INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The effect of clinical intervention on the speech of individuals with PTSD: features and recognition performances

Alexander Kathan[1], *Andreas Triantafyllopoulos*[1], *Shahin Amiriparian*[1], *Sabrina Milkus*[2], *Alexander Gebhard*[1], *Jonas Hohmann*[2], *Pauline Muderlak*[2], *Jürgen Schottdorf*[3], *Björn W. Schuller*[1,4], *Richard Musil*[2]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Germany
[3]Zentrumspraxis Friedberg, Germany
[4]Group on Language, Audio & Music, Imperial College London, UK

alexander.kathan@uni-a.de

Abstract

Post-traumatic stress disorder (PTSD) is an anxiety disorder that can occur as a response to traumatic experiences, such as catastrophic events, and can have a detrimental influence on mental wellbeing. Furthermore, PTSD is present in $5-10\,\%$ of the population, making it a prevalent disorder in our time, thus necessitating a timely diagnosis and proper treatment. In this paper, we present results for PTSD detection based on speech recordings on a newly collected dataset consisting of 15 participants, including speakers with PTSD and a control group. Moreover, the dataset includes speech data immediately before and after a clinical intervention (i.e., acupuncture-supported psychotherapy), allowing us to examine the effect of a single treatment session. In our experiments, we achieve a best area under the curve (AUC) of 82 % using solely pre-treatment data. Finally, we analyse prominent acoustic patterns of individuals with PTSD compared to the control group.

Index Terms: digital health, PTSD, machine learning

1. Introduction

Worldwide more than 70 % of all people encounter at least one traumatic experience in their lifetime, and up to 31 % face four or more of them [1]. Examples of such events might be life-threatening accidents, experienced violence, abuse, terrorist attacks, war, or natural disasters [1, 2]. To make matters worse, these experiences can have lasting effects on individuals, resulting in a Post-traumatic Stress Disorder (PTSD), which is the fourth most prevalent mental disorder of our time [3]. Moreover, people exposed to such events also have a higher likelihood of developing other psychiatric disorders, such as major depression [3]. PTSD occurs in $5-10\,\%$ of people and can have a detrimental influence on mental wellbeing, while occurring twice as often in women than in men [4]. PTSD encompasses numerous symptoms, ranging from bad memories and nightmares to a constant preoccupation with the traumatic experience. In addition, PTSD can also be associated with social distancing (e. g., by avoiding places or activities that might evoke memories of the experience). All in all, PTSD is a serious mental illness with a heavy burden for both the person suffering from it as well as for their social environment, making it an emerging issue to be detected and resolved as early as possible [4, 5].

To this end, previous work addressed the challenge of detecting PTSD using various approaches and data modalities, enabling an earlier detection and therefore an earlier clinical intervention [6, 7, 8, 9]. For determining the PTSD diagnosis status, typically self-reporting questionnaires as well as structured clinical interviews, such as the Clinician-Administered PTSD Scale (CAPS) interview are used [5, 10]. However, these interviews are time-consuming and require appropriate psychological experts [11]. To cope with the increasing need, approaches are necessary that allow for a time-efficient PTSD detection. With the rise of machine learning (ML) methods in recent years, previous work explored ML-based approaches for automatically recognising mental disorders (e. g., [12]), enabling an earlier intervention with improved outcomes to a broader population.

Lekkas et al. [6] experimented with passively collected smartphone data for PTSD detection. In doing so, they utilised features derived from GPS data, such as the daily time spent away or the maximum travelled distance. In their experiments, they achieve an area under the curve (AUC) of 81.6 %, making GPS features a promising digital biomarker for PTSD detection [6]. Scherer et al. [7] explored audio-based features as indicators for depression and PTSD. In their experiments, they yield a performance between 52 % - 72 % using only four acoustic features that capture the tenseness of speakers' voice [7]. Vergyri et al. [11] used an extended acoustic feature set consisting of frame-level spectral-, longer-range prosodic-, and lexical features, resulting in a best accuracy of 77 %. Furthermore, they examined the feasibility of performing a speech-based assessment in a military population [11]. More recent work explored the effectiveness of applying transfer learning approaches to the task of speech-based PTSD diagnosis [5] and also started to explore multimodal approaches using audio, video, and text data [9]. However, despite the numerous advances in the automatic assessment of mental disorders in recent years, the effect of treatments still received insufficient attention. Recent work focused solely on PTSD detection, uncoupled from any treatment session that might take place for an individual, lacking the ability to monitor changes during a clinical intervention and the effect on the performance of the model.

To address this shortcoming, we introduce a novel dataset consisting of 15 participants. Our dataset comprises audio data from people with and without PTSD, but, in contrast to previous studies, also includes data directly before and after a treatment session for each participant, i. e., enabling an analysis of the effect of the intervention on acoustic features of patients and therefore also on the predictive power of the prediction models by a) using both pre- and post-treatment data, b) using only post-treatment data, and c) using only pre-treatment data. Furthermore, we analyse important acoustic features that have the biggest impact on the models' decision, revealing insights into prominent acoustic patterns of individuals with PTSD.

10.21437/Interspeech.2023-1668

2. Dataset

For our experiments, we use a newly collected dataset consisting of 15 participants (all female as we had none male participants with PTSD), including speakers with PTSD and a control group. The data collection was approved by the ethics committee of the Ludwig-Maximilians-University (LMU) Munich and takes place at the Psychiatric Clinic of the University Hospital LMU since April 2022 and is still ongoing. Thus, this paper provides first insights based on data from the 15 persons available so far. The inclusion criterion for the PTSD group comprises a type I PTSD diagnosis (ICD-10: F43.1)[1]. Patients taking psychotropic drugs or affected by other mental disorders were excluded from the study.

Subsequently, audio recordings before and after a clinical intervention were collected for each study participant using a Zoom H5 audio recorder. First, participants were asked to read the (German) text *Der Nordwind und die Sonne* [*The Northwind and the Sun*] (NuS) prior to the treatment session. To prevent a habituation effect, participants were asked to read data excerpts from another text *Das tapfere Schneiderlein* [*The Valiant Little Tailor*] (DtS) post-treatment, but showing similar difficulty and length to the pre-treatment text NuS. In addition to the pre- and post-treatment text read aloud by the participants, they were also asked questions in a partially scripted interview before and after the treatment session, which they answered spontaneously.

The clinical intervention made between pre- and post-treatment audio recordings is twofold. 50 % of all the PTSD group received a treatment involving ear-acupuncture according to the NADA-protocoll [13], a five ear point standardised protocol used in psychiatric and PTSD patients worldwide. The second clinical intervention (applied to the remaining 50 % of PTSD patients) comprises a novel treatment approach of psychotherapy in combination with acupuncture [14]. The treatment was divided into three steps: First, the general condition of the study participants was treated with acupuncture to relieve the tension in the body and prepare them for the next steps. Secondly, study participants were confronted with a traumatic picture of a traumatic situation (e. g., car crash) and were tightly monitored for any sign of physical or in the third step emotional discomfort. Occurring negative sentiments were treated with the appropriate acupuncture points. As a result, the negative sentiments were dissolved, converting the traumatic picture only into a memory without any emotional discomfort. As all PTSD patients undergo treatment (and examining the particular type of intervention is not the focus of our experiments) we do not distinguish between the two treatment approaches in the remainder of our work.

For each participant, all steps of the study procedure were conducted during one day (pre-treatment recordings, treatment session, and post-treatment recordings). For our experiments, we use only the audio data from reading before and after the clinical treatment session, but not the semi-scripted interview data. Details about the collected dataset are shown in Table 1.

3. Experimental setup

In our approach, we first process the raw audio recordings and split them into smaller segments (cf. Section 3.1). Subsequently, we extract handcrafted (i. e., expert-designed) audio features as well as audio representations obtained using Transformer models (cf. Section 3.2). For modelling, we use a Support-Vector-Machine (SVM), of which the used parameters are explained in more detail in Section 3.3. Finally, we outline the overall evaluation protocol in Section 3.4.

Table 1: *Statistics of the Post-traumatic Stress Disorder (PTSD) dataset used in our experiments, showing the number (#), mean (μ), and standard deviation σ for several parameters.*

Parameter	$\#/\mu$	σ
Participants	15	-
with PTSD	7	-
without PTSD	8	-
Age [years]	29.1	7.5
Audio data per participant [sec]	85.5	12.2
Mean duration per phrase [sec]	2.0	0.7

3.1. Preprocessing

The pre- and post-treatment texts have an average duration of 42.8 seconds. To create sentence-level speech segments, we use the Munich AUtomatic Segmentation (MAUS) toolkit [15, 16], which has the advantage of taking prosodic characteristics into account. In doing so, we do not split the whole text into arbitrary small time windows, but divide the text into single prosodic phrases (20 phrases for NuS and 17 phrases for DtS), resulting in audio segments with a mean duration of 2.0 seconds per phrase.

Overall, we obtain a total of $15 \times 20 = 300$ segments as pre-treatment and $15 \times 17 = 255$ segments as post-treatment data, respectively.

3.2. Feature extraction

For our experiments, we use four different feature sets. In doing so, we apply both, traditional expert-designed features obtained via the OPENSMILE toolkit [17], as well as deep audio representations utilising Transformer models.

EGEMAPS [18] consists of 88 handcrafted features (e. g., F0, spectral or filter features) and thus has the advantage of being interpretable. For each previously segmented phrase, we obtain an 88-dimensional feature vector. Similarly, we extract an extended set of acoustic parameters (COMPARE [19]) which comprises 6 373 features, including, e. g., prosodic, temporal, or spectral features along with different statistically derived features. Both feature sets have been shown to be effective for related digital health tasks [12].

Furthermore, we experiment with using Transformer-based models as feature extractors. In doing so, we extract two different variants of WAV2VEC2.0 features. On the one hand, we use *wav2vec2-l-xlsr-53-german* [20], a variant of the *wav2vec2-l-xlsr* model [21] fine-tuned on data with German speakers. As all speakers in our dataset are German, we assume a better performance compared to the original version trained on 53 different languages. On the other hand, we utilise *w2v2-l-emotion-msp-dim* [22] representing a version of WAV2VEC2.0 which is fine-tuned for emotion recognition tasks. Since previous studies revealed that there is a relationship between emotion regulation difficulties and PTSD [23], we expect emotion-based features to be promising. Both WAV2VEC2.0 variants result in a 1 024-dimensional feature vector for each phrase-based segment obtained by averaging the second-to-last layer.

[1] https://icd.who.int/browse10/2019/en#/F43.1

3.3. Classifier

In our experiments, we utilise SVMs due to the limited amount of data. To compensate for the slightly unequal data distribution of people with PTSD and the control group (7 with and 8 without PTSD), we choose a balanced class weight. Furthermore, we optimise the cost parameter C $\in$ {.0001, .001, .1, 1, 5, 10}, the SVM kernel $\in$ {radial basis function (RBF), linear} as well as gamma $\in$ {auto, scale}. The optimisation is performed using grid search in a nested (3-fold) cross-validation setup.

Moreover, we min-max normalise the input features for the SVM to a range of $[0, 1]$. We also experimented with mean-std normalising the data, but do not report results as min-max normalisation showed superior performance on the collected dataset.

3.4. Evaluation protocol

As our final dataset comprises 15 participants, we use a leave-one-speaker-out cross-validation setup. A separate SVM model is trained for each of the 15 speakers, whereas testing is conducted on one hold-out subject at a time, while using all remaining 14 speakers for training. The SVM parameter optimisation is performed separately for each model on the train set (consisting of 14 subjects) using nested cross-validation.

For evaluating the model performance, we choose AUC as evaluation metric, splitting the participants into a positive (i. e., people with PTSD) and a negative (i. e., control group) class. Moreover, we calculate the AUC in two different ways. First, we calculate the performance per unit (i. e., phrase), indicating how well the model can distinguish single phrases between people with PTSD and the control group. Second, we average all predictions for each speaker, resulting in one session-based prediction per subject. The latter one can be interpreted as how well the SVM can distinguish not single phrases, but each individual into the two groups. Furthermore, we provide a 95 % confidence interval (CI) calculated over 1 000 bootstrap samples.

4. Results and discussion

The results achieved in our experiments are depicted in Table 2. In particular, we discuss the model performance w.r.t. the different features and (pre-/post-treatment) datasets (cf. Section 4.1) and provide an in-depth analysis of acoustic features related to PTSD detection (cf. Section 4.1).

4.1. Post-traumatic stress disorder detection

Table 2 summarises the prediction performance for the three variants a) using both pre- and post-treatment data, b) using only post-treatment data, and c) using only pre-treatment data. Overall, using both pre- and post-treatment data works better than using solely post-treatment data, but is being outperformed by utilising solely pre-treatment data for which we yield the best performance. The best result for Unit $_{\text{AUC}}$ is achieved using only pre-treatment data in combination with EGEMAPS features, leading to an AUC of 73 % compared to 58 % achieved using only post-treatment data. Moreover, it can be observed that session-based predictions outperform their unit-based counterparts in almost all cases. The best result for session-based prediction is yielded using only pre-treatment data with *w2v2-l-emotion-msp-dim* features, resulting in an AUC of 82 % compared to 66 % using only post-treatment data. Even though the number of participants in our dataset is limited, this large performance gap may indicate that important features present in PTSD patients decrease after a treatment session. To gain insights into which features are most important for the model, we analyse them in more detail in Section 4.2.

Table 2: *Mean AUC results with 95 % CI for PTSD detection in a leave-one-speaker-out cross-validation setup using five different random seeds.*

Feature	Unit $_{\text{AUC}}$[%]	Session $_{\text{AUC}}$[%]
Using pre- and post-treatment data		
w2v2-l-xlsr	54 (52-56)	55 (41-68)
w2v2-l-emotion-msp-dim	60 (58-62)	65 (52-79)
ComParE	54 (51-56)	57 (43-72)
eGeMAPS	**61 (59-63)**	**70 (58-82)**
Using only post-treatment data		
w2v2-l-xlsr	55 (52-58)	51 (36-66)
w2v2-l-emotion-msp-dim	**58 (55-61)**	**66 (52-78)**
ComParE	49 (45-52)	50 (35-64)
eGeMAPS	44 (40-47)	42 (30-56)
Using only pre-treatment data		
w2v2-l-xlsr	54 (51-57)	58 (44-73)
w2v2-l-emotion-msp-dim	68 (66-71)	**82 (72-91)**
ComParE	56 (53-59)	56 (43-71)
eGeMAPS	**73 (70-75)**	75 (63-85)

Furthermore, it can be observed that feature sets, which are often used for emotion recognition tasks (i. e., EGEMAPS and *w2v2-l-emotion-msp-dim*), perform best, confirming the assumption that there might be a link between emotional characteristics, such as arousal and PTSD. To further examine this relation, we apply the pre-trained emotion recognition model *w2v2-l-emotion-msp-dim* [22] and extract the corresponding arousal value for each audio segment (i. e., for all phrases) for all study participants. Overall, the results show that people with PTSD tend to have a lower mean arousal value ($.36 \pm .09$ compared to $.38 \pm .09$ in the control group), which may be reflected in a more monotone pronunciation.

4.2. Acoustic feature analysis

We further analyse the features that have the biggest impact on the model decision. In doing so, we utilise the SHapley Additive exPlanations (SHAP) toolkit [24] which is based on the idea of building surrogate models. First, all potential feature subsets are determined. Subsequently, the model performance is determined for all feature subsets, once with and once without the target feature, giving insights into how important a specific feature is. Finally, an interpretable global importance value can be derived for each feature [24].

We begin our analysis with the 10 most important features (of the EGEMAPS feature set) for the model decision, determined using SHAP. Furthermore, we apply the Mann-Whitney U test (features are not normally distributed) to gain insights into differences between the PTSD and control group with regard to the distribution of their acoustic features. Due to space constraints, we focus from all previously identified important features on the ones with an effect size $\geq 25\,\%$ and $\rho < .001$ calculated with the Mann-Whitney U test. Furthermore, we remove duplicates of features that show a similar trend and are, e. g., merely functionals, leading to five remaining key features

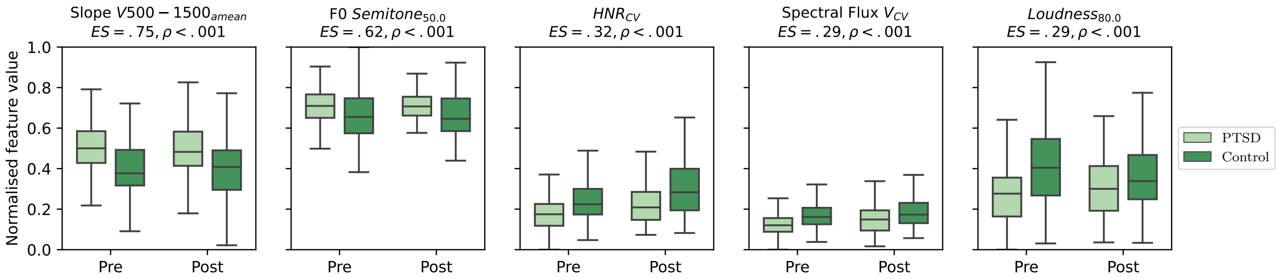

Figure 1: *Normalised feature values for the* 5 *(out of* 10*) most important acoustic features determined using SHAP and a Mann-Whitney U test (Spectral slope, F0, Harmonics-to-Noise-Ratio (HNR), Spectral flux, and Loudness). The features are ordered by their effect size (ES) and distinguished for participants with PTSD and the control group as well as for pre- and post-treatment data.*

which are depicted in Figure 1 and which we discuss in the remainder of this work.

Spectral slope: The first feature is known as the linear regression slope fitted to the log-magnitude spectrum and can be significantly influenced by angry/sad or loud/quiet speech [25]. Furthermore, previous work explored that an increased spectral slope correlates with higher breathiness, which in turn is often accompanied by a reduced loudness [26, 27]. Participants with PTSD show a higher spectral slope, which therefore may indicate a more breathy phonation with reduced loudness compared to the control group.

F0: This prosodic feature captures pitch on a logarithmic semitone scale. Figure 1 represents that people suffering from PTSD exhibit a higher pitch, for which previous studies explored that this might be correlated with an increased stress level of the participants [28].

Harmonics-to-Noise-Ratio (HNR): HNR describes the relationship between the harmonic and noise-like component energy. Studies from related fields, such as depression detection, showed that a decreased HNR goes along with an overall increased depression severity [29]. In addition, a decreased HNR indicates a more breathy or hoarse phonation, which can also be observed in the speech of depressed people [27, 30]. Similarly, it can be observed in the case of PTSD patients that HNR and its coefficient of variation, respectively, is lower compared to the control group, indicating similar patterns as people with depression.

Spectral flux: Another spectral feature is spectral flux, which describes the difference between the spectrum of two consecutive frames. In our experiments, PTSD participants tend to have a reduced spectral flux variation (i. e., smaller difference between the two spectrums of consecutive time windows), which may indicate more monotone speech [31]. Moreover, this observation is reinforced by the previously identified finding that PTSD participants exhibit a less mean arousal value than people without PTSD.

Loudness: Finally, we explore the loudness (i. e., perception of sound volume) of the audio recordings for both groups. Figure 1 shows that participants of the control group speak louder than the PTSD group on average. This aligns with the interpretation of the spectral slope and HNR feature as well as studies from other mental disorders such as depression [32, 33].

To summarise our findings: Participants with PTSD tend to speak more monotonously with a decreased loudness compared to the control group. These characteristics are consistent with observations of other mental disorders such as depression, but should still be validated specifically for PTSD on a larger dataset. With regard to differences between pre- and post treatment data, no clear trend can be observed. In some cases, the difference between the means of the two study groups is smaller using solely post-treatment data (e. g., spectral slope or loudness), but there are also cases where the difference increases (e. g., HNR). To be able to examine the two variants (using only pre- or post-treatment data) more closely, more data is needed. However, based on the data distribution of the two groups, it appears that in both cases the same features play a similarly important role in the decision process of the model.

5. Conclusions

In our study, we demonstrated the feasibility of speech-based PTSD detection before and after a clinical intervention. In our experiments, we achieved a best AUC of 73 % for unit-based and 82 % for session-based prediction, respectively, using solely pre-treatment data. Moreover, we showed that the emotion-based feature sets EGEMAPS and *w2v2-l-emotion-msp-dim*, yield the best results, which strengthens the assumption that there might be a correlation between PTSD and acoustic emotional characteristics. Furthermore, we analysed the most important features and concluded that participants suffering from PTSD exhibit a more quiet and monotonous speech in both cases, before and after a treatment session. However, our study also comes with limitations. On the one hand, the sample size of our collected dataset is still relatively small. On the other hand, we only explored a binary classification (into patients with and without PTSD). Thus, whether the models can separate between PTSD and other psychiatric disorders (e. g., borderline PD) cannot be judged. Therefore, future work should focus on collecting more data, allowing a deeper analysis between experiments using solely pre- or post-treatment data, also on an individual-level. Furthermore, future studies should target collecting data from people with mental disorders other than PTSD, enabling a multiclass-classification task. Moreover, additional data enable more complex deep neural networks to be trained complementary to SVMs. Finally, personalisation approaches that showed already promising results in the depression domain by taking individual characteristics of speakers into account and adapting the model to one person accordingly, should be explored [34, 35].

6. Acknowledgements

This work was funded by DFG project No. 461420398 (Leader Humor), and the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS).

7. References

[1] A. Shalev, I. Liberzon, and C. Marmar, "Post-traumatic stress disorder," *New England Journal of Medicine*, vol. 376, no. 25, pp. 2459–2469, 2017.

[2] J. Mylle and M. Maes, "Partial posttraumatic stress disorder revisited," *Journal of Affective Disorders*, vol. 78, no. 1, pp. 37–48, 2004.

[3] R. Yehuda, "Post-traumatic stress disorder," *New England journal of medicine*, vol. 346, no. 2, pp. 108–114, 2002.

[4] R. Yehuda, C. W. Hoge, A. C. McFarlane, E. Vermetten, R. A. Lanius, C. M. Nievergelt, S. E. Hobfoll, K. C. Koenen, T. C. Neylan, and S. E. Hyman, "Post-traumatic stress disorder," *Nature Reviews Disease Primers*, vol. 1, no. 1, pp. 1–22, 2015.

[5] D. Banerjee, K. Islam, K. Xue, G. Mei, L. Xiao, G. Zhang, R. Xu, C. Lei, S. Ji, and J. Li, "A deep transfer learning approach for improved post-traumatic stress disorder diagnosis," *Knowledge and Information Systems*, vol. 60, pp. 1693–1724, 2019.

[6] D. Lekkas and N. C. Jacobson, "Using artificial intelligence and longitudinal location data to differentiate persons who develop posttraumatic stress disorder following childhood trauma," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[7] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd." in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 847–851.

[8] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proc. Audio/Visual Emotion Challenge and Workshop*. Nice, France: ACM, 2019, pp. 3–12.

[9] M. A. L. Sawadogo, F. Pala, G. Singh, I. Selmi, P. Puteaux, and A. Othmani, "Ptsd in the wild: A video database for studying post-traumatic stress disorder recognition in unconstrained environments," *arXiv preprint arXiv:2209.14085*, pp. 1–12, 2022.

[10] D. D. Blake, F. W. Weathers, L. M. Nagy, D. G. Kaloupek, F. D. Gusman, D. S. Charney, and T. M. Keane, "The development of a clinician-administered ptsd scale," *Journal of traumatic stress*, vol. 8, pp. 75–90, 1995.

[11] D. Vergyri, B. Knoth, E. Shriberg, V. Mitra, M. McLaren, L. Ferrer, P. Garcia, and C. Marmar, "Speech-based assessment of ptsd in a military population using diverse feature classes," in *Proc. INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 3729–3733.

[12] S. Amiriparian and B. Schuller, "Ai hears your health: Computer audition for health monitoring," in *Proc. ICT for Health, Accessibility and Wellbeing (IHAW)*. Larnaca, Cyprus: Springer, 2022, pp. 227–233.

[13] B. Cole and M. Yarberry, "Nada training provides ptsd relief in haiti," *Deutsche Zeitschrift für Akupunktur*, vol. 54, no. 1, pp. 21–24, 2011.

[14] J. Schottdorf and R. Musil, "Psychological trauma therapy with 3-steps acupuncture-based exposition," *Deutsche Zeitschrift für Akupunktur*, vol. 60, no. 4, pp. 6–12, 2017.

[15] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proc. International Congress of Phonetic Sciences (ICPhS)*, San Francisco, USA, 1999, pp. 607–610.

[16] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM International Conference on Multimedia (ACM MM)*. Ottawa, Canada: ACM, 2010, pp. 1459–1462.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[19] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 1–5.

[20] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in German," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german, 2021.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. INTERSPEECH*. Virtual Conference: ISCA, 2020, pp. 2426–2430.

[22] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, pp. 1–25, 2022.

[23] L. Shepherd and J. Wild, "Emotion regulation, physiological arousal and ptsd symptoms in trauma-exposed individuals," *Journal of behavior therapy and experimental psychiatry*, vol. 45, no. 3, pp. 360–367, 2014.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Neural Information Processing Systems (NeurIPS)*, Long Beach, USA, 2017, pp. 1–10.

[25] K. W. Godin, T. Hasan, and J. H. Hansen, "Glottal waveform analysis of physical task stress speech," in *Proc. INTERSPEECH*. Portland, USA: ISCA, 2012, pp. 1–4.

[26] A. Triantafyllopoulos, M. Fendler, A. Batliner, M. Gerczuk, S. Amiriparian, T. M. Berghaus, and B. W. Schuller, "Distinguishing between pre-and post-treatment in the speech of patients with chronic obstructive pulmonary disease," in *Proc. INTERSPEECH*. Incheon, Korea: ISCA, 2022, pp. 1–5.

[27] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: relevant features and relevance of gender," in *Proc. INTERSPEECH*. Singapore: ISCA, 2014, pp. 1248–1252.

[28] A. Eriksson and M. Heldner, "The acoustics of word stress in english as a function of stress level and speaking style," in *Proc. INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 41–45.

[29] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Proc. INTERSPEECH*. Portland, USA: ISCA, 2012, pp. 1–4.

[30] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Communication*, vol. 75, pp. 27–49, 2015.

[31] J. De Boer, A. Voppel, S. Brederoo, H. Schnack, K. Truong, F. Wijnen, and I. Sommer, "Acoustic speech markers for schizophrenia-spectrum disorders: a diagnostic and symptom-recognition tool," *Psychological Medicine*, pp. 1–11, 2021.

[32] J. Wang, L. Zhang, T. Liu, W. Pan, B. Hu, and T. Zhu, "Acoustic differences between healthy and depressed people: a cross-situation study," *BMC Psychiatry*, vol. 19, pp. 1–12, 2019.

[33] J. K. Darby, N. Simmons, and P. A. Berger, "Speech and voice parameters of depression: A pilot study," *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.

[34] A. Kathan, M. Harrer, L. Küster, A. Triantafyllopoulos, X. He, M. Milling, M. Gerczuk, T. Yan, S. T. Rajamani, E. Heber, I. Grossmann, D. D. Ebert, and B. W. Schuller, "Personalised depression forecasting using mobile sensor data and ecological momentary assessment," *Frontiers in Digital Health*, vol. 4, p. 964582, 2022.

[35] M. Gerczuk, A. Triantafyllopoulos, S. Amiriparian, A. Kathan, J. Bauer, M. Berking, and B. Schuller, "Personalised deep learning for monitoring depressed mood from speech," in *Proc. E-Health and Bioengineering Conference (EHB)*. Iași, Romania: IEEE, 2022, pp. 1–5.