Computer-Based PTSD Assessment in VR Exposure Therapy

Leili Tavabi, Anna Poon, Albert Skip Rizzo, and Mohammad Soleymani(✉)

Institute for Creative Technologies, University of Southern California,
12015 Waterfront Drive, Los Angeles, CA 90094, USA
{ltavabi,soleymani}@ict.usc.edu
http://ict.usc.edu

Abstract. Post-traumatic stress disorder (PTSD) is a mental health condition affecting people who experienced a traumatic event. In addition to the clinical diagnostic criteria for PTSD, behavioral changes in voice, language, facial expression and head movement may occur. In this paper, we demonstrate how a machine learning model trained on a general population with self-reported PTSD scores can be used to provide behavioral metrics that could enhance the accuracy of the clinical diagnosis with patients. Both datasets were collected from a clinical interview conducted by a virtual agent (SimSensei) [10]. The clinical data was recorded from PTSD patients, who were victims of sexual assault, undergoing a VR exposure therapy. A recurrent neural network was trained on verbal, visual and vocal features to recognize PTSD, according to self-reported PCL-C scores [4]. We then performed decision fusion to fuse three modalities to recognize PTSD in patients with a clinical diagnosis, achieving an F1-score of 0.85. Our analysis demonstrates that machine-based PTSD assessment with self-reported PTSD scores can generalize across different groups and be deployed to assist diagnosis of PTSD.

Keywords: Post-traumatic stress disorder · Virtual reality · Exposure therapy · Human behavior · Multimodal machine learning

1 Introduction

Post-traumatic stress disorder (PTSD) is a psychiatric disorder that can occur in people who have experienced or witnessed a traumatic event such as a natural disaster, a serious accident, a terrorist act, war/combat, rape or other violent personal assault [19]. Despite the growing number of people experiencing mental health disorders including PTSD, there is still a large gap between available clinical resources and the need for accurate assessment, detection, and treatment. This problem highlights the importance for devising automated methods for assessing mental health disorders in order to augment clinical resources. As a result, there is growing interest in using automatic human behavior analysis for

C. Stephanidis et al. (Eds.): HCII 2020, LNCS 12428, pp. 440–449, 2020.
https://doi.org/10.1007/978-3-030-59990-4_32

computer-aided mental health diagnosis. Such computer-based diagnosis methods are based on the detection, quantification, and analysis of behavioral cues such as verbal content, facial and body expressions and speech prosody [27].

In this work, we trained a machine learning model for PTSD assessment by leveraging data collected during semi-structured interviews between a Virtual Human (VH) agent and subjects from the general population. We then evaluate the model on the clinical data collected from PTSD patients as a result of sexual trauma. We demonstrate how the weak labels that are based on self-reported questionnaire assessments can be used to train a classifier to inform the diagnosis of PTSD due to sexual trauma using behavioral data from patients. During the test phase, we evaluate our model's performance on real patients with expert clinical diagnosis. To this end, we use verbal and nonverbal data obtained from human-VH agent interactions during semi-structured interviews with probing cues relevant to mental health disorders. The semi-structured interviews consist of either a wizard-controlled or AI-controlled agent, asking a set of predefined questions along with follow-up questions and continuation prompts from the VH depending on the user's responses. We use verbal, speech and visual data from the user during the interactions for performing machine learning with the goal of PTSD detection.

The remainder of this paper is organized as follows. The related work on computational analysis of PTSD is given in Sect. 2. The datasets are described in Sect. 3. The feature extraction and machine learning methods are given in Sect. 4. Experimental results are presented and discussed in Sect. 5. Finally, the work is concluded in Sect. 6.

2 Related Work

Verbal and nonverbal communicative behaviors have been shown to be effective indicators for assessing mental health disorders [7,13,25]. Previous studies indicate that mental health disorders can result in consistent changes in patterns of behavior, including reduced vocal pitch, slower speaking rate or less intense smiles [9,23].

Compared to automated depression diagnosis, for example [1,16,18], there have been fewer studies on automatic detection of PTSD. Recent research with this purpose has leveraged the data recorded during interaction with the VH agent interviewer that is part of the SimSensei Kiosk [21]. Stratou *et al.* [27], analyzed the association between behavioral responses related to affect, expression variability, motor variability and PTSD. They found that correlations were gender dependent. For example, expression of disgust was higher in men with PTSD while the opposite was true for women. These findings motivated additional attention investigating gender-based PTSD recognition, which resulted in higher accuracy compared to the gender-independent models. Scherer *et al.* [22] reported that subjects with PTSD and depression had a significantly lower vowel space in their speech. In a previous study Scherer *et al.* [24], investigating voice quality features, identified that PTSD patients spoke in a more tense voice.

More recently, Marmar *et al.* [17] studied audio features for automatic recognition of PTSD using a dataset of patients' voices, recorded during psychotherapy sessions, with PTSD patients without comorbidity. After analyzing audio descriptors they found that PTSD patients spoke more slowly in a more affectively flat, monotonous tone with less activation.

3 Data

We used two datasets consisting of clinical interviews from research participants and a VH agent interviewer. The first dataset is the Extended Distress Analysis Interview Corpus (E-DAIC), a subset of the larger DAIC database [12], which contains data captured during clinical interviews with the SimSensei Kiosk VH agent, referred to as "Ellie". The second dataset also includes similar interviews with Ellie, recorded from patients who participated in a virtual reality (VR) exposure therapy program to address PTSD due to military sexual trauma (MST) [15].

3.1 Distress Analysis Interview Corpus

The Distress Analysis Interview Corpus is an audiovisual dataset of human-agent interactions captured during a VH (Ellie) clinical interview [12]. The interviews were designed to investigate the occurrence of user's behavioral signals that were hypothesized to be related to psychological distress conditions such as anxiety, depression, and PTSD. DAIC was recorded using SimSensei Kiosk [21], an interactive system with multimodal behavior recognition and generation components, including dialogue management, body tracking, facial expression analysis, agent visualization, and behavior realization. In this paper, we use a portion of DAIC dataset, *i.e.*, E-DAIC, that was used in the Audio/Visual Emotion Challenge and Workshop (AVEC 2019), depression detection sub-challenge [20]. The Extended Distress Analysis Interview Corpus (E-DAIC) [10] is an extended version of a dataset developed in former AVEC challenges, called DAIC-WoZ [12].

For the purpose of the challenge, the E-DAIC dataset was partitioned into training, development, and test sets while preserving the overall speaker diversity – in terms of age, gender distribution, and the eight-item Patient Health Questionnaire (PHQ-8) scores – within the partitions. Whereas the training and development sets include a mix of WoZ and AI scenarios, the test set is solely constituted from the data collected by the autonomous AI. E-DAIC data also includes PCL-C (PTSD Checklist-Civilian version) self-reported questionnaire scores that are indicative of PTSD [4]. The E-DAIC dataset includes recordings from 275 subject, from the general population (Fig. 1).

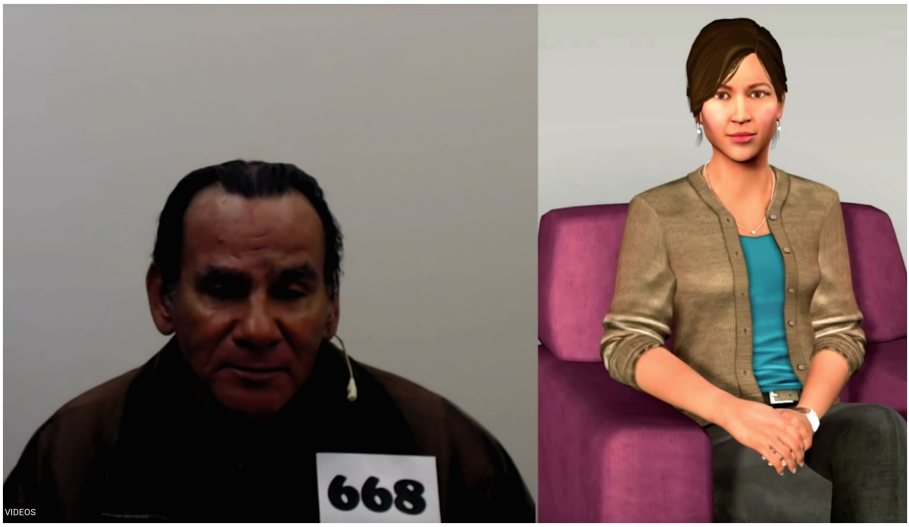

Fig. 1. A participant and the virtual agent, Ellie, during a clinical interview.

3.2 PTSD Patients

The data used for evaluation consists of verbal and nonverbal data collected from PTSD patients participating in a clinical trial investigating the safety and efficacy of VR exposure therapy for PTSD due to MST [14]. The patients are all military veterans who were victims of sexual assault. Patients went through the same interview as in DCAPS with the same virtual agent (Fig. 2).

For more information regarding the VR exposure therapy, experimental protocol and details of clinical intervention, we refer the reader to [14].

4 Method

We analyzed three modalities for automatic behavioral assessment of PTSD, namely, language (verbal content), vision (face and head behavior) and voice (speech prosody). A machine learning model was trained to recognize PTSD from each modality and their results were fused at decision-level to perform multimodal classification.

Textual Features. The textual data is transcribed from the recorded audio and using the Google Cloud's speech recognition service. Google's pretrained Universal Sentence Encoder [5] is then used to extract sentence embeddings (feature vectors) of size 512. We input every speech turn from the participant to the encoder and as a result, a sequence of $N \times 512$ vectors was generated for each participant, where N is the number of subject's speech turns.

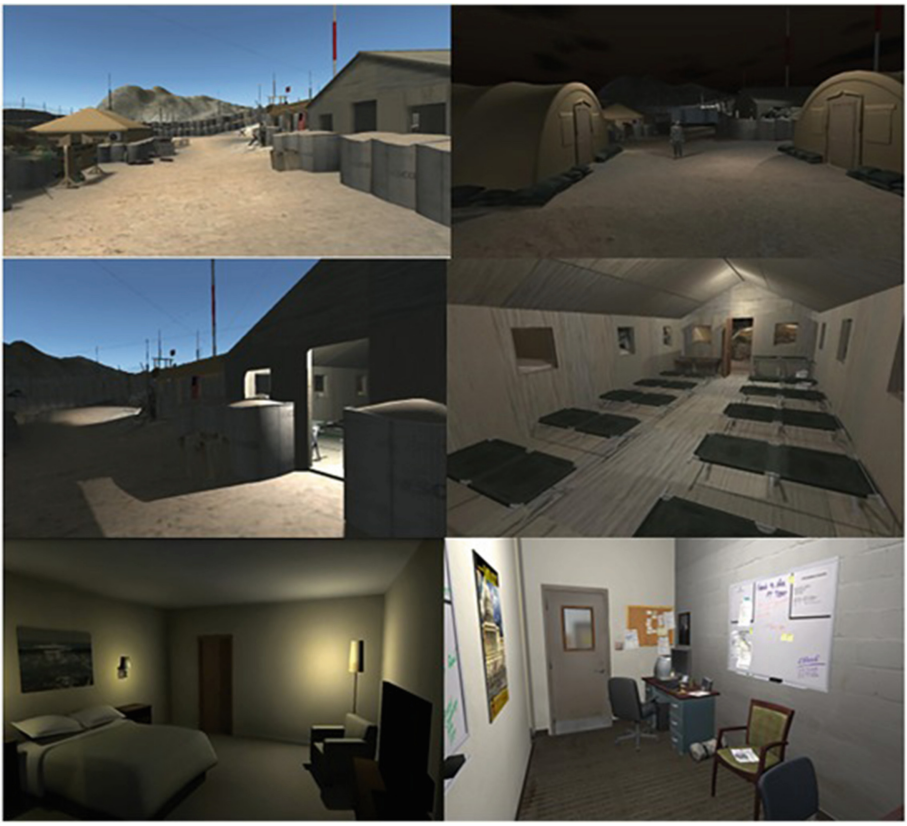

Fig. 2. Snapshots of content in the Bravemind Military Sexual Trauma (MST) exposure therapy.

Audio Features. We use VGG-16 Deep Spectrum[1] features as our audio feature representation [2]. These features are inspired by deep representation learning paradigms common in image processing: spectrogram images of speech frames are fed into pretrained CNNs and the resulting activations are extracted as feature vectors.

To extract deep spectrum features, the audio signals were transformed to mel-spectrogram images with 128 mel-frequency bands with a temporal window of 4 s and hop size of 1 s.

This feature set is extracted from a VGG-16 pretrained CNNs [26] to obtain a sequential representation of vector size 4096. As a result, for each participant, we obtain a sequence of $T \times 4096d$ vectors as speech features, where T is the duration in seconds.

[1] https://github.com/DeepSpectrum/DeepSpectrum.

Visual Features. For the visual features, we extract action units and head pose angles per frame using OpenFace [3], a computer-vision based open-source software for head pose tracking and facial action unit detection. OpenFace is used to extract the intensity of 17 facial action units (FAU), based on the Facial Action Coding System (FACS) [11] along with 3d head pose angles per frame, therefore providing a $25 \times T \times 20$ representation, where T is the duration in seconds and 25 is the video frame rate.

4.1 Model Architecture

Our multimodal learning model relies on fusing the three mentioned modalities, namely, vision (facial expressions and head pose), speech prosody and verbal content for PTSD recognition. We use deep encoders to map voice and verbal modalities to embeddings. Visual features were extracted per-frame by OpenFace [3]. Sequences of features for all three modalities were fed to a one-layer recurrent neural network followed by a linear layer for classification. The results of unimodal classifiers are in turn fused for multimodal PTSD recognition.

Since all the embeddings retain the temporal dimension of the modalities, we use a single-layer Gated Recurrent Unit (GRU) that maps each representation to a fixed-size embedding, keeping only the last state. GRU is a variant of recurrent units able to capture long and short term dependencies in sequences [6]. The dimensionality of the hidden layer is adjusted based on the size of the representation for each modality. Therefore the input feature space for the text, audio and visual representations are mapped to 128-d, 256-d and 8-d embedding space respectively. The resulting embedding space is then mapped to a single output using a linear layer. We use a sigmoid activation function for the last layer and train the model with a binary cross entropy loss. To alleviate the class imbalance problem in E-DAIC (large number of healthy vs PTSD), we use a weighted loss whose weights for each class is inversely proportional to the number of samples in that class, in each training batch.

5 Experimental Results

The neural networks are implemented in PyTorch. The network training is optimized using Adam, with a batch size of 15 and a learning rate of 10^{-4}. The models are trained for 80 epochs and the best performing model on the validation set is selected. The evaluation results are computed using the area under the curve (AUC) of the Receiver Operating Characteristics (ROC) curve. The training and validation sets are fixed and obtained from the E-DAIC dataset, consisting of 219 and 56 subjects respectively. The test set is from the VR exposure therapy dataset and consists of 14 sessions (from seven unique patients). The results of the binary classification from each modality are fused together using a majority-vote late fusion to obtain the multimodal fusion results.

PTSD recognition is evaluated using AUC of ROC curves and F1-score. F1-score is a harmonic mean of recall and precision scores. The recognition

performances are given in Table 1. Unimodal results demonstrate that features extracted from nonverbal behavior, including head movement, facial action units and speech prosody perform slightly better than verbal features for this task. Moreover, multimodal fusion achieve superior results compared to unimodal results.

The recognition performance with a relatively simple machine learning pipeline is promising. The results demonstrate how the machine learning models trained on a limited experimental dataset, from the general population and with self-reported scores can generalize to a different population with clinically validated labels.

Table 1. Recognition performance for multimodal PTSD classification on clinical data. AUC stands for area under curve for ROC.

Modality	Features	AUC	F1-score
Verbal	USE	0.53	0.67
Vocal	DS-VGG	0.68	0.67
Visual	AU+Pose	0.54	0.76
Multimodal fusion	-	0.70	0.85

Even though the obtained results are fairly accurate, there is still room for improvement. Marmar *et al.* [17] argued that since a large number of PTSD patients are also diagnosed with other mental health disorders, such as depression, that can alter behavior, the training samples should not be from comorbid patients. We did not follow this approach, since we did not have a large enough sample to do so. The data recorded form patients were also of much lower quality, as both video and audio were captured with a camcorder, as opposed to high quality webcams and wearable microphones, in DAIC. We believe this might have also had a negative effect on the performance.

6 Conclusions

In this paper, we reported on our efforts for detecting behaviors relating to PTSD using verbal and audiovisual data captured during a clinical interview between a patient and a VH agent. Our analysis shows that our model trained on a diverse population can be applied to a population of actual patients for automatic clinical behavioral assessment to inform diagnosis. We trained and evaluated PTSD recognition models on verbal, audio and visual modalities. The results show that audio has the highest ROC score, outperforming verbal and visual modalities and the visual modality appears to have a higher recall on PTSD patients. Late fusion of the three modalities results in improved F1-score of the PTSD patients, as well as the AUC score.

In the future, we will perform domain adaptation [8] to reduce the effect of using different types of recording apparatus. Training gender-based models, such as the one proposed in [27], and training samples without comorbidity, as proposed in [17] will be also explored in our future work.

Automated mental health assessment methods have the potential to augment and improve mental health care. We hope the advancement of such technologies combined with novel treatment methods, such as VR exposure therapy, improves access and quality of care for mental health disorders.

Acknowledgments. The work of Poon was supported by the National Science Foundation under award 1852583, "REU Site: Research in Interactive Virtual Experiences" (PI: Ron Artstein). This work was supported in part by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Breakspear, M., Parker, G.: Detecting depression: a comparison between spontaneous and read speech. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7547–7551. IEEE (2013)
2. Amiriparian, S., et al.: Snore sound classification using image-based deep spectrum features. In: Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, pp. 3512–3516. ISCA, Stockholm, Sweden, August 2017
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 59–66. IEEE (2018)
4. Blanchard, E.B., Jones-Alexander, J., Buckley, T.C., Forneris, C.A.: Psychometric properties of the PTSD checklist (PCL). Behaviour Res. Therapy **34**(8), 669–673 (1996)
5. Cer, D., et al.: Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018)
6. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
7. Cohn, J.F., et al.: Detecting depression from facial actions and vocal prosody. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. IEEE, Amsterdam, Netherlands (2009). 7 pages
8. Csurka, G.: Domain adaptation for visual applications: a comprehensive survey. arXiv preprint arXiv:1702.05374 (2017)
9. Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F.: A review of depression and suicide risk assessment using speech analysis. Speech Commun. **71**, 10–49 (2015)
10. DeVault, D., et al.: SimSensei Kiosk: a virtual human interviewer for healthcare decision support. In: Proceeding of the International Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 201, pp. 1061–1068. ACM, Paris, France (2014)
11. Ekman, P., Friesen, W.: The Facial Action Coding System (FACS). Consulting Psychologists Press, Stanford University, Palo Alto (1978)

12. Gratch, J., et al.: The Distress Analysis Interview Corpus of human and computer interviews. In: Proceeding of the 9th International Conference on Language Resources and Evaluation, LREC 2014, pp. 3123–3128. ELRA, Reykjavik, Iceland, May 2014
13. Joshi, J., et al.: Multimodal assistive technologies for depression diagnosis and monitoring. J. Multimodal User Interfaces **7**(3), 217–228 (2013). https://doi.org/10.1007/s12193-013-0123-2
14. Loucks, L., et al.: You can do that?!: Feasibility of virtual reality exposure therapy in the treatment of PTSD due to military sexual trauma. J. Anxiety Disorders **61**, 55 – 63 (2019). https://doi.org/10.1016/j.janxdis.2018.06.004, http://www.sciencedirect.com/science/article/pii/S0887618517304991, virtual reality applications for the anxiety disorders
15. Loucks, L., et al.: You can do that?!: Feasibility of virtual reality exposure therapy in the treatment of PTSD due to military sexual trauma. J. Anxiety Disorders **61**, 55–63 (2019)
16. Low, L.S.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B.: Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans. Biomed. Eng. **58**(3), 574–586 (2010)
17. Marmar, C.R., Brown, A.D., Qian, M., Laska, E., Siegel, C., Li, M., Abu-Amara,D., Tsiartas, A., Richey, C., Smith, J., Knoth, B., Vergyri, D.: Speech-based markers for posttraumatic stress disorder in us veterans. Depression Anxiety **36**(7), 607–616 (2019). https://doi.org/10.1002/da.22890, https://onlinelibrary.wiley.com/doi/abs/10.1002/da.22890
18. Meng, H., Huang, D., Wang, H., Yang, H., Ai-Shuraifi, M., Wang, Y.: Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. pp. 21–30 (2013)
19. Organization, W.H., et al.: Guidelines for the management of conditions that are specifically related to stress. World Health Organization (2013)
20. Ringeval, F., et al.: Avec 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. AVEC 2019, pp. 3–12. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3347320.3357688
21. Rizzo, A., et al.: Detection and computational analysis of psychological signals using a virtual human interviewing agent. J. Pain Manag. **9**, 311–321 (2016)
22. Scherer, S., Lucas, G.M., Gratch, J., Rizzo, A.S., Morency, L.P.: Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews. IEEE Trans. Affect. Comput. **7**(1), 59–73 (2015)
23. Scherer, S., et al.: Automatic behavior descriptors for psychological disorder analysis. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face & Gesture Recognition (FG). IEEE, Shanghai, P. R. China, April 2013. 8 pages
24. Scherer, S., Stratou, G., Gratch, J., Morency, L.P.: Investigating voice quality as a speaker-independent indicator of depression and PTSD. In: Interspeech, pp. 847–851 (2013)
25. Scherer, S., et al.: Automatic audiovisual behavior descriptors for psychological disorder analysis. Image Vis. Comput. **32**(10), 648–658 (2014)

26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556 (2014). 14 pages
27. Stratou, G., Scherer, S., Gratch, J., Morency, L.: Automatic nonverbal behavior indicators of depression and PTSD: exploring gender differences. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 147–152, September 2013. https://doi.org/10.1109/ACII.2013.31