When it Sounds like a Duck and it Looks like a Dog… Auditory icons vs. Earcons in Multimedia Environments

Myra P. Bussemakers, Ab de Haan
Nijmegen Institute for Cognition and Information
University of Nijmegen
Montessorilaan 3,
6525 HR Nijmegen, The Netherlands
+31 24 3612648
bussemakers@nici.kun.nl

ABSTRACT
In this research a categorization paradigm is used to study the multimodal integration processes that take place when working with an interface. Redundant auditory icons are used with visual information to investigate their influence on the categorization. Results are related to earlier experiments with redundant earcons and show that reaction times are faster in conditions with auditory icons than conditions with no sound, and conditions with earcons are slower than conditions with no sound. Having sound does not always lead to faster responses. It seems that the type of sound and its congruency with visual information can have an effect.

Keywords
Multimodal integration, auditory icons, earcons, categorization

INTRODUCTION
When attempting to optimize the interaction between a user and his/her computer, (interface) designers have a variety of modalities they can apply. Information can be presented through vision, touch, sound, or any combination of these sensory modalities. An interface designer has the difficult task of choosing the best combination of information streams for a particular situation or function. 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 cannot perform as well without having the audio on, although this perception has never been scientifically validated [6].

In most multimedia applications, both sound and visual information are used to convey the message. Vision, in these implementations, can be combined with different types of sound, like speech, real-life sounds (auditory icons, e.g. [7]) or abstract musical sounds (earcons, e.g. [1]). Each of these types of sound have advantages and disadvantages. The benefit of speech for instance, is that, most of the time, the meaning of the message is relatively unambiguous. However, users need to listen to the whole message to understand the meaning. Real-life sounds on the other hand, have the ability to convey complex messages in a single sound, provided that the sound is easy to identify and it is a good conceptual mapping of the function [9]. Yet users report them to be annoying after prolonged use [11,13]. Earcons do not posses that intuitive mapping and therefore have to be learned, but users find them in general appropriate for applications [11]. In interfaces they have furthermore shown to 'steer the emotional reaction of the user in support of a certain response' [1].

Most studies that look at sound in interfaces, use sound as a substitute for visual information that is incomplete or unavailable. It seems interesting to see if the same effects are found when the information is redundant, i.e. also available in the other modality. When information is presented both visually and auditory, i.e. in multiple streams, users need to integrate these informational elements into a single experience unity [4]. Like in the real world, when you see a colored light in the sky and you hear a bang for example, the integrated information suggests that there are fireworks. Both instances happen (almost) at the same time and therefore you assume that they are related. The integration is expected to take place at different levels of abstraction, where one is more concrete, based on examples of similar experiences, and one is more abstract, based on general rules, that are formulated from those examples. This abstract, conceptual level can only be formed when there is an example-level. It is possible that this integration in some functional situations is better, in this case meaning faster and with fewer errors, than in other situations. To what extent the information from different modalities can and will be integrated, could depend on the combination of relevant aspects of contingent information in the signals, such as for example color, location, loudness and mood [5]. When you see the same fireworks for instance in front of you, and you hear a bang behind you, the location of these instances is so different that you assume they are not coming from the same source. As a result, both information streams are most probably not integrated, but treated as separate. Mood as another example, is often manipulated in cinematography, where the dramatic, high tempo music combined with pictures of a chase can lead to great excitement for the viewer. Without the music, or with a calm, happy tune, the perception of the same scene can be quite different.

The concept of integration, with the aspects mentioned before, are investigated in our project, of which a study is reported here. In this study auditory icons are tested, presented together with visual pictures in a visual categorization task, to investigate whether the integration that is assumed to take place, can assist users in their task. Having information with the same contingencies available may lead to faster response times and fewer errors, when users have to decide whether the picture they see is of a certain class or not. This type of task is used, because of its analogy with an interface situation. Icons on a desktop belong to different categories of documents, folders or applications and users have to determine what the category is in order to decide what to do with the icon (for instance drag to an application in the case of a document). (the experimental conditions and stimuli will be presented at the conference)

The result of the experiment presented here, related to earlier experiments with earcons, will contribute to the development of a general theoretical model of multimodal integration. This theory can assist developers in defining functions and situations in visually oriented environments that are suitable for auditory additions and may help decide what types of sound to use.

EARLIER EXPERIMENTS

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. [7, 2, 10]). 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, i.e. extra, not present in the visual modality, regardless of the auditory nature of that information, could account for some of the effects reported [6].

In our studies, we try to avoid using complementary information in order to investigate the effect of different types of sound on the same task. The information that is present in the sound signal is therefore redundant with information presented in the visual modality. Users should be (and are) able to perform the task without having the sound present. To test that in this experimental paradigm, control trials are present with just the visual information available. The task users have to perform is a visual categorization task. Pictures, i.e. line drawings of animals (for instance a dog) and non-animals (for instance a candle), are presented to subjects on a computer screen. Subjects have to determine whether or not the picture they see is of an animal or not by pressing a button labeled ‘yes’ or ‘no’.

In our earlier experiments, the pictures were accompanied by an earcon or by silence. However, subjects were not instructed to use the sound when it was present in the task they had to perform. Four different earcons were used (differing in pitch), two major chords (C) and two minor earcons (Cm). Major key tones are generally associated with positive feelings, like happiness, and minor key tones are associated with negative emotions, like sadness (e.g. [8]). Other experiments have shown that subjects, when asked to press a button labeled ‘yes’ or ‘no’ on the auditory stimuli alone, more often respond ‘yes’ on major key tones and ‘no’ on minor key tones. In the visual categorization task the pictures in combination with the earcons were presented to the subjects in blocks. All pictures within a block were also presented with no sound, so subjects saw all pictures two times in each block, once with a sound and once without a sound present. The experiments consisted of four blocks. In one block, all pictures of animals were presented with a major chord and all pictures of non-animals were accompanied by a minor chord. Because the mood of the sound and the category of the picture both suggest a similar ‘answer’, either positive or negative, this block was labeled *congruent*. In a similar fashion there was one block where the pictures of animals were accompanied by minor earcons and non-animals by major earcons. This block was labeled *incongruent*. Finally, two blocks where both types of pictures were presented with either a major or a minor earcon were called *neutral* (see Table 1). Subjects saw all blocks in a randomized order.

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

Table 1. Experimental conditions earlier experiments

Results show that there is a significant overall delay in reaction times for all trials with an earcon (see Figure 1). It seems that having redundant information in another modality 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 auditory information seems to suggest another response than the visual information, meaning the *incongruent* trials. When the pictures of animals are presented with a minor earcon or the non-animals with a major earcon, the delay is significantly greater than in the other conditions with sound(see Figure 1). This effect has been validated in several experiments [3,4,5].

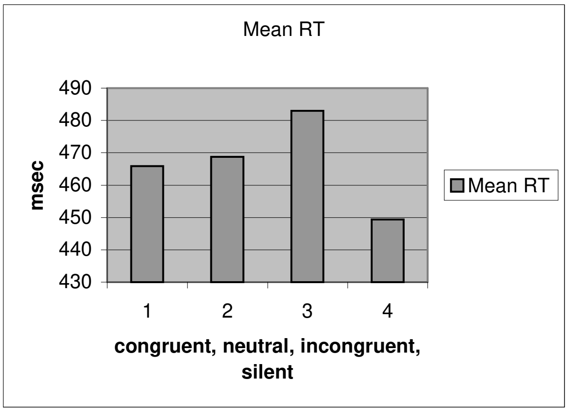

Figure 1. Earlier experimental results

Since these results involve abstract, musical sounds, it seems interesting to study, whether a similar effect occurs when more concrete, real-life sounds are used. Theories on cognition indicate that it is possible that there are two different levels of processing in memory that could be linked to this distinction in types of sound (e.g. [12]). The first level is called the exemplar-similarity-level and involves categorization based on the remembrance of instances. For instance you know that what you see is a cat, because you have seen similar cats before. The second, conceptual level is mediated by rules that are abstracted from objects that fall into the category, i.e. after experiencing multiple examples of a certain category it is possible in some cases to formulate a rule for that category. It almost seems like an imaginary line is drawn and you have to determine on which side of the line the object falls, based on rules. For instance, when you have to determine for a series of numbers whether it is an odd number or an even number, you do this by applying the rule: 'can be divided by two' [14]. Reaction times experiments with exemplar-similarity based stimuli are faster than reaction times with conceptually based stimuli. If such a distinction between exemplar-similarity and conceptual categorization is applicable here, it is possible that a difference in reaction times or errors can be shown when comparing the results of the experiments described in this paper with the earlier results involving earcons. Since the conceptual level is derived from the exemplar-similarity level, that is much more concrete, it is expected that auditory icons, that represent concrete aspects of the objects falling into the category, will lead to faster responses than earcons, that are only in an abstract manner related to the target category.

EXPERIMENT

Subjects

In this study 20 subjects participated, that 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.

Material

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 watch the 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, Monacor BH-004 with a microphone. The microphone was not used during the experiment. 16 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 the picture 'produces' 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 take samples of music that were closed, so subjects didn't 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.

Procedure

Subjects first heard a simple tone (F) to alert them to fixate on a displayed fixation point on the screen ('x'). 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 and the quacking of a duck. 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 were no blocks). 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'. Again, subjects were not 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 length of the rest-period was determined by the subject. 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.

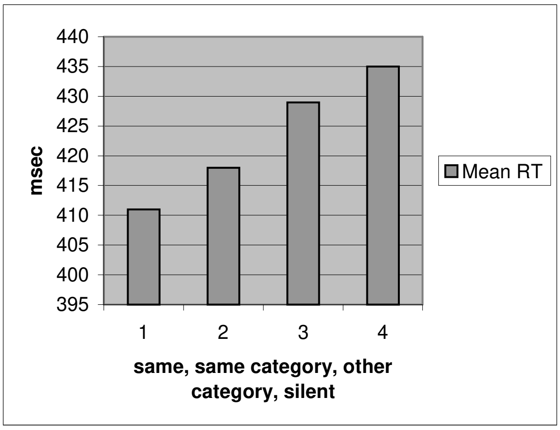

Figure 2. Visual categorization task

The mean reaction times per condition are presented in Figure 2. 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 (average of 411 msec). The condition where the sound presented with the picture was of the same category was a little slower. This condition had an average reaction time of 418 msec. The condition with a sound from another category and the silent condition were slowest (average 429 msec and 435 msec). A repeated measurements analysis was conducted on the means, that showed that the differences in mean reaction times is significant ($F(3,17)=9.713$, $p=0.001$). When comparing the conditions that have the same category (*same and same category*) there is no significant difference in mean reaction times ($F(1.19)=1.481$, $p=0.238$). One seems to be a subset of the other, since both sounds are from the same category as the pictures, and when we combine the data from both conditions (see figure 3) and compare this to the *other category* condition in a contrast analysis, there is a significant difference in mean reaction times ($F(1,19)=6.109$, $p=0.023$). The reaction times in the conditions with the same category are faster than the reaction times in the conditions with another category. However when looking at the condition *other category* compared to the condition without sound, there is no significant difference in mean reaction times ($F(1,19)=0.482$, $p=0.496$).

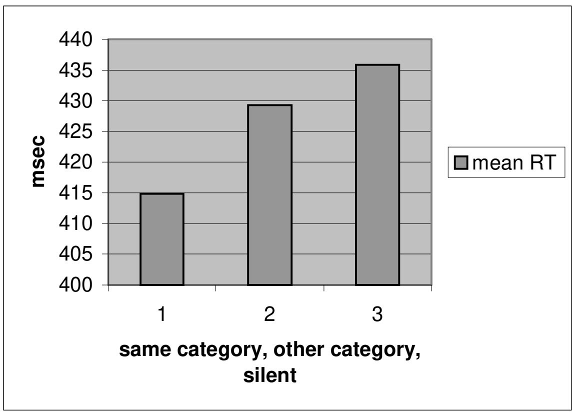

Figure 3. Results combined

DISCUSSION

Auditory icons

The results show significantly faster reaction times in the conditions with auditory icons. 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 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 congruent with the picture information. When what you see and what you hear suggests the same type of response, subjects are able to react faster.

Auditory icons vs. earcons

Looking at the data of the study presented here in comparison with the results on the experiments conducted earlier with earcons as sound stimuli, some interesting similar effects can be found. Figure 4 displays the data from both the earcon-experiments described earlier and the auditory icon data from this experiment. The mean reaction times per category are presented. The mean of the silent condition is based on the reaction times of both experiments.

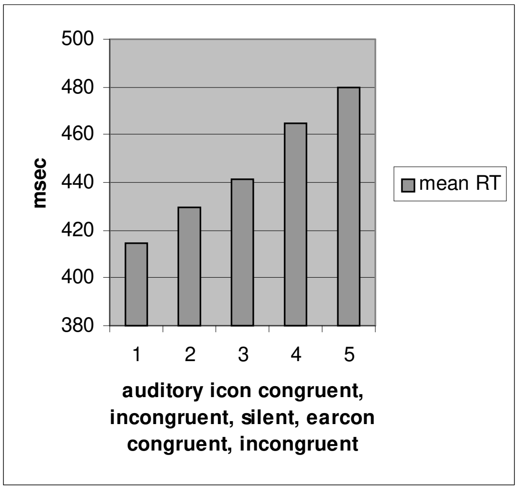

Figure 4. Earcon and auditory icon data combined

The difference between having a sound of the same ‘category’, or a major chord on the one hand, and having a sound of another ‘category’ on the other hand, is similar for both types of sound. In both series of experiments the reaction times for the ‘congruent’ condition(s) are faster than the ‘incongruent’ condition. This seems to indicate that when doing a visual task, having information in the other modality only aids when that information is *congruent* with the visual information. When looking at the Figure 4, also some differences between the two types of experiments can be observed. Firstly, there is a larger difference between the mean reaction times of the *incongruent* condition and the silent condition in the earcon experiment, than in the auditory icon experiment. In the earcon experiment, it seems that having information that contradicts the response that is suggested by the picture, leads to significant increases in reaction times. In the experiment with auditory icons, there is no significant difference in mean reaction times between the condition with a sound in another category and the silent condition. To determine whether this apparent contrast in results is an indication of an effect, further research is needed.

The most important difference between the two types of experiments is that all conditions with auditory icons are faster than the silent condition and the all conditions with earcons are slower than the silent condition. This seems to suggest that users are able to categorize pictures faster when they are accompanied by a real-life sound, then when they are presented with an abstract, musical sound. However, the silent trials were embedded in the experiment and therefore could be influenced by the context. Further studies in purely visual categorization are needed to validate this finding. The earlier observation that when having redundant information in another modality, users seems to respond slower needs to be specified: It seems that when having exemplar-based redundant information in the other modality results in users responding faster.

The results seem to suggest that there is a general difference in mean reaction times in experiments with auditory icons and experiments with earcons. This could mean that the proposed distinction between exemplar-similarity and conceptual categorization are reflected in this reaction time data. Furthermore the results could indicate that categorization on the basis of conceptual information is slower than categorization based on exemplar-similarity information. Intuitively this seems reasonable, since the conceptual categorization is more abstract. It seems plausible that categorizing on the basis of more abstract information takes more time.

IMPLICATIONS FOR INTERFACE DESIGN

The study presented here reiterates earlier findings that care should be taken when applying sound. Having sound in a visual task does not always lead to faster reaction times. The type of sound, whether it is an auditory icon or an earcon, seems to influence the time it takes to come to a response. On the basis of reaction times alone it seems that auditory icons are preferred in situations where performance is important. However, as noted earlier, performance should not be the only criterion when defining sounds. Since users find real-life sounds annoying when they often hear them, frequency of use should be taken into account as well. Although reaction times are slower in the experiment with earcons, it seems that users are able to extract information from these sounds and use it. Again it is shown that using abstract sounds is a good alternative to auditory icons. Nevertheless this should not be done arbitrarily. Even the abstract meaning of a sound may influence the processing of a concrete picture. Also, the results show that when having any type of additional sound present, auditory icon or earcon, it should be congruent with the visual information to assist the subjects in their task. This means that not any sound is suitable for application in a certain function.

Finally, the research presented here indicates that there are different levels or types of categorization. Whether using earcons or auditory icons, the meaning of the sound in relation to the visual information should be carefully considered.

REFERENCES

1. Blattner, M.M., Sumikawa, D.A., and Greenberg, R.M., Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction* 4(1), 1989, 11-44.
2. Brewster, S.A. Providing a structured method for integrating non-speech audio into human-computer interfaces. *Doctoral dissertation* (York, UK, 1994).
3. Bussemakers, M.P., and de Haan, A. Getting in touch with your moods: using sound in interfaces, submitted.
4. Bussemakers, M.P., and de Haan, A. Using earcons and icons in categorisation tasks to improve multimedia interfaces, in *Proceedings of ICAD'98* (Glasgow, UK, November 1998) British Computer Society.
5. Bussemakers, M.P., de Haan, A., and Lemmens, P.M.C. The effect of auditory accessory stimuli on picture categorisation; implications for interface design, in *Proceedings of HCII'99* (München, Germany, August 1999) 436-440.
6. Edworthy, J. Does sound help us to work better with machine? A commentary on Rauterberg's paper 'About the importance of auditory alarms during the operation of a plans simulator'. *Interacting with Computers,* 10*,* (1998), 401-409.
7. Gaver, W.W. The sonicfinder: an interface that uses auditory icons. *Human Computer Interaction* 4(1), 1989, 67-94.
8. Hevner, K. The mood effects of the major and minor modes in music, in *Proceedings of the midwestern psychological assiciation* (1933), 584.
9. Mynatt, E.D. Designing with auditory icons: how well do we identify auditory cues? in *Proceedings of CHI'94* (Boston, US 1994), 269-270.
10. Rauterberg, M. About the importance of auditory alarms during the operation of a plant simulator. *Interacting with computers,* 10, (1998), 31-44.
11. Roberts, L.A., and Sikora, C.A. Optimising feedback signals for multimedia devices: Earcons vs Auditory icons vs Speech, in *Proceedings of IEA'97* (Tampere, Finland, 1997), 224-226.
12. Shanks, D.R. Representation of categories and concepts in memory, in *Cognitive models of memory.* Martin A. Conway (eds). 1997. Psychology Press, UK. 111-146.
13. Sikora, C.A., Roberts, L.A., and Murray, L. Musical vs. Real world feedback signals, in *Proceedings of CHI'95* (Denver, US, 1995), 220-221.
14. Smith, E.E., Patalano, A.L. and Jonides, J. Alternative strategies of categorization. *Cognition, 65,* (1998), 167-196.