The Smile is The New Like: Controlling Music with Facial Expressions to Minimize Driver Distraction

Michael Braun
BMW Group Research, New Technologies, Innovations
Garching, Germany
michael.bf.braun@bmw.de

Sarah Theres VÃűlkel
LMU Munich
Munich, Germany
sarah.voelkel@ifi.lmu.de

Gesa Wiegand
fortiss GmbH
Munich, Germany
wiegand@fortiss.de

Thomas Puls
LMU Munich
Munich, Germany
t.puls@campus.lmu.de

Daniel Steidl
LMU Munich
Munich, Germany
d.steidl@campus.lmu.de

Yannick Weiß
LMU Munich
Munich, Germany
yannick.weiss@campus.lmu.de

Florian Alt
Bundeswehr University
Munich, Germany
florian.alt@unibw.de

Abstract

The control of user interfaces while driving is a textbook example for driver distraction. Modern in-car interfaces are growing in complexity and visual demand, yet they need to stay simple enough to handle while driving. One common approach to solve this problem are multimodal interfaces, incorporating e.g. touch, speech, and mid-air gestures for the control of distinct features. This allows for an optimization of used cognitive resources and can relieve the driver of potential overload. We introduce a novel modality for in-car interaction: our system allows drivers to use facial expressions to control a music player.

The results of a user study show that both implicit emotion recognition and explicit facial expressions are applicable for music control in cars. Subconscious emotion recognition could decrease distraction, while explicit expressions can be used as an alternative input modality. A simple smiling gesture showed good potential, e.g. to save favorite songs.

Author Keywords

Affective Computing; Automotive User Interfaces; Driver Distraction; Face Recognition; Multimodal Interaction.

CCS Concepts

•Human-centered computing → HCI design and evaluation methods;

MUM '18,, November 25–28, 2018, Cairo, Egypt
ACM 978-1-4503-6594-9/18/11.
https://doi.org/10.1145/3282894.3289729

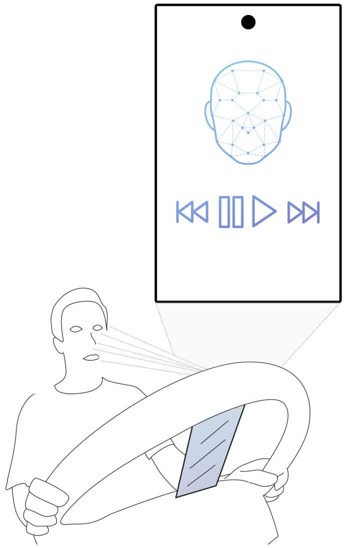

Figure 1: When designing automotive user interfaces, one point of focus lies in finding new ways to minimize driver distraction while optimizing usability. We present a first study on facial expressions as a modality for automotive UIs: our system allows users to select songs by smiling and skip them by frowning. Facial expressions show potential as an alternative modality to classic controls, especially when the connected cognitive resources are exhausted. Icons © Mungang Kim & Ralf Schmitzer.

Introduction

Modern automotive user interfaces often allow the user to choose between multiple input modalities depending on their preferences and momentary demands. The suitability of specific input modalities depends heavily on the function to be controlled. Text input for instance is easiest to complete with speech interaction, while spatial input such as the manipulation of navigation maps is best achieved using manual control [9]. By providing multiple input modalities for single functionalities, drivers can rely on alternatives when the preferred input modality is occupied with a more important driving task [6].

In-vehicle systems in recent production cars (e.g. 2016 BMW 7series, 2018 VW Golf) have introduced new modalities with limited versatility but high benefit, such as mid-air gestures to accept or deny incoming calls or adjust volume. These concepts offer an easy input alternative for much used functionalities with high distraction potential, increasing road safety by limiting distraction. We propose the usage of facial expression recognition as an additional input modality for simple interactions, e.g. skipping songs during music playback.

Related Work

Drivers utilize varying cognitive resources when operating a vehicle, most importantly visual attention and manual control [10]. Auditory and speech resources are used to a lesser extent for driving tasks and more for entertainment purposes like listening to music and conversation. Using these free resources as input modalities is a good way to decrease the driver's cognitive load. However, in certain driving situations these resources may also be required to operate the vehicle, rendering them less appropriate for other tasks [8].

With new sensing technologies like driver camera systems entering the car, novel approaches to input modalities such as explicit facial expressions and implicit signs of emotions can be utilized to control systems. A widely used method to assess facial expressions is the *Facial Action Coding System (FACS)* introduced by Ekman & Friesen [3]. This framework can be used with computer vision to assess single expressions, which can then be utilized to derive an estimation of the driver's emotional state [5].

Building upon this previous work, we propose an in-vehicle entertainment system which allows the user to control music playback with facial expressions in a multimodal input approach.

User Study

We conducted a user study to investigate two fundamental questions regarding the control of in-vehicle systems with facial gestures:

RQ: Are explicit facial expressions a user-friendly modality to control music while driving?

RQ: Can a system deduce if a played song is liked or disliked solely by facial emotion recognition?

Study Design

The first part of the study used a within-subject design. Each participant listened to four different songs while they were driving. The system captured their facial expressions with a frontal camera and calculated values for emotion classification using the state-of-the-art recognition toolkit Affdex [5]. After each song, participants were asked to rate the song on a scale from 1 (very bad) to 5 (very good) and decide whether they would have skipped the song.

In the second part of the user study, the participants were instructed to use explicit facial expressions to control the music. A between-group design was chosen for this section. 10 of the participants were given the task to skip songs they did not like by making a disgusted face, the other 9 were instructed to continue songs they liked by smiling.

Participants

The study was conducted among 19 volunteers aged 17 - 64 of whom 10 stated they were university students and 9 working as professionals. 10 participants were female, 9 male. All participants were active drivers, 11 of them owned a vehicle of their own, and 16 of them stated they listened to music daily.

Study Tasks

Participants listened to 2 rounds of 30 second snippets from well-known songs. We chose this duration as literature demonstrates immediate emotional responses to musical stimuli within the first seconds [2, 7]. The processing of music against preexisting expectations takes the human brain another single-digit duration of seconds [4]. Thus 30 seconds is an adequate duration for assessing a user's reaction to music.

Implicit Emotion Recognition

In the first part of the study, an emotion recognition software classified the data feed from a frontal camera into values for the dimensions *joy*, *engagement*, *surprise*, and *valence*. These emotions where chosen based on previous observations regarding music-induced emotions and experiences with the used software. In this step the user listened to the songs but did not actively interact with the system except to rate each song after playback.

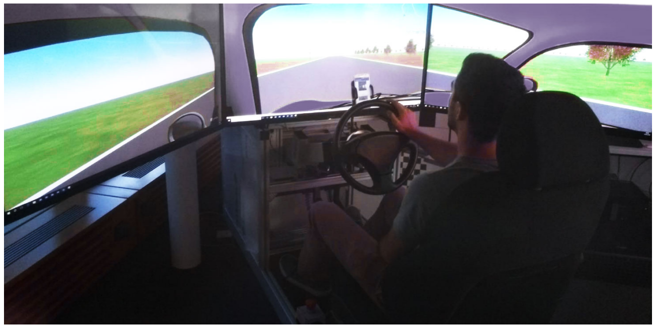

Figure 2: User study setup: participants experienced a driving simulation while they interacted with the face recognition system. A smartphone positioned on top of the steering wheel showed the music player and direct feedback.

Explicit Facial Expressions

In the second part of the study, participants actively controlled the music playback with facial expressions. One group could skip a song if they did not like it by frowning or wrinkling their nose as if they smelled something unpleasant. The other group could prevent a song from automatically skipping by smiling (cf. FM scan mode in common car radios). Each facial gesture was to be held for a dwell time of 2 seconds to be accepted by the system, the dwell time was indicated by a progress bar.

Apparatus

The experiment was conducted in a static medium fidelity driving simulator (System Experience Platform[1]). The simulation consisted of three screens and could be controlled by a steering wheel, a brake pedal and an accelerator pedal. It simulated a racetrack without obstacles. An android device

[1] http://www.ipg-automotive.com/products-services/test-systems/driving-simulators/

was positioned behind the steering wheel, fixated by a cellphone holder. It was placed high enough for the camera to have an unobstructed view, without significantly impairing the user's field of vision (see Figure 2). The recognition system was built as an android application using the *Affectiva Emotion SDK*[2] for the recognition of facial expressions and emotion classification, which was verified on a set of 10,000 non-optimized images [5]. Music playback and control was implemented using the *Spotify Android SDK*[3].

Procedure

The experiment was divided into an short interview, a ca. 15 minute driving part, and a final questionnaire, adding up to a duration of approximately 30 minutes. First a preliminary interview was conducted where general demographics and informed consent were determined. Then participants stepped into the driving simulator and had some time to get accustomed to the controls and the primary driving task. The android application was started and calibrated on the user, whereby participants learned about the operating principle of the facial recognition toolkit in a short live preview.

In the first round, participants listened to 30 seconds snippets of songs and after each were asked to rate the song and indicate whether they would have skipped this song if they could. After this, each participant was assigned to a group for the explicit facial expression assessment. Participants in the first group were told to skip songs they did not like by frowning. The task for the second group was to confirm the playback with a smile if they liked the song, otherwise the song would fade out and skip ahead. Finally, the experimenter led the participants into another room to answer a concluding questionnaire.

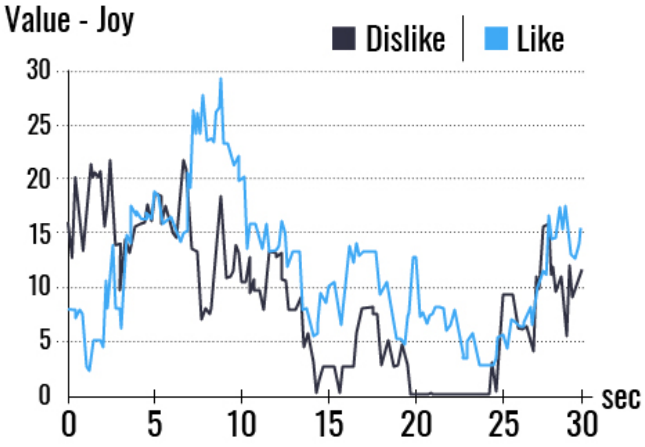

Figure 3: Average joy value of all songs combined, separated into dislike and like group (value range: 0 - 100).

Results

During the first part of the user study, the participants' emotions were tracked. Those values were matched to the scores and answers the participants gave which allowed us to assign them to two groups: A *like*-group, which liked the particular song and would not skip it; and a *dislike*-group which would have preferred to skip the song. The computed mean values of these groups for the joy parameter combining all four songs are displayed in Figure 3. It clearly shows that during the songs the joy value was predominantly higher for the participants who liked the song. However, it is noticeable that at the beginning of the song (up to about 4 seconds) the *dislike*-group had a stronger facial expression of joy.

[2] http://developer.affectiva.com

[3] http://developer.spotify.com

We observed the same pattern for the engagement, valence and surprise values. The mean values over each 30 second track can be seen in Table 1. The differences between the liked and disliked group fluctuate between the different songs. Valence shows the highest mean difference between the groups (3.97 or a 7.2 times increase), but the values for track 3 and especially track 1 show inconsistency. Joy has a mean difference of 2.79, engagement of 3.09, while surprise only has a difference of 0.74 between the groups. Note that the values for all tracks combined are not equal to the mean of each of the track's means. It is the true mean of all data points. Since the face recognition can delay or fail to get an emotion, the songs do not contain the exact same amount of data points and therefore cannot be simply used to compute the mean between them.

The mean values for each song combined are additionally shown in Figure 4. A difference in values can be clearly made out for each emotion, with significances in pairwise comparison for all dimensions but surprise (t-Test, $p < .05$). However, it is important to note that both surprise and valence have rather small values. In contrary to the emotion recognition results, the results of the facial expressions method rely solely on the subjective evaluation of the participants.

Answers from the questionnaires show that participants prefer the smiling gesture over the frowning expression (see Table 2). Smiling was reported to work better than frowning, as it is easier to express and the systems works better at detecting it. Participants were also open to use the smiling expression to control user interfaces in the future. Frowning on the other side was rated more disturbing and users described the gesture as less natural than the smile. A Mann-Whitney U Test found these statements to be significant only for "Would Use" ($p < .05$).

		Joy	Engage	Surprise	Valence
Track1	liked	8.05	16.91	1.77	-0.45
	disliked	5.94	11.23	0.66	0.32
Track2	liked	13.03	20.76	3.81	7.93
	disliked	8.95	17.84	1.84	-2.65
Track3	liked	9.09	20.89	2.35	0.32
	disliked	4.19	10.71	2.16	-0.72
Track4	liked	27.51	32.86	3.60	22.10
	disliked	13.30	23.87	2.14	3.44
All	liked	11.90	20.77	2.68	4.61
	disliked	9.11	17.68	1.94	0.64

Table 1: Mean values of tracked emotions for each song. Values range from 0 to 100, valence from -100 to 100.

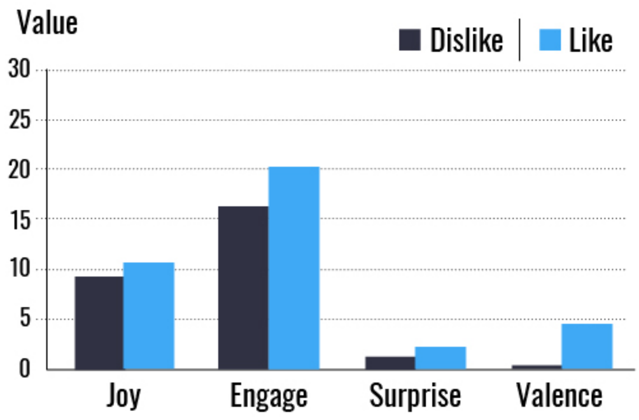

Figure 4: Mean value of the recognized emotions over all four tracks, grouped into the like and dislike group.

	Smiling	**Frowning**
Disturbing	2.44 ± 1.34	3.3 ± 1.42
Worked Well	4.22 ± 1.03	3.2 ± 1.25
Would Use	4.22 ± 0.42	3.1 ± 1.14

Table 2: Mean and standard deviation of user ratings on the questions whether the modality was found to be disturbing, whether they worked well, and if they would use them in the future on a scale from 1 (not at all) to 5 (very much).

Discussion and Future Work

Concluding by the acquired data, it seems to be possible for an intelligent system to track a driver's emotional response to music. However, the differences in values can be very small and highly dependent on the particular song. In addition to the relative margin between the groups (liked/disliked), we need to consider the absolute values. Surprise and valence have overall low values which makes them harder to classify and more prone to errors through outliers. The values of engagement and joy show more reliable results. It is important to note that although for each individual song the mean values of the liked-group were always higher than the disliked-group, the absolute values of the dislike-group can exceed the like-group's values on a different song (see Table 1). This means that there cannot be a simple threshold rule to separate the two groups programmatically.

Another interesting insight is that even disliked songs can trigger positive emotion values in the beginning (see Figure 3). We observed initial smiles for songs which later triggered negative emotional expressions or were voted to skip. Participants accounted this to ironic smiles when they recognized a song they disliked, which goes hand in hand with findings by Abdić et al. who found brief smiles during interactions to correlate with momentary frustration [1]. This potential ambiguity could pose the biggest obstacle for the utilization of facial gestures in automotive user interfaces.

Regarding the results from the participants' questionnaires, the impression arises that using the explicit facial expressions were generally accepted and worked well, but still felt uncomfortable or distracting for some participants (see Table 2). The smiling was clearly preferred to the frowning gesture. Especially the participants with glasses showed difficulties using the latter gesture. Facial gestures might not be more practical than a button press, though they could have a supporting role to conventional modalities.

Thinking about future applications, an implicit emotion recognition could be a good solution as it has zero potential to distraction. The utilization of facial expressions as a fallback could be useful, for instance, to detect the user's dislike of a song. The system could also interact with the driver, who can answer by using facial expressions, instead of everything being controlled autonomously. This combination could decrease the proneness to errors of the emotion recognition and increase the transparency of the system.

All these previously explained results show a potential benefit using face recognition in the car. A next step would be to implement and examine a machine learning approach to autonomously decide over music control using the recognized emotions. We can also imagine a system which asks the user if they want to change the music when lower values of defined emotions are observed. The smiling expression, which was the preferred way of explicit interaction by most participants, could be integrated to interact with the system e.g. to confirm recommendations or to generate playlists based on smiles.

REFERENCES

1. Irman Abdić, Lex Fridman, Daniel McDuff, Erik Marchi, Bryan Reimer, and Björn Schuller. 2016. Driver Frustration Detection from Audio and Video in the Wild. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16)*. AAAI Press, New York, NY, USA, 1354–1360. http://dl.acm.org/citation.cfm?id=3060621.3060809

2. Emmanuel Bigand, Suzanne Filipic, and Philippe Lalitte. 2005. The time course of emotional responses to music. *Annals of the New York Academy of Sciences* 1060, 1 (2005), 429–437. DOI: http://dx.doi.org/10.1196/annals.1360.036

3. Paul Ekman and WV Friesen. 1978. Facial action coding system: A technique for the measurement of facial action. *Manual for the Facial Action Coding System* (1978).

4. Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences* 31, 5 (2008), 559–575. DOI: http://dx.doi.org/10.1017/S0140525X08005293

5. Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 3723–3726. DOI:http://dx.doi.org/10.1145/2851581.2890247

6. C. Muller and G. Weinberg. 2011. Multimodal Input in the Car, Today and Tomorrow. *IEEE MultiMedia* 18, 1 (Jan 2011), 98–103. DOI: http://dx.doi.org/10.1109/MMUL.2011.14

7. Carlos Silva Pereira, João Teixeira, Patrícia Figueiredo, João Xavier, São Luís Castro, and Elvira Brattico. 2011. Music and emotions in the brain: familiarity matters. *PloS one* 6, 11 (2011), e27241. DOI: http://dx.doi.org/10.1371/journal.pone.0027241

8. Florian Roider, Sonja Rümelin, Bastian Pfleging, and Tom Gross. 2017. The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '17)*. ACM, New York, NY, USA, 94–102. DOI: http://dx.doi.org/10.1145/3122986.3122999

9. Christopher D. Wickens, Diane L. Sandry, and Michael Vidulich. 1983. Compatibility and Resource Competition between Modalities of Input, Central Processing, and Output. *Human Factors* 25, 2 (1983), 227–248. DOI: http://dx.doi.org/10.1177/001872088302500209 PMID: 6862451.

10. Walter W. Wierwille. 1993. Demands on driver resources associated with introducing advanced technology into the vehicle. *Transportation Research Part C: Emerging Technologies* 1, 2 (1993), 133 – 142. DOI:http://dx.doi.org/https://doi.org/10.1016/0968-090X(93)90010-D