstract, 'musical' sounds (earcons, e.g. Blattner, Sumikawa & Greenberg, 1989).

Auditory icons are based on the concept of everyday listening (e.g. Gaver, 1989; Mynatt, 1994). People tend to describe sounds in terms of the objects and events that cause the sound. For instance when crossing the road we hear a car approaching instead of a repetitive sound with a large bandwidth, increasing in intensity. An auditory icon is linked to an object or concept in the real world. Whether or not these real-life sounds also can be applied in human-computer interfaces, i.e. whether or not they are 'usable' for users depends on four factors (Mynatt, 1994). First of all there is the *identifiability* of a sound. An auditory icon can be defined in terms of how often a user is likely to have been in contact with the sound (its 'ecological frequency') versus its uniqueness. Obviously it is more difficult to link a relatively unknown sound to an action or object in the interface. Secondly there is the *conceptual mapping* of the object or action to those in the real world. It may be effective to use the sound of breaking glass as the auditory icon to the action of emptying the trashcan on the desktop.

However, most users are aware that the file that is being deleted from the hard-drive of the computer is not made of glass. This can lead to annoyance or even misinterpretation of the sound. The third factor of influence on the usability of auditory icons consists of the *physical parameters* like the duration of the sound, the intensity, the quality, the bandwidth, etc. Finally, the *user preference* is important. Some sounds may have certain connotations that need to be taken into account when designing auditory icons.

In short, the overall advantage of using auditory icons is that users know what the sound is referring to in the real world, if the function is represented well. Because it is a 'real world' sound, users do not have to learn the icon. On the other hand, there are some disadvantages that complicate the use of this type of information, one of them being that users are annoyed by auditory icons after prolonged use (see also Sikora, Roberts & Murray, 1995; Roberts & Sikora, 1997).

Earcons, the second type of possible non-speech sounds in feedback, are abstract sounds (i.e. not event or object related), commonly referred to as musical. They presumably can be used to steer the emotional reaction of the user in support of a certain response (Blattner, Sumikawa & Greenberg, 1989). For instance, when creating a new file on the computer a sound is played that increases in loudness, indicating that something is appearing. An earcon can consist of a single note or a structured combination of notes, called a 'motive' (e.g. Dix et al, 1998). Like an auditory icon, an earcon can represent both actions and objects. Differences in representation can be indicated by differences in pitch, timbre, volume and tempo. Earcons can be used in isolation or as part of a compound earcon or a family of earcons. A compound earcon creates 'sentences' to convey information by combining motives. For example if there is an earcon for the function 'delete' and there is an earcon for the object 'file', playing the two in succession, could represent the compound earcon: 'delete file' (e.g. Blattner, Sumikawa & Greenberg, 1989). When designing the earcons for an entire interface, it is important to create a family of sounds. Similar types of information, error-messages for example, should sound related as well. A way of realizing such a relationship is by using the same motive, but with, for instance, a different timbre.

Because earcons are abstract, they lack the advantage for users of not having to be learned as to how the sound is related to the function the earcon is representing (e.g. Gaver, 1986; Hereford & Winn, 1994). On the other hand, users report them to be less annoying than auditory icons (Roberts & Sikora, 1997). Besides, it is possible to create an appropriate mapping between an earcon and the function in the interface that users judge to be better than the mapping with auditory icons (Roberts & Sikora, 1997).

Major and Minor key sounds

Some sounds seem to have particular connotations that could be useful in creating earcons. Possibly these sounds are easier for users to learn than sounds without this predefined connotation. One emotive connotation that has been the focus of research for centuries is the distinction between major and minor key tones.

Historically there have been many studies that showed the existence or non-existence of a link between major tones and 'happy' and minor key tones and 'sad', and recently the work by Crowder (1984, 1985a, 1985b, 1991) has validated this relation.

Crowder, looking at the association between major/happy and minor/sad does not attempt to answer questions about the genetic or physical basis for the effect. It is only the result of the connotation and its possible use that is under investigation.

In his historical overview, Crowder (1984) describes and reanalyzes an experiment conducted by Heinlein (1928), where 48 isolated, major and minor chords were played to 30 subjects. Half of the subjects were musically skilled and half were unskilled. The chords were played one by one to the subjects and they were asked to select a single adjective from a list of 15 adjectives (bright, cheerful, doleful, dark, etc.) indicating the connotation. Although Heinlein focused on the errors and inconsistencies in the data, Crowder reanalyzed the percentage of positive adjectives that were responded to major chords and the percentage of negative adjectives that were responded to minor chords. These results indicated that "There cannot be the remotest doubt that for both trained and untrained listeners, isolated chords, played in random order and out of all musical context, produce connotative judgments in accord with the conventional happy/sad dimension, whatever Heinlein and careless readers of his work have claimed to the contrary" (Crowder, 1984).

An important question that results from these studies is why major is associated with happy and minor with sad. Currently, there are three ideas on the subject. Firstly, it is possible that, because major chords are more common in nature, they have a positive connotation (see also Zajonc, 1980). Secondly, there could be a preference of consonant tones over dissonant tones. Looking at all the possible combinations of tones in a chord, there are more tones that sound like 'beats' or 'roughness' (dissonance) in minor chords, explaining the relation with sad. Thirdly, it could be a cultural convention to relate major to happy and minor to sad. Furthermore it is possible that a combination of the ideas mentioned lead to the connotation. Perhaps at first there was a preference, based on the higher partials, which evolved to the happy/sad dimension through musical socialization (Crowder, 1984).

In this dissertation major and minor earcons are used to study the effect on visual categorization. It is expected that the above-described connotation will influence users when responding to pictures.

Visual Perception and Feature Integration

When users respond to pictures presented to them, they first have to perceive the visual information. When a ray of light reaches our eye, it first comes into contact with our cornea. Then, shining further into the eye light passes the iris and reaches the lens, where it is converted to have its focal point exactly on the retina, in many cases even on a specialized region of the retina called the fovea, because it provides the best detail-oriented vision (highest spatial resolution) (Schwartz, 1999).

The visual information that reaches the retina is passed on to the optic nerve that relays the messages to the visual cortex in the brain. There it is combined with additional information from memory and the other senses. The mind creates an image of the world from the light rays (Gibson, 1966).

But how is a picture recognized?
When sensations of an image arrive at the visual cortex the separate elements of that picture are perceived (e.g. Marr, 1982). There are two schools of thought on how visual perception is achieved. Constructivist theorists, like Marr (1982) believe that information from the real world is actively constructed by combining it with what is already stored in our memory (Preece, 1994). Ecological theorists, like Gibson (1966) believe that vision is actively explored, not constructed, by picking up special features, either combined or in isolation, that are representative of a certain shape. These 'invariants' (a kind of 'schemas') in a pictorial array are for instance the information *about* the dog, cat, man, house or car. This exploration assumes that vision is a precise reflection of reality, which does not seem to be true.

A more cognitive approach to visual perception is advocated by Neisser (1994), who distinguishes different levels of processing for stimuli, based on the neuroanatomic distinction between 'where' and 'what'. The first type of processing is direct perception (where). It enables us to perceive and act fast on an environment, almost like a reflex. Another type of processing is representation or recognition, which enables us to identify and respond to familiar objects and situations (what) (see also Van Galen, 1974). It depends on strategies and uses conceptual knowledge. It is the second level that enables us to deal with ambiguous or abstract stimuli.

Vision in Human-Computer Interaction
In the interaction with human users, vision is still the most important output channel. Information is presented on the basis of the characteristics of objects in our real world (Preece, 1994), so that we can understand the meaning.

In order to present these characteristics different attributes of the visual system are used. An example of that is monocular depth vision, where aspects like the size of objects, the relative position to each other (interposition), the contrast, clarity, and brightness of an image, shadows that objects cast on its surroundings, and different types of textures, are used to create the 'illusion' of a 3-dimensional environment.

How much of an object or image is perceived not only depends on our visual system, but also on our attention.

Attention and Privileged Loops
When we talk about attention, it is still not entirely clear what that means. Attention is not a unitary concept (Styles, 1997). It is a term that is used to describe a variety of psychological phenomena. In today's research aspects are taken into account that involve biological, neuropsychological, computational and functional considerations.

Attention is important in relation to multimodal experiments, because in these studies attention needs to be shifted, or shared between two or more input modalities. This shifting or sharing could influence the behavior or responses of the subjects.

What happens to attention when two or more tasks need to be completed at the same time? Is your attention divided between the tasks, or are you able to perform two separate actions at the same time? According to the old filter theory (e.g. Broadbent, 1958), completing multiple tasks at the same time is achieved by switching back and forth between actions. It is questionable whether this is the case, because studies investigating continual (like for example shadowing, where a continuing text is immediately repeated out loud by a subject) dual-task situations show that there is no loss of performance, which could be expected if this rapid shifting would occur (e.g. Styles, 1997).

Some researchers believe that there is a pool of attention, from which each task takes the necessary amount. This theory is called the capacity-theory (see also Knowles, 1963; Kahneman, 1973). The size of the attention-pool can vary, depending on someone's motivation, stress, information processing capacities, etc. However, the key idea is that when a task needs to be completed, attentional resources from the pool are allocated, which leaves a diminished capacity for other tasks, i.e. there is less attention available.

Posner and Boies (1971), and more specifically McLeod and Posner (1984) show that this pool of attention does not always seem to be shared by tasks. A subject is able to complete certain tasks, like shadowing for example, at the same time with other tasks, like playing the piano, seemingly without interference. McLeod and Posner assume that this effect occurs, because there is a privileged loop, an automatic and shielded link between the auditory input and a vocal response, which is always active. It is possible that there are more types of privileged loops, resulting in more channels relating their input to actions directly. In these cases interference only occurs when the shared resource is allocated within the same modality or channel, or the same task needs to be accomplished at the same time in different modalities.

When subjects are working in a single task and they have focused their (endogenous or intention-related (Posner, 1980)) attention to a certain location on the computer screen for example, other stimuli, which are presented at the same location, are also processed (Eriksen & Eriksen, 1974; Spence & Driver, 1996). This means that within this 'spotlight' of attention even distracters are processed. When having to respond to the stimuli and distracters, a conflict arises, if both stimuli lead to a contradicting reaction. The resulting interference can be seen in longer reaction times or more errors in the data. It is likely that a similar effect can be expected in multimodal experiments, where an auditory stimulus is presented at the same temporal location or spotlight as a visual stimulus.

Intersensory Integration

Often when environments are discussed that offer information from multiple information streams, like visual and auditory, these are called multimedia environments or multimodal environments. However, there is a difference and it is important to define the two terms for the rest of this dissertation. A medium is a means through which information is presented, like for instance a computer, a radio, a newspaper etc. Consequently in a human-computer interaction environment, the media are: the computer screen for visual information, the loudspeakers for auditory information etc.

Multimodal on the other hand, involves the 'modes' of interaction; information that is available to the senses that are used to perceive the information. In multimodal human-computer interaction there is auditory information, visual information, tactile information etc. In this thesis only the auditory and visual modalities will be discussed, so whenever the term 'multimodal' is used, that is what it refers to.

One of the characteristics of multimodal (auditory and visual) representations within the interface is that task relevant information in both modalities needs to be perceived and integrated. From studies in perception we know that observation in one modality, or sensory system, can influence another and even that a particular modality can substitute for another modality (Stein & Meredith, 1993). An example of that is *the ventriloquism effect* (Howard & Templeton, 1966). Although the ventriloquist is speaking it seems the sound is coming from the puppet, because the mouth is being moved. In this case vision influences the auditory perception in the sense that it seems that the puppet is speaking. Although this effect can occur in different modalities, the idea seems to be that it is most dominant in vision. Stein and Meredith (1993) state that when there are no great differences in the intensities of the stimuli, the effect of the visual stimuli on the stimuli in other modalities is greater than their influence on visual perception. This would indicate that the contribution to the perceptual integration in the parallel information processing that seems to occur in multimodal perception is not equal.

However, visual dominance seems to depend on the task that needs to be performed. Welch and Warren (1980) defined *the modality appropriateness hypothesis*. It seems that when a modality is better suitable for a certain task, it dominates over other modalities. In choice reaction time tasks, for example, where subjects have to press a response key if a tone is presented and another key when a light is presented, a clear dominance to respond to the light was shown, when stimuli from both modalities were presented at the same time (Colavita & Weisberg, 1979). Subjects often only noticed the light and did not even perceive the tone. In spatial perception tasks, vision is also a more precise and accurate modality than audition. On the other hand, in tasks that involve temporal acuity, for instance adjustments to flickering light and fluttering sound, audition seems to influence vision more than vision seems to affect audition (e.g. Welch, DuttonHurt & Warren, 1986). When subjects adjust the sound, perceptually the flickering of the light seems to change also, although in frequency it stays the same. This is referred to as *the driving effect* (Gebhard & Mowbray, 1959).

In the case of the visual and auditory modality especially, the two sensory systems have learned to work together in the perception of information (Stein & Meredith, 1993). For instance in a noisy room where many people are talking at the same time and someone is trying to get a message across to you, it helps when you can also see the other person's lip movements along with the auditory signal. Generally, when an auditory signal and a visual signal are presented together, carrying the same information, there is a faster reaction time, than when either input is presented alone. This is even more true, when the visual stimulus is presented 40-60 ms before the auditory input. Visual information takes longer to respond to than auditory information, because of the longer processing time in the retina compared to the inner ear (Stein & Meredith, 1993).

There are several different reasons mentioned in the literature on why a redundant multimodal stimulus is faster than a unimodal stimulus. Some researchers believe, that since stimuli with a higher intensity lead to faster reaction times, the intensity of both stimuli are summed, leading to an overall higher intensity (e.g. Colavita & Weisberg, 1979). Others believe that there is an alerting effect of one stimulus on the other, leading to a faster response (Nickerson, 1973). Some even believe, that it is a combination of both mechanisms that can explain the observed results (Welch & Warren, 1980). Studies looking at evoked potential recordings seem to suggest that the same information in two modalities increases the magnitude of the evoked potential, increasing the salience of the stimulus. When a stimulus is more salient, this means that it is less ambiguous to the subject. The subject can therefore respond more quickly (e.g. Stein & Meredith, 1993).

Most of these studies are based on experiments with non-semantic stimuli, like a flash of light or a burst of noise. The same effect may also be found however in situations where the stimuli are of a more semantic nature and it seems relevant in that respect to mention experiments on categorization and more specifically, the Stroop paradigm and the Simon paradigm, but first it will be discussed what happens when subjects categorize pictures.

Categories and Concepts

Most information in our memory is either categorical or conceptual[2] in nature. Categorical knowledge, which is perceptual in nature, enables us to generalize about what we have learned about an object or an event to other similar objects or events. It enables us to adapt our behavior to our environment. When we try to categorize an object, we are essentially assuming that this object is comparable to other objects we know, contrary to the process of discrimination where we try to determine that an object is different (Shanks, 1997).

Apart from categorical knowledge we are also able to acquire and use conceptual knowledge. Concepts are a representation of a class of objects.

[2] This distinction is analogous to the previously mentioned where/what distinction, or direct perception versus representation (Neisser, 1994).

Where the usefulness of categorical knowledge may be limited, because it depends on the perceptual similarity between the object and previously encountered objects, conceptualization can assist in setting the boundaries of what belongs to the category and what does not by using principles or rules. Concepts are based on deeper and more abstract properties of objects (Goldstone, 1994). For instance to determine whether a number is odd or even, we apply the rule 'can be divided by two'. We need these mental representations, which are a combination of categories and concepts, or we would not be able to infer or make decisions. If we for instance hear a dog barking we use our categorical knowledge more than our conceptual knowledge to determine what we hear. However when we hear a piano playing a chord our conceptual knowledge is used more than our categorical knowledge to recognize it.

But where do categories and concepts originate from?
Rosch (1973) believes that a prototype of a certain category is abstracted from our experiences, by learning from experiencing many instances in the class. For example a prototype 'bird' is derived from our experience with many birds (Shanks, 1997) and consists of a combination of features from all the birds we know. How quickly we can recognize a bird would then depend on how related this new bird is to our prototype.

This idea of a prototype is invalidated by experiments of Whittlesea (1987), who shows that it is not likely that there is a prototype that is constructed from our experience, but that there is instance memorization when learning categories. The category bird consists of remembered instances of birds we have encountered. Each of these instances again is connected to the category (see also Mareschal, French & Quinn, 2000).

More recent research has suggested that both categorical and conceptual information is stored in neural networks (also known as connectionistic networks) in the brain. The information is retained in a distributed fashion by weighted connections between neurons. Categories are represented by mental associations between the elements of the stimuli and the categories, which are incremented or decremented according to an adaptive rule (e.g. Shanks, 1997). Information is passed by an excitatory or inhibitory connection from one neuron to another. This system is extremely flexible, because neurons can influence each other in parallel. Our memory is therefore content-addressable, meaning that even with incomplete or false details still the right representation or category can be retrieved and the object can be categorized (e.g. McClelland & Rumelhart, 1985). Another question that has not yet been fully answered is how concepts are represented within a connectionistic network (see also Robertson & Murre, 1999).

After discussing categories and concepts the two paradigms studied in this dissertation are briefly mentioned.

Stroop Effect

To study attention and interference, Stroop (1935) aimed at discovering what the effect would be of the different aspects of a 'compound' stimulus on the attempt to name the other aspect. Furthermore he wanted to study the effect of practice on interference (MacLeod, 1991). In his classical studies, he tested the effect of ink colors on reading aloud. Subjects were presented with five words that describe a certain color (red, blue, green, brown and purple) and the colors themselves. First the subjects had to read the words presented in all possible ink colors aloud. In the control condition, all color words were presented in black ink. Results showed that the overall response time was longer for the condition where the words were in different ink colors, although this result was not significant. MacLeod (1991) later replicated this result. In a second study, subjects had to name the color of the ink the word was written in aloud. In the control condition, subjects named solid color squares. Results here showed that subjects were 74% slower in naming the colors of the words when the color was incongruent with the written word, than naming the squares.

A possible variation of this task is presented in this thesis. In a visual categorization task with a manual response, in some trials redundant auditory icons are presented at the same time with the visual information. In this task there are four types of conditions. The first condition is in those trials where the sound that is played is congruent with the visual information. For example, subjects see the picture of a cat and hear the sound of a cat meowing. In a second condition, the auditory icon does not represent the same information as the picture, but the category of the two is the same, for instance when the sound of a dog barking is presented with the picture of the cat. In the third condition, the information presented auditorily is incongruent with the visual information, for example when the picture of the cat is presented with the sound of a piano playing. Finally, there are control trials, where the picture is presented in isolation with no sound.

Apart from concrete sounds as a redundant source of information, abstract sounds are also studied in relation to visual categorization. The possible different type of effect the earcons might have on categorization is mentioned in the next section.

Simon Effect

In traditional studies looking at the Simon effect (in a way a special kind of Stroop effect), the consequence is studied of an irrelevant distracter, like the location of a stimulus on the screen, on the response (e.g. Simon, 1990). Results show that irrelevant information cannot fully be ignored, because there is facilitation in those trials where the location on the screen is congruent with the location of the response button. More recently this paradigm has been extended to look at a semantic irrelevant factor to test if a semantic Simon effect also can be found.

DeHouwer (1998) presented subjects with English and Dutch words of animals and occupations. Half of the words were in English and the other half was in Dutch. Half of the subjects were instructed to indicate the language of the word by saying 'dier' (animal) when they saw a Dutch word and 'beroep' (occupation) for every English

word. The other subjects were instructed to say 'dier' to an English word and 'beroep' (occupation) to a Dutch word. This experiment is especially interesting, because of its relationship with the experiments described in this dissertation. The results show that in those trials where the response is congruent with the irrelevant stimulus, for instance when a Dutch word of an animal is shown, and the subject is supposed to say 'dier' (animal), facilitation occurs compared to trials where the response is incongruent with the irrelevant stimulus, and for instance an English word of an animal is shown and the subject is supposed to say 'beroep' (occupation).

With a Simon effect three issues are of importance. First of all, the relevant stimulus feature, i.e. that what the subject needs to respond to. In the experiment by DeHouwer (1998), the relevant stimulus feature is the language of the word (Dutch or English). Secondly the irrelevant stimulus feature plays a critical part, since it is this feature that distracts the subject when responding. Again in the experiment by DeHouwer, this is the category of the word (animal or occupation). Lastly the relevant response feature, meaning the response the subject needs to give to the relevant stimulus. In this case it is 'dier' (animal) to a Dutch word and 'beroep' (occupation) to an English word.

In these types of experiments, the irrelevant stimulus feature is related to the relevant response feature: the category of the word is what needs to be responded. However, the relevant stimulus feature is not related to the relevant response feature. At the same time the relevant stimulus feature is not related to the irrelevant stimulus feature (Kornblum, 1992; Simon, 1990). This experimental approach has advantages over traditional Stroop-like tasks, because in some of the Stroop-experiments the targets not only have a semantical relation with the distracters, but also are related in other non-semantic dimensions (associative or perceptual for example). It is therefore difficult to assess whether the results of these experiments are a consequence of the automatic semantic processing or the non-semantic processing (e.g. LaHeij, 1988; Shelton & Martin, 1992; Williams, 1996).

In this dissertation a similar semantic Simon effect is studied in one of the experimental setups. In the visual categorization tasks with redundant earcon distracters the relevant stimulus feature is the category of the word (animal or non-animal). The irrelevant stimulus feature is the connotation of the earcon, meaning the association between major and positive and minor and negative. The relevant response feature is the button labeled 'yes', that should be pressed when the subject sees a picture of an animal, and the button labeled 'no' for pictures of non-animals. Similar to the previous example, there is no relationship between the relevant stimulus feature (for example the category animal) and the relevant response feature (for example the 'yes' button) and between the relevant stimulus feature and the irrelevant stimulus feature (for example a major chord which has a positive connotation).

Summary
In designing multimedia products, little is known about the influence of sound on visual perception. Commercial products like games for example offer both types of information, because it is what the consumer wants, but what exactly is the effect on responses of integrating sound and images? After studying both senses separately, it no longer seems appropriate to study them in isolation, but to find out within specific task environments how information is combined. In this thesis the influence is observed of concrete and abstract auditory information on a visual categorization task. On a memory level, this distinction can be viewed as the influence of categorical auditory knowledge versus conceptual auditory knowledge. The result can assist on the one hand in the choice whether or not to use sound or on the other hand what type of sound to use in the development of multimedia devices in general, and interfaces in particular.

Overview of the thesis
In chapter 2 several experiments are described, looking at the effect of conceptual auditory information, i.e. major and minor third earcons, on a visual categorization task. Different experimental setups are reported, that are tested to see whether the effect found is due to the complete randomization of the trials or to the time between the presentation of the sound and the presentation of the picture (Stimulus Onset Asynchrony). Finally a paradigm is studied, where the subject is presented with trials grouped in blocks. Within each block the relationship between the type of earcon and the intended response is kept constant.

The results found in Chapter 2 are validated in Chapter 3 by a second experiment with the same paradigm, where again the trials are grouped in blocks. Furthermore, a distinction is made between subjects that are musically experienced, meaning that they have played an instrument for more than 6 years, and subjects that are less experienced, meaning that they have no experience in playing a musical instrument. Results show similar findings as in Chapter 2, with no significant difference between the groups of subjects.

Since the experiments in the previous two chapters are based on a single task experiment, in Chapter 4 a dual task setting is tested. The single task experiments do not lead to many errors in performance. Because errors can be another important source of information besides response times, introducing a secondary task, i.e. a cumulative addition of single-digit numbers, induced more errors. Errors can be of interest, because if a subject responds more quickly, but also makes more errors, it may seem judging from just the reaction time data, this situation is an improvement. It is by also taking into account the number of errors, that a full picture can be drawn of a subject's strategy in the experiment.

Earcons are a representation of conceptual information. More concrete categorical auditory information can be found in auditory icons. In Chapter 5 a similar experiment is run as in Chapters 2 and 3, but with auditory icons as accessory auditory stimuli. From the results it can be seen that there are differences in the way users respond to conceptual or to categorical auditory information in a visual task.

Similar to Chapter 4, a dual task experiment is described in Chapter 6. Auditory icons are again used as the redundant auditory information in the visual categorization task. Results of the error data indicate similar findings as in the dual task setting with the earcons. It seems that the difference between categorical and conceptual auditory information is mainly found in the reaction time data.

In the final chapter conclusions are drawn. Differences and similarities between the findings in the previous chapters are discussed against the previously mentioned theories. Furthermore suggestions for further research are given.

Chapter 2

Earcon Experiments with Single Tasks[3]

Abstract

In the three experiments described in this chapter, it is investigated what the effect is of short musical earcons, i.e. conceptual (abstract) auditory information, in multimodal interaction. The results indicate that the addition of sound in a visual categorization task introduces a delay in response times, which was greatest when the multimodal information was presented simultaneously. Secondly, a further delay was obtained when the assumed connotation of the earcon contradicted the correct response. For instance when a minor earcon was added to a picture of an animal, which was linked by the instruction to a positive, 'yes' response, this created a further significant increase in reaction time.

[3] Parts of the research reported in this chapter have appeared in separate publications as:
Bussemakers, M.P., & de Haan, A. (1998). Using earcons and icons in categorisation tasks to improve multimedia interfaces. *Proceedings of the International Community for Auditory Display* (pp. 152-157). Glasgow (UK): The British Computer Society.
Bussemakers, M.P., & De Haan, A. (in press) Getting in touch with your moods: using sound in interfaces. *Interacting with Computers.*

Introduction

In an environment where information is presented in more than one modality, the auditory and visual messages reach the user simultaneously. It is up to the user to create a perceptual unity, i.e. an unequivocal interpretation, in order to come to a correct response. A good example of this phenomenon in our everyday world is related to speech. When you are at a party and you are talking to someone, it is much easier to understand what the other is saying when you are able to see that person's face. The visual cues coming from the movements of the lips are combined with the auditory information to facilitate recognition (e.g. McGurk & MacDonald, 1976).

In the design of multimedia devices non-speech sounds are used as well as speech to convey a message. In dealing with non-speech information, there are two types of sounds that are representative for the distinction between the 'where' and 'what' system within our mind (Neisser, 1994; Shanks, 1997). Auditory icons (Gaver, 1989) are concrete, perceptual sounds that represent categorical knowledge of the world around us (where). There is little doubt to the meaning of the sound itself. However, in multimodal interaction sometimes this categorical meaning does not fully map with its function. For example the sound of water dripping when a file is being copied (Gaver, 1989). The intended function of the sound is clear, yet most users know that a computer file is not made of water. Perhaps this is why users report auditory icons to be annoying *after* prolonged use (Roberts & Sikora, 1997).

Earcons (Blattner, Sumikawa & Greenberg, 1989) are abstract, often musical sounds that are a representation of conceptual knowledge (what). The meaning of an earcon in most cases needs to be learned, but users report the mapping of the sound onto functions in the interface to be better than the mapping of auditory icons (e.g. Roberts & Sikora, 1997).

In earlier research earcons were mostly intended to assist the user in a task where the visual information by itself may not be sufficient. In those cases sound has a clear positive effect on reaction times, errors and preference (e.g. Blattner, Sumikawa & Greenberg, 1989; Brewster, Wright & Edwards, 1994; Brewster, et al, 1995; Brewster, 1997, 1998). However in situations where there are few errors in the visual task and the sound is considered redundant, the effect of the extra non-speech auditory information is not yet established.

Speech as Accessory Stimulus

An area where the influence of (auditory or visual) accessory stimuli on a visual task has been extensively studied is language. Lupker (1976) for instance investigated the effect of words and non-words on picture naming. Subjects have to name the picture out loud, while they also see a word or a non-word displayed. Results showed that besides words, also non-words that are presented visually with a picture cause a significant delay in the naming response compared to a situation where the picture is presented in isolation.

However this is within a single modality. Ishio (1990) tested a multimodal situation where spoken non-words and words were presented with pictures. The results of that study showed that there is even interference on the naming response compared to just the picture naming, when spoken non-words accompany the pictures. On the other hand he did not find interference of visual non-words compared to the condition where just the pictures were presented. A possible explanation for this effect according to Ishio (1992) is, that the interference in the multimodal naming task occurs at the response level. In the unimodal task the visually presented words do not reach this level, because earlier in the process the decision can be made that it is a non-word.

How would these results in language studies translate to non-speech sounds? A picture (and possibly a concrete non-speech sound) can be understood in a single glance (Kennedy & Ross, 1975). A word on the other hand is a very different type of graphical symbol and semantic information on a word in our memory is harder to access than semantic information on a picture (Smith & Magee, 1980; Glaser & Glaser, 1989).

An example of a type of non-speech sound, i.e. white noise, which was used in a picture-naming task is an experiment by Schriefers and Meyer (1990). They used white noise besides different types of words in their experiments. The results showed that the effect of noise on the response times was similar to the effect of silence on the response times; there was no interference of noise with the naming of the pictures.

A different type of task that is used to study interference effects is categorization. Subjects have to determine for a presented stimulus whether or not it is part of a certain category. This type of experimental paradigm is different from naming, because instead of saying the verbal label of a stimulus, subjects need to indicate a positive or negative response to a question about the categorical or conceptual meaning of the stimulus. In the case of pictures, a subject is able to perform the task without verbalizing what he or she sees.

Ishio (1992) besides the earlier mentioned experiments with a naming task, also investigated the interference of spoken words in a categorization task. The results did not show an effect of spoken words. Similar findings were also reported in the visual domain by Smith and Magee (1980). According to them this lack of interference is caused by a more rapid access to semantic information by pictures compared to words. Subjects are able to come to a response before the semantic information on the word is accessed. This theory seems to be supported by studies where words need to be categorized and pictures are presented as accessory stimuli. In picture-word studies interference can be observed indicating that the semantic information of the picture interferes with the response that needs to be given to the word (Smith & Magee, 1980).

In this dissertation a categorization paradigm will be used to study the effects of auditory accessory stimuli on pictures. Categorization was used instead of naming, because users do not have to use language to articulate a response, but can manually indicate the category. Whether or not there will be interference of the accessory stimulus seems to depend on the ease of access to semantic information. If an earcon is accessed more quickly than a picture, interference can be expected. If however accessing the semantic information of the earcon takes longer than accessing the semantic information of the picture, the earcon may not have an effect on the categorization times.

Earcons as Accessory Stimuli
An example of a group of earcons with a certain connotation is major and minor chords. Crowder (1984, 1985a, 1985b) tested and verified the association subjects seem to have between on the one hand 'major' and 'happy, positive' and on the other hand 'minor' and 'sad, negative'.

In this chapter three studies are reported investigating a visual categorization paradigm with auditory accessory stimuli. The connotation is related to the manual response that needs to be given. The manipulation is similar to a paradigm used by Simon (1990), where a task-irrelevant distracter, for instance the location of a stimulus on a screen, has a facilitation effect on the reaction times when the location of the response-button is the same as the location of the stimulus on the screen.

Predictions
It is expected that since the connotation information is of a higher abstraction level, that seems to require semantic integration, there will be interference resulting in longer reaction times for the trials with sound. Furthermore it is expected that in the cases where the connotation of the sound is incongruent with the response, more interference will occur than when the connotation is congruent with the response.

Experiment 1: Complete Randomization of Trials

Subjects
In this experiment 20 subjects participated. Ten of the subjects were male and the other 10 were female. All participants were students at the Catholic University of Nijmegen. They received course credits or were paid for their participation.

Materials and Design
Sixteen line drawings were used as visual stimuli, 8 were of animals and 8 of non-animals (see table 2.1). In addition there were 6 practice drawings that were not part of the test-set. All pictures were selected from a pool of line drawings that were used for other experiments, so that it was known that these stimuli usually were categorized as animal or non-animal.

Animal	Non-animal
Squirrel	Balloon
Dog	Airplane
Cat	Bread
Frog	Violin
Cow	Bed
Lion	Candle
Bird	Shoe
Butterfly	Boat

Table 2.1 Line drawings used in the experiments.

In the experiment 5 types of sounds were presented with the pictures. They could differ in pitch (high or low) and major or minor key (C4major, C4minor, C5major and C5minor). Furthermore a single tone (F4) was included as another condition in the experiment. The sounds had a duration of 2500 ms. All pictures were also presented without sound (silence), resulting in a total of 6 conditions and 96 trials. The order of the trials was randomized completely.

A Macintosh Quadra 840 AV was used, with a 14 inch screen to present the visual stimuli. The screen was raised to eye-level. The subjects wore Monacor BH-004 headphones for the presentation of the auditory stimuli. The manual responses were registered by using a button-box with a voice-key. The assignment of the buttons was randomized between subjects.

Procedure

Before the experiment started the experimenter read the instruction to the subject. After ensuring that there were no questions regarding the experiment, the subject put headphones on and the experiment could begin. The categorization task that had to be performed was whether or not a displayed picture was of an animal (see for example figure 2.1).

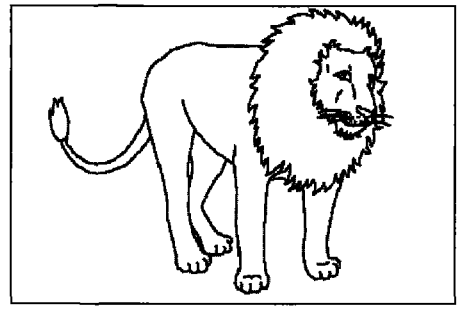


Figure 2.1 Line drawing of a lion.

In every trial, participants first heard the single frequency tone (F4). At the same time a plus ('+') was displayed for 500 ms in the center of the screen to fixate on.

Then the fixation-point disappeared and after 1000 ms a line drawing (picture) was shown for 300 ms, together with a sound or with silence. Subjects were asked to respond as quickly as possible by pressing the button labeled 'yes' or the button labeled 'no' with respectively their left or right index finger. After 2.5 seconds another trial was started. The subjects were able to take a break halfway through the experiment.

Results

The reaction times were measured from the onset of the critical stimulus, i.e. the moment the picture appeared. Only reaction times to correct responses were included in the analysis. In only 2.6% of the trials an error was made, so they were not analyzed further. The errors were evenly distributed across conditions. The mean reaction times per condition and the standard error of the mean are presented in figure 2.2.

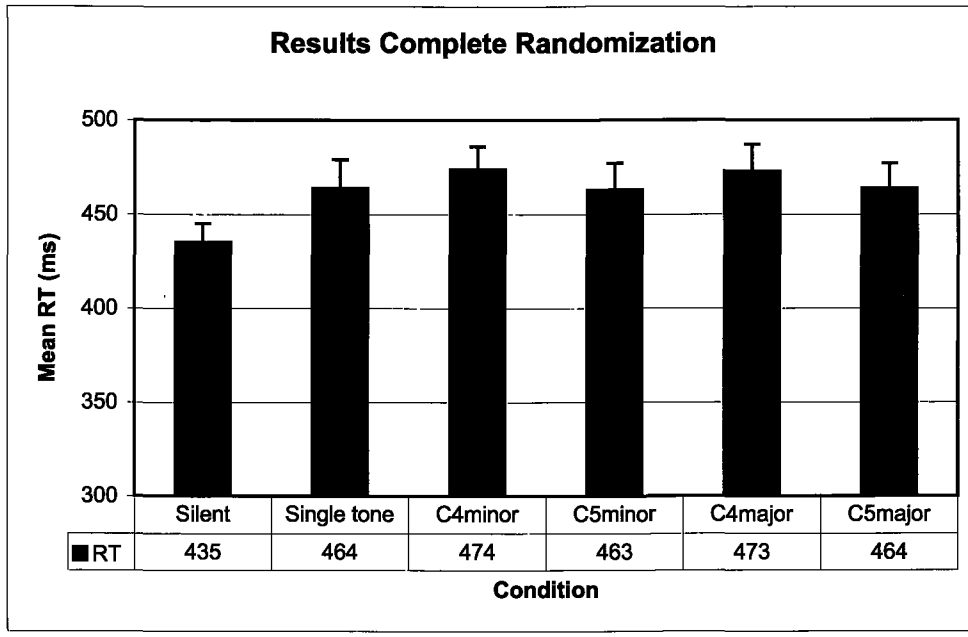


Figure 2.2. Mean reaction times (RT) per condition of experiment 1.

Every subject participated in every condition, therefore a repeated measurement analysis was conducted that showed that a statistical significant difference in mean reaction times between conditions was obtained ($F(5,15) = 10.091$, $p < .001$). More specifically, a contrast analysis showed that the difference between the conditions with sound and the condition without sound was significant ($F(1,19) = 29.429$, $p < .001$).

However, the reaction times on the types of sounds did not differ significantly from each other ($F(4,16) = 0.941$, $p > .1$). The difference in frequency, i.e. higher and lower tones, between C4 and C5 shows a trend where the lower tones result in longer reaction times than the higher tones, but this is also not statistically significant ($F(1,19) = 1.923$, $p > .1$).

Discussion
The results indicate that the mean reaction times in the conditions with sound are slower than the mean reaction times in the conditions without sound. Having conceptual (i.e. 'what') auditory information in a visual categorization task seems to significantly slow down the response times.

However, there is no effect of the connotation of the sound. The mean reaction times of the conditions with sound did not differ from each other. One reason for this could be that the processing of earcons takes longer than the processing of the picture by itself. If this were the case most of the effect would not be due to the connotation of the sound, but would be due to the longer processing of the earcon.

Therefore a second experiment was conducted where the earcon was presented earlier than the visual information to allow for a longer processing time, before the integration with the visual information can take place. To test the timing of the auditory and visual information only a single earcon is used. It is expected that the interference effect of the earcon will become increasingly greater when it is presented earlier than the picture.

There is another possible reason that there is no effect of the connotation of the sound. Because the relation between the sound and the response changes every trial, this switching between relationships could influence the effect of the connotation of the sound. Therefore a third experiment was conducted where a constant relationship was created per condition between the type of response and the type of sound that was played.

Experiment 2: Time Course Effects

Subjects
In the second experiment 20 new subjects participated, who received course credits or were paid for their participation. They were all students at the Catholic University of Nijmegen.

Materials and Design
The 16 line drawings from experiment 1 were again used as visual stimuli. The auditory stimulus was C5minor with a duration of 2500 ms. The sound was presented relative to the picture at different time intervals (Stimulus Onset Asynchrony) of –500 ms, -250 ms, -100 ms, or at the same time with the picture (SOA of 0 ms), resulting in 4 conditions. Within a condition all pictures were presented twice, once with the sound and once with no sound. In the case of no sound the inter-trial-interval was increased with the same SOA, so that the time

between the pictures was constant. The order of the conditions was randomized between subjects to cancel out any order effects. Furthermore the order of the trials within a condition was randomized. The total experiment consisted of 128 trials, 32 trials per condition.

Similar to the previous setup, the Macintosh Quadra 840 AV was used with the screen raised to eye-level. Subjects wore the Monacor BH-004 headphones and their manual responses were registered by the button-box with voice-key.

Procedure
The same procedure was used as in experiment 1. Subjects first heard the instruction and could ask questions. Then the headphones were put on and the experiment was started. Subjects first saw the fixation point and heard the single tone. Then, depending on the condition the sound was played or there was a silence and after the SOA the pictures was presented. Subjects responded manually to the question whether or not it was a picture of an animal via the button-box. After every condition, subjects could take a break, before going on to the next set of trials.

Results
The reaction times were again measured from the moment the picture appeared.
Only reaction times to correct responses were included in the analysis.
In 1.5% of the trials an error was made. Because there were so few errors, they were not analyzed further. The mean reaction times per condition and the standard error of the mean are presented in figure 2.3.

Again a repeated measurements analysis was conducted, that showed a significant difference in mean reaction times between the conditions ($F(4,16) = 8.947$, $p < .001$). The difference between the conditions with sound and without sound was the most apparent at an SOA of 0 ms. In table 2.2 the analysis of the difference between sound and no sound is displayed for every SOA.

SOA (ms)	F(1,19)	*p*	
0	20.418	0.000	*
-100	7.054	0.016	*
-250	5.461.	0.031	*
-500	0.466	0.512	

Table 2.2: Contrast results sound vs. silence per SOA (* significant with alpha of .05).

Furthermore there is a significant difference between the silent condition at an SOA of 0 ms and an SOA of –500 ms ($F(1,19) = 5.219$, $p < .05$).

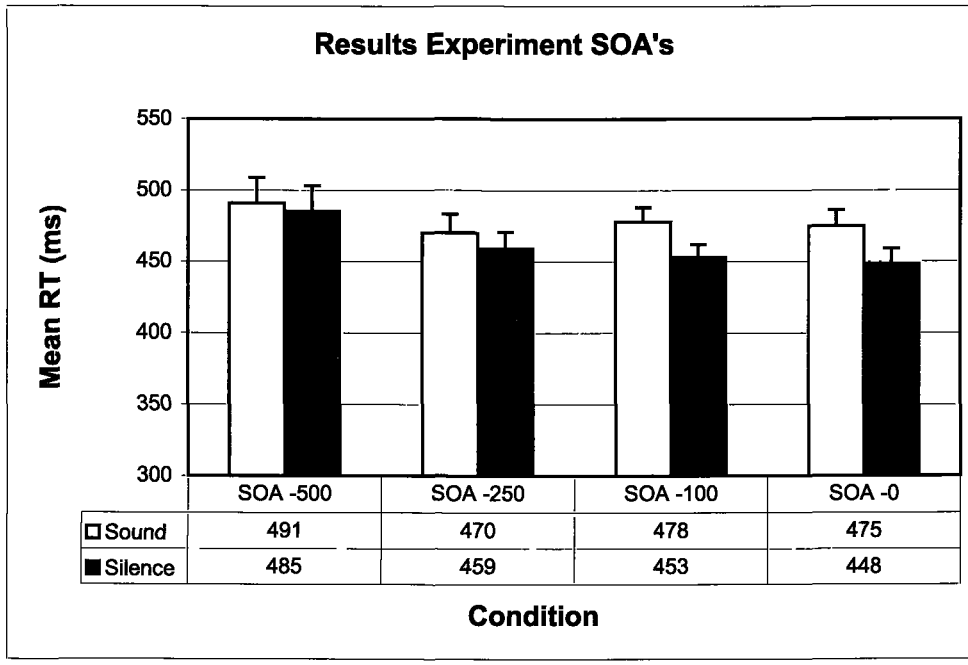


Figure 2.3. Mean reaction times (RT) per condition of experiment 2.

Discussion

From a simultaneous presentation until an SOA of –250 ms the results of the first experiment are validated. The trials where the sound is presented with the picture are slower than the trials where only the picture was presented.

Furthermore, the difference in means between the condition with sound and the condition without sound is the greatest at an SOA of 0 ms. These results are similar to the findings of Schriefers and Meyer (1990) in their experiments on picture-word naming. Apparently the hypothesized greater difference in mean reaction times between the condition with sound and without sound as the SOA increased did not occur. On the contrary, the effect of the auditory conceptual information decreases. It seems that the longer the time between the sound and the picture, the less the two are integrated. At an SOA of –500 ms, the effect of the sound has almost disappeared.

Finally there seems to be an effect of the delay on the mean reaction times of the conditions without sound. Subjects are slower in responding when there is a delay in the presentation of the picture of 500 ms compared to no delay. Subjects' attention levels not being as high when there is more time between the fixation point and the presentation of the picture could cause the effect.

As mentioned before, in experiment 1 no effect of the connotation of the sound was found. To test a situation where there is a more constant relationship between a category and a type of response a third experiment was run, where a Simon-like procedure was used and the consequence was studied of an irrelevant distracter (the earcon) on the response ('yes' or 'no').

Experiment 3: Relation Between Connotation of Sound and Response; Simon Effect

Subjects
Eleven subjects participated in the experiment, who received course credits or were paid for their participation. They were all students at the Catholic University of Nijmegen.

Materials and design
The same 16 pictures were used as in experiment 1. Furthermore the same auditory stimuli were used, i.e. C4major, C4minor, C5major, C5minor and F4. The sounds were exhaustively combined in pairs and then related to the categories of the experiment. For example, in one of the conditions (see table 2.3, first line) the pictures of animals were accompanied by a high-major sound (C5major) and the pictures of non-animals were accompanied by a high-minor sound (C5minor).

Relation	Animal	Non-animal
Congruent	C5major	C5minor
Congruent	C4major	C4minor
Congruent	C5major	C4minor
Congruent	C4major	C5minor
Incongruent	C5minor	C5major
Incongruent	C4minor	C4major
Incongruent	C5minor	C4major
Incongruent	C4minor	C5major
Neutral	C5major	C5major
Neutral	C5minor	C5minor
Neutral	C4major	C4major
Neutral	C4minor	C4minor

Table 2.3. The conditions of experiment 3.

The order of the conditions was randomized between subjects. For every condition the trials were randomized as well. Within the 12 conditions every picture was presented twice, once with the sound and once without sound. This resulted in 384 test-trials in total. Furthermore there were 10 practice trials. The same equipment was used as in experiment 1 and 2.

Procedure
Again the same procedure was used as in experiment 1 and 2. However, by using different conditions, the relation between the sound and the intended response was

tested. Similar to experiment 1, the picture and the sound were presented at the same time (SOA is 0 ms). After every condition subjects could take a break.

Results

The reaction times were measured from the moment the picture appeared. Only reaction times to correct responses were included in the analysis. In 3.1% of the trials an error was made. Since the number of errors was low, they were not analyzed further. The mean reaction time per relation and the standard error of the mean are presented in figure 2.4.

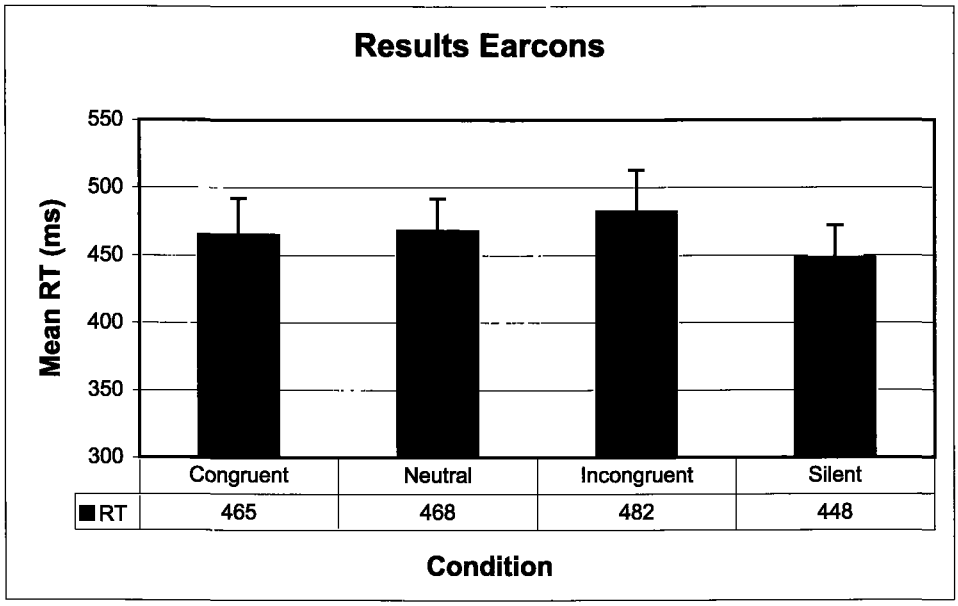


Figure 2.4. Mean reaction times (RT) per condition of experiment 3.

Similarly as in experiment 1 and 2, a repeated measurements analysis was conducted (with an alpha level of .05), that showed a significant difference in mean reaction times between conditions ($F(2,9) = 4.569$, $p < .05$). Furthermore, there is a significant difference in mean reaction times between the congruent and the incongruent condition ($F(1,10) = 5.186$, $p < .05$). In the congruent condition the reaction times are faster than in the incongruent condition. The mean reaction times of the congruent condition and the neutral condition do not differ from each other ($F(1,10) = 0.190$, $p > .1$).

Discussion

The results indicate that when there is a fixed combination between a type of sound (major or minor) and a response ('yes' or 'no'), a negative Simon effect occurs: when the type of sound is incongruent with the response, there is inhibition. It seems that the integration of contradicting information from both modalities takes longer than the integration of confirming or neutral information.

Although there is a facilitation effect of congruent information compared to the incongruent information, there is no difference found between the neutral and the congruent condition. This would indicate that it does not matter whether the auditory information is irrelevant or in agreement with the visual information. It seems that in this task, there is only an effect when the information is contradicting.

Similar to the first two experiments, a general inhibitory effect of earcons is found; every condition where the sound is played with the visual information is slower than the silent condition.

General Discussion

The results from the experiments described in this chapter confirm the hypothesis that it takes longer for conceptual, auditory information (earcons) to be integrated with visual information in a visual categorization task compared to a situation with no additional auditory information.

Secondly, the inhibitory effect of the earcons seems to be greatest when the auditory information is presented simultaneously with the visual information. The results suggest that when there is an SOA between information in the two modalities that is greater than 500 ms, the information is no longer integrated and the auditory information is disregarded when completing the visual categorization task.

When the conditions with sound are studied more closely, there seems to be a negative Simon effect. The trials where the sound is contradicting the type of response that needs to be given, are slower than the other conditions. So apart from the general inhibition effect of the auditory information that is present, there is a further delay in mean reaction times when the information is suggesting the opposite response.

Apart from reaction times, errors are also a good indication of the cognitive processes that take place during a certain task. In the experiments mentioned here there were too few errors to take them into account. It seems that the task is relatively simple and subjects have no difficulty completing it. A possible way of increasing the difficulty of the task would be to instruct subjects to complete a second task at the same time, like for instance to calculate for every category the cumulative sum of digits presented visually with the pictures. Since subjects now have to divide their resources to complete both tasks two things can happen. On the one hand they could make more errors and on the other hand they could take more time to complete the task. As a result the reaction times would increase. What strategy is selected and in what conditions more errors are made can provide more insight in the mental processes involved.

Lastly, in these experiments only earcons were tested. Further research is needed comparing these results to the effect of concrete, categorical auditory icons on visual categorization.

Chapter 3

Musical Experience and Earcons[4]

Abstract

Earlier experiments on visual categorization with major and minor earcons as accessory stimuli showed a negative Simon effect for incongruent combinations of stimuli and accessory stimuli. In this study musically experienced and musically inexperienced subjects were distinguished and an SOA of 500 ms between the presentation of the sound and the picture was used to test whether this would increase the differences between conditions. Results confirm that both musically experienced and inexperienced subjects respond slower in conditions with earcons compared to the silent condition. The SOA however did not lead to increased differences between conditions. It seems that subjects are more affected by incongruencies between stimuli than congruency or neutrality.

[4] This chapter is a revised version of a separate publication:
Bussemakers, M.P., de Haan, A., & Lemmens, P.M.C. (1999). The effect of auditory accessory stimuli on picture categorisation; implications for interface design. *Proceedings of the 8th Human Computer Interaction International conference* (pp. 436-440). Munich (Germany): Lawrence Erlbaum Associates.

Introduction

In multimodal interaction users need to integrate information from separate modalities to process the full message (e.g. Bussemakers & de Haan, 1998). An example related to speech where sound and visual information are integrated can be observed when listening to a lecture in a big hall. The lecturer is speaking into a microphone and the sound is transmitted through loudspeakers that are located on either side of the hall. Although the sound actually is coming from your left and right, you integrate the movement of the lecturer's lips with the sound in such a way that it seems as if the sound is actually coming from the lecturer (i.e. the ventriloquism effect (Howard and Templeton, 1966)).

When designing non-speech audio that is going to be presented together with visual information, two types of sounds can be used. Auditory icons (e.g. Gaver, 1986, 1989) are concrete sounds that are perceptually similar to an object or event in the real world. An example of an auditory icon is the slamming of a door. Earcons (e.g. Blattner, Sumikawa & Greenberg, 1989) on the other hand are often musical and their connotation refers to a more abstract concept of an event or object. An example is a major or minor key chord. Historically studies have reported that major chords have a positive connotation and minor chords have a negative connotation (Hevner, 1933; Crowder, 1984, 1985a, 1985b).

Although both auditory icons and earcons are valuable sources of information, the nature of the message they are carrying is similar to the distinction in memory between categorical and conceptual information (e.g. Shanks, 1997).

Categorical information allows an organism to generalize about what it has learned about an object or an event to a similar object or event and to adjust its behavior to its surroundings. For instance when you have a pet cat and you see the neighbor's pet, you know that it is a cat because of your encounters with your own animal. Categorical knowledge is limited because it depends on the perceptual similarity between in this case your cat and your neighbor's animal. Based on perceptual similarity, we might be tempted to not only categorize our neighbor's pet as a cat, but also classify other animals as cats, like for instance tigers, cheetah's and lions when in fact they are not. Conceptual knowledge, a mental representation of the class of objects and events, allows us to make the distinction between cats, lions and cheetah's by principles instead of perceptual similarity. The benefit of conceptual knowledge is that decisions can be made on the basis of deeper and more abstract properties of objects (Gelman & Markman, 1986, Goldstone, 1994). Furthermore inferences can be drawn for all entities of the concept. For instance if you learn that cats live with people, you know that this most likely will be the case for your neighbor's cat as well, but not for lions or cheetah's.

In this chapter part of a research project is reported, that aims to study and predict the effect of categorical (auditory icons) and conceptual (major and minor earcons) auditory information as additional stimuli in a visual categorization task.

Previous Results

Earlier experiments with these earcons showed a negative Simon effect. In a traditional Simon paradigm, the (positive) influence is studied of an irrelevant distracter on the response, like for example the location of a stimulus on a screen in relation to the location of a response-button (e.g. Simon, 1990). Subjects respond faster when the stimulus location is congruent with the button location, than when the stimulus location is incongruent with the location of the button.

In the project mentioned here, the connotation of major and minor earcons is used as an irrelevant distracter that is related to the response ('yes' or 'no') in a visual categorization task. Subjects are instructed to press a button labeled 'yes' when they see a picture of an animal and 'no' when they see a picture of something other than an animal (for instance a boat).

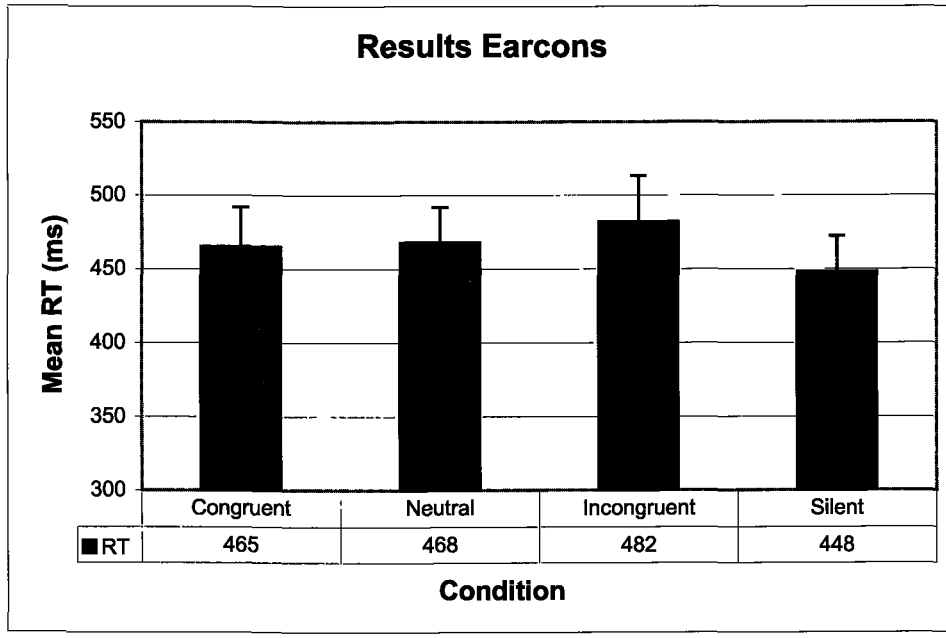


Figure 3.1: Results of categorization task of first series of experiments.

The reaction-time results indicate that, compared to both conditions where the earcon is congruent with the response and to the silent condition, subjects are significantly slower in the conditions where the earcon is incongruent with the type of response (Bussemakers & de Haan, in press), i.e. for instance a condition where a minor chord is played with pictures of animals and a major chord is presented with the pictures of non-animals (see figure 3.1). This suggests a negative Simon effect for the incongruent condition. There is no difference in reaction times between the congruent condition (animal with major, non-animal with minor) and the neutral conditions (both categories with the same sound, either major or minor).

Experiment: Difference in Musical Experience

In the experiment described in this chapter an attempt was made to validate and further understand the earlier results. Earcons are musical stimuli, so it is possible that the degree of musicality of the subject influences the responses.

It is known from neurological studies that both attentively and pre-attentively there is a difference in the cognitive and automatic processing of major and minor chords by musicians and non-musicians (Siegel & Siegel, 1977; Koelsch, Schroeger & Tervaniemi, 1999). When chords are presented to subjects pre-attentively, there is even a mechanism observed where musicians only 'notice' the stimulus if a presented chord is slightly diminished. In the Koelsch et al (1999) study violinists were instructed to perform another task, like reading, while listening to musical chords. Some of these chords were not in tune. EEG measurements indicated that some areas of the brain became active only when the diminished chords were played even though the subjects were not instructed to pay attention to the music. They concluded that this difference seems to be noticed on a more subconscious level.

The results of studies like the one by Koelsch et al (1999) could indicate that even when we are not explicitly paying attention, for instance when we are busy with a visual task, we only notice the sounds if they do not correspond with our expectations. It is possible that this pre-attentive 'check' can also be observed cross-modally in situations where the sound is incongruent with the response to the visual information. If so, an effect of incongruence and not so much of congruency would be found in the results.

To test this and to allow ample time for the subjects to process the auditory stimulus and relate the sound to the pictorial stimulus, the previous experiments were extended by adding a Stimulus Onset Asynchrony (SOA) of 500 ms (similar to the second experiment in chapter 2). It was expected that if the previously found similar mean reaction times between the congruent and the neutral condition were caused by a lack of time to process the auditory information fully, the SOA in this experiment would lead to greater differences in mean reaction times between the conditions. If on the other hand the lack of difference in mean reaction times between the two conditions was caused by a pre-attentive check to see if the stimuli are not incongruent, the SOA would not have an effect. It would then be expected that the mean reaction times of the congruent and the neutral conditions would not differ from each other.

Subjects

The participants in this study were students of Psychology or Cognitive Science at the Catholic University of Nijmegen. Of the 14 participants, 7 participants were musically experienced and 7 were inexperienced. The criterion that was used to determine whether or not a subject was musically experienced was 6 years or more of experience with playing a musical instrument vs. no experience with playing a musical instrument. Subjects were paid or received course credits for their participation.

Materials and Design
Similar to earlier experiments, the same 16 line drawings were used as visual stimuli as well as the same chords as auditory stimuli (C4major, C4minor, C5major, C5minor) (e.g. Bussemakers & de Haan, 1998).

Relation	Animal	Non-animal
Congruent	C5major	C5minor
Congruent	C4major	C4minor
Congruent	C5major	C4minor
Congruent	C4major	C5minor
Incongruent	C5minor	C5major
Incongruent	C4minor	C4major
Incongruent	C5minor	C4major
Incongruent	C4minor	C5major
Neutral	C5major	C5major
Neutral	C5minor	C5minor
Neutral	C4major	C4major
Neutral	C4minor	C4minor

Table 3.1. The conditions of the experiment.

Similar to a previous experiment the sounds were combined in pairs and then related to the categories, i.e. animal or non-animal (see table 3.1). Participants were presented with all conditions. The order of the conditions was randomized between subjects. Within every condition the trials were presented in random order as well.

The materials that were used were the same as in the previous experiments. A Macintosh Quadra 840 AV with a 14-inch screen that was raised to eye-level presented the stimuli. Furthermore a pair of Monacor BH-004 headphones was used to present the sounds. The manual responses were registered by a button-box with voice-key. The assignment of the buttons was randomized between subjects.

Procedure
Subjects were first asked for their level of musical experience. Then the instruction was read and when there were no more questions regarding the experiment, the subject put the headphones on and the experiment was started. After the practice trials, the subject had a break where he or she could ask additional questions. Then the test-trials were started. After every condition the subject had time to rest.

Every trial started with a single frequency tone (F4) together with a fixation point ('+'), which was presented in the center of the screen for 500 ms. Subjects were instructed to focus their eyes on the fixation point. When the fixation point disappeared, after 1000 ms a sound was presented or there was a silence. After 500 ms a picture was shown for 300 ms and subjects could indicate their response to the question whether or not the picture they saw was of an animal by pressing a button labeled 'yes' or 'no'. After 2.5 seconds another trial was started.

Results

Only correct responses were included in the analysis. In 2.2% of the trials an error was made. Because of the low percentage of errors, they were not analyzed further. The resulting general mean reaction time per condition is presented in figure 3.2.

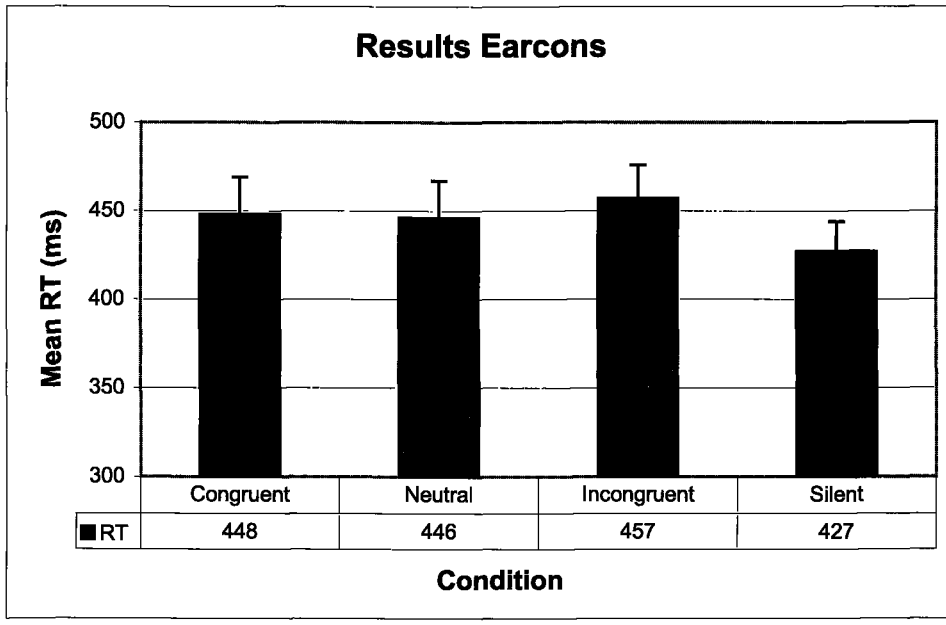


Figure 3.2: Overall results categorization with SOA.

Similar to the earlier findings, the results show a statistical significant difference in mean reaction time per condition ($F(3,11) = 21.992$, $p < .001$). The conditions with sound have higher mean reaction times than the silent condition. Furthermore, the congruent and the neutral condition differ significantly from the incongruent condition ($F(1,13) = 5.349$, $p < .05$). Reaction times in the incongruent condition are significantly slower. However, the addition of the SOA did not result in a difference between the congruent condition and the neutral condition ($F(1,13)$ 0.109, $p > .5$)[5].

The differentiation on musical experience showed a trend, though not significantly, in differences in reaction times between musically experienced and musically inexperienced participants ($F(3,10) = 0.171$, $p > .1$) (see figure 3.3).

[5] These findings are similar to the results of earlier studies (Bussemakers & de Haan, in press).

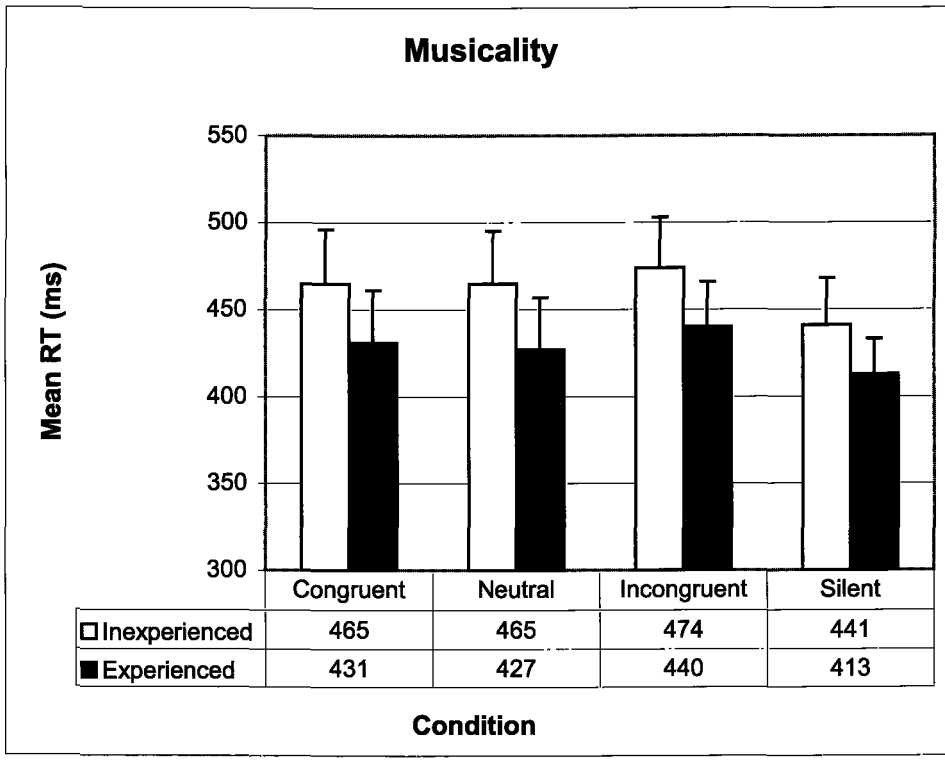


Figure 3.3: Results separated for musically experienced and inexperienced people.

Conclusions

The results confirm the findings from previous studies. First of all, adding earcons to a visual categorization task leads to longer reaction times both for musically experienced and musically inexperienced subjects. Apparently it takes longer for conceptual auditory information to be processed and interpreted when compared to a situation with no accessory information.

Secondly, a negative cross-modal Simon effect was again observed; subjects respond significantly slower when the auditory information is incongruent with the type of response that needs to be given, compared to the mean reaction times in the congruent and neutral conditions.

Between the congruent and the neutral condition no significant difference in mean reaction times was found. The longer processing time, by introducing an SOA of 500 ms did not seem to have an effect on the mean reaction times.

As can be observed in the results, the differentiation on musical experience only showed a trend, but did not lead to significant differences in mean reaction times. Perhaps if more subjects were tested, the differences would be greater. It is also possible, that the level of 6 years playing a musical instrument was not a suitable

criterion. A third possibility is that since most children in school in The Netherlands are taught music for at least a year, there is also musical knowledge present in the inexperienced group. The differences in mean reaction times between conditions could substantiate this, because the relative effects between conditions seem to be similar between groups. Both groups seem to show the same pre-attentive 'check' for discrepancies that is reported with musicians in neurological studies (Koelsch, Schroeger & Tervaniemi, 1999). Only the check in this study is not unimodal but across modalities. In further studies the distinction in musicality could be made more explicit, for instance by using professional musicians versus Psychology students.

Another suggestion for further research would be to conduct a double-task experiment, where subjects have to accomplish another task while performing a visual categorization task. A double-task could lead to longer reaction times and/or more errors, which could provide new insights on the effects observed. For instance a certain multimodal situation where the response times are higher may be acceptable, because the number of critical errors is low. The opposite might hold true as well: if in a certain situation the reaction times are fast, but there are many errors, it might not be the best solution.

Chapter 4

Earcon Experiment with Dual Task[6]

Abstract

An experiment was conducted with two picture categorization tasks to study the effect of abstract, redundant auditory information (earcons). In the first task participants had to perform just the categorization. The second task consisted of the categorization, plus an extra cumulative addition task. The dual-task situation was expected to lead to longer reaction times and more errors than the single-task situation. The results indicated that mean reaction times increased for the dual-task situation. There was no difference in proportion of errors between tasks or between conditions, except for a lower proportion of errors in the congruent condition compared to the silent condition in the dual-task situation. It seems that congruent conceptual auditory information can assist in dual-task situations where a low number of errors is critical.

[6] This chapter is a revised version of the paper that has been published as: Lemmens, P.M.C., Bussemakers, M.P., & de Haan, A. (2000). The effect of earcons on reaction times and error-rates in a dual-task vs. a single-task experiment. *Proceedings of the International Conference on Auditory Display* (pp. 177-183). Atlanta (USA):International Community for Auditory Display.

Introduction

One of the areas in human-computer interaction and interface design that is more and more under investigation is multimodal interaction. Researchers as well as designers want to learn what the consequences are of communicating information to the user through the visual, auditory or even the haptic modality.
When looking at the combination of visual and auditory information, earlier research has shown that if auditory information is used to supplement visual information, for instance because the eyes are otherwise occupied or if it is difficult to perform the task with only the visual information, sounds can lead to faster response times and fewer errors (e.g. Brewster, 1999).

In this project, the effect of redundant auditory information is investigated on a primary visual categorization task. This means that the information that is presented through audio is also available visually and users are able to perform the task without listening to the sound. Redundancy can be functional especially when users have to perform two tasks at the same time. The information is present in two modalities, so if for instance the visual modality is occupied by a task, the user can get the necessary information for another task via the auditory modality.

Within our project two types of sounds are used, representing two types of information. Auditory icons (Gaver, 1989; Mynatt, 1994) are concrete, real-life sounds that are perceptual and categorical in nature. The meaning of such a sound seems to be determined by earlier experiences with similar sounds. Auditory icons are mostly described by the source of the sound: the principle of everyday listening (Gaver, 1993a, 1993b). An example is the sound of a door closing.

The other type of sounds that is used in the experiment presented here is an abstract, musical, major and minor chord, an earcon (Blattner, Sumikawa & Greenberg, 1989). This type of sound represents conceptual, rule-based information, i.e. their connotation refers to a more abstract concept. Historically major chords have a positive connotation and minor chords have a negative connotation (Hevner, 1933; Crowder, 1984).

In the experiment users have to categorize line drawings of animals and non-animals (e.g. a chair) by indicating whether the picture they see is that of an animal, 'yes' or 'no'. At the same moment they see the picture the earcon is played.

Earlier results have shown that adding conceptual information through the auditory modality can slow down the performance if the sound does not coincide with the intended response. For instance when pictures have to be categorized as animal or non-animal a minor chord presented together with a picture of an animal leads to slower responses (e.g. Bussemakers & de Haan, 1998; Bussemakers & de Haan, in press). However, only looking at the speed of the responses possibly does not give us a full idea of what happens in these multimodal situations. Errors also are a source of information on what processes take place on a cognitive level.

Because in this categorization experiment, subjects make very few errors, a second task was added to perhaps induce more errors. The results of these tasks are presented in this chapter.

The assumption that a secondary task can increase the cognitive load for participants is based on the theory of a shared pool of mental resources which participants (or people in general) have available (McLeod & Posner, 1984; Logie, 1996; Pashler & Johnston, 1998). Carrying out two tasks simultaneously sometimes requires a larger amount of resources than carrying out only a single task. This could lead to longer reaction times and/or more errors in performance.

There are tasks however that can be carried out at the same time without interference, because of a so-called 'privileged loop' (McLeod & Posner, 1984), an automatic link that leads incoming perceptual information directly to actions. There is no loss of performance if the resource is allocated to another modality or if it is not the same task that needs to be carried out simultaneously.

The experiment described in this chapter uses picture categorization as a primary task and addition in one head of single-digit numbers per category as a secondary task, i.e. subjects need to remember and add to the cumulative sum of the digits in the pictures of animals and the pictures of non-animals separately. Here there does not seem to be a privileged loop, because both the mental addition and the processing of the sounds are a cognitively mediated task. It is thus expected that interference will occur and that there will be more errors and longer reaction times compared to a single task experiment.

When selecting subjects their experience level in music was noted, although earlier experiments have shown that there was little influence of the level of musical experience on the effects observed when the criterion was 6 years of playing a musical instrument (Bussemakers, de Haan & Lemmens, 1999).

Method

Subjects
Participants were 20 students (with an average age of 23 years) of the Catholic University of Nijmegen. Ten were considered musically experienced because they had been playing a musical instrument for at least 8 years and 10 were considered musically inexperienced. They were paid or received course credits for their participation.

Design
In the experiment a within-subjects design was used, in which in each of the four conditions the relationship between the type of earcon (major or minor) and the type of response (yes or no) was manipulated.

The design is similar to previous experiments (e.g. Bussemakers & de Haan, 1998):

- Congruent condition: all animal pictures were combined with the major earcon; all non-animal pictures were combined with the minor earcon.
- Incongruent condition: all animal pictures were combined with the minor earcon and all non-animal pictures with the major earcon.
- Neutral condition: both the animal and non-animal pictures were exhaustively and randomly combined with the major as well as the minor earcon.
- Silent condition: no sounds are played in conjunction with the pictures.

Procedure

With this design the participants carried out two different tasks. In one task the subjects only performed the categorization, whereas in the other task apart from the categorization task the extra, cumulative addition task had to be completed. The order of the dual-task versus single-task situation was randomized between subjects. Furthermore, to prevent effects of a preferred hand, the position of the 'yes' and 'no'-buttons on the button-box was varied between subjects. Finally, the musically experienced participants were spread equally over both task sequences.

Materials

The experiment was carried out on a Macintosh Quadra 840AV. A button-box with voice-key was connected to the computer as well as a pair of headphones, through which the earcons were presented. In this experiment the 16 line drawings earlier used (e.g. Bussemakers & de Haan, 1998) were modified to include a single-digit number in a position close to the center of the drawing (see figure 4.1). The earcons used were a C4 and C5 chord both major and minor with a duration of 2500 ms. The alert sound was a single, F4-tone.

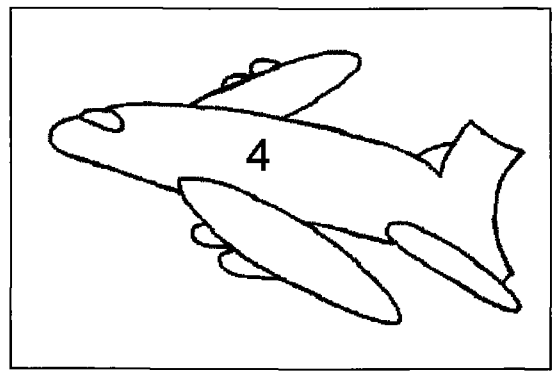


Figure 4.1. Example of non-animal with digit to accumulate.

Results: Reaction Times (RT's)

For the statistical analyses of the reaction times, all incorrect responses and null-reactions were disregarded.

The average reaction times for both tasks without differentiation on musical experience are shown in the figure below. A repeated measurements analysis showed a significant difference in mean reaction times between the conditions with the cumulative addition task and the conditions with just the categorization ($F(1,19) = 31.196$, $p < .001$) (see figure 4.2). The dual-task conditions are slower than the single-task conditions.

Within the addition task a significant difference in reaction times between conditions was found ($F(3,17) = 6.429$, $p < .005$). A contrast analysis showed that the conditions with sound were significantly slower than the condition without sound ($F(1,19) = 16.255$, $p < .001$). However, between the conditions with sound no significant difference in mean reaction times was found ($F(1,19) = 0.014$, $p > .5$).

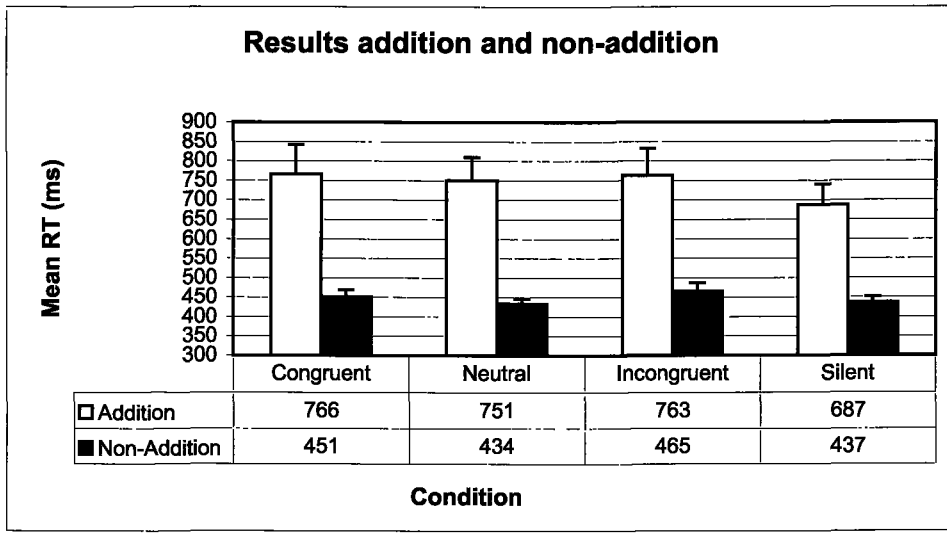


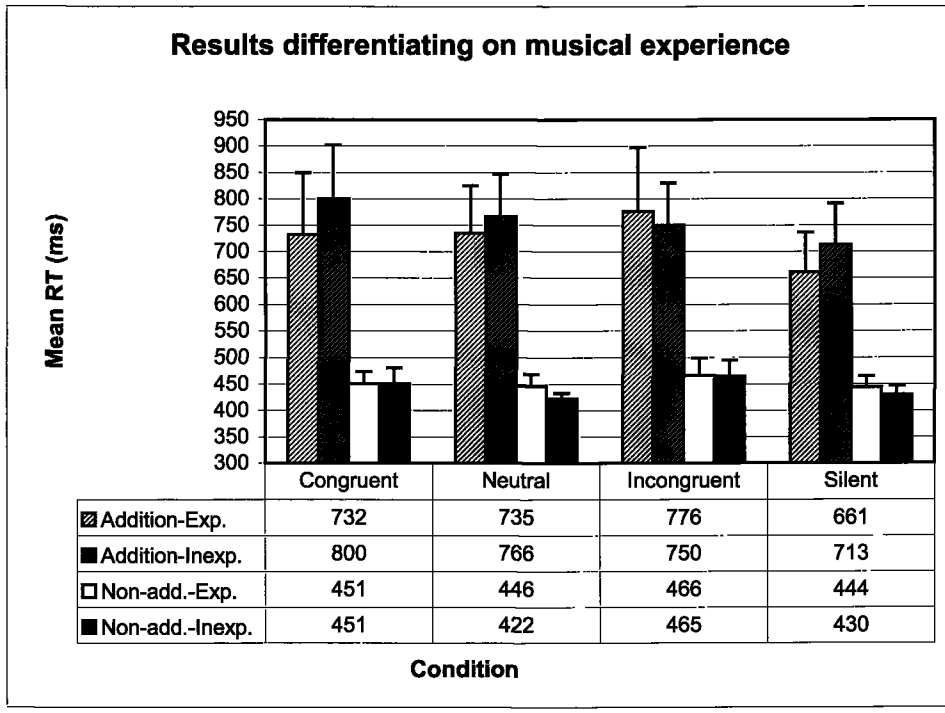
Figure 4.2. Average reaction times in ms for the dual-task situation versus the single-task situation.

For the non-addition task, the difference between the conditions with sound and the condition with no sound was not significant ($F(1,19) = 3.622$, $p > .05$). However the results showed a trend that is similar to the addition task: the conditions with sound were slower than the condition without sound. Similarly, the incongruent condition was slower than the other conditions with sound, though the result was not significant ($F(1,19) = 3.804$, $p > .05$). More specifically, there was a significant difference, between the incongruent and the neutral condition ($F(1,19)$ 8.325, $p < .01$), but no significant difference between the congruent and neutral condition ($F(1,19)$ 2.275, $p > .1$).

Musically experienced vs. inexperienced

When distinguishing between the subjects with musical experience and with less musical experience, there is no difference in mean reaction times ($F(1,18)$ 0.143, $p > .5$) (see also figure 4.3).

In the task with the cumulative addition, within the group of musically experienced participants, a difference in mean reaction times between conditions with sound and the condition without sound was found ($F(1,9) = 7.687$, $p < .05$). The conditions with sound were slower than the condition without sound. Between the conditions with sound however, no significant difference in mean reaction times was found ($F(1,9) = 0.414$, $p > .5$).



	Congruent	Neutral	Incongruent	Silent
Addition-Exp.	732	735	776	661
Addition-Inexp.	800	766	750	713
Non-add.-Exp.	451	446	466	444
Non-add.-Inexp.	451	422	465	430

Figure 4.3. Average Reaction Times in ms for all tasks.

For the musically inexperienced group of participants there was also a difference in mean reaction times between the conditions with and the condition without sound in the addition task ($F(1,9) = 9.840$, $p < .05$). Again the conditions with sound are slower than the condition without sound. Between the conditions with sound there is no significant difference in mean reaction times ($F(1,9) = 0.863$, $p > .1$).

The non-addition task both for experienced and inexperienced subjects showed the same trends as found in the addition task, but the differences in mean reaction times were not significant.

Error analysis

The error analysis showed that 4.1% of the total number of responses was incorrect. Of these errors about half (49.75%) were made in the addition task, the other half in the non-addition task.

The mean proportion of errors in all tasks is presented in the table below. A repeated measurements analysis did not show the differences in average number of errors between both tasks to be significant ($F(1,19) = 0.051$, $p > .8$).

	Addition	**SE**	**Non-addition**	**SE**
Congruent	0.021	0.007	0.035	0.013
Neutral	0.037	0.009	0.041	0.013
Incongruent	0.038	0.012	0.035	0.014
Silent	0.045	0.009	0.038	0.007

Table 4.1. Mean proportion of errors and the standard error of the mean (SE) per task and condition.

A contrast analysis showed only the difference between the congruent and the silent condition in the addition task to be significant ($F(1,19) = 6.216$, $p < .05$), showing that there are more errors made in the silent condition compared to the congruent condition.

Discussion

This experiment investigated the effects of redundant conceptual auditory information on visual categorization in the context of a dual-task paradigm, where subjects had to accumulate digits presented in the pictures for every category.

The reaction time data from the addition task shows the same effects as earlier studies: having earcons in a visual categorization task leads to longer reaction times than when only the visual information is presented (the silent condition). This seems to be a robust effect, which is not influenced by a dual-task setting. In the non-addition task there is a trend that seems to confirm the earlier observed negative Simon effect. In the incongruent condition the reaction times are slower when compared to the congruent and the neutral condition.

Furthermore earlier findings have been confirmed that there is no significant difference in mean reaction times between experienced and inexperienced musicians (Bussemakers, de Haan & Lemmens, 1999), at least in the definition as it is used here. Perhaps if professional musicians were tested against for instance Psychology students the results would be different.

Finally, the results show that the reaction times in the dual task are significantly higher than in the single task. Subjects seem to 'select' a strategy to slow down the response in order to make as few errors as possible. The number of errors increased compared to earlier studies (Bussemakers & de Haan, 1998; Bussemakers & de Haan, in press). Between tasks or between conditions, generally the proportion of errors did not differ. Only in the addition task, there were fewer errors in the congruent condition than in the silent condition. It seems that when there is a higher cognitive load the subjects are able to use congruent auditory information to reduce the number of errors, compared to a situation with no auditory information. Furthermore, in all conditions there is an inhibition of the reaction times because of the additional auditory information that is present.

This inhibition does not depend on the relationship between the sound and the picture. It is even observed in the neutral condition, where for all pictures the same sound is played. It seems that this is a general effect as a result of the additional information that needs to be processed.

Further research is needed to look at the effects, both in a single task and a dual task, of *categorical* auditory information (auditory icons) on reaction times and errors in a visual categorization task.

Chapter 5

Auditory Icon Experiment with Single Task[7]

Abstract

In this research a categorization paradigm is used to study the multimodal integration processes that take place when working with for example a computer interface. Redundant auditory icons are used with visual information to investigate their influence on the categorization of pictures. Reaction time results indicate that responses are faster in conditions with auditory icons (categorical information) than conditions with no sound. Earlier experiments with redundant earcons (conceptual information) showed that reaction times in conditions with earcons are slower than conditions with no sound. These findings suggest integration at different stages of processing for categorical or conceptual information.

[7] This chapter is a revised version of the paper that has been published as: Bussemakers, M.P., & de Haan, A. (2000). When it sounds like a duck and it looks like a dog... Auditory icons vs. earcons in multimedia environments. *Proceedings of the International Conference on Auditory Display* (pp. 184-189). Atlanta (USA): International Community for Auditory Display.

Introduction

When we interact with devices, from a microwave to a computer, both visual information and audio is used, even if we are sometimes not aware of it. Although many computer users report that they turn the audio feedback from their system off because it becomes annoying after a while this does not mean that they do not continue to use audio feedback from their system. For example they still listen to the sound of the hard disk to determine whether a file is being saved or a data transfer is complete (e.g. Buxton, 1989), although the meaning of the sound needs to be learned. Apparently, when the mapping of the sound onto its function seems natural to the user and the sounds are non-intrusive, this type of feedback is useful and preferred.

Earlier research has confirmed that audio can be helpful in user-system interaction not only as feedback but also in a number of other situations (Buxton, 1989). First of all sound can assist to alarm users and display warning messages. Auditory perception in the brain is more directly connected to arousal systems in the nervous system than visual perception. Furthermore auditory perception cannot be easily suppressed (Jones, 1989). Even if our eyes are occupied or our attention is directed towards another modality, it seems impossible to ignore for example a siren going off.

Secondly sound can be used to present the status of a system or monitor for incoming messages. An example of this is the sound of a cup of liquids filling up to indicate the process of copying a file. The sound may hardly be noticed while it is there, but as soon at is gone the user notices its absence and realizes the system has finished copying the file.

Lastly encoded messages can be presented in audio (Buxton, 1989). It is possible to present complex data in a single sound, by using timbre, loudness and pitch variations. Gaver (1989, 1993a, 1993b) even takes data presentation one step further by defining a theory of 'sources' instead of mapping the attributes of the data onto the parameters of a sound. For those sounds the meaning is not based on its musical qualities, like loudness, pitch or timbre, but on the relationship between the sound and the object or event in the real world that is related to the sound: the principle of everyday listening (see also Vanderveer, 1979). An example of this is the sound of a slamming door or the sound of skidding car tires. Sounds like these are described by referring to its source, and not by its psychophysical properties.

From the definition of everyday listening a distinction becomes apparent between sounds that can be associated with a source or an event (auditory icons, (e.g. Gaver, 1989; Mynatt, 1994)) and sounds that are more abstract or musical in nature (earcons, (e.g. Blattner, Sumikawa & Greenberg, 1989)).[8]

[8] This differentiation is analogous to a distinction between categorical and conceptual information by Shanks (1997).

Earcons are abstract, because they do not refer to an object directly, but often have a connotation that is related to an abstract concept, like for example major and minor chords. Studies have shown that major chords have a positive connotation and minor chords have a negative connotation (Crowder, 1984, 1985a, 1985b). The meaning of these conceptual sounds within a certain context needs to be learned, but it is possible to create earcons that are informative and non-intrusive, and that assist users in their task.

Auditory icons are perceptual, categorical sounds. The meaning of an auditory icon is often intuitively clear, but it seems to be more difficult than with earcons to find an exact mapping between the sound and its function. In the example of the glass filling up with liquids, it seems clear what the sound is trying to tell us, but most users are aware that a file is not made of liquids. After a while, as mentioned before, this mapping can even become annoying. Especially in cases where the information is not critical but redundant, i.e. the same data is presented visually and through auditory icons, frequent users have a tendency to turn the sounds off. How well an auditory icon can be mapped onto a function or object depends on several factors (Mynatt, 1994). First of all it is important how often someone has heard the sound. This is referred to as its ecological frequency and determines the *identifiability* of the sound. Secondly its *conceptual mapping* onto a function is of importance. Thirdly the *physical parameters* like the duration, intensity and quality determine its usability. Lastly there is also an emotional response to a sound that can determine the *user preference*.

But how well do users respond to auditory icons in terms of speed and accuracy? Beltz et al (1999) tested auditory icons as well as more traditional tonal warnings in a car collision-prevention experiment. Subjects were instructed to drive in a simulator and to prevent accidents by either hitting the breaks (front collision prevention) or by moving to another lane (side collision prevention). Speed and accuracy was measured in multimodal conditions with a visual display as well as sound or unimodal conditions with audio only. The auditory icons were the sound of a car tire skidding for the front collision and a long horn honk for the side collision.

These sounds were developed through a methodology developed earlier (Beltz et al, 1997, 1998) to design auditory icons as warnings, where a particular sound is selected because of ratings on perceived urgency, meaning and level of association with the intended meaning. Having such a methodology is necessary according to Beltz et al (1999), because the implementation of auditory icons in complex systems was not successful in the past as a result of a mismatch between the signal and the event that it represents.

The results from this study showed a clear improvement both in performance and in preference in the conditions with the auditory icons compared to the tonal warning; reaction times in conditions with auditory icons were 80-100 ms faster than conditions with tonal warnings. Furthermore the responses improved more in the multimodal conditions versus the unimodal conditions, except for the conditions with auditory icons.

In those conditions there was no improvement in performance if there was also visual information present. It seems that the auditory icon in itself carries enough information to prevent a collision and no additional visual information is needed.

These types of studies are beginning to show that having redundant auditory information could assist users in their task. A more fundamental question remains however: Why do auditory icons work better than tonal sounds in this environment? Or, more general, what is the difference between conceptual auditory information (earcons) and categorical auditory information (auditory icons) on speed and accuracy of performance? To test and compare these different multimodal situations a paradigm is used, where users have to categorize pictures of animals or non-animals (for instance musical instruments) while hearing sounds. Response times are measured as well as error rates. Subjects indicate by pressing a button labeled 'yes' or 'no' whether or not the picture they see is of an animal.

Experimental Studies with Earcons
Earlier experiments with major and minor key earcons have indicated that there is a significant delay in reaction times for all trials with sound. It seems that having redundant conceptual information causes subjects to respond slower, although they were not instructed to pay attention to the sound.

Furthermore the delay is greatest in trials where the connotation of the earcon (positive or negative) does not correspond with the intended response (yes or no), i.e. the *incongruent* trials, where minor chords were played with pictures of animals and major chords were played with pictures of non-animals. This effect has been validated in several experiments (Bussemakers & de Haan 1998; Bussemakers, de Haan & Lemmens, 1999; Bussemakers & de Haan, in press) (see for the experimental conditions table 5.1 and for the results figure 5.1).

Relation	Animal	Non-animal
Congruent	Major	Minor
Incongruent	Minor	Major
Neutral	Major	Major
Neutral	Minor	Minor

Table 5.1 Experimental conditions earlier experiments.

Since these results involve abstract, musical sounds, it seems interesting to study, whether a similar effect occurs when categorical sounds, i.e. auditory icons are used in the same experimental setup.

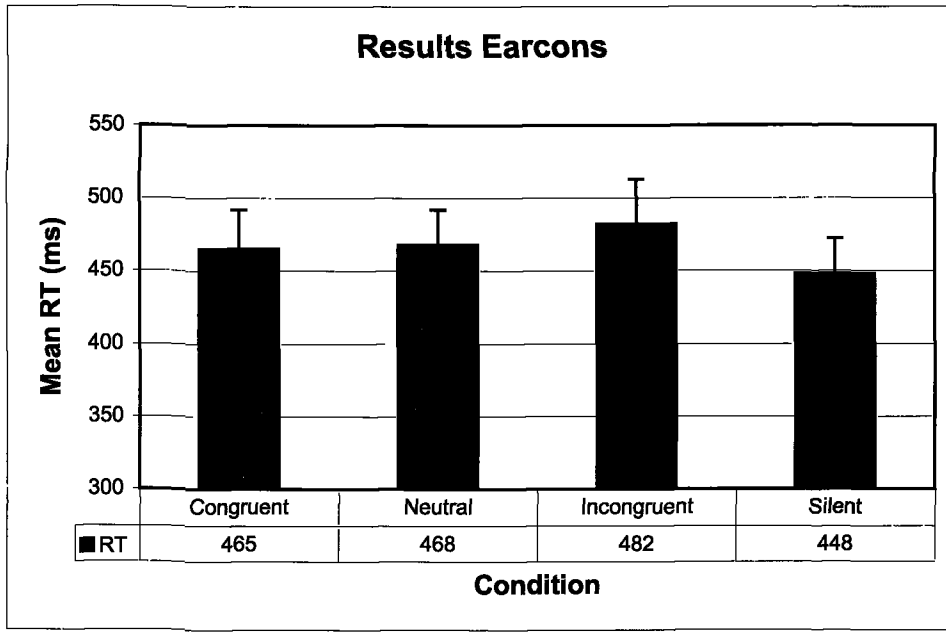


Figure 5.1. Experimental results earcons.

Experiment

Subjects
In this study 20 subjects participated, who were all students at the Catholic University of Nijmegen. They were paid for their participation. 18 Participants were female and 2 were male. The average age was 22 years.

Materials
A Macintosh Quadra 840AV was used with a 256 color screen, with a diagonal of 32 cm. The screen was raised so that subjects could see the visual stimuli at eye-level. Furthermore, a button-box was used with three buttons; one for each response category and a final one for starting each set of trials. The sounds were presented through a stereophonic headphone, a Monacor BH-004. Sixteen line drawings were used as visual stimuli in the experiments, 8 of animals and 8 of musical instruments.

The pictures were selected from a database used in other experiments, by taking the distinctiveness of the real-life sound associated with the picture as a selection criterion. The sounds that were used were wav-samples of animals and musical solo-pieces. The duration of each sample was normalized to 1.226 sec. Care was taken to choose samples of music that were closed, so subjects did not feel the music stopped in the middle of an expression. In a pilot-test the distinctiveness was tested with a larger pool of stimuli, by asking two subjects for each sound and each picture separately what they heard/saw. The 16 stimuli that they identified quickly and without errors were used in this experiment (see for example figure 5.2).

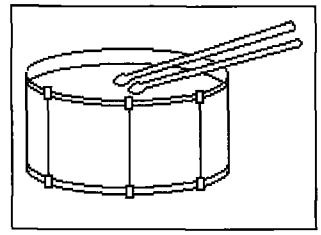


Figure 5.2. Example of stimulus.

Procedure
Subjects first heard a simple tone (F) to alert them to fixate on a displayed fixation point on the screen ('+'). Then a drawing was shown and in some cases a sound was played at the same time (Stimulus Onset Asynchrony of 0 ms). Subjects were instructed to respond as quickly as possible by pressing the button indicating their response, 'yes' or 'no'. After 2.5 seconds another trial was started and the alert sound was played again.

Three conditions were distinguished in this experiment. When the picture was accompanied by the correct real-life sound, this was called *same*. An example of this is a picture of a duck with the quacking of a duck. When another sound was played of the same category, this was labeled *same category*. For instance, the picture of a dog was shown and the quacking of a duck was played. The last sound condition is referred to as *other category*. This for instance when with the picture of a duck, an excerpt of guitar music is played. Finally, all pictures were also shown with no sound (*silent*).

The trials per condition were presented to the subjects in a randomized order (there was no need to present the trials in blocks), because the relation between the sound and the picture was at a trial level and not a condition level. In the instruction it was explained that this was a study to investigate the effect of sound on a task. They would see pictures and would have to indicate whether it was a picture of an animal or not. Subjects were instructed to respond as quickly as possible by pressing the button indicating their response, 'yes' or 'no'. Subjects were not specifically instructed to do something with the accompanying sound.

First, as practice session, subjects saw all pictures accompanied by the matching real-life sound. Participants also had to indicate in this session whether the picture was of an animal or not. Then subjects could ask questions and the experiment started. After a number of trials, subjects could take a break. The total experiment took about 25 minutes.

Results
The practice trials were excluded from the analyses. Also, error-responses or no-responses were left out. Since the number of errors and no-responses was small (less than 2%), they were not analyzed further.

The mean reaction times per condition are presented in Figure 5.3. The conditions with sound (*same, same category* and *other category*) are faster than the condition without sound. The shortest reaction times were observed in the *same* condition, where the picture corresponds with the sound. The condition where the sound presented with the picture was of the same category was slower. The condition with a sound from another category and the silent condition were slowest. A repeated measurements analysis showed a significant difference in mean reaction times between conditions ($F(3,19) = 9.713$, $p < .001$). The conditions with sound are significantly faster than the silent condition ($F(1,19) = 5.392$, $p < .05$). Table 5.2 shows a comparison in mean reaction times between the different conditions.

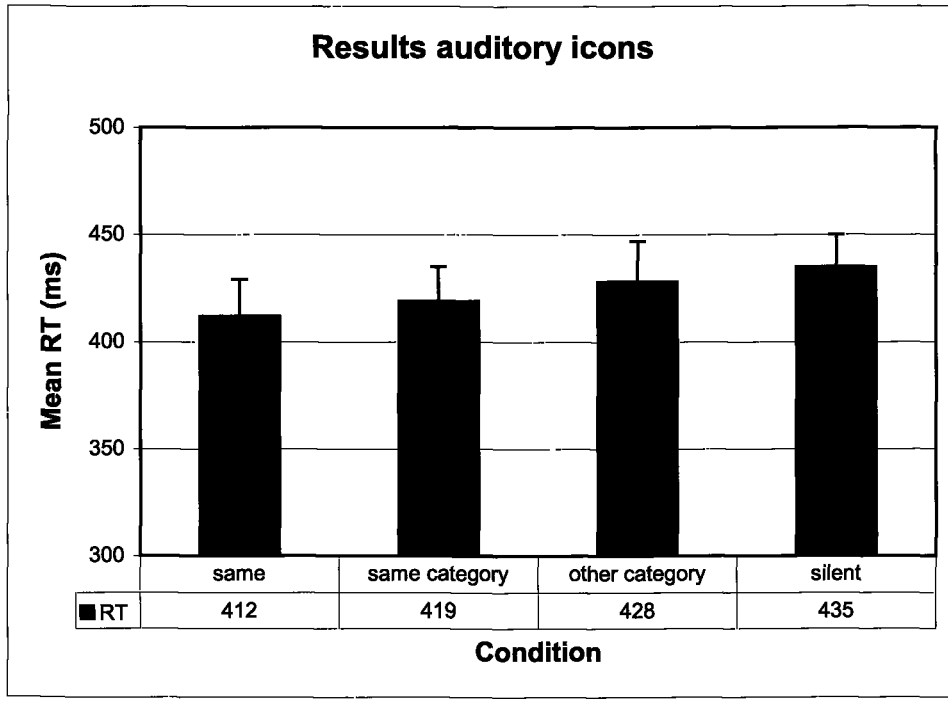


Figure 5.3. Mean reaction times per condition.

When comparing the mean reaction times between conditions, most differences are significant, except for the difference in mean reaction times between the other category condition and the silent condition. Also, the same condition consists of only a few trials (each picture has only a single corresponding sound), so it is hard to compare the results to other conditions (see table 5.2).

When comparing the conditions that have the same category (*same and same category*) there is no significant difference in mean reaction times (see table). This is not surprising, because one seems to be a subset of the other, since both sounds are from the same category as the pictures.

Combining the data from both conditions (see figure 5.3) and comparing this to the *other category* condition in a contrast analysis, there is a significant difference in mean reaction times ($F(1,19) = 6.109$, $p < .05$). The reaction times in the conditions with the same category (including trials with the same sound and picture) are faster than the reaction times in the conditions with another category.

Condition	F(1,19)	*p*	
Same vs. Same category	1.481	0.238	
Same vs. Other category	6.367	0.021	*
Same vs. Silent	13.334	0.002	*
Same category vs. Other category	3.189	0.090	
Same category vs. Silent	6.795	0.017	*
Other category vs. Silent	0.482	0.496	

Table 5.2: Contrast results comparisons between conditions (* significant with alpha of .05).

Discussion

The results show significantly faster reaction times in the conditions with auditory icons compared to the silent condition. It seems that in this setting, having information in another modality assists in a categorization task. Users are able to respond faster when they not only see a picture, but they hear an accompanying sound as well. This result is interesting, because as already mentioned users were not instructed to pay attention to the auditory stimulus. The task was to categorize the visual pictures. It seems that subjects do not shut out the auditory information to focus entirely on the pictures. Instead they seem to use the information in both modalities to come to a faster response. Nevertheless the mean reaction times between the *other category* condition and the *silent* condition do not differ significantly. This could indicate that, when the information is not of the same category as the pictures subjects have to categorize, for instance in the case where you see a violin and you hear a dog barking, the sound does not facilitate the response. It seems, that having sound present only contributes to the categorization when this information is more or less congruent with the picture information. When what you see and what you hear suggest the same type of response, subjects are able to react faster.

The effects observed are different from those observed in earlier studies with earcons. There, the mean reaction times in the conditions with sound were significantly slower than the conditions without sound: a modulating delay-effect. It seems that there is a difference in response times between a situation with additional categorical auditory information and a situation with additional conceptual information.

Categorical information that is perceptual in nature is (according to earlier statements in this dissertation) integrated at a response level. Having such multimodal information can speed up the response because of an alerting effect of one stimulus on the other (e.g. Welch & Warren, 1986).

It is known that auditory information is processed faster than visual information (Stein & Meredith, 1993), so it seems that the auditory information could alert to the visual information, decreasing the time that is needed to come to the response. In the case of the more abstract conceptual, rule-based information in the earcons, both stimuli are analyzed separately before the integration takes place. This increases the cognitive resources that are needed to process the information, which could be responsible for the increased response times.

The duration of the auditory icons was shorter than the earcons, because in the earcons after 1.226 sec. there was a long decay time of the signal. In other experiments with shorter earcons however, similar effects were found (Lemmens, Bussemakers & de Haan, submitted).

In the experiments presented in this chapter subjects made very few errors. However besides reaction times errors are a valuable source of information on the cognitive processes involved in tasks. Earlier tests with earcons have shown that a dual-task paradigm can induce more errors. In a dual-task experiment subjects have to cumulate digits for every category that were presented with the pictures. Further studies investigating auditory icons as redundant information in a dual-task situation are needed.

Chapter 6

Auditory Icon Experiment with Dual Task[9]

Abstract

A dual-task experiment was conducted where subjects had to categorize pictures presented with auditory icons, while at the same time they had to add to and remember per category the sum of a single-digit number presented in every picture. Results showed, similar to earlier findings that subjects respond faster in conditions with meaningful sounds than in the silent condition. With the extra task the reaction times increased and compared to the silent condition reaction times were slower. It seems that if the task becomes more complex, having additional auditory icons inhibits the categorization.

[9] Parts of this chapter are submitted for publication as: Lemmens, P.M.C., Bussemakers, M.P., & De Haan, A. The effects of auditory icons and earcons on visual categorization: the bigger picture. Submitted to proceedings of the 2001 International Conference on Auditory Display.

Introduction

We may not always notice it, but we often use auditory information in our everyday life to provide us with valuable information on what is going on around us. For example when attempting to cross a street, we listen to a car approaching behind us to determine how fast that car is going, so that we can make sure that we can safely cross (Buxton, 1989). When working with machines sounds can also help us. For instance when drilling a hole in a piece of wood, we listen for a change in the sound of the drill to determine whether we have passed through the material or not.

These are examples of real-life sounds that are used when our eyes cannot provide us with necessary information. It is obvious that sound can assist us in those cases. However, what happens if the sound is present in addition to ample visual information? Is it still useful to have the extra information in another modality, or does it hinder us in the tasks that we are trying to accomplish?

In the study presented here, the effect is determined of a concrete, categorical, real-life sound, an auditory icon (e.g. Gaver, 1989; Mynatt, 1994), as additional auditory information in a visual categorization task. Auditory icons are based on the principle of everyday listening (Vanderveer, 1979, Gaver 1993a, 1993b). This principle states that we do not attach a meaning to a sound based on musical qualities like the timbre or pitch for example, but based on the relationship between that sound and an event or object in the real world. An example of an auditory icon is the sound of a door closing. Auditory icons can be very useful, because users don't have to learn the meaning of the sound, but designers should take caution when applying them in user interfaces. After hearing the same auditory icon for a while, users find them annoying (Roberts & Sikora, 1997), most likely because the metaphor of the sound and its function in the interface is not an exact match (for example the sound of a closing door when closing a file).

Besides auditory icons, also abstract, conceptual sounds, namely earcons (e.g. Blattner, Sumikawa & Greenberg, 1989) can be used as auditory information. Earlier studies have shown that having earcons in the same categorization task leads to longer reaction times when compared to a situation with just the visual information (e.g. Bussemakers & de Haan, 1998; Bussemakers, de Haan & Lemmens, 1999).

A previous experiment with auditory icons has shown that, instead of leading to longer reaction times, also presenting the categorical auditory information leads to shorter reaction times when compared to a situation with no auditory information. It seems that there is an alerting effect of the auditory stimulus to the visual stimulus. Furthermore if the information is of the same category as the visual information, subjects benefit from having the extra information as is shown in a further facilitation effect.

The effect of the sounds on the reaction times may suggest that auditory icons lead to an improvement, but it is also important to take error rates into account. What if the reaction times are shorter, but subjects also make more errors?

In certain situations where a low number of errors is critical, it can be important to also know the effect of the concrete sounds on error rates. Since the error rates in a single task visual categorization experiment are low, a dual task paradigm is used in the experiment described in this chapter. Besides the categorization users have a secondary task where they have to mentally add single-digit numbers that are presented with the pictures. Per category the subject needs to remember the sum of the numbers on the screen. A task like the addition takes up extra resources of the working memory. Researchers like McLeod and Posner (1984) believe that there is a pool of resources available to everyone. Each task takes up a part of the resources available and if a lot of the capacity is used, this could lead to longer reaction times and higher error rates (see also Logie, 1996; Pashler & Johnston, 1998). If there are more errors, it might be possible to find out if there are differences in number of errors, for instance between the conditions in the experiment with auditory icons and the condition with just the visual information.

Experiment

Subjects
20 Students of the Catholic University of Nijmegen participated in the study. They were paid or received course credits for their time.

Materials
The experiment was carried out on a Macintosh Quadra 840 AV. The visual stimuli were presented on a 14-inch screen that was raised to eye-level. The auditory stimuli were presented via a pair of Monacor BH-004 headphones. The responses were registered through a button-box. The 14 line drawings of animals or musical instruments were used from previous tests (Bussemakers & de Haan, 2000) (see table 6.1). The pictures were modified to include a single-digit number in a position close to the center of the drawing (see figure 6.1).

Animal	Musical instrument
Duck	Flute
Donkey	Guitar
Cat	Harp
Chicken	Organ
Horse	Drum
Lion	Trumpet
Cow	Violin

Table 6.1 Visual stimuli.

Three sounds were used as auditory stimuli. As a representation of the 'animal' category the sound of a dog barking was presented. For the 'musical instrument' category the sound of a piano playing was used. As a neutral stimulus, the sound of water dripping was played. None of the sounds were represented in the pictures. The duration of the sounds was normalized to 1227 ms.

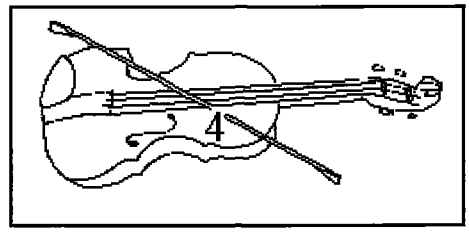


Figure 6.1 Picture of a violin with number 4.

Design

To allow the subject to take full advantage of the relationship between the picture and the sound, 6 conditions were defined (see table 6.2). Furthermore because of the similar setup, the results could be compared to earlier studies with earcons (see also Bussemakers & de Haan, 1998; Bussemakers, de Haan & Lemmens, 1999). In the congruent condition the pictures of animals were presented with the sound of a dog barking. With the pictures of musical instruments the sound of someone playing the piano was presented. Both the sound and the picture suggest the same response, because they are of the same category. In the incongruent condition the opposite was the case: with pictures of animals the subjects heard the piano sound and with pictures of musical instruments the sound of a dog barking. The visual and the auditory information suggest a different response. In the silent condition the pictures were presented with no additional auditory stimulus.

Condition	Picture	Auditory Icon
Congruent	Animal	Dog
	Musical instrument	Piano
Incongruent	Animal	Piano
	Musical instrument	Dog
Neutral	Animal	Dog
	Musical instrument	Dog
Neutral	Animal	Piano
	Musical instrument	Piano
Neutral	Animal	Water
	Musical instrument	Water
Silent	Animal	-
	Musical Instrument	-

Table 6.2 Experimental conditions.

There were three neutral conditions. In previous studies the neutral condition was defined as the condition where for both categories of visual stimuli the same auditory stimulus is presented. However in the case of auditory icons, the perceptual, categorical nature of the sound provides information not only on a condition-level, but also on a trial-level. It is possible that in the neutral condition with the sound of the dog, the pictures of animals benefit more from the additional auditory information than the pictures of the musical instruments. Therefore a third neutral condition was included with a sound that is not related to any of the categories.

To control for any order effects a digram-balanced Latin Square was used which ensured that every condition followed every other condition an equal amount of times (e.g. Wagenaar, 1969). Furthermore the position of the buttons was varied across subjects.

Procedure
Subjects participated in two tasks of which the order was controlled. The first task was a visual categorization task with additional auditory stimuli. In the second task, subjects were also instructed to cumulate per category the numbers that were presented in the pictures. Both tasks followed a similar procedure.

After the instructions were read the experiment was started. A trial was started with a single frequency tone (F4) that was presented together with a fixation-point ('+') for 500 ms in the center of the screen. Subjects were instructed to focus their eyes on the fixation-point. Then the fixation-point disappeared and after 1000 ms a sound was presented together with a picture, which was displayed for 300 ms. Subjects could then indicate their response to the question whether or not they saw a picture of an animal by pressing a button labeled 'yes' or a button labeled 'no'. A new trial was started after 2500 ms.

Results: Reaction Times (RT's)
In the statistical analyses error-responses and no-responses were excluded. The mean reaction times for both tasks are shown in figure 6.2.

A repeated measurements analysis showed that the conditions with the extra addition task were significantly slower than the conditions without the extra addition task ($F(1,19) = 15.624$, $p < .001$).

Within the non-addition task there was no significant difference in mean reaction times between the conditions with sound and the silent condition ($F(1,19) = 1.385$, $p > .1$). However when comparing the results in the congruent condition together with the incongruent condition to the results in the silent condition, the first two are significantly faster ($F(1,19) = 6.711$, $p < .05$). It seems that when the sounds are different for each category, the reaction times are faster than when there is no sound present.

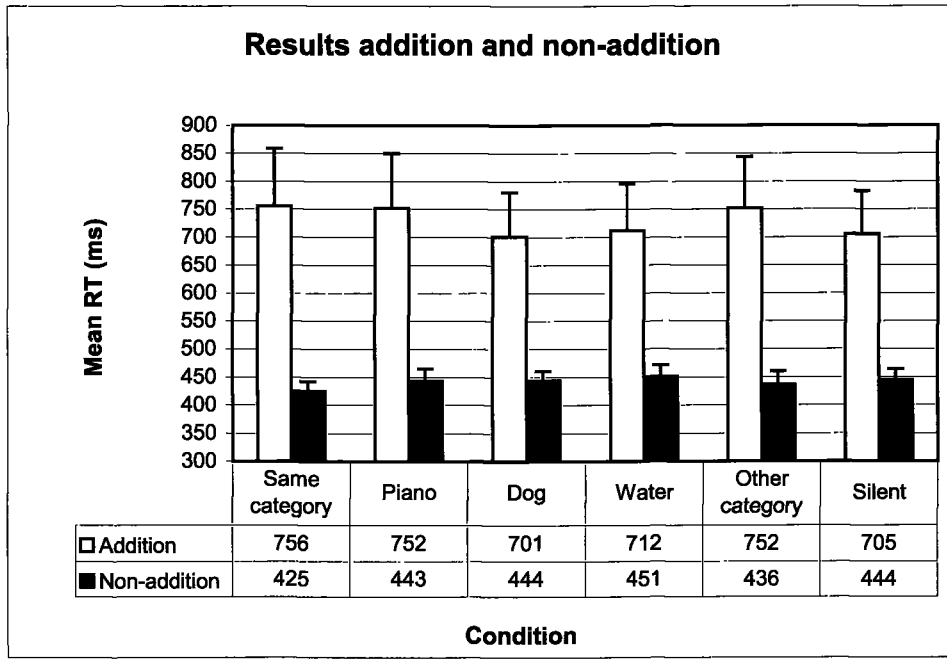


Figure 6.2 Mean reaction times in ms for the task with and without the extra addition.

The mean reaction times in the neutral conditions do not differ significantly from either the incongruent condition or the silent condition ($F(1,19) = 1.298$, $p > .1$ and $F(1,19) = 0.195$, $p > .5$). When comparing the neutral conditions to the congruent condition, the congruent condition is significantly faster ($F(1,19) = 9.180$, $p < .01$).

Within the dual task situation, the mean reaction times on the conditions with sound are significantly slower than the silent condition ($F(1,19) = 4.792$, $p < .05$).

Comparison between the mean reaction times on conditions in the dual task situation showed trends, but no significant differences.

Error analysis

Analyzing the errors showed that in 3.1% of the trials an error was made. Furthermore 53% of those errors were made in the addition-task and 47% of the errors were made in the non-addition task. In the table below the mean proportion of errors per task and per condition is displayed (see table 6.3). A repeated measurements analysis did not show a significant difference in mean proportion of errors between tasks ($F(1,19) = 0.479$, $p > .4$).

Within each task, none of the differences in proportion of errors between conditions were significant.

	Addition	SE	Non-addition	SE
Congruent	0.021	0.009	0.021	0.007
Piano	0.028	0.009	0.028	0.009
Dog	0.039	0.012	0.032	0.010
Water	0.025	0.011	0.018	0.007
Incongruent	0.025	0.009	0.018	0.007
Silent	0.025	0.005	0.025	0.004

Table 6.3. Mean proportion of errors and the standard error of the mean (SE) per task and condition.

Discussion

In this dual-task experiment the effect of concrete, categorical auditory information (auditory icons) was studied on a visual categorization task, when subjects at the same time had to add for each category single-digit numbers that were presented in the pictures.

Earlier studies (Bussemakers & de Haan, 2000) have shown that having auditory icons in a visual categorization task can lead to faster response times, if the sounds are different for each category. Especially when the auditory information is of the same category as the visual information, subjects respond fastest.

The results from the non-addition task in this study confirm these findings. The reaction times on the condition where the sound is of the same category as the pictures and the condition where the sound is of another category as the pictures are faster than the reaction times on the silent condition. It seems that having categorical auditory information, even if it is not the same as the visual information, leads to faster responses, because one stimulus seems to alert to the other. Between the neutral conditions and the silent condition there is no significant difference. Having the same auditory information with every picture within a block, does not seem to influence the response times, regardless of the category of the picture. Whether this information is related to one of the categories, like for instance in the case of the dog sound or the piano sound, or an entirely different sound like the water, does not seem to matter. It seems that the auditory information needs to be *different* per category to assist in the response to the visual information.

In the addition task there is no significant difference between the conditions with sound, but there is a significant difference between the conditions with sound and the silent condition. Having extra auditory information while having to add numbers seems to slow the response down, but it does not matter what kind of information is in the sound. Possibly just the fact that the auditory information needs to be processed and this interferes with the mental addition can explain the findings. Subjects need to first understand what the sound is before they can disregard it in the context of the task.

The error data shows that there is no difference in mean proportion of errors between conditions. It seems that the different types of additional auditory information do not lead to differences in error rates.

From these findings it can be concluded that having auditory icons in a visual categorization task leads to shorter reaction times when compared to a situation where there is no sound. However if there is a secondary cognitive task, the effect changes and reaction times are slower in the conditions with sound compared to the silent condition. It seems that in a more complex situation, having additional information in another modality needs to be processed, which leads to a decrease of the reaction times. Comparing these results to earlier findings on earcons, there is a clear difference. The reaction time data shows that contrary to auditory icons, earcons, both in a single-task and a dual-task setting, lead to longer reaction times.

These findings are restricted to categories of animals, non-animals and musical instruments. It seems interesting to find out if the results would be similar if other categories were used. Also, the categories that are used here are concrete and perceptual. It is possible that the results would be very different for abstract categories, for instance with a category like emotions.

Chapter 7

Conclusions and Future Research

Goal and method

The main goal of the research described in this dissertation was to gain more insight into the integration of multimodal information. In many predominantly visual tasks, like for example computer games, auditory information is added without experimentally validating the effect of the additional, redundant information. In the experiments described here, auditory icons and earcons are presented as redundant auditory information in a visual categorization task, because these types of sounds are most commonly implemented to provide information in human-computer interaction. Earcons provide abstract, conceptual information and auditory icons provide concrete, perceptual information.

The experiments were conducted using two paradigms. In one series of experiments, earcons were tested in a Simon paradigm (e.g. Simon, 1990), where the assumed positive or negative connotation of the sound was related to an affirmative or negative *response* to the question whether or not the pictures were of an animal. The accessory auditory information is independent of the visual information and makes it possible to test whether this leads to interference or facilitation at a response level.

In another type of experiments involving the Stroop paradigm (Stroop, 1935), auditory icons were related to the *categories* to which the pictures belonged. Subjects again had to determine for every picture whether or not it was a picture of an animal. This paradigm enables us to test which perceptually based attributes of the auditory stimulus have an influence on the central component of the processing of both visual and auditory information.

Both paradigms were used in a single-task situation where subjects just had to categorize the pictures and in a dual task situation, where subjects also had to remember the cumulative sum per category of single-digit numbers displayed in the center of the pictures.

Conclusions and Theoretical Framework

The influence of the auditory icons and earcons within the two paradigms is different. The abstract information of the earcons has an effect that is greatest when the subjects are presented with the trials in constant blocks. Since the information in the sound is related to the response that needs to be given, the impact is best observed if the information is constant across a group of trials. Within the Stroop task, the information has an effect on a trial-by-trial basis and it is the level of matching that is of influence here. When the sound is not only of the same category as the picture, but actually the same object (for instance a picture of a violin and the sound of a violin playing), the influence is greater, than when it is only of the same category (for instance a picture of a violin and the sound of a trumpet).

The nature of the influences of both types of sound is very different however. Auditory icons have a facilitating effect on the reaction times, where earcons have a inhibitory effect, as described below.

Earcons
The experimental results show that within our Simon paradigm, unlike in other Simon experiments where a facilitation of a more concrete aspect of the irrelevant stimulus is observed, it takes longer to integrate the conceptual auditory information, i.e. the earcons, with the visual information, when compared to a unimodal visual situation. Apparently, having an abstract sound that is not related to the visual information, but to the response, leads to a separate perceptual set, that needs to be processed. More cognitive resources seem to be needed to process two perceptual sets, instead of a single set and therefore in those trials a delay can be observed.

The integration can be seen as follows. Both stimuli are perceived and analyzed. It is not until both the task-related and the response-related information are analyzed separately, that the two are integrated. Therefore the overall response times for the trials where the additional auditory information is present (multimodal trials) are slower than the response times in the trials where only the visual information is present.

When the response-related information and the task-related information are integrated, a further delay is observed, if the response-related information is contradicting the task-related information. This is mentioned as a negative Simon effect. If a subject hears a sound with a negative connotation, but sees a picture that is part of the category ('yes'), it takes longer to come to a response, than when both the response-related information and the task-related information suggest the same response.

Within a design characterized by a complete randomization of trials the overall delay-effect of the earcons is greatest when the multimodal information is presented simultaneously. If, in such a design, the auditory information is presented before the visual information with an SOA of 500 ms, the delay-effect is no longer observed. It seems that the auditory and visual information is no longer perceived as related to each other and the multimodal information is no longer integrated.

If the trials are presented to the subject in constant blocks, where a certain earcon is always played with a category of the pictures (and thus is related to a certain response), the overall delay effect of the trials with sound as well as the negative Simon effect can still be observed at an SOA of 500 ms. It seems that when subjects are able to rely on the fact that the auditory information stays the same for a certain number of trials, both types of information are integrated, even if there is a time difference in presentation. It seems that contrary to a design with complete randomization, where the integration seems more automatic, here the integration is more intentional, so that differences in time of presentation are less of influence.

A further finding of the presented results is that the observed effects are independent of the musical experience of the subject. There is no difference in reaction times between subjects that are considered to be experienced with music and subjects that are less experienced with music.

Findings, studying the role of attentional demands and multimodal integration show that in a dual-task situation, where the cognitive load is greater than in a single-task experiment, again there is a modulating delay-effect. It seems that the effect, that in terms of reaction time differences is considered to be small, is so robust, that the increased load does not have an influence.

The modulating effect of a delay however is greater in a dual-task setting. Generally subjects unconsciously adopt one of two strategies. Either they increase the reaction times in order not to make any more mistakes, or they make more mistakes in order to be as fast as in a single-task setting. In the studies described in this dissertation a delay in reaction times is observed in the dual-task situation when compared to the single-task situation. The overall proportion of errors does not differ between the tasks, and within the tasks the proportion of errors is equally distributed over conditions.

Auditory Icons
Within the Stroop paradigm, the results show that there is a facilitation-effect when the categorical auditory information, i.e. the auditory icon, is presented with the visual information, when compared to a unimodal visual situation. This effect is observed even if the auditory information is incongruent with the visual information, meaning that both stimuli do not indicate the same response. It seems that if the sounds are different per category they lead to faster reaction times. From studies on non-semantic stimuli it is known that having information in two modalities can speed up the response, because of an alerting effect of one stimulus on the other (Nickerson (1973); Welch & Warren (1986)). Auditory information can be processed faster than visual information (Stein & Meredith, 1993), so it is possible that although both types of information are presented simultaneously, the auditory information alerted to the visual information, leading to an overall facilitation of the response. On the other hand it is also possible that the multimodal information leads to a spreading of activation, enabling a faster response.

The results of the experiments described in this dissertation furthermore suggest that there is an increased facilitation-effect if the same *semantic* information is presented in two modalities, for instance if the subject hears a dog barking and sees a picture of a dog as well. Apart from the alerting effect mentioned earlier (Nickerson (1973); Welch & Warren (1986), there seems to be a mechanism that leads to faster response times if the information in both modalities is congruent. Studies reported in literature suggest, that the intensities of both stimuli are summed, leading to a higher activation (Colavita & Weisberg (1979). Some suggest, that the same information in two modalities increases the salience of the stimulus, making it less ambiguous to the subject and therefore leading to shorter response times (Stein & Meredith, 1993).

It seems that in the studies described in this dissertation both the alerting effect of information in another modality has an effect, as well as the increased salience of the information.

The effect of the perceptual multimodal information seems to be greatest if the information is exactly the same in both modalities. If the information is of the same category, for example the subject sees a dog, but hears a cat, the overall facilitation-effect of the perceptual multimodal information is observed, but the increased facilitation-effect of the same semantic information disappears.

In a dual-task situation, when the multimodal trials are compared with the unimodal trials (with just the visual information), a delay-effect can be observed, similar to the experiments with earcons. It seems that the additional auditory information that needs to be processed interferes with the increased cognitive load of the dual-task setting and instead of making more errors, the subjects adopt the strategy to increase the reaction times.

Limitations and Future Research

In trying to find general mechanisms that can be applied in many practical situations, it is sometimes difficult to extrapolate from the experimental findings to more general 'guidelines'. The categorization tasks that were studied in the experiments described here have some relationship with practical human-computer interaction design in the sense that icons on a computer desktop can be viewed as belonging to different categories like files or programs. However in order to apply the results and conclusions from this dissertation, it is always wise to validate the design by testing it with users.

Furthermore, it would be interesting to broaden the kinds of visual categories that have been tested. This can be done in several ways. First of all, the picture categories in the described experiments were perceptual and categorical. It seems interesting to study more abstract, conceptual visual categories to see if the effects of both perceptual and abstract sounds on the categorization of these categories are similar to the effect on animals and musical instruments.

Secondly, there are other types of sounds that can be used as auditory information, like impact sounds, i.e. for instance the sound of a ball bouncing. The placement along the perceptual-concrete continuum of kinds of auditory information can assist designers in their choice for certain sounds. The impact sounds again as an example, seem to be rather perceptual, but in some cases lack the metaphor that is so often associated with auditory icons. It seems that in the metaphorical sense they can be considered as more abstract than auditory icons, but certainly more concrete than musical motives, since they have a reference in the real world. It would be interesting to see what their effect would be on performance as well as preference of subjects.

References

American Standards Association. (1960). *Acoustical Terminology* (No. S1.1). American Standards Association, New York.

Beltz, S.M., Winters, J.J., Robinson, G.S., & Casali, J.G. (1997). A methodology for selecting auditory icons for use in commercial motor vehicles. In *Proceedings of the Human Gactors and Ergonomics Society 41st Annual Meeting* (pp. 939-943). Danta Monica, CA: Human Factors and Ergonomics Society.

Beltz, S.M., Winters, J.J., Robinson, G.S., & Casali, J.G. (1998). Auditory icons: A new class of auditory warning signals for use in intelligent transportation subsystems. *Society of Automotive Engineers 1997 Transactions-Journal of Commercial Vehicles/Section 2, 106*, 431-439.

Beltz, S.M., Robnison, G.S., Casali, J.G. (1999). A new class of auditory warning signals for complex systems: Auditory icons. *Human Factors, 41* (4), 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(1),* 11-44.

Bower, G.H. (1981). Mood and memory. *American Psychologist, 36,* 129-148.

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge: The MIT Press.

Brewster, S.A. (1994). *Providing a Structured Method for Integrating Non-Speech Audio into Human-Computer Interfaces*. Unpublished doctoral dissertation, University of York, United Kingdom.

Brewster, S.A., Wright, P.C. & Edwards, A.D.N. (1994). The design and evaluation of an auditory-enhanced scrollbar. In B. Adelson, S. Dumais, & J. Olson (Eds.), *Proceedings of CHI'94*, Boston, Massachusetts: ACM Press, Addison-Wesley, pp. 173-179.

Brewster, S.A., Wright, P.C., Dix, A.J. & Edwards, A.D.N. (1995). The sonic enhancement of graphical buttons. In K. Nordby, P. Helmersen, D. Gilmore, & S. Arnesen (Ed.), *Proceedings of Interact'95*, Lillehammer, Norway: Chapman & Hall, pp. 43-48.

Brewster, S.A. (1997). Using non-speech sound to overcome information overload. *Displays, Special issue on multimedia displays, 17,* pp 179-189.

Brewster, S.A. (1998). Sonically-enhanced drag and drop, *in Proceedings of ICAD'98* (Glasgow, UK), British Computer Society.

Brewster, S.A. (1991). Sound in the Interface to a Mobile Computer, in *Proceedings of HCII'99* (Munich, Germany, August 1999), 43-47.

Broadbent, D.E. (1958). *Perception and Communication.* London: Pergamon Press.

Bussemakers, M.P., & de Haan, A. (1998). Using earcons and icons in categorisation tasks to improve multimedia interfaces, in *Proceedings of ICAD'98* (Glasgow, UK) British Computer Society.

Bussemakers, M.P., de Haan, A., & Lemmens, P.M.C. (1999). The effect of auditory accessory stimuli on picture categorisation; implications for interface design, in *Proceedings of HCII'99* (Munich, Germany, August 1999) 436-440.

Bussemakers, M.P., & de Haan, A. (2000). When it sounds like a duck and it looks like a dog... Auditory icons vs. earcons in multimedia environments. Proceedings of the International Conference on Auditory Display 2000. Atlanta, USA: International Community for Auditory Display. Pp. 184-189.

Bussemakers, M.P., & De Haan, A. Getting in touch with your moods: using sound in interfaces. Accepted for publication for Interacting with Computers.

Buxton, W. (1989). Introduction to this special issue on nonspeech audio. *Human-Computer Interaction, 4,* 1-9.

Colavita, F.B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception and Psychophysics, 25,* 345-347.

Cook, P.R. (1999). *Music, Cognition and Computerized Sound.* Cambridge: The MIT Press.

Crowder, R.G. (1984). Perception of the major/minor distinction: I. Historical and theoretical foundations. *Psychomusicology, 4(1-2),* 3-12.

Crowder, R.G. (1985a). Perception of the major/minor distinction: II. Experimental investigations. *Psychomusicology, 5(1-2),* 3-24.

Crowder, R.G. (1985b). Perception of the major/minor distinction: III. Hedonic, musical, and affective discriminations. *Bulletin of the Psychonomic Society, 23(4),* 314-316.

Crowder, R.G., Reznick, J.S., & Rosenkrantz, S.L. (1991). Perception of the major/minor distinction: V. Preferences among infants. *Bulletin of the Psychonomic Society, 29(3),* 187-188.

DeHouwer, J. (1998). The semantic Simon effect. *The Quarterly Journal of Experimental Psychology, 51A(3),* 683-688.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (Eds.) (1998) *Human-Computer Interaction.* London: Prentice Hall Europe.

Edworthy, J. (1998). Does sound help us to work better with machines? A commentary on Rauterberg's paper 'About the importance of auditory alarms during the operation of a plans simulator'. *Interacting with Computers, 10,* 401-409.

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

Fiedler, K., & Stroehm, W. (1986). What kind of mood influences what kind of memory; The role of arousal and information structure. *Memory and Cognition, 14(2),* 181-188.

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(1),* 67-94.

Gaver, W.W. (1993a). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology, 5* (1), 1-29.

Gaver W.W. (1993b). How do we hear in the world? Explorations in ecological acoustics. *Ecological Psychology, 5* (4), 285-313.

Gebhard, J.W., & Mowbray, G.H. (1959). On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology, 72,* 521-528

Gelman, S.A., & Markman, E.M. (1986). Categories and induction in young children. *Cognition, 23(3),* 183-209.

Gibson, J.J. (1966). *The senses considered as Perceptual Systems.* Boston: Houghton Mifflin Company.

Glaser, W.R., & Glaser, M.O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General, 118(1),* 13-42.

Goldstone, R.L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition, 52,* 125-157.

Heinlein, C.P. (1928). The affective characters of the major and minor modes in music. *Journal of Comparative Psychology, 8,* 101-142.

Hereford, J., & Winn, W. (1994). Non-speech sound in human-computer interaction: a review and design guidelines. *Journal of Educational Computing Research, 11(3),* 211-233.

Hevner, K. (1933). The mood effects of the major and minor modes in music, in Proceedings of the midwestern psychological association, 584.

Hofman, P.M., Van Riswick, J.G.A., & Van Opstal, A.J. (1998). Relearning sound localization with new ears. *Nature Neuroscience, 1(5),* 417-421.

Howard, I.O., & Templeton, W.B. (1966). *Human Spatial Orientation.* London: Wiley.

Isen, A.M., Shalker, T.E., Clark, M., & Karp, L. (1978). Affect, accessibility of material in memory, and behavior: A cognitive loop? *Journal of Personality and Social Psychology, 36,* 1-12.

Ishio, A. (1990). Auditory-visual Stroop interference in picture-word processing. *Japanese Journal of Psychology, 61,* 329-335.

Ishio, A. (1992). Picture categorizing processing in a cross-modal interference task. *Japanese Psychological Research, 4,* 117-125.

Jones, D. (1989). The sonic interface, in M. J. Smith and G. Salvendy (Eds.), *Work with Computers: Organizational, Management, Stress and Health Aspects (Amsterdam: Elsevier),* pp. 382-388.

Kahneman, D. (1973). *Attention and effort.* New York: Prentice Hall.

Kastner, M.P, & Crowder, R.G. (1990). Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Perception, 8(2),* 189-202.

Kennedy, J.M., & Ross, A.S. (1975). Outline picture perception by the Songe of Papua. *Perception, 4(4)*, 391-406.

Knowles, W.B. (1963). Operator loading tasks. *Human Factors, 5,* 151-161.

Koelsch, S., Schroeger, E., & Tervaniemi, M. (1999). Superior pre-attentive auditory processing in musicians. *NeuroReport, 10,* 1309-1313.

Kornblum, S. (1992). Dimensional overlap and dimensional relevance in stimulus-response and stimulus-stimulus compatibility. In G.E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior*, Vol. 2 (pp. 743-777). Amsterdam: North Holland.

LaHeij, W. (1988). Components of Stroop-like interference in picture naming. *Memory and Cognition, 16*, 400-410.

Lupker, S.J. (1976). The semantic nature of response competition in the picture-word interference task. *Memory and Cognition, 7(6)*, 485-495.MacLeod, C.M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109(2),* 163-203.

Logie, R.H. (1996). The Seven Ages of Working Memory, in *Working Memory and Cognition,* Richardson, J.T.E. (Eds.), 1996, 31-65. Oxford University Press, New York, USA.

MacLeod, C.M. (1991). Half a century of research on the Stroop effect: An integrative view. *Psychological Bulletin, 109(2)*, 163-203.

Mareschal, D., French, R.M., & Quinn, P.C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology, 36(5)*, 635-345.

Marr, D. (1982). *Vision: a Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco: W.H. Freeman and Company.

McClelland, J.L. & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159-188.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

McLeod, P., & Posner, M.I. (1984). Privileged loops from percept to act. In H. Bouma & D.G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes.* Hove: Lawrence Erlbaum Associates Ltd.

Mesulam, M.M. (1998). From sensation to cognition. *Brain, 121,* 1013-1052.

Mynatt, E.D. (1994). Designing with auditory icons: how well do we identify auditory cues? in *Proceedings of CHI'94* (Boston, US), 269-270.

Moore, B.C.J. (1989). *In Introduction to the Psychology of Hearing.* London: Academic Press Limited.

Murch, G.M. (1973). *Visual and Auditory Perception.* New York: The Bobbs-Merrill Company Inc.

Neisser, U. (1994). Multiple systems: A new approach to cognitive theory. *European Journal of Cognitive Psychology, 6(3),* 225-241.

Nickerson, R.S. (1973). Intersensory facilitation of reaction times: Energy summation or preparation enhancement? *Psychological Review, 80(6),* 489-509.

Pashler, H., Johnston, J.C. (1998). Attentional Limitations in Dual-task Performance, in *Attention*, Pashler, H. (Eds.), 155-189. Psychology Press: Hove, UK.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3-25.

Posner, M.I., & Boies, S.J. (1971). Components of attention. *Psychological Review, 78,* 391-408.

Preece, J. (1994). *Human-Computer Interaction.* New York: Addison-Wesley Publishing Company.

Rauterberg, M. (1998). About the importance of auditory alarms during the operation of a plant simulator. *Interacting with computers, 10,* 31-44.

Roberts, L.A., & Sikora, C.A. (1997). Optimizing feedback signals for multimedia devices: Earcons vs. auditory icons vs. speech. *Proceedings of the IEA'97, Tampere*, 224-226.

Robertson, I.H., Murre, J.M.J (1999). Rehabilitation of brain damage: Brain plasticity and principles of guided recovery. *Psychological Bulletin, 125(5)*, 544-575.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T.E. Moore (Ed.), *Cognitive development and the acquisition of language* (pp. 111-144). New York: Academic Press.

Rossing, T.D. (1989). *The science of sound.* New York: Addison-Wesley Publishing Company.

Schriefers, H., & Meyer, A.S. (1990). Experimental note: Cross-modal visual-auditory picture-word interference. *Bulletin of the Psychonomic Society, 28,* 418-420.

Schwartz, S.H. (1999). *Visual Perception.* Stamford: Appleton & Lange.

Siegel, J.A., & Siegel, W. (1997). Absolute identification of notes and intervals by musicians. *Perception and Psychophysics, 21(2),* 143-152.

Shanks, D.R. (1997). Representation of categories and concepts in memory. In M.A. Conway (Ed.), *Cognitive Models of Memory* (pp. 111-146). Hove: Psychology Press.

Shelton, J.R., & Martin, R.C. (1992). How semantic is semantic priming? *Journal of Experimental Psychology: Learning, Memory and Cognition, 18,* 1191-1210.

Sikora, C.A., Roberts, L.A., & Murray, L. (1995). Musical vs. real world feedback signals. *Proceedings of CHI'95, Denver*, 220-221.

Simon, J.R. (1990). The effects of an irrelevant directional cue on human information processing. In R.W. Proctor & T.G. Reeve (Eds.), *Stimulus-response compatibility: An integrated perspective* (pp. 31-86). Amsterdam: North Holland.

Simon, P. & Garfunkel, A. (1966). *The Sounds of Silence*. Columbia Records.

Smith, M.C., & Magee, L.E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General, 109,* 373-392.

Snyder, M., & White, P. (1982). Mood and memories: Elation, depression, and the remembering of the events on one's life. *Journal of Personality, 50,* 149-167.

Spence, C., & Driver, J. (1996). Audio-visual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1005-1030.

Stein, B.E., & Meredith, M.A. (1993). *The Merging of the Senses*. Massachusetts: The MIT Press.

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

Styles, E.A. (1997). *The Psychology of Attention*. Hove: Psychology Press, Ltd.

Vanderveer, N.J. (1979). Ecological acoustics: human perception of environmental sounds. *Dissertation Abstracts International*, 40/09B, 4543.

Van Galen, G.P. (1974). Ambient versus focal information processing and single-channelness. Doctoral dissertation.

Wagenaar, W.A. (1969). Note on the construction of digram-balanced Latin squares. *Psychological Bulletin, 72 (6)*, 384-386.

Welch, R.B., DuttonHurt, L.D., & Warren, D.H. (1986). Contributions of audition and vision to temporal rate perception. *Perception and Psychophysics, 39,* 294-300.

Welch, R.B., & Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638-667.

Whittlesea, B.W.A. (1987). Preservation of specific experiences in the representation of general knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 3-17.

Williams, J.N. (1996). Is automatic priming semantic? *European Journal of Cognitive Psychology, 8,* 113-161.

Williams, S.M. (1992). Perceptual Principles in Sound Grouping. In Kramer, G. (Ed.). *Auditory display, sonification, audification and auditory interfaces. The Proceedings of the First International Conference on Auditory Display* (pp. 96-125). Reading, Massachusetts: Santa Fé Institute, Addison-Wesley.

Windows interface guidelines for software design. Microsoft (4-12-98).

Zajonc, R.B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35(2)*, 151-175.

Summary

Nowadays most computers and computer games provide a true multimedia experience. In combining images and sound, the user is shown what the effect is of their manipulations. Generally, users report that they appreciate this multimodal information stream, but empirically the effect of presenting the same information in pictures and sound has never been determined (Edworthy, 1998).

This dissertation tries to provide more clarity on the issue, by attempting to answer the question: how are different types of visual and auditory information integrated?

To investigate this question a visual categorization task was used, where concrete and abstract sounds were presented as redundant information. In a categorization task, subjects are asked to determine for every stimulus, in this case a picture, if it is part of a certain category. For instance, for every picture the subject sees he or she needs to press a 'yes'-button or a 'no'-button answering the question: is this a picture of an animal? How fast a button is pressed is an indication of the time it took to come to a decision.

In the experiments described in this dissertation, apart from the picture also a sound was presented to the subjects. In some cases the subjects heard the sound of an animal or a musical instrument. These concrete, everyday sounds are also known as auditory icons (e.g. Gaver, 1989; Mynatt, 1994). In other cases, subjects heard major or minor chords. It is known that major chords have a positive connotation and minor chords are associated with something negative (e.g. Hevner, 1933; Crowder, 1984). These abstract, more conceptual sounds are called earcons (e.g. Blattner, Sumikawa & Greenberg, 1989).

The first chapter of the dissertation explains what processes play a part in the integration of information from two modalities. First of all, there are the auditory and visual stimuli by themselves that need to be processed by the visual and auditory systems. Furthermore, it depends on our attention how much we notice of our environment. If we are performing more than one task at a time or we are receiving information from multiple sources, the capacity of our attention is important, i.e. how much we can register and process at that time (Knowles, 1963; Kahneman, 1973).

But what happens if auditory and visual information is presented together? From earlier research it is known that if a stimulus is presented, accompanied by a redundant second stimulus, this leads to faster responses (Nickerson, 1973; Colavita & Weisberg; 1979; Welch & Warren, 1986; Stein & Meredith, 1993). It is assumed that this on the one hand is the result of higher attention levels: one concrete stimulus has an alerting, warning effect on the other stimulus. On the other hand presenting the same concrete information in two modalities causes the activation of that information in memory to be greater and thus the information to be more salient. The more salient the information is, the less ambiguous and the faster the subject can respond to the question.

These earlier studies are mostly based on stimuli without semantics, like a flash of light or auditory noise. In our experiments a more complex categorization task was used, where either more concrete sounds, like auditory icons, were presented as accessories, or more abstract, conceptual sounds like earcons.

In the second chapter a paradigm is mentioned that is used in the categorization experiments, namely the Simon paradigm (e.g. Simon, 1990). In the classical Simon experiment an irrelevant feature of a stimulus is combined with a relevant feature of the response and this irrelevant feature is independent of the relevant feature of the stimulus. In our studies the connotation of the earcon (positive in the case of major and negative with minor) is combined with the response to the categorization (yes or no).

A first experiment shows that with a complete randomization of trials an inhibitory effect occurs if the earcon is present, compared to trials were just the visual stimulus is available. Furthermore, a second experiment shows that this effect is greatest if the picture and the sound are presented at the same time. If the sound is presented 500 ms before the visual stimulus, the effect disappears. In a third experiment, apart from the inhibitory effect of the earcons, a negative Simon effect is revealed if the stimulus is presented grouped by connotation. If the sound has the opposite connotation of the intended response to the picture, for instance when the subject sees a cat and hears a minor chord, then the response is delayed more than when the subject sees a cat and hears a major chord, and the sound has the same connotation as the intended response.

In the third chapter these findings are confirmed and furthermore it is established that musical experience (actively playing an instrument for more than 6 years) has no statistically significant influence on the effects found.

Apart from the reaction times, errors are also an important source of information on what happens when information is integrated. During the experiments in chapter 2 and 3 few errors were made, probably as a result of the simplicity of the task. To complicate the task a dual-task experiment was conducted, that is described in chapter 4. Apart from the categorization task, subjects had to remember per category the sum of digits that were presented in every picture. The results show that there are longer response times than in earlier experiments and that there are more errors. The proportion of errors however is divided evenly across the conditions. Furthermore, a delay effect of the earcons is again observed compared to trials where only the picture is presented.

In the fifth chapter a second paradigm is mentioned, that is used in categorization experiments, namely the Stroop paradigm (Stroop, 1935; MacLeod, 1991). Stroop wanted to study attention and interference by testing the effect of different aspects of a compound stimulus on naming another aspect of the stimulus. In the traditional task, subjects for instance had to name the ink color of written names of colors.

In the experiments in chapter 5, analogous to this idea, the effect is studied of concrete auditory information (auditory icons), in this case sounds of animals and sounds of musical instruments, on the categorization of pictures of the same categories. What happens for instance if a subject has to indicate for a picture of a dog, whether or not it is an animal, while at the same time hearing the whinnying of a horse? Just as with the earcons two kinds of effects are found, but the nature of the effects is quite different from those with the earcons. First of all, subjects respond faster in the multimodal trials compared to trials where just the picture is presented. Apparently it does not matter what kind of auditory icon is presented with the picture. The reaction times are always faster if there is a sound present. This is similar to the previously mentioned idea that one stimulus can have an alerting effect on another stimulus and as such can lead to a higher attention level. Furthermore subjects respond fastest if the picture and the sound are exactly the same, for instance if the picture of a cat is presented with the sound of a meowing cat. It seems that how much the auditory and visual information represent the same instance is important. The more the multimodal information is congruent, the faster the reaction times are.

Comparable to the Simon paradigm and the earcons, in chapter 6 a dual task is described, where subjects again apart from the categorization have to remember the sum of numbers that are presented in the pictures. The auditory icons that are used are related to the categories of the pictures, but are not visually represented. For the category animal for example, a barking dog is used as an auditory icon, but there is no visual stimulus of a dog. The results indicate that in the dual task, contrary to the experiment with the single task, subjects respond slower in the multimodal trials than in the trials with just the pictures. This larger amount of information, whether it is of the same category as the pictures or not, simply takes more time. The number of errors is larger than in the experiment of chapter 5, but similar to the dual task experiment with the earcons, the errors are equally divided across conditions.

General conclusions

The experiments described here show that there is a different effect of concrete and abstract auditory information in a visual categorization task. The concrete auditory icons in the single task lead to facilitation of the reaction times, where the abstract earcons lead to inhibition of the reaction times. Furthermore, the extent of the correspondence between the information in the sounds and the visual information is important. If the abstract earcons do not correspond with the intended response, the response is delayed further. For the concrete auditory icons the opposite holds: if the concrete sounds are an exact match of the visual stimuli, then the response is facilitated further. Finally, the results indicate that in a dual task, where apart from the categorization there is a higher cognitive load by also having to remember the sum of a series of numbers, the addition of both concrete or abstract sounds leads to a delay in the response times.

What do these results mean for the development of multimodal interfaces? First of all, the results show that care should be taken when adding redundant information in another modality. There is not always a positive effect in terms of productivity (response times). If auditory signals need to be added, the results of the experiments seem to suggest that for a concrete situation concrete sounds are recommended. However, subjective measures like preference and annoyance of subjects were not measured and should be studied. Furthermore, the results show that if the task is more complex, which is most often the case in reality, the situation can be quite different from a simple task. Both with concrete and abstract sounds the presentation with visual information leads to a delay in response times. In a time-critical, complex situation it seems that multimodality would not be the best option.

Samenvatting

De meeste computers en computerspelletjes bieden tegenwoordig een ware multimedia-ervaring. Door combinaties van beeld en geluid wordt aan gebruikers getoond wat het effect is van hun handelingen. Over het algemeen geven gebruikers aan, dat zij deze multimodale informatiestroom waarderen, maar het effect van het aanbieden van dezelfde informatie in beeld en geluid is nooit empirisch vastgesteld (Edworthy, 1998).

Dit proefschrift probeert hier meer duidelijkheid over te verschaffen door de specifieke vraag te beantwoorden: hoe worden verschillende vormen van visuele en auditieve informatie geïntegreerd?

Om deze vraag te onderzoeken werd gebruik gemaakt van een visuele categorisatietaak, waarbij concrete en abstracte geluiden als redundante informatie werden aangeboden. In een categorisatietaak krijgen proefpersonen de opdracht om voor elke stimulus, in dit geval een plaatje, te bepalen of het tot een bepaalde categorie behoort. Bijvoorbeeld, voor elk plaatje dat de proefpersoon ziet, moet hij of zij op een 'ja'-knop of een 'nee'-knop drukken als antwoord op de vraag: is dit een plaatje van een dier? Hoe snel er gedrukt wordt is vervolgens een indicatie van de tijd die het duurde om tot een besluit te komen.

In de experimenten die in dit proefschrift beschreven worden, werd naast het plaatje ook nog een geluid aan de proefpersonen aangeboden. In sommige gevallen kregen de proefpersonen het geluid van een dier of een muziekinstrument te horen. Deze concrete, alledaagse geluiden worden ook wel auditory icons genoemd (e.g. Gaver, 1989; Mynatt, 1994). In andere gevallen kregen de proefpersonen majeur- of mineur-akkoorden te horen. Van majeur-akkoorden is bekend dat ze een positieve connotatie hebben en van mineur-akkoorden weten we dat ze met iets negatiefs geassocieerd worden (e.g. Hevner, 1933; Crowder, 1984). Deze abstracte, meer conceptuele geluiden worden earcons genoemd (e.g. Blattner, Sumikawa & Greenberg, 1989).

In het eerste hoofdstuk van het proefschrift wordt uitgelegd welke processen een rol spelen bij het integreren van informatie van twee modaliteiten. Allereerst zijn er natuurlijk de auditieve en visuele stimuli op zich die verwerkt moeten worden door het visuele en auditieve systeem. Daarnaast hangt het af van onze attentie hoeveel we van onze omgeving opmerken. Op het moment dat we meerdere dingen tegelijk aan het doen zijn of vanuit meerdere bronnen informatie ontvangen, dan is de capaciteit van onze attentie, hoeveel we op dat moment kunnen registreren en verwerken, daarbij van belang (Knowles, 1963; Kahneman, 1973).

Maar wat gebeurt er als de auditieve en visuele informatie gezamelijk aangeboden wordt? Uit eerder onderzoek is gebleken dat als een stimulus gepresenteerd wordt, vergezeld van een redundante tweede stimulus, dit leidt tot snellere reacties (Nickerson, 1973; Colavita & Weisberg, 1979; Welch & Warren, 1986; Stein &

Meredith, 1993). Aangenomen wordt dat dit aan de ene kant komt door een verhoogde attentie: de ene concrete stimulus heeft een alarmerend, waarschuwend effect voor de andere stimulus. Aan de andere kant zorgt het aanbieden van dezelfde concrete informatie in twee modaliteiten dat de activatie van die informatie in het geheugen groter wordt en dus de informatie saillanter. Hoe saillanter de informatie is, des te minder ambigu en des te sneller kan de proefpersoon reageren op de vraag.

Deze eerdere studies zijn voornamelijk gebaseerd op stimuli zonder semantiek, zoals een lichtflits of auditieve ruis. In onze experimenten werd een meer complexe categorisatietaak gebruikt, waarbij ofwel meer concrete geluiden in de vorm van auditory icons als accessories aangeboden werden, ofwel meer abstracte, conceptuele geluiden in de vorm van earcons.

In het tweede hoofdstuk komt een paradigma aan de orde, dat gebruikt wordt binnen de categorisatie-experimenten, namelijk het Simon-paradigma (e.g. Simon, 1990). In het klassieke Simon-experiment wordt een irrelevant kenmerk van een stimulus gekoppeld aan een relevant kenmerk van een respons en dit irrelevante kenmerk is onafhankelijk van het relevante kenmerk van de stimulus. In onze studie wordt de connotatie van het earcon (positief in het geval van majeur en negatief bij mineur) gekoppeld aan de respons op de categorisatie (ja of nee).

Een eerste experiment laat zien dat er bij een complete randomisatie van de aanbiedingen een vertragend effect optreedt als het earcon aanwezig is, ten opzichte van de aanbiedingen waarbij alleen de visuele stimulus beschikbaar is. Verder wordt in een tweede experiment aangetoond dat dit effect het grootst is als het plaatje en het geluidje tegelijk worden gepresenteerd. Op het moment dat het geluid 500 ms eerder wordt aangeboden dan de visuele stimulus, verdwijnt het effect. In een derde experiment komt naast het vertragende effect van de earcons een negatief Simon-effect naar voren, als de stimuli gegroepeerd naar connotatie worden aangeboden. Als het geluidje de tegenovergestelde connotatie heeft van de beoogde respons op het plaatje, bijvoorbeeld als de proefpersoon een kat ziet en een mineur-akkoord hoort, dan is de reactie meer vertraagd dan wanneer de proefpersoon een kat ziet en een majeur-akkoord hoort en dus het geluidje dezelfde connotatie heeft als de beoogde respons.

In het derde hoofdstuk worden deze bevindingen nog eens bevestigd en wordt verder vastgesteld dat muzikale ervaring (het meer dan 6 jaar actief bespelen van een instrument) geen statistisch significante invloed heeft op de gevonden effecten.

Naast reactietijden vormen fouten ook een belangrijke bron van informatie over wat er gebeurt bij het integreren van informatie. Tijdens de experimenten in hoofdstuk 2 en 3 werden weinig fouten gemaakt, waarschijnlijk ten gevolge van de eenvoudigheid van de taak. Om te taak moeilijker te maken werd een dubbeltaak-experiment gedaan, dat wordt beschreven in hoofdstuk 4. Naast de categorisatie-taak moeten proefpersonen per categorie de som onthouden van cijfers die in elk plaatje worden weergegeven.

De resultaten laten zien dat er langzamer wordt gereageerd dan in eerdere experimenten en dat er meer fouten worden gemaakt. De aantallen fouten zijn echter gelijkmatig verdeeld over de condities. Verder wordt er wederom een vertragend effect vastgesteld van de earcons ten opzichte van de aanbiedingen waar alleen het plaatje wordt getoond.

In het vijfde hoofdstuk komt het tweede paradigma aan bod, dat gebruikt wordt in categorisatie-experimenten, het Stroop-paradigma (Stroop, 1935; MacLeod, 1991). Stroop wilde attentie en interferentie onderzoeken door te testen wat het effect zou zijn van verschillende aspecten van een samengestelde stimulus op het benoemen van een ander aspect van de stimulus. In de traditionele taak moesten proefpersonen bijvoorbeeld de kleur van de inkt van geschreven namen van kleuren benoemen. In de experimenten in hoofdstuk 5 wordt analoog aan dit idee het effect onderzocht van concrete auditieve informatie (auditory icons) in de vorm van dierengeluiden en geluiden van muziekinstrumenten op de categorisatie van plaatjes van dezelfde categorieën. Wat gebeurt er bijvoorbeeld als een proefpersoon bij een plaatje van een hond moet aangeven of het een dier is, maar tegelijkertijd het hinniken van een paard hoort? Net als bij de earcons worden twee soorten effecten gevonden, maar de aard van de effecten verschilt nogal van die bij de earcons. Allereerst reageren proefpersonen sneller in de multimodale aanbiedingen vergeleken met de aanbiedingen waar alleen het plaatje wordt aangeboden. Het doet er blijkbaar niet toe wat voor auditory icon bij het plaatje wordt aangeboden. De reactietijden zijn altijd sneller als er geluid aanwezig is. Dit komt overeen met de eerder genoemde gedachte dat de ene stimulus een alarmerende werking kan hebben op de andere stimulus en zo de attentie kan verhogen. Verder reageren proefpersonen het snelst als het plaatje en het geluidje precies overeenkomen, bijvoorbeeld als het plaatje van een kat gepresenteerd wordt met het geluid van een miauwende kat. Blijkbaar is de mate waarin de auditieve en visuele informatie hetzelfde representeren ook van belang. Hoe meer de multimodale informatie overeenkomt, hoe sneller de reactietijden.

Vergelijkbaar met het Simon-paradigma en de earcons, wordt in hoofdstuk 6 een dubbeltaak beschreven, waarbij proefpersonen wederom naast de categorisatie de som moeten onthouden van getallen die in de plaatjes worden weergegeven. De auditory icons die hierbij gebruikt worden zijn gerelateerd aan de categorieen van plaatjes maar worden niet als zodanig visueel gerepresenteerd. Voor de categorie dier wordt bijvoorbeeld een blaffende hond gebruikt als auditory icon, maar er is geen visuele stimulus van een hond. De resultaten geven aan dat nu, in tegenstelling tot het experiment met de enkelvoudige taak, proefpersonen in de multimodale aanbiedingen langzamer reageren dan in de aanbiedingen met alleen de plaatjes. Dit is waarschijnlijk het geval omdat er door het tellen meer informatie verwerkt moet worden. Deze grotere hoeveelheid informatie, ongeacht of het overeenkomt met de categorie van de plaatjes of niet, kost meer tijd. Het aantal fouten is groter dan in het experiment uit hoofdstuk 5, maar net als bij de dubbeltaak met de earcons gelijk over de verschillende condities verdeeld.

Algemene conclusies

De experimenten die hier beschreven werden laten zien dat er een verschillend effect is van concrete en abstracte auditieve informatie op een visuele categorisatie-taak. De concrete auditory icons leiden in de enkelvoudige taak tot een facilitatie van de reactietijden, terwijl de abstracte earcons een inhibitie van de reactietijden tot gevolg hebben. Verder is de mate waarin de informatie in de geluiden overeenkomt met de visuele informatie ook van belang. Als de abstracte earcons niet overeenkomen met de beoogde respons, dan wordt de reactie verder vertraagd. Voor de concrete auditory icons geldt het tegenovergestelde: als de concrete geluiden precies overeenkomen met de visuele stimuli dan wordt de reactie verder versneld. Als laatste laten de resultaten zien dat bij een dubbeltaak, waarbij naast de categorisatie een extra cognitieve belasting optreedt door het moeten onthouden van de som van een reeks getallen, de toevoeging van zowel concreet als abstract geluid leidt tot een vertraging van de reactietijden.

Wat betekenen deze resultaten echter voor de ontwikkeling van multimodale interfaces? Allereerst laten de gegevens zien dat voorzichtig omgesprongen dient te worden met het toevoegen van redundante informatie in een andere modaliteit. Het is zeker niet zo, dat hiermee altijd een positief effect verkregen wordt in termen van productiviteit (responstijden). Als toch auditieve signalen toegevoegd worden, lijken de uitkomsten van dit onderzoek voor concrete situaties ook meer concrete geluiden aan te bevelen, echter subjectieve maten als preferentie en irritatie bij de gebruikers zijn hier niet in meegenomen en zouden verder onderzocht moeten worden. Verder tonen deze resultaten aan, dat als de taak meer complex is, hetgeen in de praktijk meestal het geval is, de situatie heel anders kan liggen dan wanneer de taak eenvoudig is. Zowel bij concrete als abstracte geluiden, leidt de aanbieding naast de visuele informatie in een complexe taak tot een vertraging van de respons. In een tijdcritische complexe situatie lijkt multimodaliteit niet de voorkeur te genieten.

Personal Publications

Bussemakers, M.P., and de Haan, A. (1998). Using earcons and icons in categorisation tasks to improve multimedia interfaces. *Proceedings of the International Community for Auditory Display* (pp. 152-157). Glasgow (UK): British Computer Society.

Hoffman, M.S, Cohen, S.M., Bussemakers, M.P., & de Gruil, R. (1998). Influence of organizational culture in Europe on the perceptions of the role of POS technology. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 954-958). Chicago (USA): Human Factors and Ergonomics Society.

Bussemakers, M.P, de Haan, A. (1999). Guidelines for using non-speech sounds in human-computer interaction. NCR white paper.

Bussemakers, M.P., de Haan, A., and Lemmens, P.M.C. (1999). The effect of auditory accessory stimuli on picture categorisation; implications for interface design. *Proceedings of Human Computer Interaction International* (pp.436-440). Munich (Germany): Lawrence Erlbaum Associates.

Bussemakers, M.P., & de Haan, A. (2000). When it sounds like a duck and it looks like a dog... Auditory icons vs. earcons in multimedia environments. *Proceedings of the International Conference on Auditory Display* (pp.184-189). Atlanta (USA): International Community for Auditory Display.

Lemmens, P.M.C., Bussemakers, M.P., & de Haan, A. (2000). The effect of earcons on reaction times and error-rates in a dual-task vs. a single-task experiment. *Proceedings of the International Conference on Auditory Display* (pp.177-183). Atlanta (USA): International Community for Auditory Display.

Bussemakers, M.P., & Psihogios, J. (2000). Cultural, psychological and ergonomic considerations involving changes in the workplace. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual meeting.* San Diego (USA): Human Factors and Ergonomics Society.

Bussemakers, M.P. (2000). Single, multiple or complex scanner tones. NCR white paper. (Patent pending.)

Bussemakers, M.P., & De Haan, A. (in press). Getting in touch with your moods: using sound in interfaces. *Interacting with Computers.*

Lemmens, P.M.C., Bussemakers, M.P., & De Haan, A. The effects of auditory icons and earcons on visual categorization: the bigger picture. Accepted for the proceedings of the 2001 International Conference on Auditory Display (July 2001).

Bussemakers, M.P. & de Haan, A. Auditory Icons and Earcons: Categorical and Conceptual Multimodal Interaction. Accepted for the proceedings of the 1st International Conference on Universal Access in Human Computer Interaction (August, 2001).

Curriculum Vitae

Myra van Esch-Bussemakers was born in Nijmegen on February 27th, 1973. In 1991 she graduated from the Elshof College[10] in Nijmegen, completing her pre-university studies (VWO). That same year, she started her studies of Psychology at the Catholic University of Nijmegen (Katholieke Universiteit Nijmegen), where, after finishing her first 'propedeuse' year, she decided to graduate in Cognitive Science, specializing in Cognitive Ergonomics and Cognitive Psychological Research. Her thesis was called: 'Als je begrijpt wat ik bedoel. Interactie in een multimodale omgeving (If you know what I mean. Interaction in a multimodal environment)'. Even before reaching her Masters in 1996 she was approached by NCR, an Atlanta-based multinational that develops products for retail and banking environments, like for instance product-scanners and ATM-machines. NCR asked Myra to work after her graduation for part of her time as a consultant for NCR's Human Factors Engineering Department within Europe, the Middle East and Africa. In that role she has worked with many retailers throughout the region. For the other part of her time Myra worked as a researcher at the Nijmegen Institute for Cognition and Information (NICI). This dissertation is the result of the research that was set up as a joint effort between the NICI and NCR. From the 1st of June 2001 Myra accepted a new position within the Usability Engineering Group at TNO in Soesterberg.

[10] In terms of this dissertation the author's college performance of "The Sounds of Silence" (Simon & Garfunkel, 1966) is to be categorized as a contradictory non-inhibited multimodal integration of both auditory and visual stimuli in a dual-task experiment.