Using Earcons and Icons in Categorisation Tasks to Improve Multimedia Interfaces

Drs. Myra P. Bussemakers

Nijmegen Institute for Cognition and Information, Catholic University of Nijmegen
The Netherlands

Dr. Abraham de Haan

Nijmegen Institute for Cognition and Information, Catholic University of Nijmegen
The Netherlands

Abstract

In this study, the modality appropriateness hypothesis that originated from experiments in perception is tested for human computer interaction situations. In multimodal information processing users need to integrate the data coming from various sources into one message. In a visual and auditory categorisation task with accessory stimuli in the other modality, containing a mood, it was shown that in tasks where choices need to be made based on the meaning of the stimuli, the visual modality seems more appropriate. From the results can be concluded that users do not always benefit from having information in more than one modality.

1 Introduction

When users interact with multimedia oriented interfaces, there are different information streams that users have to process. Besides receiving data visually, they can listen to their computers and in some cases they can even feel the desktop they are working on.

Usually, these types of information reach the user simultaneously and it is up to him or her to make sense out of all this, i.e. to create a perceptual unity from this multimodal information, in order to come to a correct response. A perceptual unity means that users come to a unequivocal interpretation. A good example of this phenomenon in our everyday world is 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 (e.g. [16]).

When we look at the modalities that are used in interfaces, we clearly observe the auditory and visual modality. The research reported here will focus on these information streams. Graphical representations, icons, are used for example in a desktop environment to assist in the selection and manipulation of files and programs. A file when selected changes colour and gives feedback on what happens in the system. Auditory information, an abstract musical sound (earcon) that is played for example, is used in a similar sense to give feedback on operations performed by the user or can even assist in the representation of complex data.

The representations within the interface have different characteristics. On the one hand they are perceptual stimuli that need to be perceived to have effect, but they are more than that. An icon for instance graphically represents a function and therefore seems to have meaning for the user. A user can tell which program will be activated when the icon is selected.

On the other hand in some cases the representation also consists of a distinct emotional tone, a mood. An auditory example of a representation of a negative mood is the sound that is played when an error occurs. The purpose of the sound is to alert the user that (s)he has made a mistake. Often, the interaction with the interface will take place involving multiple modalities. The user will receive information both visually and through sound.

Assuming that the human perceptual system is capable of processing these different information streams in parallel, at some point these streams are expected to be integrated [22]. The role that each modality plays in the integration can be complex. Both the picture and the sound can carry necessary information. It is also possible that one modality carries the 'primary message' and the other carries additional characteristics, co-messages, such as a mood. The aim of this research is to model and predict what combinations of primary messages and co-messages in different modalities for interface-related tasks could support each other and what combinations could inhibit each other, so that multimedia interfaces can be created more optimally.

2 Intersensory Integration in Perception

One of the characteristics of multimodal representations within the interface is that both modalities need 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 system can substitute for another system [21]. An example of that is *the ventriloquism effect* [13], where vision influences the auditory perception in the sense that it seems that the puppet is speaking. Although this effect can occur in different modalities, it seems to be most dominant in vision. 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 [21]. This indicates that the contribution to the perceptual unity in the parallel information processing that seems to occur in multimodal perception, is not equal.

However, this visual dominance seems to depend on the task that needs to be performed. Welch and Warren [24] 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 [4]. Subjects often only noticed the light and did not even perceive the tone. In spatial perception tasks, vision is also a much 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. [23]). 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* [10].

The type of research mentioned here seems to confirm the modality appropriateness hypothesis for reactions to perceptual stimuli, like a flash of light for example, or a simple tone. Because in human-computer interaction the presented stimuli have more characteristics, it is questionable if the same results will occur in more interface related situations.

3 Special Types of Earcons and Icons in Interfaces

Another aspect that results from the integration of various modalities can be a mood (e.g. [22]). When we see a movie, for example, the music that is played gives a special flavour to the images that are presented. Music in this sense can be understood as the interpretative co-message, that indicates whether the images should be seen as funny, frightening or otherwise (e.g. [2]). It seems that blending occurs automatically, of both the images and the music into a single, coherent perceptual experience that is associated with a distinct emotional tone.

In interfaces this co-message can assist the user in making the right choices where to go or what to do next. The emotional tone, or mood, either visually or auditory, can be used to guide the user in the appropriate direction. A negative mood for example can indicate that the computer did not 'understand' the input that was given. A positive mood on the other hand can indicate that a process was successfully completed. It is our aim to get more insight in how the co-message may blend with the situation.

3.1 'Moody' Icons

When we observe human-human interaction, like a conversation for example, emotions are expressed visually by body posture and gestures, but mostly by facial expressions. These emotions seem to be universal across literate and even preliterate cultures (e.g. [14, 8, 6]) and in both free-choice and fixed-choice studies [18]. From poses or still pictures it is possible for subjects to clearly distinguish at least 7 categories of emotions: happiness, surprise, fear, anger, sadness, disgust/contempt and interest [7]. Because of these categorical distinctions facial expressions seem a good way to visually represent moods in interfaces.

Not only is it possible to 'read' emotions from a face, when subjects take on a certain emotional facial expression, they report to actually feeling the corresponding emotion. This effect occurred when subjects took on the expression upon request as well as by a more subtle muscle manipulation. From these and other studies, it seems that facial expressions of emotions are not only correlated with the experience of emotions, but could also control or initiate it (e.g. [1]). The relationship between the expression and the experience of the emotion is multidimensional and categorical, meaning that a certain expression increases the corresponding feeling of emotion significantly [5].

In deciding what abstract representations of human faces to use in interfaces, the discussion rises whether or not the different areas in the face, the upper face (the area above the eyebrows), the eyes and the lower face (nose, mouth and chin), play an equally important role in the expression of emotions. Matsumoto [15] looked at the eyes when attempting to convey anger and fear. It seems that the eyes not so much indicate what emotion is expressed, but are the prime markers of the *intensity* of a certain emotion. With Caucasian people, for example, the widening of the eyes causes an increase in the perceived emotion; it seems more intense. The lower area of the face seems to be more able to convey what type of emotion is meant. This is especially the case with happiness and sadness, although less distinct for the latter (e.g. [12]). Therefore it seems possible in interfaces through the use of facial icons, to indicate what type of emotion is represented by varying the area around the mouth and to indicate the intensity of the emotion by varying the eyes.

3.2 'Moody' Earcons

Emotions in human-human communication can be expressed in sound, apart from speech, through attributes like intonation and loudness. In interfaces there seem to be two ways to apply non-speech sounds as a co-message.

The first is through the use of auditory icons, that represent an event or object by using 'natural sounds' [9]. An example of this is the sound of breaking glass when you empty the trashcan on your desktop. The advantage of auditory icons is that users know what the sound represents because of it's relationship with instances in the real world. One of the disadvantages of their use however is that users report them to be annoying after prolonged use [20].

A second way of applying non-speech sound is by using earcons. Earcons are abstract, i.e. not event or object related, musical sounds, that can be used 'to steer the emotional reaction of the user in support of a certain response' [3]. Earcons do not share the relationship with events or objects in the real world that auditory icons have, so users have to learn them. Yet it seems to be possible to create a mapping between the earcon and a function in the interface that users judge to be better than the mapping with auditory icons [17]. An example of an earcon is a chord, played on a piano.

So far we have defined a possible way of expressing some of the characteristics of stimuli, their perceptibility and mood. Visually faces can be used that express emotions and in audition earcons can be used. In the interaction with an interface a specific task-setting is involved. Decisions need to be made in order to select or manipulate items. It is this setting that gives meaning to the icons and earcons.

4 Multimodal Integration Involving Icons/Earcons

An example of a task that is somewhat similar to what seems to happen in human-computer interaction is a categorisation task. Subjects have to decide whether or not the presented stimuli are of a certain category by pressing a button labelled 'yes' or a button labelled 'no'. In order to make that decision, they have to determine the meaning of each presented stimulus. In an interface, icons and earcons also belong to a group or category in the sense that they represent a function, i.e. have a meaning.

In the study presented here, there was one modality that was carrying the primary message, the information necessary to decide on the category. The accessory stimuli in the other modality had additional information that in some cases could assist in the categorisation judgement and in other cases could distract from it. The stimuli that subjects were instructed to categorise were animals, like a dog for example, or non-animals, like a plane.

Because the processing of auditory information is about 40-60ms faster than the processing of visual information, the visual stimulus needs to be presented earlier than an auditory stimulus so that the processing is expected to occur at about the same time. Gielen et al [11] found that presenting a visual stimulus a little earlier than an auditory stimulus can lead to significantly shorter reaction times than the reaction times to one of the stimuli alone.

Because we wanted to keep both experimental manipulations as similar as possible and wanted the processing of both stimuli to occur at about the same time, the visual accessory stimulus was presented earlier than the auditory target stimulus. In the case of the auditory accessory stimulus earlier experiments showed that the effect of that stimulus was the greatest when were presented simultaneously with the visual target stimulus at an Stimulus Onset Asynchrony of 0 [19, 2].

4.1 Visual categorisation

In the first type of experiments the target stimuli containing the primary message, were presented visually through line-drawings. With the line-drawings in some cases earcons were played. As a positive emotion, a major chord (Cmaj) was played and as a negative emotion a minor chord (Cmin) was played. Representing the neutral emotion, the same simple tone was played that was used as a attention-sound (F). The duration of the sounds was 2500 ms each. To alert subjects in both modalities to the next trial a cross ('x') was displayed on the screen and the neutral sound was played.

Subjects were instructed to press a button labelled 'yes' as quickly as possible when they saw a picture representing an animal and to press a button labelled 'no' when they saw a picture that did not represent an animal. The reaction times were measured to come to a correct response. All subjects participated in all the conditions in random order. In the three experiments run with this design, a total of 55 subjects participated. They were all students at the Catholic University of Nijmegen and were paid for their time. (For further explanation of the experiments see [2]).

The three experiments all show a similar general result. In all cases where a sound was played with the pictures, the reaction times were significantly slower than in cases where no sound was played. This negative effect differed in the third experiment when looking at the mood of the sound in combination with the type of pictures that was shown. In this design there was a direct relationship between the type of picture that was shown and the type of sound that was played. This way the subjects were able to 'hear' when a sound was played what type of picture was presented, although they were not explicitly instructed to do so.

The results indicated that when the mood of the sound was incongruent with the response that was needed, the inhibitory effect on the reaction times was even greater. This means that it took subjects longer to respond to a picture of a dog, suggesting a positive answer, together with a minor chord (a negative emotion, suggesting a negative answer), than to a picture of a dog with a major chord. In the latter, congruent case both stimuli suggest a positive answer. In trials where both the animals and the non-animals were accompanied by either a minor chord or a major chord (the 'same' condition), the reaction times were similar to the congruent situation (see figure 1).

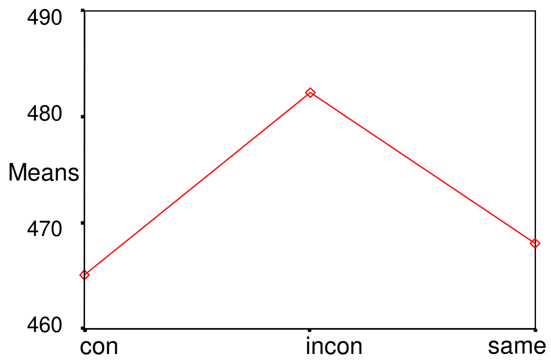

Figure 1: Type of picture vs. Type of sound

4.2 Auditory Categorisation

Similarly experiments were run where the stimuli containing the primary message were presented through spoken words. In some cases, at a Stimulus Onset Asynchrony (SOA) of –100ms, icons were presented, i.e. abstract representations of a human face, that were used as a co-message with an emotional tone. The upper area of the face, represented by the eyes were constant over conditions. The only manipulation occurred in the lower face, meaning that the mouth was varied across conditions. The abstract representation that was chosen is often used in visually oriented communications, such as e-mail for example, as 'smiley's', to express emotions. For this study the smiley's were rotated 90 degrees (see figures 2, 3 and 4). To alert subjects in both modalities to the next trial a cross ('x') was displayed on the screen and a neutral sound was played.

Figure 2, 3 and 4: positive, neutral and negative icons

Subjects were instructed to press a 'yes' button as quickly as possible if they heard a word that represented an animal or to press a 'no' button when they heard another word. Again the time was measured for subjects to come to a correct response. All subjects participated in all conditions in random order.

In the two experiments run with this design, 48 subjects participated. They all were students at the Catholic University of Nijmegen and were paid for their time.

Both experiments show a similar result. The reaction times to the trials with no icons were significantly slower than the trials with icons. In the second experiment a design was used where the relationship between the type of words and the type of icons was constant. Within a condition, words of animals were always played together with happy faces and non-animals words with sad faces for example, or sad faces with animal words and neutral faces with non-animal words. All possible combinations of word-types and faces were used. In a different condition all words were combined with the same icon, either happy, sad or neutral.

In this experiment the positive effect was different for the combinations between the type of word the subjects heard and the icon they saw. When the mood of the face was either congruent or incongruent with the response that was needed on the word, there was a significant positive effect on the reaction times when compared to the 'same' condition (see figure 5). This means that it took subjects less time to respond in trials where the icons were different for the animals and the non-animals.

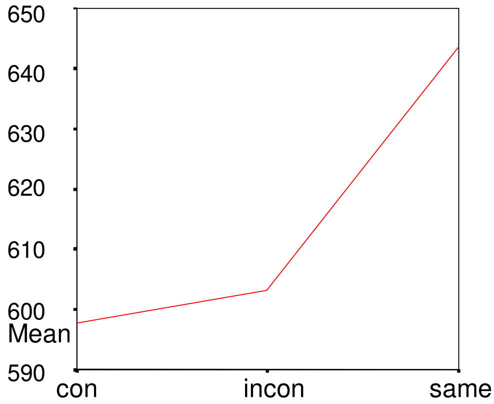

Figure 5: Type of word vs. Type of face

5 Conclusions and Implications

The results presented in the visual categorisation task, suggest that when the primary message is presented visually, any accessory information that is provided in the auditory modality causes a delay, even when this information can assist in the choice for a correct response. There seems to be a preference to only have the information presented visually, where speed of responding is concerned. The error-rate during all experiments was very low, so it is impossible to judge whether or not the accessory stimulus more often assisted in formulating a correct response.

The results from the auditory categorisation task suggest a somewhat different effect. When the primary message is presented auditorily, in all cases the accessory visual information speeds up the response. Apparently in this situation the extra visual information can help quicken the response times. Furthermore the reaction times to the conditions where different icons were used with each word-type, the reaction times were even faster. This could mean that subjects learned the relationship between the face and the word-type in stead of using the mood it was representing. More research is needed to study these findings.

Summarising the results it seems that the modality appropriateness hypothesis is confirmed. In other than purely perceptual tasks there seem to be modalities that are better suitable for certain tasks than others. In this categorisation task the visual modality seems better suitable than the auditory, which leads to a dominance of vision over audition. It will be interesting to study other types of tasks to see whether the modality appropriateness hypothesis also applies there. A possible task in this area could be key-entry.

When we generalise these outcomes to a specific task-set within human-computer interaction, especially graphical interfaces, it seems that great care must be taken when using sound as an accessory stimulus to primary visual information. It is shown here that it is possible that there are operations where sound may even hinder the interaction. It is even the question if in some situations sound should be included at all. In primary auditory task-settings however where categorisation is required, the user can benefit from having additional visual information. Here having information in both modalities could lead to a better, i.e. faster interaction.

It is our goal to theoretically model this multimodal integration in human-computer interaction and to try and predict from this model, in what situations users can benefit from having information in more than one modality. Above all, these experiments show that using multiple modalities in interfaces to get a message across may not always be the best option.

6 References

1. Adelmann PK, Zajonc, RB. Facial efference and the experience of emotion. Ann Rev Psychol 1989; 40: 249-280

2. Bussemakers MP, De Haan A. Getting in touch with your moods: using sound in interfaces. In press.

3. Blattner M, Sumikawa D, Greenberg R. Earcons and icons: their structure and common design principles. Hum Comput Interact 1989; 4: 11-44.

4. Colavita FB, Weisberg D. A further investigation of visual dominance. Perc Psychoph 1979; 25: 345-347

5. Duclos SE, Laird JD, Schneider E, Sexter M, Stern L, Van Lighten O. Emotion-specific effects of facial expressions and postures on emotional experience. J Person Soc Psychol 1989; 57: 100-108

6. Ekman P. Facial expressions of emotion: New findings, new questions. Psychol Sci 1992; 3: 34-38

7. Ekman P, Friesen WV, Ellsworth P. Does the face provide accurate information. In: Ekman P, Scherer KR (eds) Studies in Emotion and Social Interaction. Cambridge University Press, Cambridge, 1982, pp 56-98

8. Ekman P, Sorenson ER, Friezen WV. Pan-cultural elements in facial displays of emotions. Science 1969; 164: 86-88

9. Gaver W. The SonicFinder: an interface that uses auditory icons. Hum Comput Interact 1989; 4: 67-94

10. Gebhard JW, Mowbray GH. On discriminating the rate of visual flicker and auditory flutter. Am J Psychol 1959; 72: 521-528

11. Gielen SCAM, Schmidt RA, Van den Heuvel PJM. On the nature of intersensory facilitation of reaction time. Perc Psychoph 1983; 34: 161-168

12. Hanawalt NG. The role of the upper and the lower parts of the face as the basis for judging facial expressions: II. In posed expressions and “candid camera” pictures. J Gen Psychol 1944; 31: 23-36

13. Howard IO, Templeton WB. Human Spatial Orientation. Wiley, London, 1966

14. Izard CE. The face of emotion. Appleton-Century-Crofts, New York, 1971

15. Matsumoto D. Face, culture, and judgements of anger and fear: do the eyes have it. J Nonverb Behav 1989; 13: 171-188

16. McGurk H, MacDonald J. Hearing lips and seeing voices. Nature 1976; 264: 746-748

17. Roberts LA, Sikora CA. Optimising feedback signals for multimedia devices: Earcons vs. Auditory icons vs. Speech. Proceedings of IEA’97, Tampere, 1997

18. Rosenberg EL, Ekman P. Conceptual and methodological issues in the judgment of facial expressions of emotion. Motiv Emo 1995; 19: 111-138

19. Schriefers H, Meyer AS. Experimental note: Cross-modal visual-auditory picture-word interference. Bull Psychon Soc 1990; 28: 418-420

20. Sikora CA, Roberts LA, Murray L. Musical vs. Real world feedback signals. Proceedings of CHI’95, Denver, 1995, pp 220-221

21. Stein BE, Meredith MA. The Merging of the Senses. MIT, Massachusetts, 1993

22. Teasdale JD, Barnard PJ. Affect, cognition and change: Re-modelling depressive thought. Lawrence Erlbaum Associates Publishers, London, 1993

23. Welch RB, DuttonHurt LD, Warren DH. Contributions of audition and vision to temporal rate perception. Perc Psychoph 1986; 39: 294-300

24. Welch RB, Warren DH. Immediate perceptual response to intersensory discrepancy. Psychol Bull 1980; 88: 638-667