Interspeech 2024
1-5 September 2024, Kos, Greece

Real-world PTSD Recognition: A Cross-corpus and Cross-linguistic Evaluation

Alexander Kathan[1,2], *Martin Bürger*[1,2], *Andreas Triantafyllopoulos*[1,2], *Sabrina Milkus*[3], *Jonas Hohmann*[3], *Pauline Muderlak*[3], *Jürgen Schottdorf*[4], *Richard Musil*[3], *Björn W. Schuller*[1,2,5], *Shahin Amiriparian*[1,2]

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

alexander.kathan@uni-a.de

Abstract

Post-traumatic Stress Disorder (PTSD) is a mental condition that develops as a result of catastrophic events. Triggers for this may include experiences, such as military combat, natural disasters, or sexual abuse, having a great influence on the mental wellbeing. Due to the severity of this condition, early detection and professional treatment is crucial. For this reason, previous research explored prediction models for recognising PTSD at an early stage. However, when these models are transferred from research to real-world applications, they face heterogeneous environments (e. g., different recording settings, various dialects or languages). To analyse this effect, we develop a speech-based PTSD recognition model and subsequently analyse its cross-corpus and cross-linguistic performance. Our experiments indicate that there are cross-cultural factors influencing PTSD and leading to a best area under the ROC curve (AUC) of $70.1\,\%$ evaluated cross-corpus.

Index Terms: PTSD, machine learning, digital health

1. Introduction

Traumatic events, such as ones that involve a real risk of death or a serious injury can have chronic repercussions in the form of a Post-traumatic Stress Disorder (PTSD) [1, 2]. With a prevalence of $5 - 10\,\%$, PTSD is a serious condition, to which women are twice as susceptible as men [3]. The triggers for PTSD can be multifaceted and range from war operations to serious accidents, or sexual abuse [4, 5, 6] . The effects of this condition on the daily lives of those affected are equally complex. In addition to constant fears, nightmares, and reliving the trigger event, patients with PTSD can also develop further mental disorders such as Major Depressive Disorder (MDD) [1], which further deteriorate the overall condition of the patient.

To counteract this, it is important to recognise PTSD at an early stage, providing professional treatment in time. To this end, many different approaches have been explored in previous work with the aim of recognising PTSD earlier and thus enabling a timely clinical treatment. From a clinical perspective, self-reporting questionnaires and structural clinical interviews [7, 8, 9] are typically used to determine whether a person suffers from PTSD. However, even if they enable reliable detection, these clinical assessment methods have weaknesses. First, they are time-consuming. Second, they can only be carried out by properly trained psychologists (i. e., the assessment is not quick and cannot be automated) [10]. Machine learning-based approaches can help to overcome these limitations, utilising data, e. g., obtained from smartphone sensors or audio.

The former was explored, e. g., by Lekkas et al. [11]. In their study, they used GPS data collected with a smartphone and subsequently derived movement patterns, leading to an area under the ROC curve (AUC) of $81.6\,\%$. Vergyri et al. [10] investigated the latter one and experimented with prosodic- and lexical audio features. In their work, they demonstrated the feasibility of a PTSD assessment solely based on audio data, resulting in a best classification accuracy of $77\,\%$. Moreover, Kathan et al. [12] explored the effect of a clinical treatment session on the speech of people with PTSD. In their experiments, they showed that a best performance of $82\,\%$ is achieved using only data before the clinical intervention, indicating the positive effect of a single treatment session. Finally, Sawadogo et al. [13] investigated multimodal approaches, considering audio, video, and text data.

Despite the numerous researches in the field of PTSD detection, it has been neglected to analyse the effect and performance when being transferred into real-world applications, where trained models face different recording devices, recording environments, local dialects or even a different language. To solve this limitation, we present in this work a cross-corpus and cross-linguistic analysis. Our contribution is threefold. First, we explore four self-supervised learning (SSL) [14] approaches for PTSD recognition, improving the accuracy on the English-language dataset *PTSD In The Wild* (ITW) [13]. Second, we freeze these models and apply them to a German-language PTSD dataset (on which they were never trained), *LMU PTSD* (LMU) [12], and analyse the cross-corpus and cross-linguistic performance. Third, we examine common acoustic markers that are present in both datasets (i. e., languages), but also discuss those features that reveal different trends in the two datasets.

2. Datasets

For our study, we consider two PTSD datasets, the English-language ITW [13] and the German-language LMU [12] dataset, enabling a cross-corpus and cross-linguistic analysis.

The English ITW dataset consists of 634 videos (317 people with PTSD and 317 healthy controls) downloaded from YouTube. For the PTSD group, the authors chose people who experienced a traumatic event, giving an interview and answering questions about their story with PTSD. For the control group, they include celebrities who have been interviewed. The average video dura-

10.21437/Interspeech.2024-493

tion is approximately 2 minutes, leading to a total of 21 hours of available raw video and audio data. Further details about the dataset can be found in [13].

The German LMU dataset, recently introduced in [12], comprises 20 participants (7 people with and 13 people without PTSD). Study participants of the PTSD cohort were required to have a type I PTSD diagnosis (ICD-10: F43.1). For the audio recording, the participants were asked to read the two given texts, *Der Nordwind und die Sonne* [*The Northwind and the Sun*] and *Das tapfere Schneiderlein* [*The Valiant Little Tailor*], respectively. Furthermore, the LMU dataset includes data before and after a clinical treatment session. However, since the aim of this study is not to distinguish between pre- and post-treatment sessions, we will not differentiate between these two categories in our study, but only focus on distinguishing PTSD patients from the control group.

As the ITW dataset is significantly larger, we use it for training and evaluating our SSL models for recognising PTSD. Subsequently, these trained models are then applied (without further training) and evaluated on the LMU dataset, revealing whether there are acoustic markers in the speech of people with PTSD that work cross-linguistically.

3. Experimental setup

We begin our experiments with preprocessing both datasets, applying speaker diarisation, splitting the audio recordings into smaller chunks with a length of 5 and 30 seconds, respectively, followed by several data augmentation methods (Section 3.1). Subsequently, we extract audio representations using four SSL models and utilise these features with a simple decision head to perform binary PTSD classification (Section 3.2). Furthermore, we do not only use the four models as feature extractors, but also fine-tune them, leading to more specialised models for PTSD recognition (Section 3.3). Finally, we apply these models to a novel unseen dataset with another language (Section 3.4).

3.1. Preprocessing and data augmentation

In addition to the patient, some videos of the PTSD cohort in the ITW dataset also include additional speakers. Therefore, we determine for each audio file all speakers using speaker diarisation[1] [15, 16] and remove for the PTSD cohort all non-PTSD speakers as well as segments without any speech at all.

The total available audio duration per participant is about 2 minutes for ITW and 1.5 minutes for LMU. In a subsequent step, we therefore split the audio files into smaller segments of 5 and 30 seconds, respectively. In cases of speech recordings that are shorter than the specified chunk length (5 or 30 seconds), the audio is not splitted any further.

To further enrich the data, we apply three data augmentation methods on the training data. First, we perform random cropping and zero padding, to extend the amount of available data. Second, we randomly shift the pitch by up to 3 semitones up or down with a likelihood of 30 % during training. Third, we add White Gaussian Noise with a probability of 40 % and an applied signal-to-noise ratio which is randomly sampled from the range $[35; 55]$ (for BYOL-A as it is more sensitive to Gaussian Noise compared to wav2vec2.0 (w2v2) models [17]) and $[15; 35]$ (for w2v2 models), respectively.

[1]https://huggingface.co/pyannote/speaker-diarization

3.2. Frozen experiments

In our experiments, we apply four different pretrained SSL models as frozen feature extractors, followed by a simple decision head which we train on the ITW dataset.

1) BYOL for Audio (BYOL-A)[2] [17] comprises a dual-network architecture, pretrained on AudioSet [18]. The first network tries to predict the representation of the second network which is an augmented version of the same audio clip of the first network. The augmented counterpart could be altered by, e. g., adding background noise, changing the pitch or time-stretching, enabling the model to learn strong features that remain consistent and unaffected by these transformations. *2) W2v2-base*[3] [19] is a version of w2v2, including a convolutional feature encoder for handling raw audio signals as well as a Transformer-based context network. The base model comprises 95m parameters and is pretrained on Librispeech. *3) W2v2-large-lv60*[4] [19] represents another variant of w2v2, which is pretrained on the larger LibriVox dataset. Furthermore the model is bigger compared to the smaller base model, comprising 317m parameters. *4) W2v2-emo-msp*[5] [20] forms a third variant, which is fine-tuned for emotion recognition tasks and has proven to be efficient for PTSD recognition [12].

For our frozen experiments, we apply all four SSL models as feature extractors, perform pooling across the time dimension and add a simple multilayer perceptron (MLP) as decision head for the classification task. In the case of BYOL-A, we perform average- and max-pooling in parallel and subsequently sum up both tensors, resulting in a 2 048-dimensional feature vector. In the case of the w2v2 models, average pooling is applied, leading to a 1 024-dimensional feature vector. The MLP consists of two fully-connected layers, GELU as activation function, Layer Normalisation, as well as a dropout of $[0.5 - 0.7]$.

For training the model, we use the data split described in the ITW paper [13], splitting the recordings into 80 % for training, 10 % for validation, and 10 % for testing. As loss function, we apply BCE-Loss. Moreover, we use Adam as optimiser, a learning rate of 3e-6, and train the model for a maximum of 100 epochs, including early stopping when not improving for 4 consecutive epochs. For BYOL-A, we apply a batch size of 256 for the 5 s and 128 for the 30 s segments, respectively. For w2v2-base, we use 128 and 64, for w2v2-emo-msp 64 and 16, respectively. For w2v2-large-lv60, we use for both segment lengths a batch size of 16. As evaluation metric, we choose the area under the ROC curve (AUC) . In doing so, the AUC is calculated in two variants. First, it is calculated on the basis of single chunks. Second, the mean prediction probabilities for all chunks per speaker are determined, resulting in a speaker-level performance measurement.

3.3. Fine-tuning experiments

Even if upstream models can already produce good representations, fine-tuning them on the downstream classification task has been shown to improve performance [21]. For this reason, we use the previously introduced pretrained models and fine-tune them on the ITW dataset while updating all parameters. Furthermore, we apply the same decision head, as well as the same experimental setup as outlined in Section 3.2. However, due to

[2]https://github.com/nttcslab/byol-a

[3]https://huggingface.co/facebook/wav2vec2-base

[4]https://huggingface.co/facebook/wav2vec2-large-lv60

[5]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

the increased computational demand of computing gradients for the pretrained models, we reduce the batch sizes for w2v2-base to 64 for the 30 s and 16 for the 5 s segments, respectively. For w2v2-emo-dim, it is reduced to 64 and 4. For w2v2-large-lv60, we reduce the batch size for the 5 s experiments to 8. We do not conduct experiments for the 30 s experiments, due to its memory consumption, which exceeded our GPU capacity.

3.4. Cross-corpus and cross-linguistic experiments

The experiments outlined in Section 3.2 and Section 3.3 only involve data from the ITW dataset. Although the English ITW and the German LMU dataset differ in many aspects (e. g., language, environment, circumstances of recording, content of speech), they should both exhibit paralinguistic features associated with PTSD in individuals affected by this condition. If the previously trained models successfully recognise such peculiarities, they should also be – to some degree – effective on the LMU samples. The examined models (BYOL-A and the w2v2 variants) are therefore also evaluated on all samples of the LMU dataset (after training them on the ITW dataset).

4. Results and discussion

Table 1 shows the results achieved in our experiments. In Section 4.1, we discuss them especially w. r. t. the model prediction performance when being applied cross-corpus. In Section 4.2, we conduct an extensive feature analysis, discussing cross-linguistic acoustic markers present in both datasets.

4.1. Binary PTSD prediction

The binary PTSD prediction results are depicted in Table 1. In all cases, the best result is achieved using the w2v2-emo-msp model. The frozen experiments (in which we used the model solely as feature extractor) yield a best chunk-level AUC of .947 and speaker-level result of .981 when trained and evaluated on the ITW dataset. Furthermore, it can be observed that mostly the longer chunks with a duration of 30 s perform better compared to the 5 s chunks.

For the fine-tuned (on ITW) experiments, we obtain a best chunk-level AUC of .983 and a best speaker-level AUC of .997 using w2v2-emo-msp. Similar to the frozen experiments, the w2v2-emo-msp model outperforms the other models by a large margin. Moreover, the averaged speaker-level predictions improve the results in most of the cases, except for the BYOL-A model in conjunction with a segment length of 30 s. The fine-tuned experiments show that the SSL model benefits from updating the parameters of the pretrained models during training and not only using them as feature extractors, resulting in a higher prediction performance, consistent with previous work showing that fine-tuning layers is important for improving performance [21].

Finally, we take the best models (pretrained and fine-tuned solely on the English-language ITW dataset) and utilise them for binary PTSD prediction for all audio chunks of the German-language LMU dataset to test their generalisation on a novel unseen cross-linguistic dataset. The results show that a best overall performance is achieved using 5 s chunks on a speaker-level in conjunction with the w2v2-emo-msp model, leading to an AUC of .701. The best chunk-level result is obtained for the 30 s length in combination with the w2v2-emo-msp model. Whereas the BYOL-A and w2v2-base results are mostly around chance-level, the w2v2-large-lv60 and w2v2-emo-msp models are able to clearly outperform chance-level, indicating that there

Table 1: *AUC results for PTSD prediction using the English-language PTSD In The Wild (ITW) and the German-language LMU PTSD (LMU) dataset. We report results for chunk- and speaker-level, and for different segment lengths (Len). The frozen results are formed using the SSL models as feature extractors. The fine-tuned results are obtained by fine-tuning the models. Cross-corpus results are achieved evaluating the fine-tuned (on the ITW dataset) models on the LMU dataset.*

[AUC]		ITW [Frozen]		ITW [Fine-tuned]		LMU [Cross-corpus]	
Model	**Len**	**Chunks**	**Speaker**	**Chunks**	**Speaker**	**Chunks**	**Speaker**
BYOL-A	5 s	.825	.869	.825	.871	.469	.500
	30 s	.858	.846	.875	.857	.529	.531
w2v2-base	5 s	.693	.726	.884	.931	.506	.554
	30 s	.745	.731	.885	.957	.539	.531
w2v2-large-lv60	5 s	.646	.692	.751	.843	.584	.637
	30 s	–	–	–	–	–	–
w2v2-emo-msp	5 s	.865	.942	.925	.956	.632	**.701**
	30 s	**.947**	**.981**	**.983**	**.997**	**.688**	.678

are paralinguistic acoustic markers for people suffering from PTSD that are similarly prominent in both, German and English language.

4.2. Cross-linguistic feature analysis

To further explore which acoustic markers are important in both languages, we conduct a cross-linguistic feature analysis. In doing so, we only analyse female speakers in both datasets (as the majority of study participants in the LMU dataset is female) to eliminate a sex-based influence of the voice. Subsequently, we use the OPENSMILE toolkit [22] to extract the eGeMAPS features [23], which comprise functionals over a set of low-level descriptors (LLDs), and apply the Mann-Whitney-U test (features are not normally distributed). In the next step, we filter all features based on their effect sizes and ρ values, leading to a list of acoustic markers that only contains features that show a significant difference between participants with PTSD and the control group in both datasets. In our analysis, we consider an effect size of $\geq 30\,\%$ and $\rho < .001$. Due to space limitations, we remove functionals of the same LLD that show a similar trend.

Figure 1 depicts 10 acoustic markers that are important in both, the English ITW and the German LMU dataset. The upper row shows relevant features that show the same pattern in both datasets. The feature with the highest effect size is *MFCC2*. In both datasets, *MFCC2* is higher for people with PTSD compared to the control group. A similar observation was made in the related field of MDD. Taguchi et al. [24] explored that *MFCC2* was significantly higher for study participants with MDD, concluding that this might reflect a change of the quality of voice. Another acoustic marker is the variation of *F0* (in semitones). This marker seems to decrease in both datasets for participants of the PTSD cohort, resulting in a speech that can be perceived more monotonous and sad [25, 26]. This is also the case in depression and schizophrenic research, which show a similar pattern. An early study by Nilsonne [25] came to the conclusion that the variation of *F0* is lower for depressed people and increases once they recover. *Harmonics-to-Noise-Ratio (HNR)* quantifies the relationship between harmonic sounds (e. g., responsible for the tone or pitch) to noise sounds (e. g., aperiodic components such as breathiness or hoarseness). Previous PTSD studies showed that HNR and its coefficient of variation are lower for the PTSD cohort [12], which can also be observed in our case for both the English-language and German-language

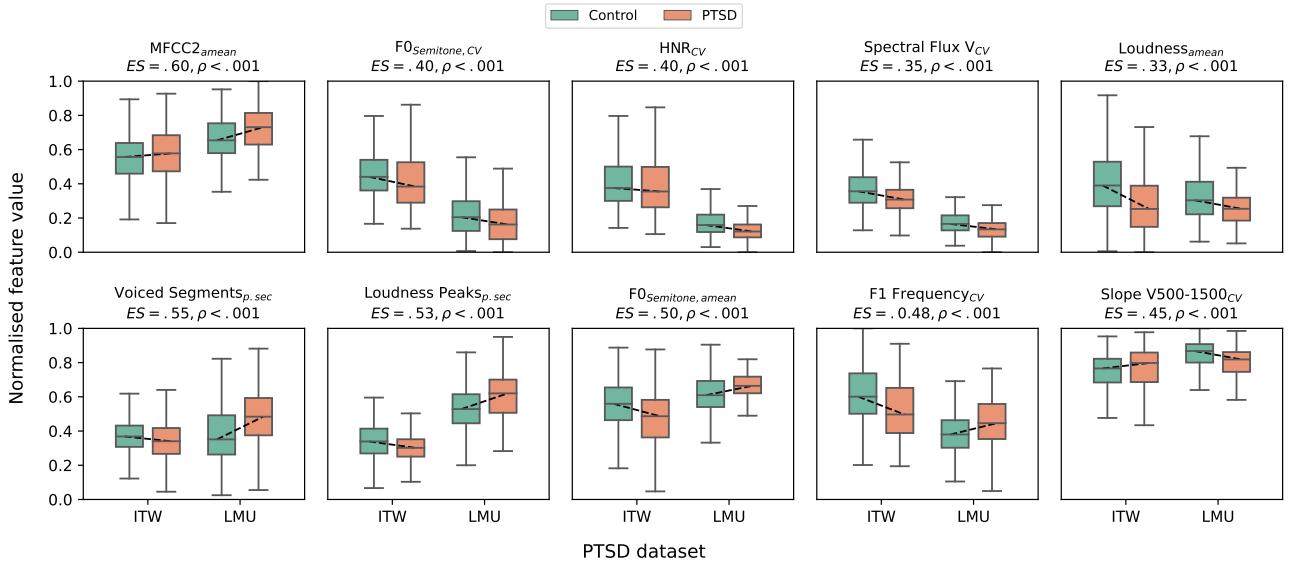

Figure 1: *Normalised feature values for acoustic markers that are important in both, the English-language dataset PTSD In The Wild (ITW) as well as in the German-language dataset LMU PTSD (LMU). The features are calculated separately for people with PTSD and the control group and are ordered by their effect size (ES). The upper row shows important features that have a similar trend in both datasets whereas the bottom row displays acoustic markers revealing opposing trends among both datasets.*

dataset. Similarly, the *Spectral flux*, describing the change of two consecutive spectrum frames, seems to be a promising marker for determining PTSD [12]. In both datasets, people with PTSD exhibit a lower spectral flux variation, indicating a more monotone speech [27]. Finally, *Loudness* appears to be a generally valid cross-linguistic marker for PTSD. In the ITW as well as in the LMU dataset, we observe that the control group speaks louder than the PTSD cohort, which is also visible in several mental disorder studies [28, 29, 30].

The bottom row of Figure 1 in turn reveals important acoustic features with a high effect size, but showing opposing patterns among both datasets (e. g., increased value for PTSD in the ITW dataset, but decreased value for the PTSD group in the LMU dataset). Some of the opposing trends, e. g., *Loudness peaks per second* may arise due to phonetic and prosodic differences across the German and English language [31, 32]. In particular, German may have a more forceful or abrupt articulation compared to English, because of consonants such as the hard 't', which could result in more loudness peaks. Furthermore, there are acoustic markers that show differences between PTSD patients and the control group in single datasets and studies, but that are often not significant. Nilsonne [25], for example, investigated that people with a related mental disease (i. e., depression) have a minimally lower mean *F0* value (similar to the ITW dataset), but which was not significant. In contrast, Kathan et al. [12] explored an increased *F0* which might be correlated with a higher stress level of PTSD patients.

To summarise the results of our feature analysis: On the one hand, there are several acoustic markers, such as *MFCC2* or *Loudness* that are significant in both, the ITW and the LMU dataset, showing similar patterns across both languages, and therefore appear to be valid cross-linguistic markers. Moreover, people with PTSD seem to speak more monotonous with a reduced loudness compared to healthy controls. On the other hand, there are acoustic markers that may be relevant in individual datasets or in a specific language (e. g., mean *F0*), however, do not seem to be generalisable cross-linguistically and therefore may not be reliable valid cross-corpus markers for PTSD.

5. Conclusions

When machine learning models are transferred from research to real-world applications, they are typically confronted with heterogeneous environments compared to the training data (e. g., different recording settings, background noise, or various dialects or languages). Our experiments show that there are robust paralinguistic features in the speech of PTSD patients, enabling the model to cope with these differences, leading to a reliable cross-corpus and cross-linguistic PTSD detection with a best AUC of .701. Moreover, our analysis indicates that there are well generalisable cross-linguistic markers in patients with PTSD, such as *MFCC2* or *Loudness*. At the same time, however, there are acoustic features that only have a certain relevance in individual datasets or languages (e. g., mean *F0*).

Nevertheless, our study also comes with limitations. First, we only considered two datasets (one per language). Therefore, future work should integrate more datasets, including several languages. Second, we only explored a binary PTSD classification in our machine learning experiments, making it impossible to judge whether the models can only distinguish between the PTSD and healthy control cohort, or also are able to separate between PTSD and other psychiatric disorders. To overcome this limitation, future work should also collect multi-linguistic datasets, including different mental conditions as well as exploring personalisation strategies, proven to be successful in depression recognition [33, 34, 35], to adapt the model to the individual characteristics of the people in a new dataset.

6. Acknowledgements

This work was supported by the MDSI – Munich Data Science Institute and the MCML – Munich Center for Machine Learning.

7. References

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

[2] U. Hepp, H. Moergeli, S. Buchi, H. Bruchhaus-Steinert, B. Kraemer, T. Sensky, and U. Schnyder, "Post-traumatic stress disorder in serious accidental injury: 3-year follow-up study," *The British Journal of Psychiatry*, vol. 192, no. 5, pp. 376–383, 2008.

[3] 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.

[4] E. J. Bromet, L. Atwoli, N. Kawakami, F. Navarro-Mateu, P. Piotrowski, A. King, S. Aguilar-Gaxiola, J. Alonso, B. Bunting *et al.*, "Post-traumatic stress disorder associated with natural and human-made disasters in the world mental health surveys," *Psychological medicine*, vol. 47, no. 2, pp. 227–241, 2017.

[5] M. C. Mobbs and G. A. Bonanno, "Beyond war and ptsd: The crucial role of transition stress in the lives of military veterans," *Clinical psychology review*, vol. 59, pp. 137–144, 2018.

[6] R. J. Ursano, C. S. Fullerton, R. S. Epstein, B. Crowley, T.-C. Kao, K. Vance, K. J. Craig, A. L. Dougall, and A. Baum, "Acute and chronic posttraumatic stress disorder in motor vehicle accident victims," *American Journal of Psychiatry*, vol. 156, no. 4, pp. 589–595, 1999.

[7] 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.

[8] 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.

[9] R. C. Kessler, S. Aguilar-Gaxiola, J. Alonso, C. Benjet, E. J. Bromet, G. Cardoso, L. Degenhardt, G. de Girolamo, R. V. Dinolova, F. Ferry *et al.*, "Trauma and ptsd in the who world mental health surveys," *European journal of psychotraumatology*, vol. 8, no. sup5, p. 1353383, 2017.

[10] 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.

[11] 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.

[12] A. Kathan, A. Triantafyllopoulos, S. Amiriparian, S. Milkus, A. Gebhard, J. Hohmann, P. Muderlak, J. Schottdorf, B. W. Schuller, and R. Musil, "The effect of clinical intervention on the speech of individuals with ptsd: features and recognition performances," in *Proc. INTERSPEECH*. Dublin, Ireland: ISCA, 2023, pp. 4139–4143.

[13] 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," *Multimedia Tools and Applications*, pp. 1–23, 2023.

[14] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *Patterns*, vol. 3, no. 12, 2022.

[15] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *Proc. ICASSP*. Barcelona, Spain: IEEE, 2020, pp. 7124–7128.

[16] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. INTERSPEECH*. Brno, Czech Republic: ISCA, 2021, pp. 1–5.

[17] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *Proc. International Joint Conference on Neural Networks*. Virtual Conference: IEEE, 2021, pp. 1–8.

[18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*. New Orleans, USA: IEEE, 2017, pp. 776–780.

[19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[20] 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.

[21] A. Triantafyllopoulos and B. W. Schuller, "The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case," in *Proc. ICASSP*. Toronto, Canada: IEEE, 2021, pp. 7268–7272.

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

[23] 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.

[24] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana, M. Nishimura, and T. Arai, "Major depressive disorder discrimination using vocal acoustic features," *Journal of affective disorders*, vol. 225, pp. 214–220, 2018.

[25] Å. Nilsonne, "Acoustic analysis of speech variables during depression and after improvement," *Acta psychiatrica scandinavica*, vol. 76, no. 3, pp. 235–245, 1987.

[26] K. R. Scherer, "Speech and emotional states," *Speech evaluation in psychiatry*, pp. 189–220, 1981.

[27] 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*, vol. 53, no. 4, pp. 1302–1312, 2023.

[28] 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.

[29] 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.

[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] W. Strange, A. Weber, E. S. Levy, V. Shafiro, M. Hisagi, and K. Nishi, "Acoustic variability within and across german, french, and american english vowels: Phonetic context effects," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 1111–1129, 2007.

[32] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[33] 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.

[34] M. Gerczuk, A. Triantafyllopoulos, S. Amiriparian, A. Kathan, J. Bauer, M. Berking, and B. W. Schuller, "Zero-shot personalization of speech foundation models for depressed mood monitoring," *Patterns*, vol. 4, no. 11, 2023.

[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. EHB Conference*. Iași, Romania: IEEE, 2022, pp. 1–5.