Sonification of Facial Expressions

Ioulia Guizatdinova
TAUCHI Computer-Human Interaction Unit
Department of Computer and Information Sciences
University of Tampere
FIN-33014 Tampere, Finland
+358 3 215 8556
ioulia_gizatdinova@pochtamt.ru

Zhiguo Guo
Department of Computer and Information Sciences
University of Tampere
FIN-33014 Tampere, Finland
+358 3 041 440 9668
Zhiguo.Guo@uta.fi

ABSTRACT
Visual attributes specifically describing human facial expression can be inaccessible when visual perception is blocked or an image is hidden for observer. Current computer programs can process facial expressions with speech output or alternatively by adding tactile feedback cues. However, facial traits and expressions are still not sonificated. The goal of this project was development and usability evaluation of the sonification system for alternative access to facial expressions through eARmoticons. eARmoticons should evoke the emotional feelings of a listener like those a visual image could produce. The results of our study showed that after some training time the auditory imaging facial expressions could also be accessible for a listener. The proposed technique, which can briefly display an array of related attributes like facial traits, could facilitate communication and interpretation of visual images for people with special needs.

KEYWORDS: emotional expression, facial traits, sonification, earcons, eARmoticons.

INTRODUCTION
The lack of visual information can become a great problem when visual perception is temporary blocked or even unavailable. For example, for blind people an access to visual information has always been an obstacle. Nowadays some computer programs with text-to-speech and video processing can recognize simple pictures and sonificate some attributes with auditory output or alternatively by adding tactile feedback cues. Both an emotional content of the images and facial expressions are still not accessible for the blind or when an image is hidden for observer.

In the development of alternative communication and access to visual information, sonification, as a method, has been used to map the communication between two modalities. Sonification is the use of non-speech audio to convey information; more specifically sonification is the transformation of data relations into perceived relations in an acoustic signal for the purposes of facilitating communication or interpretation [3].

Some sonification methods suppose to set only quantitative conformity of sound parameters to the visual ones. It is supposed too that a person may learn any sound code (like Morse code) and substitute visual mental notions (percepts) by hearing complicated sound patterns [8]. Other investigators are relying on cognitive transfer, like as synesthesia phenomenon, or intermodal sensations [2].

P. Meijer [6] proposed the sonification method of arbitrary pictures registered by a video camera. His approach is based on agreement that amplitude of a sinusoidal oscillator is proportional to pixel gray level, frequency is dependent on a vertical pixel position, and a horizontal pixel position is translated into time-after-click. After that image is scanned through one column at a time, and the associated sinusoidal oscillator outputs are presented as a sum (chord) followed by a click before presentation of a new sequence of columns. The example of smiling face which is sonificated using “The vOICe Java Applet” [9] and its spectrogram are shown in Figure 1. The method requires a lot of training to recognize the complicated earcons.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

New Interaction Techniques '03, Spring, 2003, University of Tampere.
Copyright 2003 University of Tampere, Finland.

Figure 1: An original image (on the left) and image-to-sound mapping (spectrogram) on the right. [6].
However, in contrast to emoticons, present methods of transformation cannot be used to communicate meaningful visual attributes such as facial traits or expression without long training, based on previous experience of a listener. And the question remains how to convey the facial traits with emotional tinge of some person presented like any visual image into the sounds, for instance, earcons or music chords.

In a paper on auditory emotional access to information [8], Harald Schwende writes that sonification system enables an emotional access using the similarity of feeling, by watching a picture or hearing music and transforming visual information in music compositions.

The goal of our work was designing and usability evaluation of the sonification system for facial expressions, which should allow a transformation of the visual information (related attributes specifically describing emotional expression/state) the into sound patterns. This kind of sounds ("eARmoticons") should evoke the similar emotional feelings of a listener, that is, to provide an emotional access through “audio snapshot” when visual perception is blocked or an image is hidden for observer.

SOUND MAPPING
There are two main blindness categories, they are: congenital and adventitious. The adventitious blind people have experience with real images, whereas the congenital blind people have not any associations to visual image [5], but extensive tactile and auditory experience with surround objects.

There are also some facial features, which are hardly transformed by sounds, for instance, congenital or non-classified attributes.

Let us assume that a visual image of facial expression (video image, picture) or related attributes specifically describing facial expression were processed and classified. Then, the problem is how to convey an array of known visual traits into brief and informative sounds – stimuli, which will provoke similar behavioral patterns and feelings (or emotions), or evoke a mental association with emotional content.

First, two approaches to image-to-sound mapping could be defined. One way is a transformation of each facial expression as a whole. We called this mode in the future system – direct sonification (DS). The second technique could be based on sonification of the remarkably sparse set of facial landmarks like forehead, eye, nose bridge and mouth defining a human face, called as indirect sonification (IS). The conceptual model of both approaches is shown schematically in the Figure 2. The model is obvious but it is a good start point for evaluating sound parameters, which could be adopted or rejected through empirical investigation.

Six basic emotions, which could reliably be identified through corresponding six universal expressions proposed by Ekman and Friesen, were selected and marked by the following labels: surprise, fear, anger, disgust, sadness, and happiness [1]. According to this classification, there are generally three types of emotions and facial expressions: positive, neutral and negative. Anger, disgust, fear and sadness are classified into the negative type; happiness and surprise are classified into the positive type; and additionally there is one neutral type. Then, the simplest approach to sound mapping is to divide a frequency range according emotional state (Figure 3).

Figure 2: Conceptual model of the image-to-sound-mapping.
The positive emotions (happiness and surprise) could be sonificated through high-frequency sounds, the negative emotional states (sadness, anger and disgust) could be conveyed into sounds with low frequencies, and the neutral emotional state could be presented by sounds of middle-frequency range. The expressive performance of sounds includes different spectral shapes of the signals. Probably, the timbre is more correct sound parameter describing eARmoticons by taking into account spectrum of MIDI tools used in the synthesis. We will discuss this parameter in the next sections. Meanwhile the fear belongs to the negative type, but in our study we considered the fear to be the positive type regarding sound frequency.

To design eARmoticons for DS and IS modes we have used different methods and MIDI tools by creating the sounds [4]. With IS technique it is supposed that facial landmarks should have well-recognized timbres which are determined by emotional state. The tools, used to change sound envelope and timbre, were organ, piano, guitar and trumpet (Appendix 1). When we used DS mode to sonificate the emotion as a whole, we used Orchestra Hit (Appendix 2). The instruments for both sonification modes are shown in Figure 4.

![Figure 3: MIDI Frequency (Hz) for sonification of the facial expressions for both IS and DS modes.](image)

![Figure 4: Modes and instruments: Indirect Sonification (the left image); Direct Sonification (the right image).](image)

**EXPERIMENTAL EVALUATION**

**Participants**

13 volunteers (7 males and 6 females) from staff and students at the University of Tampere participated in the study. The ages of the subjects ranged from 21 to 33 years with a mean age of 27. All of the participants had a normal vision and hearing. None used hearing aid.

**Apparatus and Procedure**

The experiments have taken place in the usability laboratory of Tampere Unit for Computer-Human Interaction (TAUCHI) during two weeks.

A desktop PC (ASUS A7V133) was equipped with AMD Duron Processor, 256MB RAM, VIA AC’97 Audio Controller (WDM) and stereo system. The 15” LG Flatron Monitor, 575 LM Multimedia Speakers were used.

The experimental software was written by Ioulia Guizatdinova in VC++ running in Windows 2000 operating system.

In order to test both techniques through DS and IS, we carried out two experimental sessions. Each experimental session consisted of two parts, training and the test. During training images were displayed in the following order: neutral, happiness, surprise, fear, anger, disgust and sadness. The subjects could manipulate sounds by clicking on the whole image or on the different facial landmarks. After presentation of the facial expressions and sounds the pictures were masked. The subject could make a decision by clicking on the mask (the same layout of the facial landmarks) or just by listening sounds of the whole face. During the test, hidden emotional expressions were presented in a random order. Figure 5 shows a snapshot of the experimental software during IS and DS modes correspondingly.

Both the size and optical parameters (brightness and contrast) of images and sound volume were not changed throughout testing. There was a 5 minutes demonstration about the system for the subjects, they also were instructed to click on the image, listen to the sounds, and try to memorize them. After that, eight trials were performed for each subject.

While test time was not limited, each trial lasted about 30 minutes, 15 minutes per session. A five-minute break was allowed between experimental sessions. The test took a total time of 4 hours for each participant. To motivate a performance of the subjects after each trial, spent time and recognition rate per image were displayed. At the end of the test subjects were asked to fill a questionnaire.

During training and the test we investigated learning and retention of the sounds. In particular, we recorded a number of clicks and the time needed to associate and memorize eARmoticons regarding the image.
RESULTS

Data of two experimental sessions were analyzed separately.

Indirect Sonification

The time required to memorize and recognize eARmoticons was analyzed regarding each facial expression. Before experiment, it was supposed that the subjects could be confused whether eARmoticons with high or low frequencies belong to positive or negative type, or when their frequency or spectrum are similar (Appendix 1). Nevertheless, experimental results showed that the subjects spent more time to investigate neutral eARmoticons, which were sonified in middle-frequency range (Figure 6).

A number of clicks was used to estimate learning and retention of the eARmoticons. A number of clicks was averaged over facial expressions and facial traits. The results demonstrated that neutral expression required a greatest number of clicks in training mode. The measure of ANOVA (analysis of variance) revealed the statistically significant differences between neutral expression and others ones.

Figure 7 shows that the greater values of differences were detected for sadness in relation to neutral eARmoticons when the subjects tried to memorize them. For example, \( F_{\text{forehead}} = 10.96; F_{\text{eyes}} = 7.19; F_{\text{nose bridge}} = 9.35d; F_{\text{mouth}} = 8.34, \alpha = .05. \)

Eyes’ landmarks required the greatest number of clicks in training mode when visual image was available (Figure 7). Moreover, five participants chose strategy to remember only eARmoticons of the eyes and ignored other facial landmarks. Seven subjects noted that eARmoticons well designed with piano timbre and these sounds were more familiar and pleasant than organ, guitar or trumpet ones. They told also that it was hard to understand differences between guitar and trumpet regarding their timbre. They assumed that these eARmoticons might be easily distinguishable if duration would be longer.

While all the participants reported that they have a normal hearing, recognition of eARmoticons depend on individual sensitivity to sounds of high and low frequency. The overall recognition picture was good for neutral, anger and sadness eARmoticons, which had low frequencies (Figure 10). Recognition of eARmoticons, which had high frequency, was worse. However, some of the subjects showed very good recognition rate of high-frequency eARmoticons (Figure 11), in particular, women.

Direct Sonification

No facial features were investigated in this experiment. Sounds described facial expression as a whole. The results for a number of clicks and elapsed time were averaged over type of expressions (Figure 8 and Figure 9). We did not find of specific features in behavior of the subjects, for instance, through a number of clicks in both training and the test.
Figure 7: The mean number of clicks (and ANOVA) needed to memorize and recognize facial expression or facial traits through eARmoticons; the top image – in training mode, the bottom image – during the test.

Figure 8: The mean number of clicks required to memorize and recognize facial expression through eARmoticons. Only elapsed time to recognize neutral expression revealed significant difference. The maximum difference was observed between a repetition of eARmoticons that symbolized sadness and neutral expression: in training mode $F = 13.49$, during the test $F = 10.60$. The results of this experimental session showed that recognition of low-frequency eARmoticons was significantly better (Figure 10).

eARmoticons that symbolized anger and disgust were recognized correctly in 100% of cases. The overall recognition performance of high-frequency eARmoticons remained low, with an exception of the fear eARmoticon. As in a case of IS, some of the subjects have demonstrated an excellent recognition of high-frequency sounds (Figure 11). But we did not take into account musical experience of some participants, as the sonification should satisfy needs of all people independently of individual skills.
DISCUSSION

There is incontestable evidence about vision domination under other attention mechanisms of human beings [10]. On the other hand, it is found that eyes are the most informative regions on the face [11]. As the test consisted of not only of audio stimuli (eARmoticons) but also visual presentation of the pictures of verified emotional expressions, forming of mental associations took place under dominant influence of visual analyzer. Thus, the subjects might subconsciously give preference for the investigation of the eyes only.

The received data gave us information that sonification of eyes should be made more accurately as this trait captures more attention of an observer. Other facial features had similar recognition rate through eARmoticons when visual expression was hidden.

Data of both IS and DS sessions showed that in comparison to the neutral expression, sonification of other ones had required fewer repetitions of eARmoticons for memorizing. We have considered several reasons for these phenomena. First, it should be noted that the concept of a neutral face is a problematic one. Some psychologists state that there are no perfectly neutral faces. A neutrality of the exposed face may depend on the individual experience of the observer. It has been shown [7] that reactions to expressive displays can be a joint function of the type of the display and the observer’s attitudes. Perception of the neutral face may also depend on the context of other (previously) presented facial expressions. For example, in the context of smiling faces a neutral face may be classified as sad. Thus, ambiguity in visual perception of the facial neutrality can essentially influence onto the auditory perception of eARmoticons. Second, due to some similarity of spectrums between neutral, disgust and anger eARmoticons (Appendix 2) the subjects have felt confusion during recognition in IS mode. Subjects could be confused whether eARmoticons with middle-frequencies belong to the negative type. The same distraction was observed for eARmoticons that symbolized surprise and happiness both with IS and DS techniques. Third, short duration of eARmoticons (<600 ms) might also present some difficulties in recognition of sounds.

CONCLUSION

Overall, the results of our experiments showed that facial expressions could be briefly presented through sounds with emotional content called eARmoticons. In presented project there were used two techniques. The first method was sonification of facial expression as a whole through one sound. The second technique was sonification of emotional expression with the help of the set of eARmoticons that symbolized traits like forehead, eye, nose bridge and mouth. In spite of good overall recognition picture through eARmoticons and both techniques, several remarks could be formulated.

The further work could concentrate on designing more complicate brief patterns. But, similar to conventional menu alternatives, sound parameters of eARmoticons should be unambiguous, mutually exhaustive, exhaustive, and non-overlapping. The results demonstrated that designers of sonification system should take into account destination of symbolic sound system, whether sonification system is intended for the visually impaired or users with normal vision.

ACKNOWLEDGMENTS

The research presented here was performed at TAUCHI Computer-Human Interaction Unit of Department of Computer Science of University of Tampere. We would like to thank every volunteer participating in our usability test. We greatly acknowledge Grigori Evreinov who gave us this interesting topic and guidance during our work.

REFERENCES

4. GoldWave. Product information is available at:
http://www.goldwave.com


APPENDIX 1

Sound signals (wav) and spectrograms (frequency spectrum) of neutral eARmoticons (185 Hz) for imaging facial traits with Indirect Sonification technique.

Sound processing was performed by GoldWave digital audio editor [http://www.goldwave.com].
Sound signals (wav) and spectrograms (frequency spectrum) of eARmoticons for imaging facial expressions with Direct Sonification technique. Sound processing was performed by GoldWave digital audio editor [http://www.goldwave.com].