Using Expressive Avatars to Increase Emotion Recognition: A Pilot Study

Natalie Hube
Mercedes-Benz AG
Stuttgart, Germany
natalie.hube@mercedes-benz.de

Krešimir Vidačković
Hochschule der Medien - University of Applied Science
Stuttgart, Germany
vidackovic@hdm-stuttgart.de

Michael Sedlmair
University of Stuttgart
Stuttgart, Germany
michael.sedlmair@uni-stuttgart.de

ABSTRACT

Virtual avatars are widely used for collaborating in virtual environments. Yet, often these avatars lack expressiveness to determine a state of mind. Prior work has demonstrated effective usage of determining emotions and animated lip movement through analyzing mere audio tracks of spoken words. To provide this information on a virtual avatar, we created a natural audio data set consisting of 17 audio files from which we then extracted the underlying emotion and lip movement. To conduct a pilot study, we developed a prototypical system that displays the extracted visual parameters and then maps them on a virtual avatar while playing the corresponding audio file. We tested the system with 5 participants in two conditions: (i) while seeing the virtual avatar only an audio file was played. (ii) In addition to the audio file, the extracted facial visual parameters were displayed on the virtual avatar. Our results suggest the validity of using additional visual parameters in the avatars' face as it helps to determine emotions. We conclude with a brief discussion on the outcomes and their implications on future work.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Virtual reality**.

KEYWORDS

virtual reality, avatars, emotion, lip synchronization

ACM Reference Format:
Natalie Hube, Krešimir Vidačković, and Michael Sedlmair. 2022. Using Expressive Avatars to Increase Emotion Recognition: A Pilot Study. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3519822

1 INTRODUCTION

Non-verbal communication is an important part of social interaction in addition to spoken words. Humans can grasp a state of mind of another person based on their non-verbal behavior, more precise, body language and facial expression [35]. More and more researchers dedicate their work to examine the influence of the non-verbal communication channel on virtual humans [6, 27, 34], as it goes mostly unattended in current avatar-supported Virtual Reality (VR) collaboration tools [2, 36, 44] which allow body language in a flattened form with full body avatars. Yet, an increasing number of tools already acknowledged the importance of non-verbal facial communication and implement facial expressions or voice mimicking based on different input methods. For example, VRChat [20] uses the user's voice or audio tracks to animate the virtual characters' face and Vive Sync [10] allows using eye tracking and facial tracking [9]. Both applications allow using avatars to meet with other people in an immersive virtual environment.

Our work is limited to interpersonal communication that is not conveyed through verbal language or body posture [21]. More specifically, we address using voice to map facial expressions on avatars. Voice does not only contain important semantic meanings [5, 19], but with verbal pronouncement emotions can be expressed, named and conveyed. Neuroscientists suggest that emotions are essential for human cognition [38]. The majority of our processes of perception, thinking and acting with other human beings is determined through emotions. Thus, the general context of social communication, particularly in collaborative virtual environments (CVE), has an important role in shaping successful communication between virtual participants.

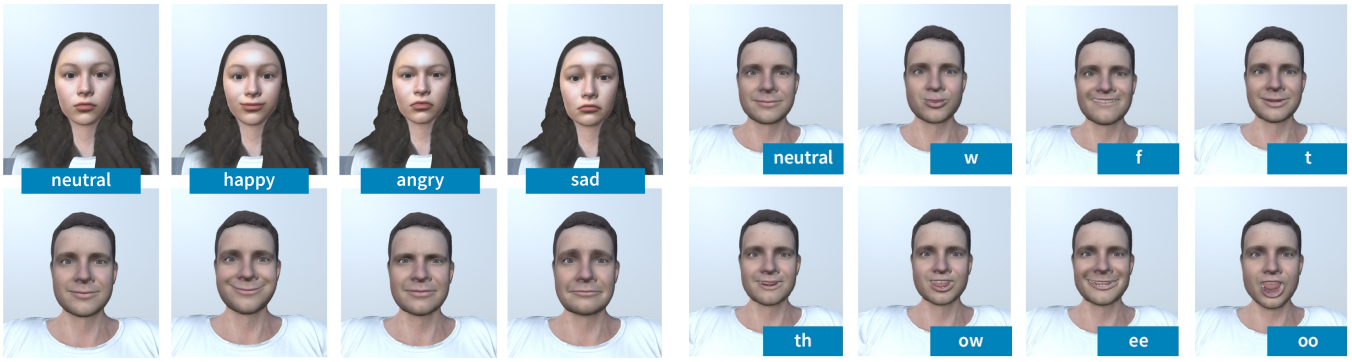
Previous work [7, 23, 24] offers promising approaches to the use of facial expressions on avatars. Building up on these results, in this work, we examine the recognizability of emotions mapped on an avatar through facial expressions extracted from audio tracks in an immersive setting. The animations of an avatars facial expression are inquired by two different factors: the visualization of lip movement and the display of emotions on the face of a virtual character, that can both be extracted from the user's voice or an audio track. To investigate both, we conducted a pilot study with 5 participants in a virtual environment to compare the accuracy and confidence when deciding on emotions. Our goal is to compare whether the use of facial animations makes it easier to determine specific emotions in VR. Our pilot evaluation provides interesting preliminary evidence of the efficacy of using additional facial visual parameters such as emotional expressions and lip synchronization that we want to look closer at in future studies.

2 BACKGROUND & RELATED WORK

When studying virtual avatars and their impact on social interaction and the quality of that interaction, we need to consider different

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04…$15.00
https://doi.org/10.1145/3491101.3519822



(a) (b)

Figure 1: Close-up screenshots of the female and male avatar used during the pilot study. Figure 1a displays the emotions on both avatars: *neutral, happy, angry* and *sad*. Figure 1b shows the most used visemes on the male avatar.

aspects. Besides the degree of avatar realism, embodiment and behavioral realism [37], facial expressions, gaze direction and posture play a decisive role in order to not fall into the uncanny valley [8, 21], which describes the eeriness when looking at a humanly avatar. Movement can have an increasing impact on the uncanny valley [30].

Communication, with regard to verbal and non-verbal expressions, is a critical factor for successful teamwork [41]. Users awareness increases through communicative indicators when collaborating on a virtual platform in the group. Researchers [14, 29] found that presenting users' emotions can even increase the performance within a team, while the lack of it can make it difficult to assess a person's state of mind [18].

2.1 Determining Emotions From Speech

The recognition of emotions in a virtual environment can be achieved through voice [22, 40]. Using voice as acoustic signal can convey information that goes beyond semantics [3]. Paralinguistic features such as tone of voice and intonation, as well as volume, speed, voice quality, hesitation or non-verbal sounds such as sighs and groans are crucial for understanding an emotion [15, 21, 43]. Studies have shown that people may have varying degrees of ability to understand and express emotions through voice. Spackman et al. [40] found that test subjects developed a specific strategy for recognizing emotions, through the characteristic pitches or intensities of a speaking person.

In general, speech recognition system extract important features from the recorded speech signal. These features relate to different emotions within a speech signal, known as feature extraction. If necessary, the set of features can be reduced to a more manageable level at this point. The characteristics of the language are used in classifiers, the core element of artificial intelligence, which are then assigned to specific emotions [39, 42].

In order to classify emotions using algorithms, mathematical models are required which can classify emotions. One methodology for classifying emotions is the discrete emotion theory. It focuses on the cognitive evaluation processes that are necessary to evoke the full spectrum of emotions in adults [28, 32]. Here, one approach is to focus on statistical models and data sets of qualitatively assessable emotions such as anger, happiness and sadness [1]. How emotions are then categorized in number and type, however, varies.

2.2 Lip Synchronization

By examining the speech signal, conclusions can be drawn from the visual information of the lip movement and mouth opening. A synchronized animation on a speaking avatar creates a realistic lip synchronization [33]. The aim of this mapping is to achieve a precise synchronization of a spoken word [25]. For methods of lip synchronization, the decision for the used characteristic to classify can be decisive for the achievable display accuracy. The facial movement of the speaking person creates a specific face image called viseme [16], which is used to classify lip synchronization [4].

For us, the question that arises from the resulting process is, on the one hand, how emotions and speech animation can be adequately identified and what influence the representation of this information on an avatar has in a virtual environment. A general understanding formula has not yet been found in research, which is why numerous features that can be found within an audio track, such as the speed of speech, pitch, intensity still have to be taken into account.

3 PILOT EVALUATION

Virtual avatars in CVEs often lack social interaction. Yet, users are reluctant to use additional hardware, as it makes virtual collaboration more complicated [18], making it necessary to recalibrate devices for each user. One approach to bridge this gap is to transfer a user's facial expression on the avatar that would help to build meaningful interpersonal relations and without the necessity to use additional hardware. We set out to examine if the mapping of facial expressions on a virtual avatar extracted from audio tracks makes it easier to determine a user's emotional state compared to only using the audio tracks together with the virtual avatar. In order to do so, we use our own natural audio data set consisting of 17 audio files.

In our controlled pilot study, we display emotions and lip synchronization on a virtual avatar while playing the corresponding

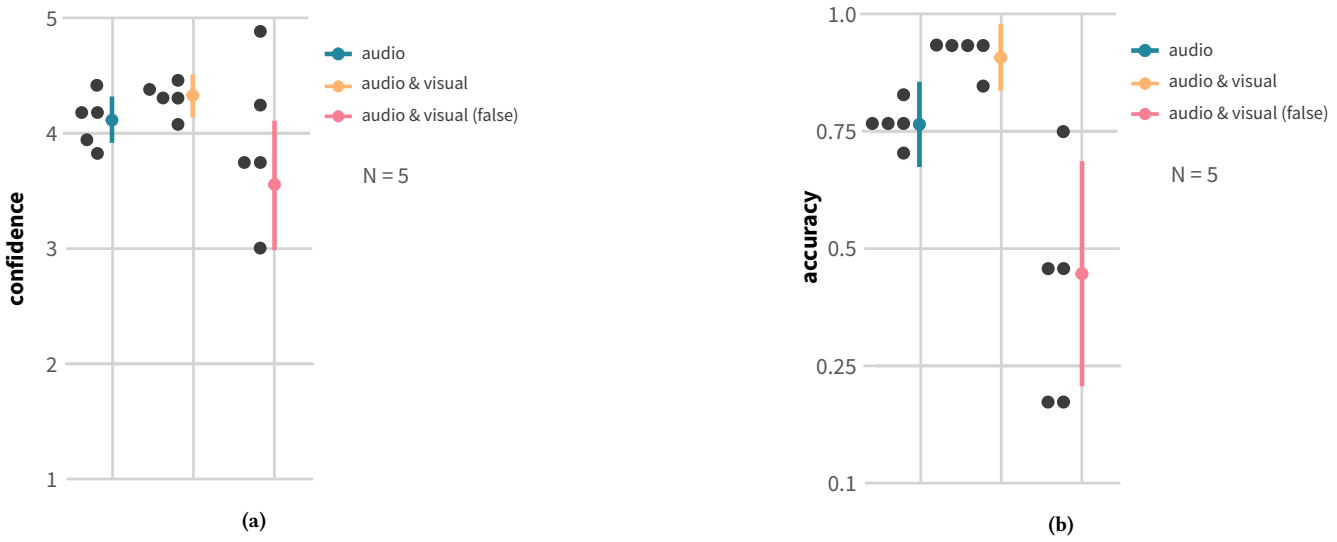


Figure 2: The accuracy (see Figure 2b) and recognition rates (see Figure 2b) are plotted with 95%-confidence intervals for the condition *audio* and *audio & visual*. The results for the falsified emotions' *audio & visual (false)* are plotted in a separated interval. We added data points to the plot to display the individual results for each participant.

audio track. We examine the recognizability of the visual parameters, as well as the accuracy and confidence of each participant. A realistic avatar was chosen to present the visual parameters. Our aim is to provide preliminary evidence of the efficacy of enriching virtual avatars with emotion and lip synchronization to improve social interaction.

3.1 Participants

Due to the ongoing COVID-19 pandemic, we restricted the number of participants and decided to set up a pilot study design for preliminary results. The pilot study involved 5 participants (2 female/3 male) with a mean age of 24.4 years. 80% of the participants are familiar with VR applications, and 2 of 5 participants are working with VR on a regular basis. Three participants stated to have experienced being misunderstood by a peer in a virtual collaboration before due to missing non-verbal cues.

3.2 Setup

The participant is equipped with the VR head-mounted display HTC VIVE Pro. To avoid any bias from artificial, non-reproducible social or behavioral cues such as appearance, postures, facial or gaze displays, participants were put in an empty virtual room where the user is standing in front of an avatar. Depending on the condition, the avatar is underlaid with an audio track or displayed with additional visual parameters, facial emotions and lip synchronization. Our audio tracks have a duration from 3 to 9 seconds. The emotional expressions and lip synchronization were determined through audio tracks in advance to increase comparability between both conditions to anticipate illumination issues, false recognition and performance-wise fps drops. Figure 1 displays the male and female avatars used during the user study.

3.3 Design & Measures

Due to the small sample size, we used a within-subject design to investigate two conditions (audio only, audio & visual) in random order, while wearing a VR head-mounted display. We asked participants to judge the emotional expression of a virtual avatar by choosing between four different emotions. As we used our 17 audio tracks for both conditions, a total of 34 judgements were performed in a random order. The two conditions differ as follows:

- **audio only** The user is presented with a virtual avatar that does not inhabit any facial expression while playing an audio track.
- **audio & visual** The user is presented with a virtual avatar that is displaying the visual facial parameters (emotions & lip synchronization) while playing an audio track.

Three basic emotions (*happiness*, *sadness* and *anger*) were used to represent the emotions on the avatar's face, as these are known and predictable by each human [13]. Additionally, we used a *neutral* expression as our baseline. Each audio track was displayed in random order. Although, the same emotions were used for both conditions, participants did not know how often an emotion appears, therefore avoiding tactical choices. By limiting answer options, we aim to increase comparability as no synonyms appear.

3.3.1 Independent Variables. We considered one independent variable: the animation of visual parameters in the virtual avatars face. The main independent variable distinguishes between audio only and audio with visual parameters. Additionally, we displayed *false* emotions on the avatar in the audio & visual condition during the study to measure its influence on the recognition of emotions. Here, we substituted the emotion displayed on the face with another (see Figure 2). In addition, accuracy and confidence rates were measured.

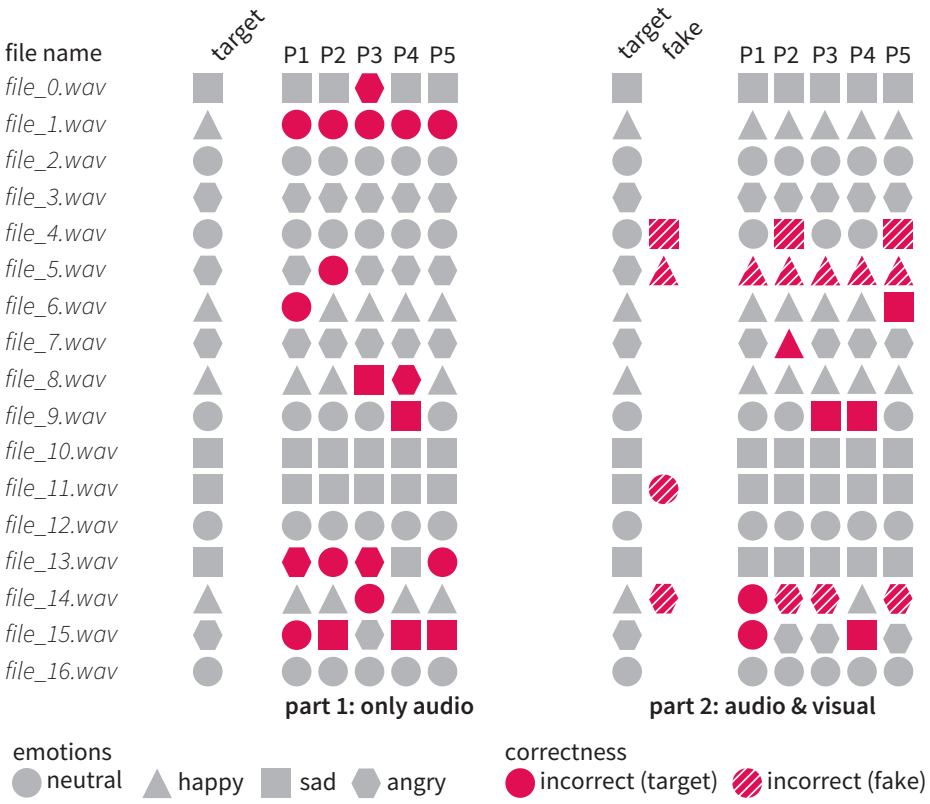


Figure 3: The chart displays emotions selected by each participant during the study as well as the *target* emotion and *fake* emotions in the second set of the study. The wrongly identified emotions are additionally visualized by the color *red* ■. Emotions that were falsely identified when showing the fake emotions are visualized through *white hatching*.

3.3.2 Dependent Variables. The measured data includes the estimation of the recognized emotion. Our questionnaire includes a set of demographic questions, the NASA-TLX [17] and a set of questions regarding the emotion recognition. Additionally, we collected the participants verbal feedback during the study. However, the key question is whether the mapping of additional visual parameters (emotions and lip synchronization) is more suitable to determine emotions in a verbal communication when using a virtual avatar.

3.4 Task & Procedure

Our user study consisted of two steps:

(Step 1) First, we introduced the system and the experiment. Then participants filled out a demographic questionnaire. Before each condition, participants were allowed to adjust to the immersive environment as needed.

(Step 2) The avatar was presented to the participants in a randomized order for each condition and 17 audio tracks representing different emotions. Participants judged each presented facial expression and were asked to rate how confident they were when deciding on the emotion. The study leader then selects the given answers and proceeds with the next expression. Participants were allowed to re-watch the current expression, but had to judge the emotion before the next expression. After each batch of expressions, participants filled out the NASA-TLX. There was a short break between batches.

Table 1: Recognition rates in percentage for each condition without false emotions. Every emotion was recognized better in the *audio + visual* condition, except the baseline emotion *neutral*.

	audio	audio + visual
happy	46.7%	93.3%
angry	73.3%	80.0%
sad	66.7%	100%
neutral	95.0%	90.0%

3.5 Prototype Description

Our developed prototype ran on a Windows 10 computer with 32 GB RAM and a NVIDIA Geforce 1060 GPU. The HTC VIVE Pro was used as the head-mounted display for the immersive environment. In practice, different approaches exist to create lip synchronization. In our prototype, we used SALSA, a real-time lip synchronization framework, which does not support a phoneme or viseme classification, instead uses the dynamic energy of the audio signal using

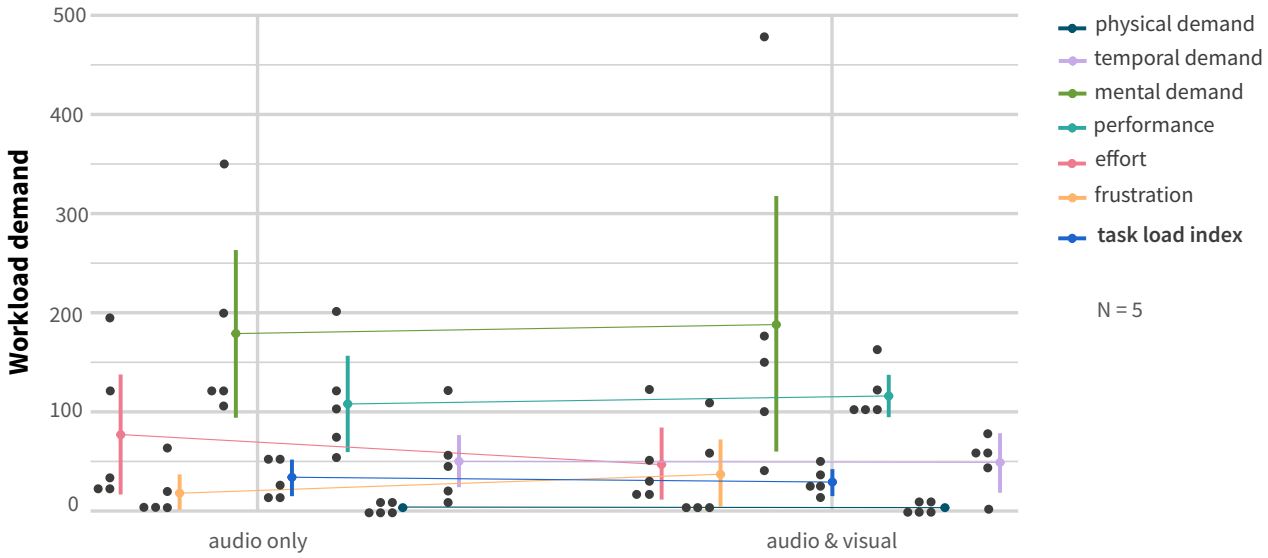


Figure 4: The NASA-TLX measures the perceived workload. The scale is rated within a 500 points rage *(0: very low - 500: very high)* and displayed with 95% confidence intervals. All ratings are combined to the task load index ■. We added data points to the plot to display the individual results for each participant.

an approximation algorithm [11]. To determine the emotional state of our audio data set, we used a framework that chooses the best estimators to run the speech emotion recognition [45].

4 RESULTS

In the following, we present our results including computed means, standard deviations and individual data points. Effect sizes are shown graphically with 95% confidence intervals. The results of our user studies were statistically evaluated using R[1].

4.1 Recognition Rates of Emotions

After presenting an audio track, participants judged the emotion by choosing one of four emotional states (*neutral*, *happy*, *sad* and *angry*). The recognition rate was determined by correctly identified emotions for each condition (see Figure 3). Participants received lower recognition rates for the *audio* condition (M = 70%, SD = 20%) than for the *audio with visual* condition (M = 91%, SD = 8%) (see Figure 2b & Table 1). Additionally, to the recognition rate, we measured the confidence regarding the given answers using a Likert scale raging from 1 - 5 for each condition (see Figure 2a).

4.2 Visualization Quality of Visual Parameters

In a subsidiary questionnaire, we asked participants to rate the visualization aspects of the virtual avatar and visual parameters. On a scale ranging from 1 (*very bad*) to 5 (*very good*) participants rated the lip synchronization with $M = 3$ ($SD = 1.2$). The authenticity of the emotions was rated with $M = 3.4$ ($SD = 1.3$). Opposed to that, participants rated that having a lip synchronization on virtual avatars would make virtual collaboration easier with $M = 4.6$ ($SD = 0.9$) and the importance of having emotionally expressive avatars in a virtual collaboration was rated with $M = 4.4$ ($SD = 0.5$). Regarding the uncanny valley, we asked participants which representation would be more natural to them in a virtual collaborative setting. The majority (4 of 5) found the representation combining lip movements and expressive emotions to be most natural. In addition, participants (4 of 5) found the display of facial animation to be essential for communication between colleagues, although some (3 of 5) also referred to situations of miscommunication in VR from the past.

[1]https://www.r-project.org/

4.3 NASA-TLX

After each condition (*audio* or *audio + visual*), we asked participants to till out a weighted NASA-TLX. Weighted subjective items were then checked individually and plotted on a scale of 0-500 (see Figure 4). In our case, we used a 95%-confidence interval to show tendencies for evaluating pilot study results with small sample groups. Due to the small sample size, the intervals of the plot in Figure 4 is rather high, thus, we added individual data points collected from each participant to show the distribution of results.

5 DISCUSSION

The results of our pilot user study suggest that complementing a virtual avatar with visual facial parameters helps to determine the emotional state of an avatar. Although this finding does not seem surprising, we wanted to provide preliminary evidence as similar studies [7, 23, 24] exist, but none addresses the usage of supporting visual parameters on talking avatars in VR to support further studies. In a recent study, Mukashev et al. [31] already found that including the voice itself enhances the correct recognition of an emotion. Our results show, that based on the *accuracy* of the emotion recognition (see Table 1), it is more likely for a user to determine an emotion correctly, when the virtual avatar is additionally displayed with

a meaningful facial expression and lip synchronization. However, when using false emotions on a corresponding audio track, users tended to misjudge emotions (see Figure 3). With regard to the small sample size, we interpret this tendency as the importance to display emotions truthfully as it might otherwise influence a user's ability to detect an emotion correctly, though this finding needs further investigation pointing us in an interesting direction. The highest recognition rate was determined for the emotions *happy* and *sad*.

The confidence of the participants was also higher for the *audio with visual* condition, which had less variance within results. This could indicate that the interpretation of an emotion according to Spackman et al. [40] is highly person-dependent. The representation of further indicators of the emotion according to Darwin [12], for instance additional representation of facial expression, increases the recognition accuracy, which usually not exist in virtual environments. Our approach uses emotional expressions and speech synchronization as additional visual parameters to increase the recognition accuracy. In a further investigation, we want to examine to what extent further indicators, such as gestures on the body, can improve the recognition of emotions, to have a full set of non-verbal cues.

In addition to a change in accuracy and confidence, we expected that the task load index would decrease due to the representation of the emotion and lip movement. The statistical comparison of the task load index and its sub-categories based on the confidence intervals shows a minimal decrease in the index, which can be primarily related to the decrease in effort. Additionally, we see a slight increase in frustration, which can be related to the fact that participants were frustrated by falsified emotions and less certainty in the classification. As participants did not know that we concealed falsified emotions, we could not collect an isolated NASA-TLX. However, we need to regard these findings with caution due to the small sample size. In further studies we plan on asking more explicit question to differentiate between falsified and correct emotions.

Furthermore, we investigated to what extent participants get confused by falsified emotional expressions on the avatar in their assessment (see Figure 2a and Figure 2b). This investigation shows the importance of correctly extracted emotions and that errors in the classification of an emotion via an AI framework can influence the emotion recognition based on an audio track. From the participants verbal feedback, we found that when the emotions are falsified, the participants relate more to the representation on the face than to the representation in the audio signal. With the expressions *happy* and *angry*, when we switched the emotion, resulted in incorrect recognitions by our participants. We expected errors at the linguistic level, since, as Lugger and Yang [26] also found, that characteristics of speech signals are very similar in *happy* and *angry* emotions. Yet, we need to find options to increase emotion recognition for these specific emotions.

6 CONCLUSION & FUTURE WORK

In this work, we presented the results of a pilot study to examine the influence of facial visual parameters on virtual avatars in VR. Therefore, we extracted emotions and lip synchronization from audio files to then display the information on a virtual avatar in a virtual environment. Our preliminary results suggest that extracting additional non-verbal cues and lip movement may help users to correctly identify emotions compared to mere verbal communication through an audio channel in a virtual setting pointing to further interesting research directions. To create meaningful expressive avatars, we plan to add more non-verbal cues, such as body language to increase identification of a user's state of mind as well as looking into non-verbal features users rely on when determined a specific emotion. However, it should be noted that further testing will be operated with a larger, more diverse population to see if the preliminary results from this work have the desired effects. This is an ongoing project, and work towards the approach mentioned above is currently underway.

ACKNOWLEDGMENTS

We thank Natalie Linya Arnold and Jonas Vogelsang who supported our research and helped to build the prototype. Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21, 4, 1249. https://doi.org/10.3390/s21041249

[2] Autodesk, Inc. 2022. VRED. https://www.autodesk.de/products/vred. Online; accessed 12th January 2022.

[3] Jo-Anne Bachorowski. 1999. Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science* 8, 2 (1999), 53–57. https://doi.org/10.1111/1467-8721.00013

[4] Helen L Bear and Richard Harvey. 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication* 95 (2017), 40–67. https://doi.org/10.1016/j.specom.2017.07.001

[5] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Conf. on Empirical Methods in Natural Language Processing*. 1042–1047. https://doi.org/10.18653/v1/D16-1110

[6] Aleksandra Cerekovic, Oya Aran, and Daniel Gatica-Perez. 2014. How do you like your virtual agent?: Human-agent interaction experience through nonverbal features and personality traits. In *International Workshop on Human Behavior Understanding*. Springer, 1–15. https://doi.org/10.1007/978-3-319-11839-0_1

[7] Saverio Cinieri, Bill Kapralos, Alvaro Uribe-Quevedo, and Fabrizio Lamberti. 2020. Eye Tracking and Speech Driven Human-Avatar Emotion-Based Communication. In *2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 1–5. https://doi.org/10.1109/SeGAH49190.2020.9201874

[8] Tara Collingwoode-Williams, Marco Gillies, Cade McCall, and Xueni Pan. 2017. The effect of lip and arm synchronization on embodiment: A pilot study. In *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 253–254. https://doi.org/10.1109/VR.2017.7892272

[9] HTC Corporation. 2021. VIVE Facial Tracker. https://www.vive.com/us/accessory/facial-tracker/. Online; accessed 12th January 2022 2022.

[10] HTC Corporation. 2021. Vive Sync. https://sync.vive.com/. Online; accessed 12th January 2022 2022.

[11] LLC Crazy Minnow Studio. 2021. SALSA LipSync Suite. https://crazyminnowstudio.com/docs/salsa-lip-sync/. Online; accessed 20th December 2021.

[12] Charles Darwin. 1872. *The Expression of the Emotions in Man and Animals by Charles Darwin.* John Murray.

[13] Paul Ekman and Erika L Rosenberg. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS).* Oxford University Press, USA. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

[14] Ulises Xolocotzin Eligio, Shaaron E Ainsworth, and Charles K Crook. 2012. Emotion understanding and performance during computer-supported collaboration. *Computers in Human Behavior* 28, 6, 2046–2054. https://doi.org/doi.org/10.1016/j.chb.2012.06.001

[15] Ivan Fonagy and Klara Magdics. 1963. Emotional patterns in intonation and music. *STUF-Language Typology and Universals* 16, 1-4, 293–326. https://doi.org/10.1524/stuf.1963.16.14.293

[16] Mahesh Goyani, Namrata Dave, and NM Patel. 2010. Performance analysis of lip synchronization using LPC, MFCC and PLP speech parameters. In *Int. Conf. on Computational Intelligence and Communication Networks*. IEEE, 582–587. https://doi.org/10.1109/CICN.2010.115

[17] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[18] Natalie Hube, Katrin Angerbauer, Daniel Pohlandt, Krešimir Vidačković, and Michael Sedlmair. 2021. VR Collaboration in Large Companies: An Interview Study on the Role of Avatars. In *Proc. IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 139–144. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00037

[19] Natalie Hube, Oliver Lenz, Lars Engeln, Rainer Groh, and Michael Sedlmair. 2020. Comparing Methods for Mapping Facial Expressions to Enhance Immersive Collaboration with Signs of Emotion. In *Proc. IEEE Int. Symp. on Mixed and Augmented Reality (ISMAR)*. IEEE, 30–35. https://doi.org/10.1109/ISMAR-Adjunct51615.2020.00023

[20] VRChart Inc. 2022. VRChart. https://hello.vrchat.com/. Online; accessed 12th January 2022 2022.

[21] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal Communication in Human Interaction*. Wadsworth Cengage Learning.

[22] Petri Laukka, Patrik Juslin, and Roberto Bresin. 2005. A dimensional approach to vocal expression of emotion. *Cognition & Emotion* 19, 5 (2005), 633–653. https://doi.org/10.1080/02699930441000445

[23] Jieun Lee, Jeongyun Heo, Hayeong Kim, and Sanghoon Jeong. 2021. Fostering Empathy and Privacy: The Effect of Using Expressive Avatars for Remote Communication. In *International Conference on Human-Computer Interaction*. Springer, 566–583. https://doi.org/10.1007/978-3-642-15892-6_8

[24] Sangyoon Lee, Gordon Carlson, Steve Jones, Andrew Johnson, Jason Leigh, and Luc Renambot. 2010. Designing an expressive avatar of a real person. In *International Conference on Intelligent Virtual Agents*. Springer, 64–76.

[25] Wentao Liu, Baocai Yin, Xibin Jia, and Dehui Kong. 2004. Audio to visual signal mappings with HMM. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, Vol. 5. IEEE, 885 – 888. https://doi.org/10.1109/ICASSP.2004.1327253

[26] Marko Lugger and Bin Yang. 2007. The relevance of voice quality features in speaker independent emotion recognition. In *Proc. Int. Conf. on Acoustics, Speech & Signal Processing*, Vol. 4. IEEE, IV–17. https://doi.org/10.1109/ICASSP.2007.367152

[27] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. 2020. "Talking without a Voice" Understanding Non-verbal Communication in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25. https://doi.org/10.1145/3415246

[28] Javier Marín-Morales, Carmen Llinares, Jaime Guixeres, and Mariano Alcañiz. 2020. Emotion recognition in immersive virtual reality: From statistics to affective computing. *Sensors* 20, 18, 5163. https://doi.org/10.3390/s20185163

[29] Gaëlle Molinari, Guillaume Chanel, Mireille Betrancourt, Thierry Pun, and Christelle Bozelle Giroud. 2013. Emotion feedback during computer-mediated collaboration: Effects on self-reported emotions and perceived interaction. In *To see the World and a Grain of Sand: Learning across Levels of Space, Time, and Scale*. International Society of the Learning Sciences.

[30] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2, 98–100. https://doi.org/10.1109/MRA.2012.2192811

[31] Dinmukhamed Mukashev, Merey Kairgaliyev, Ulugbek Alibekov, Nurziya Oralbayeva, and Anara Sandygulova. 2021. Facial expression generation of 3D avatar based on semantic analysis. In *2021 30th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*. 89–94. https://doi.org/10.1109/RO-MAN50785.2021.9515463

[32] Andrew Ortony, Gerald L Clore, and Allan Collins. 1990. *The Cognitive Structure of Emotions*. Cambridge University Press.

[33] Junho Park and Hanseok Ko. 2008. Real-time continuous phoneme recognition system using class-dependent tied-mixture hmm with hbt structure for speech-driven lip-sync. *IEEE Transactions on Multimedia* 10, 7, 1299–1306. https://doi.org/10.1109/TMM.2008.2004908

[34] Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Communication* 51, 7 (2009), 630–639. https://doi.org/10.1016/j.specom.2008.04.009

[35] Deepika Phutela. 2015. The importance of non-verbal communication. *IUP Journal of Soft Skills* 9, 4 (2015), 43.

[36] Protics. 2022. Engineering Hub. https://www.daimler-protics.com/landing-pages/index-2.html. Online; accessed 12th January 2022 2022.

[37] Daniel Roth, Jean-Luc Lugrin, Dmitri Galakhov, Arvid Hofmann, Gary Bente, Marc Erich Latoschik, and Arnulph Fuhrmann. 2016. Avatar realism and social interaction quality in virtual reality. In *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 277–278. https://doi.org/10.1109/VR.2016.7504761

[38] Monika Schwarz-Friesel. 2012. On the status of external evidence in the theories of cognitive linguistics: compatibility problems or signs of stagnation in the field? Or: why do some linguists behave like Fodor's input systems? *Language Sciences* 34, 6, 656–664. https://doi.org/10.1016/j.langsci.2012.04.007

[39] Akash Shaw, Rohan Kumar Vardhan, and Siddharth Saxena. 2016. Emotion recognition and classification in speech using artificial neural networks. *International Journal of Computer Applications* 145, 8, 5–9. https://doi.org/10.5120/ijca2016910710

[40] Matthew P Spackman, Bruce L Brown, and Sean Otto. 2009. Do emotions have distinct vocal profiles? A study of idiographic patterns of expression. *Cognition and Emotion* 23, 8, 1565–1588. https://doi.org/10.1080/02699930802536268

[41] Pina Tarricone and Joseph Luca. 2002. Successful Teamwork: A Case Study. *Higher Education Research and Development Society of Australasia*.

[42] A Tickle, S Raghu, and Mark Elshaw. 2013. Emotional recognition from the speech signal for a virtual education agent. In *Journal of Physics: Conference Series*, Vol. 450. IOP Publishing, 012053. https://doi.org/10.1088/1742-6596/450/1/012053

[43] Gary R VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association.

[44] Virbela. 2021. Virbela: A Virtual World for Work, Education & Events. https://www.virbela.com/. Online; accessed 12th January 2022 2022.

[45] x4nth055. 2022. Building and training Speech Emotion Recognizer that predicts human emotions using Python, Sci-kit learn and Keras. https://github.com/x4nth055/emotion-recognition-using-speech. Online; accessed 12th January 2022 2022.