IberSPEECH 2022
14-16 November 2022, Granada, Spain

Detecting Gender-based Violence aftereffects from Emotional Speech Paralinguistic Features

Emma Reyner-Fuentes[1], *Esther Rituerto-González*[2], *Clara Luis-Mingueza*[2], *Carmen Peláez-Moreno*[2], *Celia López-Ongil*[1]

[1]Department of Electronics, University Carlos III of Madrid, Spain
[2]Department of Signal Theory and Communications, University Carlos III of Madrid, Spain
ereyner@pa.uc3m.es, erituert@ing.uc3m.es, cluis@pa.uc3m.es, carmen@tsc.uc3m.es, celia@ing.uc3m.es

Abstract

Speech is known to provide information regarding the person speaking, such as their gender, identity, emotions, and even disorders or trauma. In this paper we aim to answer the following question, can women who have suffered from gender-based violence (GBV) be distinguished from those who have not, just by using speech paralinguistic cues?

In this work, we intend to demonstrate whether there exist measurable differences between the emotional expression in the voice of GBV victims (GBVV) and non-victims (Non-GBVV). The present study was carried out in the framework of the project EMPATIA-CM, whose aim is to understand the reaction of GBVV to dangerous situations and develop automatic mechanisms to protect them. For this purpose, we use data collected and partly published from the WEMAC Database, a multimodal database containing physiological and speech data from women who have and have not suffered from GBV while visualizing different emotion-eliciting video clips. With the performed analysis, it is proven that such differences exist indeed and, therefore, that suffering from GBV alters the way women react to the same emotion eliciting stimulus in terms of physical variables, specifically certain voice features.

Index Terms: Gender-based violence, speech emotion recognition, speech features, post-traumatic alterations, paralinguistics

1. Introduction

According to UNICEF and WHO, about 1 in 3 women and girls worldwide will suffer from gender-based violence (GBV) at least once in their lifetime [1]. Only in Spain, from 2003 to 2021 more than 1000 women have been murdered due to GBV. There is, thus, a need to fight this violence. The numbers prove that more effort needs to be made to seek solutions from the educational, legislative and technological spheres [2].

EMPATIA-CM[1], the framework of the present study, is a project carried out by the UC3M4Safety[2] team that seeks prevention and protection for the victims of this GBV by the use of Artificial Intelligence (AI). The general goal of the project is to understand the reactions of GBV victims (GBVV) when they face risky situations, to generate automatic detection mechanisms for these situations and to study how to efficiently react to protect them. The project includes the development of wereable devices which are able to use physical and physiological variables to automatically detect certain user's emotions, present when a potentially dangerous situation is happening. The lack of proper data to train such AI system lead to the creation of the WEMAC Database [3] where emotional reactions were captured by physical and physiological data collected while receiving a set of emotion-related audiovisual stimuli [4]. Women volunteers belonging to two different main groups: gender-based violence victims (GBVV) –with low levels of Post Traumatic Stress Disorder (PTSD)[3]–, and non-victims (Non-GBVV); have participated in the creation of this database and it is this condition, which distinguishes between the two groups, that we want to find out if we can classify using speech data..

[1]https://www.uc3m.es/instituto-estudios-genero/EMPATIA

[2]https://www.uc3m.es/instituto-estudios-genero/UC3M4Safety

2. Related Work

It has been widely accepted [5] [6] that PTSD causes impairments when it comes to both the correct interpretation of emotions coming from others, and also the expression of the emotions of oneself – also named as emotional functioning. However, such studies include both men and women, which significantly reduces the sample for each of them, although it has been proved [7] [3] [8] that the gender influences the way emotions are expressed and the changes they cause in human bodies. Several research works [9] [10] present the study of the psychological effects of domestic violence, both in children and women, but the ones concerning emotional functioning mainly concern children and prove that it is severely affected. There is, therefore, a gap in research regarding the emotional functioning of women after experiencing GBV abuses.

One non-invasive way of assessing the emotion expression is by using SER tools (Speech Emotion Recognition tools), a topic whose literature is growing increasingly fast [11]. It is possible to classify a person's emotions in several categories, although the exact number of them differs between studies [12]. Particularly speaking about fear, –the emotion that gets the highest interest within this study– it is known that such emotion has an impact on the paralinguistic features of the speaker, according to [13] and [14], which state that the fundamental frequency of the voice can be used as an indicator of anxiety; and a higher pitch, lower pitch variation, and lower levels of spectral energy over 500 Hz, are associated with fear. Surprisingly, none of this prosodic features are selected as the best features for GBVV/non-GBVV distinction, as it will be further presented. These belong to the field of paralinguistics, a term that means *alongside linguistics*, as defined in [15], is

[3]Women with high levels of PSTD were discarded given the risk of revictimization they presented.

 10.21437/IberSPEECH.2022-20

the discipline dealing with those phenomena that are modulated onto or embedded into the verbal message, be this in acoustics or in linguistics. Thus, everything that can be found in the speech signal.

It was observed by the professionals of the IEG (Institute of Gender Studies) –who have interviewed several experts and GBVV– that GBVV react differently to emotion-eliciting stimuli. Their response is usually longer in time and activates different factors than that of the Non-GBVV. However, this was only an observation and no studies have been made in order to prove such statement. Thus, the present study intends to determine –by using paralinguistic features– whether there exists a measurable difference between the expression of emotions in general, and of fear in particular, within the voice of the two main groups of volunteer women participating on WEMAC Database (victims and non-victims of GBV), and to evaluate such difference based on statistical analysis and machine learning methods.

3. Methodology and Materials

3.1. Database

In this study the WEMAC Database was used [3], a novel multi-modal affective computing database with realistic emotions[4]. It consists of self-reported annotations, physiological and speech data, and emotional annotations; all of them captured in laboratory conditions from both GBVV and Non-GBVV. The collection of data was done while the participants were presented several audiovisual stimuli with a Virtual Reality (VR) Headset in an immersive setup for emotion elicitation. When the visualization finished, two questions were to be answered aloud by the participant and were recorded for collection of speech data. The database is partially released, with the data of 47 out of 104 non-victims available for use.

3.2. Feature Extraction and Selection

The feature extraction process conducted for the speech data in the WEMAC dataset is coded in Python[5] and includes the features presented in Table 1:

Table 1: *Features extracted from the speech data in the WEMAC Database.*

Feature set	Number of Features	Kind of Features
librosa	38	Frequency, Energy, Temporal features (MFCC,...)
eGeMaps	88	Frequency, Energy, Temporal features
ComParE	6,373	Frequency, Energy, Voicing features
DeepSpectrum ResNet50	2,048	Visual representation of the audio (spectrogram, chromogram) features
DeepSpectrum VGG19	4,096	Spectral features
VGGish	128	Spectral features

However, all the feature sets extracted would add up to more than 12000 features, which would make a difficult and heavy statistical analysis. As this is the first work of its kind –to the knowledge of the authors–, only the `librosa` [16] feature set was chosen due to adequate number of features, to their physical and biological meaning, and to its proof of validation of use in previous works for emotions and speaker recognition [17], [18], [19]. We extracted 19 features from `librosa` with a window size of 20ms and a 10ms overlap and then computed their mean and standard deviations every second resulting in 38 features. Moreover, these features are more amenable for implementation in low resource devices, one of the goals in EMPATIA.

3.3. Preprocessing

Since the number of GBVV and Non-GBVV differs, in order to make a balanced comparison, it is important to select as many Non-GBVV as GBVV available. There are 26 available GBVV volunteers that belong to the following age groups:

Table 2: *Age groups of the volunteers.*

Group	Age	Participants
G1	<25	3
G2	25-35	5
G3	35-45	8
G4	45-55	7
G5	>55	3

Therefore, the data from 26 random Non-GBVV belonging to the same age-groups as the available GBVV were selected. Such randomization was performed 10 times for statistical validation.

After normalizing the data (Z-score/Mean and std normalization performed to all users together), it was necessary to choose one set of values as the representative per speech recording since each of them have different durations. For simplicity and to further allow for a more efficient implementation in wearable devices, 3 different alternatives were chosen for representing each complete speech signal:

- Mean of each feature sequence.
- Mean and the maximum and minimum values of certain features. Since fear is the most important emotion for the framework of the study (and it causes a higher pitch, lower energy and lower pitch variance) the mean of the rest of the features was maintained while the maximum value of the pitch and minimum of both the energy and the pitch variance were taken.
- Bland-Altman (B-A) difference, which is a measurement of the dispersion of the data. The width of the 95% confidence interval of the Bland-Altman plot [20] was chosen to represent the audio in this case.

3.4. Statistical Tests and Classification

T-tests and Wilcoxon's-tests were applied between both groups (GBVV and Non-GBVV) to check if there were any features that presented significant differences between groups, and thus allowed for distinction between them. The tests were applied depending on whether the distribution of the data was normal or not, and returned a *p value* that was considered significant if <0.05 [21]. Afterwards, a shallow neural network, a Multilayer Perceptron (MLP) –coded in Python with the `sci-kit learn` library– was created in order to validate the statistical results. It was used to determine whether it is possible to classify and predict the GBV condition –victim or not victim– of a user being evaluated only from the speech paralinguistic features extracted.

[4] The database is partially released, with the data of 47 out of 104 non-victims available for use [3].

[5] `https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing`

4. Results

The statistical analyses conducted led to the discovery of the following features (Table 3) as significant:

Table 3: *Statistically significant features, selected with Wilcoxon's and t-test per type of representative value.*

Mean	Mean + Max/Min	B-A Difference
Mean Flatness	Mean Flatness	Mean Flatness
Std MFCC5	Std MFCC5	
Std MFCC6	Std MFCC6	
Std MFCC9	Std MFCC9	
Std Roll off	Std Roll off	

In all three cases there are features that allow for differentiation between groups. However, introducing the maximum and minimum values does not imply any differences on the resulted relevant features; and using the B-A difference reduces the amount of significant features by 80%, which is the reason why only the *mean* was kept as representative value for the following experiments.

As regards to the Multilayer Perceptron classification model used, it consisted of 5 hidden layers with 100 hidden units each. The training was done for 250 epochs, and the data was split in 5 folds for cross validation. Keeping in mind that the framework of the study intends to develop wereable devices with low computational and memory resources , in order to assess the performance against the number of features, different number of features were used:

A. 38 features. Mean and std of a subset of 19 features extracted from `librosa`. Starting point for B, C, D.

B. 5 significant features (5-SF). The ones which remained significant (regarding the statistical tests in Table 3) in all the 10 trials, no matter which random selection of participants was made.

C. PCA. In the line of dimensionality reduction for lightweight models, a Principal Component Analysis (PCA) was performed in order to see the efficiency of the model when using 5 principal components (5-PCA) instead of the 5 significant features.

D. 5 non-significant features (5-NoSF). The 5 features which never came out as significant, no matter which random selection of participants.

For each of the above features subset groups, two approaches were implemented: a random 5-fold cross validation and thus subject-dependent strategy, and a LOSO (Leave One Subject Out)[6] cross validation strategy, using subject-independent (SI) training leaving all the instances of each user out each time. No intra-subject inter-session variability is reported due to the small amount of data available per subject and session.

The results obtained for each approach and experiment are depicted in Table 4 and complemented with Figures 1-4.

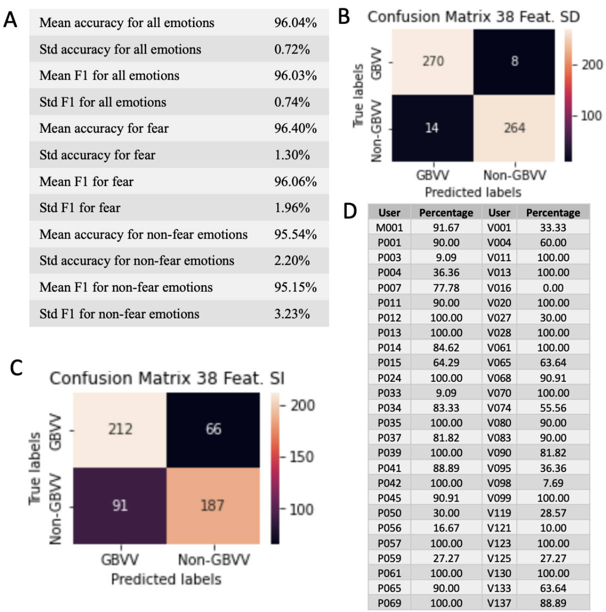
They show that the MLP model is capable of distinguishing between GBVV and Non-GBVV with a subject-dependent approach better than with the subject-independent approach. The use of 38 features gives the best scores, followed closely by the PCA dimensionality reduction to 5 principal components. When 5 significant features were used, the scores diminished, although the MLP was still capable of performing well, The use of the 5 less significant features, as expected, gave the worse score.

[6]Performed with the *LOGO* (Leave One Group Out) function of `sk-learn`.

Table 4: *Performance scores (%) for the Subject Dependent (SD) and Subject Independent (SI) experiments for each feature subset.*

		38 features	5-SF	5-PCA	5-NoSF
Subject Dep. (SD)	Acc	**96.04 ± 0.72**	78.05 ± 1.90	95.87 ± 1.66	71.40 ± 2.63
	F1	**96.03 ± 0.74**	79.31 ± 1.58	95.84 ± 1.74	71.25 ± 3.00
Subject Indep. (SI)	Acc	71.53 ± 32.85	59.38 ± 32.78	55.65 ± 33.50	52.96 ± 30.87
	F1	40.61 ± 44.93	34.75 ± 41.19	29.56 ± 39.12	31.57 ± 38.11

Another experiment was made for each of the approaches: testing the model when using only the *fear* emotion vs. when using only non-fear emotions. The scores obtained for both cases, as shown in Figure 1-A for the 38 features approach, overlap. This means that separating fear from all the other emotions does not necessarily allow for a better separation between GBVV and Non-GBVV groups. This prevails in all the experiments. However, we need to take into account that separating both types of emotion, leaves less instances to train the models and thus are not fully comparable.



Mean accuracy for all emotions	96.04%
Std accuracy for all emotions	0.72%
Mean F1 for all emotions	96.03%
Std F1 for all emotions	0.74%
Mean accuracy for fear	96.40%
Std accuracy for fear	1.30%
Mean F1 for fear	96.06%
Std F1 for fear	1.96%
Mean accuracy for non-fear emotions	95.54%
Std accuracy for non-fear emotions	2.20%
Mean F1 for non-fear emotions	95.15%
Std F1 for non-fear emotions	3.23%

User	Percentage	User	Percentage
M001	91.67	V001	33.33
P001	90.00	V004	60.00
P003	9.09	V011	100.00
P004	36.36	V013	100.00
P007	77.78	V016	0.00
P011	90.00	V020	100.00
P012	100.00	V027	30.00
P013	100.00	V028	100.00
P014	84.62	V061	100.00
P015	64.29	V065	63.64
P024	100.00	V068	90.91
P033	9.09	V070	100.00
P034	83.33	V074	55.56
P035	100.00	V080	90.00
P037	81.82	V083	90.00
P039	100.00	V090	81.82
P041	88.89	V095	36.36
P042	100.00	V098	7.69
P045	90.91	V099	100.00
P050	30.00	V119	28.57
P056	16.67	V121	10.00
P057	100.00	V123	100.00
P059	27.27	V125	27.27
P061	100.00	V130	100.00
P065	90.00	V133	63.64
P069	100.00	V137	88.89

Figure 1: *Classification results of the MLP model trained with 38 features. A) Accuracies and F1 scores for SD approach separated by emotion groups. B) Confusion matrix for SD approach with all emotions. C) Confusion matrix for SI approach. D) Accuracies per user for SI approach*

5. Discussion

The statistical analysis from Table 3, already pointed to the conclusion that it is indeed possible to differentiate between GBVV and Non-GBVV only by using paralinguistic cues, due to the discovery of certain features as significant. This led to the design of the MLP that, as shown in Figures 1 A and B, when trained with all 38 features is capable of discriminating between groups with a really high accuracy in a SD experiment.

In order to test a SI experiment, a LOSO cross validation was performed, with the results shown in Figures 1 C and D. In this case, although the mean accuracy decreases, it is

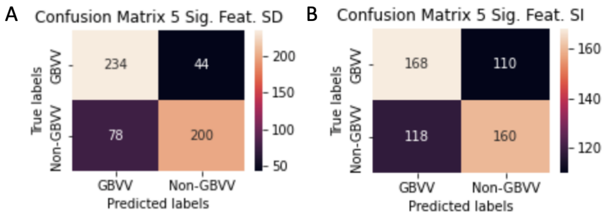


Figure 2: *Classification results of the MLP model trained with 5 significant features. A) Confusion matrix SD B) Confusion matrix SI*

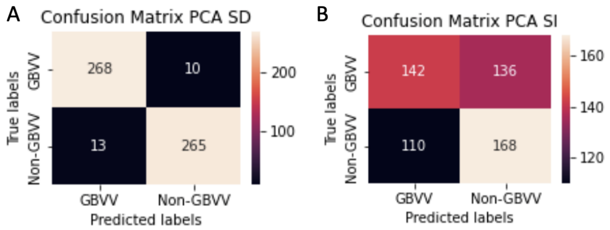


Figure 3: *Classification results of the MLP model trained with 5 dimensions of the PCA. A) Confusion matrix SD. B) Confusion matrix SI.*

still around 70%. However, the standard deviation is really high. Our hypothesis to explain such phenomenon was that there were some subjects that were outliers within our model, that is, subjects that did not follow the same patterns as the majoritarian group in our sample and, therefore, that our classifier was not capable of generalizing well on them. This was verified when observing the accuracies per user (Figure 1-D), where some users' scores were similar to the ones from the subject-dependent experiment; and other users had rather low scores. This observation prevails in all the experiments, as it is possible to see in Figs. 2, 3, 4 – B.

In Fig. 1-A we divided the scores in fear and non-fear emotions. Another aim of the study was to see whether it was easier to distinguish between victims and non victims when the participant was experiencing a certain kind of emotion. However, the results show that there is no actual difference, since the confidence intervals for the scores of both groups overlap. Again, this result prevails in all approaches, which is why it is not showed in the rest of the figures.

When reducing the 38 initial features to the 5 significant ones –a reduction of the 86.84% of the data– the results in Fig. 2 show that, although there is a reduction in the accuracy of 18.7%, accuracy and f1-score are of almost 80%, which is still a good result for classification.

The model trained with the 5 principal components of the PCA (Fig. 3), gives a much better score, close to that of the model trained with all the data. This result validates the main conclusion of the study, that it is possible to distinguish between GBVV and Non-GBVV; and also supports the idea that the classification can also be made with lower dimensional data. One may think, on a first sight, that this result reduces the importance of the previous one, that shows how the classification can also be made with 5 features. However, in the context of the project EMPATÍA-CM, which works with embedded, computationally constraint systems, it is interesting to know that only extracting 5 features it is possible to get good results, instead of needing to extract all of them and then performing the PCA transformation.

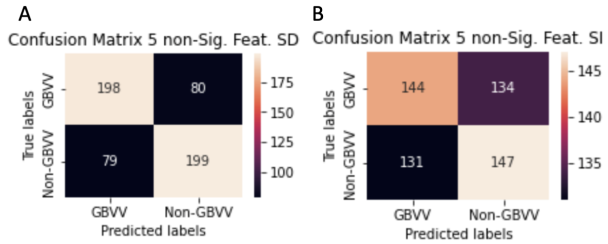


Figure 4: *Results of the MLP model trained with 5 least significant features. A) Confusion matrix SD. B) Confusion matrix SI.*

Finally, to prove that the 5 selected features are more relevant for the classification than other features, the same model was trained with 5 non- significant features. The results, depicted in Figure 4, show how the accuracies are reduced by 8.52% in the SD case, and by 10.81% in the SI case, a reduction when using the same amount of data. We hypothesize that redundancy between the 5 selected features might be the cause to not finding a big reduction, which is something to be checked in future lines of work.

Nevertheless, it needs to be taken into account that this is preliminary work to be continued, and that it has a main limitation, which is the fact that the sample contains only 26 GBVV and 26 Non-GBVV, 52 subjects in total. To enlarge this sample, more GBVV data is in process of collection, and part of the planned future work is to re-perform this analysis with a bigger sample once it is available. It would also be interesting to keep more than one representative value per video for the classification to have more data per user, to use embeddings from sequence models or try other different kinds of features. It is important to mention too that some attention must be drawn in the future to the fact that false positive and negative errors may not have the same relevance. This is why confusion matrixes were presented alongside with scores, but some ponderation in the scores may as well be considered.

6. Conclusions

The main goal of this study was to determine if, by using the paralinguistic cues of their speech when experiencing emotions, it was possible to distinguish between GBVV and Non-GBVV.

First, a statistical analysis was conducted, which evidenced that it is possible to do so. To corroborate such hypothesis, a MLP was designed and trained with all the available data; and it turned out to be capable of classifying the condition of GBVV with a really high accuracy in a subject-dependent experiment. When removing such dependency, the scores diminished significantly, which we believe it is explained by the existence of outliers and the small amount of data. Because of the EMPATIA-CM framework and its use of computationally constraint embedded systems, a dimensionality reduction was made by taking the 5 features found most significant within the statistical analysis. The classification scores in this case are slightly lower, but still capable of performing well with a really big reduction of the dimensionality of the feature set. To sum up, this preliminary work indicates that it is possible to distinguish between GBVV and Non-GBVV by using their paralinguistic cues. These promising results open the way for GBVV assistance in mental health diagnosis and therapy.

7. Acknowledgements

This work has been supported by the Dept. of Research and Innovation of Madrid Regional Authority, in the EMPATIA-CM research project (reference Y2018/TCS-5046), SAPIENTAE4Bindi Project from Grant PDC2021-121071-I00 funded by MCINAEI10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR", grant PID2021-125780NB-I00 funded by AEI, and the Spanish Ministry of Science, Innovation and Universities with the FPU grant FPU19/00448. The authors thank all the members of the UC3M4Safety for their contribution and support.

8. References

[1] UNICEF, "Gender-based violence in emergencies," Mar. 2021. [Online]. Available: https://www.unicef.org/protection/gender-based-violence-in-emergencies

[2] Ministerio de Igualdad, "Mujeres Víctimas Mortales por Violencia de Género en España a manos de sus Parejas o Exparejas. Datos Provisionales." Jan. 2022. [Online]. Available: https://violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales/fichaMujeres/pdf/VMortales_25012022_Act2021.pdf

[3] 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, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/2203.00456

[4] M. Blanco Ruiz, L. Gutiérrez Martín, J. A. Miranda Calero, M. F. Canabal Benito, E. Romero Perales, C. Sainz De Baranda Andújar, R. San Segundo Manuel, D. Larrabeiti López, C. Peláez Moreno, and C. López Ongil, "UC3M4Safety Database - List of Audiovisual Stimuli (Video)," 2021, type: dataset. [Online]. Available: https://edatos.consorciomadrono.es/citation?persistentId=doi:10.21950/LUO1IZ

[5] C. R. Brewin and E. A. Holmes, "Psychological theories of posttraumatic stress disorder," *Clinical Psychology Review*, vol. 23, no. 3, pp. 339–376, May 2003, number: 3. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0272735803000333

[6] E. Poljac, B. Montagne, and E. H. de Haan, "Reduced recognition of fear and sadness in post-traumatic stress disorder," *Cortex*, vol. 47, no. 8, p. 974–980, 2011.

[7] D. L. Robinson, "Brain function, emotional experience and personality," *Netherlands Journal of Psychology*, vol. 64, no. 4, pp. 152–168, Dec. 2008, number: 4. [Online]. Available: http://link.springer.com/10.1007/BF03076418

[8] M. Blanco-Ruiz, C. Sainz de Baranda, L. Gutierrez Martin, E. Romero-Perales, and C. Ongil, "Emotion elicitation under audiovisual stimuli reception: Should artificial intelligence consider the gender perspective?" *International Journal of Environmental Research and Public Health*, vol. 17, p. 8534, 11 2020.

[9] C. M. Chemtob and J. G. Carlson, "Psychological Effects of Domestic Violence on Children and Their Mothers." *International Journal of Stress Management*, vol. 11, no. 3, pp. 209–226, 2004, number: 3. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/1072-5245.11.3.209

[10] D. E. Logan and S. A. Graham-Bermann, "Emotion Expression in Children Exposed to Family Violence," *Journal of Emotional Abuse*, vol. 1, no. 3, pp. 39–64, Sep. 1999, number: 3. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1300/J135v01n03_03

[11] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *2011 Defense Science Research Conference and Expo (DSR)*. Singapore, Singapore: IEEE, Aug. 2011, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/6026823/

[12] Analytics Insight, "Speech Emotion Recognition (SER) through machine learning," Jul. 2020. [Online]. Available: https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/

[13] J. W. Weeks, C.-Y. Lee, A. R. Reilly, A. N. Howell, C. France, J. M. Kowalsky, and A. Bush, ""The Sound of Fear": Assessing vocal fundamental frequency as a physiological indicator of social anxiety disorder," *Journal of Anxiety Disorders*, vol. 26, no. 8, pp. 811–822, Dec. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S088761851200093X

[14] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual Cues in Nonverbal Vocal Expressions of Emotion," *Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1080/17470211003721642

[15] B. Schuller and et al., "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *The INTERSPEECH 2014*, 2014.

[16] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 01 2015, pp. 18–24.

[17] E. Rituerto-González, A. Mínguez-Sánchez, A. Gallardo-Antolín, and C. Peláez-Moreno, "Data augmentation for speaker identification under stress conditions to combat gender-based violence," *Applied Sciences*, vol. 9, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2076-3417/9/11/2298

[18] E. Rituerto-González, C. Luis-Mingueza, and C. Pelález-Moreno, "Using Audio Events to Extend a Multi-modal Public Speaking Database with Reinterpreted Emotional Annotations," in *Proc. IberSPEECH 2021*, 2021, pp. 61–65. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2021-13

[19] J. A. Miranda, E. Rituerto-González, C. Luis-Mingueza, M. F. Canabal, A. R. Bárcenas, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil, "Bindi: Affective internet of things to combat gender-based violence," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[20] D. Giavarina, "Understanding Bland Altman Analysis," *Biochemia medica*, vol. 25, no. 2, pp. 141–151, 2015, ISBN: 1330-0962.

[21] J. Armat Rodrigo, "Test de wilcoxon mann whitney como alternativa al t-test." 2017. [Online]. Available: {https://www.cienciadedatos.net/documentos/17_mann-whitney_u_test}