INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Prediction of the Gender-based Violence Victim Condition using Speech: What do Machine Learning Models rely on?

Emma Reyner-Fuentes[1], *Esther Rituerto-González*[2], *Isabel Trancoso*[3], *Carmen Peláez-Moreno*[2]

[1]Department of Electronics, University Carlos III of Madrid (UC3M), Spain
[2]Department of Signal Theory and Communications, University Carlos III of Madrid (UC3M), Spain
[3]INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal

e.reyner@pa.uc3m.es, erituert@ing.uc3m.es, isabel.trancoso@inesc-id.pt, carmen@tsc.uc3m.es

Abstract

Women who have experienced gender-based violence (GBV) are at an increased risk of developing mental illnesses such as depression, anxiety, and post-traumatic stress disorder (PTSD). Recently, Artificial Intelligence (AI) has provided new tools to assist in mental health clinical diagnosis, including speech-based detection. However, there is not much work done on the GBV victim (GBVV) condition detection. This study aims to identify specific speech features that aid this detection, analyse the relationship of such results with the user's psychological evaluation, and evaluate whether the models rely on the speaker identity or self-reported emotions to predict the GBVV condition. Our results indicate that it is possible to distinguish GBVV with controlled sequelae from non-victims, which may suggest that such differentiation for GBVV with more severe mental aftereffects–such as PTSD–may be even more meaningful. We believe that our work can help future mental health AI therapy assistants.

Index Terms: Gender-Based Violence, Speech Paralinguistics, Psychological Conditions, AI Therapy Assistants

1. Introduction

Physical or sexual aggressions are considered manifestations of GBV, a pervasive problem that affects mostly women and girls[1] in our society. Different forms of GBV consistently lead to a range of mental illnesses globally, including anxiety, depression, suicide, PTSD, and substance abuse [1, 2, 3]. This pervasiveness of mental disorders is high among women who have suffered from some type of GBV [4, 5]. Their symptoms can be extremely distressing and can deteriorate a person's ability to work, socialize, and carry out daily activities. Notably, PTSD is the most common sequela among GBVV [3, 6, 7].

Conventional medical assessments to detect mental conditions often include psychosocial questionnaires and are generally diagnosed, or confirmed, by a professional. Recently, AI has given rise to a whole new set of tools to assist in clinical diagnosis [8], including mental health disorders [9]. Novel studies address the detection of mental conditions via automatic speech-based detection [10, 11], being speech a non-invasive data modality where the state or perception of the patient could be less influenced by the diagnosis being performed. Even though these studies are on the rise, there are no studies that use speech particularly in the detection of the GBVV condition (GBVVC), a data modality which could therefore be used in early and non-invasive detection.

[1]European Institute for Gender Equality. 2023. *What is gender-based violence?* Available in: https://eige.europa.eu/gender-based-violence/what-is-gender-based-violence

Thus, in the present study, we aim to nswer to the following questions: What is the prevalence of the GBVVC in the speech? Are there specific speech features that help in such prediction? Is additional information, such as the speaker identity or emotions, influencing such detection? By extracting different feature sets, we aim to determine a suitable one for the detection of the GBVVC, as well as the possible influence of emotions or speaker voice traits in such detection.

2. Related work

Research in the use of speech as a source of data and AI algorithms–such as Machine Learning (ML) or Deep Learning (DL) models–for the detection of mental health disorders has skyrocketed in the last decade [11, 12]. Yet, there is a scarce number of open speech corpora from patients with mental disorders [12] since clinical data has often privacy restrictions [11]. A notable exception is the DAIC database [13], that contains clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and PTSD–which are often present in GBVV. But due to speech from GBVV being very sensitive data, there are no large databases to work within the literature. Given that many GBVV suffer from PTSD, techniques to capture data that make the patients re-live traumatic experiences can only be applied with great care, as re-living past events can have a negative impact in the subjects [14]: the so-called *re-traumatization*. In GBV, this is called *re-victimization* [15].

There is only one study [16] to date for the detection of the GBVVC via speech. On it, we show preliminary results and ablation studies where we used a subset of data from the WEMAC Database [17] (26 GBVV + 26 non-GBVV) and a total of 756 feature samples (1 averaged value per audio), obtaining a score of $71.53 \pm 32.85\%$ in a Leave-One-Subject-Out (LOSO) scheme. The high variability among users made us hypothesize that clusters within GBVV and non-GBVV could exist in the feature domain and–since that sample was really small–those positive results could be obtained for subjects whose clusters were well represented in the sample. With our present work, we extend [16] coping with its limitations and determining what information the models rely on.

3. Methodology and Materials

3.1. Database

In this study, we use the WEMAC Database [17], a multi-modal affective computing dataset, which includes Spanish speech data and self-reported emotional annotations, captured in laboratory conditions from volunteer women. The participants were diverse women (ages, education levels, and nationalities) including GBVV. The collection of data was done after the

10.21437/Interspeech.2023-1771

participants experienced 14 audiovisual stimuli with a Virtual Reality (VR) Headset in an immersive setup for emotion elicitation. Right after each stimulus, two questions were answered aloud and recorded. The protocol for volunteer recruiting in WEMAC [17] was different for non-GBVV than for GBVV. The first group only surveyed an initial general questionnaire, whereas the second underwent a more extensive psychological evaluation that included a PTSD Symptom Severity Scale-Revised (*EGS-R*) test [18]. The score for such test ranges between 0 and 63. For the participants in WEMAC, only those below 20–i.e. considered as recovered from PTSD– were allowed to participate, to avoid the risk of re-victimization.

3.2. Preprocessing

WEMAC[2] contains data from 39 GBVV and 104 non-GBVV. We use a balanced dataset of 39 GBVV and 39 non-GBVV, split into five age-matched age groups, divided equally in GBVV and NGBVV in each group. In the first age group, defined as G1 $(18-24)$, we use data from 12 volunteers, in G2 $(25-34)$ from 14, in G3 $(35-44)$ from 20, in G4 $(45-54)$ from 24, and in G5 (≥ 55) from 8. The speech signals are down-sampled at 16 kHz and normalized per user using a z-score normalization. The distribution of 1s windows per group is shown in Table 1.

	GBVV	NGBVV	**Total**
Fear	9,954	7,296	17,250
Non-Fear	15,549	10,104	25,653
Total	25,503	17,400	**42,903**

Table 1: *Distribution for data samples of* 1 s *windows used, according to the self-reported emotional labels.*

3.3. Feature Extraction and Selection

The feature extraction process is coded in Python[3].

- *librosa* [19] [20]: 19 features are extracted with the `librosa` Python toolkit (13 Mel-Frequency Cepstral Coefficients, Root Mean Square or Energy, Zero Crossing Rate, Spectral Centroid, Spectral Roll-off, Spectral Flatness, and Pitch). The mean and standard deviation for each feature are aggregated resulting in 38 speech features.
- *eGeMAPS* [21]: 88 functionals from 13 Low-Level Descriptors (LLDs) related to speech and audio are extracted through the `openSMILE` Python toolkit [22] on its default configuration, *e.g.* $f0$, harmonic features, HNR (Harmonics-to-Noise Ratio), jitter, shimmer loudness, spectral slope, formants, harmonics, Hammarberg Index, Alpha ratio, etc.
- *VGGish*: 128-dimensional embeddings from the output layer of the VGG-19 network trained for AudioSet [23].
- *PASE+* [24]: 256-dimensional features from the PASE+ (*Problem Agnostic Speech Encoder+*) encoder network, used as a speech feature extractor.

3.4. Experiments

All the experiments are performed with a Multilayer Perceptron (MLP)–coded in Python using `sklearn`– which yielded good results for the task. It consists of 5 hidden layers with 100 hidden units each, found to be optimal after a delimited hyperparameter search with layers $\{5, 7, 10\}$ and neurons $\{50, 100, 150, 200\}$. The training uses 250 epochs maximum–early stopping when the loss did not improve over 0.001 for 10 epochs–, with the following data splitting strategies:

1. 5-fold cross-validation (5FCV). Random split of the data in 5 folds, using 1 of the 5 splits for testing in each iteration.
2. Leave One Subject Out (LOSO). The data of one subject is left out during the training phase and then used for testing. This means that, since we have 78 users, 78 iterations are made, each with the data of one user as the testing set. This strategy helps us evaluate how much the model relies on speaker identification to make its GBVVC prediction.
3. Leave One Video Out (LOVO). The data corresponding to the speech recorded after the same stimulus for all volunteers is separated for testing. We have then 14 iterations, one per speech recorded after each video used in the test set. This strategy helps us evaluate how much the model relies on the self-reported emotions to make its GBVVC prediction.

All these training strategies are used in order to predict the following outputs: 1. Binary GBVV vs. non-GBVV. 2. Binary fear vs. non-fear emotion–as the dataset is balanced for such emotion–. 3. Multi-class User ID, to identify the subject (note that the LOSO split cannot be carried out for this output since the model cannot predict a label never seen before).

Due to the exceptional results obtained in the 5FCV strategy (see Table 2), one extra experiment is performed only for the GBVVC detection, with the `librosa` feature set and 5FCV and LOSO strategies, to check for consistency: User ID-consistent label randomization, that is, giving all the samples of the same user, the same random label (GBVV/non-GBVV).

Besides that, as the self-reported emotion in WEMAC was labelled at audio signal level, a majority voting system is implemented based on the `librosa` LOSO experiment to give one final label (GBVV or Non-GBVV) to each new subject the model has never seen. This is in line with developing subject-independent future AI therapy assistants.

4. Results

4.1. Librosa

The results for this feature set (Table 2) for the 5FCV are acutely high, with scores over 90% in all three tasks. Given that the data split is random, this may result from temporally-contiguous samples being both in train and test sets.

In the LOSO and LOVO splits, it can be guaranteed that the model is not trained and tested with parts of the same audio signal, and regarding the GBVVC, the results are fairly acceptable. We follow our hypothesis from [16] of the existence of different clusters of similar users among the GBVV and non-GBVV. Since the new sample is bigger, we expect to achieve better generalization and improve the previous score. When performing the same experiment but using one averaged value per speech signal instead of values per second, the overall score is $73.40 \pm 27.78\%$ for the 52 users present in [16], and $67.38 \pm 34.17\%$ for all 78 users. That means our result is improved 2.61% while reducing *std* by 15.43% relative, although the high variability within the new users included implies a lower score in the new group. This result is shown per subject in Figure 1. The first part of the graph corresponds to the accuracy obtained with the newly added subjects. Despite the high variability, the overall results show an improvement as we mentioned before.

[2]The database is in process of being fully released by in `https://edatos.consorciomadrono.es/dataverse/empatia`.

[3]Code for feature extraction available in: `https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing`.

[4]Since for half of the iterations there is no positive label on the test set (the

		GBVV vs. Non-GBVV			Fear vs. Non-Fear			User ID	
		5FCV	LOSO	LOVO	5FCV	LOSO	LOVO	5FCV	LOVO
librosa	**Acc**	$\mathbf{99.89 \pm 0.04}$	$\mathbf{65.14 \pm 30.67}$	$\mathbf{96.80 \pm 1.41}$	$\mathbf{94.21 \pm 0.41}$	53.61 ± 9.09	$\mathbf{54.28 \pm 8.32}$	$\mathbf{96.68 \pm 0.23}$	$\mathbf{81.25 \pm 4.54}$
	F1	$\mathbf{99.91 \pm 0.03}$	$\mathbf{70.80}$ [4]	$\mathbf{97.28 \pm 1.22}$	$\mathbf{92.70 \pm 0.47}$	31.69 ± 17.41	39.64 [4]	$\mathbf{98.68 \pm 0.23}$	$\mathbf{79.53 \pm 4.96}$
eGeMAPS	**Acc**	70.52 ± 0.34	51.11 ± 14.8	71.38 ± 0.96	57.69 ± 0.76	$\mathbf{59.49 \pm 7.43}$	53.68 ± 12.07	32.50 ± 0.27	28.50 ± 1.35
	F1	75.62 ± 0.42	60.74 [4]	76.39 ± 1.45	28.74 ± 1.16	24.85 ± 7.72	37.05 [4]	31.09 ± 0.59	25.40 ± 1.45
VGGish	**Acc**	67.36 ± 0.35	52.91 ± 14.51	66.80 ± 1.41	55.08 ± 0.91	54.47 ± 5.77	51.28 ± 7.01	21.75 ± 0.54	19.21 ± 1.07
	F1	72.91 ± 0.49	62.96 [4]	72.57 ± 1.72	37.41 ± 1.97	$\mathbf{34.00 \pm 8.84}$	37.76 [4]	19.49 ± 0.48	15.63 ± 0.97
PASE+	**Acc**	89.39 ± 0.30	53.01 ± 21.79	86.70 ± 1.30	61.64 ± 0.48	56.78 ± 8.63	52.72 ± 6.83	65.00 ± 0.79	54.86 ± 2.11
	F1	91.16 ± 0.27	63.58 [4]	88.84 ± 1.27	42.77 ± 1.13	30.03 ± 11.95	39.58 [4]	64.88 ± 0.75	53.17 ± 2.12

Table 2: *Metrics (mean % ± std %) for the* `librosa`*,* `eGeMAPS`*,* `VGGish` *and* `PASE+` *feature sets with 1s windows as input.*

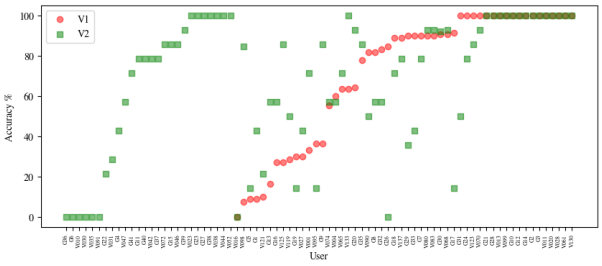

Figure 1: `librosa` *classification results per user, LOSO split and averaged samples. In green, the accuracy scores obtained per user with the proposed experiments–78 users–(V2); in red, those obtained in [16]–52 users–(V1) aligned to the right.*

As to the fear/non-fear emotions classification, our model is slightly above chance in the splits in which it does not observe samples from the same audio signal in training and test sets, so we can conclude that the model and features used are not well suited for such task.

The LOVO experiment eliminates the possibility of the same audio being in train and test sets simultaneously while keeping a subject-dependent strategy as in 5FCV. This experiment draws metrics almost as high as the 5FCV when predicting the GBVVC, while decreasing a 16.06% w.r.t. user ID prediction. This suggests that this model relies both in speaker information as well as in the differences between GBVV and non-GBVV in order to draw a prediction. It is relevant to mention, as well, that the little variability shown when detecting the GBVVC in this experiment proves that the emotion that the subject is experiencing–which varies from one video to another–does not affect the prediction of the GBVVC.

4.1.1. Label Randomization

The results shown in Table 3 display the experiments in which all data from the same subject is given a random label (either GBVV/non-GBVV) no matter their real label. These results show how, when the model can train on samples of the user/audio tested (5FCV), it can still predict the GBVV label even though it is a random fake one. However, when the model cannot rely on those similarities (LOSO), it cannot find any relation between the evaluated user and those of its group, now that those groups are not real.

These results imply that there are indeed underlying characteristics in speech that distinguish GBVV and Non-GBVV and can be captured by ML models.

user is non-GBVV or the video non-fear), the F1-score cannot be calculated per iteration and averaged; thus it was calculated on the concatenated predictions of all iterations and, therefore, *standard deviation* (*std*) cannot be computed.

	5FCV	LOSO
Acc	99.83 ± 0.06	42.37 ± 30.08
F1	99.85 ± 0.05	49.38 [4]

Table 3: *Performance scores (%) for the* `librosa` *feature set when randomizing the labels for the task of GBVVC prediction.*

4.1.2. Majority Voting (MV)

Aiming to settle foundation for speaker-agnostic AI therapy assistants, all 1 s predictions are combined in a MV system (MVS) which gives a final label (GBVV or not) per user. Its accuracy for all 78 users is 73.08%, improving the results when the input was both 1 s samples and the averaged value.

4.1.3. Correlation with Psychological Evaluation

With the aim of explaining the reason behind the variability shown in Figure 1, the correlation between the accuracy of the presented model per user with their psychological evaluation is explored (Figure 2)[5]. The correlation map is done between the users of the central region of Figure 1–those with high variability, without extreme values of accuracy below 10% or over 90%. Given the limitations of this result–the psychological evaluation was only available for GBVV who have their trauma and aftermath under control–the matrix shows some correlation (0.36) between the EGS-R score and the accuracy obtained with the model, implying that it is possible that pre-clinical PTSD symptoms are, somehow, reflected in the GBVV's voice and that it helps the model to classify them with a higher accuracy. This suggests that if we can detect GBVV who are not traumatised and have controlled sequelae, then the differentiation between victims with actual PTSD and non-victims may even be more significant.

4.2. eGeMAPS

The features from eGeMAPS feature set were chosen by their authors because of their potential with affective physiological changes in voice production, their theoretical meaning and their proven value in former studies [21]. Regarding emotion classification, these features provide the best results in the LOSO split according to the accuracy metric. However, *F1* results show that the model could be using the *a priori* distributions to maximise accuracy. For speaker identification the model decays abruptly, so `eGeMAPS` cannot identify the user either in this problem (nor was this intended by its authors).

In the GBVV vs non-GBVV classification, the results are lower compared with `librosa`, but still interesting, given the

[5] The *psychological consequences* category includes and implies the presence of one or more of the following categories: past dissociation, self-harm, depression, anxiety, eating disorder.

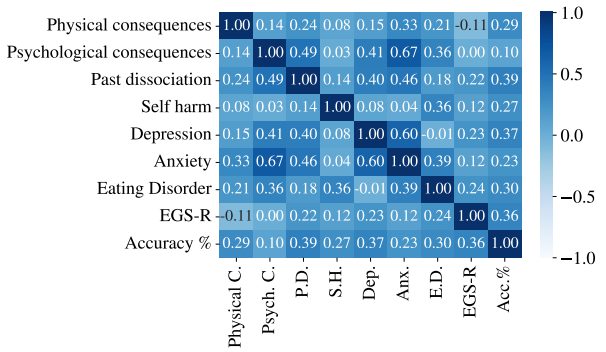

Figure 2: *Correlation map between physical and psychological conditions and the model's accuracies presented in Figure 1.*

fact that the features are not meant for identifying the speaker and yet the performance is high ($\sim$ 70%) in both 5FCV and LOVO settings. This supports the hypothesis of noticeable differences in the voice of those 2 groups.

4.3. VGGish

Since this is a feature set intended for general audio, and not speech specific [23], it makes sense that it provides the lowest metrics: it does a rather random classification between fear and non-fear emotions and gives poor results for user ID. However, it still finds a distinction between GBVV and non-GBVV within the 5FCV and LOVO splits, backing up once again the hypothesis of the differences between those groups.

4.4. PASE+

This feature set, designed to be *problem agnostic*, works better than `eGeMAPS` and `VGGish`, but not than `librosa`. However, it still gets its highest score for the GBVV vs. non-GBVV task, just as the previous feature sets.

5. Discussion

In the experiments presented in this paper, summarized in Table 2, our model is relatively capable of distinguishing between GBVV and non-GBVV. The scores obtained depend highly on the feature set as well as on the data split chosen:

- `librosa` achieves the highest scores. They suggest that this model relies both in speaker identification and differences between GBVV and Non-GBVV for such classification.
- `eGeMAPS` and `VGGish`, although unable to identify the speaker and, thus, to rely on that information; still manage to give good predictions of the GBVVC in the 5FCV and LOVO experiments.
- `PASE+` outperforms the previous two, but not `librosa`. It scores high in the prediction of the GBVVC, presumably relying as well on speaker identification.

Comparing the presented results with our prior work in this topic [16], confirms the hypothesis on the existence of clusters within groups that may not be well represented in the sample due to the lack of data. It does so by increasing their overall accuracy by 2.61% when enlarging the sample. However, a user-by-user comparison makes evident that some of them improve whereas others, in fact, worsen. This suggests that the use of an ensemble model with some kind of *bootstrapping* that uses several models trained with different subgroups of the population could improve the obtained metrics in the LOSO strategy, which will be considered for future work. Also, ID-consistent randomization of the labels was implemented to rule out the possibility that the model was relying on the identification of the speaker to decide about the GBVVC.

Working towards an AI therapy assistant capable of predicting the GBVVC even when new subjects do not identify as such, an MVS was implemented. Such system was capable of outputting a final label for each user without prior information on her speech (LOSO split) with a $\sim$ 73% confidence.

To understand what the model was basing its decision on, for the GBVV detection, the correlation between accuracy and physical and psychological conditions per GBVV was explored. Although such correlation is very limited due to the previously discussed characteristics of the GBVV, it shows some correlation between the pre-clinical symptoms of PTSD (EGS-R) and the accuracy of the prediction, which constitutes an important branch for future work: exploring correlations of actual PTSD-GBVV. However, this poses ethical constraints and thorough care should be put to avoid re-victimizing them. Besides, non-GBVV evaluations would also be relevant so as to be sure they do not have PTSD or other psychological conditions due to any other traumatic event.

6. Conclusions

Our current study shows that it is possible to distinguish GBVV from Non-GBVV by the extraction of features of their speech. To that purpose, the `librosa` feature set has been proven the best and, although the emotions felt by the users at the moment of speech do not affect the GBVVC recognition, the experiments suggest that the model does rely slightly on user identification. However, the model is still capable of making an acceptable distinction when the ID factor is removed and a MVS was implemented to classify new users with a confidence of 73.08%. Regarding the prevalence of psychological conditions in the speech of GBVV and being this the reason to be able to classify them, the results of the correlations point to such direction. However, they are not fully conclusive since the GBVV we are working with are those with controlled aftermath, so we should take into account that our scope is limited. For future work, we leave to explore multi-task learning and adversarial training strategies to disentangle the speaker information from the GBVVC, in order to develop speaker-agnostic GBVVC detection; as well as the use of personalization techniques to study whether different speakers have particular ways to express the GBVVC in speech. We will also explore the use of linguistic features alongside the acoustic ones and the impact of the silence windows in the classification.

This paper presents a novel study in a topic which has not been deeply studied before, to the knowledge of the authors, and gives promising results that open the way for GBVV assistance in mental health therapy and early diagnosis in a non-invasive, non-re-victimizing way.

7. Acknowledgements

This work has been partially supportedby the SAPIENTAE4Bindi Grant PDC2021-121071-I00 funded by MCINAEI10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR", PID2021-125780NB-I00 funded by AEI, the Spanish Ministry of Science, Innovation and Universities with the FPU grant FPU19/00448, Portuguese national funds through FCT (UIDB/50021/2020), and project C645008882-00000055. The authors thank all the members of UC3M4Safety for their contribution and support.

8. References

[1] S. Oram, H. Khalifeh, and L. M. Howard, "Violence against women and mental health," *The Lancet Psychiatry*, vol. 4, no. 2, pp. 159–170, 2017.

[2] V. Escribà-Agüir, I. Ruiz-Pérez, I. Montero, C. Vives-Cases, J. Plazaola-Castaño, and D. Martin-Baena, "Partner violence and psychological well-being: Buffer or indirect effect of social support," *Psychosomatic medicine*, vol. 72, pp. 383–9, 04 2010.

[3] G. Ferrari, R. Agnew-Davies, J. Bailey, L. Howard, E. Howarth, T. Peters, L. Sardhina, and G. Feder, "Domestic violence and mental health: A cross-sectional survey of women seeking help from domestic violence support services," *Global health action*, vol. 7, p. 25519, 10 2014.

[4] A. B. Ludermir, L. B. Schraiber, A. F. D'Oliveira, I. França-Junior, and H. A. Jansen, "Violence against women by their intimate partner and common mental disorders," *Social Science & Medicine*, vol. 66, no. 4, pp. 1008–1018, 2008.

[5] S. Rees, D. Silove, T. Chey, L. Ivancic, Z. Steel, M. Creamer, M. Teesson, R. Bryant, A. C. McFarlane, K. L. Mills, T. Slade, N. Carragher, M. O'Donnell, and D. Forbes, "Lifetime Prevalence of Gender-Based Violence in Women and the Relationship With Mental Disorders and Psychosocial Function," *JAMA*, vol. 306, no. 5, pp. 513–521, 08 2011.

[6] J. Chandan, T. Thomas, C. Bradbury-Jones, R. Russell, S. Bandyopadhyay, K. Nirantharakumar, and J. Taylor, "Female survivors of intimate partner violence and risk of depression, anxiety and serious mental illness," *The British journal of psychiatry : the journal of mental science*, vol. 217, pp. 1–6, 06 2019.

[7] S. Shen and Y. Kusunoki, "Intimate partner violence and psychological distress among emerging adult women: A bidirectional relationship," *Journal of Women's Health*, vol. 28, no. 8, pp. 1060–1067, 2019.

[8] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–28, 2022.

[9] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim, and D. V. Jeste, "Artificial intelligence for mental health and mental illnesses: an overview," *Current psychiatry reports*, vol. 21, pp. 1–18, 2019.

[10] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is speech the new blood? recent progress in ai-based disease detection from audio in a nutshell," *Frontiers in Digital Health*, vol. 4, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2022.886615

[11] D. Low, K. Bentley, and S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, 01 2020.

[12] Y. Li, Y. Lin, H. Ding, and C. Li, "Speech databases for mental disorders: A systematic review," *General Psychiatry*, vol. 32, 2019.

[13] J. Gratch, R. Arstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L. Morency, "The distress analysis interview corpus of human and computer interviews." 01 2014.

[14] M. P. Duckworth and V. M. Follette, *Retraumatization: Assessment, treatment, and prevention*. Routledge, 2012.

[15] A. Jaffe, D. DiLillo, K. Gratz, and T. Messman, "Risk for revictimization following interpersonal and noninterpersonal trauma: Clarifying the role of posttraumatic stress symptoms and trauma-related cognitions," *Journal of Traumatic Stress*, 02 2019.

[16] E. Reyner Fuentes, E. Rituerto González, C. L. Mingueza, C. Peláez Moreno, and C. López Ongil, "Detecting Gender-based Violence aftereffects from Emotional Speech Paralinguistic Features ," in *Proc. IberSPEECH*, 2022, pp. 96–100.

[17] J. A. Miranda, E. Rituerto-González, L. Gutiérrez-Martín, C. Luis-Mingueza, M. F. Canabal, A. R. Bárcenas, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil, "WEMAC: Women and Emotion Multi-modal Affective Computing dataset," 2022. [Online]. Available: https://arxiv.org/abs/2203.00456

[18] E. Echeburúa, P. J. Amor, B. Sarasua, I. Zubizarreta, F. P. Holgado-Tello, and J. M. Muñoz, "Escala de Gravedad de Síntomas Revisada (EGS-R) del Trastorno de Estrés Postraumático según el DSM-5: propiedades psicométricas," *Terapia Psicológica*, vol. 34, pp. 111 – 128, 07 2016.

[19] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa, audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 01 2015, pp. 18–24.

[20] B. McFee *et al.*, "librosa/librosa: 0.10.0," Feb. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7657336

[21] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, pp. 1–1, 01 2015.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile – the munich versatile and fast open-source audio feature extractor," 01 2010, pp. 1459–1462.

[23] J. F. Gemmeke, D. P. W. 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. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[24] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," 2020.