Multimodal Affective Computing in Wearable Devices with Applications in the Detection of Gender-based Violence

by

Esther RITUERTO GONZÁLEZ

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor in

Multimedia and Communications

UNIVERSIDAD CARLOS III DE MADRID

Advisor:
Dr. Carmen PELÁEZ MORENO

December 15, 2022

Dedicated to all women that aim to achieve a better world

Acknowledgements

El castellano es mi idioma materno, y con él me siento más cerca de las personas a las que quiero dar las gracias.

En la Universidad Carlos III, quiero agradecer a todas las personas que me han alentado en mis estudios y me han motivado en mi carrera. En especial, a mi directora de tésis Carmen Peláez Moreno, por su dedicación y su supervisión durante estos años. Le agradezco sinceramente su guía, apoyo y motivación, por proporcionarme la orientación y el consejo que he necesitado para completar este programa de doctorado. También quiero dar las gracias a Celia López Ongil, por dirigir el primer proyecto en el que se enmarca esta tésis; a Rosa San Segundo, por la formación que me ha proporcionado en feminismo; y a Jose Miranda Calero, por ser un ejemplo de apoyo, dedicación y buen trabajo.

Por acogerme durante mi estancia en Alemania, quiero dar las gracias al Prof. Björn Schuller, por su valiosa orientación, y a las personas dentro y fuera del equipo de investigación que hicieron de la experiencia una de las más enriquecedoras de mi vida. En especial a Adrià Mallol Ragolta, por las gratificantes charlas sobre la investigación y el futuro; a Meishu Song, por su inestimable ayuda; a Alice Baird, por ser un modelo a seguir de entrega, constancia y éxito; y a Manuel Milling, por su incalculable apoyo y por enseñarme lo que es el verdadero trabajo en equipo.

En cuanto a mi vida personal, no puedo dejar de agradecerles a mis amigas y amigos su apoyo incondicional en la amistad que nos une. En especial a Alba, quien siempre ha creído en mí y me inspira a querer ser mejor persona. También a mi psicóloga, Pilar, que es una de las grandes responsables de que ésta tesis haya salido adelante, por creer siempre en mí y ayudarme en mis momentos más difíciles. Además quiero dar especialmente las gracias a mi famila, por educarme en los valores de la empatía, la responsabilidad y la gratitud, por proporcionarme una base segura y alentarme en esta carrera, por darme la posibilidad de estudiar, por todo su apoyo, y por los tápers, mamá.

Y, finalmente, quiero dar las gracias a todas las mujeres que batallan cada día por conseguir un mundo mejor. La fuerza de la lucha colectiva viene de cada una de nosotras.

Funding Acknowledgements

I would like to thank the following institutions for the financial support received to complete this thesis:

- Department of Research and Innovation of Madrid Regional Authority, for the EMPATIA-CM research project (reference Y2018/TCS-5046).
- Spanish State Research Agency, for the SAPIENTAE4Bindi Project (reference PDC2021-121071-I00) funded by MCINAEI10.13039/501100011033 and by the European Union ”NextGenerationEU/PRTR”.
- Community of Madrid YEI Predoctoral Program, for the Predoctoral Research Personnel in Training (PEJD-2019-PRE/TIC-16295) scholarship.
- Spanish Ministry of Universities, for the University Teacher in Training ["*Formación de Personal Universitario (FPU)*"] grant FPU19/00448 and the Supplementary Short-stay Mobility 2020 Grant for beneficiaries of the University Teacher in Training program ["*Ayudas complementarias de movilidad Estancias Breves 2020 destinadas a beneficiarios del programa de Formación del Profesorado Universitario (FPU)*"].
- German Academic Exchange Service (DAAD), for the Short Term Grant Scholarship 2020.

Published and Submitted Content

The published and submitted content works are enumerated here. Some parts of the following publications are fully or partially included in this thesis:

Journal articles

[1] Jose A. Miranda, Esther Rituerto-González, Clara Luis-Mingueza, Manuel F. Canabal, Alberto Ramírez Bárcenas, Jose M. Lanza-Gutiérrez, Carmen Peláez-Moreno, Celia López-Ongil. (2022). BINDI: Affective Internet of Things to Combat Gender-Based Violence. In IEEE Internet of Things. (JCR Q1). Vol. 9, no. 21, pp. 21174-21193. DOI:10.1109/JIOT.2022.3177256

This publication is partially included in Abstract, Extended Summary, Chapter 1 (Secs. 1.1.3 and 1.2.2), Chapter 5 (all Secs. but 5.4.1, 5.4.2 and 5.6.1), Chapter 6 (Sec. 6.1.1) and Chapter 7 (Sec. 7.1). The material from this source included in this thesis is not singled out with typographic means.

[2] Esther Rituerto-González, Carmen Peláez-Moreno. (2021) End-to-end Recurrent Denoising Autoencoder Embeddings for Speaker Identification. In Neural Computing and Applications, Springer. (JCR Q1). Vol. 33, no. 21, pp. 14429–14439. DOI:10.1007/s00521-021-06083-7

This publication is fully included in Extended Summary and Chapter 4 (Secs. 4.2, 4.4 and 4.6). The material from this source included in this thesis is not singled out with typographic means.

[3] Esther Rituerto-González, Alba Mínguez-Sánchez, Carmen Peláez-Moreno, Ascensión Gallardo-Antolín. (2019). Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-based Violence. In Applied Sciences Journal: Special Issue "IberSPEECH 2018 Speech and Language Technologies for Iberian Languages", MDPI. (JCR Q2). Vol. 9, no. 11, pp. 2076-3417. DOI:10.3390/app9112298

This publication is fully included in Abstract, Extended Summary, Chapter 3 (Sec. 3.2.1) and Chapter 4 (Secs. 4.1 - 4.3). The material from this source included in this thesis is not singled out with typographic means.

Conference articles

[4] Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez Moreno. Affective Acoustic Scene Analysis (2022). In TECNIACUSTICA 2022 Conference. Available in: http://www.sea-acustica.es/fileadmin/Elche22/ID-65.pdf

This publication is partially included in Chapter 6 (Sec. 6.1.2). The material from this source included in this thesis is not singled out with typographic means.

[5] Clara Luis-Mingueza, Esther Rituerto-González, Carmen Peláez Moreno. Bridging the Semantic Gap with Affective Acoustic Scene Classification: an Information Retrieval-based Approach. (2022). In IBERSPEECH 2022 Conference. DOI:10.21437/IberSPEECH.2022-19

This publication is partially included in Chapter 6 (Sec. 6.1.2). The material from this source included in this thesis is not singled out with typographic means.

[6] Emma Reyner-Fuentes, Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez Moreno. Detecting Gender-based Violence Aftereffects from Emotional Speech Paralinguistic Features (2022). In IBERSPEECH 2022 Conference. DOI:10.21437/IberSPEECH.2022-20

This publication is partially included in Chapter 6 (Sec. 6.3). The material from this source included in this thesis is not singled out with typographic means.

[7] Andreas Triantafyllopoulos, Sandra Ottl, Alexander Gebhard, Esther Rituerto-González, Mirko Jaumann, Steffen Hüttner, Valerie Dieter, Patrick Schneeweiß, Inga Krauß, Maurice Gerczuk, Shahin Amiriparian, and Björn W. Schuller. Fatigue Prediction in Outdoor Running Conditions using Audio Data. In International Conference on Engineering in Medicine and Biology Conference (EMBC) 2022. DOI:10.1109/EMBC48229.2022.9871225. Available in: [Deep AI Online]

This publication is partially included in Chapter 6 (Sec. 6.2). The material from this source included in this thesis is not singled out with typographic means.

[8] Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez-Moreno. (2020). Using Audio Events to Extend a Multi-modal Public Speaking Database with Reinterpreted Emotional Annotations. In IBERSPEECH 2020 Conference. DOI:10.21437/IberSPEECH.2021-13

This publication is partially included in Chapter 3 (Secs. 3.2.2 and 3.2.3), Chapter 4 (Sec. 4.5), Chapter 5 (Secs. 5.6.1 and 5.7) and Chapter 6 (Sec. 6.1.1). The material from this source included in this thesis is not singled out with typographic means.

[9] Esther Rituerto-González, Jose A. Miranda; Manuel F. Canabal; José M. Lanza-Gutierrez; Carmen Peláez-Moreno. (2020). A Hybrid Data Fusion Architecture for BINDI: a Wearable Solution to Combat Gender-based Violence. In International Conference on Multimedia Communications, Services and Security (MCSS) 2020. DOI:10.1007/978-3-030-59000-0_17

This publication is partially included in Abstract, Extended Summary, Chapter 3 (Sec. 3.2.1), Chapter 4 (Sec. 4.2.1) and Chapter 5 (Secs. 5.1, 5.2.1, 5.2.4, 5.4 and 5.7). The material from this source included in this thesis is not singled out with typographic means.

[10] Esther Rituerto-González, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno. (2018). Speaker Recognition under Stress Conditions. In

IBERSPEECH 2018 Conference. DOI:10.21437/IberSPEECH.2018-4

This publication is partially included in Chapter 3 (Sec. 3.2.1) and Chapter 4 (Secs. 4.1 - 4.3). The material from this source included in this thesis is not singled out with typographic means.

Datasets

[11] Jose A. Miranda, Esther Rituerto-González, Laura Gutiérrez-Martín, Clara Luis-Mingueza, Manuel F. Canabal, Alberto Ramírez Bárcenas, Jose M. Lanza-Gutiérrez, Carmen Peláez-Moreno, Celia López-Ongil. (2022). WEMAC: Women and Emotion Multi-modal Affective Computing dataset. Preprint in arXiv. DOI:10.48550/arXiv.2203.00456

This publication is partially included in Chapter 3 (Sec. 3.3). The material from this source included in this thesis is not singled out with typographic means.

[11.1] 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", 2021. DOI:10.21950/CXAAHR

[11.2] 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. DOI:10.21950/LUO1IZ

[11.3] J. A. Miranda Calero, L. Gutiérrez Martín, E. Martínez Rubio, M. Blanco Ruiz, C. Sainz de Baranda Andújar, E. Romero Perales, R. San Segundo Manuel, and C. López Ongil. "UC3M4Safety Database - WEMAC: Biopsychosocial questionnaire and informed consent", 2022. DOI:10.21950/U5DXJR

[11.4] J. A. Miranda Calero, L. Gutiérrez Martín, M. F. Canabal Benito, A. Paez Montoro, A. Ramírez Bárcenas, J. M. Lanza Gutiérrez, E. Romero Perales, and C. López Ongil. "UC3M4Safety Database - WEMAC: Physiological signals", 2022. DOI:10.21950/FNUHKE

[11.5] E. Rituerto González, J. A. Miranda Calero, C. Luis Mingueza, L. Gutiérrez Martín, M. F. Canabal Benito, J. M. Lanza Gutiérrez, C. Peláez Moreno, and C. López Ongil. "UC3M4Safety Database - WEMAC: Audio features", 2022. DOI:10.21950/XKHCCW

[11.6] J. A. Miranda Calero, L. Gutiérrez Martín, E. Martínez Rubio, M. Blanco Ruiz, C. Sainz de Baranda Andújar, E. Romero Perales, B. Alboreca Fernández-Barredo, R. San Segundo Manuel, and C. López Ongil, "UC3M4Safety Database - WEMAC: Emotional labelling", 2022. DOI:10.21950/RYUCLV

Open Access Archive

[12] Björn W. Schuller, Alican Akman, Harry Coppock, Yi Chang, Alexander Gebhard, Alexander Kathan, Andreas Triantafyllopoulos, Esther Rituerto-González, Florian B. Pokorny. (2022). Climate Change and Computer Audition: A Call to Action and Overview on Audio Intelligence to Help Save the Planet. Preprint in arXiv. DOI:10.48550/arXiv.2203.06064

This publication is partially included in Chapter 6 (Sec. 6.4). The material from this source included in this thesis is not singled out with typographic means.

Other Research Merits

The following contributions were part of the research conducted during this Ph.D. degree, but they are not included in this thesis:

Communications

[13] Esther Rituerto-González. "Analysis of Biases in Artificial Intelligence Reflecting Gender Stereotypes and Social Patterns of Structural Inequality". VII International Congress of Young Researchers with a Gender Perpective at University Rey Juan Carlos (URJC), Madrid. June 2022. [Conference Program]

[14] M. Blanco Ruiz, L. Gutiérrez Martín, J. A. Miranda Calero, M. F. Canabal Benito, E. Rituerto-González, C. Luis Mingueza, J. C. Robredo García, B. Morán González, A. Páez Montoro, A. Ramírez Bárcenas, E. Martínez Rubio, E. Romero Perales, C. Sainz de Baranda Andújar, R. San Segundo Manuel, D. Larrabeiti López, c. Peláez Moreno, C. López Ongil. "UC3M4Safety Database Description", 2021. Available in: http://hdl.handle.net/10016/32481

[15] Esther Rituerto-González, Carmen Peláez-Moreno. "Characterising Fear through Acoustic Scene Classification". II International Congress R+D+I Technology for Gender Equality: Solutions, Perspectives and Challenges, at UC3M, Getafe. May 2022. [Conference Program]

[16] Esther Rituerto-González. "Open Science: A Revolutionary concept". (2021, June 28). [Blog post]

[17] Esther Rituerto-González. "The Influence of Sex and Gender in Artificial Intelligence". VI International Congress of Young Researchers with a Gender Perpective at UC3M, Getafe. June 2021. [Conference Program]

[18] Esther Rituerto-González, Carmen Peláez-Moreno. "Fear Detection in Speech". I International Congress R+D+I Technology for Gender Equality: Solutions, Perspectives and Challenges, at UC3M, Getafe. April 2021. [Conference Program]

[19] Esther Rituerto-González. "Recognizing the victim's voice in gender-based violence situations: A machine learning approach". V International Congress of Young Researchers with a Gender Perpective at UC3M, Getafe. June 2020. [Conference Program]

[20] Esther Rituerto-González. "Wearable Device to Combat Gender-based Violence". IV International Congress of Young Researchers with a Gender Perpective at UC3M, Getafe. June 2019. [Conference Program]

Abstract

According to the World Health Organization (WHO), 1 out of every 3 women suffer from physical or sexual violence in some point of their lives, reflecting the effect of Gender-based Violence (GBV) in the world. In particular in Spain, more than 1,100 women have been assassinated from 2003 to 2022, victims of gender-based violence.

There is an urgent need for solutions to this prevailing problem in our society. They may involve the appropriate investment in technological research, among legislative, educational and economical efforts. An Artificial Intelligence (AI) driven solution that made a comprehensive analysis of aspects such as the person's emotional state, plus a context or external situation analysis (e.g.: circumstances, location) and therefore automatically detect when a woman's security is in danger, could provide an automatic and fast response to ensure women's safety.

Thus, this PhD thesis stems from the need to detect gender-based violence risk situations for women, addressing the problem from a multidisciplinary point of view by bringing together Artificial Intelligence (AI) technologies and a gender perspective. More specifically, we direct the focus to the auditory modality, analysing speech data produced by the user given that voice can be recorded unobtrusively, can be used as a personal identifier and indicator of affective sates reflected in it.

The immediate response in a human being when in a situation of risk or danger is the fight-flight-freeze response. Several physiological changes affect the body: breathing, heart rate, muscle activation including the complex speech production apparatus and vocalisation characteristics, affecting our speech production. Due to all these physical and physiological changes and their involuntary nature as a result of being in a situation of risk, we considered relying on physiological signals such as pulse, perspiration, respiration, and also speech, in order to detect the emotional state of a person with the intention of recognising *fear*, which could be a consequence of being in a threatening situation. For such, we developed "Bindi". This is a an end-to-end, AI-driven, inconspicuous, connected, edge computation-based, and wearable solution targeting the automatic detection of GBV situations. It consists of two smart devices that monitor the physiological variables and the acoustic environment including voice of an individual, connected to a smartphone and a cloud system able to call for help.

Ideally, in order to build a Machine Learning or Deep Learning Artificial Intelligence system for the automatic detection of risk situations from auditory data, we would like to count on speech recorded under realistic conditions belonging to the target user.

In our first steps, we found the difficulty of the lack of suitable data available, as there were non-existent (or non-available) speech datasets of real *fear* (not acted) currently in the literature. Real, original, spontaneous, in-the-wild and emotional speech are the ideal categories we needed for our application. Therefore, we decided to choose stress as the closest emotion to the target scenario possible for data collection to be able to flesh out the algorithms and acquire the knowledge needed. Thus, we describe and justify the use of datasets containing such emotion as the starting point of our investigation. Additionally, we describe the need for the creation of our own set of datasets to fill such literature niche.

Then, members of our UC3M4Safety team captured the UC3M4Safety Audiovisual Stimuli Database, a dataset of 42 audiovisual stimuli to elicit emotions. Using them, we contributed to the community with the collection of WEMAC, a multi-modal dataset, which comprises a laboratory-based experiment for women volunteers that were exposed to the UC3M4Safety Audiovisual Stimuli Database. It aims to induce real emotions by using a virtual reality headset while the user's physiological, speech signals and self-reports are collected.

But recording emotional speech in fearful conditions that is realistic and spontaneous is very difficult, if not impossible. To get as close as possible to these conditions and hopefully record fearful speech, the UC3M4Safety team created the WE-LIVE database. With it we collected physiological, auditory and contextual signals from women in real-life conditions, as well as the labelling of their emotional reactions to everyday events in their lives, using the current Bindi system (wristband, pendant, mobile application and server).

In order to detect GBV risk situations through speech, we first need to detect the voice of the specific user we are interested in, a speaker recognition task, among all the information contained in the audio signal. Thus, we aim to track the user's voice separating it from the rest of the speakers in the acoustic scenario, trying to avoid the influence of emotions or ambient noise on the identification of the speaker as these factors could be detrimental for it.

We study speaker recognition systems under two variability conditions, 1) speaker identification under stress conditions, to see how much these stress conditions affect speaker recognition systems and, 2) speaker recognition under real-life noisy conditions, isolating the speaker's identity, among all additional information contained in the audio signal.

We also dive into the development of the Bindi system for the recognition of *fear*-related emotions. We describe the architectures in Bindi versions 1.0 and 2.0, the evolution from one another, together with their implementation. We explain the approach followed for the design of a cascade multimodal system for Bindi 1.0, and also the design of a complete Internet of Things system with edge, fog and cloud computing components, for Bindi 2.0; specifically detailing how we designed the intelligence architectures in the Bindi devices for *fear* detection in the user.

We then perform monomodal inference first by targeting the detection of realistic stress through speech. Later, as core experimentation, we work with WEMAC for the task of *fear* detection using data fusion strategies. The experimental results show an average accuracy of *fear* recognition of 63.61% with the Leave-hAlf-Subject-Out (LASO) method, which is a speaker-adapted subject-dependent training classification strategy. To the best of the UC3M4Safety team's knowledge, this is the first time that a multimodal fusion of physiological and speech data for *fear* recognition has been given in this GBV context. Besides, this is the first time a LASO model considering *fear* recognition, multisensorial signal fusion, and virtual reality stimuli has been presented. We even explored how the gender-based violence victim condition could be detected only by speech paralinguistic cues.

Overall, this thesis explores the use of audio technology and artificial intelligence to prevent and combat gender-based violence. We hope that we have lit the way for it in the speech community and beyond and that our experimentation, findings and conclusions can help in future research. The ultimate goal of this work is to ignite the community's interest in developing solutions to the very challenging problem of GBV.

Extended Summary

MOTIVATION

According to the World Health Organization (WHO), 1 out of every 3 women suffer from physical or sexual violence in some point of their lives, reflecting the effect of Gender-based Violence (GBV) in the world. In particular in Spain, more than 1,100 women have been assassinated from 2003 to 2022, victims of gender-based violence.

GBV, in all its forms, leads to psychological trauma, having behavioural and physical consequences; survivors may struggle with depression and are at a higher risk of suicide. Therefore, there is an urgent need for solutions to this prevailing and widespread problem in our society in the short and medium terms, the latter being the purpose of the projects in which this thesis is framed.

Solutions to GBV may involve the appropriate investment in technological research, among legislative, educational and economical efforts. But despite the technological efforts, several GBV experts question the existing solutions to date and regard them as outdated, as they present different research gaps. These experts demand more advanced research in technology for GBV solutions. And in spite of the impressive advances of Artificial Intelligence (AI), there are no technological solutions for the automatic detection of life-threatening situations for women that incorporate intelligence.

An AI-driven solution that made a comprehensive analysis of aspects such as the person's emotional state, plus a context or external situation analysis (e.g., circumstances, location) and therefore, detect automatically if a woman's integrity is in danger, could provide an automatic and fast response to ensure her safety.

The Bindisystem described in this thesis is conceived as an AI-driven, inconspicuous and and wearable solution that targets the automatic detection of GBV situations. It consists of two smart devices concealed inside jewelry that monitor physiological variables and the acoustic scene, including the voice. These are connected to a smart phone with an application with an AI-driven core that can produce different kinds of alerts, and also encrypt and send the information to a securitized server. Bindi is a cutting-edge technology that combines intelligent Affective Computing (AC) and IoT with physical and physiological multisensorial signal acquisition and fusion and a secure server infrastructure to autonomously detect risky situations, flagging alarms, and recording data for further legal actions.

Thus, this PhD thesis is based on the detection of gender-based violence risk situations for women, addressing the problem from a multidisciplinary point of view by bringing together Artificial Intelligence (AI) technologies and gender perspective. More specifically, we direct the focus to the auditory modality, which we capture with Bindi, analysing speech data produced by the user as voice can be recorded unobtrusively and can be used as a personal identifier and indicator of affective state reflected in it.

BASIS OF EMOTIONS

In this thesis, we also give an overview of the basis of affect, mood and emotion and from where they emerge, the different theories of emotion in cognitive sciences, their current applications and some important ethical considerations. We explain

its relationship with the field of Affective Computing, which is the one in charge of studying and developing technological devices and systems that can recognize, process, simulate and interpret human emotions and affects.

The immediate response in a human being being threatened is the fight-flight-freeze response. This response triggers automatically a physiological reaction that occurs when an event is recognized as frightening. It is an active defense response where the person either fights, flees or stays. The body is affected by physiological changes in order to prepare the person to act appropriately and rapidly.

In response to perceived danger, the Autonomic Nervous System starts a chain reaction that implies a whole series of changes to the heart rate, breathing and muscle activation, including the complex speech production apparatus and therefore characteristics of vocalisations, to meet the challenge of the moment. Muscle tension can lead to having a constricted throat and vocal chords and resulting in a person's voice becoming high pitched, low voice or even absence of voice entirely. Muscle constriction can also cause increased speech speed, jaw and tonge tension, hindering intelligibility, and the shutting down of salivation makes the mouth feel dry and can produce a hoarse voice.

Due to all these physical and physiological changes and their involuntary nature – the person has no control over them – that occur in a person as a result of being in a situation of risk, we considered relying on physiological signals such as pulse, perspiration, respiration, and also speech, in order to detect the emotional state of a person in Bindi with the intention of recognising fear, which could be a consequence of being in a threatening situation. An example of a life-threatening situation that could trigger the fight-flight-freeze responses in women are gender-based violence situations, those in which a women suffers a physical or sexual assault.

Yet there is still no scientific consensus on a single valid theory of the fundamental nature of emotion. This is because emotions are very subjective, and there is no objective way to categorize and quantify them. Thus its inference is a challenge. Emotion labeling depends on, first, the intrinsic difficulty in interpreting innermost feelings of oneself. Second, how much we externalize an emotion. And third, how we interpret certain situations, which can give rise to different emotions.

Another challenge to take into account is the gender personalization challenge. There seems to be clear differences in the expression of emotions according to sex. It has been found that gender-stereotypical expressions, – arising from gendered socialization –, are displayed differently in men and women, as most frequently men express anger and contempt than women, who most frequently express *fear* than men.

DATA CHARACTERIZATION

Ideally, in order to build our Machine Learning (ML) or Deep Learning (DL) AI system for the automatic detection of risk situations from audio data, we would like to count on speech recorded under realistic conditions belonging to the target user.

In our first steps, we found the difficulty of the lack of suitable data available, as speech datasets of real *fear* (not acted) were unavailable or non-existent in the literature. Real, original, spontaneous, in-the-wild and emotional speech are the ideal categories we needed for our application. Therefore, we decided to choose stress as the closest emotion to the target scenario possible for data collection to be able to flesh out the algorithms and acquire the knowledge needed. Thus, we

describe and justify the use of datasets containing such emotion as the starting point of our investigation.

Additionally, and as a consequence of the previous problem, we describe one of the main contributions by the UC3M4Safety team that is the creation of our own set of datasets to fill such literature niche. This has the intention of triggering and collecting human variables to emotional stimuli that could serve in AI or ML/DL systems to distinguish emotions automatically and in real time, specially the emotions of *fear* or panic in women.

First, our team captured the "UC3M4Safety Audiovisual Stimuli Database". It is a high-quality dataset of audiovisual stimuli to trigger up to 12 different emotions in women – including *fear* – under a controlled scenario. It contains a dataset of 42 audiovisual stimuli validated with a discrete and continuous emotional categorization by more than 50 raters each in a crowd-sourcing setting with high level of agreement.

Second, we contributed to the community with the collection of "WEMAC", a multi-modal dataset which comprises a laboratory-based experiment conducted with women volunteers that visualize the UC3M4Safety Audiovisual Stimuli Database. It aims to induce real emotions by using a virtual reality headset while the user's physiological, speech signals and self-reports are collected. Virtual reality is used to maximize the immersive experience and, consequently, achieve a better emotion elicitation.

The database consists of 101 women volunteers who never suffered from GBV and 43 gender-based violence victims (GBVV) women volunteers. The latter group performed the experiment under the supervision of a psychologist. The 28 audio-visual stimuli are selected from UC3M4Safety Audiovisual Stimuli Database to be presented, some of them are stereoscopic 360° videos. Right after every emotional video clip visualization, the volunteers are asked to answer out loud two questions about the video stimuli, to make the volunteers relive the emotions felt during the video visualization, aiming to capture at least traces of emotion in their voice.

In addition to the voice, volunteers label their emotional reactions after the visualization with a joystick. They use the "Modified Self-Assessment Manikins (SAM)" to annotate continuous labels of emotion (Valence/Pleasure, Arousal, and Dominance) and one discrete emotion label out of a total of 12 emotional categories.

We published the first release of the WEMAC database with the aim of sharing it with the research community, encouraging the improvement of the baseline results and advancing the research of multi-modal emotion analysis in general and, in gender equality, in particular. However, we cannot publish or realease the raw speech signals due to ethical and privacy issues, so we have processed the speech signals and extracted low-level and high-level descriptors widely used in literature so that the research community can analyze and work with them.

WEMAC is still far from real-life conditions as recording emotional speech in fearful conditions that is realistic and spontaneous is very difficult, if not impossible. To get as close as possible to these conditions and perhaps record fearful speech, the UC3M4Safety team created the "WE-LIVEd̈atabase. With it we collected physiological, auditory and contextual signals from women in a relevant and uncontrolled environments (real-world conditions), as well as the labelling of their emotional reactions to everyday events in their lives, using the current Bindi system (wristband, pendant, mobile application and server). The database is composed of 13 women volunteers, some of them being GBVV.

SPEAKER RECOGNITION

For our goal of detecting GBV risk situations through speech, we first need to detect the voice of the specific user we are interested in, a speaker identification task, among all the information contained in the audio signal. Thus, we aim to track the user's voice separated from the rest of the speakers in the acoustic scenario, trying to avoid the influence of emotions or ambient noise on the identification of the speaker as these factors could be detrimental for it. But the performance of ML models for detecting speakers through the voice drops a lot when they are under emotional conditions. So the fact that the voice of a GBVV could be influenced by her emotional state constitutes a challenge for a speaker identification system.

We study speaker identification systems under two variability conditions, 1) under stress conditions, to see how much these stress conditions affect the SR systems — in the absence of databases of speech in conditions of realistic *fear* at that time in literature –-, and 2) under real-life noisy conditions, isolating the speaker's identity, among all additional information contained in the audio signal.

With our studies, we verified that stressed speech in the testing stage affects negatively when SI systems use an MLP model and are trained only with neutral speech. As for the case of match and mismatch conditions, in the mixed setting – using neutral and stressed original utterances for both training and testing – the SI system achieves a very satisfactory rate for this type of tasks, demonstrating that the set of features chosen for the task is adequate.

Regarding our experiments in which we augment the data by means of generating artificial stress, we can conclude that the generation of different synthetically generated stressed utterances of speech by modifications in pitch and speed, and their addition to the used database, enlarges meaningfully the instances to work with, improving substantially the results achieved by the Speaker Identification system with a 99.45% of accuracy.

In the line of the speaker identification field under real-life conditions, we studied how speech recorded in real conditions including environmental noise is detrimental for SR systems, and so we explored how to eliminate it with effective denoising methods in order to achieve the best SR performances.

We use robust speaker discriminator oriented embeddings extracted from a Recurrent Denoising Autoencoder combined with a Shallow Neural Network acting as a back-end classifier for the task of Speaker Identification. The proposed end-to-end architecture used a feedback loop to encode information regarding the speaker into low-dimensional representations extracted by a spectrogram denoising autoencoder. We employed data augmentation techniques by additively corrupting clean speech with real life environmental noise in a database containing real stressed speech.

Our proposed architecture achieves reliable results for the whole range of SNRs contaminated signals, being a more robust approach than the rest of the tested architectures, specially in lower SNRs. In the resulting tables, lower SI rates were observed when performing inference in stressed utterances, showing the difficulties induced by stress. This suggests the need to specifically cater for distortions caused by emotional speech for speaker identification tasks.

EMOTION DETECTION

In this thesis, we also dive into the development of the Bindi system for the recognition of *fear*-related emotions, their detection and classification in a highly

multidisciplinary approach as there are many contributions supported by other members of the UC3M4Safety team.

First we describe the architectures in Bindi versions 1.0 and 2.0, the evolution from one another, together with their implementation. We explain the approach followed for the design of a cascade multimodal system for Bindi 1.0, and also the deployment of a complete Internet of Things system with edge, fog and cloud computing components, for Bindi 2.0; specifically detailing how we designed the intelligence architectures in the Bindi devices for *fear* detection in the user. In an additional study using Biospeech data, we demonstrated that extending our database with stressful acoustic events is even beneficial for the recognition of stress in speech and audio.

We then describe our monomodal experimentation with speech for the detection of *fear*-related emotions, first by targeting the detection of realistic stress. Later, as core experimentation, we work with WEMAC for the task of *fear* detection with data fusion strategies. There is a strong multimodal component, since we work on the part of emotions recognition from speech together with data from physiological signals, in the same way Bindi's two wearable devices would work. We use three multimodal data fusion strategies which are evaluated and validated. The experimental results show an average accuracy of *fear* recognition of 63.61% with the Leave-hAlf-Subject-Out (LASO) method, which is a speaker-adapted subject-dependent training classification strategy.

To the best of the UC3M4Safety team's knowledge, this is the first time that a multimodal fusion of physiological and speech data for *fear* recognition has been given in this GBV context. Besides, this is the first time a LASO model considering *fear* recognition, multisensorial signal fusion, and virtual reality stimuli has been presented.

ADDITIONAL RESEARCH DIRECTIONS FOR AUDIO AND GENDER-BASED VIOLENCE

We end this thesis with some parallel and complementary lines of research which opened up in collaboration with other members of the research group and could be of help in the prevention of gender-based violence. We explain the work carried out in the field of "Acoustic Scene Analysis" and the importance of audio events analysis for the detection of risk situations. We define the term of "*Affective Acoustic Scene Analysis*" and with it the need to unify the work carried out on acoustic scenes and emotions under the same title, in order to lift the research and move it forward. We study robust and interpretable acoustic embeddings that characterize emotions in the UC3M Audiovisual Stimuli Dataset.

Additionally, we perform a brief study of fatigue expression, observing the results by gender, and in future research it would be interesting to characterize it to see the differences between stress, *fear*, and fatigue on physiological variables and their effects on the voice. Furthermore, we explore how the gender-based violence victim condition could be detected only by speech paralinguistic cues. Finally, and following another objective aligned with the social good, we briefly explore the relationship between gender-based violence and climate change.

Much of the work in this thesis – just as that of the rest of the members of the UC3M4Safety team– is carried out with a gender perspective in mind, which is to the best of the team's knowledge the first time this is done in research, so we can consider this investigation to be in a preliminary stage in which we are laying the foundations, and in which we aim to continue doing future work. Overall, this thesis explores the use of technology and artificial intelligence to prevent and combat gender-based violence. We hope that we have lit the way for it in the speech community and beyond and that our experimentation, findings and conclusions can help in future research. The ultimate goal of this work is to ignite the community's interest in developing solutions to the very challenging problem of GBV.

To conclude the thesis we consider some options for future work, such as using WEMAC and WE-LIVE in more complex deep learning architectures for the disentanglement of the speaker's identity and the emotional information with (e.g., adversarial model) into different embeddings – low-dimensional space vectors –, to make the detection of speaker and emotion jointly.

In the general terms of the development of Bindi we also have to consider that many women remain in a state of shock when assaulted or becoming victims of an aggression, instead of producing fearful speech. We must take into account this fact for further developments in the Bindi system, or by analyzing the occurrence of silences in the audio, together with the other variables that have already been explored.

Regarding the analysis of acoustic events and acoustic context within Bindi, also of special interest it is the detection of vocal bursts such as grunts, growls, heavy-breathing, squeals or shrieks, and also acoustic events such as hits, bumps or impacts, which would likely denote that a dangerous situation is happening.

The analysis of emotions, particularly *fear*, and the condition of gender-based violence discussed in this thesis could also help health-oriented audio AI research, in particular with applications in mental health care and psychotherapy.

Contents

List of Figures

List of Tables

List of Abbreviations

AC	**A**ffective **C**omputing
AED/C	**A**coustic **E**vent **D**etection / **C**lassification
AI	**A**rtificial **I**ntelligence
ANS	**A**utonomic **N**ervous **S**ystem
ASD	**A**cute **S**tress **D**isorder
ASR	**A**utomatic **S**peech **R**ecognition
BT	**B**lue**t**ooth
BVP	**B**lood **V**olume **P**ulse
CNN	**C**onvolutional **N**eural **N**etwork
CP	**C**omputational **P**aralinguistics
CVAWG	**C**yber **V**iolence **A**gainst **W**omen and **G**irls
DAE	**D**enoising **A**uto-**e**ncoder
DNN	**D**eep **N**eural **N**etwork
EEG	**E**lectro**e**ncephalo**g**ram
EIGE	**E**uropean **I**nstitute for **G**ender **S**tudies
EMG	**E**lectro**m**io**g**raphy
EU	**E**uropean **U**nion
FFT	**F**ast **F**ourier **T**ransform
GBV	**G**ender-**b**ased **V**iolence
GBVV	**G**ender-**b**ased **V**iolence **V**ictim(s)
GRU	**G**ated **R**ecurrent **U**nit
GSR	**G**alvanic **Sk**in **R**esponse
HR	**H**eart **R**ate
IUCN	**I**nternational **U**nion for the **C**onservation of **N**ature
IoB	**I**nternet **o**f **B**odies
IoT	**I**nternet **o**f **T**hings
IPV	**I**ntimate **P**artner **V**iolence
LEAs	**L**aw **E**nforcement **A**gencie**s**
LSTM	**L**ong **S**hort-**t**erm **M**emory
MAE	**M**ean **A**bsolute **E**rror
MFCC	**M**el **F**requency **C**epstral **C**oefficients
ML	**M**achine **L**earning
MSE	**M**ean **S**quared **E**rror
NLP	**N**atural **L**anguage **P**rocessing
PAD	**P**leasure **A**rousal **D**ominance
PNS	**P**arasympathetic **N**ervous **S**ystem
PTSD	**P**ost-**t**raumatic **S**tress **D**isorder
RDAE	**R**ecurrent **D**enoising **A**uto-**e**ncoder
RMS	**R**oot **M**ean **S**quare
SD	**S**peaker **D**ependent
SDG	**S**ustainable **D**evelopment **G**oals
SER	**S**peech **E**motion **R**ecognition

SI	**S**peaker **I**dentification
SI	**S**peaker **I**ndependent
SKT	**Sk**in **T**emperature
SNN	**S**hallow **N**eural **N**etwork
SNR	**S**ignal to **N**oise **R**atio
SNS	**S**ympathetic **N**ervous **S**ystem
SR	**S**peaker **R**ecognition
SV	**S**peaker **V**erification
TTS	**T**ext-**t**o-**S**peech
UN	**U**nited **N**ations
UTC	**U**niversal **T**ime **C**oordinated
WHO	**W**orld **H**ealth **O**rganization

Chapter 1

Introduction to Gender-based Violence

This chapter introduces the motivation and context of this Doctoral Thesis. The first part gives answers to questions regarding the problem we want to solve, and the justification of the relevance of this work. The second part encompasses aspects such as the challenges we face both from the technical and societal points of view.

1.1 Motivation

According to the World Health Organization (WHO), 1 out of every 3 women suffer from sexual or physical violence in some point of their lives, reflecting the effect of Gender-based Violence (GBV) in the world [21]. In particular in Spain, more than 1,100 women have been assassinated from 2003 to 2022, victims of gender-based violence [22]. Gender discrimination and its violence manifestation, are a pervasive problem in our society that affects 50% of the worldwide population.

According to the European Institute for Gender Equality (EIGE) the term of *gender-based violence* is defined as "the violence directed towards a person by reason of their gender". They also state that "both women and men experience gender-based violence but the vast majority of victims are women and girls, and most of the offenders are men" [23]. Thus, throughout this thesis, we will indistinctly use the terms *gender-based violence* and *violence against women* as we believe that using the term 'gender-based' puts the focus on the existing power inequalities between men and women, which is the origin of gender-based violence.

Gender-based violence is demonstrated under many distinct not mutually exclusive manifestations, various incidences of violence can happen at the same time and reinforce each other. Acts of violence can be driven towards people who experience inequalities, such as related to their age, race, disability, religion, social class or sexuality. Thus, the violence and discrimination that women face is not only based on gender, but they also experience diverse and interlinked forms of violence [24].

Violence against women can fall under four key forms of violence, this promotes an exhaustive comprehension of what is considered as gender-based violence. These forms are: physical, sexual, psychological and economic [25].

- Physical violence: Any unlawful physical force that results in any kind of physical harm. Physical violence can be manifested as serious and minor assault, limitation of liberty and ultimately homicide, among others.

- Sexual violence: When any sexual act is performed on a person against their will, either when consent cannot be given or when they do not give explicit

consent – either because the person has a mental disability, is a child, or is unconscious or intoxicated as a result of drugs or alcohol – [26]. It can take the shape of sexual assault or rape.

- Psychological or Emotional violence: Any action or behaviour that causes psychological harm to a person, causing *fear* by intimidation or sabotaging a person's sense of self-worth through continuous criticism. Psychological violence can be manifested as, for instance, intimidation, defamation, harassment, humiliation or verbal insult.

- Economic violence: All actions or behaviours that cause economic harm to a person, making the person financially dependent, keeping partial or total control over their financial resources. Economic violence can be manifested as restricting access to education, to financial resources or to the labour market; property damage, or not meeting with economic responsibilities – such as the maintenance allowance – [27] among others.

In this era of the digital space, new types of discrimination and violence against women have emerged recently, currently defined as cyber violence against women and girls (CBAWG). This type of cyber-violence includes actions such as non-consensual pornography (also called 'revenge porn'), cyber stalking, 'slut-shaming', 'doxing', unrequested pornography, rape and death threats, gender-based slurs and harassment, 'sextortion', gender-based slurs and harassment, and electronically enabled trafficking [28]. Cyber VAWG is a continuation of the violence that ocurrs offline. For instance, cyber stalking by an ex-partner or partner has similar consequences to offline stalking and is considered the same type of intimate partner violence. The only simple difference it is that it is facilitated by technology. Thus CVAWG can be manifested as multiple forms of violence, including psychological and sexual violence. The rising trendencies also point out that economic violence is on the rise, which happens for instance when the victim's employment status (or future employment) is endangered by information that is released online. The importance for violence in the cyber-space to be manifested also psychically should be taken into account too [29]. In the last couple of years, the COVID-19 pandemic has aggravated the risks of cyber-violence against girls and women. According to [30], "Internet use has increased between 50%-70% from the levels that it was used before the pandemic, and this increased vulnerability has led to a ghost pandemic of online gender-based violence".

Structural inequalities is one of the causes gender-based violence is normalised and reproduced. These are the societal norms, attitudes and stereotypes around gender in society. Therefore, it is important to recognize structural and institutional violence when trying to explain the pervasiveness of gender-based violence in our society. This is defined as "the subordination of women in economic, social and political life" [24]. Much violence against women is made invisible, which is barely reported due to the shame and stigmatization suffered by the victims, and the impunity enjoyed by the perpetrators. GBV is not an individual problem but a social phenomenon intersecting many different areas of life, in which invisible violence is the basis for sustaining the most life-threatening forms of violence. Gender inequality is what lies below the surface, thus it is essential that society acknowledges and recognizes types of visible and invisible violence, represented in the iceberg of violence in Fig. 1.1, in order to disarm the social and cultural framework which perpetrates such violence.

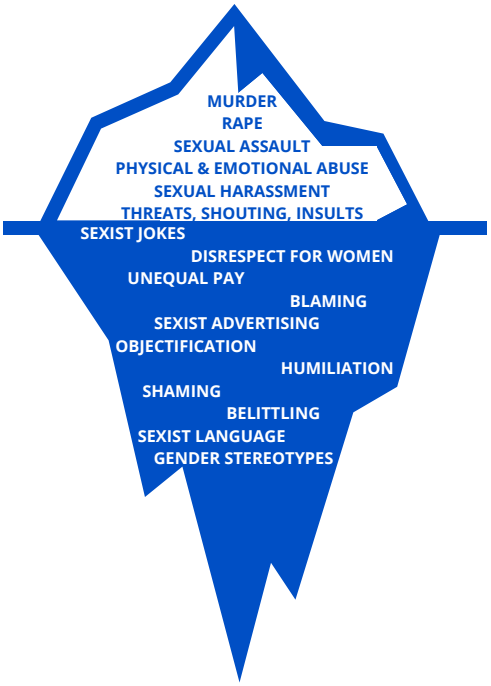


FIGURE 1.1: Iceberg metaphor of visible and invisible forms of Gender-based Violence. Illustration based on [31].

GBV, in all its forms, leads to psychological trauma, having behavioural and physical consequences; survivors may struggle with depression and are at higher risk for suicide. Therefore, there is an urgent need for solutions to this prevailing and widespread problem in our society in the short and medium terms, the latter being the purpose of the projects in which this thesis is framed.

1.1.1 Economic Framework for Gender-Based Violence in Europe

According to the European Institute for Gender Equality (EIGE), the European Union (EU) spends 366 billion euros a year on the consequences of gender-based violence [32]. This updated study based on the 2014 titled report *Estimating the costs of gender-based violence in the European Union* [33], provides "revised estimates of the costs of gender-based and intimate partner violence in the EU".

By extrapolating the results from the UK case study to the European Union, "the estimated cost of gender-based violence against women in the EU-27 was more than €290 billion, representing 79% of all costs of GBV against both women and men". Besides, "the estimated cost of intimate partner violence against women in the EU-27 was nearly €152 billion, representing 87% of all costs of intimate partner violence against both women and men" [33].

The different costs of GBV in the report are broken down as follows, with the greatest cost being the emotional and physical impact on victims (56%) understood as the harm they suffered as a consequence of the crime, followed by criminal justice services (21%) (justice system and police) and loss of economic output (14%) explained as a general increase in incidence data for both women and men. Other costs to consider can include civil justice services (e.g., divorce and child custody procedures) and financial support for housing and child protection services. Specifically, during the Covid-19 pandemic and as one of the consequences of lockdown restrictions, "intimate partner violence spiked, accounting for almost half

(48%, €174 billion) of the cost of gender-based violence". From these, "intimate partner violence against women makes up 87% of this sum (€151 billion)".

Even when the amounts taken into account are in the order of billions of euros, the money that goes to support gender-based violence victims (GBVV) is not enough, as services such as shelters for women in situations of violence account for only 0.4% of the cost of GBV.

The United Nations (UN) proclaimed in 1993 the "Declaration on the Elimination of Violence against Women" [34], and since then, "violence against women and domestic violence are considered forms of discrimination, matters of criminal law and violations of human rights". The 2030 Agenda for Sustainable Development [35] was adopted by the UN Member States in 2015 and it provides a "shared plan for prosperity and peace for the planet and the people, in the present and in the future" [1]. On it they proclaim the 17 Sustainable Development Goals (SDGs), among which they recognize "the goal to achieve gender equality and empower all women and girls" (SDG 5). The SDGs present an urgent call for action in a global cooperation by all countries.

Moreover, combating GBV is part of the European Commission's activities to protect the main EU values and guarantee that the EU Charter on Fundamental Rights is maintained. As for March 2022, the European Commission proposed new rules applying to the whole EU to combat gender-based violence and domestic violence[36], including the criminalisation of female genital mutilation, rape based on lack of consent, and cyber violence, as well as strengthening the access of victims to justice, among others.

While it is impossible for human pain and suffering, even a human life, to have a "price tag" [33], being aware of the cost of violence can guide countries conduct the money to where it is actually necessary; most life-saving-efficient and cost-effective, which are both a moral imperative and an intelligent use of economy.

1.1.2 Eradication of Gender-Based Violence

The underlying cause of gender-based violence is related to structural gender inequality, based on the patriarchal system of society and the imbalance of power between men and women. The concept of intersectionality recognizes that systemic inequalities are shaped by the overlapping of different social factors, such as sexual orientation, gender identity, ethnicity, race, disability, and economic class, among other factors of discrimination. All these intersect to create particular dynamics and effects of crossed discrimination. All forms of inequality are jointly reinforcing and should be analysed and addressed at the same time to prevent inequalities from reinforcing one another [37]. Therefore in particular, there are certain groups of women who are more vulnerable to violence, because they already suffer from gender discrimination, these are the ones that have more difficulties to resist when confronted with this threatening phenomenon. This includes young and elder women, children, disabled, racial, ethnic, migrant or indigenous female persons, perceived to be LGBTIQ+ (defined as "people who have identified themselves as lesbian, gay, bisexual, transgender, intersex, or questioning"), as well as substance abusers and those with family or economic difficulties, these are at a higher risk of suffering from GBV.

This leads to the conclusion that the profile of a GBVV is not homogenous, and to ensure effective help and support for victims, diverse intervention strategies,

[1] https://sdgs.un.org/2030agenda

prevention, education and therapy programs are needed. GBVV usually have individual/special protection needs, in which some are especially vulnerable to repeat victimization, mostly based on psycho-social-cultural aspects that must be taken into account when when targeting solutions to this prevailing problem.

Solutions to GBV may involve the correct allocation of economic resources for legal and social means (e.g.,for the protection of victims and their children), for education and, in addition, for investment in technological research.

At the general population level, it is necessary to extend prevention from an integral perspective as awareness-raising measures, based on respect for human rights, teaching the rejection of all types of violence and including specific actions against gender-based violence.

In the educational field it is fundamental to go beyond the development of one-off materials and programs so that prevalent educational measures are used. Through collaborative experiences between girls and boys in the classroom, based on mutual respect, great progress could be made in overcoming two of the main conditions that underlie gender violence: the resistance to change that this situation produces and the unequal distribution of power in society [38]. Additionally, institutions such as schools and high schools, ought to develop protocols on how to act in the event that it becomes aware of violence among students or their families, by means of educational intervention. Because it is not enough to provide information, but it is necessary to build equality from practice from the awareness and education of the new generations.

In the legislative field, there are state regulations in Spain [39], Europe [40] and Internationally [41] that take care of regulating aspects such as the judiciary, or the labor field regarding GBV. In Spain, the Law 1/2004 for Integral Protection Measures against Gender Violence [42] would establish "the essential judicial mechanisms to avoid a double victimisation of suffering women, involving the unification of a framework of assistance and protection for all women, whatever is their personal situation". This law also establishes "all-inclusive protection measures, whose purpose is to prevent, eradicate and punish gender-based violence and to provide assistance to women and minors under their custody, also victims of this violence".

In the European Union, all EU Member States have signed up to the "main human rights mechanisms", which oblige them to "combat violence against women as it is considered a violation of human rights, and a specific form of gender-based violence linked to discrimination against women". In such way, Member States are obliged to terminate impunity and prohibit all kinds of violence, to provide adequate protection to survivors, to take measures to prevent it, and to ensure help [40]. Examples of the interest in the prevention of violence and gender equality is the creation in 2006 of the EIGE[2] – already mentioned in Sec. 1.1.1 – which is in charge of the collection, analysis and dissemination of information on equality and GBV; as well as the establishment of the Istanbul Convention [43] by the Council of Europe in 2011 on the fight and prevention of domestic violence and gender-based violence.

From the technological and research fields, the fight against violence in the European Union has made the EU fund research and innovation projects to fight Gender-based Violence for more than two decades[3]. And their research findings have been translated into recommendations tailored to the different sectors involved in the protection of victims, namely, police, health and social sectors, needing for multi-agency cooperation.

[2] https://eige.europa.eu/

[3] https://eige.europa.eu/topics/research?ts=technology

1.1.3 Technological solutions to combat Gender-Based Violence

Technology has achieved many medical and scientific breakthroughs, but it has also given rise to new types of online discrimination, hate speech, and virtual human rights violations, including online gender-based violence or cyberviolence.

Technology can be a tool for empowerment and security for women, to make them active participants escaping violent relationships. Various target groups can benefit from the use of technology: not only recovering victims but also social agents involved in prevention of gender-based violence and protection against it such as therapists, secutity forces and health services. Technical solutions improve the efficiency of professional intervening GBV. That results in a better quality of services offered to citizens and greater security.

All women are at risk of suffering GBV, that's why effective strategies to prevent and reduce violence must aim at its roots. "The development of safe technology to address gender-based violence requires leadership by and collaboration with women and girls", UNICEF states [44]. As the causes of GBV significantly differ from other types of violence, we need to systemize the existing knowledge and then use it as a standard to adapt technological tools in a way that responds to it. The rights, needs, and requirements of survivors of this particular violence, are key for its design. Because UNICEF also states that "technology and its tools should not expose girls and women to any more harm; solutions have to be built with an extensive and strong foundation of ethical protocols an standards of the GBV community, while ensuring digital safety and privacy standards – e.g., anonymity and data protection –, in order to prevent discrimination and victimization".

In recent years, digital technology growth has benefited the development of novel web and smartphone applications aimed to fight against GBV. Together with the advent of the Internet of Things (IoT), these technologies have triggered the development of several solutions that range from Law Enforcement Agencies (LEAs) to mapping sexual violence exposure within a location [45].

Applications based on geolocation features can increment awareness and reduce a user's risk of violence, supporting prevention [44]. For instance, Ec Shlirë (Walk Freely)[4], is an app developed by Girls Coding Kosova, enables users to report instances of sexual harassment in a discreet manner, that are shared with authorities. A similar app for smartphones, Safetipin[5], crowdsources and maps real-time user's data — mainly women and girls -– to provide location safety scores in order to improve public safety.

Another innovative tech use to facilitate access to information and services without the need of on-site attendance, in a way that is safe, culturally suitable and with high user accessibility, is the use of interactive chatbots or dissemination apps. Some examples are Project Caretas[6] in Brazil, Maru[7] by Plan International NGO, Virtual safe spaces (VSS)[8] and Springster[9] by UNICEF. These apps provide resources and real advice from activists and experts, supplying with information about self-care and empowerment, gender-based violence and reproductive and sexual health for women and girls. However, in the case of chatbots, some are AI-driven and only a few include human interaction in the background. Recent

[4] http://iwalkfreely.com/

[5] https://safetipin.com/

[6] https://www.unicef.org/brazil/projeto-caretas

[7] https://plan-international.org/news/2020/11/25/new-chatbot-to-tackle-online-harassment-faced-by-girls/

[8] https://www.unicef.org/media/111806/file/UNICEF-Virtual-Safe-Spaces-21.pdf

[9] https://global.girleffect.org/products-showcase/big-sis-chatbot-springster/

studies are concerned about the potential and efficacy of such chatbots to provide effective online emotional support to humans, and conclude that "users seemingly consider human-generated support more reliable than machine automated support" [46]. That is why the inclusion of *human-on-the-loop* [47] it is crucial in such apps.

In the case of apps for GBVV, technology also offers improved delivery of gender-based violence services and reaction quality. Primero/GBVIMS+[10] is a technological solution open source-abed for the management of GBV cases. The system improves quality of care for survivors and remote collaboration between supervisors and workers working on such cases. Another app, ROSA[11] provides essential education and the exchange of knowledge for staff to support people who suffer from GBV. Medicapt[12] collects forensic evidence – which is court-admissible – from survivors of sexual violence, and it can securely transmit these data to police, jugdes and lawyers. And, VictimsVoice[13], which is an app that enables GBV survivors to annotate incidences of abuse in a legally admissible, secure and safe manner [44].

Still, these solutions are lacking important features when it comes to GBV live protection, ensuring women's safety from physical GBV, that is, attacks and aggressions. Thus, other solutions aim to target such situations, such as SAFER PRO[14], which is a wearable device developed by a company in New Delhi, India, that contains a chip built into a wristband that sends alerts to emergency contacts, when the device is activated by the user, informing of an emergency situation. Additionally, India issued a directive related to the mandatory inclusion of a panic button on every mobile phone sold as of 2017. However, panic buttons present significant limitations regarding women's safety, such as the requirement of an active role in their self-protection –certainly not possible under some types of aggression–, their lack of inconspicuous design –that can lead to stigmas in GBVV–, or even worse, the lack of infrastructure support [48].

Particularizing the GBV technological solutions developed to date in national territory, Spain is pioneer in GBV technology. Some institutional solutions include technological tools to support and protect GBV such as the following:

- VioGén[15], a protocol followed by police officers who take a statement from the complainant of a gender-based violence victim. They fill in a specific questionnaire which results in a risk rating which, if high, activates police protection measures. These can range from making follow-up calls to placing a patrol car 24 hours a day at the door of the victim's home.

- ATENPRO[16], a mobile telephone and telecommunications device that allows users to contact a Call Centre staffed by personnel specifically trained to provide an appropriate response to their GVB situation at any time.

- AlertCops [49], it is a service to help citizens in dangerous situations, with the aim of sending warnings, including geolocalised data, with photographs or

[10] https://www.gbvims.com/primero

[11] https://www.rescue-uk.org/perspective/why-we-need-go-mobile-protect-women-violence

[12] https://phr.org/issues/sexual-violence/medicapt-innovation-2/

[13] https://victimsvoice.app/

[14] https://theindexproject.org/award/nominees/3198

[15] http://www.interior.gob.es/web/servicios-al-ciudadano/violencia-contra-la-mujer/sistema-viogen

[16] https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/servicioTecnico/home.htm

audio, to police officers to warn them of a witness or the presence of a crime. AlertCops incorporates an SOS button to reinforce the protection of victims of gender violence and healthcare personnel[17].

- More specific to the protection of GBVV, the COMETA[18] Centre offers the monitoring, operation and installation services of devices that monitor both the GBVV and the aggressor. The system is designed to trigger an alarm in cases the GBVV is at risk, such as if the offender gets too close to her, or in case of the manipulation of the strap or breakage of the bracelet, among others.

A research study performs a thorough analysis of the technological solutions [50] and states that "the objective must be to achieve an holistic solution, as the proper integration of diverse approaches could lead to a multi-strategy proposal that could improve women's safety and contribute to the end of this kind of violence".

Despite the technological efforts, the solutions existing to date present different research gaps questioned by several GBV experts who demand more advanced research [51] and technology for some solutions regarded as outdated. And in spite of the impressive advances of Artificial Ingelligence (AI), there are no solutions that incorporate intelligence for the automatic detection of a risk situation that could endanger women's lives. We have already mentioned that panic buttons, or telematic help centres, are solutions that involve the engagement of the victim in her own safety. And in cases where women are attacked by an offender or aggressor, they may not have the resources to carry out these actions.

Novel paradigms are emerging in which new AI tools are proposed, but not as replacing the existing ones, but rather as complementary, including advantages in comparison to traditional ones. AI-powered predictive analysis is able to collect user data, analyze it, and draw valuable insights from it. Predictions regarding women's safety can be accurately estimated with a correctly trained AI algorithm, by analysing the data (i.e.: users state, context) with which it has been fed. AI provides features such as analytical study of user state and context, custom-made predictions and decisions, a dynamic cycle of personalized and actionable insights, real-time accurate data-driven output, and elimination of irrelevant information, providing a comprehensive solution that would ensure women's safety.

An AI-driven solution that made a comprehensive analysis of aspects such as the person's emotional state, plus a context or external situation analysis (e.g., circumstances, location) and therefore, detect automatically whether a woman's integrity is in danger, would avoid the person's requirement for active involvement their self-protection –certainly not possible under some types of aggression– and could provide an automatic, faster response to ensure her safety.

1.1.4 A cutting-edge AI technology solution to combat GBV: Bindi

In response to the requirements discussed above and after a socio-psychological study of the advantages and disadvantages of the technology currently employed in Spain, the multidisciplinary UC3M4Safety team was born in 2016 – to which both the author and the supervisor of this thesis belong – to develop an innovative AI solution called Bindi.

This thesis project was framed in EMPATIA-CM (*ProtEcción integral de las víctimas de violencia de género Mediante comPutación AfecTIva multimodAl*), a project

[17] https://alertcops.ses.mir.es/mialertcops/en/index.html

[18] https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/dispositivosControlTelematico/home.htm

of the *Convocatoria 2018 de Proyectos Sinérgicos de I+D en Nuevas y Emergentes Áreas Científicas Comunidad de Madrid* that was awarded to the multidisciplinary UC3M4Safety team from April 2019 until June 2022. Currently, the group has been awarded a second continuation project until June 2023 namely S4B (*SistemA ciberfísico Para el seguimIENTo y prevencIón de cAsos de violencia de género*: SAPIENTAE4Bindi)) to increase the technological maturity level (TRL) of the research results of EMPATIA-CM and this thesis is framed in both projects, EMPATIA-CM and S4B.

The UC3M4Safety team set up from the need to join efforts in the fight against Gender Violence (GBV) from multiple disciplines. The project has 42 researchers from the Institute of Gender Studies (IEG) and the departments of Electronic Technology, Telematics and Signal Theory and Communications of the University Carlos III of Madrid (UC3M). In addition, the group collaborates closely with the Centro de Electrónica Industrial (CEI), Universidad Politécnica de Madrid (UPM). And this multidisciplinarity of the team refers to the involvement of diverse researchers providing knowledge from several disciplines on it, each one contributing to the project from their own area.

The UC3M4Safety team started its journey in the Anu and Naveen Jain Women's Safety XPrize[19] competition in which it was a semi-finalist. It has registered a utility model application [52] and has obtained several grants and awards.

The main objective that EMPATIA-CM had and now continues in S4B is to improve the protection that society offers to women in situations of GBV aggression, generating a reliable and robust protocol to detect, prevent and solve these crimes. The mission of the innovative technologies proposed in the projects carried out is to help prevent GBV by means of: 1) the early detection of risk situations, 2) the interconnection of potential victims and protective agents, 3) the secure and accurate collection of evidence of the alleged crime, as well as 4) to have sufficient data. All of these means will help us to study the problem of GBV in a comprehensive and multidisciplinary way. To this end, the team proposes the use of cyber-physical systems with Affective Computing (AC). In particular, wearable devices (wearables) with intelligent sensors that monitor in real time, detect the circumstances in which the user is and the emotions experienced in risk situations and connect with protective agents, governmental and / or non-governmental, warning in real time. The expected impact is the notable improvement of women's vulnerability by providing tools that improve women's safety and favor their personal and professional development.

This is the multidisciplinary environment of social commitment and cutting-edge research in which this thesis project is developed. The human and material resources dedicated to the project are optimal thanks to the adequacy of the profiles of the researchers involved, the technological infrastructure available and the funding obtained in the EMPATIA-CM and S4B projects.

Regarding Bindi's technical operation, this system is conceived as an end-to-end, smart, inconspicuous, connected, edge-computing, and wearable solution targeting the automatic detection of GBV situations.

As can be observed in Fig. 1.2, the GBVV wears two smart devices hidden inside jewelry that monitor the physiological variables and the acoustic environment including voice. These are connected to a smart phone with an application with an

[19] https://portal.uc3m.es/portal/page/portal/inst_estudios_genero/proyectos/Women_Safety_XPrize_2018

FIGURE 1.2: Outline of Bindi's operation [53]. Reproduced with permission of the copyright owner, UC3M4Safety team.

AI-driven core that can produce different kinds of alerts, and also encrypt and send the information to a securitized server.

Bindi is a cutting-edge technology that combines intelligent Affective Computing and IoT with physical and physiological multisensorial signal acquisition and fusion and a secure server infrastructure to autonomously detect risky situations, flagging alarms, and recording data for further legal actions. More specifically, Bindicaptures metadata and data regarding the user and her context (e.g., custom routines, geolocation, physiological variables, speech, acoustic events,...) and determines the affective state of the user considering the circumstances in which it finds herself. With such data, Bindi uses its AI core to evaluate each situation, and has the ability to detect automatically when a situation could be life-threatening for the user, triggering its alarms, alerting emergency services, and offering support and help on-the-fly. In Fig. 1.3 we represent the evolution of the wearable devices from Bindi version 1.0 to 2.0.

1.2 Context: Technical Challenges

After having discussed the motivation for this work, we consider it essential to explain the context in which it is framed. In this subsection we will discuss the technical challenges arising from this work.

The thesis focuses on the analysis of the affective state of a person by means of different input or data modalities, but more specifically of the auditory one, and in particular from speech. Thus, we review some of the challenges regarding the bias problem in AI algorithms, pointedly on the gender aspect. Moreover, this technological solution is intended to be embedded in wearable devices and smartphones, so we expose the concern on the computational challenges of such devices. Lastly, our ultimate goal with it is to detect when a woman's life is in danger due to a situation of gender-based violence, therefore, we briefly present the great challenge that is to be able to provide solutions to all women/people.

1.2.1 Research, Data & Biases

Intersecting and overlapping social hierarchies found in power, religion, race, ethnicity, gender, age, sexual orientation, or class, result in the unequal distribution

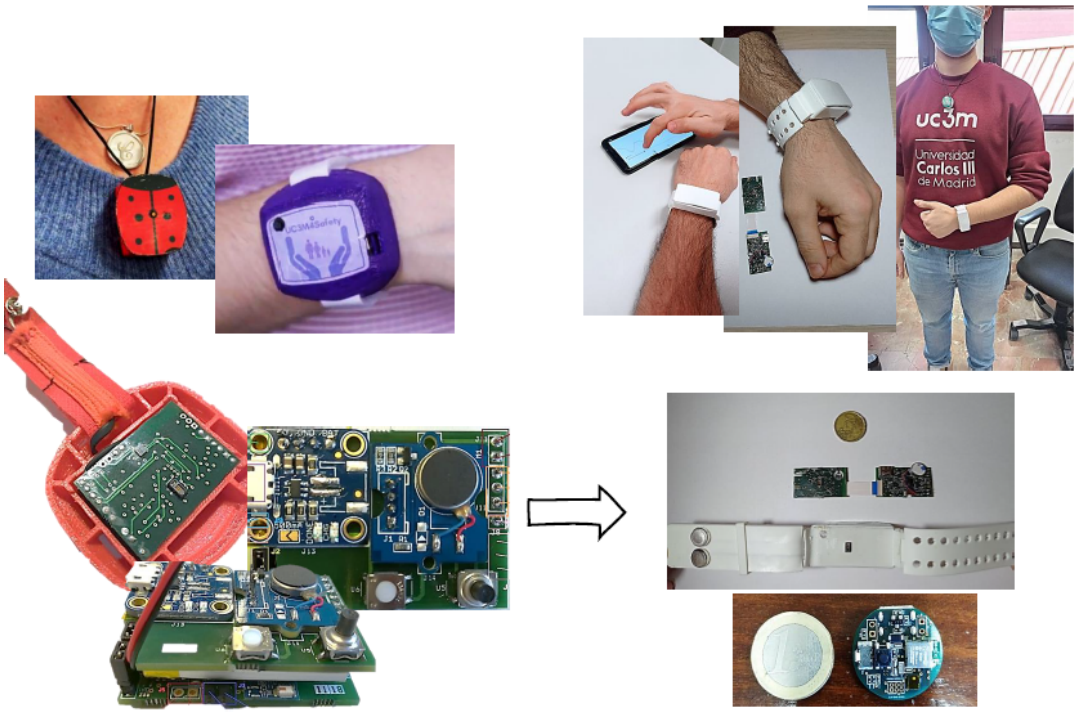


FIGURE 1.3: Evolution of Bindi wearable devices [53]. Reproduced with permission of the copyright owner, UC3M4Safety team.

access to resources and rights, constituting social inequality. And from an historical perspective, research advances –and more specifically, in science and technology– have been oriented, focused and directed towards a specific profile of people due to the prevailing and historical social inequality.

Social groups to which research, science and technology have not been historically oriented towards, have been demonstrated to have difficulties in getting such developments applied and working for them. A clear example of such social groups are women.

In the medical field, there was a near doubling of the rate of death in women than in men due to heart disease [54] until the year 2000. There were clear research gaps in myocardial infarction, carried out taking into account mostly male patients. This resulted in less adequate diagnosis and treatment received by women than men, as a result of the education and information received by the physicians. Up to that date, women were thought to experience a greater variety of symptoms, being called 'atypical syndromes', because they did not correspond to the ones experienced by men and thus myocardial diseases were seldom properly identified in women.

In the field of technology and AI, there are countless examples of such discrimination, some of the most recent are described next. A well known car brand was forced to request for return of one of its car models form the market because male drivers in Germany did not trust the female voice giving directions in the car's navigation system [55]. Another study [56] that evaluated the accuracy of YouTube's automatically-generated captions across the two standard genders showed robust differences in accuracy for both genders, with significantly lower accuracy for women's voices, displaying the need for sociolinguistically-stratified validation of systems. An additional study [57] evaluated 3 commercial facial classification systems showing considerable discrepancies in "the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification

systems". Females with dark skin were the most misclassified group – "error rates of up to 34.7%" [58] –, whereas "the maximum error rate for lighter-skinned males was 0.8%". We can find a further example of the research gap between social groups in [59], where "wearable devices are be less accurate for detecting heart rate in individuals with darker skin tones". Investigators concluded that "wearables are less accurate for detection of heart rate in such participants, being a possible cause the lack of research on darker-skinned subjects at the moment of such wearables being developed".

We would also like to highlight that, in the speech recognition field, it is known for a long time that models oriented towards voice recognition of men on the one hand and of women on the other hand work better if designed separately due to the known and different speech characteristics each gender has [60].

These are just a few of the many examples that evidence that AI is generally biased towards specific social groups, discriminating others, as it is the case of women [61], where AI is historically clearly lacking gender perspective. This points out that specific attention is needed if research is to build absolutely fair, transparent and accountable AI algorithms. The role of AI in achieving United Nations (UN) Sustainable Development Goals (SDG) is controversial: it may enable the accomplishment of 134 targets across the goals, but also inhibit 59 [62]. Along with the huge explosion of advances in AI in recent decades, an ignited activism for social rights has also emerged as algorithmic systems have been criticized for perpetuating bias, unfair discrimination, and contributing to inequality [63].

Despite the breakthroughs uncovering biases in subfields of AI such as Natural Language Processing (NLP) [64], Computer Vision (CV) [65] and Medicine and Healthcare [66], from the Affective Computing point of view – when the task requires recognizing, or simulating human affects –, there is little evidence in literature of the biases of AI algorithms (e.g., using psychobiological features [67]) when the data used for the task are speech signals. Nevertheless, this does not mean that biases do not exist in such cases, but that it is such a recent field that literature has not yet been systematically reviewed on the topic.

Thus, extrapolating analysis from other research fields, we provide a brief review of some of the possible biases in Affective Computing, and more specifically in the task of Speech Emotion Recognition (SER) that result in possible gender-unbalanced solutions, and possible ways to mitigate them:

- Sampling bias in datasets creation: Increased amount of data gathered from male subjects or speakers. There is an active tendency, in emotion speech recognition databases, to provide the same amount of male and female speaker data, but for the vast majority of databases that have an imbalance, it is due to a greater number of male over female speaker data [68].

- Negative set bias (or Closed World Fallacy): Refers to not having enough samples to make a reliable representation of the whole world in the available dataset. Datasets may be unbalanced if we want to represent all the categories or types of data that exist in the world and the dataset does not contain them. This is a very common bias and factually challenging to address.

- Labelling bias: Labels that come along with data are usually generated by humans (e.g., expert annotators, *crowdsourcing*,...) and this bias refers to the fact that various labelers may annotate the same data differently. Each annotator bases on their background for labelling, according to their training, origin, context or trajectory. Therefore the annotation is a consequence of the

previous perceptions and experience of the annotators. For example, in the case of emotion annotation, whose character is inherently subjective, the label may vary according to characteristics of the labeler, such as his or her culture, emotional intelligence, etc.

- Human evaluation bias: Consists of drawing conclusions based on the experience or trajectory of the person or persons conducting the research, who will have perception and analysis biases depending on their context.

The solution to all these biases goes through including a balanced, rich and diverse number of: data, annotators for such data, and evaluators; in order to faithfully capture reality, characterize content and data reliably, and draw appropriate conclusions taking such biases into consideration.

It is important to look out for biases in data and AI algorithms and mitigate them, because AI can exacerbate and reinforce social biases, with all the negative consequences that this entails. Data captures reality and reality is biased, if such biases feed AI algorithms, they will ultimately reinforce and perpetrate discrimination. Researchers should take biases into account and work towards fairness and achieving equality in society with their systems, but all of us should go beyond debiasing, and tackle the discrimination problem on its roots, by taking social responsibility against structural inequality.

1.2.2 Hardware: Computational complexity and battery

The success of AI algorithms relies in many factors, such as the amount or quality of the data used for training, the complexity of the models and the accuracy of the labels, among others. Complex and deep computational models have proven to be successful due to their good generalization capabilities, and so are more suitable when the training phase contains lots of data, compared to using smaller datasets and shallow machine learning algorithms [69]. Thus one of the main disadvantages of AI models these past few years is this tendency to grow towards bigger architectures and schemes [70], in order to achieve better performances. But bigger is not always better for machine learning. However groundbreaking they are, the consequences of bigger models are severe for both budgets –need of more computational power and energy supply– and the environment –the more the energy consumption, the worse for the environment in terms of pollution– [71].

In Bindi the IoT architecture designed considers a three-layer division, i.e., edge, fog, and cloud computing [1]. In the system, the edge-computing layer is conceived as a smart cyber-physical network composed of two devices (a pendant and a bracelet). The fog computing layer in Bindi is conceived as a smartphone application. Finally, the relevant information obtained throughout the edge and fog layers is securely stored in specific computing services in the cloud. And power consumption management is a requirement for the design of such wearable systems. If we want to employ AI algorithms in such devices, an accurate measure of the state of the battery charge and autonomy of the hardware devices involved in each layer is essential to ensure that the system works when needed.

As part of the UC3M4Safety team's work, quantitative consumption analysis on the wearable devices is studied [72], measuring the most energy-demanding actions through the monitoring part [1]. The team measured the power consumption due to sensor data communication and acquisition, as they are essential for the system and are intrinsically related to the specific hardware design of the devices [1].

1.2.3 A solution for all women? The generalization challenge of AI in GBV

In spite of current expectations about the role of Artificial Intelligence in our society and the impressive advances we have witnessed in the last decade, the robustness of AI is of great concern. Indeed, the European Union has developed the Artificial Intelligence Act [73], providing a set of guidelines to protect European citizens from possible misuses and errors with an emphasis on trustworthiness and the avoidance of all kinds of biases regarding the demographic characteristics of the users.

There are different approaches to improve AI robustness in the literature, but most of them coincide in the diagnosis of the roots of the problem: the mismatch between the mathematical models obtained from training with laboratory data (captured under controlled conditions) and used then in the real-world where the conditions are uncontrolled. The more complex the reality, the more data is needed to correctly obtain accurate models. All data collected needs to capture the diversity and complexity of the phenomenon to be modeled.

We are aware, as previously stated, that individual vulnerability to GBV is related to psycho-socio-cultural aspects. Therefore, it is a massive challenge to develop oneAI support tool for all women, and more specifically for GBVV. Such solutions can be non acceptable, nor appropriate, and/or non available, or even dangerous in some circumstances. This is the reason for which we shall consider a broader dimension. Because this technological solution is very important to save lives and assure women's safety. We are aware that we should be able to provide solutions to all women/people, but it is still a great and complex challenge.

One of the drawbacks of this type of technological solutions is their limited generalization abilities. This means that the technologies are not currently capable of automatically adapting to the diversity of GBVV and their changing situations, for example in daily routines, cultural habits, diversity of familiar situations, etc. This could impact the performance severely and the rate of false alarms triggered to raise significantly, which would make the LEA's resources needed to attend them completely unaffordable.

As we have mentioned, the complexity of the problem of GBV is dynamic and hard to measure since the measures taken to combat them, together with the raise of public awareness and educational efforts, that the different countries apply, modify its aspect and prevalence. From the point of view of automatically modelling it by using technological tools and different sensory devices, it is clear that we need to split the problem, first by understanding the different socio-cultural realities, and second, the psychological situation of the victims, by means of expert knowledge and quantitative methodologies from the social sciences.

The observed balance between collective and individual behaviors has to be translated into a methodology to collect and model data, and to articulate the relationships between the mathematical sub-models obtained by AI. But not only the virtuous circle closes when the technologies obtained with the help of the social sciences are used to enhance the socio-psychological understanding of GBV, but when the from of the smart integration and aggregation of data captured make quantitative methods improve substantially.

1.3 Objectives and Relevance

Having given the motivation and context of the challenges to be faced, in this section we break down the objectives of this thesis and their justification and relevance.

We believe that the solutions to the GBV problem from the social sciences could come hand in hand with technology and Artificial Intelligence. We believe that technology is an enabler for eradicating GBV, but not the only solution itself. Thus, this PhD thesis aims at detecting situations of risk of gender-based violence for women, addressing the problem from a multidisciplinary point of view by bringing together AI technologies and a gender perspective, needing from techniques of several disciplines.

We want to analyse the emotional state of woman through the auditory and physiological modalities, – being auditory, speech and acoustic –, but more specifically, we direct the focus to the auditory modality, analysing the speech produced by the user, for the detection of emotions in the voice and the ways it can be combined with physiological information.

Thus, we aim to research on Affective Computing, particularizing on Speech Technologies and its applications, from a gender perspective, to give a technological solution that can protect women from gender-based violence risk situations. Fig. 1.4 presents this definition in a conceptual map of this thesis.

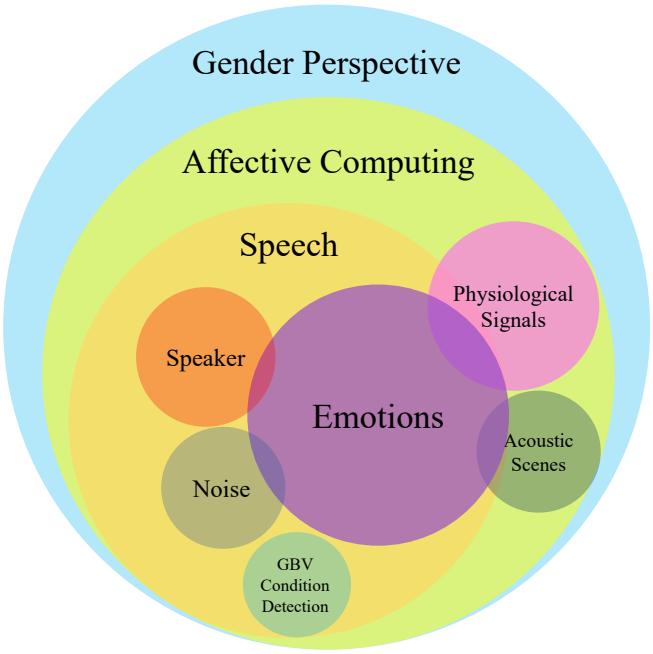


FIGURE 1.4: Conceptual Map framing this PhD Thesis.

There are several key aspects why the voice was selected as a variable to be recorded with Bindi for the protection of the users. First, in order to capture the voice, there is no need to carry heavy or complex devices like a helmet, a headband or a chest band. Voice is one of the signals that can be recorded unobtrusively, just by using a smartphone or a lavalier microphone; thus it is easy to capture and use voice data, and also users can easily access to the technology that can record such.

Second, voice is an unique identifier of a person. Any other body signal like skin temperature (SKT), will not include information regarding the person, without being able to make a direct link between the person and the signal. Whereas, by capturing voice, we do capture a lot of relevant information regarding person speaking [74], voice can be used as a personal identifier, e.g., it can be used for judicial evidence if stored correctly and protected to avoid modifications and hacks. And third, because emotions are reflected in the voice, as we will further see in Chapter 2, especially *fear*

– for which the absence of voice is also important (further discussion regarding the aid of physiological variables is done in Sec. 2.2.1).

As stated previously, this thesis was born framed within the EMPATIA-CM project, more precisely under the *Objective 2: To research, design and verify algorithms to automatically detect risk situations in victims of Gender Violence*. Towards the end of the thesis, it has also taken shape from SAPIENTAE4Bindi, under *Objective 1: To reach a level of maturity of the Bindi system equivalent to a TRL (Technology Readiness Level) 7-8 to be able to implement it with guarantees in the mechanisms of protection and attention to the Victims of Gender Violence in public administrations*, continuing the analysis of the input modalities of the Bindi portable devices and their processing for the automatic detection of situations of risk.

To define the focus of this thesis, we describe the general and specific objectives as follows:
General Objective (GO):

(**GO.1**): To understand the reactions of women – including victims of gender-based violence – to situations of risk or danger to the point of being able to generate automatic detection mechanisms for these situations from the auditory modality.

Specific Objectives (SO):

(SO.1): To **identify the voice of the speaker** from among all the information contained in the audio signal, coping with the influence of emotions or ambient noise. This is needed first to then perform emotion recognition in speech.

(SO.2): To develop robust machine or deep learning computational models based on the **speech signal** to **detect *fear*** or *panic* – or in its absence, its close relative: *stress* – in the voice of the target speaker, reflecting the emotional state the user and giving insight of the context in which the user is in.

(SO.3): To investigate and develop robust multimodal computational models with the same function as SO.2 where different modalities are combined in an intelligent way according to the constraints of the Bindi wearable devices. This thesis project will explore methods of merging the SO.2 models with those of other modalities contributed by other UC3M4Safety team. This objective is interdisciplinary, as it requires from **data fusion techniques**, performing a very thorough, conscientious and precise work.

(SO.4): Research and develop methods to customise the SO.2 and SO.3 models to the individual user, in particular with regard to the auditory modality. This **personalization** step is crucial to increase the performance of the generic models.

(SO.5): To **ignite the interest of the research community** in developing solutions for the prevention of the very challenging problem of **GBV**.

1.4 Contributions and Structure of the thesis

In this section we present the contributions of this thesis and the structure to be followed throughout the chapters. The main scientific contributions of this thesis are as follows:

- A comprehensive analysis of the pervasive problem of gender-based violence and the use of AI as a technological solution, including challenges, existing strategies and solutions, and limitations. Particularly, the ethical considerations and challenges of the use of Affective Computing from the gender perspective.
- A study of realistic stressful speech databases in literature that would be suitable for our goal of *fear* recognition through speech and its constraints, including a robust and stable method to label emotions when annotated continuously by a limited amount of expert raters.
- A novel multimodal laboratory database for the elicitation of realistic emotions – including *fear* – by means of virtual reality called WEMAC together with other UC3M4Safety team. It includes the physiological and speech variables of the participants, together with different emotional annotations.
- A novel multimodal database in real-life conditions captured with Bindi called WE-LIVE, that includes physiological and speech variables, geolocation and wrist accelerometers, together with different emotional annotations also in close collaboration with other UC3M4Safety team.
- A data augmentation strategy to cope with the negative effects of stress conditions in speech for the task of speaker recognition by means of synthetically generated stressed speech. This technique could be extrapolated to the problem of *fear* detection through speech when data are limited.
- A robust machine learning model for the task of speaker recognition under noisy conditions that denoises speech at the time that identifies the speaker, including speech under stress conditions. We prove the model to be more robust and stable than other methods for the detection of the speaker in conditions of loud noise.
- A design of a cascade multimodal system for Bindi 1.0 and its consequent evolution to an asynchronous hybrid fusion system of the physiological and speech modalities for *fear* detection.
- A design of an overarching Internet of Things system with edge, fog and cloud computing components for of BINDI 2.0 again together with other UC3M4Safety team. Specifically detailing how we designed the intelligence architectures in the Bindi devices for *fear* detection in the user and the experimental validation of such data pipelines.
- A design of low computational complexity models for the detection of realistic stress through speech.
- A design and validation of a system for the detection of realistic *fear* using monomodal – speech – and multimodal – speech and physiological signals – data systems, emulating the live operation of Bindi's two wearable devices, using different data fusion approaches and a speaker-adaptation strategy.
- A methodology and an use case in a preliminary research on the field of *Affective Acoustic Scene Analysis*, to study the relationship between an acoustic scenes and the emotions that they can provoke in people immersed on them, in collaboration with other UC3M4Safety team.
- A preliminary study on gender-based violence condition detection through speech and paralinguistic cues also together with other UC3M4Safety team.

As we have introduced before, this thesis is a comprehensive study of how we can use the auditory modality from the gender perspective for women's protection against GBV. Now, we briefly present the structure of the rest of the dissertation.

In Chapter 1 we describe what is GBV and its consequences, serving as the motivation and background for this thesis.

Chapter 2 presents an introduction to affects, emotions and how they emerge and their effects in the human body. It also introduces to the topic of Affective Computing, the AI research field aiming to give the ability of emotional intelligence to machines, including to simulate empathy.

Chapter 3 describes and and justifies the use of datasets containing stress as the starting point of our investigation. Additionally, and as a consequence of lack of realistic fearful speech databases in literature, we describe one of the main contributions of our UC3M4Safety team that is the creation of our own set of datasets to fill such literature niche: UC3M4Safety Audiovisual Stimuli Dataset, WEMAC and WE-LIVE.

Chapter 4 and Chapter 5 are task-focused and experimental, each introducing the topics of Speaker Recognition – studying topics such as speech denoising and speaker identification under stress conditions –, and Emotions Recognition – particularity focused on negative ones, e.g., *stress*, *fear* –, respectively, including the works carried out on this thesis for each field. A comprehensive overview and discussion about the operation and significance of Bindi is also offered in Chapter 5.

In Chapter 6 we address other complementary research works to this thesis, such as *Affective Acoustic Scene Analysis* or gender-based violence victim condition detection from speech.

Finally, Chapter 7 presents the conclusions from the research works conducted on this multimodal and multidisciplinary thesis and what we aim to continue doing after it as future work.

Chapter 2

A Multidisciplinary Perspective on Affective Computing

The present chapter consists of a definition of the field of Affective Computing. It covers the basis of affect, mood and emotion and from where do they emerge, the different theories of emotion in affective sciences, their current applications and a few important ethical considerations.

2.1 Affect, Emotions and Mood

The research field of Affective Computing (AC) comprises "the study and development of systems that can recognize, interpret, process, and simulate human affects and emotions" [75]. It is a multidisciplinary field that involves the fields of computer science, psychology, and cognitive sciences, focused in allowing robots and computers to respond in an intelligent way to natural human emotional feedback.

Affect is the unified term to describe states of feeling, such as moods and emotions. In [76] the authors define that "affective states vary in several ways, including their intensity, duration, and levels of arousal and pleasantness" – which we will describe further in Sec. 2.3.2 –. The study also declares that "emotions play an important role in regulating cognition, behavior, and social interactions, and affect is considered the experiential state of feeling. Even though in everyday language, terms like affect, emotion and mood are often used interchangeably, affect is conceived as the superior category to which emotions and moods belong" [76].

Moods and emotions are mostly differentiated by their duration in time and they are triggered by specific cause. Emotions are rather intense and ephemeral experiences that can happen for two reasons. On the one hand, they can be triggered in response to a particular external stimuli (for instance, events, actions or objects), and might emerge somewhat unconsciously. In the other hand, they can follow a cognitive judgement of a stimulus happening in the moment (e.g., How personally relevant is this stimulus?, Does this stimulus have any relationship with my goals?) [77].

Moods moreover, have a longer duration in time than emotions, and have a more diverse nature. For example, a generalized feeling of sadness without a definite origin could be understood or interpreted as a mood state. These experiences of affect happening recurrently over a prolonged time period can denote people's subjective well-being, for instance their global satisfaction with life, or be a sign of depression.

Affect has essential cognitive functions, we use it as a supply of information when making deductions about objects or people, priming agreeable memories, and influencing information processing and decision making [77].

In this thesis we focus on the detection of emotions rather than moods. Emotions can be described as "the brain's best guesses of what the bodily sensations mean, guided by the past experience" [78].

2.2 Neurophysiological basis of Affects and Emotions

The limbic system of the brain comprises a group of structures which are in charge of regulating emotions and behavior. Located deep within the brain, it's the part responsible for behavioral and emotional responses [79]. Some of the structures that are implicated with the actions of the limbic system are located beneath the cerebral cortex and over the brainstem [80]. Some of them are the thalamus, the hypothalamus which is in charge of the production of principal hormones and managing thirst, hunger, and moods among others; and basal ganglia, which reward processing, habit formation, leaning and movement. But the two major and most important for emotional processing structures are the hippocampus and the amygdala [81].

- **Hippocampus**: It is in essence the memory centre of the brain. It is where episodic memories are formed, catalogued and archived in long-term storage crosswise several parts of the cerebral cortex. Connections created in the hippocampus help to associate the senses with memories. Spacial orientation has also some origin in the hippocampus [80].

- **Amygdala**: It is key for the generation of emotional responses and it is located right next to the hippocampus, specially responsible of feelings like *pleasure*, fear, *anxiety* and *anger*. Emotional content attached to the memories is due to the amygdala. The amygdala modifies the intensity and emotional content of memories and plays a crucial role in creating new memories, especially ones related to *fear*. Fearful memories are created only after a few reoccurrences, which makes 'fear learning' a well-known method to research on memories formation and consolidation, and recall [80].

- **Hypothalamus**: The hypothalamus is one part of the brain that takes charge of growth, metabolism, sexual differentiation, emotional responses, and the desires and drives which are necessary so that an individual can survive [82]. The hypothalamus, together with the pituitary gland, administers the blood pressure, emission of hormones, force and rate of the heartbeat, body temperature, and electrolyte and water levels. The hypothalamus is also the core for the administration of the activity of the two parts of the Autonomic Nervous System (ANS), the sympathetic and parasympathetic nervous systems. Emotional expression depends largely on the sympathetic nervous system, and it is controlled by regions of the brain hemispheres above the hypothalamus and by the midbrain below it.

The limbic system, particularly the amygdala, are key controlling different emotional behaviors, such as anxiety, rage and fear [83]. And from a biological point of view, fear is a key emotion, as it assist the body to answer accordingly to threatening situations that could be of harm for an individual. The *fear* response

is caused by the stimulation of the amygdala. Initially the amygdala triggers the hypothalamus, which initiates the fight-or-flight response.

2.2.1 *Fight-flight-freeze* response

The fight-flight-freeze response –also known as fight-or-flight response– occurs as a consequence to an event that is recognized as stressful or frightening, it is an automatic physiological reaction of the body [84]. Examples of such can be seeing an oncoming vehicle getting fast in the way, getting spooked by someone, or acknowledging someone walking behind you while walking down a street. The fight-or-flight response, from an evolutionary perspective, is considered an adaptive instinct that humans evolved when environmental stimuli or predators endangered the survival of humans.

Specifically, fight-or-flight is an active defense response where the person either fights or flees. The body is affected by physiological changes to get the person ready to act appropriately and rapidly. Freezing can happen before the brain decides fighting or fleeing, being fight-or-flight on hold momentarily. It's also called reactive immobility or attentive immobility. But the moment in which the mind and the body are conscious, by a process called neuroception, that fighting or running are no longer alternatives to deal with the perceived threat, this response switches to the option of remaining still during the entire threatening situation as a last alternative to save itself [85]. Fight-flight-freeze is an automatic – non-concious – reaction of the brain and the body, which can't be triggered or controlled.

The fight-flight-freeze response can not only be triggered by an event, but also by a psychological fear. The brain associates negative experiences with a specific situation, which means that *fear* is conditioned. What can cause fear is called a perceived threat, or something the brain considers to be dangerous, which are different for each person. When facing a perceived threat, the brain thinks the person is in danger, as it acknowledged the situation to be a threat for one's life. Thus, the body unconsciously reacts with the fight-flight-freeze answer to preserve one-self's life [86].

In such cases, the fight-flight-freeze response is called to be overactive. Which means it happens when situations from normal life that are not actually threatening trigger the reaction. These overactive responses are frequent in people who have experienced traumatic events or suffer from an anxiety disorders. The example of seeing somebody walking behind you on the street alone need not be dangerous per se, but can trigger the fight-or-flight response if you are a woman, the person is a corpulent man, and it is past midnight.

Unlike males, who mostly experience the fight-or-flight response under a threatening situation – first proposed by Bradford Cannon in 1915 –, women seem to have two equally likely responses to stressful conditions, fight-or-flight and tend-and-befriend presented by Shelley Taylor in 2000 [87]. This tend-and-befriend response produces similar biochemical changes in the body to the fight-or-flight response.

Due to the exclusion of women from clinical trials in research – as we already described in subsection 1.2.1 –, the tend-and-befriend theory was discovered just two decades ago. It states that when faced with a perceived threat, females will tend to the protection of their offspring (tending) and to look up for social group to get mutual defense (befriending) [88]. It is believed that, due to natural selection, humans have a biological system that manages social interactions in the same way it regulates basic needs as thirst or hunger. And it appears to have its roots in

this instinctual need to protect children and affiliate with others for greater safety. Additionally to the basic needs for physical wellbeing, humans are social creatures that rely on their instincts to interact with others. The females are often inclined to protect and take care of offspring.

Consequences in Physiology

When faced with a dangerous or threatening situation, the emotional state of the person can be fear, panic, stress, nervousness, shock, insecurity, worry, ... Afterwards, a series of reactions triggered by the fight-flight-freeze mechanism involve physical and physiological changes occurring in the human body.

The first reaction to a threatening situation is generated by the amygdala on the brain, more specifically in the limbic system [77]. The amygdala is responsible for regulating the fight-or-flight response and plays a key role in processing of fear. In the moment that danger is perceived, the amygdala triggers a signal to the hypothalamus, responsible for hormone release, and connects the endocrine and nervous systems. The latter then informs the rest of the body via the autonomic nervous system, which is in charge for controlling the involuntary mechanisms of the body, and stimulates the sympathetic nervous system (SNS).

When the amygdala triggers a distress signal, the hypothalamus turns on the sympathetic nervous system by transmitting signals to the adrenal glands [89]. These glands react by injecting adrenaline through the bloodstream.

This causes a series of physiological changes that retrace energy from areas of the body which are associated with resting or passive processes – such as the digestive system – to areas of the body that support the individual to be ready for action in case of an emergency so that it can avoid harm [90]. The heart starts then to beat faster than in a normal state, providing blood to the muscles, heart and other vital organs. Heart rate and blood pressure also increase. The blood vessels in the muscles dilate and muscle tension increases to give the body greater speed and strength [77]. The breathing rate increases and the airways of the lungs open so that they can get in the maximum amount of oxygen in each breath. All the extra oxygen available is sent to the brain so that alertness is increased. Senses such as hearing and sight are also sharpened [89] [91].

During the fight-or-flight response, besides causing vasodilation in skeletal muscle for speed and strength, the SNS also causes vasoconstriction in the skin, to allow blood to reach the major organs, leaving the skin looking paler. As blood vessels narrow, the body can heat up very quickly, so fight-or-flight response also increases perspiration (sweating). The body is cooled to prevent overheating by evaporating sweat, and thus allows to continue fleeing or fighting from harm without feeling exhausted from the heat.

In addition, epinephrine – also known as adrenaline – triggers the release of fats and blood sugar, at it is these nutrients which flood the bloodstream, providing with energy to all parts of the body. As long as the threat remains, the hypothalamus will continue to signal the SNS to keep secreting adrenaline and cortisol to maintain the body's activation [92]. The release of excessive adrenaline can cause sympathetic Acute Stress Disorder (ASD), also known as state of shock.

Specifically, the (ANS) – composed of the sympathetic and parasympathetic systems – is responsible for regulating involuntary physiological processes. Within the limbic system, the hippocampus – responsible for the creation of memories with emotional context – forms emotionally meaningful representations and interpretation of events. The amygdala – responsible for associating fear

with threatening situations – is also particularly important in the processing of fear-related emotions. This connection between the amygdala, the hypothalamus and the ANS is closely related to reflexes and fight-or-flight responses, fear expressions in the body, and the activation of neurotransmitters such as adrenaline and cortisol, which are linked to stress responses.

When the response is neither fight or flight, but freezing, the parasympathetic nervous system (PNS) takes charge, and takes its role of relaxation to the extreme by activating the unmyelinated vagus nerve, the main nerve of the PNS [93]. This neural network works as a vagal brake on the heart, because it slows it to a lower beat than the one at rest state, it also numbs senses and muscles by dispensing chemicals into the bloodstream [85].

From the fight-flight-freeze responses, the resulting outcome depends on how the body has learnt by experiences to deal with each kind of threat, along with the inborn fight-or-flight plan in the brain, which determines the most favourable reaction in order to overcome the threat [94] [95].

Consequences in Speech Production

As a reaction to perceived or real danger, the ANS performs many different changes in the body, regarding the heart rate, muscle activation and vocalisation, breathing, just to be able to deal with the threatening situation. Depending on the situation and the individual, each experience different variations of the flight, fight and freeze responses [96].

One of the main responsible for these changes is the 10th Cranial nerve or the Vagus nerve. It is the longest nerve of the ANS and goes through the mouth, tongue, larynx, heart, lungs and digestive system [97]. The vocal tract specifically – formed by the vocal folds, larynx and pharynx – has a complex nerve system which includes input from the SNS.

Stress, anxiety and fright can have a profound affect on vocal performance when triggering the fight-or-flight response. Muscle tension can lead to having a constricted throat and vocal chords and resulting in a person's voice becoming higher pitched, quietening voice or even loss of voice entirely [96]. Muscle constriction can also cause increasing speech speed, tension in the jaw and tongue hindering intelligibility, and the shutting down of salivation makes the mouth feel dry and can produce a raspy voice. The increased breathing rate can also lead to loss of breath while speaking [96].

When freezing, the heart slows to a beat lower than that at rest, and the muscles feel numb due to the release of chemicals into the bloodstream. This reaction also locks the vocal folds apart to keep the oxygen flowing into the lungs. This is the reason why in the freezing mode a person may feel it physically impossible to speak, scream or call for help [85].

Due to all these physical and physiological changes and their involuntary nature – the person has no control over them – that occur in a person as a result of being in a situation of risk, we considered relying on physiological signals such as pulse, perspiration, respiration, and also speech, in order to detect the emotional state of a person – with the intention of recognising fear –, which could be a consequence of being in a threatening situation. An example of a life-threatening situation that could trigger the fight-flight-freeze responses in women are gender-based violence situations, those in which a women suffers a physical or sexual assault. Finding oneself in a situation of potential danger can entail the aforementioned physical and physiological changes in the body.

2.3 Theories of Emotion in Science

After examining where the emotional response arises in the brain and how it affects the body, – especially fear –, now we present the two main different perspectives of emotional theory in the field of emotion perception. Whereas it is the discipline of affective neuroscience the one that aims to develop an deep understanding of emotions, moods, and feelings and how they are integrated within the brain, yet there is still no scientific consensus on that there is only one valid theory of the fundamental nature of emotion. This is where the two apparently opposed theories that govern the field of emotion perception arise: the categorical theory and the dimensional theory.

2.3.1 Emotions as Discrete Categories

The categorical theory of emotions posits the existence of six well-defined universal emotions: "*happiness*, *anger*, *sadness*, *surprise*, *disgust*, and *fear*" [98].

These emotions are basic to humans as we are equipped with biological instruments to react to universal life situations – such as successes or losses –, and each emotion guides for a reaction that, during evolution, worked more effectively than other solutions in similar relevant circumstances for human survival [99]. Each one of the basic emotions is not an individual physiological or affective state but somewhat a family of related states that people from all cultures may have experienced due to similar adaptive problems, so these emotions are described as universal.

Even though the categorical perspective of human emotion does not essentially need an evolutionary explanation of its origins, humans and animals experience discrete categories of each emotion as each is believed to come up from an adaptation that was developed to solve a singular adaptive problem [99]. As an example, the discrete emotion of *fear* was thought to be developed as a mechanism to enhance the survival of the individuals by avoiding dangers over time via evolution [100]. Table 2.1 provides the six categorical basic emotions and next to each, the adaptive difficulty the emotion might have evolved to solve [101].

Discrete Emotion	Adaptive Problem
Happiness	Seeking valuable mates
Anger	Managing physical threat in the environment
Sadness	Strengthening social bonds by inducing compassion [102]
Surprise	Awaring of a schema-discrepancy signal [103]
Disgust	Avoiding or expelling poisonous food
Fear	Avoiding danger

TABLE 2.1: Adaptive problems solved by the basic 6 emotion categories, from an evolutionary perspective [101].

Basic emotion theory, by applying the Darwinian theory, suggests that the adaptability feature of emotions increased our gene survival by improving reproductive options (e.g., joy encourages people to explore and meet new possible mates) or by by dealing with threats to reproduction (e.g., disgust assists us in avoiding death) [101] [100]. Yet, some AI applied research is investigating emotions beyond the big six [104], and the ultimate theory of emotions has not yet been agreed upon in the research community.

2.3.2 Dimensional Space of Emotions

Dimensional theories are opposed to the theories of discrete, basic emotions due to the latter not fully explaining some observations in empirical studies of affective neuroscience. The circumplex model of affect [105] is a dimensional theory that suggests that "all affective states derive from cognitive perceptions of neural impressions that are the product of minimum two independent neuro-physiological systems: one related to arousal or alertness, and the other related to valence – a pleasure-displeasure continuum –" [106].

These models based in continuous dimensions – dimensional models – think of affective experiences as a continuous range of well interconnected and indefinite states [105]. In the end, emotions are seen as "the product of an intrincate communication between cognitions, probable of occurring initially in neocortical structures, and neurophysiological changes related to these valence and arousal systems" [105]. There is one system associated generally with pleasure and reward, the mesolimbic dopamine system, and it might represent a neural substratum for the dimension of valence [105]. In addition, the reticular formation is believed to adjust arousal balance of the central nervous system over its connections with the limbic system, thalamus and amygdala [106].

However, since 1974, there are psychologists discussions around this theory, about the specific interpretation of the dimensions connected to affect and cognition And from the point of view of categorization of emotions, using the arousal-valence representation or space, emotions such as fear or *anger* would lay very close to each other, when in fact they have different physiological consequences and the sensation of each is different [107].

The PAD space dimensional theory adds one axis to the arousal-valence space [108]. The PAD space, pleasure (valence), arousal and dominance, is formed by three independent emotional dimensions that are thought to describe human emotions [109]. Pleasure – which is the valence – is believed to be as a continuum ranging from intense happiness to extreme unhappiness or pain; it comprises extremes such as happy-unhappy, satisfied-unsatisfied and pleased-annoyed, in order to determine the level of pleasure of an individual. Arousal is known to be the amount of mental activity along a single dimension describing emotions, which ranges from sleep to extreme excitement. On each end, the words that describe arousal are stimulated-relaxed, excited-calm and awake-sleepy. Additionally, dominance was thought to be related to feelings of control and restriction, expressing how much an individuals dominates the emotion from his behaviour. The degree of dominance lies in a continuum range from complete dominance to submissiveness, with descriptors such as autonomous, influential and controlling [109].

Emotional databases can be labeled either w.r.t discrete emotions or continuous emotions, thus recent works propose a discrete-continuous mapping in AC research [110], [111], [112] and some even suggest the need for a fourth axis to accurately represent discrete emotions in a 4D continuous space [113].

Focusing on dominance, it refers to the feeling of influence and control over other people or/and the surroundings or environment versus feeling controlled or influenced by others or the situation (e.g., *anger*, *power*, versus *anxiety* and *fear*) [109]. And since in this thesis we have a focus in detecting the fear provoked by a GBV situation, the dominance axis is of special relevance for the labeling of the emotions, clearly explaining when a person feels completely overwhelmed by the emotion and controlled in such situation.

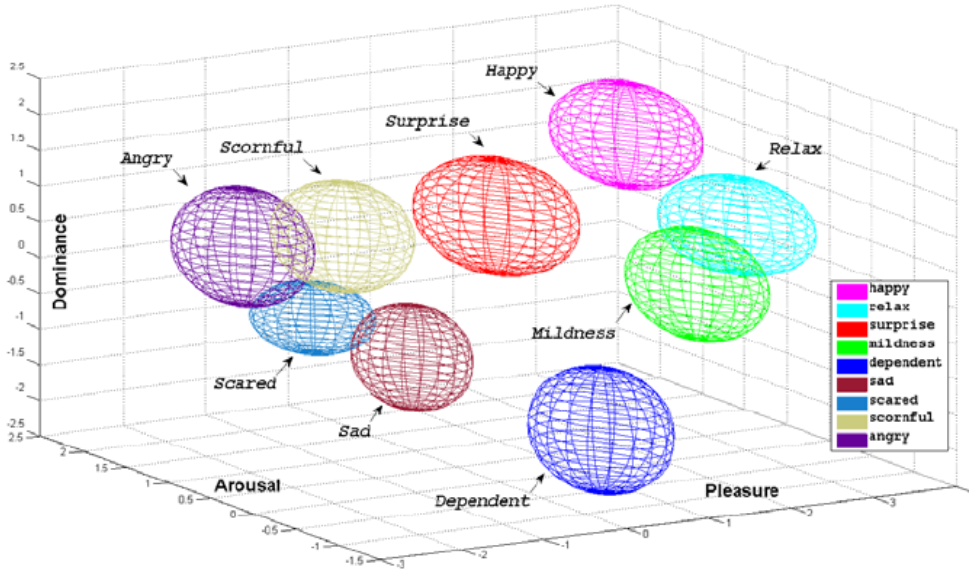


FIGURE 2.1: Emotional mapping from discrete to continuous abbreviated PAD emotions space [112]. Reproduced with permission of the copyright owner, Springer Nature.

2.4 Interpretation and Understanding in Affective Computing

After describing how emotions arise and what theories there are on how to classify them, in this section we explain more in depth about the comprehension of this branch of artificial intelligence that deals with emotion-related tasks.

Affective Computing gives an account of, appears from, or impacts emotion [114]. It is a multidisciplinary field that is continuously growing. It investigates how machines can get to interpret human affect, and how communication between humans and machines can be embedded by affect, how we can design systems with affect so that their capabilities are enhanced, and how computer interaction can be transformed by sensing and affective strategies [115]. It encompasses several disciplines such as psychology, engineering, cognitive science, education, sociology, and more.

Affective Computing is based on Machine Learning and Deep Learning. According to [116], "Machine Learning is the subfield of computer science and a branch of artificial intelligence, which aims to develop techniques that enable computers to learn". An agent is considered to learn when its efficiency is improved by experience with the use of data. In the process of machine learning, first a computer software is fed with data an it observes it, then it builds a model based on such data, and uses the model in two ways, first as an hypothesis about the world and second as a method to solve problems and make inferences and predictions about new data.

On the other hand, deep learning is a set of machine learning algorithms with the same purpose, they aim to model high-level representations in data using complex computational architectures that support non-linear transformations of data [117], thus having the most flexible ability to model real-world problems with more generalisability. Deep learning is part of a wider group of machine learning methods which are based on comprehending representations of data. Research in this area

attempts to define which representations are optimal to be understood and how to create models to recognise and interpret these representations.

Artificial Intelligence disciplines can be classified according to the type of data they employ, e.g., the Computer Vision field analyses images and videos, Speech Technologies work with the processing of speech data, Natural Language Processing (NLP) utilizes textual data, ... But Affective Computing is a discipline of AI defined by the task to be achieved, not restricted by the type of data used.

One categorization that has been made in the field [118] establishes four areas on it, which are namely, " 1. the analysis and characterization of affective states –identification or detection–, 2. the automatic recognition of affective state –recognition–, 3. the expression of affective states –generation and elicitation–, 4. and adaption of response to user's affective state".

Thus, within the target tasks involving emotions in AC, helped by dictionary definitions [119], we can then describe the following:

Detection - "*to notice something that is partly hidden or not clear or to discover something, especially using a special method*". In AC, it can be also known as identification, and it applies to finding an alteration of the emotional state of a person, by a discovery of something happening that it is altering the neutral basal state in which the person was inertially. Among the applications of this area are the identification of emotions or triggers that provoke them.

Recognition - "*to know someone or something because it has been seen or heard or experienced before*". It goes a step further than the detection in AC, because it analyses the new detected state and categorizes it, in some category –either continuous or discrete– of already known emotional states. This area is a consequence of the previous one, as in addition to identifying the moment in which an emotion occurs, it is also in charge of classifying it into a known type.

Generation - "*the production or creation of something*". To this area belongs the ability to create and synthesize emotions by machines. Examples such as the generation of emotions in speech –Emotional Text-to-Speech (TTS)– or the generation of text with emotional content.

Elicitation - "*to produce something, especially a reaction*". This area is in charge of inducing, triggering or evoking an emotional state in a person, which ultimately can influence their actions or reactions. With applications such as inducing comfort in telephone customer support.

The convergence of the four aforementioned areas results in the ability of **adaptation** in machines, mirroring *empathy*. Adaptation is defined as "*something produced to adjust to different conditions or uses, or to meet different situations*". Then it can be said that they simulate empathy, when they are able to identify an emotional state in a person, and then recognise it, to then generate a synthetic affect response, and be able to elicit other emotional reactions in people, imitating the mechanism of human *empathy*.

But this so-called empathy is an holistic process for which machines need to understand not only the emotions but also the context and situation that led the person to that affective state. In our specific case, we aim to detect risk situations for women, and this can't only by achieved by recognizing the affective state of a woman by means of her user data –i.e., physiological variables, speech– but also together with the situational data –i.e., GPS location, environmental sounds, time of the day–. All these data together give contextual information to perform an holistic interpretation and understanding of the situation, which could, ultimately, determine when the life of a person is in danger.

Interpreting and understanding is classic in other AI fields, i.e., in Speech Recognition, phonemes of spoken speech are identified – voice-unvoiced decision –, and recognized – vowels and consonants – which words and sentences that ultimately have a meaning, interpreted by our higher cognitive understanding. In this work we are pursuing a similar goal, to make it possible to identify, for instance, a sudden rise in heart rate, along with the recognition of hurried footsteps and panting from an audio signal, and that we can interpret together those – at first glance, isolated – events, making a holistic understanding of the situation, determining its level of risk to the user.

That said, with all this information we aim to take one step forward with Bindi on Affective Computing, towards a higher cognitive level. We aim not only at analysing data from an user in order to detect and recognise isolated emotions from a basic computational level, but going beyond to *interpreting* and *understanding* such affective states, together with the situational information. This will lead us to comprehend its context and circumstances, to finally be able to detect a risk, threatening or dangerous situation for a woman.

2.5 Challenges: Subjectivity, Annotations and Gender

Rosalind Picard, who coined the term Affective Computing [75], describes the term *emotion* as "the relations among external incentives, thoughts, and changes in internal feelings; just like weather is a superordinate term for the changing relations among wind velocity, humidity, temperature, barometric pressure, and form of precipitation" [120]. She defines a weather metaphor, stating that "a unique combination of meteorological qualities creates a storm, a tornado, a blizzard, or a hurricane, events that are analogous to the temporary but intense emotions of *fear*, *joy*, *excitement*, *disgust*, or *anger*. But wind, temperature, and humidity vary continually, and not necessarily produce such extreme combinations. Thus meteorologists do not ask what *weather* means, but determine the relations among the measurable qualities and later name whatever coherences they discover" [120]. In the end, Rosalind Picard states that it is difficult to expect researchers to be successful matching human labels when those labels might not specifically exist, comparing the problem to not having specific terms for most of the states of weather but only names for its extreme states, and so the same applies to emotions.

This metaphor makes it clear that emotions are not objective, they are not digits that can be clearly recognized and differentiated. Emotions are imbued with subjectivity, and this particularity is bi-directional, as it has two parts/directions.

First, there is an intrinsic difficulty in labeling or categorizing the innermost feelings of oneself, even though we have better access to them than anyone else. Still many people do not know how to connect with their own feeling state and recognize them, although people have feelings permanently [121]. Many Affective Computing databases are labeled by self-annotations of the subjects participating; sometimes w.r.t. discrete emotion categories and sometimes referring to continuous axis – PAD space –. It also depends on the emotional training of each person, i.e., if the person has not been taught to identify their own emotions or has no experience recognizing them – lack of emotional intelligence, still a pending subject in many schools [122] – their own emotions, they may have different perceptions of what the Likert scales of Arousal or Valence mean. This can vary very much depending on the person's background and culture. Just as 2mm of precipitation per hour is a weak rain in Spain, in the Philippine Islands it may not even be considered rain.

Second, some other databases are labeled by external annotators, independent from the subject experimenting the emotion. And in the path of an emotion from generation to externalisation, the person can have some degree of control of such externalization – we already explained some of the uncontrollable consequences of fear from the autonomic nervous system, but not all of the externalization is automatic, straightforward or unavoidable –, making it difficult for the annotator to ascertain correctly the emotion being presented by the subject if the person does not openly externalise it.

To these two is added a third subjectivity. In the particular case in which the aim is to annotate the elicited emotions in a subject by means of the visualisation of an audio-visual –or other sensory (i.e., olfactory, gustatory or tactile)– stimulus, or an real-life experience of a specific kind, aiming for a target emotion, the situation can be perceived differently by one person than by another. For one person the visualization – or the experience – of the nightlife in a crowded city can be exciting, appealing, thrilling, but for others it can be stressful, disturbing or tense. This would mean that even when trying to elicit a certain emotion in the viewers –i.e., for the generation of a database– it may not be able to elicit the target emotion because not all people react in the same way to the same stimuli.

All this subjectivity of emotions means that in the field of Affective Computing, where AI models are trained with data and labels, they do not have "black-white", objective labels as in other areas of AI. Then, either self-annotated or externally annotated emotional labels should not be taken as absolute gold standard labels. People can react diversely to the same stimuli, including during different moments of time according to many variables, such as the state of mind, past experiences, culture and background. That makes Affective Computing, a field of Artificial Intelligence subjective and slightly elusive, in which all of these nuances must be taken into account.

In line with the interpretation and understating of situations as a whole, there are to date no databases – to the best of our knowledge – that specifically serve to identify and understand emotional situations. And as described in Sec. 2.4, decisions should be taken away from the theoretical framework in which isolated emotions are analysed and processed, to understand emotional or affective situations in an holistic manner, taking the context into account, to fully understand why a situation elicits a certain emotion in a person. This is crucial in the detection of risk situations.

On a different note in parallel to Section 1.2.3, and in line with the subjectivity of emotions that depend on the person, another challenge that arises is the gender personalization challenge. There seems to be clear differences in the expression of emotions according to sex [123]. It has been found that men and women more precisely display gender-stereotypical expressions, – arising from gender socialization –, as men more correctly express *anger* and *contempt*, while women more exactly express *fear* and *happiness* [124] [125]. Specifically regarding the elicitation of emotions, when visualizing gender-based violence videos, the identification of the viewers with the protagonist in the video affects directly to the labelling as women favour to label mainly *fear* while men label the emotion as *anger* or *sadness* [126]. It could be that societal restrictions on the emotional expression in men were a reason for the high rates of violence against women perpetrated by men. These masculine ideals, such as the pressure to meet to expectations of dominance that society imposes, might increase the potential for boys to involve themselves in general acts of violence, as assaults, bullying and/or physical and verbal aggressions [127].

Regarding such gender differences on the data used to train Machine Learning (ML) models, some studies show that taking age and gender into account in AC can convey an improvement in the accuracy of emotion recognition tasks [67], stating that subject-specific variables should not be overseen in AC analysis, such as gender, personality and age. It is well known that gender-dependent emotion recognition systems perform more efficiently than gender-independent ones, thus some studies improve the discerning quality of gender-dependent features [128], or model gender information for more robust emotional representation [129], in order to achieve better accuracies. Results on this topic can be found on Sec. 3.3.1.

2.6 Ethical, Practical and Legal Application Considerations

Affective Computing has made great social impact since its emergence. Some ethical concerns to be discussed in AC are related to, generically discrimination and biases, abuse of influence and manipulation, mental health and safety, and sensitive data privacy.

Yet, there is not guideline of principles to contemporary research ethics protocols and standards, but some studies aim to gather the most common ones, such as [130], [131]: "1. informed consent, which implies the avoidance of covert or secret participant observation 2. privacy of participants (confidentiality and anonymity) 3. avoiding harm (including psychological effect) and doing good 4. cognisance of vulnerable groups 5. participants' right to withdraw or terminate 6. restricted use of data 7. due care in the storage of data 8. avoidance of conflicts of interest".

Rosalind Picard [132] raises the concern that "a computer that can express itself emotionally will someday act emotionally". In the case of the aforementioned *adaptation* of machines, the downside is the ability of machines of manipulating humans. For instance in the case of companies further understanding their clients' needs and wants, making possible to create a new type marketing, targeting emotional attachment and control.

In [133], there is a comprehensive discussion on the topic of privacy preservation and technologies in speaker and speech characterisation tasks. In [134], they give an overview of the paralinguistic phenomena that can be or even is used to obtain personal information by means of speech signals. In [131] the authors suggest guidelines for good practice in Computational Paralinguistics (CP) and AC, such as choosing the proper performance metric, and accounting for interpretability and representativity [135].

From the point of view of **our application** – developing a wearable device that is able to detect risk situations for the user and alert emergency services automatically if needed – there are some key ethical concerns to consider.

It is possible that such system – specially in its first stages of development – can make mistakes. It could trigger up false alarms and even miss out risk situations. And this possibility of failure could have dangerous consequences. There is a need to strike a balance between having false negatives and not having false positives, being preferable to have never any of the former at the expense of having some of the latter. An approach called *passive-aggressive* learning takes charge of fixing the false positives and reducing alerts, in machine learning models [136].

We have previously discussed ML biases in Sec. 1.2.1. On many occasions the problem of biases in ML comes from the fact that most algorithms are considered black boxes, i.e., they provide the desired outputs in response to the inputs that are introduced but are not able to explain how they achieved that conclusion. When

we let such algorithms make decisions that have great significance for the people who make those decisions, as in the case of risk situations detection, we should be able to explain the reasons for those decisions. For both the supervisory technical staff and the user using Bindi, having a highly explainable system would provide everyone with confidence in its use, as well as being able to see more clearly how the developers' changes to the system affect the decisions Bindi makes.

A couple points to consider, also mentioned in Sec. 1.2.3 are PTSD and diversity.

The fact that users using Bindi would necessarily include victims of gender-based violence, needs special attention because the violence they suffered has post-traumatic stress consequences on them. Bindi has to take into account the specific needs that this type of users may need from women who have never suffered it, somehow taking into account an assessment of post-traumatic symptomatology.

There is a great cultural diversity in Spain, and even more so in Europe and the World. Culture is an orientation system for a nation, society, organization or group. Culture is also a subconscious action-influencing system of values and norms. And all these aspects affect the way we interact with the world, including how we express emotions. Bindi needs to take into account the target group that to which it is oriented and which will use at any given time, in order to be able to provide a system that can protect all women, taking into account their individual differences.

In the legal aspect, the European and National Regulatory Framework emphasizes the challenges of AI linked to the need to process real data including the trade-off between privacy and data protection and the tensions between explanation and prediction. Addressing these challenges is a task that requires the existence of appropriate legislation. There is already European legislation related to data protection, the General Data Protection Regulation (GDPR) [137], but in addition, work is also underway to specifically regulate AI: the Artificial Intelligence Act by the European Commission [73] is a draft that is expected to be approved shortly. The Secretary of State for Digitalization and Artificial Intelligence (*Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA))* proposes to test in Spain the new European Regulation on Artificial Intelligence through a pilot project that will test a new agency: the Algorithm Oversight Agency (*Agencia de Supervisión de Algoritmos*) which is expected to be set up by the end of 2022. The new European regulation will come into force on January 1, 2024.

2.7 Literature Review on Affective Computing and Gender-based Violence

Affective Computing makes use of AI – Machine Learning and Deep Learning – models for the aforementioned tasks: detection, recognition, generation and elicitation, of emotions. And there is not much literature on AC and GBV together due to the lack of data from gender-based violence victims, as explained in Section 1.2.1. But it is an emerging field as the world and the research community are are seeing it as the threat to human life and human rights that it is, and research on the merger of these two fields is growing.

In the NLP field, researchers [138] captured a dataset for the identification of feminicide (the murder of women because they are women) from 400 written media reports, together with its labels. They also trained a machine learning model using these data, achieving an accuracy on the test data set of 81.1%, with 400 samples (written articles).

In [139], the authors examine the fundamentals of activist and civil society efforts to collect counter data about feminicide and gender-related killings, reviewing on the efforts of activists to monitor and challenge gender-related violence.

In [140], they use a database of data collected over two decades of GBV in Spain. They use feature selection, predictive algorithms are applied and compared to predict quite successfully the number of GBV complaints to be presented to a court within the next six months in the country. The same team [141] has a study on a biosensors-based surveillance solution for the protection of GBVV, similar to our contribution [1], serving as a statement which shows that technology is increasingly being accepted and used as a solution to combat and mitigate gender-based violence.

With the rise of social networks in the last decade, and together cyber-activism and cyber-bullying, some works [142], [143] explore neural network models to identify gender-based violence on messages from Twitter in Spanish language based in Mexico, and discriminate among manually labelled GBV-intentioned tweets.

From the point of view of the other side of the violence, i.e., the perpetrator, in [144] they analyze the most influential variables and also predict the chance of perpetration of GBV by using questionnaires data from homeless youth in Los Angeles. Several supervised machine learning algorithms are used to build an intimate partner violence (IPV) perpetration triage tool to detect which young people are at high-risk for engaging in violence perpetration.

Regarding mental health, GBV leads to traumatic disorders such as Acute Stress Disorder (ASD) and Posttraumatic Stress Disorder (PTSD), and there is current literature on the usage of machine learning methods in the estimation of subjects with ASD and PTSD [145] where multiple levels of biological data – clinical, neuro-endocrine, psycho-physiological– or other data sources – i.e., demographic information– are used to predict early symptoms or identify risk factors related to PTSD or ASD.

There is also an interest in the research community in generating gender-neutral voices for voice assistants and eliminating gender bias [146]. But in general in the field of speech technologies there is little or no work to combat or prevent gender-based violence. Thus, this thesis aims to fill that niche and to explore and investigate the use of speech technologies for the prevention of gender-based violence, also igniting the interest of the research community in developing solutions for the prevention of the very challenging problem of GBV.

Chapter 3

Data Characterization for the Detection of GBV Situations

The way data is captured influences the methodology that can be applied to it, thus the methodology has to go hand in hand with the data capture process. In this chapter we want to bring together the methodological effort from the databases point of view, explaining the decisions taken with respect to the data used in chronological order for the research on this thesis. We detail the difficulties found to achieve our objectives due to the lack of suitable data available, as speech datasets of real *fear* (not acted) were are unavailable or non-existent in the literature. The closest realistic emotion to it is *stress*, so in this chapter we describe and justify the use of datasets containing such emotion as the starting point of our investigation. Additionally, and as a consequence of the previous problem, we describe one of the main contributions by the UC3M4Safety team that is the creation of our own set of datasets to fill such literature niche.

The design and collection of the datasets described in Secs. 3.3 and 3.4 has involved a huge effort of the members of UC3M4Safety team participating in the EMPATIA-CM project. As part of this effort, the following contributions were made to this thesis: the design of speech and audio data collection, protocol technical assistance and support, processing of speech data pipeline as well as assistance with its capture and user tracking both in WEMAC and WE-LIVE.

In line with a modern Data-Centric AI conceptualization [147], we want to focus in the use of appropriate data for our task, prioritizing the importance of *good*, i.e.fitting, data. This unique and recent technique involves constructing AI systems with quality data, with an emphasis on ensuring that the data clearly expresses what the AI must learn, rather than focusing on writing code. It emerged due to former more costly AI solutions adopted by improving AI models over the years – resources and economically wise –, and this approach bets for a necessary fundamental shift to truly unleash AI's full potential, by providing a systematic method for improving data, reaching a consensus on the data, and cleaning up inconsistent data.

3.1 Challenges of Auditory Data when used for GBV Detection

In our application we want to detect risk GBV situations through the auditory modality. We define in Fig. 3.1 an outline of the auditory data, tasks and conditions in which it can be recorded, to be used in the detection of GBV situations. There are 3 major components into which the tasks that can use audio data can be divided: *speech* – in which speaker and emotion tasks can be performed –, *audio (other than speech)* – for the detection of acoustic events and the classification of sounds,

among others –, and *background noise* – which could add beneficial or detrimental information, depending on the task–.

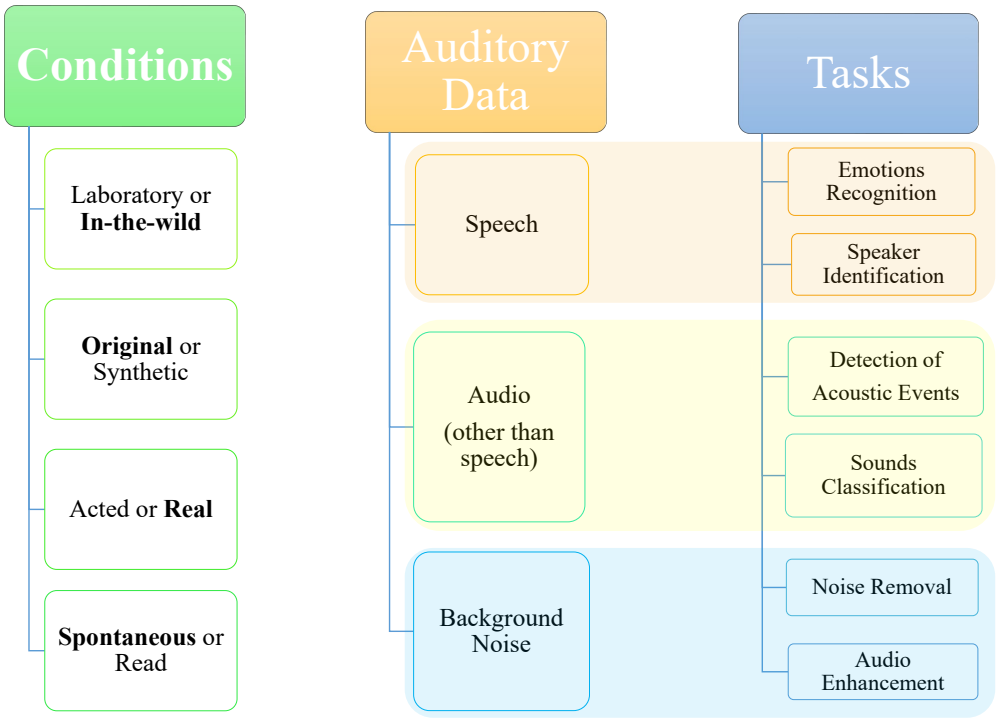


FIGURE 3.1: Auditory Data outline to be used in the detection of GBV situations.

Ideally, in order to build our ML or DL AI system for the automatic detection of risk situations from auditory data, we would like to count on speech belonging to the user concerned, plus background audio that could give us the acoustic context in which the user is set. The former ideally would be clean (not noisy) in-the-wild and spontaneous speech, including neutral and emotional – and fearful – speech. The latter ought to include acoustic events and background noise – which is not always detrimental but can add situational information – so that the situation can be fully comprehended. Additional sources of information such as physiological biosensors measuring information from the user would also be desirable, as those would give a more reliable insight of the circumstances, for a full and comprehensive understanding of the situation; and the use of such fitting data would lead to the development of successful ML models for the detection of risk situations.

In this thesis we follow a bottom-up approach, focusing on identifying and understanding the speaker and emotional content first, and then complement the conclusions gathered with the additional auditory info, to get a complete overview of the situation. As a first step, we need to recognise the user (Speaker Identification) and then the appearance of *fear* in speech (Emotion Recognition), and for this purpose we make an analysis of available databases in literature that can be used and afterwards we provide a full description of our efforts to create a suitable database for our purposes.

Audio databases are usually recorded under laboratory conditions but increasingly there are some that are recorded in-the-wild. This metaphor, alike to the dictionary definition of in-the-wild – beings living free and in a natural state, not

looked after [20] –, invokes realistic, in real-life, natural conditions and with natural characteristics. However, the performance of models evaluated in-the-wild is still unreliable partly because of diversity and often unknown contextual factors such as recording conditions that can be encountered in data in-the-wild.

The laboratory conditions mentioned usually have the advantage of coming from studio work where there is a target speech to record, exactly intended to capture, which is highly helpful. The counterpart is that these laboratory conditions are usually far from real-life conditions because they do not include everything that can appear when recording in the wild. Laboratory audio recordings are usually clear and clean – with non-existent or very little noise –, whilst in-the-wild those conditions are not met (noise from vehicles, domestic, outdoor and indoor sounds, even bumping sounds in microphone or rubbing if it is being wore). These two setting conditions are vastly different, the in-the-wild setting is affected by a lot of variability that can worsen the performance of the tasks.

Synthetically generated audio data cannot perfectly emulate original and real audio data, but this artificial generation is a way to test how the models work under specific conditions by using only a few resources – without using the great amount of resources needed (personnel, material, time, ...) for a database recording –. When we refer to *synthetically generated* data we are not describing *synthesized* data which results from text-to-speech models, voice-overs or voice generators; instead we refer to augmented auditory data (in terms of artificial modifications such as speed, pitch, combination or sum of two signals, etc). These synthetically generated data could serve as a preliminary way to test ML models or to obtain similar or indicative results of what we could expect with data recorded in such specific conditions.

The databases with emotional and neutral speech are either recorded by actors simulating speech under those emotions, or by people under actual real emotions, which are previously induced. Due to the characteristics of our task at hand, we prefer to prioritize the use of real emotional databases rather than acted ones, since actor's databases may be overacted, which calls into question the value of using actors to research actual emotions [148]. The difference between spontaneous or read speech gives rise to another type of categorisation. For our use case, we are looking for speech that is preferably spontaneous, again, more like real-life speech.

An ML model that could be trained with this type of data would be good for inference and predictive ability in real-world situations with spontaneous and real emotional speech. But in the literature there are very few databases that are original, real and spontaneous including emotional speech – most importantly including *fear* – from different speakers, specially female. Thus, taking into account these three limitations, we initially work with the most suitable databases found in literature with original, real and spontaneous speech, which include *stress*, a close relative of *fear*.

Because of its importance along the thesis, we believe it is necessary to mention the subject of *stress*. Despite the fact that *stress* is not considered as a recognized emotion, *anxiety* and *nervousness* are closely tied to it [3]. It is described as a condition of *stress* one brought on by challenging or unfavorable situation. Both internal and external variables, such as workload, noises, vibrations, lack of sleep, fatigue, etc., can cause *stress*. This literature work provides a very comprehensive overview of stress detection system [149], including the role of machine learning in emotion detection systems, feature selection methods, different evaluation measures, tasks

[20] https://dictionary.cambridge.org/dictionary/english/wild?q=in+the+wild

and applications. In addition to studying the connection between the biological nature of people's emotions and mental *stress*.

Stress has several physiological consequences, such as respiratory changes (faster breathing), increased heart rate, more sweating (skin perspiration), including increased muscle tension which is also reflected in the vocal cords and vocal tract, affecting speech production. All of these factors may, directly or indirectly, negatively affect the quality of speech [150] and help us discriminate between stressed or neutral speech when using machine learning algorithms [3].

3.2 Compatible and Available Speech Databases in Literature

Referring to the categorization of speech data in Fig. 3.1, the bold categories are ideally the ones we need for our *fear* speech data and our application, but due to the unavailability of such open databases in literature, we tried to find the most suitable for our objectives, and we prioritized the use of real speech over acted the most. The closest alternative there is to real *fear* are datasets recorded by actors [151], [152]. These include databases with movie clips such as SAFE [153] with a focus on the emotion of *fear*. The recording of original, real, spontaneous *stress* is in turn difficult to find in literature, since there are very few datasets in which stressed speech is either simulated or recorded under real conditions. Some examples are SUSAS database [154], a collected speech data for speech recognition analysis and the design of robust algorithms to *noise* and *stress*; or UT-Scope [155] which provides automatic and perceptive estimation of Lombard speech from built-in speaker recognition – the Lombard effect is the unconscious tendency of speakers to increase the volume of their voice to improve intelligibility when speaking in a noisy environment –; or the VOCE Corpus [156] – a database in neutral and *stress* conditions of realistic, read and spontaneous speech –. Another read *stress* database we used is Biospeech [157], to which their authors granted us access.

Some of the work found for *stress* or real *fear* use proprietary databases that have not been released to the public. For *fear* we found some such as [158], where they present speech data recordings from emergencies (real urgent and fearful situations) from an emergency call center; or [159], where speech is recorded from *fear*-induced users with agoraphobia. Regarding *stress*, some realistic state-of-the-art databases are not fully available to use, such as [160] which include Russian voice recordings (words, phrases, and sentences) recorded by witnesses of in adverse events experiencing stress; or [161, 142], where they use virtual environments to induce *stress* in the participants; or three German corpora – the FAU, Ulm- and Reg-TSST – which were all collected following the well-known Trier Social Stress Test (TSST) protocol [163]. While, it is positive that there is work done and described in the field, we cannot fully benefit from it since it is not openly available for research.

3.2.1 VOCE Corpus Database

Since Bindi will be used in real-life to detect dangerous situations it is necessary to 1. work with databases containing speech in real-life conditions and 2. that those include real *fear*, *panic*, or *anxiety* feelings, which could be evoked in the type of situations to be detected in the use case. The VOCE Database was used in works [2], [3] and [10].

The first condition is relatively easy to obtain in the literature, but not the second one. As a result, we chose to select a dataset generated in real-life conditions but studying a relatively close feeling such as *stress*. Specifically, we selected the VOCE Corpus [156] because of three main reasons, 1) it includes data captured in real *stress* conditions and 2) some sensors used during the capturing stage are similar to those present in the bracelet for getting additional heart rate measurements, and 3) due to the existence of previous studies [164] confirming the feasibility of relating heart rate metrics with *stress* in speech.

VOCE [156] comprises 45-speaker's recordings in neutral and *stress* conditions of realistic, read and spontaneous speech [3]. The last updated version of this dataset includes a total of 135 voice recordings that result from a set of 45 students (21 men, 17 women and 7 unidentified) from the University of Porto, with ages between 19 and 49 years. For each user, speech was recorded on three different scenarios: *pre-baseline*, *baseline* and *recording*, which were acquired as the speaker is reading a paper 24 hours before the public speech, as the speaker reading the same paper only 30 minutes but before the public speaking setting, and in a public speaking setting where the speaker is under *stress* conditions respectively. The heart rate (HR) was also acquired every second for the three recordings.

Together with these audio files, 117 files containing 2 measured physiological variables are provided and used to estimate the Heart Rate (HR). These measurements, taken with a Zephyr HxM BT2 device, are 1. (i) Z_{ecg} representing an averaged and filtered HR value with a sampling period of 1s; and 2. (ii) Z_{ts} values that refer to the instants of time in which R peaks occur in the electrocardiogram obtained with the device, measured with an internal clock of 16 bits. Each of these values is accompanied by the Universal Time Coordinated (UTC) corresponding moment. Furthermore, the database contains a metadata file that includes gender, age, health information, experience in public speaking, STAI (State-Trait Anxiety Inventory) [165] test scores and information about the quality of the recordings (energy level, saturation...). Unfortunately, this is only provided for 38 out of the 45 individuals in the database and the database only gathers complete information (the 3 audio files and its corresponding HR values) from 21 individuals.

We divided these 21 speakers into two sets, Set 1 was composed of 10 speakers whose HR were coherent with the recordings – in the sense that, when a speaker was reading the heart rate remained stable, but on the public speaking setting the HR rose –. Set 2 was made out of the other 11 remaining speakers. In Table 3.1 the number of samples per setting are specified, each sample representing 1s audio frames.

Samples	**Neutral**	**Stressed**	**Total**
Set 1	1.389	3.989	5.378
Set 2	1.716	4.858	6.574
Total	3.105	8.847	11.952

TABLE 3.1: Number of speech utterances (samples) of the preprocessed VOCE Corpus Database [10].

Data pre-processing

For its use in this thesis, we process the speech data as well as the the HR signals. For simplicity, we begin with a conversion from stereo to mono of the audio recordings,

followed by a downsampling from 44.1kHz to 16kHz to reduce the computational cost without loosing too much precision. Then we continue by performing a z-score normalization. Finally, the signals go through a voice activity detector (VAD) [166] that removes silent audio frames as those do not include valuable information to our task. This specific VAD algorithm is designed for improving speech detection robustness in noisy environments, by removing one-second length chunks of non-speech audio where no decision about *stress* or speaker can be taken. As for the HR measures collected in the database, the original signed Z_{ecg} values were converted to unsigned ones from 0 to 255. The Z_{ts} sequences were discarded since they were considered too noisy and Z_{ecg} already provided the HR information needed with a reasonable temporal resolution

Labelling

Labelling an audio signal to determine *stress* presence is a delicate matter since there is not a prescribed way to do so given *stress* is non binary and very subjective. Taking a pragmatical perspective, once more we relied on previous work [164] where the recordings of this corpus were labeled according to each user's heart rate (HR). Instead of the labels included in the original VOCE Corpus to each recording situation (0 for the full *prebaseline* or *baseline* sequences and 1 for *recording*) we generated the labels from the HR sequences. Every 1s audio utterance is labelled as stressed or neutral using a speaker dependent HR threshold established for each of the speakers using their respective *prebaseline* recordings. Two different HR thresholds were compared: the *prebaseline* HR average plus the standard deviation and the 75% percentile of the HR value, and finally the former one was discarded.

Balancing and Data Augmentation

The fact that the data instances were not balanced – i.e., there are speakers with significantly more samples than others – led us to perform an adjustment for each set and condition to get consistent estimates. Then, all classes – in this case, speakers – need to be seen as equally important from the point of view of a speaker recognition classifier to minimize the loss accordingly in the training phase. Nevertheless, the use of a purely statistical over-sampling technique would have a big drawback in our case since the imbalance is very severe and the amount of artificial data created would be too large. To cope with this problem, we first under-sampled the set of neutral data admitting a maximum of 120 samples per speaker in both sets (1 and 2) as well as the stressed set using a threshold of 300 samples. Applying an over-sampling technique (in particular, SMOTE [167]) to the under-sampled data resulted in sufficient new samples achieving a balanced data set but without including a disproportionate amount of artificial data. Furthermore, we experimented with applying modifications in the locution speed and the pitch on the original database, to produce synthetically generated stressed samples of speech, and measure its effect with ML classifiers. This process is detailed in Sec. 4.3.1

3.2.2 Biospeech

Biospeech (BioS-DB) [157] is a multimodal public speaking database which includes continuous-time emotional annotations. It consists of 55 speakers reading two texts, one in German and one in English, while their physiological variables – Blood Volume Pulse (BVP), Skin Conductance (SC) – and speech are being recorded [8].

This database responds to the idea that performance *anxiety* can happen when speaking aloud, and can be reflected in the physiological variables and speech. Three annotators with previous training use a joystick to obtain continuous time labels for the emotional state of the speaker in a 2D space, of which their axis represent the aforementioned described in Sec. 2.3.2 arousal and valence. Biospeech was used in [8].

The aim of using these data for the thesis is twofold, 1. detecting *stress* in speech and 2. recognizing the speaker even when the speech is under *stress* conditions. For such, we perform classification with the data, rather than regression. Both regression and classification tasks make predictions about data, but the difference is that regression aims to predict continuous values, and classification predicts discrete values among a limited number of classes to which each data point belongs.

In order to create a gold standard for the emotional labels from the three individual time-continuous annotations, the authors of BioS-DB used the evaluator weighted estimation (EWE) metric [168]. The EWE is reliable when the number of annotators is rather large, but in this case we only count with 3 evaluators, which makes the possibility of disparity in the ratings very high.

The background of each annotator affects their ratings, besides the bias of the possible comparisons between consecutive speakers. These factors can induce variability and discrepancies in the ratings, and a weighted combination of the labels of each annotator may not be the optimal merging method. This was detrimental for our classification purposes, that are different from those of the creators of the dataset, which was regression.

Reinterpretation of Labels for a Classification approach

Thus, as part of this thesis, we propose a re-labelling of BioS-DB values of arousal and valence by quantizing them into 4 categorical quadrants [169]. This is crucial to define a classification task instead of using a regressor. These four quadrants are:

- High Valence, High Arousal (HVHA): $Q1$
- Low Valence, High Arousal (LVHA): $Q2$
- Low Valence, Low Arousal (LVLA): $Q3$
- High Valence, Low Arousal (HVLA): $Q4$

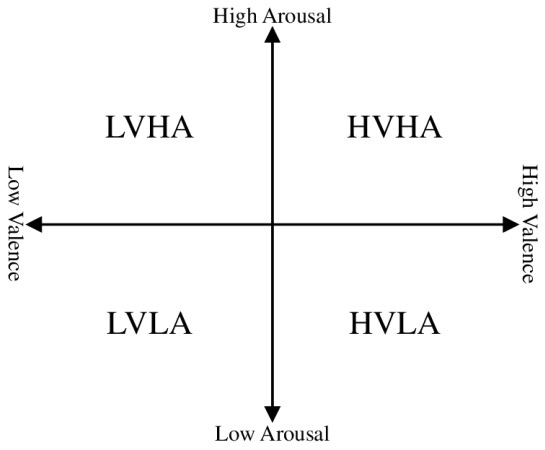


FIGURE 3.2: Four Quadrants of Valence-Arousal space [169]. Reproduced with permission from the copyright owner © 2012 IEEE.

We also believe that although BioS-DB has a very precise temporal resolution in the labelling, a coarser time resolution for capturing the underlying emotions in speech is more suitable in classification tasks such as ours. In particular, the raw annotations in BioS-DB from each annotator were originally sampled at $2Hz$ and their range was $[-1000, 1000]$. Therefore for our purposes, we downsample the signals to $1Hz$ to obtain one label per second, which will be our baseline working frequency for future data fusion schemes. To compute a combined final label for each second, we chose the two annotators that had labelled closer in the 2D space, and based on the sign of the arousal and valence values, we convert these into a categorical label in each of the four quadrants. If the quadrant where the two labels considered coincides, it is chosen as the aggregated label, otherwise, we assign a provisional undetermined value, x.

Then, we analyze several cases for the undetermined labels, as shows Fig. 3.3. If x is due to a transition between quadrants (one annotator has crossed the boundary but the other has not yet), we randomly choose any of the two quadrants. Otherwise, we consider whether two annotators fall into the same quadrant even though they are not the closest in the 2D space. If so, the aggregated label is the corresponding to that quadrant. This process solves a great amount of undetermined labels. For the rest and those cases where we found several x in a row, we used a 5-second window and replaced the unknown labels with majority voting. Our process takes into account the proximity of the labels of the raters, which provides confidence about the resulting label since the annotators interpret the 2D space in terms of the quadrants meaning.

lgorithm 1 Aggregated label value

```
.: procedure SOLVEINDETERMINACY
):     x_t ← quadrant label to determine in instant t
):     ann1_t ← quadrant label from annotator 1 in instant t
):     ann2_t ← quadrant label from annotator 2 in instant t
):     ann3_t ← quadrant label from annotator 3 in instant t
):     if not (ann1_t == ann2_t == ann3_t) then
':         (annA_t, annB_t) ← argmin(euclideanDistanceCoord2D(ann1_t, ann2_t, ann3_t))
  //   Computes the Euclidean Distance between the 2D coordinates
  for each pair or labels
):         annC_t ← the annotator left
):         if (annA_t == annB_t) then
):             return x_t ← annA_t
.:         else if (annA_t == annB_{t+1}) or (annA_{t+1} == annB_t) then
):             return x_t ← random(annA_t, annB_t)
):         else if annC_t == annA_t then
l:             return x_t ← annA_t
):         else if annC_t == annB_t then
):             return x_t ← annB_t
':         else return x_t ← majorityVoting(x_{t-2}, x_{t-1}, x_{t+1}, x_{t+2})
):
```

FIGURE 3.3: Proposed procedure to determine new combined quadrant label for Biospeech.

Transitions between quadrants are considered carefully since people do not leap from one emotional state to another suddenly. The smoothing window provides a smooth label signal by avoiding sharp changes between quadrants. Finally, for our task of automatic detection of gender-based violence situations, the second quadrant $Q2$ where emotions related to *stress*, *anxiety* and *fear* lie, will be chosen as target.

Thus, we considered two types of labellings for our tasks: quadrants and binary (considering $Q1$, $Q3$, $Q4$ as the negative label, and $Q2$ as the positive), and the result of the relabelling can be observed in Table 3.2.

	Q1 (HVHA)	**Q2 (LVHA)**	**Q3 (LVLA)**	**Q4 (HVLA)**
Original	29.22	22.56	8.53	39.67
Reinterpreted	22.16	39.04	8.56	30.24

TABLE 3.2: Percentage (%) of labels in each PAD quadrant for the relabelling of Biospeech [8]. Reproduced with permission from the copyright owner, ISCA.

3.2.3 Biospeech+

As stated in previous sections, our ultimate goal is to develop an autonomous tool to detect gender-based violence risk situations. Regarding speech and audio, we aim at tracking and identifying the user's voice and then use it to detect *fear* or *panic* – or its close relative, *stress* –. To improve the precision of the system, we aim to contextualize it – in line with understanding and interpreting the situation 2.4 – by the analysis of the acoustic scene (background sounds and noises) by using an Acoustic Event Detection and Classification (AED/C) system. BioS-DB is being used as a proxy to our problem. However, for our specific purposes it is key to complement the spoken information with knowledge about the events present in the acoustic scene: in many cases, *panic* could cause a GV victim to remain silent. That is why environmental sounds, that is, the characterization of the acoustic scene, may provide useful information for the detection system. Together with other members of UC3M4Safety team, we introduced a preliminary procedure to extend BioSpeech into Biospeech+, consisting of the original speech files synthetically enriched with environmental sounds [8].

There, we make use of AudioSet [170], a large-scale collection of human-labeled 10-second sound clips captured from YouTube. Audioset provides $2,084,320$ samples containing 527 weak annotations at clip level of sound events. We have selected a subset of $2,108$ samples from Audioset, belonging to 83 classes, to extend the original BioS-DB. To choose classes related to events that induce *fear*, we selected violent events and employed the audiovisual stimuli collection [126], [11.1], selected for the development of the WEMAC dataset [11]. The initial selection was made by experts in VG and later on validated by more than 1300 volunteers [126].

At the preprocessing stage, the audio signal is normalized, and converted into $16kHz$ mono. Then a log-mel spectrogram of 64 bins is computed to extract a time-frequency representation of the audio signal as an image.

Regarding the synthetic mixing, the process is based on the data augmentation pipeline followed in Task 4 of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 challenge [171]. Scaper [172] allows us to define probability distributions for the occurrence and duration of the sound events. Thus, the system generates as many synthetically generated mixes as desired from audio previously classified as foreground or background. In our particular case, foreground events are the original BioS-DB samples and background events are the samples of the Audioset subset. The number of generated mixes has been set to 110: we generate one mix per BioS-DB file, considering recordings captured by the lavalier microphone, i.e. 55 German and 55 English-speaking audio recordings.

```
for each lang {'de' or 'en'} do
    for file in lang_foreground_path do
        compute file duration;
        define Scaper object {sample rate = 16 kHz, n_channels = 1, set ref_db (loudness level)};
        reset previous event specifications;
        groupby: sequential Q2 labels (binary) from correspondent .csv file;
        for each Q2 group do
            define event_duration and start_time from Q2 labels;
            if binary_label == 1 then
                add background event fixing {event_duration, start_time};
            end
        end
        add foreground event fixing {file};
        synthesize defined mix;
    end
end
```

FIGURE 3.4: Procedure of generation of Biospeech+, mixing BioSpeech and Audioset samples with Scaper [8]. Reproduced with permission from the copyright owner, ISCA.

The algorithm detailing the mixing procedure, taking into account the new binarized labels explained in Sec. 3.2.2, is presented in pseudocode format in Fig. 3.4. The rationale for this methodology for the augmentation of the dataset is to provide a non-deterministic relationship between stressful or potentially frightening sounds and the appearance of *stress* in the speaker. In addition to managing probability distributions and timing of the events, Scaper allows pitch shifting and time stretching operations over foreground samples, and both could be used for further augmenting the dataset.

3.3 WEMAC: Women and Emotion Multimodal Affective Computing Database

So far we have discussed the lack of labelled databases in literature – up to the time of the work presented here – on real speech under conditions of *fear*. It seemed clear that the next step was to contribute with the collection of a database that would serve exactly our goal of detecting risk situations through voice. As we said in Sec. 1.1.4, this thesis is part of the EMPATIA-CM project that aims to develop a multimodal wearable device for automatic and inconspicuous detection of these situations, so the databases collected and explained below are also multimodal.

WEMAC is a multimodal dataset consisting of laboratory experiments on female volunteers exposed to audiovisual stimuli validated to evoke real emotions using virtual reality headsets, capturing and collecting physiological, vocal, and self-reported state variables to which I have contributed jointly with others. For its collection we have used the validated audiovisual stimuli collection, part of the UC3M4Safety Audiovisual Stimuli Dataset but its creation is due to the other members of the UC3M4Safety team. It arises from the need to elicit realistic emotions, especially *fear*, which is key for the detection of GBV risk situations. We believe this database will serve and assist research on multimodal Affective Computing using physiological and speech information, and it to be specially effective for the task of GBV risk situations detection.

3.3.1 UC3M4Safety Audiovisual Stimuli Database

Other members of the UC3M4Safety team conducted the [126] study to obtain a comprehensive, high-quality dataset of audiovisual stimuli to elicit emotions in controlled scenarios. This dataset is designed to collect additional human responses (physiological variables and speech) that can be used by AI ML/DL systems aimed at automatic and real-time emotion identification. Although the primary goal is to recognize *fear* or *panic*, we use carefully curated video clips and a comprehensive 12 emotion range tagging system to categorize the range of emotions people experience.

The paper presents the identification of emotions elicited after the visualization of the audiovisual stimuli collected. In addition, the authors conducted a statistical study of gender differences in emotional responses on 1,332 volunteers (811 women and 521 men). The research study produced a dataset of 42 audiovisual stimuli – referred to as the UC3M4Safety Audiovisual Stimuli Database [11.1] [126] – that triggers a range of 12 emotions in the viewers. Each stimuli has a high level of agreement and one discrete emotional categorization, as well as a continuous emotional categorization in the Pleasure-Arousal-Dominance (PAD) Affective Space.

The selection of the series of audiovisual stimuli was performed in five steps, as shown in Fig. 3.5. Each blue coloured box reflects the step-by-step process and criteria used for the clip selection, while the white boxes denote the supervisors involved in the process.

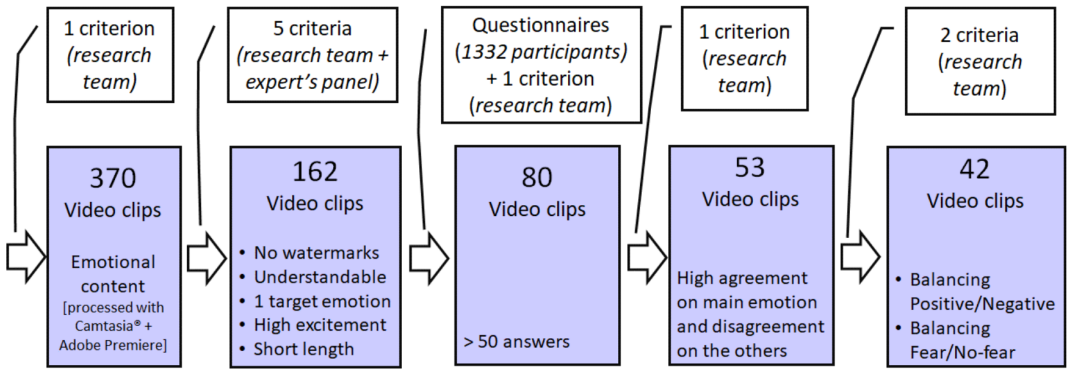


FIGURE 3.5: Video clips' processing in the creation of the UC3M Audiovisual Stimuli Database. Reproduced with permission from the copyright owner, the authors of [126] via Creative Commons License CC-BY 4.0 from MDPI.

Initially, five researchers collected samples of emotional content from commercial films, TV series, documentaries, short films, advertisements and Internet videos. These clips were originally tagged with a target emotion by other members of the UC3M4Safety team, with advice from a panel of experts. The discrete emotions contained in the audiovisual stimuli sought by the researchers were *joy, sadness, surprise, contempt, hope, fear, attraction, disgust, tenderness, anger, calm* and *tedium*.

Secondly, from the 370 samples obtained in Step 1, 162 clips were selected for further evaluation based on selection criteria (see Step 2, Fig. 3.5). An additional criterion was considered for films about gender-based violence: the protagonist in the films must be a woman who is the victim of some sort of violence (sexual, physical, psychological, etc.).

Thirdly, the list of 162 stimuli was labelled with discrete emotion categories in a crowd-sourcing setting by a large set of volunteers. Every clip is labelled with

the experimented emotion after its visualization. The reported emotions by the volunteers are not always equal to the ones originally expected the clip to elicit by the team and the experts. Only 80 video clips obtained enough answers to be considered for further analysis (each had more than 50 visualizations and therefore emotional labels).

Two conditions had to be met for the final selection of audio-visual stimuli. The first condition sought the highest percentage of consent among participants, meaning at least 50% of the volunteers considering both genders or at least 50% of one gender individually, who visualized each stimulus, labelled it with the same categorical emotion. The second condition also tested the uniqueness of this label by checking that all other possible emotions matched at most 30% of the time. Finally, some videos were removed to evenly balance the distribution between the target emotions considered as *fear* and *non-fear*, producing a selection of 42 clips. They obtained a percentage of 44.44% for *fear* and 55.55% for the rest of emotions, as shown in Table 3.3.

Fear	44.44%	**Tedium**	2.22%
Joy	8.89%	**Tenderness**	6.67%
Hope	2.22%	**Calm**	11.11%
Surprise	4.44%	**Disgust**	8.89%
Anger	4.44%	**Sadness**	6.67%

TABLE 3.3: Percentages of categorical emotions elicited by the UC3M4Safety Audiovisual Stimuli Dataset for the final sample of 42 clips [126].

Gender Differences for Emotional Annotations

The results obtained by the team in [126] show similar reported positive emotions in discrete values (and also in the PAD space) for both genders, while negative emotions (especially *fear* and *contempt*) reports are more different. Autobiographical memory may influence the perception of *fear* in those video clips related to gender-based violence. The gender-based violence is hard to label, and the fact that the viewer identifies herself with the main characters in the video clip may be having a big impact in her emotional state. In the clips were women mostly label *fear*, men label *anger* and *sadness*.

Proper labeling of women's *fear* will benefit our main goal of developing an automatic system to protect them from violent or sexual assaults. Considering the observed results, the gender variable should be considered both in the stimulus selection phase of the database and in the training phase of the machine learning algorithms. Although in the study there were no significant differences between some emotions and others (especially *fear* and *hope*), gender differences in reported emotions should be considered to improve emotion classification. This is particularly important in this work, because the main aim is to identify the conditions that cause *fear* to women, including victims of gender-based violence. In this case, gender must be taken into account because the emotion perceived for *fear* video stimuli differs by gender. Even key for the particular case of stimuli reproducing situations of gender-based violence, where there is a big difference in the labelling between women and men (*fear* versus *anger* and *sadness*).

These results corroborate those of other studies [173], where authors conclude that women reported more *fear* than men; and *sadness*, *compassion* and *fear* emotions

are felt more by women than men, which could be due to stronger empathetic traits and *care-taker syndrome* mostly occurring in women.

3.3.2 WEMAC Database Collection

In the database called WEMAC [11], we – the UC3M4Safety team – use a virtual reality environment – a headset and a joystick – to present immersive audio-visual stimuli (i.e., video clips) to women to elicit, label and measure realistic emotional reactions to them.

The participants are women volunteers, including women that suffered from GBV – so GBVV and non-GBVV volunteers –. They are divided into balanced age groups defined by 10-year intervals: G1 ($18-24$), G2 ($25-34$), G3 ($35-44$), G4 ($45-54$), and G5 (55 on-wards). The database consists of 104 women volunteers who never suffered from GBV (47 on the first release and 57, on the second) and 43 women GBVV volunteers. The latter group performed the experiment under the supervision of a psychologist. Fig. 3.6 shows a simplified diagram of the specific methodology followed during the experimentation for every volunteer and stimulus.

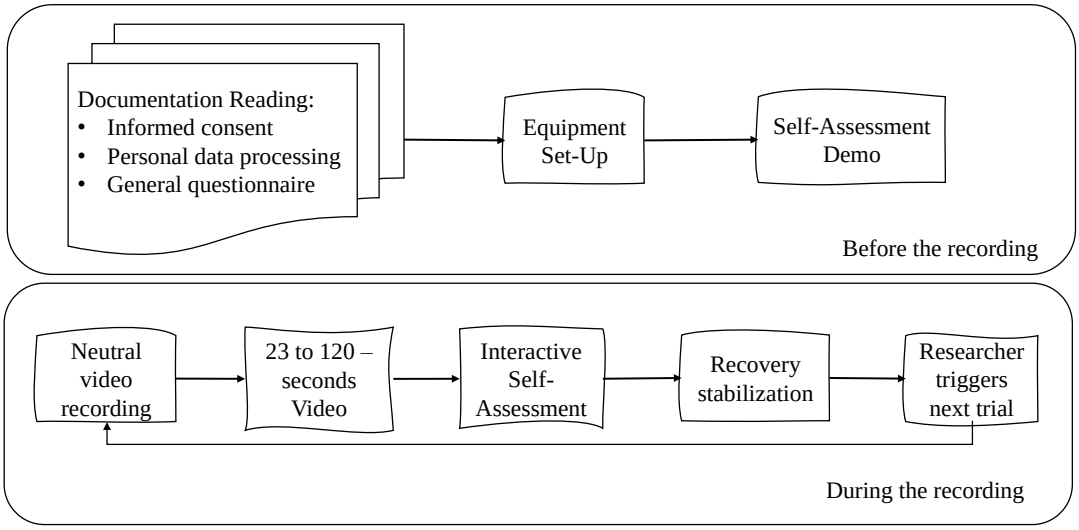


FIGURE 3.6: Experimental methodology followed during the development of the WEMAC dataset, prior and during the visualizations [11].

Volunteers were recruited through social media advertisements and internal communication channels within the University Carlos III. Prior to the experiment, all the different phases to be followed are explained to the recruited volunteers, including a set of documents with an informed consent and an initial generic questionnaire. The former is necessary for personal data processing and protection regulation. The latter collects information such as personality traits, gender, age, whether they performed any recent physical activity or if they were taking medication – the use of medication might modify the physiological response of the user –, self-identified emotional burdens due to work, economic and personal situations, mood bias (fears, phobias, traumatic experiences), among others. This information could be relevant and informative of the emotional reactions of the users captured in the experiment, affecting their perception, evaluation and attention.

In this database collection, the UC3M4Safety team followed the methodology also presented in [174], which is a study for the detection of fear using the

concentration of the hormone catecholamine in blood. The last step in the experiment preparation is an introductory demonstration where the volunteers get used to the virtual reality environment – headset and joystick – and get familiar with the labelling particularities. This environment is used to present the clips to the users, and also to annotate them according to different categories through interactive screens.

The whole process of documentation reading, equipment set-up, virtual environment demo, together with the visualization and labelling of the videos, usually takes from 60 to 100 minutes per participant.

Audiovisual Stimuli Visualization

We used an Oculus Rift-S virtual reality Headset[21] to present the audio-visual stimuli. Virtual reality is used to maximize the immersive experience and consequently, achieve a better emotion elicitation. During the recording experiment, every volunteer visualises a total of 14 audio-visual emotion-related stimuli, some of them presenting a 360° experience. These stimuli were selected from a 28 audio-visual stimuli pool resulting in two batches of 14 videos each, from the final 42 video clips of the UC3M4Safety Audiovisual Stimuli [126], as seen in Fig. 3.7. The criteria applied for the selection were the following: 1) the highest emotional discrete agreement observed in female raters, 2) an adequate laboratory experiment duration and 3) a balanced distribution of *fear/no-fear* clips in each batch [53].

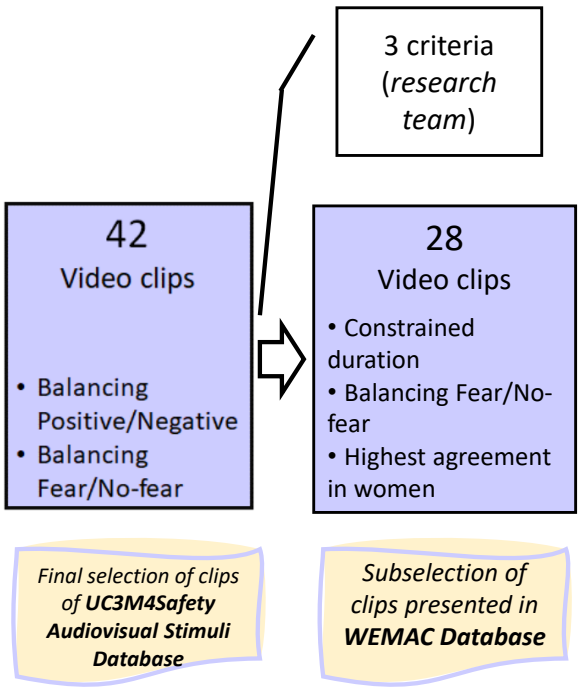


FIGURE 3.7: Schematic of subselection of clips from UC3M4Safety Audiovisual Stimuli dataset used in WEMAC database.

The stimuli average time length is 100 seconds. Both batches have 8 stimuli belonging to the second arousal-valence model of quadrants to maintain a proper

[21] https://www.oculus.com/rift-s/

balance between *fear* and *non-fear* emotions. Note that the balance premise is assessed considering the valence-arousal – or pleasure-arousal, PA – model, rather than the pleasure-arousal-dominance (PAD) space, for simplicity. Due to this fact, stimuli prelabelled as *anger* or *fear* are considered within the second quadrant, then being within the positive class for the binary ground truth labelling.

The 28 audio-visual stimuli were selected based on three main premises: the highest emotional discrete labelling agreement observed in women during the prelabelling experiment [126], targeting for an adequate laboratory experiment duration, and a balanced distribution of *fear* vs. *no-fear* categories and within the four quadrants in the arousal-valence space.

Before the presentation of each of the stimuli, a neutral video clip is displayed to set the participant in a neutral emotional state. These neutral video clips have been selected from the large pool provided by the Stanford Psycho-Physiology Laboratory [175]. Similarly, 3D recovery scenes are also shown to the volunteers after the interactive emotion labelling process. These 3D scenes were selected by unanimous consensus of the research team. The main difference between the neutral and recovery clips is that while during the display of the former no action is taken – i.e. there is no recovery monitoring –, for the latter there is a physiological monitoring through Bindi bracelet to ensure the volunteer's physiological stabilisation.

Physiological Signals Captured

During the audio-visual stimuli presentation, the physiological signals of the participants are captured[22]. The equipment used for this purpose includes the following devices and sensors:

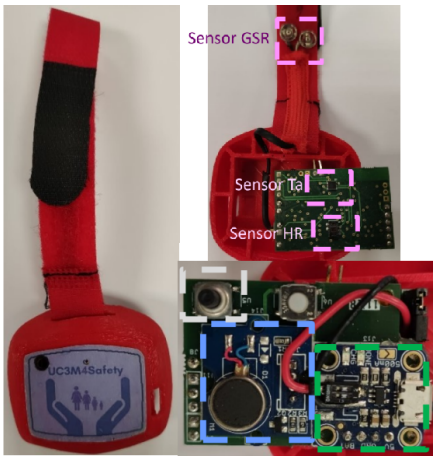
- The BioSignalPlux[23] research toolkit system, which is commonly used to acquire different physiological signals, particularly: finger Blood Volume Pulse (BVP), ventral wrist Galvanic Skin Response (GSR), forearm Skin Temperature (SKT), trapezoidal Electromyography (EMG), chest respiration, and wrist inertial movement through an accelerometer.
- The Bindi bracelet, represented in Fig. 3.8a, that measures dorsal wrist BVP, ventral wrist GSR, and SKT. The hardware and software particularities of this element are detailed in previous publications from the team [176], [177], [178].
- An additional GSR sensor to be integrated in the next version of the Bindi bracelet. Its hardware and software particularities are detailed in [179].

Note that the BioSignalPlux toolkit is employed to provide gold standard measures to be compared with the sensors included in the Bindi bracelet. In fact, BVP and GSR signals obtained from BioSignalPlux and Bindi were successfully compared and correlated in [176] and [178]. The acquisition synchronisation of all sensors together with the stages of the experiment runs on a laptop through a program based on the Unity framework[24]. In this sense, the sampling frequency of the devices sensing physiological information is 200 Hz.

Labelling Process: Speech Signals and Self-annotations

After every emotional video clip visualization, volunteers find a set of interactive screens within the virtual reality environment, developed with Unity software [180].

[22] Processing available in: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/
[23] https://biosignalsplux.com/products/kits/researcher.html
[24] https://unity.com/es



(A) Bindi 1.0 bracelet for the capture of physiological signals

(B) Bindi 1.0 pendant for the capture of auditory signals

FIGURE 3.8: Bindi 1.0 wearable devices. Reproduced with permission of the copyright owner UC3M4Safety team.

On these screens, volunteers label their emotional reactions. The annotation is done in the following order:

1. Two questions are presented to the volunteers about the video stimuli right after its visualisation – with the intention of capturing at least a speech signal of one minute's duration –, as a result, an audio signal containing the *user's emotional speech* is captured by the Oculus Rift S Headset embedded microphone. These questions were designed to make the volunteers relive the emotions felt during the video visualization, aiming to capture the last traces of emotion in their voice. Table 3.4 presents the set of questions.
2. Modified Self-Assessment Manikins (SAM) are used to annotate the values of *Valence/Pleasure, Arousal, and Dominance* by a 9-point Likert scale. Such modified SAMs appear in Fig. 3.9, and the process of redesign and assessment is detailed in [181].
3. *Familiarity* With the emotion felt and the situation displayed in the video-clip is also annotated. Both are answered using the same 9-point Likert scale as for the SAMs.
4. *Liking* of the video is annotated through a binary yes-no question.
5. Selection of one *discrete emotion* out of a total of 12, already described in Sec. 3.3.1 [126].

First question	**Second Question** *"Close your eyes and think about the situation you watched..."*
Describe what has just happened in your own words	What details can you describe?
Explain the situation you have seen in your own words	What details do you remember?
Describe what you saw in your own words	What struck you the most?
Describe what you heard in your own words	What happened at the beginning?
Describe where and when the situation happened	What would you have done if you had been there?
	What would you have done if you had been in that situation?

TABLE 3.4: Questions asked in the annotation phase of WEMAC. Two questions were asked to each participant, randomly chosen after each video visualization. These questions were originally in Spanish.

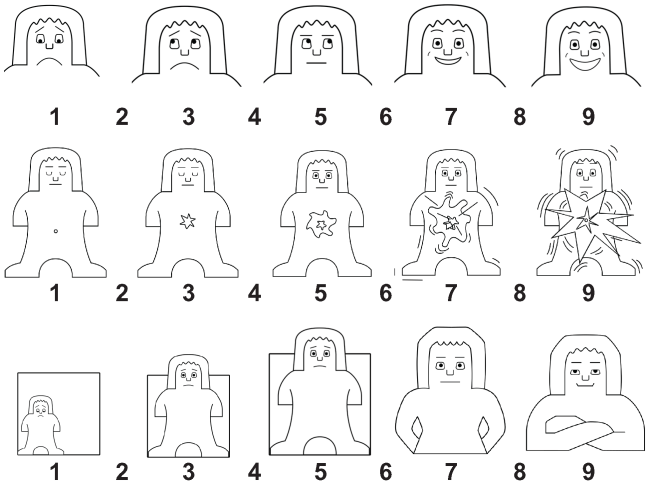


FIGURE 3.9: Modified SAM by the UC3M4Safety team [11]. Reproduced with permission from the copyright owner, the authors of [181] via Creative Commons License CC-BY 4.0 from Frontiers.

Audio Features and Embeddings Extraction

Since we cannot release the raw speech signals due to ethics and privacy issues – as this could identify the users and relate them to whether or not they have suffered gender-based violence –, we have processed the speech signals and extracted low- and high-level features so that the research community can analyze and work with them[25].

We use different Python toolkits to extract information at a window size of *1 second* and a hop size of *1 second* per audio file. We follow a similar approach to the one followed in the MuSe Challenge 2021 [182] for the feature and embeddings extraction of the audio signals. That is:

1. *librosa* [183]: we extract the mean and the standard deviation of a collection of features computed at a window size of 20*ms* and a hop size of 10*ms* through the librosa toolkit. The 38 features extracted are 13 Mel-Frequency Cepstral Coefficients (MFCC), Root-Mean-Square (RMS) or Energy, Zero Crossing Rate, Spectral Centroid, Spectral Roll-off, Spectral Flatness, and Pitch.
2. *eGeMAPS* [184]: we compute 88 features related to speech and audio through the openSMILE Python toolkit [185] on its default configuration, i.e., a window size of 25*ms* and a hop size of 10*ms*.
3. *ComParE*: we extract the 6,373 features used in the ComParE 2016 challenge [186] by using the openSMILE Python toolkit.
4. *DeepSpectrum* [187]: we extract 6,144-dimensional embeddings by this toolkit for the extraction of audio embeddings based on different deep neural network architectures (DNN) trained with ImageNet [188]. Specifically, two different configurations were considered, ResNet50 network and the output of the last Average Pooling layer (*avg_pool*), resulting in 2,048-dimensional embeddings, and VGG-19 net and the last Fully Connected layer (*fc2*), resulting in 4,096-dimensional embeddings.
5. *VGGish*: we extract 128-dimensional embeddings from the output layer of the VGG-19 network trained for AudioSet [170].

[25] Available in: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing

We published – Table 3.5 – the first release of WEMAC database with the aim of sharing it with the research community, encouraging the improvement of the baseline results – presented in [1] – through the use of fusion methods, attention models, transfer learning, semi-supervised or self-learning strategies or any other that the research community finds adequate; and advancing in the research of multimodal emotion analysis in general, and in gender-based equality in particular.

Database	Datasets	Conditions	Participants
UC3M4Safety Database [14]	Audiovisual Stimuli: Videos [11.2]	Crowdsourcing	General public and expert judges
	Audiovisual Stimuli: Emotional Ratings [11.1]		
	WEMAC: Biopsychosocial Questionnaire [11.3]	Laboratory	GBVV and Non-GBVV
	WEMAC: Physiological Signals [11.4]		
	WEMAC: Audio Features [11.5]		
	WEMAC: Self-reported Emotional Annotations [11.6]		

TABLE 3.5: Hierarchy, subdivisions and references of the UC3M4Safety Database datasets [126] [11].

3.4 Women and Emotion in Real Life Affective Computing Dataset: WE-LIVE

WEMAC is a laboratory database for detecting realistic emotions from a multimodal point of view in women, but it is still far from real-life conditions. Recording emotional speech in fearful conditions that is realistic and spontaneous is very difficult, if not impossible. To get as close as possible to these conditions and perhaps record fearful speech, the UC3M4Safety team created the "Women and Emotion in real LIfe affectiVE computing dataset" (WE-LIVE).

The objective with WE-LIVE is to collect physiological, physical and contextual signals from women in a relevant and uncontrolled environment, as well as labelling of their emotional reactions to everyday events in their lives, using the current Bindi system (wristband, pendant, mobile application and server). Through Bluetooth® connection to the mobile phone, the data captured by Bindi is sent to a protected and encrypted server. Relevant environment is understood as everyday activity within their usual routines. The devices will only perform data collection and the signal acquisition will be performed simultaneously: physiological, geolocation, audio and speech signals are temporally contextualised.

The database is composed of 13 women volunteers, including GBVV. Some of them also participated in the collection of WEMAC. As with the latter, volunteers were recruited through social media advertisements and outreach to students and researchers at the university.

First, to record the database, the information guide is explained to each participant including what the experiment will consist of. During that session the devices are given to the user and the phone application is installed in her smartphone. Their functionality and daily use is also thoroughly explained, together with certain particularities associated with them. The volunteers are also asked to fill a routine questionnaire in to collect their routine activities, to be then classified by a specialized psychologist according to activity relevance. At any time during the experiment, the volunteer can decide not to continue with it.

3.4.1 Data Captured

Each volunteers is given the devices for 7 days, extendable to 10 days, if they wish. There is a specific person – from the technical staff in UC3M4Safety team – responsible for each volunteer with whom she can contact anytime. The user lives a normal life and the devices capture relevant data. The pendant and wristband devices from Bindi capture GSR, BVP, SKT, audio, geo-location and also accelerometers, in different patterns depending on the routine the user is in at any given moment. Different patterns include longer recording during the most relevant activity slots of the day and briefly during the time slots with little activity. In Fig. 3.10 we represent the 2.0 version of Bindi wearable devices.

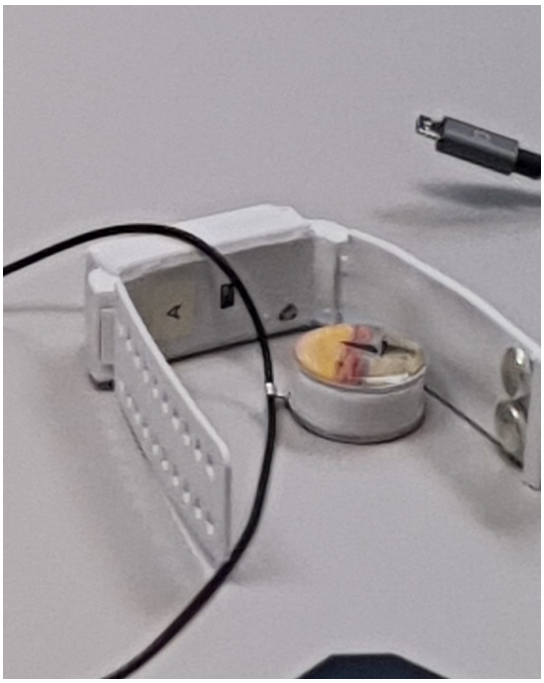


FIGURE 3.10: Bindi 2.0 wearable devices. Reproduced with permission of the copyright owner UC3M4Safety team.

The speech part is the most relevant for this thesis, and it is recorded by the microphone located in the pendant. The pendant includes a MP34DT06 omnidirectional microphone[26]. According to Spanish regulations [27] the consent of third parties is not required when the purpose of the data processing is to protect a vital interest of the data subject. Furthermore, the captured audio is not intended to be heard by any person, only analysed by ML algorithms, so the privacy of the users remains intact.

The smartphone app allows the user to control the connection of the devices with the user's mobile phone, the labelling of situations, and the deactivation of the devices. This app also allows the monitoring of the user by the technical team. There is also a *sleep* mode that allows the user to deactivate the devices at will in case she wants to deactivate them for a period of time, commonly done during the night when the user is sleeping and the devices are charging. It also has a manual *sports*

[26] https://www.mouser.es/datasheet/2/389/mp34dt06j-1387393.pdf

[27] Artículo 6.2 Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal https://www.boe.es/buscar/pdf/1999/BOE-A-1999-23750-consolidado.pdf

mode, which labels the time period while it is activated as high physical activity, due to the user doing sports.

3.4.2 Labelling

Data captured gives us very insightful information about the user's activity, but we also need emotional labelling on her side so that our ML models can work with the data and make inferences and predictions.

Users are asked to label when an emotion is happening, defined for them as *"a brief, intense reaction to a particular stimulus (internal, e.g., memory; or external, e.g., the sound of a crash), which results in bodily changes (pulse, muscle tension, facial expression, etc.) and influences our behaviour and thinking."* The volunteers are requested to characterise and define the emotional events that have occurred in a given period, as well as the context of these events through the participant's mobile app labelling screen. To facilitate the users' timely annotation, there are two **labelling modes**:

- **Triggered by prompts**: This type of labelling is prompted periodically after each of the corresponding routine activity slots and reminds the user that it is convenient for her to tag the emotions she is experiencing at each at least once.
- **On-demand**: This type of tagging is be activated when the user requires it through the smartphone app directly.

Additionally, the user is asked to complete some other labelling fields for each emotional label with the goal of making them more informative, accurate and provide as much context as possible:

1. **Arousal:** In a calm-activation 9-point scale using the modified SAM [181].
2. **Valence:** In a negativity-positivity 9-point scale.
3. **Dominance:** In a dominance-submissiveness 9-point scale.
4. **Category:** Several tags of discrete emotion categories to select one among *boredom, disgust, joy, calm, attraction, surprise, hope, contempt, gratitude, anger, fear* and *sadness*.
5. **Emotion intensity:** From low to high on a 9-point scale.
6. **Boolean experience:** A boolean switch to mark if the user thinks that the emotional reaction is related to traumatic and/or shocking experiences from the past.
7. **Context:** Discrete categories from which to select as many as the user wishes, that bring contextual information. Some are: *home, school, work, gym, restaurant, hospital, transport, coffee, party, drinks.*
8. **Audio note:** This is an optional response, where the user can record and audio signal to describe the situation, the stimulus or reasons that triggered the emotional reaction, feelings, what they attribute it to, etc.

Once all fields are complete, the emotional label is saved in the system. For each of the labels or annotations created in the app, there are 3 types of **timestamps** associated to it:

- **Created**: Manual or automatic, every time the time slot of routines changes, or right when the user creates the label.
- **Happened**: Manual, the moment the user selects the stimuli that elicited the emotion that happened.
- **Sent**: Manual, denotes when the label is complete and the users click on 'Send' or 'Save'.

During the 7 to 10 days of the experiment, a follow-up telephone call from the technical team to the user was done twice to check that everything was working

correctly and to collect feedback on how comfortable the user felt using the devices and the app, providing factors such as ease of use.

All personal data as well as the data recorded during the experiment are anonymised. That means that we generate a code for each of them so that it is impossible to associate the collected data with the person who volunteered for the experiment. The anonymised data will be kept for a maximum of 5 years. We store them in a safe encrypted server.

3.5 Conclusions, Insights and Improvements

In this chapter we have explained which databases were available in the literature that more closely meet our needs. We have detailed the modifications we made to them, and explained our purposes to do so. We have also made clear that for our specific purposes – training a GBV risk situations detector – this was not enough. Thus, we have created new datasets accordingly, describing the effort that it has taken and illustrating why we think this is an important contribution since they are a first step towards our goal. In this section we provide some brief take-home messages learnt from the databases collection and the way we are going to continue with the use of these databases for the development of better ML models for GBV risk situations detection.

The VOCE corpus database served us as a first approach to evaluate the differences between stressed and neutral speech, as well as to observe the relationship between speech and heart rate changes. Along the same research lines, we also used Biospeech. The use of the latter allowed us to change from a binary stressed setting to a more specific emotional state that included 4 arousal-valence combinations. These databases also made it possible to work towards enhancing speaker recognition thanks to the diversity of speakers available. Besides, the reinterpretation of the Biospeech labels and the addition of acoustic events resulting in Biospeech+ enabled us to study the influence of such events in the detection of *stress* in speech.

As we have previously stated, we are aware that stressed speech is not fearful speech, but the use of the former has helped us to analyze speech in similar emotional conditions, its characteristics and to look for appropriate models for our challenges and needs, while pursuing the goal of GBV risk situations detection.

Some of the challenges involved in creating databases for our specific application have already been described in Sec. 1.2. In our case, we collected two databases only with female users' data, without a need to perform a gender balance due to the use it will be given in the field of gender equality, but the data was actually balanced in terms of age: there were 5 age balanced groups, in both WEMAC and WE-LIVE. Regarding cultural background, all the participants were residents in Spain, the country in which Bindi is going to be put into practice and for which the recorded data would be used.

When it comes to the labelling process, we explained thoroughly the annotation process in WEMAC as well as WE-LIVE in order to get the most reliable and accurate labels as possible, but it is indeed possible that background bias is present in the data, both from the crowdsourced annotators in the Audiovisual Stimuli dataset, and in WEMAC and WE-LIVE. This is why we should handle these labels with care, especially those that only have self-annotations. Instead of using them as ground-truth labels, we could study new ways to use them in the future, for instance, taking into account other aspects of the background of each user that may influence

the labelling, or using an aggregate emotion label to have a more reliable label for each video in the case of WEMAC.

Focusing on speech, we expect the captured speech in WEMAC to have traces of the emotion in it, but we cannot assert this as the recording is performed right after the emotion eliciting stimuli visualization. During the database collection, we have observed that some users took a short break between viewing the stimuli and capturing the speech signal, to recover from the stimuli they had seen – which is normal as there were videos that were intended to cause real *fear*. However, this means that the emotional content of the speech signals can be variable, and that is something we also need to keep in mind when working with such data.

As for the WE-LIVE database, the recording of the database was completed in July 2022, and the processing and analysis of the captured data for further use is still a work in progress. It is necessary to do a thorough work in this regard, first of all with data cleaning, since the devices have recorded a huge amount of data that not always contains relevant information for the task; then regarding coordination and alignment of data, since it is necessary to synchronise the different modalities and solve the problem of missing data in each particular modality. Also in the case of WE-LIVE, we acknowledged some problems during the time of the capture, most failures were due to disconnections of the devices with the smartphone – via Bluetooth (BT) –, and due to malfunctioning of the bracelet when sweat came in. This is being addressed by the team and an enhanced version in the next Bindi prototype expected to be designed: Bindi 3.0.

Chapter 4

Speaker Recognition under Variability Conditions

For our goal of detecting GBV risk situations through speech, the first step to be taken seems to be to detect the voice of the specific user we are interested in, from among all the information contained in the audio signal. To consistently detect emotion, especially *fear*, in a user's speech, the user's voice must first be isolated from other speakers in an acoustic scenario. This practice also opens up interesting opportunities in situations where all speakers in a scene need to be identified, for example in forensic evidence. Thus, we dedicate this chapter to the research on the field of Speaker Recognition (SR) because after we detect the user we can make an analysis of the emotions they experience.

This field is slightly elusive in our case, because we need to detect the user's voice in order to identify them, but the performance of ML models for detecting speakers through the voice drops when they are under emotional conditions. So the fact that the voice of a GBVV could be influenced by her emotional state constitutes a challenge for a speaker identification system. We are interested in achieving good performances for the recognition of the speaker even when the voice expresses stress or fear.

Then, the contributions of this chapter rely on our study of speaker recognition systems under variability conditions, 1) speaker identification under stress conditions, to see how much these stress conditions affect the SR systems[28] and 2) speaker recognition under real-life noisy conditions, which is where our application will ultimately work, isolating the speaker's voice, among all additional environmental noises.

4.1 Introduction

Speaker Recognition (SR) refers to the task of the automatic detection of a person from the characteristics of their voices, also known as voice biometrics [189]. On it, we can distinguish two subtasks, Speaker Identification (SI) and Speaker Verification (SV). The first refers to the recognition of a particular user among a known number of users (a multi-class setting), and the second aims at identifying one user versus the rest (binary setting). It is on speaker identification where we focus along this chapter, in the ability to detect to whom the voice belongs, even under emotional conditions. The effects of emotions in SI [190] have been studied in the literature but research is scarce about the influence of stress specifically. The technological development

[28] In the absence of databases of speech in conditions of realistic fear in literature – which is explained in detail in Sec. 3.1 – in the moment this part of this thesis was conducted

of Speech Emotion Recognition (SER) is plenty, but the task of SR under emotional conditions it is still an early stage scientific field.

As for the SI system to be designed, it should be adapted to what we expect Bindi to find in a real world situation: the goal of our system is to detect the users speaking even when their voices present fearful or stressed conditions. For this reason, we could be facing a mismatch learning problem in which we may only have neutral speech utterances available for training, gathered in an initial Bindi setup – given that the possibility of forcing the user to speak while under fear or stressed conditions is difficult – whereas the real environment operating conditions would contain both neutral and stressed or fearful samples together.

More technically, in this chapter we study and design robust Speaker Identification systems to the variability conditions that could be induced by a microphone embedded in a wearable device working in a real environment, such as emotions – stress and fear, particularly – and environmental noise conditions in speech. We track these problems by means of different techniques such as data augmentation (DA) or synthetic data generation, that takes into account Bindi's computational restrictions and audio input characteristics.

4.2 Related Work

We have already mentioned the difficulties faced when looking for suitable databases to develop machine learning models appropriate for our task in Sec. 3.1. There are very few databases in which stressed speech is recorded under real conditions – and there is none in the case of real fear –, in addition to the challenge of labelling process.

Handcrafted Features

Various strategies of hand-crafted or hand-selected features are used for speech-related applications in the literature., [191], [192], [193]. Speaker Identification systems work with speech signals and try to use acoustic features that differ between individuals to discriminate among them. Some of the features that exhibit good performance when used in neutral or emotionless conditions in speech-based systems are Mel-Frequency Cepstral Coefficients (MFCC) [194]. The MFCCs model the human auditory system to capture the phonetically important speech features, by distributing frequency mel-filters almost linearly at low frequencies and logarithmically at high frequencies. Although many of coefficients can be calculated, usually the first 12 or 13 coefficients are usually calculated. Prosodic features are widely used as well.

Prosodic features are suprasegmental characteristics that appear when sounds are put together in connected speech. Some of the domains or phenomena for which features are implemented are intonation, stressed syllables and rhythm. Besides, phonetic features are also used. They also model acoustics by capturing pronunciation variation between speakers and tessellating the acoustic space, enabling the modeling of longer-term patterns such as detection of the phonemes and their statistics. Along the same line, Linear Prediction (LP) is used in audio and speech processing for defining the so-called spectral envelope of speech signals in compressed form, using the optimized coefficients of a linear predictive model. It is also a powerful speech analysis technique to provide estimates of speech parameters like pitch, duration and energy [195].

Although for speaker identification under stress conditions there is hardly any previous work, MFCCs [196] together with prosodic features as the pitch, energy and word duration [197] are used and achieve good performances[198]. However, their generalization abilities and robustness against variability are limited.

Automatic Features

Beyond the previously mentioned hand-crafted features, learned features (automatically obtained from the algorithms) extracted from raw data by DNN (Deep Neural Networks) is a novel trend achieving very innovative results [199], [200]. In the last decade, it has been found that when sufficient data is available, automatically learnt feature representations or DNN-based embeddings are usually more effective than hand-crafted or manually designed features, allowing to develop better and faster predictive models [201], [202]. Most importantly, automatically learnt feature representations are usually more flexible and powerful. Representation learning consists in yielding abstract and useful features usually from the signal waveform directly or from relatively sophisticated low-dimensional representations, by using autoencoders and other deep learning architectures often generalizing better to unseen data [203], [204]. Nevertheless, in our case, the use of complex DNN approaches should be handled with care due to their high computational load, delay and the availability of sufficiently big training datasets, which are three very important limitations within the Bindi system, detailed in Sec. 1.2.2. Besides, automatic features can be hard to understand or interpret by humans, having low "explainability", which is detrimental for the "transparency" of the machine learning models.

Due to the sequential nature of speech signals, their temporal context is of great relevance for classification and prediction tasks [205]. Besides, the sequential character of its frequency contents carries very relevant information of speech [206]. Recurrent Neural Networks (RNN) are powerful tools to model sequential data [207], having become the state of the art due to their improved performance and generalization capabilities. However, the availability of larger databases is, again, of paramount importance for training such networks. Unfortunately, this is not the case for data of real stressing situations in particular, such as the ones we are facing.

Deep neural networks are even able to condense efficiently the information related to the identity of the speaker, being able to exclude the rest of irrelevant info for tasks of SR, into what are called speaker embeddings. In a wide sense, all of the neural embeddings which include some form of global temporal pooling and are trained to identify the speakers in a set of training recordings are unified under the term *x-vectors* according to [208], [209]. Variants of *x-vector* systems are characterized by different encoder architectures; pooling methods and training objectives [210] and in this sense all of the embeddings tested in this section could be consider such.

Data Augmentation

DA is a key ingredient of state of the art speech technologies as it is a common strategy adopted to increase the amount of training data. It can also act as a regularizer to prevent overfitting [211] and to improve the performance in imbalanced class problems [167]. This makes the whole process more robust achieving a better performance. Due to the scarcity of data we mentioned in Sec. 3.1 and even if it is not fully realistic, this is a good match for our case as the databases we use can be quite small and may have data imbalances. By using DA, we can

increase the amount of data available and deal with the lack of balance between classes [212], approaching how the system would work in a real case for which we would not yet have data.

Together with the wealth of research to cope with the widespread problem of variability of speech signals, DA is a widely applied technique to enlarge databases [213], for instance, by adding noise or applying transformations to the speech signals, similarly to the ones introduced by transmission channels.

Speech enhancement techniques are also used to improve the overall perceptual quality of speech, specifically intelligibility [214], [215], [216]. Remarkably, these techniques can be modified towards a speaker recognition objective, instead of audio quality [210].

Additionally, in order to alleviate the intrinsic variation mismatch and specifically the one caused by emotions for the tasks of speaker identification, literature considers several solutions, such as eliciting emotions in speakers in a way that accomplishes similar effects as spontaneous [217] due to the difficulties of recording authentic emotions – both in terms of privacy and labelling –. Likewise, statistical estimations and domain adaptation methods are also used [218], [219], [220]. This lack of datasets containing real and natural – not acted – negative emotions in speech, as the ones a user could experience in a risk or violent situation, is indeed a challenge.

Classification Models

Regarding models, algorithms such as Gaussian Mixture Models (GMM) were generally employed for speaker recognition [221] and Support Vector Machines (SVM) are also widely applied [222],[223]. Other studies suggest the use of DNN for speaker recognition [224], [225] The improved accuracy achieved by DNNs, as compared to other state-of-the-art systems, is the result of their ability to extract discriminating representations from data that are robust to the variability particularly in speech signals. In recent years Deep Learning algorithms have skyrocketed in many scientific fields thanks to the availability of large amounts of data.

However in the research conducted in this chapter, we aim to keep a balance between computational complexity and accuracy due to the hardware constraints of the device, where the battery consumption is critical – that our targeted device hardware imposes – and the scarcity of training data originally available (see challenges in Sec. 1.2).

Noise Variability

Recently, to deal with the problems of ambient noise that arises in real-life situations, DA with additive and convolutional noise with neural networks has risen to be one of the best approaches in SR. Then, the use of models to effectively denoise – or dereverberate – speech samples maintaining specific speaker information using DNNs is a flourishing field with emerging work [226]. Current research includes two-stage models showing improved speaker intelligibility [227], Long-Short Term Memory (LSTM) architectures exploiting speech sequential characteristics [228], unsupervised feature enhancement modules robust to noise unconstrained conditions [229], and specially targeted speech enhancement modules with the joint optimization of speaker identification and feature extraction modules [230],[204],[215].

4.2.1 Challenges of Variability in Speaker Recognition

Speech in real life is commonly noisy and under unconstrained conditions that are difficult to predict and complicate their recognition and understanding. Speaker Recognition (SR) systems need high performance under these 'real-world' conditions. This is extremely difficult to achieve due to both extrinsic and intrinsic variations and is commonly referred to as Speaker Recognition *in-the-wild*. Usually, this problem of variability affects speech systems due to their reliance on probabilistic models trained from clean training corpora. That means that there is a need to develop robust systems that can handle variability without a degradation in performance. Extrinsic variations encompass background chatter and music, environmental noise, reverberation, channel and microphone effects, etc. On the other hand, intrinsic variations are the inherent factors present in speech from the speakers themselves, such as age, accent, emotion, intonation or speaking rate [231].

Automatic speech recognition (ASR) systems aim to extract the linguistic information from speech in spite of the intrinsic and extrinsic variations [210]. However, speaker recognition (SR) takes advantage of the intrinsic or idiosyncratic variations to find out the uniqueness of each speaker. Besides intra-speaker variability (emotion, health, age), the speaker identity results from a complex combination of physiological and cultural aspects. Still, the role of emotional speech has not been deeply explored in SR. Although it could be considered an idiosyncratic trait, it poses a challenge due to the distortions it produces on the speech signal. It influences the speech spectrum significantly, having a considerable impact on the features extracted from it and deteriorating the performance of SR systems.

At the same time, some examples of extrinsic factors are noise, music or the reverberation present in the environment or in the transmission channel. Some examples of extrinsic factors are the speaker's accent, emotions, speaking rate, and style. The mismatch problem between the statistical features of the training utterances and those of real-life can lead to very different characteristics on the speaker's voice, causing speaker recognition models to lose some of their precision and predictive power. Extrinsic variations have been a long standing challenge affecting the basis of all speech technologies. Deep Neural Networks have given rise to substantial improvements due to their ability to deal with real-world, noisy datasets without the need for handcrafted features specifically designed for robustness. One of the most important ingredients to the success of such methods, however, is the availability of large and diverse training datasets.

4.3 Effects of Stress in Speaker Recognition Rates

In this section we detail the experiments conducted in our own contributions published in [3] and [10]. In them, we aim to analyze how stress in speech affects speaker recognition rates. We aim to find techniques for strengthening speaker recognition systems, either by neutralizing the effects of stress – and ultimately fear – or by being able to synthesize it from neutral speech, to strengthen the training of the machine learning inference models. We propose the use of DA techniques using synthetically generated speech under stressed conditions by modifying speech's pitch and speed, together with an analysis of the best feature extraction methods to create robust SR inference models to emotions variability by adapting the data while maintaining a lightweight architecture.

The block diagram for the methodology followed is represented in Fig. 4.1. The characteristics of the database employed, the automatic labelling process based

on heart rate measurement, the two stages feature extraction, the data augmentation and normalization techniques are described in detail next.

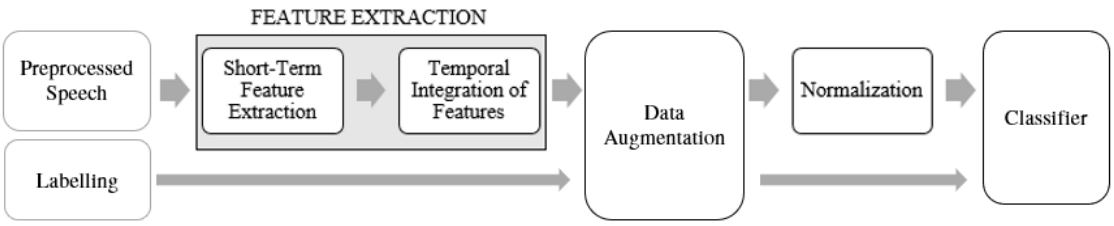


FIGURE 4.1: Block diagram of the speaker recognition under stress conditions methodology with VOCE Corpus. Reproduced with permission from the copyright owner, the authors of [3] via Creative Commons License CC-BY 4.0 from MDPI.

On this part of the thesis we used the VOCE Corpus Database [156] which we already described in detail, together with the preprocessing and the data augmentation techniques used in Sec. 3.2.1. Regarding the labelling, we work with two types of labels for each audio utterance: boolean *stress labels* that indicate the presence or absence of stress, and *speaker labels*, taking values from 1 to n, representing the speaker id of each the audio sample, where n is equal to the total number of speakers, different in each set; in Set 1 n is equal to 10, and in Set 2 n is equal to 11, with a total of 21 speakers – see Sec. 3.2.1 –.

After the speech pre-processing block, we extract handcrafted acoustic features from it. These should reflect the anatomy of the speech production system (e.g., size and shape of the larynx and mouth) and the learned behavioral patterns that shape the speaking habits (e.g., voice pitch, speaking style). For the feature extraction we use a window of 20ms with 50% of overlap – very common values used to analyze the temporal evolution of the signals in literature [232] –. To convert the feature vectors into the resolution of 1s and align them with the stress labels, we perform the mean and standard deviation of the acoustic characteristics over segments of one second. Thus we obtain one feature vector per second of audio. These features add up a total of 34 and are the mean and standard deviation of MFCC 1-13, formants 1 to 3 and pitch. They were selected according to the literature for emotions and speaker recognition [232, 213].

4.3.1 Synthetically Generated Stressed Samples

The pitch and the elocution speed were two variables we informally observed to be changing between neutral and stressed speech utterances. As a consequence, we performed an analysis to measure the differences between the mean pitch from neutral to stressed audio utterances for each speaker using the VoiceBox [166] toolbox. An estimation of the average elocution speed for each user was calculated by computing the mean number of words per second of each speaker, obtaining an automatic transcription of each of the recordings using Google Speech Recognition [234], and dividing it over the length of the audio signals after having removed silent audio frames with a VAD module.

The differences in pitch from neutral to stressed speech were in a range between a relative percentage of -2% and $+7\%$ for all speakers, increasing an average 2.2% Hz. In regard to the elocution speed, subjectively, it seemed to rise in stressed speech utterances, however our analysis gave us the opposite conclusion. The number of words per second was higher when the user was reading a text, 2.2 words/s on

average, in comparison with when the speaker was performing an oral presentation, 1.85 words/s. By listening to the signals, we determined that the words were spoken faster during the public speaking setting but there were many short pauses and pause fillers – words like '*ehm*', '*um*', '*ah*' – between them, that did not count as words for the transcription but were not removed by the VAD module either. Those causes lead to a lower elocution rate in overall.

Thus, we applied modifications in the locution speed and the pitch on the original database, to produce synthetically generated stressed samples of speech. The pitch was modified by the following relative percentages $[-6\%, -3\%, +3\%, +6\%]$ and the speech signals were slowed down – with the aim of extending the duration – by the following percentages $[-20\%, -15\%, -10\%, -5\%]$. All of these modifications were applied to the original audio signals and resulted in what we call a new *synthetically generated stressed set* per modification. In this manner, we augmented our data by a factor of 9, the original dataset plus 8 modifications.

4.3.2 Experimental Set-up and Results

Originally, for an initial experimental set-up we used the data available for Sets 1 and 2 together (21 speakers, detailed in Sec. 3.2.1, the number of samples can be observed in Fig. 4.1). This preliminary experiment is made to observe the behaviour of the speaker identification rate in mismatch conditions. First of all, we divided the data into Neutral (N) and Stressed (S) speech utterances and experimented training with one type of speech, testing with the other and then mixing both types, using a conventional Multi-Layer Perceptron (MLP) as the inference model. The **results** in terms of accuracy – the percentage of audio segments correctly classified – can be found in Table 4.2. In order to get reliable results these experiments were repeated 50 times, where for each repetition the data used for testing (30%) was chosen randomly – when in matched settings, the samples used for training were excluded.

Samples	Neutral	Stressed	Total
Set 1	1389	3989	5378
Set 2	1716	4858	6574
Total	3105	8847	11952

TABLE 4.1: Number of samples of VOCE used [3].

As we could expect at first, matched settings are better than mismatched. When training with neutral utterances and testing with stressed, accuracy decreases more than a 15% with respect to match settings, so it seems that stressed speech does have different characteristics compared to neutral speech that affect SI. On the contrary, when training with stressed utterances of speech and testing with neutral, the decrease in accuracy with respect to the matched setting is not that important (5% absolute) comparing it to the reversed case. This may indicate that stressed speech could be less homogeneous data in which neutral speech could be contained but not vice versa. Regarding the mixed conditions experiments, the accuracy reached a 96.05%, achieving a positive result for this particular task.

In the next experimental set-up, we aim at measuring the accuracy achieved by the system when training with the different synthetically generated stressed sets. We perform pitch and speed modifications to artificially stress utterances for the Set 1 of speakers, and test it with originally stressed speech. The results achieved in these experiments should reflect which modification imitates best the original

Training Set	Test Set	Mean (%)	Std (%)
Neutral	Neutral	96.73	0.33
	Stressed	79.21	0.90
Stressed	Stressed	95.87	0.28
	Neutral	90.89	0.49
Mixed	Mixed	96.05	0.12

TABLE 4.2: Accuracy results for speaker recognition under stress conditions with VOCE under matched and mismatched settings [3].

stressed samples. We kept the test set fixed for these experiments, a 30% of the samples of original stressed speech. Additionally, the same 30% in every synthetic generated stressed set was removed to achieve a more accurate comparison between experiments and guarantee that the test samples were never present in the training set even if they had been modified by our augmenting procedure.

Set compositions in Fig. 4.2 were grouped forming different combinations in order to acknowledge the differences in accuracy for each particular setting used in the training stage. On the left side, we represent the original dataset, composed by neutral and stressed samples. In this case, we used the 30% of the examples of the the stressed collection as the Test set for later experiments. On the right, we represent a diagram of one of the synthetically generated stressed sets (the original neutral speech becomes "synthetic stressed speech" and the original stressed speech becomes "synthetic *super* stressed speech"). The 30% of data used before as Test was removed to obtain more reliable results. There are several synthetically generated datasets, one per modification applied.

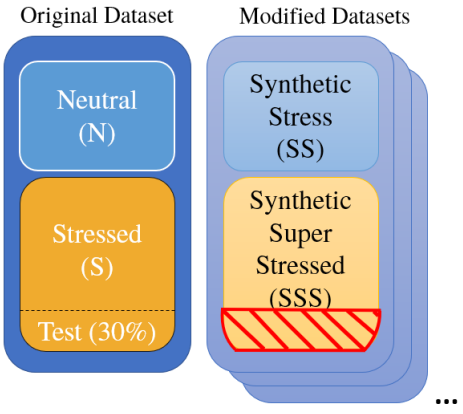


FIGURE 4.2: Schematic of Original and Modified Datasets of VOCE. The red part refers to the equivalent to the Test samples on the block in the left, meaning that they were correctly removed when SSS was used for training. Reproduced with permission from the copyright owner, the authors of [3] via Creative Commons License CC-BY 4.0 from MDPI.

In Figs. 4.3 and 4.4 we present the **results** obtained, we enumerate the data used for the training step on the X axis, the Y axis represents the accuracy achieved, and each colour bar indicates the modified set used for training. In Fig. 4.3 we can observe that the modifications that obtain the highest accuracies are *Pitch* +3% and *Pitch* −3%. When it comes to Fig. 4.4, although the speed results are very similar,

the alteration that works worse is *Speed* -20%. As for the training sets used, the SSS set works better than the SS in both cases.

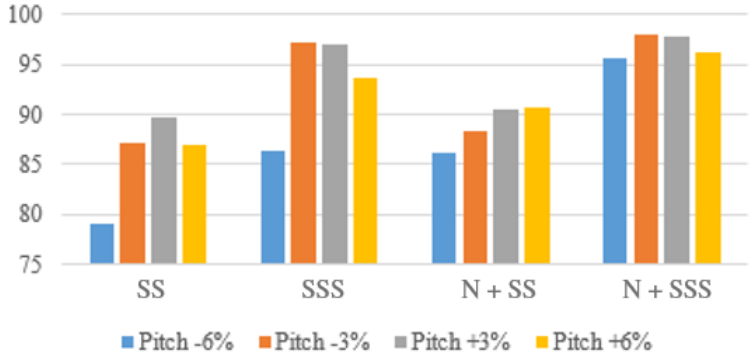


FIGURE 4.3: Accuracy results training the model with synthetically generated stressed data with pitch modifications, and testing with original stressed utterances in Set 1. Reproduced with permission from the copyright owner, the authors of [3] via Creative Commons License CC-BY 4.0 from MDPI.

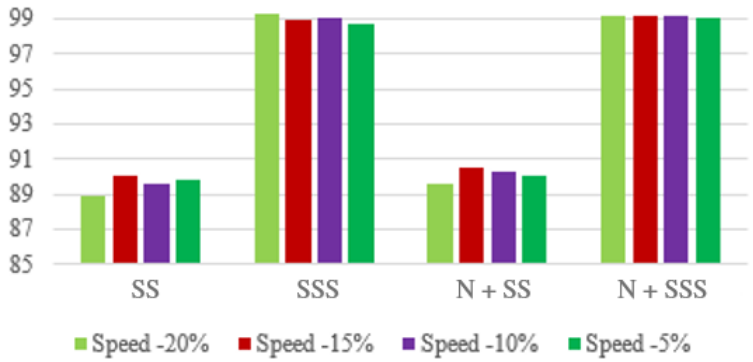


FIGURE 4.4: Accuracy results training the model with synthetically generated stressed data with elocution speed modifications, and testing with original stressed utterances in Set 1. Reproduced with permission from the copyright owner, the authors of [3] via Creative Commons License CC-BY 4.0 from MDPI.

For the next set of experiments we decided to perform the modifications to the audio recordings of Set 2 that had achieved higher accuracy rates in Set 1. These were pitch $[-3\%, +3\%]$ and signal speed $[-15\%, -10\%, -5\%]$ as mentioned. We joined Sets 1 and 2, transforming the problem in a 21-speaker SI task and combined all the synthetically generated stressed data into one dataset, augmenting in a factor of 6 the original data size, 5 modifications plus the original dataset. The same analysis were done for Set 1 and Sets $1 + 2$.

In Table 4.3 we observe two types of experiments, some in which we substitute data and others where we augment data on the training stage. These experiments were repeated 20 times for reliability. As for substituting the original set by a synthetically generated stressed one, we have experiments 6 and 7 to be compared with experiments 1 and 2 respectively. Data substitution achieves similar results to those with original data when using synthetically generated data from neutral

Exp. Num.	Case	Set 1, *Mean*	Set 1, *Std*	Set 1 + 2, *Mean*	Set 1 + 2, *Std*
1	N	89.71	0.56	78.55	0.60
2	S	98.59	0.16	97.37	0.21
3	N + S	98.48	0.23	97.21	0.26
4	N + SS	89.97	0.39	80.46	0.53
5	N + SSS	99.93	0.05	99.16	0.11
6	SS	89.72	0.53	78.19	0.71
7	SSS	99.88	0.07	99.21	0.13
8	N + S + SSS	99.91	0.07	99.45	0.08
9	N + S + SS + SSS	99.94	0.06	99.22	0.11
10	N + SS + SSS	99.91	0.07	98.97	0.14

TABLE 4.3: Accuracy results for speaker recognition under stress conditions with VOCE with synthetically generated speech using different combinations [3].

speech for training (case 1 vs. case 6) as well as better identification rates when using synthetically generated data obtained from stressed speech (case 2 vs. case 7).

Data augmentation experiments are $3, 4, 5, 8, 9$ and 10. The outcome is indeed positive, the best results are achieved in experiment 8 with a 99.45% of accuracy for Sets 1 + 2. These results show us that augmenting the data with synthetically generated stressed utterances of speech boosts the SI rate.

One of the objectives of these experiments was to determine whether experiment 4 could outperform experiment 2. This would mean that we had accomplished the task of generating appropriate synthetically generated stressed speech out of neutral utterances. However, we can see that the procedure we employed was not enough to be used as a substitute. Nevertheless, in Table 4.3 we observe that case 4 performs better than case 6, which in turn outperforms case 1. This shows that synthetically generated stressed speech and using it as training data alongside with original stressed data increases the performance of the SI system.

4.3.3 Discussion

Our goal in this section was to analyze how stressed speech utterances influenced the performance of Speaker Identification systems. We have identified a problem, stressed speech in the testing stage affects negatively when SI systems use an MLP model and are trained only with neutral speech.

As for the case of match and mismatch conditions, in the mixed setting – using neutral and stressed original utterances for both training and testing – the SI system achieves a 96.05% of accuracy, a satisfactory rate for this type of tasks, demonstrating that the set of features chosen for the task is adequate.

In the preliminary experiments for data substitution, depending on the difference between the synthetically generated data and the original one used for training, some substitutions outperform the results achieved using original data. Besides, the modifications over the pitch of the speaker work better when we include synthetically generated stressed samples for training, than when we include the modifications in speech speed. However, when we use super synthetically generated stressed samples for training, the sets modified by changes in speed achieve better results.

Regarding the experiments for augmenting the database with artificial stress, we can conclude that the generation of different synthetically generated stressed utterances of speech by modifications in pitch and speed, and their addition to the database, enlarges meaningfully the instances to work with, improving substantially the results achieved by the Speaker Identification system with a 99.45% of accuracy.

Several experiments and methods remained unexplored and are left for future work, such as shifting the paradigm to a Speaker Verification setting, which could narrow the conditions of the problem and make it more convenient for our task, as well as using speech under conditions of real fear would make the training and test conditions match completely. As Bindi works in real environments, it would be opportune to strengthen the system by degrading audios as if they had been recorded in a real environment, such as adding noise to the database used and analyze its effect. It would also be of interest to further analyse the differences between neutral and stressed speech to find new modifications to be applied to neutral speech to transform it into appropriate synthetically generated stressed speech.

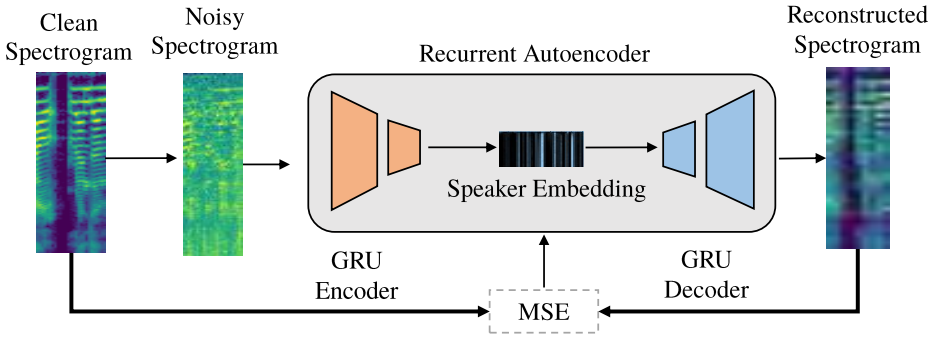
4.4 Speaker Embeddings from an End-to-end Recurrent Denoising Autoencoder

The variability that real life conditions induce in speech are indeed a handicap for speaker recognition systems, for instance the emotional state of the speaker, environmental noise,... This speech is called 'in-the-wild'. By means of the principles of representation learning, in this section we aim to detail our own contribution published in [2], on the design of a recurrent denoising autoencoder that extracts robust speaker embeddings from noisy spectrograms to perform speaker identification.

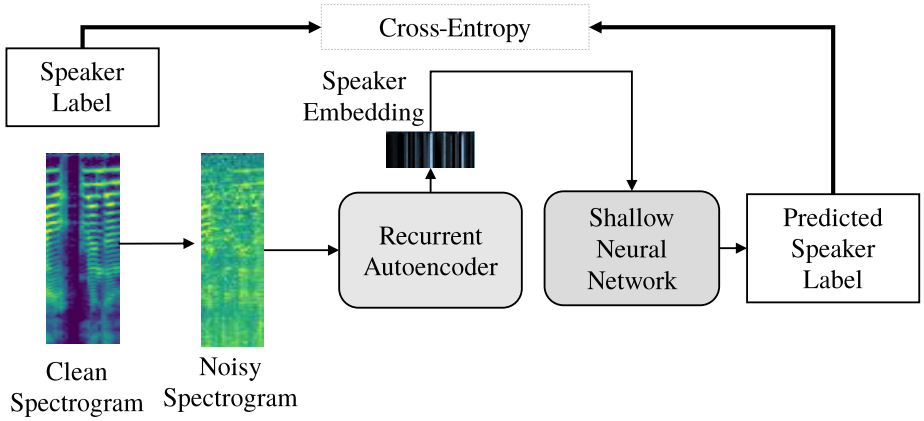
We address the combined problem of the lack of environmental noise robustness of SR systems and the effects of negative emotional speech on their performance. Our contribution capitalizes on using robust speaker discriminator oriented embeddings extracted from a Recurrent Denoising Autoencoder combined with a Shallow Neural Network – a feed-forward neural network, equivalent to a Multi-layer Perceptron (MLP) – acting as a back-end classifier for the task of Speaker Identification, as detailed in Fig. 4.5. This end-to-end architecture is designed to work under adverse conditions, both from the point of view of distorted speech due to stressing situations, and environmental noise.

We choose the VOCE Corpus Database [156] because it includes speech recorded under spontaneous stress conditions, and due to the real-life nature of it. Moreover, we augment our database with synthetic noisy signals by additively contaminating the dataset with environmental noise to emulate speech recorded in real environments.

We discuss a recurrent denoising autoencoder architecture based on Gated Recurrent Units (GRU), where the recurrent architecture targets modelling the temporal context of speech utterances. The encoder network extracts frame level embeddings from the speech spectrograms and is jointly optimized with a feed forward network whose output layer calculates speaker class posteriors. With the help of the denoising module that attempts to remove environmental noise information, and the SNN that targets recognizing the speaker, all the information that is not directly employed for speakers' identification is dismissed from the



(A) Spectrogram Enhancement



(B) Speaker Identification

FIGURE 4.5: Proposed architecture components: Recurrent Denoising Autoencoder and Shallow Neural Network [2]. Reproduced with permission from the copyright owner, Springer Nature.

embeddings. In particular, the loss function associated with this last dense network is also fed into the denoising autoencoder to guide its efforts towards the SR task.

Finally, we observe that these speaker discrimination oriented embeddings are more robust to noise and stress variability than those optimized separately by comparing the effects of automatically extracted embeddings by this two-stage connected architecture against the two modules separately, hand-crafted features previously demonstrated to be suited for this problem and a frequency recurrent alternative obtained by transposing the inputs to the GRU autoencoder.

The main difference with respect to similar works such as the mentioned in Sec. 4.2 – in particular [204] – consists of, first the shallow approach of the back-end oriented towards having a fast and real-time running system in a wearable device, seeking for a balance between computational complexity and performance; and second the use of an end-to-end system for extracting embeddings containing only speaker-relevant information together with the identification task.

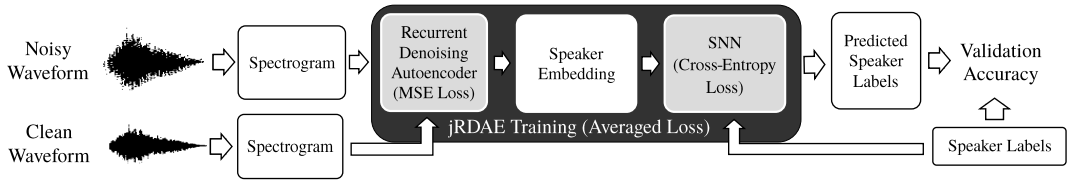
4.4.1 Model Architecture

The proposed architecture is the combination of a Recurrent Denoising Auto-Encoder (RDAE) and a Shallow fully-connected Neural Network (SNN) backend – which is equivalent to a MLP – in an end-to-end system. Autoencoders are generally unsupervised machine learning algorithms trained

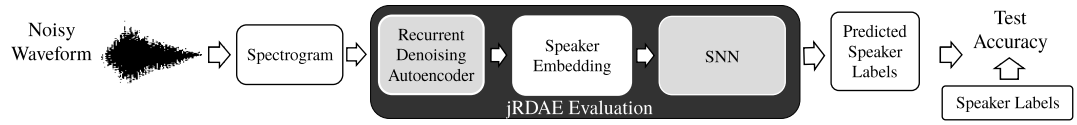
towards reconstructing their inputs through a series of layers. Denoising Auto-Encoders (DAE) take in a corrupted version of the data as the input and a clean version as the desired output and try to reconstruct the latter from the former. Our proposed RDAE is composed of a two-layer encoder and a symmetric decoder based on GRUs. The SNN includes a dense plus a dropout layers.

An autoencoder is a mathematical model trained on unlabeled data and used to convert the input data in a compressed feature representation (the so called bottleneck), and then convert that feature representation, back to the fimension of the input data. In our case, as an input the encoder takes a one-second log-scaled mel-spectrogram, and encodes it into a low-dimensional representation. Although SI systems tend to use longer temporal windows to secure their decisions, Bindi needed a real-time and quicker outcome that has motivated this *short-utterance* speaker identification architecture.

After its extraction, the embedding is fed simultaneously to the decoder and the SNN (see Fig. 4.6). First, the decoder tries to reconstruct a clean spectrogram from this embedding extracted from a noisy spectrum yielding the Mean Squared Error (MSE) between the reconstructed and clean spectrograms. Second, the SNN in charge of identifying the speaker to whom that utterance belongs to, computes the cross-entropy of the predicted speaker and the true speaker labels.



(A) Jointly optimized training procedure



(B) Testing procedure

FIGURE 4.6: Procedure for training and testing stages in the proposed architecture in [2]. Reproduced with permission from the copyright owner, Springer Nature.

Equations 4.1 and 4.2 represent the loss functions, $\mathcal{L}_d$ and $\mathcal{L}_s$, of the RDAE (mean square error) and SNN (cross-entropy) respectively

$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N}(S_i - \hat{S}_i)^2 \tag{4.1}$$

$$\mathcal{L}_s = \sum_{i=1}^{N} -logP(\hat{y}_i|y_i) \tag{4.2}$$

where S is the clean spectrogram, $\hat{S}$ the reconstructed spectrogram from the noisy one, and y and $\hat{y}$ are the original and predicted speaker labels. N represents the total number of speech samples. Finally, instead of sequentially training the RDAE and the SNN, the whole architecture is jointly optimized using an equally weighted cost function that linearly combines the previous two metrics as Equation 4.3.

$$\mathcal{L}_T = \lambda \mathcal{L}_d + (1 - \lambda)\mathcal{L}_s \tag{4.3}$$

We have empirically observed that the normalization of the spectrograms results in a normalized MSE loss that falls roughly within the same dynamic range as the cross-entropy loss. Since we did not have any *a priori* reason to think that one of the tasks could influence the result more than the other we set $\lambda = 0.5$. This showed good results in our test but further exploration of this parameter should be undergone as future work.

4.4.2 Data Augmentation

The VOCE Corpus [156] is used in this experimentation since first, it contains data taken in real stress conditions and second, it offers data from sensors similar to those present in Bindi. In order to simulate real-life environments, speech signals were additively contaminated with 5 different noises from $-5dB$ to $20dB$ in steps of 5dB Signal to Noise Ratios (SNR). Noise signals were chosen from the DEMAND dataset [235]: *DWASHING*, *OHALLWAY*, *PRESTO*, *TBUS*, *SPSQUARE* and *SCAFE*. The noises were chosen to emulate everyday life conditions similar to those envisioned for Bindi deployment. The noises were high-pass filtered to eliminate frequencies lower than $60Hz$ to remove the power line interference, specially noticeable in *Dwashing* noise.

We used a $70ms$ FFT window, an overlap of 50% and 140 mel frequency bands and extracted the spectrograms of the speech signals for each second of audio using the spectrogram extraction module in [236, 217] thus resulting in 27 time steps and 140 mel-frequency bands mel-spectrograms. These choices showed to be reasonable during a preliminary evaluation. Our choice of a higher number of mel frequency bands and longer temporal windows than typically chosen in hand-crafted feature extraction allows a balance of frequency and time resolution more suited for the recurrent networks. Although the classical choices for these values are inspired in the human auditory system, we hypothesize that machines could take advantage of their computational power when analysing data more than just what humans can hear, and therefore they could be able to overcome the human error rate given enough data is provided.

4.4.3 Experimental Set-up and Results

To measure the robustness of the system we designed a *multi-conditioning* setting in which all the contaminated speech signals at different SNRs, as well as clean speech signals, are combined. This is a more realistic scenario in which the specific SNR is not fixed *a priori* for each training. Special attention was taken to ensure that all samples belonging to the same utterance but contaminated with different noises and SNRs were grouped in the same validation fold, so that none of the various versions of the samples in the validation subset appeared in the training set.

Nested cross-validation was used to optimize the hyper parameters for the autoencoder and the SNN as speaker classifier. In nested cross-validation, an outer loop of 33% of unseen data on the training stage is used to obtain the final test results; an inner loop (3 validation folds) is used to find the optimal hyper parameters via grid search. The test set is unseen so that structural decisions made using data from the same distribution – for which final results are computed – do not undermine the validity of the conclusions reached. A block diagram of the training and testing procedures is detailed in Fig. 4.6.

The spectrograms are reduced in the frequency axis from 27×140 to 27×40. This low-dimensional image is flattened, obtaining a 1080 one-dimensional speaker embedding. The layer sizes of the architecture are shown in Table 4.4.

The number of hidden units of the dense layer of the SNN was set to $1,000$, dropout percentage to 30% and the L2 regularization parameter set to 0.01. We trained for 15 epochs with a batch size of 128 and a learning probability of 0.001. We also added a delay to the stop criterion, a patience of 5 iterations, after which if no improvements are observed, training is stopped. The model with lower validation loss during the training is selected as the optimal. The spectrograms were normalized with respect to the mean and standard deviation of their training set. Each spectrogram in the validation set was normalized in terms of the mean and standard deviation obtained from its correspondent training set in the fold.

Layer	Output
Input	(27, 140)
GRU	(27, 64)
GRU	(27, 40)
Flatten	(1080, 1)

Layer	Output
Input	(1080, 1)
Reshape	(27, 40)
GRU	(27, 40)
GRU	(27, 64)
Time Distributed	(27, 140)

Layer	Output
Input	(1080, 1)
Dense	(1000, 1)
Dropout	(1000, 1)
Dense	(21, 1)

TABLE 4.4: Output dimensions of the layers of the Autoencoder and SNN backend architectures. Encoder (left), decoder (center) and SNN (right) [2].

We compared the performance of our proposed *jointly* optimized method (jRDAE) against three different architectures. First, the same system as ours in which the RDAE and the back-end SNN have been *independently* optimized (iRDAE). Second, a transposed (frequency) Recurrent Denoising Autoencoder that differs from our approach in that the spectrograms used as input are *transposed*, as well as the GRU layers, and it is the time axis the one reduced in dimensionality. This aims at recurrently modelling the frequency domain. Finally, a system in which hand-crafted features such as pitch, formants, MFCCs and energy, were chosen based in the literature [3], are fed directly into the backend SI component, the only module to be trained.

Our results are displayed in Fig.4.7, where confidence intervals are also depicted for each of the results taken as one standard deviation on the 3-fold validation. As a metric to compare the algorithms, we chose Accuracy in terms of speaker identification as the classes were fairly balanced. For each experiment, the confidence interval is shown as a small box-and-whisker plot representing the standard deviation of the cross-validation experiments performed to indicate its statistical significance. Our aim is to achieve robustness and therefore to obtain a less degraded performance when the SNR is low.

The independently optimized cascaded architecture (iRDAE) is the algorithm that achieves the lowest results at all SNRs (with the exception of *Ohallway* at SNRs lower than $10dB$ where it is the second worst). We can conclude that the optimization of the RDAE only, towards minimizing MSE is not consistent with the needs of the SI.

The transposed architecture is the result of taking the spectrograms' axis transposed and therefore reducing the time axis in the autoencoder instead of the frequency. As can be seen in the plots, this results in an inaccurate detection of the speaker. We believe that reducing the sequential temporal characteristics of the spectrograms is a handicap for the SI system.

The handcrafted features (HC), on the other hand, achieve good results for high SNRs, since the features where chosen specifically for the task. HC works acceptably well when small amount of data is available, but its performance worsens very fast when SNR decreases.

For most of the noises, the proposed architecture (jRDAE) achieves the best results for lower SNRs and stable rates for higher ones. jRDAE achieves reliable results for the whole range of SNRs, being a more robust approach than the rest of the architectures. The exception is the *Presto* noise in which a closer look revealed that the denoised spectrograms where rather far from the clean ones.

Additionally, we stratify the results for the proposed jRDAE system (Table 4.5) to observe the differences in performance for *neutral* (N) and *stressed* (S) samples, where in the the last two columns of Table 4.5, *mean* and *std* values are provided as a summary. Clearly, lower SI rates were observed in stressed utterances, showing the difficulties induced by stress, *Presto* and *Scafe* being the most affected. This suggests the need to specifically cater for distortions caused by emotional speech.

Noise \ SNR		-5	0	5	10	15	20	Clean	**Mean**	**Std**
DWASHING	N	36.60	56.04	69.23	78.37	81.77	83.78	-	67.63	1.98
	S	28.45	45.58	58.54	68.88	74.71	78.47	-	59.11	1.14
OHALLWAY	N	49.00	68.76	78.09	81.96	83.87	85.27	-	74.49	2.42
	S	43.43	60.98	71.17	76.74	79.98	81.44	-	68.96	1.28
PRESTO	N	28.53	45.92	65.59	73.85	79.58	82.94	-	62.74	1.91
	S	20.33	38.14	56.84	68.60	75.01	78.63	-	56.26	1.02
TBUS	N	60.05	72.40	80.14	83.40	85.97	85.87	-	77.97	2.43
	S	53.46	66.37	74.47	78.39	80.34	81.12	-	72.36	1.07
SCAFE	N	41.21	61.49	75.29	80.89	84.20	85.59	-	71.45	2.05
	S	29.90	51.25	66.55	74.13	78.71	80.68	-	63.54	1.47
SPSQUARE	N	54.08	71.42	78.97	83.03	85.22	85.45	-	76.36	2.9
	S	48.05	64.05	72.82	78.11	80.46	81.70	-	70.87	1.58
CLEAN	N	-	-	-	-	-	-	86.29	-	-
	S	-	-	-	-	-	-	82.41	-	-

TABLE 4.5: Accuracy results detailed by additive noise and SNR, stratified by Stressed (S) and Neutral (N) samples for proposed jRDAE [2].

In Fig. 4.8 we show a breakdown of the results in terms of neutral and stressed speech, comparing the Handcrafted approach and the proposed jRDAE. On it we can observe a similar deterioration in the stressed cases for all additive noises. Specifically for the Handcrafted model, that follows a similar trend as the results in Table 4.5. Confidence intervals are also depicted for each of the results taken as one standard deviation on the 3-fold validation. The *std* values denote the average of the *std* values of the 3-fold validation process for the 6 SNRs. Stress accuracy results are slightly worse than neutral ones, and for lower SNRs, the results are notably worse than for jRDAE.

With a few non-significative exceptions, we observe better results for neutral speech while for stressed speech the SR achieves lower accuracy rates, for both approaches – HC and jRDAE –. This highlights that stress affects speech and deteriorates speaker recognition rates in spite of having included this particular degradation within the training set. In this section, we are not using the stress vs. neutral labels to actively work to combat stress or reduce its degrading effects and therefore we believe there is still room to improve.

4.4.4 Discussion

In this section we evaluated the performance of speaker oriented embeddings extracted with an end-to-end architecture composed of a Recurrent Denoising Autoencoder for an SR task using a Shallow Neural Network. With this approach, we aimed at mitigating the effects on SR systems caused by variability induced by ambient noise, both for neutral and stressed speech.

The end-to-end proposed architecture used a feedback loop to encode information regarding the speaker into low-dimensional representations extracted by a spectrogram denoising autoencoder. We employed data augmentation techniques by additively corrupting clean speech with real life environmental noise in a database containing real stressed speech. Our study presented that the joint optimization of both the denoiser and speaker identification modules outperformed – specially in lower SNRs – independent optimization of both components under stress and noise distortions as well as the use of hand-crafted features.

Our proposed jRDAE architecture achieves reliable results for the whole range of SNRs contaminated signals, being a more robust approach than the rest of the tested architectures. In the resulting tables, lower SI rates were observed when performing inference in stressed utterances, showing the difficulties induced by stress. This suggests the need to specifically cater for distortions caused by emotional speech.

Regarding the system's computational cost, we need to take into account that this speaker identification module is expected to be embedded into a computationally constrained device, and lightweight systems are preferred for such, in order to increase battery life. The decision of using GRU cells instead of LSTM cells was based on the fact that the number of parameters is significantly lower and therefore GRUs are fast and less computationally expensive than LSTM. With this decision the main speed bottleneck is now the SNN, with 1.1 million parameters. In the future, we aim to reduce the number of parameters of this model to develop a lightweight intelligent algorithm to be embedded within the Bindi system.

To further analyse the robustness of this speaker oriented embeddings and end-to-end architecture, it could be tested in an adversarial fashion by using an emotion – or stress – classifier as a domain adversarial module. We also intend to use richer datasets that contain real life speech, specifically WEMAC [11]. To deal with the problem of data scarcity, transfer models could be used and adapted to emotional speech other large-scale datasets for speaker identification such as the crowd-annotated VESUS [238] and VOXCeleb [239].

4.5 Speaker Recognition's Response to Acoustic Events

In the line of speaker recognition under stress conditions, additionally we perform some SR experiments with another speech database which includes realistic stress conditions. The work detailed in this section is published in [8], for which we, with the assistance of other members of UC3M4Safety team, are responsible for the contribution.

We determined that BioSpeech (BioS-DB) [157] was a good fit to our interests among other suitable databases, since it includes continuous-time annotations in the Arousal/Valence space for non-acted presumably realistic stressed speech due to its public speaking setting, and incorporates physiological data (Blood Volume Pulse – BVP – and Skin Conductance – SC –) as in Bindi which could be of great use for multimodal models in the future. Note, however, that the purpose of this

dataset collection is quite different from ours since their creators aimed at predicting bio-signals from speech.

In general, the main difficulty with emotionally labelled data relies on the proper labelling process (see Sec. 2.5). There is no universal agreement on how to categorize or measure emotions. The self-assessment annotations by a specific subject can differ from labels annotated by external evaluators observing the said subject. Moreover in Sec. 3.2.2, we already introduced our own reinterpretation of the labelling of BioS-DB, more suitable for a stressed speech classification task (further information in Sec. 5.6.1).

On the other hand, the emotional state of the subject could influence negatively the performance of any speech technology and in particular, their identification [240]. Identifying the target user's voice, separating it from the rest of the speakers, opens an interesting possibility for situations where it would be desirable to identify all the speakers involved in the scene, e.g., in case of legal evidence required.

The creation of Biospeech+ (see Sec. 3.2.3) arises from the need to find out if we could use Acoustic Event Detection or Classification (AED/C) of background events to assist the Speaker (SR) and Speech Emotion Recognition (SER) tasks. To decide whether acoustic events could most likely cause a stressful reaction loosely synchronized with the time instants where the emotional labels denote an acute stress occurrence. For Biospeech+ we combined them at different SNR[29] ratios (-5, 5 and 15 dB).

4.5.1 Experimental Set-up and Results

We extracted features from well-known libraries used for SR, SER and AED/C, respectively: librosa [241], eGeMAPS [184] from the openSMILE toolkit [185] and YAMNet embeddings [242], as we will later on explain. The size of our working window is 1 second. This is a trade-off between computational complexity and speed and a requirement in Bindi. Thus, from librosa we extracted 19 features with a window size of 20ms and a 10ms overlap and then their mean and standard deviations every second resulting in 38 features per second. Using openSMILE we extracted the eGeMAPS feature set with 88 features. For extracting features suitable for audio events we used the 1024-dimensional embeddings corresponding to the activations of the top convolutional layer of YAMNet. We used a feature selection method in which the correlation of the concatenation of the three feature sets was used to remove the features with a correlation higher than 95%. This resulted in a reduction of the 68% of the features. Examining the correlation matrices we confirmed that most YAMNet features were highly correlated with each other. All features were standardized by using z-score normalization.

With the chosen window size, BioS-DB contains approximately 5000 samples. This is a small size for the use of deep neural networks, so a simple Multi-Layer Perceptron (MLP) implemented with scikit-learn [243] and two shallow network architectures implemented with Keras [244] are tested, working towards maintaining a low computational complexity. The first of them consists of two hidden fully-connected layers with 50 and 20 neurons, respectively. The second is a combination of a convolutional 1D layer, a bidirectional Gated Recurrent Unit (GRU) layer and a fully-connected layer. This model responds to the idea that it is important to extract information from the temporal context distribution of the features. The Keras models were compiled using an Adam optimizer with a learning

[29] For the SNR measure we consider the foreground speech from Biospeech as the 'signal' and the audio events as 'noise'.

rate of 0.001, categorical cross-entropy as the loss function. All the models used F1-score as the metric to evaluate performance to mend imbalances in the dataset. For all experiments we used a 5-fold cross-validation.

Model	librosa	*p*	eGeMAPS	*p*	yamNET	*p*	L+E+Y	*p*	feat sel	*p*
Speaker Recognition										
MLP	100±0.0	*28k*	72.7±0.6	*43k*	17.8±1.4	*324k*	96.4±1.0	*361k*	98.35±0.3	*128k*
K2D	99.9±0.1	*4k*	64.3±2.0	*7k*	15.21±1.4	*53k*	95.9±0.8	*60k*	96.6±0.7	*20k*
KCGD	100±0.0	*10k*	50.9±0.7	*13k*	12.6±1.9	*73k*	90.8±1.3	*81k*	95.7±0.9	*31k*

TABLE 4.6: F1-score results for Speaker Recognition in clean Biospeech [8]. MLP refers to the Multi-Layer Perceptron, K2D refers to the 2-dense layers model in Keras and KCGD refers to the Keras model composed of a Convolutional 1D, Bidirectional GRU and Dense layers. Mean and standard deviation results are shown for a 5-fold validation.

The results for Biospeech without the audio events are shown in Table 4.6. When using Biospeech database, the speaker's samples are not equally balanced and we aim to use the F1-score metric – commonly used in inference models for unbalanced data – rather than the accuracy metric as we do when we use VOCE Database. In Table 4.6 p represents the number of parameters of each model.

For the three tasks under consideration, MLP with librosa achieves the best performance. It is worth noting that librosa features achieve the maximum score for the SR task.

The differences in performance between features can be due to multiple reasons, for instance, their nature. librosa and eGeMAPS features are manually extracted whereas YAMNet's are automatically extracted from a pretrained sound-event detection network. Also, their number – 38, 88 and 1024 respectively –, and besides, their specific potential to represent emotions or speaker information. It is worth highlighting the results for eGeMAPS with the K2D model, which is the most lightweight of the 3 models used, performing better than the KCGD with more parameters; and also achieving a performance only less than 10% lower than the MLP, having 1/6 of its parameters.

Fig. 4.9 provides the results for SR for Biospeech+ for different SNRs. We can observe again an almost perfect performance for librosa features, and good performances for eGeMAPS, l+e+y (early fusion of librosa, eGeMAPS and YAMNet features) and feature selection, but a considerable decrease in efficiency for YAMNet embeddings only. This means that acoustic events do not affect the SR task. Besides, YAMNet embeddings do not seem to capture relevant information about the acoustic cues of the speech that could help distinguish between speakers.

4.5.2 Discussion

In this section we aimed to evaluate whether adding acoustic events that were statistically loosely related to the occurrence of stress utterances could improve the performance of an SR system. We draw from the premise that detecting acoustically GBV risk situations involves taking into account speech and acoustic contexts since they could be correlated. However, there are no non-acted datasets that allow to elicit this relationship so that it could be studied, thus, we generated Biospeech+, augmenting BioS-DB with acoustic events.

Besides, in Sec. 3.2.2 we reinterpreted BioS-DB labels, including that the samples labelled $Q2$ were interpreted as those related to fear, anxiety or stress, but we should

note that without the *dominance* dimension, emotions such as anger or rage could lay in that quadrant too, generating a misinterpretation of both.

In this preliminary study we focused on the relationship between speaker, stress and acoustic events. Both the feature sets and algorithms were used with the aim of keeping low the computational complexity and taking into account the number of samples of the database used. And in conclusion, stressful acoustic events with a non-deterministic correlation to stressed speech utterances proved not to affect in the recognition of the speaker.

When it comes to the different feature extraction methods, widely used libraries for feature extraction in speaker recognition tasks – librosa and eGeMAPS – work much better than YAMNet features, which are used for acoustic events detection. The fact that the librosa feature set achieved an F1-score of 100% was thoroughly checked, as we were aware such a high accuracy could denote a training issue, but no error was found. The possible reasons for this 100% score could be because it is a feature set that works very well for the task of speaker recognition, making the model learn the voice identity patterns very well, or due to the fact that the amount of data on the test set is limited. Finally, in Sec. 5.6.1 we will re-use this database to measure the effect of acoustic events in the task of stress recognition, and we advance that for such task, the presence of acoustic events is beneficial.

4.6 Conclusions and Future Work on Speaker Recognition

In this chapter we have tackled the task of speaker recognition with the intention of first identifying the user, for later detecting emotions in their voice that may indicate a situation of risk. We focus on two variability aspects, first on the recognition of the speaker under stress conditions, to understand how and how much these stress conditions affect the speaker recognition task, and second on speaker recognition under real-life (noisy) environments, which is where our application will ultimately work.

In Sec. 4.3 we analyze how stressed speech influences the performance of SI systems and we identify that it does impact negatively to speaker recognition rates when SI systems use ML models trained only with neutral speech. We worked with matched and mismatched settings for such purpose, and created synthetically generated data to substitute the non-existent stressed data out of neutral speech, which enlarges meaningfully the instances to work with, improving substantially the results achieved by the systems. Therefore in the absence of real emotional stressed speech – and ultimately speech in fearful conditions — we can augment the data with increased pitch modifications and speech rate slowdowns to achieve data that resembles real stress and can help maintain an acceptable recognition rate in SR systems.

In Sec. 4.4 we employed data augmentation techniques by additively corrupting clean speech with real life environmental noise in a database containing real stressed speech, to study the relationship between these 3 factors – speaker, noise and emotions – based on the premise that speech *in-the-wild* is a handicap for speaker recognition systems due to the variability induced by real-life conditions, such as environmental noise and the emotional state of the speaker. The design of a recurrent denoising autoencoder that extracts robust speaker embeddings from noisy spectrograms to perform speaker identification addresses the combined problem of the lack of environmental noise robustness of SR systems, even when including stressed speech for its training. This representation learning based

method takes advantage of the joint optimization of a denoiser and SI block with a combined loss function, which is shown to work better than a general purpose denoiser. It achieves reliable results for the contaminated signals in the whole range of SNRs, being a more robust approach than the rest of the tested architectures. The performance difficulties induced by testing the architecture in stressed speech are observed when achieving lower SI rates for stressed than for neutral speech utterances. This suggests the need to keep on addressing specifically for distortions caused by emotional speech.

In Sec. 4.5 we augmented speech data with additive acoustic events, drawing from the premise that detecting risk situations involves taking into account speech and acoustic contexts since they could be correlated. But stressful acoustic events with a non-deterministic correlation to stressed speech utterances proved not to affect detrimentally in the recognition of the speaker. Thus, the relationship between stress detection and such acoustic events is left to be further studied in Chapter 5.

In closing, the lack of real emotional data is indeed a drawback for SR systems without them to train our ML models the best recognition rates – in match conditions – are difficult to achieve. It is also very important to consider that speech recorded in real conditions includes environmental noise that is also detrimental for SR systems, and so it must be eliminated with correct denoising methods in order to achieve the best SR performances. In the absence of real stressed speech and noisy conditions with which to train and test our SR inference models, we have found that augmenting the data by synthetically stressing it and adding ambient noise allows us to study and design SR models that are more robust to stress and noise.

WEMAC and WE-LIVE were created to give an answer to this problem, with which we intend to work further and apply all the knowledge obtained in the previously detailed studies performed in the field of SR for our GBV risk situations detection application. To further analyse the robustness of speaker oriented embeddings, a SR model could be tested in an adversarial fashion by using an emotion – or stress – classifier as a domain adversarial module.

The use of data augmentation techniques, together with more realistic data samples as the ones in our datasets – WEMAC and WE-LIVE – paves the way to use of more complex deep neural networks for our problem. Along the lines of foundation models and fine-tuning, the next step could be to use a pretrained neural network and fine-tune it with the data we have available in WEMAC. In order to work towards the challenging task of the detection of GBV risk situations using an inconspicuous device such as Bindi we must constantly keep in mind its constrains, already described in Secs. 1.1.4 and 1.2.2. However, it is worth taking into account that, with the rapidly changing devices, it is possible that in future Bindi versions, we are able to embed more complex and deep neural networks with more inference power and capabilities.

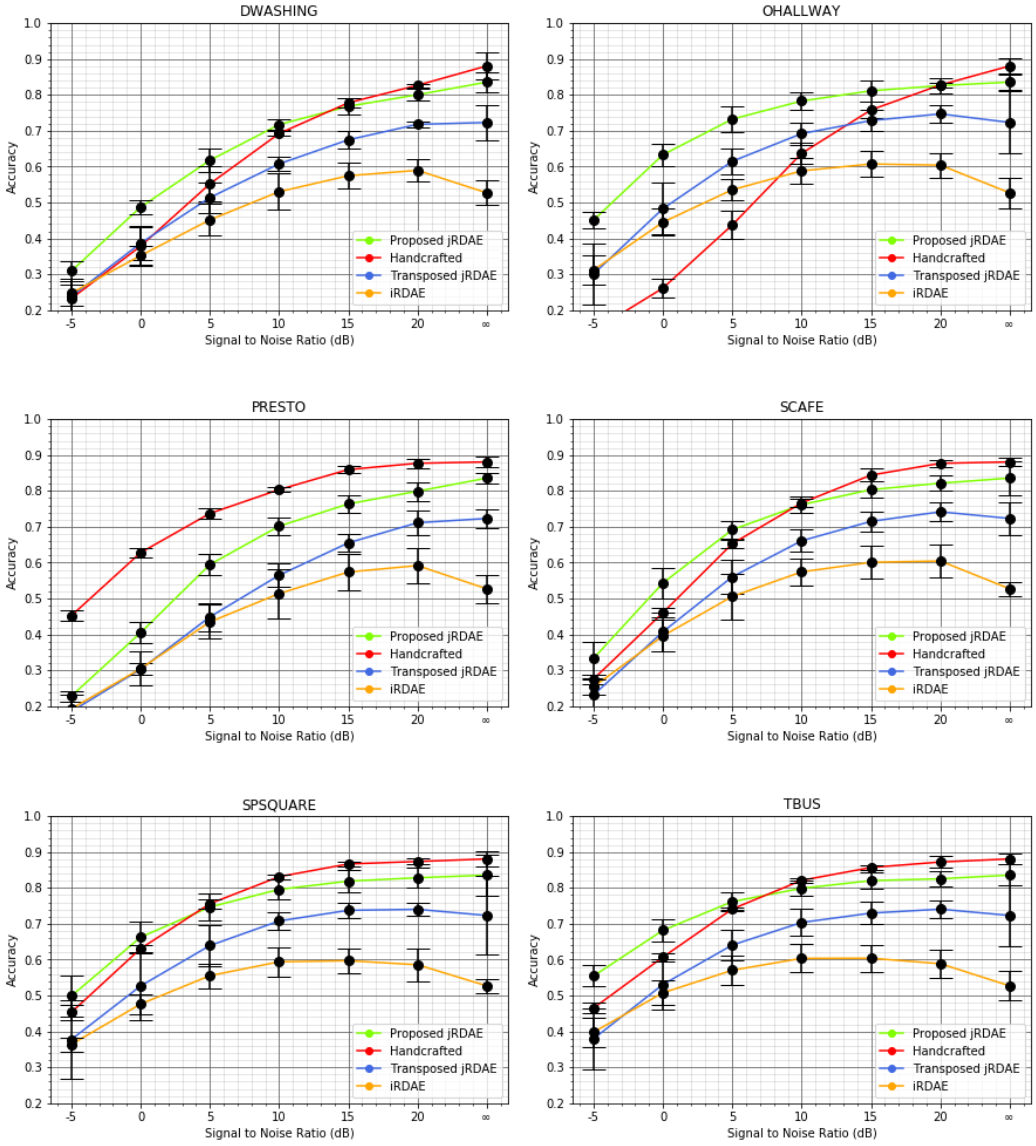


FIGURE 4.7: Results detailed by additive noise and SNR in terms of accuracy for different architectures [2]. Reproduced with permission from the copyright owner, Springer Nature.

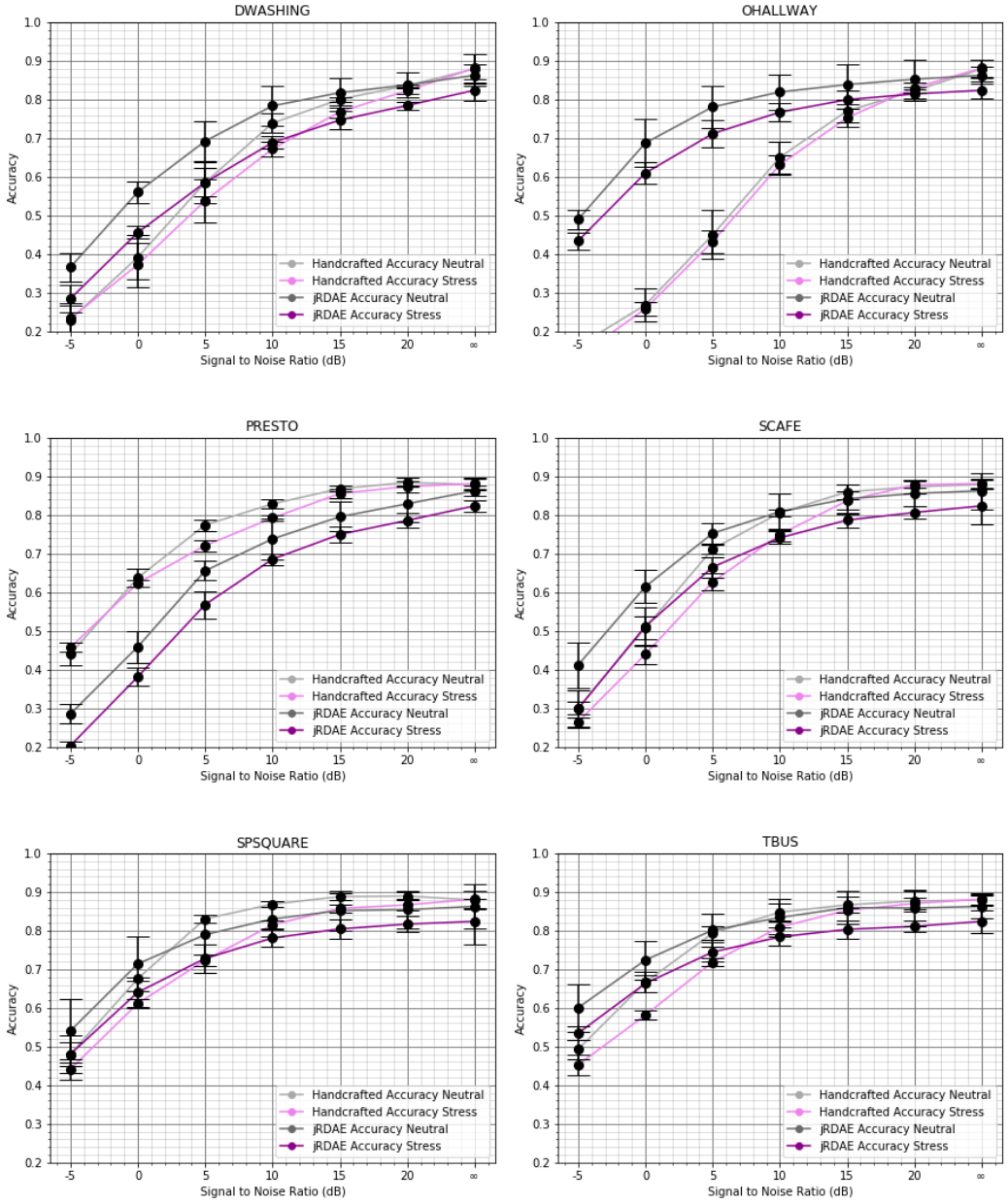


FIGURE 4.8: Results detailed by additive noise and SNR in terms of accuracy for stress and neutral samples for Handcrafted and jRDAE configurations [2]. Reproduced with permission from the copyright owner, Springer Nature.

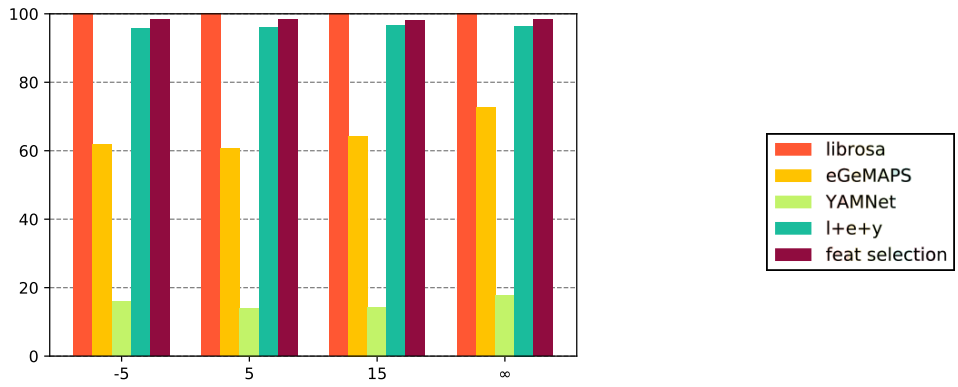


FIGURE 4.9: Speaker Recognition F1-score results with Multi-Layer Perceptron in Biospeech+ [8]. Reproduced with permission from the copyright owner, ISCA.

Chapter 5

A Multimodal Fear Emotion Recognition System for Bindi

In this chapter we dive into the development of the Bindi system for the recognition of fear-related emotions. This chapter is highly multidisciplinary as there are many contributions supported by other members of the UC3M4Safety team. This chapter recounts and reproduces in part or in full the content of the articles published in [9], [1].

First, we describe the system hardware architecture of Bindi, developing the components of the edge, fog and cloud computing systems. We then explain the approach followed for the design of the multimodal fusion strategies for an automatic alert system for Bindi, first a cascade multimodal system for Bindi 1.0, and also, the deployment of a complete Internet of Things system with edge, fog and cloud computing components, for Bindi 2.0. In the latter, we specifically detail how we designed the intelligence architectures in the Bindi devices for fear detection in the user. These contributions were made in conjunction with other members of the team.

In addition, we describe the different data processing pipelines (physiological and speech) and the monomodal experimentation with speech for the detection of fear-related emotions, first by targeting the detection of realistic stress – for which data was more easily accessible in the literature – [1].

Later, as core experimentation, jointly with other members of the UC3M4Safety team, we work with our own database designed and captured by the team – WEMAC [11] – for the task of fear detection. In this chapter there is a strong multimodal component, since we work on the part of emotions recognition from speech together with data from physiological signals, in the same way Bindi's two wearable devices would work.

Finally, a discussion about the architecture of Bindi, the results and their significance followed by conclusions and future research directions are available in the last sections.

This thesis' main contribution relies on working in the speech modality, but it also handles multimodality and the fusion of modalities. The complementary Ph.D. thesis where we find with great detail all the work carried out on the physiological modality – and in particular about the hardware components of Bindi, specifically the bracelet – developed by the member of the UC3M4Safety team Jose Miranda Calero, and can be found in [53].

5.1 Introduction

As we explained in Sec. 1.1.4, Bindi's technological solution surpasses the existing panic buttons to detect threatening situations as they can cause difficulties because victims have to use them manually even in difficult conditions. Instead, Bindi, our end-to-end fully autonomous multimodal system, relies on artificial intelligence techniques that automatically detect violent situations, based on detecting fear-related emotions, and initiates a protection protocol when necessary. To this end, "Bindi integrates modern cutting-edge technologies, such as the Internet of Bodies, Affective Computing, and cyber-physical systems, gathering i) affective IoT with auditory and physiological commercial off-the-shelf smart sensors embedded in wearable devices, ii) hierarchical multisensorial information fusion, and iii) the edge-fog-cloud IoT architecture" [1].

For the detection of risk situations in the context of GBV, we will rely on the recognition of fear-related emotions from the speech and physiological variables of a user, entering directly into the field of SER.

Speech Emotion Recognition, abbreviated as SER, is "the task of recognizing human emotions and affective states from their speech". This relies on the fact that voice often reflects underlying emotions through its characteristics and features. For the development of Bindi, we aim to detect fear emotions in the user that are the result of a risk situation for them. In the previous chapter we took the first step to recognize emotions, which is the recognition of the speaker of which the emotion is desired to be recognized. And in this chapter, we focus in the classification of emotions using speech and physiological variables – focusing on the former –, specifically fear-related ones.

There are no databases available in literature that are suitable for this task, that include speech under realistic fear conditions, so we first worked with stressed speech, being it a close relative of *fear*, and then it became available with our WEMAC Database, which was developed for closing that gap. With such data, we aim to train machine learning models to automatically detect the emotional state of a user, in particular, *fear*.

We make use of speech as well as physiological signals since Bindi aims to be an inconspicuous device – for emotion recognition – and such modalities are non-invasive and can capture data in a daily real-life setting, which is where we target Bindi to work in. These type of interconnected wearable devices belong to the field of Internet of Bodies (IoB), a subfield of Internet of Things (IoT). With the use of these two modalities we develop unimodal and multimodal fear emotion recognition systems.

Types of Fusion of Modalities in Machine Learning

Multimodality is a natural concept for living beings as a means of interacting with the world around us. In all individuals, the acquired information comes from internal and external sensors. This information is combined and fused to provide rapid responses to the external constantly changing environment.

Focusing on the field of Affective Computing, handling more than one modality is challenging because data differ in different aspects as origin, structure, and relevance. However, the diversity within a multimodal emotion recognition system (e.g., combining both physiological and auditory signals) usually allows improving the insights in a way that cannot be achieved by a single modality

[245]. In the literature, there are four main techniques for data fusion: feature-level, decision-level, model-level, and hybrid fusion [246].

In feature-level fusion or early fusion, the different synthetic metrics or features obtained from each input sensor are combined into another feature vector before the machine learning model classification. The main drawback of this method is the high dimensionality of the combined feature vector, which could lead to the well-known *curse of dimensionality*.

Unlike early fusion, decision-level fusion or late fusion requires multiple training stages, one per modality (e.g., one training stage for only physiological signals and one for only auditory signals). This fusion mechanism is based on the unimodal recognition results late combination by some criterion. In this case, each of the modalities can be modeled more precisely by their classifiers, but the system does not handle in any way the interactions or correlations between modalities. The initial version of Bindi – Bindi 1.0 – considered a decision-level fusion technique according to the unimodal inference outputs based on physiological and speech data.

Two other fusion methodologies can be applied to deal with the interaction problem from the decision-level fusion technique: hybrid and model-level fusion approaches. Both combine aspects from the two techniques already commented (early and late fusion). Model-level fusion is based on the mutual correlation between the different streams from the modalities in the system. It is usually considered to explore the temporal correlation between those streams [246]. The hybrid fusion implements more than one fusion level within the same system (e.g., combining feature- and decision-level approaches), which usually provides better recognition results than applying solely one fusion technique.

5.2 Related Work

The multidisciplinary research field aimed at recognising human emotions is the aforementioned Affective Computing [75], [247]. We find, among its applications, to provide better working conditions, entertainment, or services to people. It relies not only on smart sensors and digital signal processing but also on AI techniques, such as machine and deep learning. The collaborative research among psychology, computer science, smart sensors, and cognitive science fields [248] allows for the detection of different emotional states through the monitorization of humans' signals, such as physiological and physical signals. Some examples of physical signals include audio, speech or voice, image or video signals, tracking either the background of the scene or the user. Some examples of physiological variables include Heart Rate (HR), Galvanic Skin Response (GSR), SKin Temperature (SKT), ElectroMyoGram (EMG), and ElectroEncephaloGram (EEG).

5.2.1 Speech Perspective: Speech Emotion Recognition (SER)

Emotion detection has been widely reported in the literature with the use of speech signals [249], [250]. In recent years, the interest in detecting and interpreting emotions in speech is very extensive [251], [252]. Speech Emotion Recognition (SER) consists of the identification of the emotional content of speech signals, the task of recognizing human emotions and affective states from speech. In this field, there are three important aspects being studied and discussed in the machine learning community and literature: i) the choice of suitable acoustic features [232],

ii) the design of an appropriate classifier [253] and, iii) the generation of an emotional speech database [254], [255].

A review of the databases in literature that can be suitable for this thesis, and its challenges can be found in Secs. 3.1 and 3.2.

Speech emotion recognition has applications in human-computer interaction, as well as robots, mobile services, computer games, and psychological assessments, among others. In spite of its many applications and the substantial progress due to the advent of deep learning techniques [256], emotion recognition is still a challenging task, mainly due to the subjectivity involved in emotions (see Sec. 2.5).

The lack of existing speech corpora with strong elicited fear in real situations is a particular problem of our particular research (see Sec. 3.1). However, a few studies have managed to achieve results in this regard. For instance, Clavel et al. [153] developed an audio-based abnormal situations detection system for movie clips. Their results achieved up to 70.3% accuracy for fear detection via a Leave One Trial Out (LOTO) strategy for 30 movies. In [257], they performed emotion detection with para-linguistic cues in a dialog corpus containing real agent-client recordings obtained from a medical emergency call center. As a result, they achieved a recognition rate with up to 64% accuracy for fear recognition.

5.2.2 Emotion Recognition Using Physiological Signals

One of the Bindi devices is a smart bracelet that can track physiological signals, as for the design of an emotion recognition system, the physiological perspective is extremely informative. In physiology and emotions research, the distinction of fear – among other emotions – is not new [258].

However, to the best of our knowledge at the moment of the publication of this thesis, there are only two fear recognition systems based solely on physiological information and self-reported labels.

On the one hand, the authors in [259] used all signals available from the Database for Emotion Analysis using Physiological signals (DEAP) [260] to provide a specialized fear recognition system. They achieved a fear accuracy detection rate below 90%, although they also considered EEG, which is not currently feasible as an inconspicuous wearable device. On the other hand, in previous research from other members of the UC3M4Safety team [261], only three physiological variables available from the Multimodal Analysis of Human Nonverbal Behaviour in real-world settings dataset (MAHNOB) [262] were used, obtaining a fear recognition accuracy rate of up to 76.67% for a subject-independent approach using data from 12 women volunteers. In the latter they concluded the need for a novel data set focused on fear detection, including the usage of immersive technology, considering the gender perspective, achieving proper balanced stimuli distribution regarding the target emotions and having a greater number of participants.

5.2.3 Internet of Bodies

The growth of research on devices that monitor signals from the human body during the last years – as both edge devices in Bindi – implies an imminent extension of the Internet of Things (IoT) domain. This trend emerges concerning interconnected devices (e.g., worn, implanted, embedded, and swallowed) located in-on-and-around the human body forming a network, which is currently being called the Internet of Bodies (IoBs) [263]. This novel field has many applications, such as human activity recognition [264], user authentication [265], and even

emotion recognition [266]. This field also encompasses essential studies on the limitations of such sensors, such as time delay and energy consumption issues [267]. Thus, such in-body sensors can acquire different types of physiological information at the same time, which derives studies related to the use of multimodal data fusion techniques [268], [269].

This IoB proliferation is accompanied by advances in machine learning and deep learning technologies, resulting in an explosion of mobile intelligence and placing increasing demands on computing resources that mobile edge devices cannot meet. Consequently, edge computing capabilities are being boosted and explored to deliver better intelligence engine inference services to end-users [270]. For instance, in [271], they worked on accelerating the training process of large machine learning models in IoT to meet the hardware limitations.

Within this IoB context, the works explained in the following sections intend to provide and foster the generation of novel lightweight multimodal data fusion techniques fed by human body monitoring toward their applicability to current edge-computing devices, such as the ones in Bindi [1].

5.2.4 Multimodal Fusion Techniques

As Bindi is a multimodal system – it uses physiological and speech signals to detect the user's emotions – in this section we give a short review of multimodal theory and architectures.

Some works proposed multimodal approaches combining visual and speech data to improve and strengthen emotion recognition [272] [273]. This conceptualization is not possible in Bindi because there is no visual component. Thus, the additional information will come from physiological variables. Since Bindi is a multimodal system which consists of a bracelet that captures physiological signals (SKT, GSR, BVP) and a pendant that includes a microphone that captures audio signals (acoustic events, speech) (more details in Sec. 1.1.4), these two modalities – physiological and auditory – can work together towards the detection of fear or panic and in consequence the detection of risk situations.

When dealing with emotion recognition combining different data modalities, some comprehensive reviews can be found presenting current state-of-the-art data fusion techniques [274, 255]. These works state the need for: 1) new approaches to advance the community's understanding of multimodal casuistry, and 2) subject-independent emotion recognition models to ease the further deployments under real-life conditions. They also agree on the potential performance improvements with multimodal approaches compared to unimodal ones.

In fact, recently, research in multimodal systems is on the rise. For instance, the authors in [276] proposed a hybrid multimodal fusion emotion recognition system including facial expressions, GSR, and EEG. Their results yielded a maximum subject accuracy of 91.50% and a mean accuracy of 53.80% using a leave-one-subject-out (LOSO) strategy and a publicly available database (DEAP) for different emotion detection use cases, such as angry, disgust, afraid, happy, neutral, sad, and surprised. Moreover, they created their own dataset with which they achieved a maximum subject accuracy of 81.2% and a mean accuracy of 74.2% using a LOSO strategy for three emotion classes, i.e., sad, neutral, and happy. In [277], a weighted-based fusion strategy accompanied by transfer learning techniques was applied for multimodal emotion recognition using EEG and spontaneous spatial expression detection. The work employed a Leave-One-Trial-Out (LOTO) subject-dependent configuration and reported an average accuracy up to 69.75%

and 70.00% for the valence and arousal classification, respectively. In addition to these works, more research can be found regarding multimodal data fusion for stress-related use cases [278], [279].

Analyzing these related works, most emotion recognition systems do not target the fusion of physiological and auditory modalities nor consider vulnerable groups, such as GBVVs. Specifically regarding such bimodal fusion of physiological and vocal information, one of the few works that stands out is [280], to the best of our knowledge. This work considered different data fusion schemes and achieved an average accuracy of up to 55.00% for a subject-independent strategy using a feature fusion when targeting a valence and arousal binary classification. Consequently, there is a current need for research on these topics, which this chapter aims to deepen.

5.3 Bindi System Hardware Architecture

Early in Sec. 1.1.4 we described Bindi, in this section we go deep into its architecture design. A simplified system architecture of Bindi is presented in Fig. 5.1. The following sub-sections provide a technical overview regarding each system component. The edge devices in the Bindi architecture are the bracelet and the pendant.

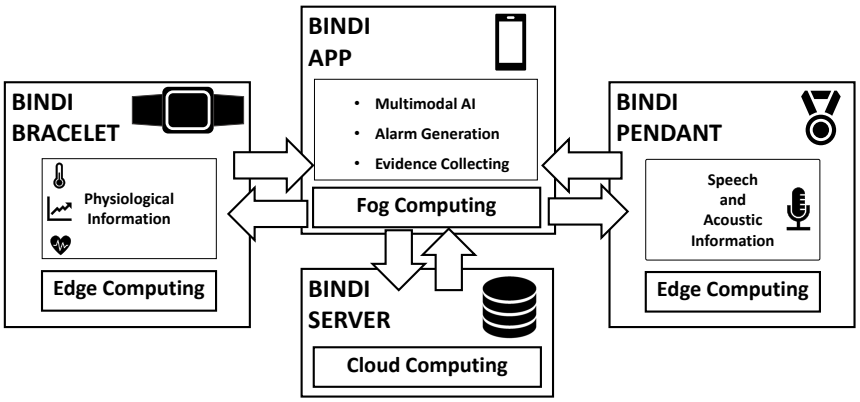


FIGURE 5.1: Simplified Bindi Hardware Architecture [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Edge Computing: Bracelet

This device runs an embedded intelligence engine for fear detection based on physiological information. Fig. 5.2 shows the hardware components integrated into this device, which can be classified into four groups: physiological sensors, actuators, power manager elements, and the microprocessor unit. For further details about them refer to [1]. Note that the radio-frequency module through Bluetooth Low Energy® communication is also integrated within this host unit.

The bracelet is equipped with a conventional electro-mechanical button for manual user activation, acting as a panic button. The physiological sensors capture the following variables:

- HR: This is based on a photoplethysmography sensor that detects Blood Volume Pulse (BVP) changes by measuring the absorption of light emitted through the skin.
- GSR: This sensor utilizes two electrodes to measure the skin conductivity through a DC exosomatic measurement.
- SKT: This integrated circuit is defined as a clinical-grade sensor for wearable applications, providing an accuracy of ±0.1 °C over a 30 °C to 50 °C temperature range.

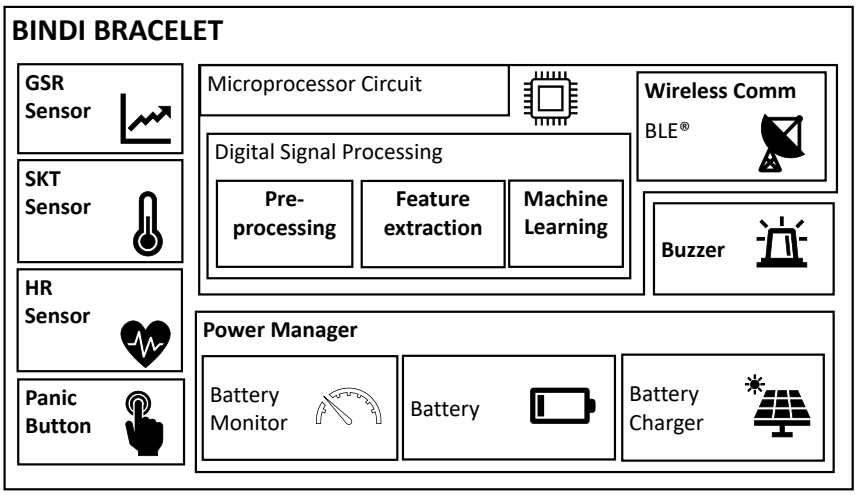


FIGURE 5.2: Simplified Bindi's Bracelet Architecture [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

The previously discussed physiological variables were chosen due to their proven strong relationship with emotion recognition [281] and their ease of implementation in wearable devices. The latter point is particularly relevant and led us to discard other typical physiological sensors used in the field (such as EEG), which do not meet the inconspicuousness requirement. The digital signal processing pipeline within the bracelet entails both, the acquisition and filtering of the physiological signals and the feature extraction and inference stages.

Edge Computing: Pendant

This device captures audio and speech information, which is fed to an intelligent engine for fear detection. The pendant has the same hardware architecture as the bracelet but integrates a microphone instead of physiological sensors. Its architecture is shown in Fig. 5.3. The microphone is based on a microelectromechanical system with an omnidirectional audio sensor. This part includes a capacitive sensing element and an integrated circuit interface, allowing a digital signal to be obtained directly. The digital signal processing pipeline within the pendant entails both, the reception and filtering of the auditory signals (audio and speech) and the wireless transmission to the smartphone. Note that, due to the limited bandwidth of the wireless communication, the audio is compressed prior to being transmitted.

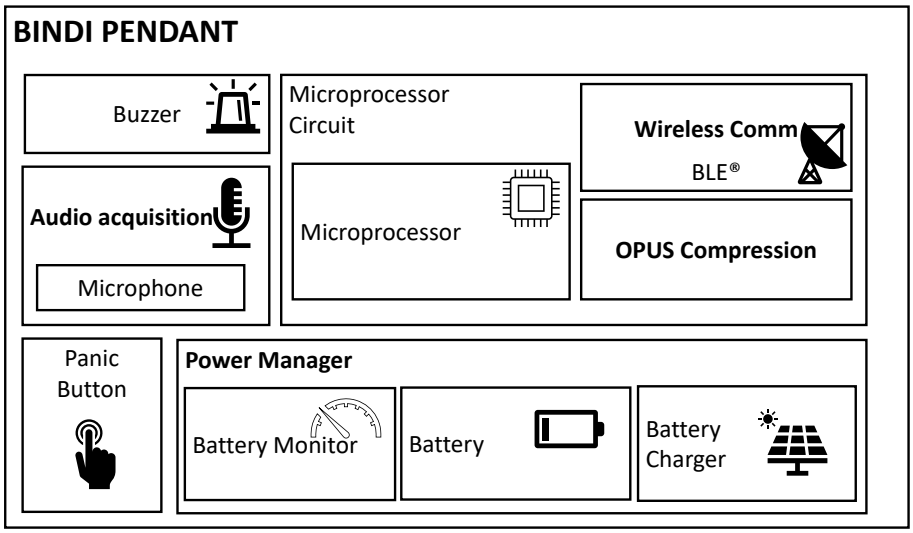


FIGURE 5.3: Simplified Bindi's Pendant Architecture [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Fog Computing

The fog computing within Bindi is represented by the Bindi App running on a smartphone. It provides an end-user graphical interface and performs the following technical functionalities:

- It requests physiological and auditory data from the bracelet and the pendant respectively, according to the data processing pipelines implemented.
- It handles the alarm triggers (SMS/Protection Unit or Emergency Services alerts) and logs them into the server based on the intelligent engine response or the manual panic button.
- It keeps track of each user's location using GPS.
- It manages secure communications with the server adapted by the current smartphone battery status.
- It collects and uploads auditory and physiological ciphered data to the cloud as evidence of an alleged crime if the alarm is triggered.
- It performs the feature extraction and inference processes for the auditory monomodal system. Moreover, it handles different data fusion strategies.

Cloud Computing

The cloud computing part is where the Bindi Server comes into operation. The Bindi Server implementation consists of a MongoDB[30] database and a NodeJS[31] web application server. This Bindi Server stores the information captured in the edge with three main goals. First, it serves as an activity monitor, indicating potential problematic situations regarding victims' long-term affective evolution for people supervising the well-being of the users. Second, it stores encrypted data, serving as digital evidence in an eventual trial. Third, it makes decisions after the alarms are triggered by following predetermined safety procedures.

[30] https://www.mongodb.com
[31] https://nodejs.org/es/

5.4 Multimodal Fusion Strategies for Bindi

Data fusion is a powerful way to improve the robustness of the multimodal intelligence engine in Bindi. Thus, this section proposes the data fusion architectures considered to strengthen the reliability and robustness of Bindi, so that physiological and audio data could be jointly considered in the decision to trigger the alarm.

We present an analysis of a multimodal late fusion strategy for combining the physiological and speech data processing pipelines to determine the best intelligence engine strategy for Bindi. Our goal is to analyze and gain a better understanding of women's responses to the fear emotion in risky situations.

5.4.1 Initial Cascaded Late Fusion: Bindi 1.0

In an initial approach in Bindi, both the speech and the physio-signals alert detection systems were fused by following a decision-level approach, also called late fusion. To this end, the authors considered a cascade approach in which the two systems run one after the other.

The system starts by running the physiological signals' system, which analyzes the data captured in the bracelet and decides if the user is in a dangerous situation or not. This physio-system is based on a KNN classification algorithm, which is run in the processor inside the bracelet. If the physio-signals system results in a positive detection, it communicates with the smartphone, which triggers a request to the speech system to analyze the current situation. The speech system captures audio data for a specific amount of time, which is sent to the smartphone with a previous compression process. In the smartphone, the audio data is analyzed and sent to an MLP model running in the smartphone's microprocessor. The prediction done by the speech system is then the global prediction reached in Bindi.

Thus, the physio-system acts as a trigger for activating the next stage in the cascade. This design decision was assumed because the energy cost of the bracelet capturing such physiological data, as well as the lightweight machine learning algorithms inside the processor, allows the device to work during hours (at least two days, at the moment of the publication of [9]). On the contrary, capturing audio data and comprising the information for sending it is costly, and then it should be reduced as much as possible. In addition, running the audio data analysis many times is also costly for the smartphone in terms of battery. For all these reasons, the speech system was decided to be in the second stage of the cascade.

5.4.2 Hybrid Fusion Approach

The work in this section is published in [9] together with other members of the UC3M4Safety team. The initial fusion architecture in Bindi previously discussed is done at the decision-level. This strategy is easy to implement, but it includes the disadvantage of not considering the possible relationships between the different modalities in the system, i.e., the possible correlations between physiological and auditory information. Moreover, another disadvantage is the heterogeneity among the confidence scores provided by the models from each modality. Before discussing other fusion architectures for Bindi some **key aspects** should be considered:

- Bindi is a distributed system composed of three devices, a smartphone and two embedded devices (a bracelet and a pendant). It means that communication between them is essentially required.

- Bindi is within a constrained cyber-physical system, meaning that both computational resources and battery are limited, especially for the two embedded devices. Focusing on battery life, data transmission consumes more energy than other usual tasks, as processing and sensing. Therefore, the less data is transmitted, the longer the battery of the devices will last.
- The initial decision-level architecture implies that signals from the two modalities are misaligned in time. Thus, the physiological signals which trigger the alarm are acquired before the audio recording.

Taking into account these key aspects, and in contrast to other methods for fusing physiological and vocal information through feature-level fusion that influenced this work [280], the authors propose a hybrid data fusion architecture by combining both the decision-level (late) and feature-level (early) approaches. As far as the authors know – at the moment of the publication of [9] –, this hybrid approach was never considered before for a multimodal physiological-audio wearable system.

The authors take two main design decisions for this hybrid architecture. First, the two embedded devices cannot perform the feature-level fusion due to constraints in computational capacity and battery. Therefore, the smartphone would be in charge of this task. Second, it is not possible to continuously send physiological and auditory information to the smartphone to perform the feature-level fusion and therefore it cannot take place at all times.

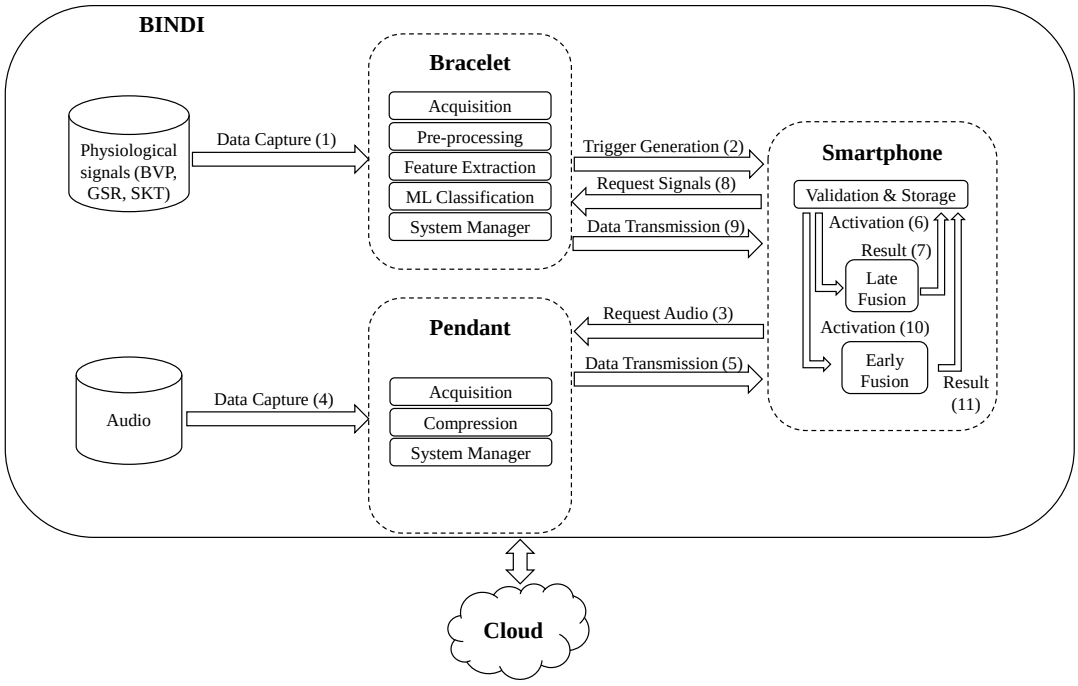


FIGURE 5.4: A Hybrid Data Fusion Architecture for Bindi 2.0 [9]. Reproduced with permission from the copyright owner, Springer Nature.

Fig. 5.4 shows the hybrid data fusion proposed for Bindi. By default, the system is performing the late fusion already included in the initial approach of Bindi 1.0 (see Sec. 5.4.1). It means that the bracelet is capturing physiological data over time (step 1). Then, the ML system in the bracelet analyzes the input data. In case it detects the targeted emotion, it generates a trigger to the smartphone (2). The smartphone requests the pendant (3) to capture audio data (4). The audio

information is compressed and sent to the smartphone (5). The smartphone runs the MLP based model (6), getting the response for the late fusion architecture (7). In case that the late fusion results in a positive detection, then the early fusion architecture is performed. In such a case, the smartphone requests the physiological data from the bracelet (8), which were captured in the past (i.e., the physiological data which generated the trigger in the late fusion and the one obtained during the time the pendant got audio data). After applying some compressing sensing techniques to alleviate the battery usage, the information requested is sent to the smartphone (9) that runs a classification algorithm combining both physiological and auditory data (10). The output of this early fusion architecture is the output of the whole Bindi system. Further processing is to be performed in the cloud [282].

5.4.3 Weighted Late Fusion Strategies: Bindi 2.0

Besides the hybrid fusion, the UC3M4Safety team also pondered other intermediate fusion architectures, and their validation and comparative is part of the team's roadmap.

The original Bindi architecture is based on a late data fusion strategy, which is executed following a two-layer cascade, where each layer has an intelligence model associated to each data modality – physiological signals acquired by the bracelet and audio and speech captured by the pendant, respectively –. The model in the first layer acts as a low-cost switch to activate a more demanding second layer, which is also related to a more powerful detection capability. This initial low-power strategy is useful for deciding when the more powerful and costly audio capture in the pendant should be carried out. However, the usage of the data captured in the bracelet only for switching purposes could imply that the intelligent decision engine is not considering all the information available.

In this section we propose three weighted late fusion strategies based on the literature (e.g., [277]) which are considered as a trade-off between low computational complexity and robustness considering the confidence of the system in the predictions [1]. These late fusion strategies are fed from the binary labels provided by the physiological and speech monomodal intelligence engines, as shown in Fig. 5.5.

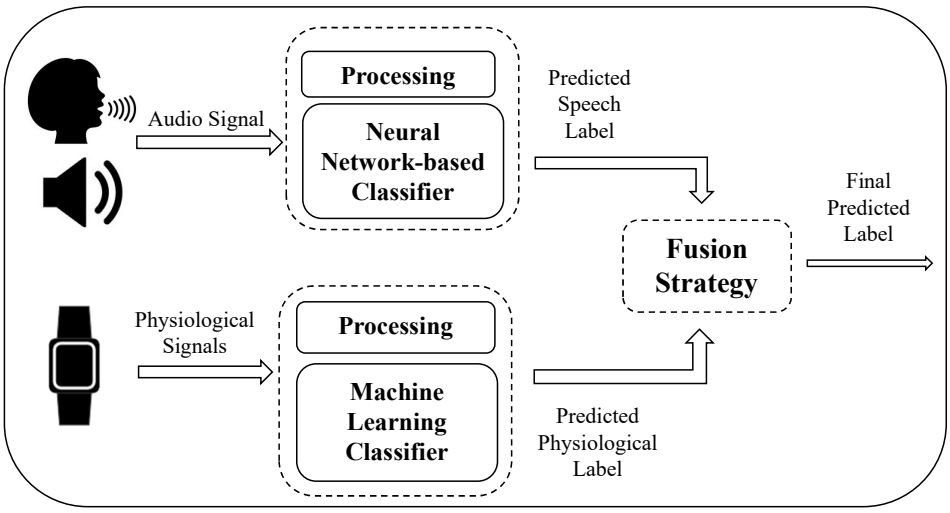


FIGURE 5.5: Bindi's Data Fusion Block Diagram [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

As discussed previously, the physiological and speech monomodal subsystems estimate a binary label, $y_k^m \in \{0,1\}$, for every time window k and modality $m \in \{\text{phy}, \text{sp}\}$, with phy and sp referring to the physiological and speech subsystems, respectively. Note that each of the modalities uses a different time window length, T_m, in seconds, due to their specific peculiarities. Bindi is intended to output a response per time period n (each one of same length L), with $n \in 1, 2, \ldots$, in seconds. Thus, an estimation of fear probability p_n^m for the n-th time period and the m-th modality is computed as

$$p_n^m = \frac{\sum_{k=1}^{K_m} y^m_{[K_m \cdot (n-1)+k]}}{K_m}, \tag{5.1}$$

where $K_m = \lfloor \frac{L}{T_m} \rfloor$, i.e., the number of time windows that we consider for each modality for the estimation of probabilities.

Thereafter, a single binary label, Y_n^m, based on p_n^m can be calculated as

$$Y_n^m = \begin{cases} 0 & \text{for } p_n^m < \text{th}_m \\ 1 & \text{otherwise} \end{cases}, \tag{5.2}$$

i.e., it will result in "1" (*fear*) if p_n^m is higher than the modality-related predefined threshold, $\text{th}_m \in \{0,1\}$, or "0" (*no-fear*) otherwise. Note that the th_{phy} and th_{sp} values are discussed in Sec. 5.6.2.

As a metric to represent how confident each monomodal system is for the class label predicted in a given period, entropy h_n^m for the n-th time period and m-th modality is calculated as

$$h_n^m = -[p_n^m \cdot \log(p_n^m) + (1 - p_n^m) \cdot \log(1 - p_n^m)]. \tag{5.3}$$

On this basis, three late fusion strategies are studied to produce fused system response Y_n^{f} for the n-th time period:

- Case 1, Lowest Entropy: The system's response corresponds to the binary label produced by the monomodal system with the smallest entropy, i.e., the most confident one. To this end, fused fear probability p_n^f for the n-th time period is calculated as

$$p_n^{\text{f}} = \begin{cases} p_n^{\text{phy}} & \text{if } h_n^{\text{phy}} < h_n^{\text{sp}} \\ p_n^{\text{sp}} & \text{otherwise} \end{cases}. \tag{5.4}$$

  Next, applying the same rationale as in Equation (5.2), a fused binary label is obtained as

$$Y_n^{\text{f}} = \begin{cases} 0 & \text{if } p_n^{\text{f}} < \text{th}_{\text{f}} \\ 1 & \text{otherwise} \end{cases}, \tag{5.5}$$

  where, for now, th_{f} is the conventional 0.5.

- Case 2, Inverse Entropy Weighted Combination: Fused fear probability p_n^f for the n-th time period is computed as a weighted sum of probabilities, as given by

$$p_n^{\text{f}} = \sum_m w_n^m \cdot p_n^m, \tag{5.6}$$

  where

$$w_n^m = \frac{1/h_n^m}{\sum_m 1/h_n^m}. \tag{5.7}$$

Next, a fused binary label is obtained according to Equation (5.5).

- Case 3, Logical OR: The system response corresponds to the logical OR computation over the binary labels for each monomodal system. That is,

$$Y_n^{\mathrm{f}} = Y_n^{\mathrm{phy}} \vee Y_n^{\mathrm{sp}}. \tag{5.8}$$

When comparing the three fusion strategies theoretically, the logical OR (Case 3) facilitates obtaining a fear class prediction without checking the subsystem confidence, which could lead to false detection. However, the lowest entropy strategy (Case 1) trusts the most confident model without considering the differences in the probabilities. Finally, the inverse entropy weighted combination (Case 2) establishes a trade-off between the probabilities and entropies for each monomodal subsystem. Thus, the confidence of this last strategy, Case 2, might be higher than that of the others.

5.5 Data Processing Pipelines

One of the key objectives of our work is to validate the data processing chain within Bindi, from data acquisition to alarm generation. Different arrangements of the system components have been applied and compared to achieve this goal. This fact has led to a design space exploration of different multimodal (physiological and auditory information) system architectures [1]. Specifically, three time arrangements have been evaluated:

1. The first version is Bindi 1.0 [177], which is based on a hierarchical or cascaded strategy. In this version (described in Sec. 5.4.1), physiological information is continuously collected by the bracelet, which runs a lightweight monomodal physiological intelligence engine. When it detects that the user is experiencing fear, it triggers a pre-alarm to the Bindi smartphone App. This action causes the pendant to start recording audio for a brief period, resulting in a low-energy consumption strategy for the microphone. The auditory signal is then sent to the Bindi App to perform fear detection using a speech-based monomodal intelligence engine. Finally, if the latter – speech – system confirms the detection, the Bindi App starts a safety procedure to help the user, triggering an alarm to the Bindi Server.

2. The subsequent version, Bindi 2.0a is based on the same two monomodal data processing pipelines in Bindi 1.0 but at the final decision stage applies a late fusion technique rather than a hierarchical agreement or confirmatory strategy [9]. It inherits the pre-alarm functionality from Bindi 1.0 for low-energy consumption for the microphone.

3. As a variation of Bindi 2.0a, Bindi 2.0b follows the late fusion scheme introduced in Bindi 2.0a but bases it on continuous physiological and auditory data acquisition, meaning that the pre-alarm functionality is not enabled.

The following subsections detail the physiological and auditory data processing pipelines. The physiological processing contribution is developed by other members of the UC3M4Safety team, the auditory pipeline is an own contribution as part of the research conducted for this thesis, and the fusion strategies are a joint contribution of the whole UC3M4Safety team.

The particular nature of the data types (physiological, speech) entails different challenges. Thus, the data processing schemes, methods, and feature extraction techniques are tailored to each signal.

Physiological Data Subsystem

In this section we will give a brief account of the physiological system in Bindi in order to help the understanding of the fusion system. For more details we refer the reader to [1] and [53].

The first physiological data processing stage is signal acquisition and windowing. In our case, the selected sampling frequencies are 100, 10, and 5 Hz for the BVP, GSR, and SKT, respectively. These frequencies are adequate to capture signal dynamics with the appropriate temporal resolution. For signal segmentation, an overlapping fixed-length strategy of 20 s windows with a 10 s overlap is used. This configuration provides a frequency resolution of 0.05Hz, which results in a good trade-off between the data storage and physiological information available to be extracted. Once the signals are captured and segmented, the filtering stage removes the out-of-band noise specifically for every signal.

Feature extraction block extracts the information contained in the physiological signals and is the next stage in the processing pipeline. Specifically, there are 25 features for BVP, 17 features for GSR, and six features for SKT. An extensive description of the features is provided in [261]. For classification, a lightweight K-Nearest Neighbors (KNN) binary supervised machine learning algorithm is used. During the training stage, cost-sensitive learning is applied by modifying the misclassification cost of KNN, which increases the sensitivity, i.e., the system will be less likely to omit a dangerous situation for the use case [283]. Finally, the physiological data subsystem output is a binary label every 10 s. This physiological pipeline has been tested in previous work using a public dataset [261].

Speech Data Subsystem

The speech data processing includes the following fundamental modules: Voice Activity Detection (VAD), frequency domain filtering, feature extraction, normalization, and a neural-network-based classifier.

A basic lightweight VAD module [284] based on spectral energy is employed to detect and remove silent parts of speech signals where the posterior feature extractor would not extract any relevant speech information due to the absence of it. Nevertheless, silence detection is crucial for correct functionality of the device, as women in dangerous situations frequently react with shock and remain silent, so it is intended to do more work in the characterisation of the silence in the future.

In combination with the VAD module, to ease the handling of the signals while keeping all significant information from the speech data, it is necessary to downsample the signals at 16kHz. Next, a low-pass filter is applied at 100Hz to remove low-frequency noise captured by the microphone and possibly caused by air-conditioning and electrical network buzzing, among other factors, as the databases we work with are recorded under laboratory conditions. Afterward, the filtered signals with a low-pass filter at 8kHz to maintain key information about speech and still maintain low complexity.

Then, the speech feature extractor computes 38 speech features dedicated to emotion detection using a 20 ms window with 10 ms overlapping, both of which are standard values from the literature. Among the features considered are pitch, Mel

frequency cepstral coefficients, formants, energy, and additional spectral features, all of which are calculated through the librosa Python toolkit [241]. The features are aggregated per second by computing their mean and standard deviation statistics to be later normalized. Preliminary ablation experiments are performed before fixing this 1 s aggregation, varying the temporal context of the aggregated speech features for 1, 5, and 10 s.

Feature normalization is done by applying the z-score mean and standard deviation values from the baseline features extracted when the user is in a resting or neutral state, named basal state normalization. Other normalization schemes (e.g., per video, per user, and traditional z-score) are informally tested before considering the basal state normalization described, but the latter was selected due to showing better performance of the system. The normalized aggregated features are fed into a user-adapted MLP neural network classifier trained for fear detection. This subsystem generates a binary label every 1 s. The labels predicted by the monomodal speech subsystem every second are smoothed in time using a 7 s window to maintain consistent and stable detection. Note that each of the modalities uses a different time window length in seconds, due to their specific peculiarities, which are fused using the different fusion strategies (see Sec. 5.4.3).

5.6 Experimental Set-up and Results on Stress and Fear Recognition

Within the field of speech-based emotion recognition we first perform voice-based stress detection experiments to assess whether acoustic events (which could define the acoustic context in which the user is) could help to detect stress in the auditory modality. Afterwards, and in order to validate the fusion strategies proposed in Sec. 5.4 we use WEMAC, a database captured by our UC3M4Safety team targeting GBV-related fear elicitation.

5.6.1 Experiments on Unimodal Stress Recognition

In this section we briefly explain our contribution on the experimentation carried out for stress classification of speech utterances, which could be understood as a fear-related emotion, in a preliminary stage previous to working with the WEMAC Database.

In our previously detailed study [8] in Sec. 4.5 we performed a speaker identification task in Bios-DB [157] and Biospeech+ (see Sec. 3.2.3), an augmented database with acoustic events based on Bios-DB. In this section we want to detail the emotions recognition task performed on the same data and its results. The methodology followed for this task is exactly the same as in Sec. 4.5, regarding the features extracted from the speech signals, and the classifiers used for the task. The main difference relies on the labels used, which now are two particular ones, i) binary labels referring to stress and neutral utterances, and ii) the reinterpreted emotions in the 4 quadrants of the PAD space as described in Sec. 3.2.2.

The results are presented in Table 5.1, where p represents the number of parameters of each model. MLP refers to the Multi-Layer Perceptron, K2D refers to the 2-dense layers model in Keras and KCGD refers to the Keras model composed of a Convolutional 1D, Bidirectional GRU and Dense layers. Mean and standard deviation results are shown for a 5-fold validation. For the two tasks under consideration, MLP with librosa achieves the best performance.

Model	librosa	*p*	eGeMAPS	*p*	yamNET	*p*	L+E+Y	*p*	feat sel	*p*
Binary Stress Recognition										
MLP	**89.1±0.9**	*12k*	65.4±1.8	*27k*	57.2±1.4	*307k*	75.3±1.7	*345k*	75.8±1.3	*111k*
K2D	82.4±1.0	*3k*	54.2±0.8	*5k*	32.7±9.0	*52k*	66.3±1.4	*58k*	65.1±1.2	*19k*
KCGD	80.9±1.8	*9k*	54.3±2.7	*12k*	30.4±5.6	*72k*	66.7±1.3	*80k*	67.2±1.3	*30k*
Speech Emotions Recognition (SER) 4-Q										
MLP	**90.0±0.9**	*12k*	45.5±1.1	*27k*	35.8±1.7	*307k*	59.5±1.0	*346k*	60.4±1.6	*112k*
K2D	73.2±1.0	*3k*	47.7±2.0	*6k*	37.6±1.0	*52k*	56.8±1.0	*59k*	57.8±1.2	*19k*
KCGD	73.2±0.9	*9k*	47.9±1.0	*12k*	37.6±0.9	*72k*	58.7±1.2	*80k*	56.9±1.7	*30k*

TABLE 5.1: F1-score results for Stress and Emotions Recognition in clean Biospeech [8].

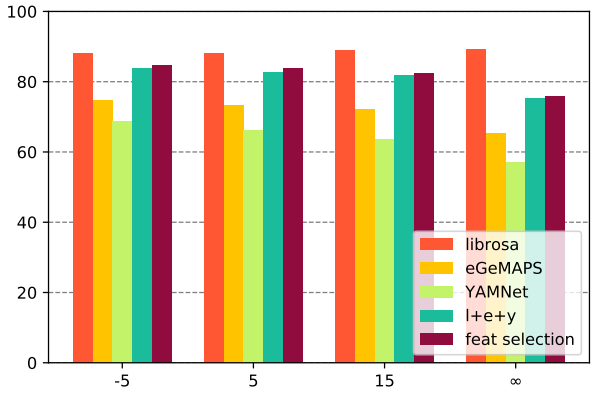


FIGURE 5.6: F1-score results for Binary Stress Recognition with Multi-Layer Perceptron in Biospeech+ [8]. Reproduced with permission from the copyright owner, ISCA.

Fig. 5.6 provides the results for different SNRs (on the horizontal axis) in Biospeech+ (see Sec. 3.2.3). Specifically, it shows the results for the binary stress labels classification for the model that performed the best (MLP). All the feature sets – except maybe librosa, which remains stable – show a trend to improve the F1-score as the SNR value gets lower, that is when the acoustic events overlay the speech[32] ratios (-5 and 5 dB) [33]. This demonstrates that extending our database with stressful events comes in handy for the recognition of stress in speech and audio. All the feature sets, in a greater or lesser extent, seem to be are able to capture information about the acoustic events which are considered stress triggers.

5.6.2 Experiments on Monomodal and Multimodal Fear Recognition using WEMAC for Bindi

In this section, we aim to validate and evaluate the different fusion architectures for Bindi for the task of *fear* recognition using WEMAC. This study is published in [1] jointly with other members of the UC3M4Safety team. This work is intended to be the first multimodal framework acting as a baseline to enable further work with real-life elicited *fear* in women. To the best of our knowledge, this is the first time

[32] For the SNR measure we consider the foreground speech from Biospeech as the 'signal' and the audio events as 'noise'.

[33] Note that the *'infinity'* symbol denotes the baseline for when no acoustic events are added to the database.

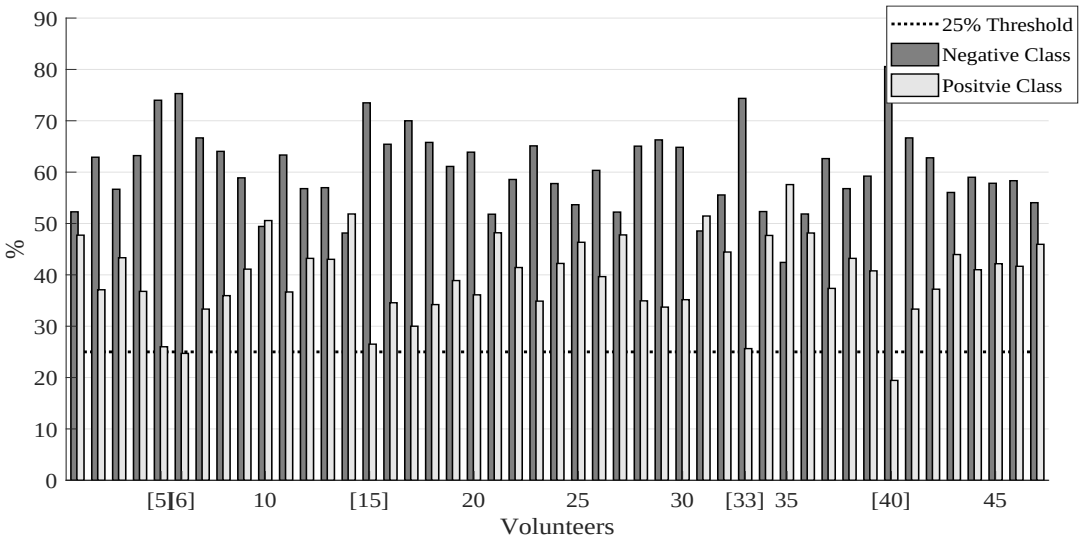


FIGURE 5.7: Statistical distributions of the positive and negative classes for the *fear*-binarized self-reporteded emotion labels in WEMAC. Volunteers in brackets are those excluded [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

that a multimodal fusion of physiological and speech data for fear recognition has been given in this GBV context.

We start working on the analysis of the labels. We first binarize the reported discrete emotions to each audiovisual stimuli by each user, to transform the modeling problem into a binary classification, where "1" (positive class) represented *fear* and "0" (negative class) any other emotion, turning the problem into a binary fear classification problem. It was observed that some particular volunteers presented a considerably unbalanced distribution in their self-reported labels, as shown in Fig. 5.7. Therefore, we decided to exclude volunteers $5, 6, 15, 33$ and 40 from the evaluation since they had only around 25% of the positive class distribution. Consequently, the evaluation was to be performed with only 42 of the 47 initial volunteers. The class distribution for these 42 volunteers was around 60% and 40% for the negative and positive classes, respectively. This distribution fits the information presented in Table 3.3 for the different emotions.

Note that the experimental results in this section are an account of the validation process performed offline, to evaluate the functionality of the data processing pipelines and fusion strategies, and later embed such modules in the architecture, balancing the trade-offs observed.

Considerations for the Experimental Set-up on the Monomodal Subsystems Training and Testing Stages

Some points had to be considered to design the training and testing strategies of the two monomodal subsystems. First, according to the WEMAC database design, it should be noted that physiological data were gathered during the stimulus visualization (and therefore, also emotion elicitation), whereas speech recording was registered during the subsequent speech annotation. That means that the physiological and speech data were not aligned in time in WEMAC. However, both data types had to be fused in Bindi 2.0b for every emotional reaction per user or experiment, unlike for Bindi 1.0 and Bindi 2.0a where the fusion was conditioned to the physiological pre-alarm (see Sec. 5.4 for the description of Bindi

fusion strategies). Therefore, we obtained a single p_n^m per experiment and modality, according to Equation 5.1; note that L is the length of the audio-visual stimuli for the physiological modality and the total length of the audio recording for the speech modality. During the labeling, the volunteers were requested to relive the emotions felt during the stimulus elicitation, so it was assumed that the correspondence was solid enough between both time instants. However, this assumption will need further validation when the rest of the subsets in WEMAC become available.

Second, for the train-test split, a LASO strategy was applied. This was a speaker-adapted subject-semi-independent (thus, subject-dependent) approach procedure for training the 42 models required, i.e., one per user. This approach was chosen due to the fact that the subject personalization provided by LASO is crucial for an emotion detection model such as ours [285]. Thus, each model was trained with all available data from the rest of the users and fine-tuned with half the instances of the subject to be tested, particularly, the data acquired from the first seven audio-visual stimuli (from a total of 14). The rest of the utterances of the last seven videos of the session were to be used as test samples. Thus, the test data were not seen during the training stage but some information about the subject was obtained by the model, as intended.

Third, regarding specific training particularities, for the physiological monomodal subsystem, the same mis-classification cost of 1.6 to the positive class to deal with the commented class imbalance was considered for all physiological models generated. This cost was fixed by an experimental parameter sweep. Moreover, the training was validated by a stratified k-fold cross-validation strategy, with $k = 5$. Finally, the normalization applied for the dataset was based on the z-score technique applied to the features extracted from all volunteers.

For the speech monomodal subsystem, the classifier consisted of a shallow lightweight neural network with input, fully-connected hidden, and fully-connected output layers. The network had 38 units in its input layer, i.e., one per feature. The number of hidden units in the dense layer was fixed to 250 to avoid largely increasing the computational cost but achieve fairly good prediction rates. The output layer yielded one predicted label as an output. All samples, except the ones from the user of interest, were used to train the model during 300 epochs, with early stopping after a 30-epoch plateau in the model validation loss, a binary cross-entropy loss function, using Adam optimizer, and a learning rate of 0.001. Then, samples from the user of interest (half of the ones available according to the LASO strategy) were used to fine-tune the model for a maximum of 100 epochs, with an early stopping approach (i.e., stopping after a 10-epoch plateau in the model loss). Regarding the z-score normalization used, the features extracted from the speech recordings of the sixth audio-visual stimuli were used as the baseline, as this video was expected to elicit a calm emotion and was assumed to evoke a neutral state in the user, so we used the aforementioned basal state normalization.

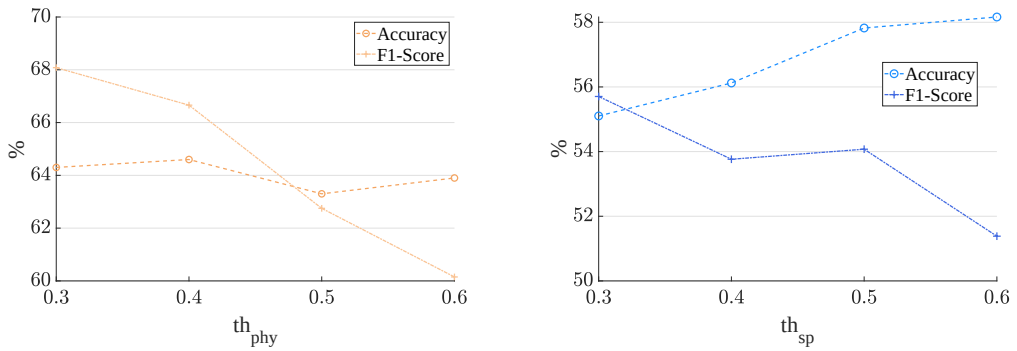
Finally, regarding the testing procedure, as discussed in Sec. 5.4, the monomodal subsystem's outputs were arrays of binary labels. Specifically for WEMAC, the length of the arrays was equal to dividing the duration of each stimulus by the monomodal sampling periods, i.e., 10 and 1 s for the physiological and speech subsystems, respectively. Afterward, those collected arrays were processed by calculating the probabilities and their corresponding binary labels by applying the physiological (th_{phy}) and speech (th_{sp}) thresholds. The data fusion strategies proposed also generated their corresponding binary labels, as described in Sec. 5.4. The evaluation metrics selected, i.e. the accuracy and F1-score, fed on the hard labels obtained. Accuracy could fairly represent the prediction rates since the class

imbalance was low, anyhow the F1-score was considered first to deal with the slight imbalance observed and second due to the higher importance of the positive class in our case of use, since the F1-score is a good metric for a detection problem in which the number of positives is lower in comparison with the negatives yet the detection of the positive class is crucial.

Results for Fear Recognition

This section presents the experimental results regarding the prediction of fear using WEMAC for the different configurations of the system discussed in Sec. 5.5. Note that this is the first time this database has been used; therefore, these results represent the first step toward real (non-acted) fear emotion detection from physiological and auditory variables for the problem of GBV and are meant as a baseline for future developments.

The first analysis concerns the performance of the physiological and speech subsystems working independently in a continuous setting, i.e., taking into account all samples. This experiment was essential to determine the thresholds, th_{phy} and th_{sp}, that convert the set of binary labels predicted during a video visualisation, into a single binary label for such period (see Equation 5.2). This step was relevant to determine whether the architecture was more or less prone to false alarms, regardless of the version of Bindi being considered. Thus, each parameter was swept in the range $[0.3, 0.6]$ with steps of 0.1 while generating the corresponding 42 monomodal subsystems following the LASO approach. In this regard, Figs. 5.8a and 5.8b show the th_{phy} and th_{sp} values versus the accuracy and F1-score average metrics for the 42 testing groups in the physiological and speech subsystems, respectively.



(A) (B)

FIGURE 5.8: Parameter sweep for the monomodal subsystems: th_{phy} in the physiological subsytem and th_{sp} in the speech subsystem [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Analyzing Fig. 5.8a, we observe how the F1-score decreases as th_{phy} grows, whereas the accuracy remains rather stable. Note that the F1-score depends to a great extent on the number of true positives (TPs) predicted but mostly disregards the true negatives (TNs). Thus, if TPs increase and the sum of false positive (FP) and false negative (FN) rates decrease, then the F1-score increases. This trade-off caused the behavior observed, where the lower the th_{phy} gets, the higher the F1-score becomes. According to this analysis, th_{phy} was fixed to 0.40, obtaining 66.66% and 64.60% for F1-score and accuracy, respectively. The reason behind choosing this value was the good compromise observed between both metrics and the fact

that missing a TP could be dramatic for the GBVV. The combined multimodal system should also refrain from triggering false alarms to avoid overwhelming the institutions in charge of protecting the users, and this is why the speech subsystem was chosen to be more conservative in this regard. Fig. 5.8b shows how the F1-score and accuracy began to diverge from 0.50 onward for the speech subsystem. Therefore, th_{sp} was fixed to this value, obtaining 54.07% and 57.82% for the F1-score and accuracy, respectively. Note that the accuracy could be increased by choosing a higher th_{sp}.

Once th_{phy} and th_{sp} were fixed, we studied the average performance prediction over the 42 testing groups for the different architecture configurations, as shown in Fig. 5.9. From left to right, the configurations are: physiological monomodal subsystem, the speech monomodal subsystem, Bindi 1.0, Bindi 2.0a with lowest entropy data fusion, Bindi 2.0a with inverse entropy weighting data fusion, Bindi 2.0b with lowest entropy data fusion, Bindi 2.0b with inverse entropy weighting data fusion, and Bindi 2.0b with logical OR data fusion. Note that Bindi 2.0a was not combined with logical OR data fusion because it is equivalent to Bindi 1.0.

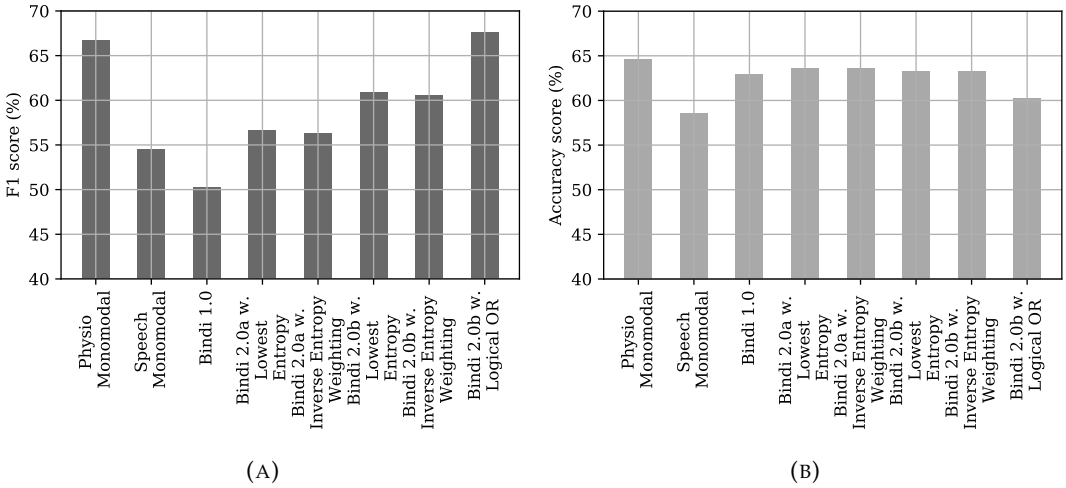


FIGURE 5.9: Average performance using the LASO strategy for the different architecture configurations: a) F1 score, b) Accuracy score, [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

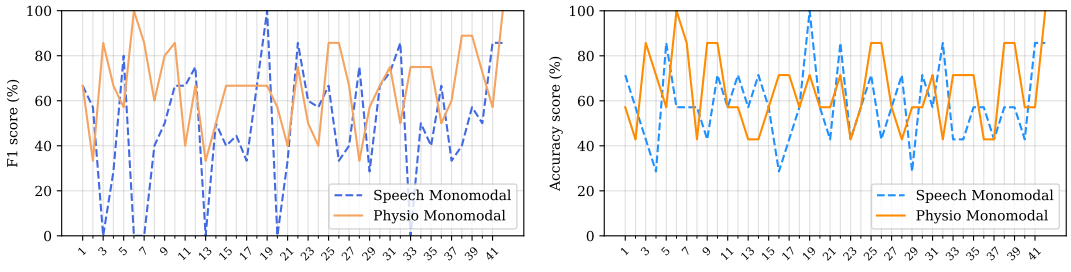
The physiological monomodal subsystem achieved the highest accuracy, a 64.63%, surpassing even the fusion schemes. For the F1-score metric, this subsystem also provided the second highest rate of 66.67%. This behavior could be first related to the bias introduced toward detecting the positive class with the misclassification cost of the classifier, and second with the parameter sweep of th_{phy}. The speech monomodal subsystem provided significantly lower metrics than the physiological subsystem. This fact could be related to the limited number of samples available to train the neural network and, possibly, some fading of the emotion felt when the samples were taken. This situation caused Bindi 1.0 to provide the lowest metrics since the final system response relies on the speech subsystem. Bindi 2.0a and Bindi 2.0b both provided similar accuracies close to those of the physiological subsystem in most cases. However, Bindi 2.0b achieved the highest F1-score in all cases, especially with the logical OR data fusion. This latter strategy provided the highest F1-score of 67.59%, although the accuracy was limited. This performance of the F1-score could be related to the positive bias contributed by the physiological

		Physiological Monomodal	Speech Monomodal	BINDI 1.0	Bindi 2.0a Lowest Entropy	Bindi 2.0a Inverse Entropy Weighting	Bindi 2.0b Lowest Entropy	Bindi 2.0b Inverse Entropy Weighting	Bindi 2.0b Logical OR
F1	mean	66.67	54.48	50.23	56.68	56.33	60.87	60.58	67.59
	std	17.31	26.73	27.64	23.91	24.05	26.63	26.98	14.27
Acc.	mean	64.63	58.50	62.93	63.61	63.61	63.27	63.27	60.20
	std	16.56	16.73	14.30	14.35	14.35	17.94	18.21	15.75

TABLE 5.2: Average performance analysis for binary fear recognition predicting over the 42 speaker-adapted subject-semi-independent testing groups [1].

subsystem due to the lower th_{phy} chosen, that introduced a conservative bias toward not missing TPs at the cost of increasing FPs. However, as for the other architectures with fusion strategies, the speech subsystem may have been slightly deteriorating the system performance in terms of the F1-score and accuracy but preventing Bindi 2.0a and Bindi 2.0b from producing too many FPs. Moreover, auditory information was expected to play an important role in detecting silences, which could mean that the user is in a state of shock caused by a GBV situation, and provide acoustic information about the environment. The meaning and consequences of these indicators over the real-life system performance should be thoroughly analyzed in the light of more robust metrics, such as in [286]. A short preview of this analysis and discussion of the confusion matrices obtained for each configuration can be found in Sec. 5.6.2.

To elaborate on the results shown in Fig. 5.9, Table 5.2 presents detailed results for the different configurations, including the average standard deviation per volunteer tested. Low standard deviation rates are good indicators of a better generalization ability as long as the results are comparable. Note for example that, although Bindi 1.0 presented the lowest standard deviation, which could be seen as a good generalization, its scores were surpassed by most of the configurations, as previously stated. Moreover, it can be observed that the standard deviation values obtained are relatively high, especially for the F1-score. The cause is shown in Fig. 5.10, where the F1-score and accuracy are provided for each of the 42 tests and monomodal subsystems. It can be noted that some volunteers had an F1-score of zero for the speech subsystem. This situation occurs because the F1-score depends on the TPs detected and there were no positive predictions for some users.

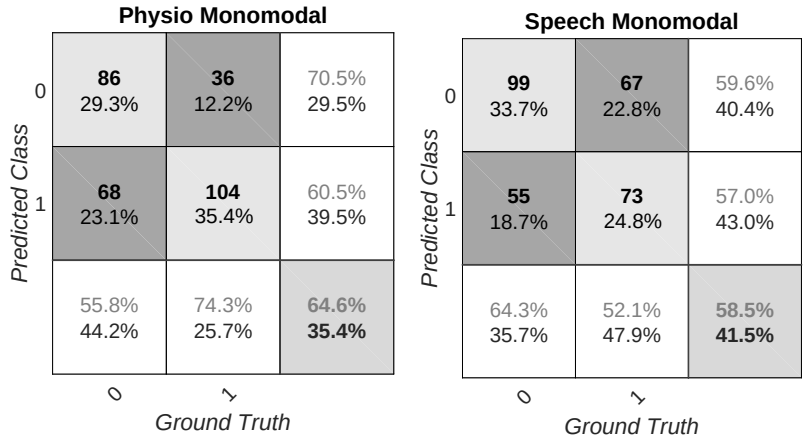


(A) F1 score (B) Accuracy

FIGURE 5.10: Individual performance analysis for binary fear recognition for the two monomodal subsystems [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Confusion Matrices for the Systems: Monomodal and Fusion

Fig. 5.11, 5.12, and 5.13 show the confusion matrices for the arrangements evaluated. In these figures, the rows correspond to the predicted class, and the columns correspond to the true class or ground truth. From left to right and from top to bottom, each confusion matrix shows the TN, FP, and false omission rates in the first row. The next row shows the FN, TP, and precision rates. The last row shows the FN rate, specificity, and overall accuracy.



(A) (B)

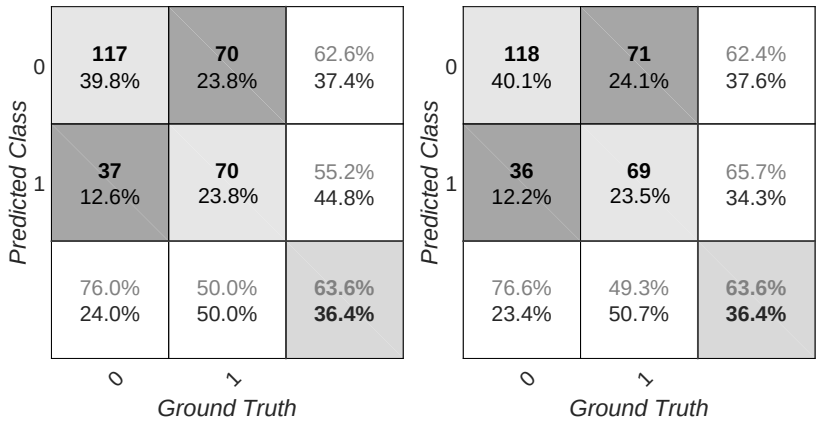
FIGURE 5.11: Monomodal confusion matrices for binary fear detection [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

The physiological subsystem confusion matrix reflects its tendency to predict the positive class at the cost of missing TNs. Meanwhile, the speech monomodal subsystem achieves lower overall rates than the others but achieves a higher TN rate. Finding a balance between these two behaviors is very important in our application, where missing alerts can be dramatic for the users, but triggering too many false alerts could overwhelm the institutions in charge of protection. Thus, the fact that the speech subsystem can hold back the FPs triggered by the physiological monomodal system looks very promising. In this line of work, the fusion strategies whose confusion matrices are shown in Figs. 5.12a, 5.12b, 5.13a and 5.13b differ only in a couple of instances but are more balanced between TNs and TPs. However, the strategy shown in Fig. 5.13c reflects much higher FP and TP rates than the others but misses more TNs than any other, and Fig. 5.12c shows how the hierarchical decision making of Bindi 1.0 performs poorly, proving that fusion is indeed essential.

Discussion

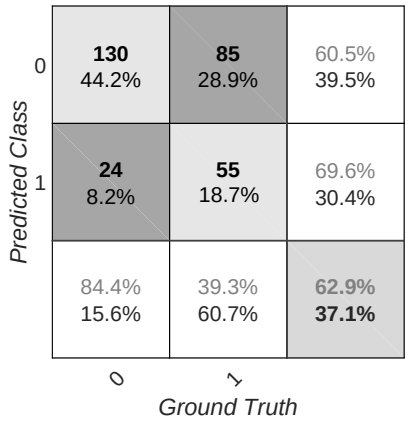
Regarding the usual IoT layer architecture (edge, fog, and cloud) considered in Bindi, a relevant system design question concerns which part of the system should be implemented in each of the layers.

First, the cloud computing layer is intended to collect and process great amounts of data without limitations regarding computing resources, energy demand, or response times [287]. This definition fits the needs of the centralized computing services of Bindi, which are therefore placed in the cloud layer to manage potential criminal evidence and historical information for victims' long-term monitoring.



(A) Bindi 2.0a. Lowest entropy fusion confusion matrix.

(B) Bindi 2.0a. Inverse entropy weighting fusion confusion matrix.

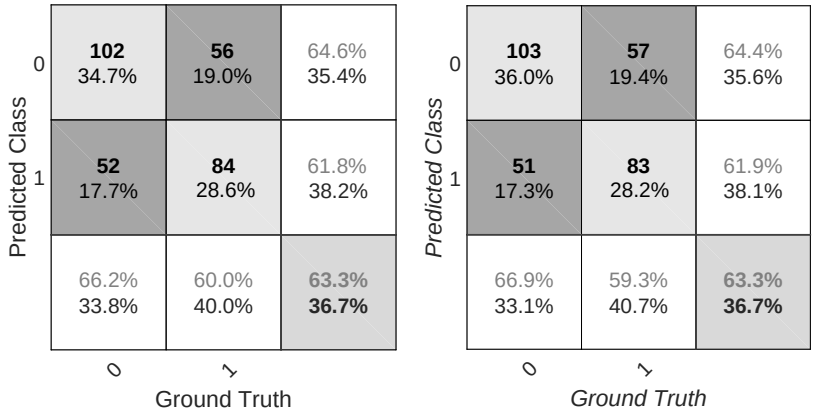


(C) Bindi 1.0 confusion matrix.

FIGURE 5.12: Confusion Matrices for Data Fusion Strategies for Bindi 2.0a and Bindi 1.0 [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

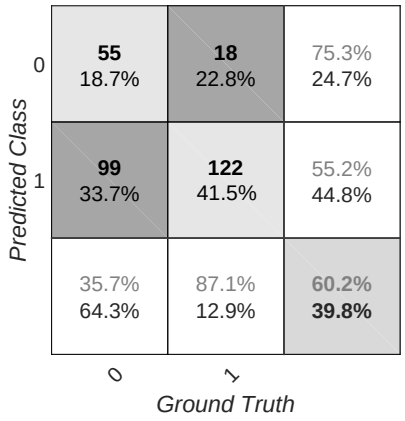
Second, edge computing takes place in the IoT nodes that capture data in the edge of the network. These devices are constrained by their computing and energy capabilities because, in most cases, they are powered by batteries or situated in hazardous environments [288]. This definition fits with the devices by which physiological and auditory data are captured over time in Bindi, i.e., a bracelet and a pendant.

Finally, the fog computing layer follows a concept similar to that of the edge computing layer. However, fog devices are less constrained in computing and energy capabilities while still remaining close to the data origin [289]. According to this description, Bindi's smartphone can be considered a fog device because it does not capture data but is close to the data origin, and both the computing and energy capabilities are less constrained than the ones in the edge devices (the bracelet and the pendant). Some authors assert that the fog does not exist, and then implement the fog layer functionalities described before, inside the edge layer [290]. Under this focus, it is still possible to structure devices in different layers inside the edge. From this point of view, the smartphone would be in an upper layer inside the edge, whereas the bracelet and the pendant would constitute the bottom layer. For further discussion about and review of the edge, fog, and cloud layers, readers are referred to [291] and [292].



(A) Lowest entropy fusion confusion matrix.

(B) Inverse entropy weighting fusion confusion matrix.



(C) Logical OR function fusion confusion matrix.

FIGURE 5.13: Confusion Matrices for Data Fusion Strategies for Bindi 2.0b [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

The proposed data fusion techniques in this work achieved a maximum of up to 63.61% average accuracy for a speaker-adapted subject-semi-independent fear recognition use case. This result was obtained using multimodal speech and physiological signals and the lowest entropy fusion strategy approach. The obtained average accuracy fell within the range of accuracy rates achieved by similar works presented in Sec. 5.2.4 and outperformed the system proposed in [280], which considered the same multimodal sources of information. It should be noted that as a differentiating feature of our system, we make use of non-invasive signal monitoring, rather than EEG headsets or face detection sensors [276, 257]. Additionally, the number of users considered (i.e., 42), provide more variability in the data and, therefore, produced a more robust model.

It is worth highlighting that the configurations described here for fear detection through physiological and speech data are just possible ways to characterize the situations and contexts in which Bindi users could be involved. These are meant as initial baselines for further developments and have allowed for the identification of important challenges. To start, finding a suitable trade-off between TPs and TNs and FPs and FNs is crucial since the cost of missing a true need for help is appalling, but we also need to avoid interfering with the everyday life of GBVVs and saturating

the protection services with false alarms.

Thus, in this work, we tried to reduce FNs as much as possible, while FPs were maintained at an adequate rate. To this end, we considered strategies based on misclassification costs and threshold parameter setting. Specifically, we fixed th_{phy} in the physiological subsystem to obtain a higher outcome of positive predictions with this system so that, in a later stage, the speech (in Bindi 1.0) and data fusion strategies (in Bindi 2.0a and Bindi 2.0b) would help in correcting the bias while trying to maintain the TP prediction. During this experimentation, the current speech monomodal system provided lower performance rates than expected. A possible explanation for this behavior could be the temporal misalignment of the physiological and speech data in WEMAC. The vanishing of the emotion elicited by the time the voice sample is collected could be behind this decrease in performance. Moreover, only classical processing and classification techniques have been used as a baseline for future exploration with this novel dataset. A similar situation applies to the fusion strategies, conceived to check the reliability of the pre-alarms triggered by the physiological model and acting as modulators to lower the FP class prediction rate.

Regarding future work, this study opens the door for further research in many directions. For example, the use of recurrent neural networks to exploit the temporal context of signals, the analysis of other fusion alternatives, or the evaluation of alternative score metrics, such as mutual information or area under the curve, could be used to continue finding a proper balance between false alarms and miss probability. Additionally, adding data acquired from more volunteers in laboratory conditions would add robustness to the models. Likewise, including GBVV data would help to better understand the GBVV activation mechanisms under fear-related situations. Lastly, it should be noted that the development of subject-adaptation techniques is critical for our GBV use case.

5.7 Conclusions

In this chapter we evaluated the detection and classification of fear-related emotions from a multimodal perspective for the development of the Bindi system. It should be noted the high level of multidisciplinary of the present work as the contributions were performed jointly with other members of the UC3M4Safety team.

In Sec. 5.3 we described in depth the components and functioning the system – bracelet, pendant, app and server –. Then, in Sec. 5.4, we described the proof-of-concept architecture evolution from Bindi 1.0 to Bindi 2.0. We analysed the monomodal data pipelines available and proposed a hybrid data fusion architecture by combining both the decision-level (late) and feature-level (early) approaches based on the combination of physiological signals and audio for detecting gender-based violence situations. This novel architecture includes a third layer (i.e., early fusion) to the beforehand implemented system in Bindi, still to be validated. Alternative fusion techniques need to be tackled, specifically tailored for this problem and able to account for its inherent limitations, such as the necessary bandwidth optimization including data compression, hardware and computational constraints, and battery consumption trade-offs. Later in Sec. 5.4 we further shape the fusion architectures for Bindi 2.0 and describe the theory behind each fusion architecture devised. We also detail each the data processing pipelines – physiological and speech – in Sec. 5.5.

Regarding Sec. 5.6.1, stressful acoustic events with a non-deterministic correlation to stressed speech utterances proved to be beneficial to some extent for the classifications of binary emotional utterances. This study leaves many open questions and future lines of work. The programming library used to create the synthetic mixes allows the definition of probability distributions for the appearance and duration of the sound events – as the procedure described in Sec. 3.2.3 –. And it is ready to perform the addition of background events when binary label is *Q2*. And so the data could also be extended by proceeding in a similar way with non-stressful events whenever binary label is not *Q2*, making the resulting mix sound more realistic. Also background sounds in the mixing process can be adapted to any kind of problem, resulting into new combinations of the BioS-DB and other datasets. As the main goal of Bindi is to detect and prevent Gender-based Violence, these background events could correspond to audio clips of movie scenes representing a GBV scenario, selected with expert knowledge and guidance.

Finally, regarding Sec. 5.6.2, we presented Bindi 2.0, an end-to-end autonomous multimodal system that leverages affective IoT throughout auditory and physiological commercial off-the-shelf smart sensors, hierarchical multisensorial signal fusion, and secure server architecture, with the final objective of providing safety for and ensuring the well-being of GBVVs. Specifically, Sec. 5.4.3 proposed three system architectures for Bindi, consisting of specific arrangements of the data processing subsystems developed, i.e., physiological, speech, and data fusion subsystems in the near future of Bindi. These architectures were validated and evaluated using the WEMAC dataset belonging to the UC3M4Safety Database. Note that the dataset was specifically built to detect fear in women in a laboratory environment.

The experimental results show an average accuracy of the fear recognition rate of up to 63.61% with the Leave-hAlf-Subject-Out (LASO) method. The obtained metrics are in line with similar multimodal-based state-of-the-art systems, such as the ones reviewed in Sec. 5.2.4. Moreover, our system outperforms the only system in the literature dealing with the same bimodal combination as in this work [280]. To the best of our knowledge, this is the first time a LASO model considering fear recognition, multisensorial signal fusion, and virtual reality stimuli has been presented. Note that the significance of the results is limited by the number of participants at the moment of the publication [1], i.e. 47 women.

This experimentation serves as an initial multimodal approach toward working with real elicited fear in women and its proper processing. Bindi is a very complex system that requires a thorough balance of many aspects, such as battery consumption, computational power, resource usage, and algorithm performance. We aimed to point out that the ultimate goal of this work is to ignite the community's interest in developing solutions to the very challenging problem of GBV.

All of this work in fear emotion recognition and the conclusions gathered are intended to pave the way and shape the next version of Bindi: Bindi 3.0.

Chapter 6

Additional Research Directions for Audio and GBV

At the same time that we carried out our research in this thesis, parallel but complementary lines of research opened up that could help in the prevention of gender-based violence using the auditory modality. In the beginning of this chapter, we talk about the affective characterization of the acoustic context, in the background of the detection of GBV risk situations, including first, the analysis of acoustic events and then the holistic analysis of acoustic scenes or scenarios. Afterwards, we superficially explore fatigue analysis on speech, seeing that it may be related to stress on speech. Next, we perform a preliminary ablation study on the detection of gender-based violence only through the use of speech utterances. And finally, the relationship between climate change and gender-based violence are broadly commented.

These lines of research do not form part of the bulk of the thesis, but we felt that they were important fields to investigate and that they could provide insight into and contribute to the prevention of gender-based violence through audio technology.

6.1 Affective Characterisation of the Acoustic Context

Within the audio signal that we capture with Bindi we have several sources of information, among them: the user's speech, the silences, environmental noise, acoustic events, auxiliary sounds, etc. By joining all of the sources together we can get an idea of the context in which the user is located in. In this thesis we have worked especially on speaker identification and the detection of emotions in the voice mainly, but it seems reasonable to think that acoustic events and noise, forming the acoustic context, could give us more information about the situation in which the user is. We are also interested in investigating the relationship between acoustic scenes and the emotions they can elicit.

We take into account this acoustic modality, because all the modalities alone are too brittle to give a reliable result of the prediction of a risk situation. So we can't look at just one but all of them contribute to a more robust and reliable prediction. This is also why it is difficult to isolate emotion detection or speaker identification from the analysis of the rest of the audio, and from each other. They are always intertwined under this challenge of gender violence.

The study of the characterisation of the acoustic events and scenes for the detection of gender-based violence risk situations is a very challenging and complex field, which would require a separate doctoral thesis itself, but we wanted to do some initial preliminary work that might light the way forward.

6.1.1 Affective Acoustic Events Characterization

The field of Acoustic Event Detection (AED) is a research field of AI in which different approaches have been developed and used for the detection of sound events, oftentimes imitating the human auditory system, and including different feature sets and detection algorithms. Sound detection can help us to emotionally characterise an audio signal, relating the acoustic events that appear with the emotion the audios tend to elicit.

Moreover, the DCASE community[34] has been releasing several datasets for the "detection and classification of acoustic scenes and events" since 2013. This has fostered a wealth of research contributions in this field. Moreover, a large-scale dataset of hand-crafted annotations of audio events, AudioSet [170], triggered the investigation in deep learning models, several opened to the research community such as YAMNet [242]. This offers a robust alternative for the representation of the acoustic environment that can be transferred to other domains and tasks.

In this section we introduce the acoustic event detection system proposed as a proof of concept for Bindi 2.0, to be included in Bindi 3.0.

Acoustic Information Subsystem in Bindi 2.0

This section describes the preliminary audio processing pipeline for acoustic scene threat detection developed by other members of UC3M4Safety team. We used it to provide an affective characterization of WEMAC in Sec. 6.1.2. This component has not yet been included in the arrangements studied in [1] and it is here conveyed as a proof of concept for further versions of Bindi. This subsystem is based on the architecture presented in [8]. Its main task is to detect whether the sound events recorded from the microphone represent a threat to the user's safety according to our use case.

The acoustic event detection system begins by processing the audio signal. First, the audio signal is normalized, just as for the speech pipeline (see Sec. 5.5). Second, a log-Mel spectrogram is computed to obtain a time-frequency representation of the signal in an image form to later feed it to the event detection network. Thus, an initial spectrogram is computed through a Short-Time Fourier Transform (STFT) with the following parameters: a window size of 25ms, window hop of 10ms, and Hanning window. The frequency dimension of the spectrogram is mapped to 64 Mel bins to cover frequencies ranging from 125 to 7500Hz and the amplitude is transformed into a log scale with an offset of 0.001.

The spectrograms taken as features are framed into examples of 0.96 seconds with an overlapping of 50%. Each example covers 96 frames of 10ms each and 64 Mel frequency bands. Therefore, the dimensions of these features are 96x64. The resulting features are fed into a pre-trained Convolutional Neural Network (CNN) to detect the audio events in a scene.

The selected model for this task is YAMNet. Specifically, the MobileNet_v1 [293] depthwise separable convolution architecture is considered. This model has been pretrained on 521 classes of the AudioSet YouTube corpus [170], a multilabel sound event classification database for general purposes, and is prepared to perform inference for the detection of acoustic events. The performance of these types of networks has been widely studied in the field of sound event detection [294].

The procedure to feed the network is as follows: First, the 96×64 patches from the feature extraction stage are transformed into a 3×2 array for the 1024 kernels

[34] https://dcase.community/

of the top convolutional layer. After being processed through the feature extraction layers, these examples are averaged to obtain a 1,024-dimension embedding. Then, a logistic layer performs the classification in 521 target classes.

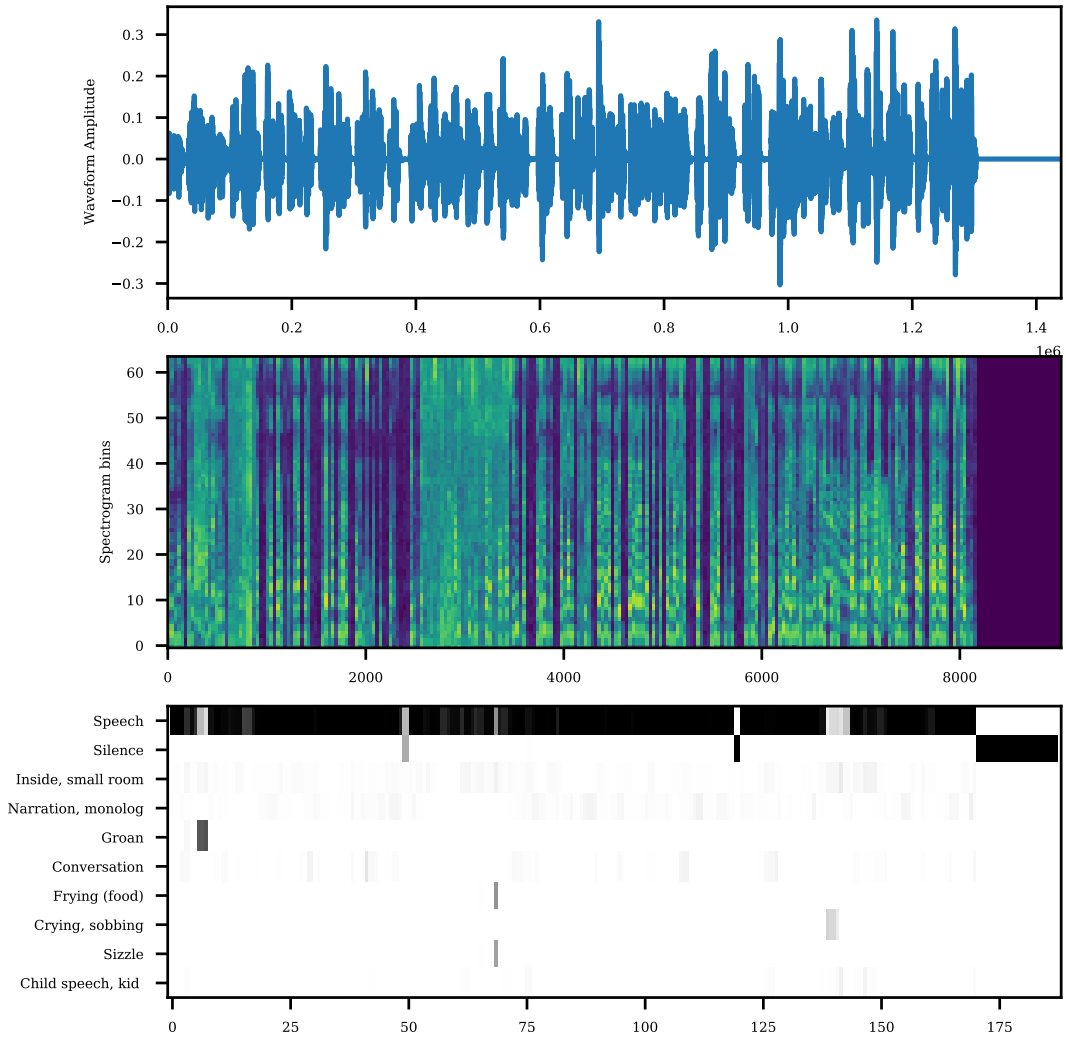


FIGURE 6.1: YAMNet processing a sample of BioSpeech+. Temporal representation (top), spectrogram with bands spanning 125 to 7500 Hz (middle), and principal events found (bottom) [8]. Reproduced with permission from the copyright owner, ISCA.

As for the Acoustic Events Detection and Classification (AED/C) task we are interested in analysing the set of stimuli used in WEMAC. In Fig. 6.1 we observe the performance of YAMNet classifying a 90*s* mixed audio as part of a small informal analysis for the identification of acoustic events present in one audio signal of the generated Biospeech+ database (see Sec. 3.2.3) [8]. The Biospeech+ dataset is pre-processed to match YAMNet's requirements ($f_s = 16KHz$, mono, amplitude normalized to $[-1, 1]$) and then fed into the model. The only free parameter is `patch_hop`, which was set to 0.48*s*.

This analysis aims to characterize the problem of GBV detection from an acoustics perspective since the development of an empirical description of the problem is important for its automatic detection. Thus, the acoustic information subsystem was applied to the audio signal of the audio-visual stimuli in WEMAC to analyse the acoustic events conforming each acoustic scene in the context of GBV.

The results obtained appear in Fig. 6.11, where all the occurrences of the YAMNet acoustic events labels in the audio-visual stimuli of WEMAC are depicted. Interestingly, some labels were exclusively found in *fear* audio-visual stimuli, such as *heartbeats*, *explosions*, and *breathing*, whereas other labels never appeared for fear, such as *tender music*, *lullabies*, and *crowds*. There were also intermediate cases in which labels appear for both types of stimuli, such as spatial-contextualization labels (indoors or outdoors-related), *animals*, *silence*, and *laughter*. Therefore, automatic classification of acoustic events seems to be promising as certain patterns can be deduced from extreme cases in which labels exclusively appear for one of the two types of audio-visual stimuli. It must be noted that YAMNet labels are very general themselves, i.e., they can appear to be related to many circumstances and scenes. Thus, they must be analyzed as a set, which is a feasible way to infer some qualities of the context of a particular scene, e.g., violence.

From this exploratory analysis, we can conclude that the information extracted from acoustic events can be very beneficial to disambiguate potential GBV situations detected automatically in Bindi with the rest of the sensors. The surrounding sound events of a scene can help infer its context, which is critical to determine whether the scene is or not violent. Thus, we expect the acoustic information subsystem to play a key role in the evaluation of WE-LIVE, where volunteers are performing everyday activities, outside of the laboratory environment.

6.1.2 Affective Acoustic Scene Characterization

After the analysis of acoustic events, in this section we present preliminary work carried out together with other members of the UC3M4Safety team in the study of acoustic scenes and emotional soundscapes. In the previous section we described the detection of acoustic events without actually relating one to one another, but in this section we want to analyse them together in order to be able to characterise an holistic acoustic scene in an affective way.

Acoustic Scene Analysis and Interpretation is a research field that aims to explain the acoustic information in the environment often captured by a multi-microphone acquisition system [295]. Although some work on the relationship between acoustic scenes and emotions exists in the literature, it has not been collectively identified or specifically defined. There is not a single title or acronym, as for example with the widely known field of Speech Emotions Recognition (SER) where a solid corpus of work is being developed. Thus, we found related work on acoustic scenes and emotions under different names: Assessments of Acoustic Environments by Emotions, Emotions in Soundscapes [296], Emotional (Acoustic) Scene Understanding, Induced Emotions in Sonification [297], Evoked Emotion Recognition by General Sound Events, Sound Design Theory [298] or Acoustic Design of Virtual Environments [299], among others. However, despite the limited number of works, there is still some promising research in the field. The motivation of such works in the literature is to provide machines with the ability of understanding what a person is experiencing from her acoustic frame of reference. This includes her acoustic contextual information, meaning the situation and auditory surroundings of the person. And our purpose with this work [4] is to provide an overarching view of this subfield, collecting it under the term of Affective Acoustic Scenes Analysis (AASA).

This innovative work [300] aimed to develop comprehensive computer models of affect in sound. In it, a high degree of coherence across domains indicated that the encoding of the two main dimensions of emotion (arousal and valence) resulted

from the evolution of voice and music together in a multimodal way, including combining nature sounds for expressive effects. However, these findings where established on the basis of acted and spontaneous emotional speech, music and general sound events [301] in isolation. Targeting a holistic model that is able to explain affect in sound, we aim to characterize the emotions elicited by being immersed in a specific *acoustic scene*, taking into account the acoustic information in the environment as a whole.

The common underlying representation of emotion triggers from sounds, music and speech is discussed in [300], but in spite of the abundant literature, pointing towards the relevance of the acoustic environment and human emotions in the cognitive sciences (e.g.,[302]), there are very few studies that investigate the relationship between acoustic events and the elicitation of emotions [301] and scarcely any investigate the relationship between *fear* and sounds [303].

As mentioned previously, we are specially interested in discussing how real-world acoustic environments can affect and influence emotions, and therefore, analyze and characterize them. These tasks could be encompassed in a subfield of Affective Computing that we term *Affective Acoustic Scene Analysis*.

To the authors' knowledge, no prior work focuses on the intrinsic emotional information of a soundscape and proposes a method to find direct and unsupervised relations between the audio events of an acoustic scene and its elicited emotion. So in the following subsection, we present a methodology for the *Affective Acoustic Scene Analysis* and then we adopt a setup based on information retrieval classical methods to produce a representation of the *affective acoustic scene* based on the well-known TF-IDF (term-frequency – inverse document frequency) algorithm [304], [305], where we build the vector space of acoustic events occurring in a scene balancing the *acoustic event frequency* and the *inverse scene frequency*.

Methodology for an Information Retrieval-based Approach

In this section we detail our proposed methodology for *Affective Acoustic Scenes Analysis* (AASA) step by step [4]. We put forward that this is a more comprehensive alternative to the classical machine learning setting that extracts features from audio signals and then plugs them directly into a machine learning model for inference, that also facilitates interpretability and accountability. Fig. 6.2 illustrates such methodology in a block diagram.

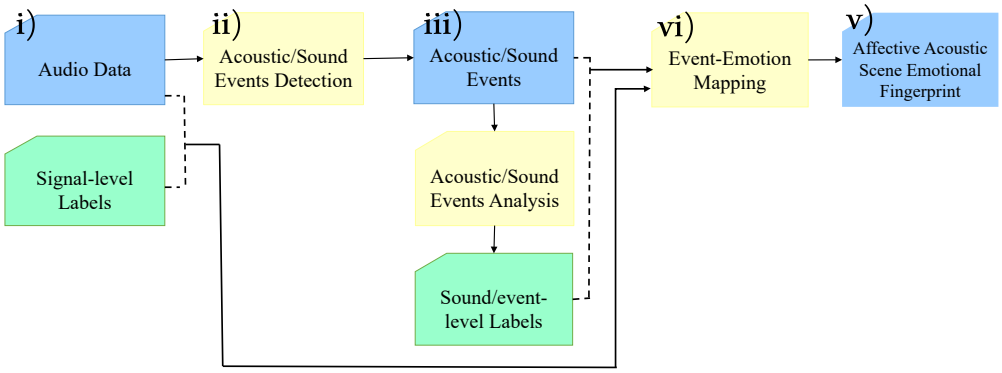


FIGURE 6.2: Block Diagram of Affective Acoustic Scene analysis methodology [4].

Starting on the blocks from the left, i) the first step is to use audio data, specially useful if recorded in realistic conditions or a synthetic mixture that imitates such (for example, a Virtual Reality Environment, movie clip, realistic video game, etc.). Ideally, such data would be labelled according to affective states or emotions perceived by the users that actively listen to it. Affect labels such as arousal, valence or dominance, pleasure, categorical emotions and liking, could be used. The aim of these labels is to reflect the emotion or affective state perceived by a person that is immersed in such acoustic environment.

As the next step, which could be optional, ii) an acoustic events detection or classification module can be applied, that identifies the acoustic events or sounds from the audio signal. Such module might be a pre-trained machine learning model with databases that include emotional labels from sounds, iii) so a relationship or alignment between the detected sounds and their emotional component annotated could be found.

Once the acoustic events or sounds have a corresponding emotional label, iv) the mapping between the two needs to be analysed, either in a supervised or unsupervised manner, with an algorithm that can evaluate the relation between the acoustic events or sounds and the emotional labels. This step can be performed with any pair or data-label, for instance, the separate acoustic events together with the whole original signal-level emotional label. Finally, v) an acoustic emotional fingerprint or embedding is extracted from the analysis, which condenses the emotional information from the analysed audio.

An important challenge that arises is related to the intensity of the emotional event, that is, its emotional saliency. It is is an adaptive biological signal that influences how events are remembered and how they are incorporated into memory.

Moreover, different sounds or acoustic environments can lead to different emotions elicited in the listeners based on their previous experience and the memory associations that the sounds evoke. Thus, there may be a majority emotional reaction, but we should not forget the individual differences in each person, specifically in the case of women who have suffered or are suffering from gender-based violence.

As we have already mentioned, the associative and relational memory component of an acoustic event can also play a role in the emotional reaction of a person. The sounds of keys opening a door may be a sound of joy meaning welcoming a loved person, but for a victim of gender-based violence it may mean that her abuser has arrived. The emotional effect can be completely different even though the acoustic event may be the same. Therefore, the need for a method that can be adapted and personalised is of paramount importance in this field with such a high level of subjectivity.

As part of the block diagram represented in Fig. 6.2, optionally we can aim to classify the acoustic events occurring in the audio data available. For such classification, we could employ pre-trained sound event classification models, able to detect acoustic events or sounds. We refer to this step as optional because a direct analysis of the complete acoustic signal and its emotional label could be also performed, but we believe this step to be key to identify the acoustic events composing an audio signal so that our later interpretation is more transparent and direct, more explainable.

The core part of the methodology is an algorithm that analyses the relationship between acoustic events or sounds and elicited emotions. Somehow, we have to extract from the audio signals the most salient or relevant moments, which allow us to condense the information, within the audio signal, that can trigger an emotion in a

listener. One way would be to extract audio features from the sound signal, as if the task were a speech emotion recognition task, for subsequent emotional classification using ML prediction models.

Another type of process that can be used and which is the one we use in our use case, is the TF-IDF algorithm, as we will explain in the next subsection.

Once we have extracted the acoustic emotional embeddings or fingerprints, they could be used as input for machine learning models. These can be supervised – re-using emotional labels as ground truth labels, as for ML regression or classification models – or unsupervised, using some kind of clustering or similarity metric to be applied. We consider it is also key that the results could be visualised, with the help of explainability models (XAI), to verify and interpret the accountability of the results gathered.

Experimental Set-up on UC3M4Safety Audio-visual Stimuli Dataset

To infer the emotions embeddings space, we use the UC3M4Safety Audiovisual Stimuli Dataset – see Sec. 3.3.1 – designed to collect the multimodal dataset WEMAC recently released [11] and specifically designed to portray the emotion of *fear*. By using the cosine similarity function, we find that the TF-IDF representation embeddings show the acoustic similarity of emotions as expressed in the dataset. Note that this emotional categorization is different (and could be complementary) to the classical Acoustic Scene Classification and Detection where scenes are typically related to the physical places to be characterized, e.g., airport, metro station or urban park.

In this research, 42 from a total of 79 videos of the UC3M4Safety Audiovisual Stimuli Dataset collection [11.1] – see Sec. 3.3.1 – are used to create a standard representation of acoustic information and events that induce certain emotions. Each stimuli lasts between $30 - 120$ seconds, and the collection consists of movie clips, ambience scenarios, and video compilations. In this subset from the first release, each video is assigned an emotion label by crowd-sourcing, corresponding to the emotion that it elicits in the viewers. Of such videos, 19 are categorized as *fear* and the 24 remaining are labeled with categories of other 9 discrete emotions.

The data we use for the work in this section is the audio component only, from the audiovisual stimuli collection. It contains different types of sounds – speech, music, sound effects – that, along with the visual information, induce in the viewers the labeled emotions. To identify the acoustic events occurring the audio data we employ a pre-trained sound event classification model: YAMNet [242].

We take the acoustic event labels predicted by YAMNet as words, and the audio stimuli eliciting emotions as documents, where our set of audio stimuli is equivalent to the collection of documents. We obtain a vector of TF-IDF scores per clip – with one value per acoustic event label – which represents the *affective acoustic fingerprint* of potential emotional triggers of each video.

TF-IDF (term frequency - inverse document frequency) [304] is a statistical method widely applied in Information Retrieval that evaluates how important a "word" is in a "document" in a "document collection". This importance is given by a score, which results of "multiplying two metrics: the number of times such word appears in a document (TF), and the inverse document frequency of the word across a set of documents (IDF)". The score increases proportionally to the number of times that a word appears in a document, but decreases when there is a high number of documents that contain such word. When theTF-IDF score of a word is high, then the more relevant the word is in that particular collection of documents.

These TF-IDF scores could be fed to machine learning algorithms as word vectors as they are a way of representing the data.

With the purpose of computing how similar each pair of TF-IDF vectors of each video of the UC3M4Safety dataset collection are, we use a similarity metric based the cosine distance (detailed in Eq. 6.1). Cosine similarity is widely used in information retrieval as a simple and effective way of providing a useful measurement of how similar two documents are likely to be, independently of the length of such documents. Thus, as our videos have different lengths, we rely on this distance to measure the similarity between the *affective acoustic embeddings* represented by the TF-IDF vectors.

Results on *Affective Acoustic Scene* classification

In order to perform the *affective acoustic scene* analysis we first extract sound events from the audio waveforms that allow us to characterize the acoustic scene with YAMNet. Our goal here is to obtain a corpus of weighted label scores that represent the occurring sound events per time window, so that we can later establish a metric that measures how close these representations are within the gathered video stimulus of the UC3M4Safety Audiovisual Stimuli Dataset. In this section we are concerned with the construction and evaluation of the vector space of acoustic events and the vectors that represent the directions of the different emotions. Thus, the following pipeline has been applied and is publicly available on GitHub[35].

Each of the 42 videos from the collection elicits one emotion, validated by more than 50 users each[36]. Both the acoustic and the visual modality are the ones inducing these emotions, thus, we first extract the audio only with the command-line tool *ffmpeg*. Apart from speech, these audios also contain information about the acoustic scene that induces such emotions. It is the acoustic scene and context what we would like to further analyze.

At the preprocessing stage, we have used the audio information subsystem of Bindi 2.0 which were already described in Sec. 6.1.1. Next we use YAMNet to detect and classify the acoustic events present in the audio signals of all the video stimuli.

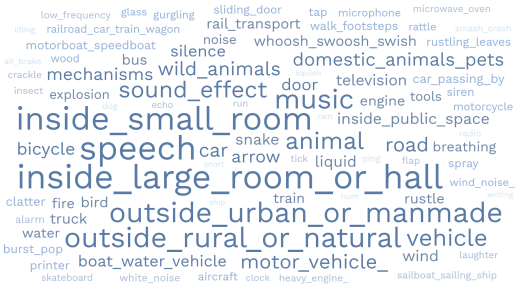


FIGURE 6.3: Word Cloud of acoustic labels output by YAMNet for audiovisual stimuli annotated as 'Fear' [5]. Reproduced with permission from the copyright owner, ISCA.

As YAMNet is a general sound event classifier, it may produce very detailed class labels which may not provide useful information to our acoustic characterization given the audiovisual stimuli used, but would only make the task and descriptions more complex. So considering the Audioset Ontology, the children classes of Music and Animal labels are filtered out, except the classes of Music Mood and Wild

[35] https://github.com/erituert/acoustic_information_retrieval

[36] The emotion chosen was the one that most annotators chose

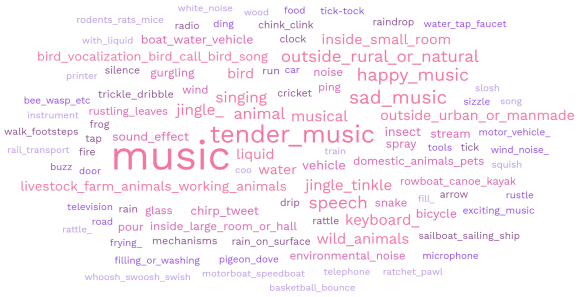


FIGURE 6.4: Word Cloud of acoustic labels output by YAMNet for audiovisual stimuli annotated as 'Tenderness' [5]. Reproduced with permission from the copyright owner, ISCA.

Animals, where all subclasses are kept. From the total of 521 classes that YAMNet classifies, the filtered ones result in 351. Figs. 6.3 and 6.4 represent the word cloud of acoustic labels output by YAMNet for audiovisual stimuli annotated as 'fear' and 'tenderness' respectively. As an example of a video chosen to elicit 'fear' as analysed through YAMNet. Fig. 6.5 represents an audio signal: at first a woman speaking can be heard, then a strong noise similar to a squeal followed by an engine sound occurs at second 18. Heartbeat sounds are present during the last part of the audio.

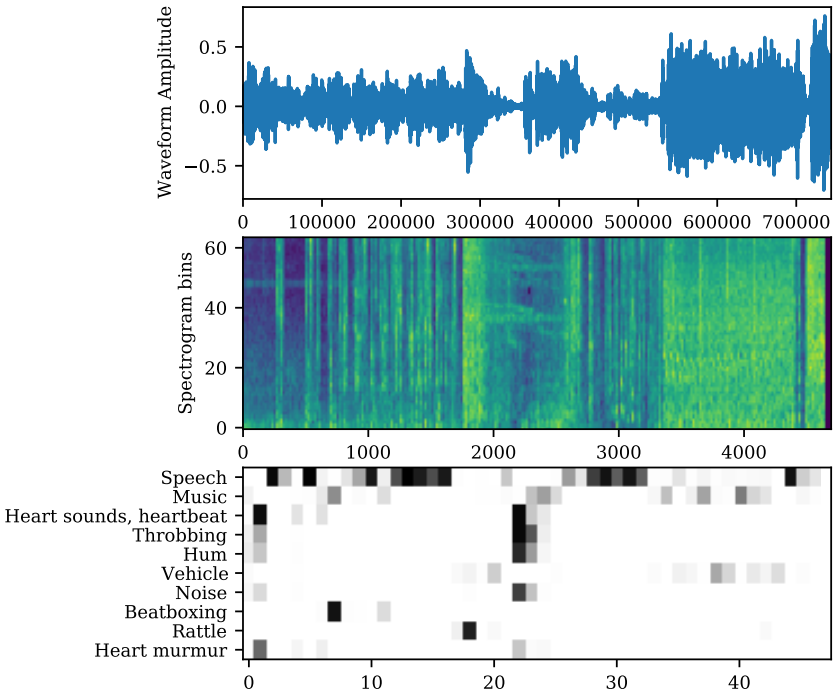


FIGURE 6.5: YAMNet processing a sample of UC3M4Safety Audiovisual Stimuli Dataset. Temporal representation (top), spectrogram with bands spanning 125 to 7500 Hz (middle), and principal acoustic events found (bottom) [8]. Reproduced with permission from the copyright owner, ISCA.

For correct comparison purposes, since the scoring ranges from YAMNet can be extremely low, all the scores are log-scaled and then binarized. The goal of the binarization is to keep only the events with an output score high enough to consider that they have occurred and are not a misinterpretation of the network. Therefore,

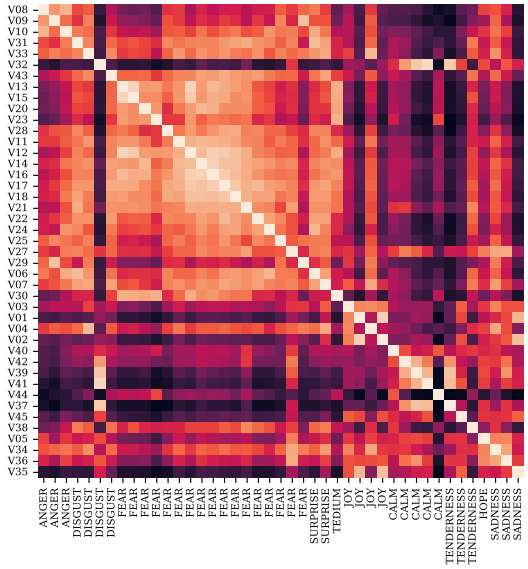


FIGURE 6.6: Original heatmap of cosine distance similarity between affective acoustic embeddings, sorted by emotions [5]. Reproduced with permission from the copyright owner, ISCA.

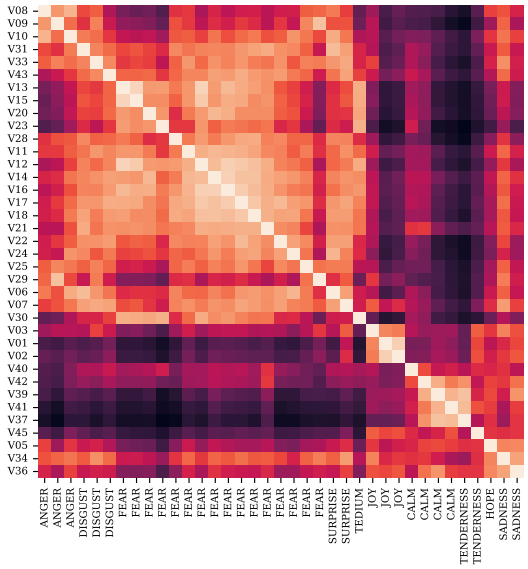


FIGURE 6.7: Heatmap affective acoustic embeddings sorted by emotions after removing outliers [5]. Reproduced with permission from the copyright owner, ISCA.

the threshold is set to the global mean value among all scores of all files, and sound events whose score is lower than the threshold are not considered.

The next step in the pipeline is to obtain a text-form corpus of the occurring events in the dataset. Thus, each audio file is treated as a text document composed by terms, which are the words of each sound event referenced through the internal identification code provided in the Audioset database (*mid*).

Finally, we use the TF-IDF [37] algorithm from the *sklearn* Python library to obtain the TF-IDF matrix for each of the 42 audiovisual stimuli in the dataset, resulting in a scores matrix of dimensions (42, 351).

With the aim of analysing the distance between the TF-IDF vectors or *affective acoustic embeddings* representing each of the instances in the dataset to understand the underlying patterns that relate the emotions, we use a similarity metric based on the cosine distance:

$$similarity(\vec{x}, \vec{y}) = 1 - cos(\theta) = 1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \tag{6.1}$$

where θ is the angle between the two vectors.

In Fig. 6.7 we represent as a heatmap the results of computing Eq. 6.1 for each audiovisual stimuli with its labeled emotion with respect to the rest of audiovisual stimuli, with a total of 37, after removal of outliers (which were still present in Fig. 6.6). Lighter colours on the heatmap represent higher similarity, and darker colours

[37] https://scikit-learn.org/stable/modules/generated/sklearn.feature_\extraction.text.TfidfVectorizer.html

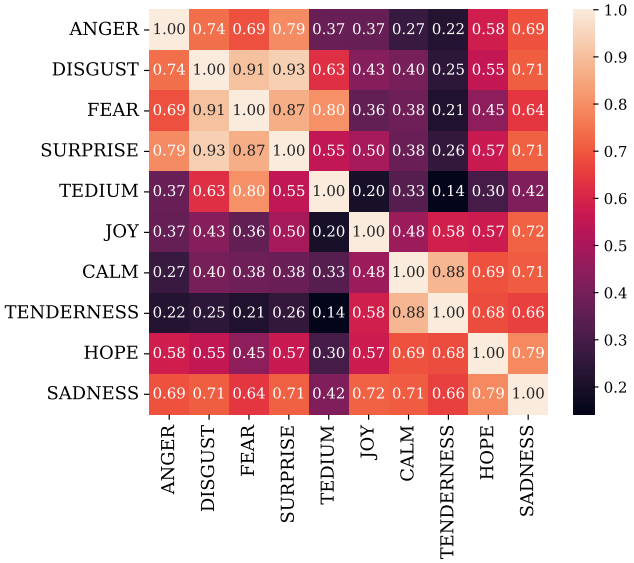


FIGURE 6.8: Heatmap of cosine distance similarities between emotion embeddings [5]. Reproduced with permission from the copyright owner, ISCA.

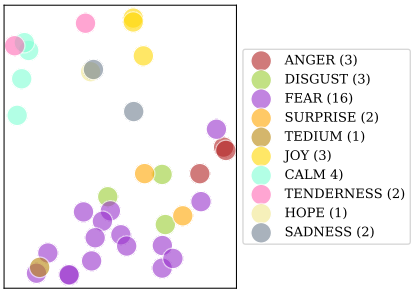


FIGURE 6.9: t-sne representation for tf-idf audiovisual stimuli embeddings [5]. Reproduced with permission from the copyright owner, ISCA.

show lower similarity, between *affective acoustic embeddings*. As each video aims to trigger a single emotion, in this manner we can understand how each video is related to the way the rest of them represent each of their corresponding emotions.

Outlier detection and removal was performed after comparing each affective acoustic embedding with the rest of embeddings of the same emotion category. For instance, $V32$ was identified as an outlier considering that its embedding had a big dissimilarity with respect of the rest of embeddings labeled with *disgust*. Further analysis reveals that the acoustic context does not match the visual information, since $V32$–which is a video compilation–contains mostly classical music, similar to videos labeled in the *calm* category, and therefore its embedding is similar to these later emotion embeddings.

We can observe that similar emotions present alike coloured clusters in Fig. 6.7, meaning that videos labeled with the same emotion have a similar acoustic characterization. On Fig. 6.7, four clusters can be roughly observed: a big cluster including *anger*, *disgust*, *fear*, *tedium* and *surprise*, another cluster for *joy*, and another cluster for *calm* and *tenderness*, and the last one including *hope* and *sadness*. These four groupings are to some extent consistent with the similarity in the PAD space on the Valence and Arousal axes [107] of these emotions.

Afterwards, we performed the mean of the TF-IDF matrix for every audiovisual stimuli labeled with the same emotion category. In that manner we can understand

how each acoustic label impacts in the classification of each emotion. In Fig. 6.8 we present the resulting heatmap, from the acoustic point of view, of emotion embeddings. We can observe how the results are promising, as similar emotions present a greater similarity between them (e.g., *calm* and *tenderness*), than emotions that humans categorize as more different (e.g., *tedium* and *joy*). In particular, the *fear* category lays close to the *disgust* and *surprise* labels, which hinders the discrimination between them if we only take into account the acoustic context.

Discussion

In Fig. 6.9 we plotted the affective acoustic embeddings using the t-sne algorithm. We can observe that the distances and clustering between them are somehow similar to the grouping happening in Fig. 6.7.

The relationship between *fear* and *anger* is peculiar, as contrary than what we would expect present a great similarity. This could be explained taking in to account the gender bias [126], that states that in certain situations, people can feel different emotions to the same stimuli depending on their gender. This deserves further investigation.

Two factors may be influencing the robustness of this analysis, first the agreement among the annotators that labeled each video and their gender, and second, the amount of videos per each emotion category. Thus, as future work, a more insightful analysis with a more in-depth study could be carried out using the original set of videos – up to 79 – or other databases of acoustic scenes with emotional annotations. The annotators agreement per gender as a variable can also be taken into account to study its relevance. Furthermore, using the TF-IDFvectors as features, machine learning models could be fed with such data and predict emotional labels in supervised learning.

As a final note, we work to try to answer the question of whether it is possible to characterize an acoustic scene or soundscape with respect to the emotions it elicits. We draw from the premise that characterizing the affective acoustic scene involves taking into account the acoustic context. And regarding the results presented achieved, we seem to have achieved a favourable emotional characterization of the acoustic scene in audiovisual material, being a first start to *affective acoustic scene* analysis in real-world environments.

We conclude that using the Affective Acoustic Scene analysis methodology is a promising method for which the results can be highly interpretable, for characterizing an acoustic scene with respect to the emotional information. Robust embeddings that acoustically characterize emotions can be used to measure the emotional load of – or the emotion to be elicited by – the acoustic information in other databases.

Other indicators besides the acoustic context – such as information from other modalities (e.g., bio-signals from the subject) – are crucial to accurately characterize a situation and to detect if the life of the user is at risk.

6.2 Intersectional Fairness Analysis on Fatigue Classification

In line with the detection of stress in the voice, we also published a study on the detection of fatigue through the voice and breathing in speech signals [7], in a joint collaboration with the University of Augsburg's Chair of Embedded Intelligence for Health Care and Wellbeing (EIHW). This study had two objectives: first, to

understand the fatigue or stress that can be caused by running or jogging exercise in order to study and characterise it, developed by members of EIHW; and second, to carry out a gender analysis in how this fatigue is observed in each gender, the part we were in charge.

We model the Borg Received Perception of Exertion (RPE) scale [306], "a well-validated subjective measure of fatigue", by means of audio signals that were captured in real outdoor environments by placing a smartphone attached in runners' arms and using machine learning models. By fine-tuning (pre-training) a convolutional neural network (CNN14 [307]) on log-Mel spectrograms, researchers at EIHW performed subject-dependent experiments and obtained a mean absolute error (MAE) of 2.35, showing that audio can be acquired more easily and non-invasively than signals from other sensors, while being effectively used to model fatigue.

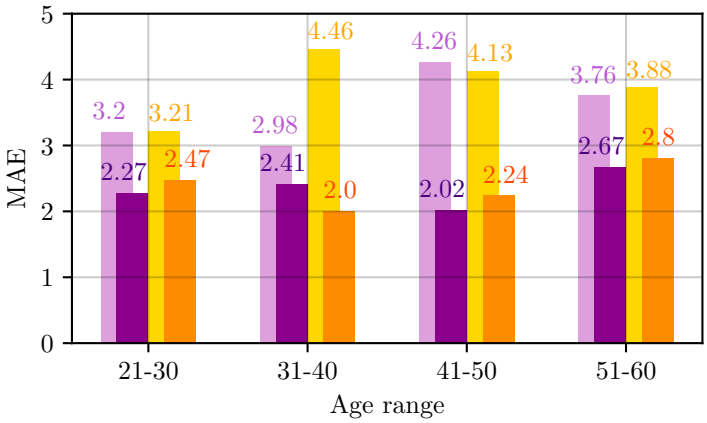


FIGURE 6.10: Results in terms of MAE stratified for Age and Gender on KIRun data. Orange color refers to female and Lilac refers to male. Dark colours refer to CNN14-pretrained and light colours refer to CNN14-random [7]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Among the results derived from this study, we performed a gender analysis, in which the model performs almost equally for male and female runners. This is despite the fact that the used data (KIRun database [38]) is biased towards females (27 female runners vs 21 male). This indicates that relative data quantity may not be the only factor causing performance imbalances. Finally, we noted that while CNN14-pretrained performance was overall better for females in most age groups, the reverse is true for the age group $(21-30)$, where females show an MAE of 2.37 compared to 1.86 for males.

Another interesting pattern is the difference of behaviour between CNN14-random and CNN14-pretrained. For some particular age groups, the two models show very different behaviour. For example, for the age group $(31-40)$ CNN14-random shows a much larger MAE for females, but the performance of CNN14-pretrained is almost the same for both sex groups. This shows that pre-trained models not only improve absolute performance, but might also change model behaviour across different strata of the dataset – which is

[38] https://www.uni-augsburg.de/de/fakultaet/fai/informatik/prof/eihw/forschung/projekte/vergangene-projekte/

an unwanted side-effect of the underspecification phenomenon observed in ML architectures [308].

To conclude, the intersectional fairness analysis performed reveals that performance differs between age groups and sex combinations, and that individual-level performances are important. This conclusion should also be recognised for the fear recognition systems, and it would be ideal to perform further work that includes data from these different but also similar affect states – fatigue, stress and fear – in order to separate them and analyse their differences, so that Bindi can be trained to detect the right conditions and not to misclassify them.

6.3 Automatic Detection of Gender-based Violence Condition in Speech

In the preliminary work based on [309] and published in [6] jointly with other members of UC3M4Safety team, we explored whether the GBVV condition could be detected from audio only by a small set of features from speech paralinguistic cues from the WEMAC Database [11]. The work in [309] addresses the use of feature selection techniques for features extracted from speech paralinguistic cues, and from that basis the present classification was made.

The data used comprises 26 non-GBVV and 26 GBVV from the same age ranges. The feature extraction process is coded in Python[39] and includes the features presented in Sec. 3.3.2. Statistical tests were applied as feature selection methods between both groups (GBVV and Non-GBVV) to check if there were any speech features that presented significant differences between groups, and thus allowed for distinction between them (further details in [309]). The statistical analyses conducted led to the use of different sets of features for its subsequent classification (see [6]).

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. Two approaches were implemented: a subject-dependent and a subject-independent strategy. The results 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. When removing the dependency, the scores diminish significantly, which we believe could be explained by the existence of outliers and the small amount of data [6]. Nevertheless, we need to take 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 52 subjects in total. Part of the planned future work is to re-perform this analysis with a bigger sample once it is available.

6.4 Climate Change and Gender-based Violence

In line with the seventeen Sustainable Development Goals (SDGs) planned and adopted by all the United Nations member states in the 2030 Agenda, the 13th SDG is "a call for action to combat climate change for a better world", and here we have briefly explored the literature linking gender-based violence and climate change.

[39] Available in: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing

The largest and most comprehensive study of the topic to date wasconducted by the International Union for the Conservation of Nature (IUCN), involving more than 1,000 research sources over two years in Climate Change and Gender-based Violence [310]. The study suggests that gender-based violence is increasing due to climate change, because increase in environmental degradation and stress on ecosystems creates scarcity of resources, which in turn creates stress for people. So, where environmental pressures increase, gender-based violence increases.

This study [311] shares an extensive plan to device Gender Based Violence Index (GBVI) in identifying the severity of abuse in relation to air pollution and vegetation coverage, such as finding a link between air pollution and green canopies with levels of aggression. There seems to be a correlation between climate change and gender-based violence, and it may be that helping with one will reduce the cases of the other.

Within the framework of this thesis, audio technologies have been used for the detection of risk situations for women in the context of GBV, and it is this same technology – computer audition – which can also be used to tackle the problem of climate change. In this work [12], which is a joint collaboration with the University of Augsburg's Chair of Embedded Intelligence for Health Care and Wellbeing (EIHW), we provide an overview of areas in which audio intelligence – a powerful but in this context so far hardly considered technology – can contribute to overcome climate-related challenges. We categorise potential computer audition applications according to the five elements: *water*, *air*, *fire*, *earth* and *aether*, proposed by the ancient Greeks in their five element theory. This categorisation serves as a framework to discuss computer audition in relation to different ecological aspects. *Earth* and *water* are concerned with the early detection of environmental changes and, thus, with the protection of humans and animals, as well as the monitoring of land and aquatic organisms. *Air* refers to aerial audio, which can be used to monitor and obtain information about bird and insect populations. Furthermore, acoustic measures can deliver relevant information for the monitoring and forecasting of extreme meteorological phenomena. Finally, the element *fire* deals with the automatic audio-based detection and classification of wildfires as well as the assessment of structural damage caused by fire. This work positions computer audition in relation to alternative approaches by discussing methodological strengths and limitations, as well as ethical aspects. We conclude with an urgent call to action to the wider community in order to collectively fight climate change.

6.5 Conclusions

In this chapter we have indicated some research lines that emerged while researching this thesis. These are highly interesting preliminary works in which we have collaborated together with other members of the UC3M4Safety team and the Chair of EIHW. Indeed they require more in-depth work in the future since their outcomes are encouraging.

We detail the work carried out in the field of Acoustic Events and Scene Analysis and the importance of audio events analysis for the detection of risk situations. From it arises the term of *Affective Acoustic Scene Analysis* and with it we call for future research under this perspective and denomination. In our study of *Affective Acoustic Scene Analysis* [5] we present favourable results, with robust and interpretable

acoustic embeddings that characterize emotions in our UC3M4Safety Audiovisual Stimuli Dataset.

Additionally we performed a brief contribution to the study of fatigue, providing a gender analysis of fatigue expression, and in future research it would be interesting to characterize it to see the differences between stress, fear, and fatigue on physiological variables and their effects on the voice. The work performed on the detection of GBV conditon from speech promisingly indicates that it is possible to distinguish between GBVV and non GBVV by using their paralinguistic cues, opening a new line of research on Affective Computing and for gender-based violence victim detection with the use of WEMAC. Finally, and following another objective aligned with the *social good*, we link climate change and gender violence.

Finally, as already mentioned, these lines of research require more attention and future work for a more holistic understanding in the context of detection and prevention of gender-based violence using audio technology.

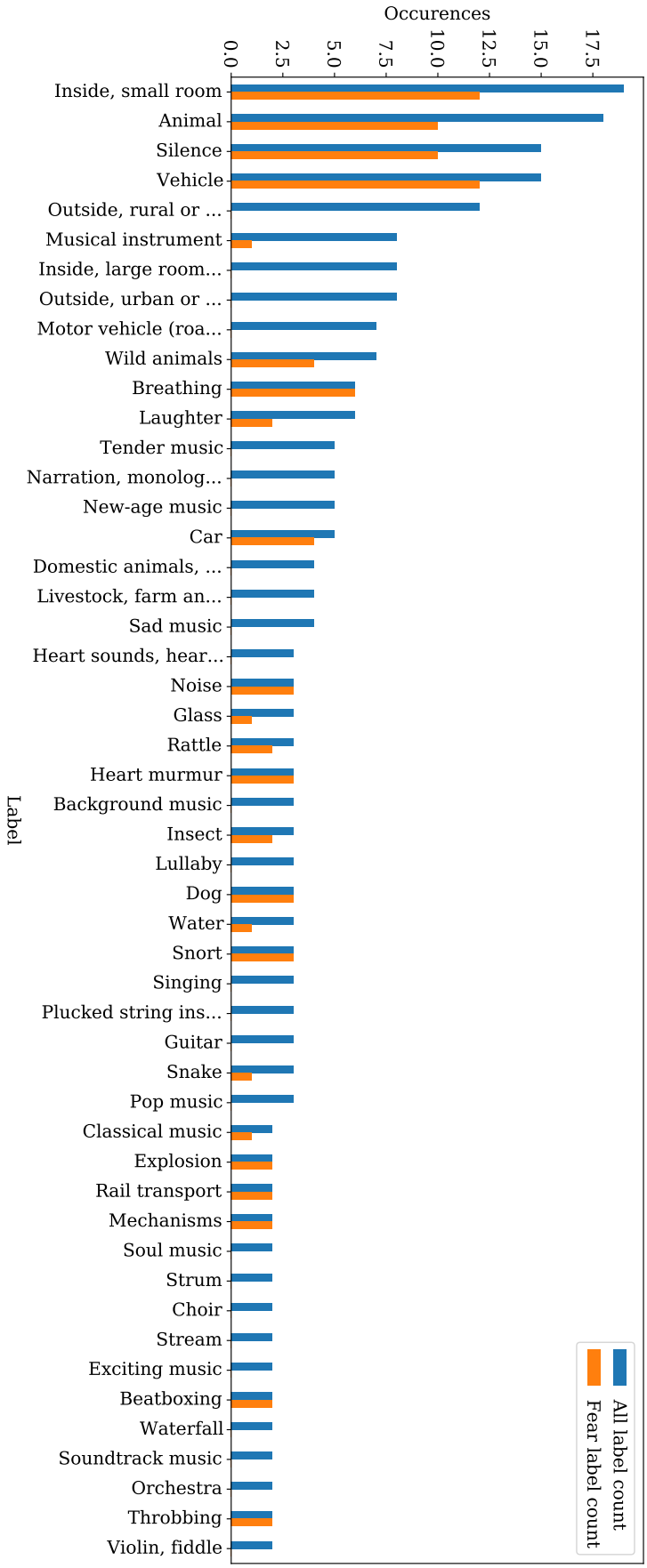


FIGURE 6.11: Absolute number of occurrences in YAMNet acoustic labels in *fear* vs. all audio-visual stimulus in WEMAC [1]. Reproduced with permission from the copyright owner, © 2022 IEEE.

Chapter 7

Conclusions and Future Work

This chapter gathers the conclusions drawn from the research works conducted on this thesis. Our main objective in this thesis was to understand the reactions of women to situations of risk to be able to detect them through automatic detection mechanisms using the auditory modality and machine learning algorithms. In this chapter we give an overview of the conclusions gathered from the work done on the analysis of gender-based violence in the Affective Computing field, the reasoning behind the databases used, and the work on the task of speaker and emotions recognition together with the evolution of inference systems in Bindi. This thesis is both multimodal and multidisciplinary because there has been a great degree of collaboration and the contributions are deeply intertwined with those of other members. We detailed each of the contributions in the introductory part of every chapter and section. A strong gender perspective guiding technology is a much needed pioneering work, so we can consider this investigation, together with that of the other members of UC3M4Safety team, to lay the foundations of this new perspective in which we aim to continue doing future work.

7.1 Conclusions

Gender-based violence is experienced by 1 in 3 women globally at some time in their lives, either physically or sexually, according to The World Health Organization (WHO) [21]. In particular in Spain, more than 1,100 women have been murdered from 2003 to 2022, victims of gender-based violence [22]. Also GBV is normalised and reproduced due to structural and social inequalities therefore, this socially-invisible problem needs to be urgently tackled in order to protect women, which comprise more than 50% of the worldwide population.

Since the creation of the UC3M4Safety team, we have been working to develop an innovative technological solution that could go hand in hand with artificial intelligence to stop violence against women. And that is how Bindi was born. Bindi consists of a system formed by a smartwatch – bracelet – a pendant, a smartphone app and a cloud server. It constitutes a system of devices that aim to detect automatically when a situation could be life-threatening for a woman, and offering support and help on-the-fly. Bindi uses as main sources of information physiological and auditory signals, in addition to other situational variables such as time of day and GPS location, to perform such analysis.

The detection and classification of emotions, within the field or Affective Computing, is a great source of information as it could inform about a person's affective state. That a person is stressed, nervous, frightened, fearful or generates the fight-flight-freeze reaction, could be an indicator that she is in a dangerous situation. And so, the use of Affective Computing and the detection of emotions, could be key

in the detection of GBV. Emotions can be tracked through physiological (see [53]) and speech variables, and in this thesis we focus of the latter.

Not only emotions, but also the context and situation in which the person finds herself can give us especially relevant information for the confirmation that a situation is of risk, as the aforementioned GPS coordinates, the time of the day, and the acoustic context (cars passing by, silence, or a scuffle). These other aspects are currently under study by other members of the UC3M4Safety team we are collaborating with.

But the detection of emotions and specifically of *fear* is very difficult. We have already talked about the difficulties of emotional labels because of their subjective nature and due to the difference in perception of the annotators who label them in Sec. 2.5. Regarding the lack of real data, we have applied data augmentation techniques generating synthetic stressed speech utterances – which we consider a realistic emotion similar to *fear* – in addition to contaminating audio signals additively with realistic environmental noise present in real-life settings, trying to mimic the conditions where Bindi would work.

However, several problems arise when the goal of a system is to work with real-life data, as Bindi is expected to do. First, the difficulty of finding realistic data, and second, the low confidence on the architectures developed if the data used to train them are acted or synthetic. This situation leads to the need to generate databases with real elicited emotions, which is highly challenging and time-consuming. This is how the datasets of UC3M Audiovisual Stimuli [11.1] [11.2], WEMAC [11] and WE-LIVE emerged, fully detailed in Chapter 3.

Particularly, working with strong negative emotion elicitation, such as that evoked in WEMAC for *fear* detection in women in a laboratory environment, can lead to ethical issues. Thus, many resources must be devoted to safeguarding the welfare of the volunteers participating in the databases collection. This particular problem is magnified when the target group of volunteers comprises women who have suffered GBV. This is because the failures of the system have critical consequences for them. Although the investment of resources to provide safety and comfortability during the recording of our databases is considerable, we were totally committed to the volunteers' well-being, providing constant psychological assistance as the probability of triggering their post-traumatic stress disorder is very high. It should be noted that the development of subject-adaptation techniques is critical for our GBV use case. Nevertheless, we consider the generation of these databases to be a great contribution, where we handled with great care the ethical issues and limitations, with the purpose of serving to pave the path for research on technology to combat gender-based violence.

But before detecting the affective state in which a person is, it is necessary to confirm that the speech within the audio signal recorded by Bindi belongs to that user, and that is the reason of our work in the Speaker Recognition or Identification field in Chapter 4, to detect the user's voice and therefore identity from among all the acoustic information present in the audio signal. In this regard, we took special care in our investigation about hardware constraints that our Bindi devices had (see Sec. 1.2.2). We addressed the SR task taking into account two sources of variability that an audio recorded in a real-life setting could include: stress conditions and environmental noise. Stress proved to affect negatively the performance of SR when only used in the testing phase, and so we augmented the database with synthetically stressed speech for the training of the ML models, which proved to improve the performance. Therefore in the absence of real emotional stressed speech we can augment the data to achieve data that resembles real stress and can help maintain

an acceptable recognition rate in SR systems. Likewise, the contamination of audio signals with environmental noise worsens the SR rates, including for systems with computational constrains such as Bindi. So we worked towards developing models robust to such noise that would indeed fit our conditions.

In Chapter 5 we detailed the development of the Bindi system. We evaluated the detection and classification of *fear*-related emotions from a multimodal and multidisciplinary perspective, from Bindi 1.0 to Bindi 2.0. We validated the use of the monomodal data pipelines and data fusion architectures combining physiological signals and audio for detecting *fear* out of speech utterances by using WEMAC and achieved a promising result of an overall *fear* classification accuracy of 63.61% for a speaker-adapted subject-dependent approach. We also described in depth the components and functioning the system – bracelet, pendant, app and server –, and described the work done in the task stress detection from speech utterances. The experimentation carried out in Chapter 5 serves as an initial multimodal approach toward working with real elicited *fear* in women and its proper processing. Bindi is a very complex system that requires a thorough balance of many aspects, such as battery consumption, computational power, resource usage, and algorithm performance. All of this work in *fear* emotion recognition is intended to pave the way and shape the next version of Bindi: Bindi 3.0.

In Chapter 6 we give voice to those lines of research that have arisen along the way while additionally investigating *fear* detection through speech with a gender-based violence perspective. We detail the work carried out in the field of affective acoustic events and scene analysis and their importance for the detection of risk situations through the analysis of the acoustic context. Additionally the brief work performed in the study of fatigue would be interesting to use to analyse the differences between stress, *fear*, and fatigue and their effects on speech. The preliminary work performed on the detection of GBV condition from speech promisingly paves the way for future work with applications in psychological therapy. And finally, and following another objective aligned with the *social good*, we link climate change and gender violence.

Overall, this thesis explores the use of technology and artificial intelligence to prevent and combat gender-based violence. We hope that we have lit the way for it in the speech modality and that our experimentation, findings and conclusions can help in future research. The ultimate goal of this work is to ignite the community's interest in developing solutions to the very challenging problem of GBV.

7.2 Future Work

Specifically in the line of the work carried out in the field of speaker identification and emotions using the speech modality, the greater amount of realistic data available due to the databases we recorded allows us to use more elaborated deep learning architectures, e.g., to be used in the disentanglement of the speaker's identity and the emotional information in the future. We could use an adversarial model that could disentangle these two branches of data into different low-dimensional vectors (embeddings) to synchronously detect the speaker and emotion together. Thereby, in every speech instance we could infer the identity of the speaker as well as her emotion at the same time. This is what we intend to work on after this thesis, using the databases collected throughout it: WEMAC and WE-LIVE.

In general terms, for the development of Bindi we also have to consider that many women remain in a state of shock when assaulted or are victims of an aggression, instead of producing fearful speech. We must take into account this fact for further developments in the Bindi system, or by analyzing the occurrence of silences in the audio, together with the other variables we have already explored.

Regarding the analysis of acoustic events and acoustic context within Bindi, it would be crucial to include a module for the study of acoustic information, with its own data processing and pipeline, and its fusion branch together with the physiological and speech modalities in order to have a more complete and holistic GBV risk situation detector. The detection of vocal bursts such as grunts, growls, heavy-breathing, squeals or shrieks, it is also of special interest for our application, as well as the detection of acoustic events such as hits, bumps or impacts, which would likely denote that a dangerous situation is happening.

Additionally, statistically speaking, it is more likely that men, rather than women, commit any form of social violence (e.g., intimate partner violence, assault, rape, murder) [312]. Thus, distinguishing male voices under dominant emotions such as anger with Bindi could denote a risk situation, as most aggression to women are perpetrated by men.

On a different note, AI-based systems are gaining popularity in healthcare, but are limited by "high requirements for accuracy, robustness, and explainability" [74]. AI in health research, a subfield of digital health, explores many human-centered approaches. There are many recent advances in the audio domain, which it has been so far understudied but it is also highly promising, with a particular focus on speech data present state-of-the-art technologies. This study [74] presents "the latest research on the automatic detection of diseases through audio signals" in a review style, "from acute and chronic respiratory diseases via psychiatric disorders to developmental and neurodegenerative disorders". The analysis of emotions, particularly *fear*, and the condition of gender-based violence discussed in this thesis, could help health-oriented audio AI research, in particular with applications in mental health care and psychotherapy.

Bibliography

References [1-20] can be found from page *vii* to page *xi*.

[21] World Health Organization. *Violence Against Women Prevalence Estimates, 2018*. Mar. 2021, p. 87. ISBN: 978-92-4-002225-6.

[22] Delegación del Gobierno contra la Violencia de Género, Ministerio de Igualdad, Gobierno de España. *Ficha estadística de víctimas mortales por Violencia de Género*. 2022.

[23] European Institute for Gender Equality. *What is gender-based violence?* Accessed: 19-11-2022. URL: https://eige.europa.eu/gender-based-violence/what-is-gender-based-violence.

[24] European Institute for Gender Equality. *Forms of Gender-based Violence*. Accessed: 23-10-2022. URL: https://eige.europa.eu/gender-based-violence/forms-of-violence.

[25] NGO Pulse. *Ending Violence Against Women*. Accessed: 23-10-2022. URL: https://www.ngopulse.org/node/75637/?mini=2022-07.

[26] *Types of violence against women and girls*. Accessed: 23-10-2022. URL: https://iran.un.org/en/102394-frequently-asked-questions-types-violence-against-women-and-girls.

[27] European Institute for Gender Equality. *Definition of Economic Violence*. Accessed: 23-10-2022. URL: https://eige.europa.eu/thesaurus/terms/1096.

[28] *Cyber Violence Against Women*. Accessed: 23-10-2022. URL: http://equality.ofetin.ro/index.php/en/introduction.

[29] European Institute for Gender Equality. *Cyber violence against women and girls*. 2017. ISBN: 978-92-9493-896-1. DOI: 10.2839/876816.

[30] *Cyber-Violence: a gendered threat. UNRIC*. Accessed: 23-10-2022. URL: https://unric.org/en/cyber-violence-a-gendered-threat/.

[31] Amnesty International. *La sociedad percibe un avance en la violencia de género, pero en la práctica la brecha continúa*. Accessed: 23-10-2022. URL: https://amnistia.org.ar/de-las-opiniones-a-los-hechos-la-sociedad-percibe-un-avance-en-igualdad-de-genero-pero-en-la-practica-la-brecha-continua/.

[32] European Institute for Gender Equality. *The costs of gender-based violence in the European Union*. Oct. 2021, pp. 1–59. ISBN: 978-92-9482-921-4. DOI: 10.2839/063244.

[33] European Institute for Gender Equality. *Estimating the costs of gender-based violence in the European Union*. 2014, pp. 1–123. ISBN: 978-92-9218-499-5. DOI: 10.2839/79629.

[34] UN General Assembly. *Declaration on the Elimination of Violence against Women*. Sexual and gender-based violence (SGBV), Document Symbol A/RES/48/104, Reference 85th plenary meeting. Dec. 1993. URL: https://www.ohchr.org/en/instruments-mechanisms/instruments/declaration-elimination-violence-against-women.

[35] United Nations. *The 2030 Agenda and the Sustainable Development Goals: An opportunity for Latin America and the Caribbean*. Santiago, 2018.

[36] *Proposal for a Directive of the European Parliament and the Council on combating violence against women and domestic violence*. Accessed: 23-10-2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX\%3A52022PC0105.

[37] *CIJ: What is intersectionality. Intersectional Justice*. Accessed: 23-10-2022. URL: https://www.intersectionaljustice.org/what-is-intersectionality.

[38] Dr. Professor María José Díaz-Aguado. "Prevenir la violencia de género desde la escuela". In: *Revista de Estudios de Juventud, ISSN 0211-4364, Nº. 86, 2009 (Ejemplar dedicado a: Juventud y violencia de género), pags. 31-46* (Jan. 2009).

[39] Inma Pastor, Angel Belzunegui Eraso, Marta Calvo Merino, and Paloma Pontón Merino. "La violencia de género en España: un análisis quince años después de la Ley 1/2004". In: *REIS: Revista Española de Investigaciones Sociológicas* 174 (2021), pp. 109–128.

[40] European Institute for Gender Equality. *EU regulations for GBV*. Accessed: 23-10-2022. URL: https://eige.europa.eu/gender-based-violence/regulatory-and-legal-framework/eu-regulations.

[41] European Institute for Gender Equality. *International regulations for GBV*. Accessed: 23-10-2022. URL: https://eige.europa.eu/gender-based-violence/regulatory-and-legal-framework/international-regulations.

[42] B.O.E. *Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género*. Dec. 2004.

[43] *Istanbul Convention Action against violence against women and domestic violence*. Accessed: 23-10-2022. URL: https://www.coe.int/en/web/istanbul-convention/home?.

[44] Unicef. *Six ways tech can help end gender-based violence*. Accessed: 23-10-2022. URL: https://www.unicef.org/eap/blog/six-ways-tech-can-help-end-gender-based-violence.

[45] Rachel Jewkes and Elizabeth Dartnall. "More research is needed on digital technologies in violence against women". In: *The Lancet Public Health* 4.6 (2019), e270–e271. ISSN: 2468-2667. DOI: https://doi.org/10.1016/S2468-2667(19)30076-3.

[46] Lenin Medeiros, Tibor Bosse, and Charlotte Gerritsen. "Can a Chatbot Comfort Humans? Studying the Impact of a Supportive Chatbot on Users' Self-Perceived Stress". In: *IEEE Transactions on Human-Machine Systems* 52.3 (2022), pp. 343–353. DOI: 10.1109/THMS.2021.3113643.

[47] Fabio Massimo Zanzotto. "Viewpoint: Human-in-the-loop Artificial Intelligence". In: *Journal of Artificial Intelligence Research* 64 (Feb. 2019), pp. 243–252. DOI: 10.1613/jair.1.11345.

[48] Naveena Karusala and Neha Kumar. "Women's Safety in Public Spaces: Examining the Efficacy of Panic Buttons in New Delhi". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2017, 3340–3351. ISBN: 9781450346559. DOI: 10.1145/3025453.3025532.

[49] Spanish Ministry of the Interior and Public Security. *alertcops.ses.mir.es*. https://alertcops.ses.mir.es/mialertcops/en/index.html. (Accessed on 04/04/2021).

[50] Ignacio Rodríguez-Rodríguez, José-Víctor Rodríguez, Aránzazu Elizondo-Moreno, Purificación Heras-González, and Michele Gentili. "Towards a Holistic ICT Platform for Protecting Intimate Partner Violence Survivors Based on the IoT Paradigm". In: *Symmetry* 12.1 (2020). ISSN: 2073-8994. DOI: 10.3390/sym12010037. URL: https://www.mdpi.com/2073-8994/12/1/37.

[51] Tania Martínez. "A travel of the Institutional System in the field of gender violence". In: *Revista de Estudios Socioeducativos. ReSed* 7 (2019), pp. 256–257.

[52] Rosa San Segundo Manuel, Clara Sainz de Baranda, Marian Blanco Ruiz, David Larrabeiti López, Manuel Urueña Pascual, Jose Carlos Robledo García, Carmen Peláez Moreno, Ascensión Gallardo Antolín, Alba Mínguez Sánchez, Teresa Riesgo Alcaide, Jose Manuel Lanza Gutiérrez, Jose Ángel Miranda Calero Rodrigo Mariño Andrés, Manuel Felipe Canabal, Marta Portela García, Isabel Pérez Garcilópez, Jose Antonio García Souto, Celia López Ongil, Emilio Olías Ruiz, and Mario García Valderas. *Utility model U202130953(3) - Sistema para Determinar un Estado Emocional de un Usuario*. Publication number: ES1269890. Publication date: 09/06/2021. Grant date: 20/09/2021 (Spain). Owner institutions: Universidad Carlos III de Madrid (85%) and Universidad Politécnica de Madrid (15%).

[53] Jose Miranda Calero. "Fear Classification using Affective Computing with Physiological Information and Smart-Wearables". PhD. Thesis. 2022.

[54] Lori Mosca, Elizabeth Barrett-Connor, and Nanette Wenger. "Sex/Gender Differences in Cardiovascular Disease Prevention What a Difference a Decade Makes". In: *Circulation* 124 (Nov. 2011), pp. 2145–54. DOI: 10.1161/CIRCULATIONAHA.110.968792.

[55] Clifford Nass. *The Man Who Lied to His Laptop*. Penguin Publishing Group, 2010. ISBN: 9781617230011.

[56] Rachael Tatman. "Gender and Dialect Bias in YouTube's Automatic Captions". In: Jan. 2017, pp. 53–59. DOI: 10.18653/v1/W17-1606.

[57] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *FAT*. 2018.

[58] *IBM Watson Visual Recognition*. Accessed: 23-10-2022. URL: https://www.ibm.com/watson/services/visual-recognition/.

[59] Daniel Koerber, Shawn Khan, Tahmina Shamsheri, Abirami Kirubarajan, and Sangeeta Mehta. "The Effect of Skin Tone on Accuracy of Heart Rate Measurement in Wearable Devices: A Systematic Review". In: *Journal of the American College of Cardiology* 79.9_Supplement (2022), pp. 1990–1990. DOI: 10.1016/S0735-1097(22)02981-3.

[60] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. "Automatic speech recognition and speech variability: A review". In: *Speech Communication* 49.10 (2007). Intrinsic Speech Variations, pp. 763–786. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2007.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0167639307000404.

[61] A. Nadeem, B. Abedin, and O. Marjanovic. "Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies". In: *Association for Information Systems (ACIS) Proceedings*. 2020.

[62] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. "The role of artificial intelligence in achieving the Sustainable Development Goals". In: *Nature Communications* 11 (2020), p. 233. ISSN: 20411723. DOI: 10.1038/s41467-019-14108-y.

[63] Daniel Greene, Anna Hoffmann, and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning". In: Jan. 2019. DOI: 10.24251/HICSS.2019.258.

[64] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. "Mitigating Gender Bias in Natural Language Processing: Literature Review". In: *CoRR* abs/1906.08976 (2019). arXiv: 1906.08976. URL: http://arxiv.org/abs/1906.08976.

[65] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. "Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[66] Davide Cirillo, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, Alfonso Valencia, María José Rementeria, Antonella Santuccione Chadha, and Nikolaos Mavridis. "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare". In: *npj Digital Medicine* 3.1 (June 2020), p. 81. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0288-5. URL: https://doi.org/10.1038/s41746-020-0288-5.

[67] Stefanie Rukavina, Sascha Gruss, Holger Hoffmann, Jun-Wen Tan, Steffen Walter, and Harald C. Traue. "Affective Computing and the Impact of Gender and Age". In: *PLOS ONE* 11.3 (Mar. 2016), pp. 1–20. DOI: 10.1371/journal.pone.0150584. URL: https://doi.org/10.1371/journal.pone.0150584.

[68] Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. *Handling emotions in human-computer dialogues*. Appendix A: Emotional Speech Databases. Springer, 2010.

[69] Yann LeCun, Y. Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521 (May 2015), pp. 436–44. DOI: 10.1038/nature14539.

[70] Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Modern Deep Learning Research". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.09 (Apr. 2020), pp. 13693–13696. DOI: 10.1609/aaai.v34i09.7123. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7123.

[71] Mark Coeckelbergh. "AI for climate: freedom, justice, and other ethical and political challenges". In: *AI and Ethics* 1 (Oct. 2020). DOI: 10.1007/s43681-020-00007-2.

[72] Alba Páez-Montoro, Mario García-Valderas, Emilio Olías-Ruíz, and Celia López-Ongil. "Solar Energy Harvesting to Improve Capabilities of Wearable Devices". In: *Sensors* 22.10 (2022). ISSN: 1424-8220. DOI: 10.3390/s22103950. URL: https://www.mdpi.com/1424-8220/22/10/3950.

[73] *Proposal for a Regulation of the European Partliament and of the Council for the Artificial Intelligence Act*. Accessed: 23-10-2022. URL: https://eur-lex.europa.eu/legal-content/en/HIS/?uri=CELEX:52021PC0206.

[74] Manuel Milling, Florian B. Pokorny, Katrin D. Bartl-Pokorny, and Björn W. Schuller. "Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell". In: *Frontiers in Digital Health* 4 (2022). ISSN: 2673-253X. DOI: 10.3389/fdgth.2022.886615. URL: https://www.frontiersin.org/article/10.3389/fdgth.2022.886615.

[75] Rosalind W. Picard. "Affective Computing". In: *Perceptual Computing Section Technical Report No. 321*. M.I.T Media Laboratory, 1995.

[76] Karen Niven. "Affect". In: *Encyclopedia of Behavioral Medicine*. Ed. by Marc D. Gellman and J. Rick Turner. New York, NY: Springer New York, 2013, pp. 49–50. ISBN: 978-1-4419-1005-9. DOI: 10.1007/978-1-4419-1005-9_1088. URL: https://doi.org/10.1007/978-1-4419-1005-9_1088.

[77] Marc Gellman and John Turner. *Encyclopedia of Behavioral Medicine*. Jan. 2013. ISBN: 978-1-4419-1004-2. DOI: 10.1007/978-1-4419-1005-9.

[78] Dr Lisa Feldman-Barrett. *We don't understand how emotions work. A neuroscientist explains why we often get it wrong*. https://www.sciencefocus.com/the-human-body/what-are-emotions/. Accessed: 2022-07-05.

[79] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition)*. New York: Worth, 2011. URL: http://www.amazon.com/Psychology-Daniel-L-Schacter/dp/1429237198/ref=sr_1_1?s=books\&ie=UTF8\&qid=1313937150\&sr=1-1.

[80] Queensland Brain Institute Australia. *The limbic system*. Accessed: 23-10-2022. URL: https://qbi.uq.edu.au/brain/brain-anatomy/limbic-system.

[81] "A review of systems and networks of the limbic forebrain/limbic midbrain". In: *Progress in Neurobiology* 75.2 (2005), pp. 143–160. ISSN: 0301-0082. DOI: https://doi.org/10.1016/j.pneurobio.2005.01.001. URL: https://www.sciencedirect.com/science/article/pii/S030100820500002X.

[82] Britannica Educational Publishing and K. Rogers. *The Brain and the Nervous System*. Human body. Britannica Educational Pub., 2010. ISBN: 9781615302567. URL: https://books.google.de/books?id=mW05ic06qdwC.

[83] *Emotion and behaviour*. https://www.britannica.com/science/human-nervous-system/Emotion-and-behaviour. Accessed: 2022-06-28.

[84] *Fight Or Flight Response. Psychologytools*. Accessed: 23-10-2022. URL: https://www.psychologytools.com/resource/fight-or-flight-response/#:~:text=The\%20fight\%20or\%20flight\%20response,body\%20to\%20fight\%20or\%20flee..

[85] Andrei Schiller-Chan. *How Stress Affects Your Voice. Blog post*. Accessed: 23-10-2022. URL: https://oratorvoice.medium.com/how-stress-affects-your-voice-the-freeze-response-1c005faecff1.

[86] Healthline. *Fight, Flight, Freeze: What This Response Means*. Accessed: 23-10-2022. URL: https://www.healthline.com/health/mental-health/fight-flight-freeze.

[87] Shelley E Taylor, Laura Cousino Klein, Brian P Lewis, Tara L Gruenewald, Regan AR Gurung, and John A Updegraff. "Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight." In: *Psychological review* 107.3 (2000), p. 411.

[88] Psychcentral. *Stress Response: What Is Tend and Befriend?* Accessed: 23-10-2022. URL: https://psychcentral.com/stress/tend-and-befriend.

[89] Harvard Health. *Understanding the stress response*. Accessed: 23-10-2022. URL: https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response#:~:text=After\%20the\%20amygdala\%20sends\%20a,as\%20adrenaline)\%20into\%20the\%20bloodstream..

[90] I. Milosevic and R.E. McCabe. *Phobias: The Psychology of Irrational Fear*. ABC-CLIO, LLC, 2015. ISBN: 9781610695756. URL: https://books.google.de/books?id=4SfroAEACAAJ.

[91] Th. Steimer. "The biology of fear- and anxiety-related behaviors". In: *Dialogues in Clinical Neuroscience* 4 (2002), pp. 231 –249.

[92] *Understanding the stress response*. https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response. Accessed: 2022-06-28.

[93] Stephen W. Porges. "The polyvagal theory: phylogenetic substrates of a social nervous system". In: *International Journal of Psychophysiology* 42.2 (2001), pp. 123–146. ISSN: 0167-8760. DOI: https://doi.org/10.1016/S0167-8760(01)00162-3. URL: https://www.sciencedirect.com/science/article/pii/S0167876001001623.

[94] Otniel E Dror. "The Cannon–Bard thalamic theory of emotions: A brief genealogy and reappraisal". In: *Emotion Review* 6.1 (2014), pp. 13–20.

[95] Psychcentral. *Fight or Flight*. Accessed: 23-10-2022. URL: https://psychcentral.com/lib/fight-or-flight#1.

[96] *Anxiety and its affects on the auditory and vocal apparatus*. https://australianvoiceassociation.com.au/2017/12/anxiety-and-its-affects-on-the-auditory-and-vocal-apparatus/. Accessed: 2022-06-29.

[97] Britannica, T. Editors of Encyclopedia. *Vagus Nerve*. Oct. 2022. URL: https://www.britannica.com/science/vagus-nerve.

[98] Paul Ekman. "An argument for basic emotions". In: *Cognition and Emotion* 6.3-4 (1992), pp. 169–200. DOI: 10.1080/02699939208411068. eprint: https://doi.org/10.1080/02699939208411068. URL: https://doi.org/10.1080/02699939208411068.

[99] Paul Ekman and Daniel Cordaro. "What is Meant by Calling Emotions Basic". In: *Emotion Review* 3 (Sept. 2011), pp. 364–370. DOI: 10.1177/1754073911410740.

[100] PSU. *Basic Emotion Theory: A Categorical Approach*. Accessed: 23-10-2022. URL: https://psu.pb.unizin.org/psych425/chapter/basic-emotion-perspective/.

[101] Michelle Yarwood. *Psychology of Human Emotion*. Accessed: 2022-06-30. Affordable Course Transformation: Pennsylvania State University.

[102] Flávia Oliveira, Rui Joaquim, Renato Salviato Fajardo, and Sandro Caramaschi. "Psychobiology of Sadness: Functional Aspects in Human Evolution". In: 7 (Nov. 2018), pp. 1015–1022.

[103] Rainer Reisenzein, Gernot Horstmann, and Achim Schuetzwohl. "The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence". In: *Topics in Cognitive Science* 11 (Sept. 2017). DOI: 10.1111/tops.12292.

[104] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements". In: *Psychological science in the public interest* 20.1 (2019), pp. 1–68.

[105] James Russell. "A Circumplex Model of Affect". In: *Journal of Personality and Social Psychology* 39 (Dec. 1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[106] Jonathan Posner, James A. Russell, and Bradley S. Peterson. "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology". In: *Development and Psychopathology* 17 (2005), pp. 715 –734.

[107] Albert. Mehrabian. *Basic dimensions for a general psychological theory : implications for personality, social, environmental, and developmental studies / Albert Mehrabian*. eng. Cambridge, Mass: Oelgeschlager, Gunn and Hain, 1980.

[108] Albert Mehrabian and James A. Russell. "The Basic Emotional Impact of Environments". In: *Perceptual and Motor Skills* 38.1 (1974). PMID: 4815507, pp. 283–301. DOI: 10.2466/pms.1974.38.1.283. eprint: https://doi.org/10.2466/pms.1974.38.1.283. URL: https://doi.org/10.2466/pms.1974.38.1.283.

[109] Iris Bakker, Theo Van der Voordt, Jan Boon, and Peter Vink. "Pleasure, Arousal, Dominance: Mehrabian and Russell revisited". In: *Current Psychology* 33 (Oct. 2014), pp. 405–421. DOI: 10.1007/s12144-014-9219-4.

[110] Mimoun Wiem and Zied Lachiri. "Emotion Classification in Arousal Valence Model using MAHNOB-HCI Database". In: *International Journal of Advanced Computer Science and Applications* 8 (Mar. 2017). DOI: 10.14569/IJACSA.2017.080344.

[111] M. Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. "A Multimodal Database for Affect Recognition and Implicit Tagging". In: *IEEE Transactions on Affective Computing* 3 (2012), pp. 42–55.

[112] Shen Zhang, Zhiyong Wu, Helen Meng, and Lianhong Cai. "Facial Expression Synthesis Based on Emotion Dimensions for Affective Talking Avatar". In: vol. 2010. June 2010, pp. 109–132. ISBN: 978-3-642-12603-1. DOI: 10.1007/978-3-642-12604-8_6.

[113] Johnny Fontaine, Klaus Scherer, Etienne Roesch, and Phoebe Ellsworth. "The World of Emotions is not Two-Dimensional". In: *Psychological science* 18 (Jan. 2008), pp. 1050–7. DOI: 10.1111/j.1467-9280.2007.02024.x.

[114] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[115] Shaundra B. Daily, Melva T. James, David Cherry, John J. Porter, Shelby S. Darnell, Joseph Isaac, and Tania Roy. "Chapter 9 - Affective Computing: Historical Foundations, Current Applications, and Future Trends". In: *Emotions and Affect in Human Factors and Human-Computer Interaction*. Ed. by Myounghoon Jeon. San Diego: Academic Press, 2017, pp. 213–231. ISBN: 978-0-12-801851-4. DOI: https://doi.org/10.1016/B978-0-12-801851-4.00009-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128018514000094.

[116] Wikipedia. *Machine Learning*. Accessed: 23-10-2022. URL: https://en.wikipedia.org/wiki/Machine_learning.

[117] *What is Deep Learning? Techtarget.* Accessed: 23-10-2022. URL: https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network.

[118] Sandra Carberry and Fiorella Rosis. "Introduction to special Issue on 'Affective modeling and adaptation'". In: *User Model. User-Adapt. Interact.* 18 (Feb. 2008), pp. 1–9. DOI: 10.1007/s11257-007-9044-7.

[119] Cambridge. *Cambridge International Dictionary of English*. Cambridge University Press, 1995.

[120] Rosalind Picard. "Affective computing: Challenges". In: *International Journal of Human-Computer Studies* 59 (July 2003), pp. 55–64. DOI: 10.1016/S1071-5819(03)00052-1.

[121] Rosalind Picard. "Affective computing: Challenges". In: *International Journal of Human-Computer Studies* 59 (July 2003), pp. 55–64. DOI: 10.1016/S1071-5819(03)00052-1.

[122] Moshe Zeidner, Richard D. Roberts, and Gerald Matthews. "Can Emotional Intelligence Be Schooled? A Critical Review". In: *Educational Psychologist* 37.4 (2002), pp. 215–231. DOI: 10.1207/S15326985EP3704_2. eprint: https://doi.org/10.1207/S15326985EP3704_2. URL: https://doi.org/10.1207/S15326985EP3704_2.

[123] Ann M Kring and Albert H Gordon. "Sex differences in emotion: expression, experience, and physiology." In: *Journal of personality and social psychology* 74.3 (1998), p. 686.

[124] Leslie R. Brody and Judith A. Hall. "Gender and emotion in context." In: *Handbook of Emotions*. 2008.

[125] Wikipedia. *Gender and emotional expression*. Accessed: 23-10-2022. URL: https://en.wikipedia.org/wiki?curid=46457885.

[126] Marian Blanco-Ruiz, Clara Sainz-de Baranda, Laura Gutiérrez-Martín, Elena Romero-Perales, and Celia López-Ongil. "Emotion Elicitation Under Audiovisual Stimuli Reception: Should Artificial Intelligence Consider the Gender Perspective?" In: *International Journal of Environmental Research and Public Health* 17.22 (2020). ISSN: 1660-4601. DOI: 10.3390/ijerph17228534. URL: https://www.mdpi.com/1660-4601/17/22/8534.

[127] June Feder, Ronald Levant, and James Dean. "Boys and Violence: A Gender-Informed Analysis". In: *Professional Psychology Research and Practice* 1 (Aug. 2010), pp. 3–12. DOI: 10.1037/2152-0828.1.S.3.

[128] Thurid Vogt and Elisabeth André. "Improving Automatic Emotion Recognition from Speech via Gender Differentiaion". In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), May 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/392_pdf.pdf.

[129] Rui Xia, Jun Deng, Björn Schuller, and Yang Liu. "Modeling gender information for emotion recognition using Denoising autoencoder". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 990–994. DOI: 10.1109/ICASSP.2014.6853745.

[130] A. Stacey and J. Stacey. "Integrating sustainable development into research ethics protocols". In: *Electronic Journal of Business Research Methods* 10 (Jan. 2012), pp. 54–63.

[131] Anton Batliner, Simone Hantke, and Björn Schuller. "Ethics and Good Practice in Computational Paralinguistics". In: *IEEE Transactions on Affective Computing* PP (Sept. 2020), pp. 1–1. DOI: 10.1109/TAFFC.2020.3021015.

[132] Shaundra B. Daily, Melva T. James, David Cherry, John J. Porter, Shelby S. Darnell, Joseph Isaac, and Tania Roy. "Chapter 9 - Affective Computing: Historical Foundations, Current Applications, and Future Trends". In: *Emotions and Affect in Human Factors and Human-Computer Interaction*. Ed. by Myounghoon Jeon. San Diego: Academic Press, 2017, pp. 213–231. ISBN: 978-0-12-801851-4. DOI: https://doi.org/10.1016/B978-0-12-801851-4.00009-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128018514000094.

[133] Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Hmani, Aymen Mtibaa, Mohamed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Matrouf Driss, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gerard Chollet, Nicholas Evans, Thomas Schneider, and Christoph Busch. "Preserving privacy in speaker and speech characterisation". In: *Computer Speech & Language* 58 (Nov. 2019). DOI: 10.1016/j.csl.2019.06.001.

[134] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. "Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference". In: *Privacy and Identity Management. Data for Better Living: AI and Privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Windisch, Switzerland, August 19–23, 2019, Revised Selected Papers*. Ed. by Michael Friedewald, Melek Önen, Eva Lievens, Stephan Krenn, and Samuel Fricker. Cham: Springer International Publishing, 2020, pp. 242–258. ISBN: 978-3-030-42504-3. DOI: 10.1007/978-3-030-42504-3_16. URL: https://doi.org/10.1007/978-3-030-42504-3_16.

[135] Anton Batliner, Simone Hantke, and Björn Schuller. "Ethics and Good Practice in Computational Paralinguistics". In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1236–1253. DOI: 10.1109/TAFFC.2020.3021015.

[136] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. "Online Passive-Aggressive Algorithms". In: *Journal of Machine Learning Research* 7 (Mar. 2006), pp. 551–585.

[137] European Comission. *Protección de datos Normas sobre protección de datos personales dentro y fuera de la UE*. Accessed: 23-10-2022. URL: https://ec.europa.eu/info/law/law-topic/data-protection_es.

[138] Catherine D'Ignazio, Helena Suarez Val, Silvana Fumega, Harini Suresh, and Isadora Cruxen. "Feminicide & Machine Learning: detecting gender-based violence to strengthen civil sector activism". In: (2020).

[139] Catherine D'Ignazio, Isadora Cruxên, Helena Suárez Val, Angeles Martinez Cuba, Mariel García-Montes, Silvana Fumega, Harini Suresh, and Wonyoung So. "Feminicide and counterdata production: Activist efforts to monitor and challenge gender-related violence". In: *Patterns* 3.7 (2022), p. 100530. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2022.100530. URL: https://www.sciencedirect.com/science/article/pii/S2666389922001271.

[140] Ignacio Rodríguez-Rodríguez, José-Víctor Rodríguez, Domingo-Javier Pardo-Quiles, Purificación Heras-González, and Ioannis Chatzigiannakis. "Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques". In: *Applied Sciences* 10.22 (2020). ISSN: 2076-3417. DOI: 10.3390/app10228244. URL: https://www.mdpi.com/2076-3417/10/22/8244.

[141] Ignacio Rodríguez-Rodríguez, José-Víctor Rodríguez, Aránzazu Elizondo-Moreno, and Purificación Heras-González. "An Autonomous Alarm System for Personal Safety Assurance of Intimate Partner Violence Survivors Based on Passive Continuous Monitoring through Biosensors". In: *Symmetry* 12.3 (2020). ISSN: 2073-8994. DOI: 10.3390/sym12030460. URL: https://www.mdpi.com/2073-8994/12/3/460.

[142] Carlos M. Castorena, Itzel M. Abundez, Roberto Alejo, Everardo E. Granda-Gutiérrez, Eréndira Rendón, and Octavio Villegas. "Deep Neural Network for Gender-Based Violence Detection on Twitter Messages". In: *Mathematics* 9.8 (2021). ISSN: 2227-7390. DOI: 10.3390/math9080807. URL: https://www.mdpi.com/2227-7390/9/8/807.

[143] Grisel Miranda, Roberto Alejo, Carlos Castorena, Eréndira Rendón, Javier Illescas, and Vicente García. "Deep Neural Network to Detect Gender Violence on Mexican Tweets". In: *Progress in Artificial Intelligence and Pattern Recognition*. Ed. by Yanio Hernández Heredia, Vladimir Milián Núñez, and José Ruiz Shulcloper. Cham: Springer International Publishing, 2021, pp. 24–32. ISBN: 978-3-030-89691-1.

[144] Robin Petering, Mee Young Um, Nazanin Alipourfard, Nazgol Tavabi, Rajni Kumari, and Setareh Nasihati Gilani. "Artificial Intelligence to predict Intimate Partner Violence perpetration". In: *Artificial intelligence and social work* (Nov. 2018), pp. 195–210.

[145] Luis Francisco Ramos-Lima, Vitoria Waikamp, Thyago Antonelli-Salgado, Ives Cavalcante Passos, and Lucia Helena Machado Freitas. "The use of machine learning techniques in trauma-related disorders: a systematic review". In: *Journal of Psychiatric Research* 121 (2020), pp. 159–172. ISSN: 0022-3956. DOI: https://doi.org/10.1016/j.jpsychires.2019.12.001. URL: https://www.sciencedirect.com/science/article/pii/S0022395619311021.

[146] Dimitrios Stoidis and Andrea Cavallaro. *Generating gender-ambiguous voices for privacy-preserving speech recognition*. 2022. DOI: 10.48550/ARXIV.2207.01052. URL: https://arxiv.org/abs/2207.01052.

[147] *Data Centric AI*. Accessed: 23-10-2022. URL: https://datacentricai.org/.

[148] Janneke Wilting, Emiel Krahmer, and Marc Swerts. "Real vs. acted emotional speech". In: *Ninth International Conference on Spoken Language Processing*. 2006.

[149] Suja Sreeith Panicker and Prakasam Gayathri. "A survey of machine learning techniques in physiology based mental stress detection systems". In: *Biocybernetics and Biomedical Engineering* 39.2 (2019), pp. 444–469. ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2019.01.004. URL: https://www.sciencedirect.com/science/article/pii/S020852161830367X.

[150] H.J.M. Steeneken and J.H.L. Hansen. "Speech under stress conditions: overview of the effect on speech production and on system performance". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. Vol. 4. 1999, 2079–2082 vol.4. DOI: 10.1109/ICASSP.1999.758342.

[151] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42 (Dec. 2008), pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

[152] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter Sendlmeier, and Benjamin Weiss. "A database of German emotional speech". In: vol. 5. Sept. 2005, pp. 1517–1520. DOI: 10.21437/Interspeech.2005-446.

[153] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. "Fear-type emotion recognition for future audio-based surveillance systems". In: *Speech Communication* 50.6 (2008), pp. 487–503. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2008.03.012. URL: https://www.sciencedirect.com/science/article/pii/S016763930800037X.

[154] John H. L. Hansen and Sahar E. Bou-Ghazale. "Getting started with SUSAS: a speech under simulated and actual stress database". In: *EUROSPEECH*. 1997.

[155] Ayako Ikeno, Vaishnevi Varadarajan, Sanjay Patil, and John H.L. Hansen. "UT-Scope: Speech under Lombard Effect and Cognitive Stress". In: *2007 IEEE Aerospace Conference*. 2007, pp. 1–7. DOI: 10.1109/AERO.2007.352975.

[156] Ana Aguiar, Mariana Kaiseler, Hugo Meinedo, Pedro Almeida, Mariana Cunha, and Jorge Silva. "VOCE Corpus: Ecologically Collected Speech Annotated with Physiological and Psychological Stress Assessments". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1568–1574. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/647_Paper.pdf.

[157] Alice Baird, Shahin Amiriparian, Miriam Berschneider, Maximilian Schmitt, and Björn Schuller. "Predicting Biological Signals from Speech: Introducing a Novel Multimodal Dataset and Results". In: Sept. 2019, pp. 1–5. DOI: 10.1109/MMSP.2019.8901758.

[158] Shin-ae Yoon, Guiyoung Son, and Soonil Kwon. "Fear emotion classification in speech by acoustic and behavioral cues". In: *Multimedia Tools and Applications* 78 (Jan. 2019). DOI: 10.1007/s11042-018-6329-2.

[159] Muriel Hagenaars and Agnes Van minnen. "The effect of fear on paralinguistic aspects of speech in patients with panic disorder with agoraphobia". In: *Journal of anxiety disorders* 19 (Feb. 2005), pp. 521–37. DOI: 10.1016/j.janxdis.2004.04.008.

[160] Alan K. Alimuradov, Alexander Yu. Tychkov, Viktoriya A. Mezhina, Ekaterina A. Fokina, Angelina E. Zhurina, Alexey V. Ageykin, Valery N. Gorbunov, and Ekaterina K. Reva. "Development of Natural Emotional Speech Database for Training Automatic Recognition Systems of Stressful Emotions in Human-Robot Interaction". In: *2020 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR)*. 2020, pp. 11–16. DOI: 10.1109/DCNAIR50402.2020.9216940.

[161] Róbert Sabo and Jakub Rajčáni. "Designing the Database of Speech Under Stress". In: *Journal of Linguistics/Jazykovedný casopis* 68 (Dec. 2017). DOI: 10.1515/jazcas-2017-0042.

[162] Róbert Sabo, Jakub Rajčáni, and Marian Ritomsky. "Designing Database of Speech Under Stress Using a Simulation in Virtual Reality". In: Aug. 2018, pp. 321–326. DOI: 10.1109/DISA.2018.8490641.

[163] Alice Baird, Andreas Triantafyllopoulos, Sandra Zänkert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, Brigitte M. Kudielka, Nicolas Rohleder, Harald Baumeister, and Björn W. Schuller. "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress". In: *Frontiers in Computer Science* 3 (2021). ISSN: 2624-9898. DOI: 10.3389/fcomp.2021.750284. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2021.750284.

[164] Alba Mínguez-Sánchez. "Detección de Estrés en Señales de voz". Bachelor Thesis. University Carlos III Madrid, Spain, 2017.

[165] C.D. Spielberger, R.L. Gorssuch, P.R. Lushene, P.R. Vagg, and G.A. Jacobs. *State-Trait Anxiety Inventory (STAI)*. 1968.

[166] M. Brookes. *"Voicebox: Speech processing toolbox for matlab [software]"*. [Online]. 2011. URL: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[167] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[168] Simone Hantke, Erik Marchi, and Björn Schuller. "Introducing the Weighted Trustability Evaluator for Crowdsourcing Exemplified by Speaker Likability Classification". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2156–2161. URL: https://aclanthology.org/L16-1342.

[169] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "DEAP: A Database for Emotion Analysis Using Physiological Signals". In: *IEEE Transactions on Affective Computing* 3 (Dec. 2011), pp. 18–31. DOI: 10.1109/T-AFFC.2011.15.

[170] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[171] Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Shah. "Sound Event Detection in Domestic Environments with Weakly Labeled Data and Soundscape Synthesis". In: Oct. 2019, pp. 253–257. DOI: 10.33682/006b-jx26.

[172] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello. "Scaper: A library for soundscape synthesis and augmentation". In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2017, pp. 344–348. DOI: 10.1109/WASPAA.2017.8170052.

[173] Antonio Maffei and Alessandro Angrilli. "E-MOVIE - Experimental MOVies for Induction of Emotions in neuroscience: An innovative film database with normative data and sex differences". In: *PLOS ONE* 14.10 (Oct. 2019), pp. 1–22. DOI: 10.1371/journal.pone.0223124. URL: https://doi.org/10.1371/journal.pone.0223124.

[174] Laura Gutiérrez-Martín, Elena Romero-Perales, Clara Sainz de Baranda Andújar, Manuel F. Canabal-Benito, Gema Esther Rodríguez-Ramos, Rafael Toro-Flores, Susana López-Ongil, and Celia López-Ongil. "Fear Detection in Multimodal Affective Computing: Physiological Signals versus Catecholamine Concentration". In: *Sensors* 22.11 (2022). ISSN: 1424-8220. DOI: 10.3390/s22114023. URL: https://www.mdpi.com/1424-8220/22/11/4023.

[175] J Rottenberg, RD Ray, and JJ Gross. *Emotion elicitation using films In: Coan JA, Allen JJB, editors. The handbook of emotion elicitation and assessment*. 2007.

[176] J. A. Miranda, M. F. Canabal, L. Gutiérrez-Martín, J. M. Lanza-Gutiérrez, and C. López-Ongil. "A Design Space Exploration for Heart Rate Variability in a Wearable Smart Device". In: *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*. 2020, pp. 1–6. DOI: 10.1109/DCIS51330.2020.9268628.

[177] Jose A Miranda, Manuel F Canabal, M Portela García, and Celia Lopez-Ongil. "Embedded emotion recognition: Autonomous multimodal affective internet of things". In: *Proceedings of the cyber-physical systems workshop*. Vol. 2208. 2018, pp. 22–29.

[178] M. F. Canabal, J. A. Miranda, J. M. Lanza-Gutiérrez, A. I. Pérez Garcilópez, and C. López-Ongil. "Electrodermal Activity Smart Sensor Integration in a Wearable Affective Computing System". In: *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*. 2020, pp. 1–6. DOI: 10.1109/DCIS51330.2020.9268662.

[179] Manuel F. Canabal, Jose A. Miranda, Alba Páez Montoro, Isabel Pérez Garcilópez, Susana Patón Álvarez, Ernesto García Ares, and Celia López-Ongil. "Design and Validation of an Efficient and Adjustable GSR Sensor for Emotion Monitoring". Manuscript submitted for publication. 2022.

[180] Laura Gutiérrez Martín. "Entorno de entrenamiento para detección de emociones en víctimas de Violencia de Género mediante realidad virtual". Bachelor Thesis. 2019.

[181] Clara Sainz-de Baranda Andujar, Laura Gutiérrez-Martín, José Ángel Miranda-Calero, Marian Blanco-Ruiz, and Celia López-Ongil. "Gender biases in the training methods of affective computing: Redesign and validation of the Self-Assessment Manikin in measuring emotions via audiovisual clips". In: *Frontiers in Psychology* 13 (2022). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2022.955530. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.955530.

[182] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W. Schuller. *The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress*. 2021. arXiv: 2104.07123 [cs.CL].

[183] Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, viktorandreevichmorozov, Keunwoo Choi, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Adam Weiss, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Matt Vollrath, Taewoon Kim, and Thassilo. *librosa/librosa: 0.9.1*. Version 0.9.1. Feb. 2022. DOI: 10.5281/zenodo.6097378. URL: https://doi.org/10.5281/zenodo.6097378.

[184] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7 (Jan. 2015), pp. 1–1. DOI: 10.1109/TAFFC.2015.2457417.

[185] Florian Eyben, Martin Wöllmer, and Björn Schuller. "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. Jan. 2010, pp. 1459–1462. DOI: 10.1145/1873951.1874246.

[186] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini. "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language". English (US). In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-September-2016 (2016). Publisher Copyright: Copyright © 2016 ISCA.; 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016 ; Conference date: 08-09-2016 Through 16-09-2016, pp. 2001–2005. ISSN: 2308-457X. DOI: 10.21437/Interspeech.2016-129.

[187] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller. "Snore Sound Classification Using Image-Based Deep Spectrum Features". en. In: *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516.

[188] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[189] Arnab Poddar, Md. Sahidullah, and Goutam Saha. "Speaker verification with short utterances: a review of challenges, trends and opportunities". In: *IET Biom.* 7 (2018), pp. 91–101.

[190] Wei Wu, Fang Zheng, Mingxing Xu, and Huanjun Bao. "Study on speaker verification on emotional speech." In: *Interspeech*. 2006.

[191] Tomi Kinnunen and Haizhou Li. "An overview of text-independent speaker recognition: From features to supervectors". In: *Speech Communication* 52.1 (2010), pp. 12 –40. ISSN: 0167-6393.

[192] R. J. Mammone, Xiaoyu Zhang, and R. P. Ramachandran. "Robust speaker recognition: a feature-based approach". In: *IEEE Signal Processing Magazine* 13.5 (Sept. 1996), p. 58. ISSN: 1558-0792.

[193] Rania Chakroun and Mondher Frikha. "Robust features for text-independent speaker recognition with short utterances". In: *Neural Computing and Applications* 32.17 (2020), pp. 13863–13883. ISSN: 1433-3058.

[194] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models". In: *Digital Signal Processing* 10.1 (2000), pp. 19 –41. ISSN: 1051-2004. DOI: https://doi.org/10.1006/dspr.1999.0361.

[195] J. P. Campbell. "Speaker recognition: a tutorial". In: *Proceedings of the IEEE* 85.9 (Sept. 1997), pp. 1437–1462. ISSN: 0018-9219. DOI: 10.1109/5.628714.

[196] G. Senthil Raja and S. Dandapat. "Speaker recognition under stressed condition". In: *International Journal of Speech Technology* 13.3 (Sept. 2010), pp. 141–161. ISSN: 1572-8110. DOI: 10.1007/s10772-010-9075-z. URL: https://doi.org/10.1007/s10772-010-9075-z.

[197] H. J. M. Steeneken and J. H. L. Hansen. "Speech under stress conditions: overview of the effect on speech production and on system performance". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99*. Vol. 4. Mar. 1999, pp. 2079–2082. DOI: 10.1109/ICASSP.1999.758342.

[198] N. Zheng, T. Lee, and P. C. Ching. "Integration of Complementary Acoustic Features for Speaker Recognition". In: *IEEE Signal Processing Letters* 14.3 (Mar. 2007), pp. 181–184. ISSN: 1070-9908. DOI: 10.1109/LSP.2006.884031.

[199] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2016, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669.

[200] S. Zhang, S. Zhang, T. Huang, and W. Gao. "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching". In: *IEEE Transactions on Multimedia* 20.6 (June 2018), pp. 1576–1590. ISSN: 1520-9210. DOI: 10.1109/TMM.2017.2766843.

[201] Guoqiang Zhong, Li-Na Wang, Xiao Ling, and Junyu Dong. "An overview on data representation learning: From traditional feature learning to recent deep learning". In: *The Journal of Finance and Data Science* 2.4 (2016), pp. 265 –278. ISSN: 2405-9188.

[202] Amir H. Hadjahmadi and Mohammad M. Homayounpour. "Robust feature extraction and uncertainty estimation based on attractor dynamics in cyclic deep denoising autoencoders". In: *Neural Computing and Applications* 31.11 (2019), pp. 7989–8002. ISSN: 1433-3058.

[203] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord. "Unsupervised Speech Representation Learning Using WaveNet Autoencoders". In: *IEEE Transactions on Audio, Speech and Language Processing* 27.12 (Dec. 2019), pp. 2041–2053. ISSN: 2329-9304.

[204] Suwon Shon, Hao Tang, and James R. Glass. "VoiceID Loss: Speech Enhancement for Speaker Verification". In: *ArXiv* abs/1904.03601 (2019).

[205] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W. Schuller. "Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends". In: *CoRR* abs/2001.00378 (2020). arXiv: 2001.00378. URL: http://arxiv.org/abs/2001.00378.

[206] J. Li, A. Mohamed, G. Zweig, and Y. Gong. "LSTM time and frequency recurrence for automatic speech recognition". In: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Dec. 2015, pp. 187–191.

[207] A. Graves, A. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. May 2013, pp. 6645–6649.

[208] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. "Deep Neural Network Embeddings for Text-Independent Speaker Verification". In: *Proc. of INTERSPEECH*. 2017.

[209] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. "X-Vectors: Robust DNN Embeddings for Speaker Recognition". In: *Proc. of ICASSP*. Apr. 2018, pp. 5329–5333.

[210] Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, Pedro A. Torres-Carrasquillo, and Najim Dehak. "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations". In: *Computer Speech & Language* 60 (2020), p. 101026. ISSN: 0885-2308.

[211] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur. "A study on data augmentation of reverberant speech for robust speech recognition". In: *Proc. of ICASSP*. Mar. 2017, pp. 5220–5224.

[212] Ilyes Rebai, Yessine BenAyed, Walid Mahdi, and Jean-Pierre Lorré. "Improving speech recognition using data augmentation and acoustic model fusion". In: *Procedia Computer Science* 112 (2017). Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France, pp. 316 –322. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.08.003. URL: http://www.sciencedirect.com/science/article/pii/S1877050917313420X.

[213] P. Y. Simard, D. Steinkraus, and J. C. Platt. "Best practices for convolutional neural networks applied to visual document analysis". In: *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* Aug. 2003, pp. 958–963. DOI: 10.1109/ICDAR.2003.1227801.

[214] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R. Hershey, and Björn Schuller. "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR". In: *Latent Variable Analysis and Signal Separation*. Ed. by Emmanuel Vincent, Arie Yeredor, Zbyněk Koldovský, and Petr Tichavský. Cham: Springer International Publishing, 2015, pp. 91–99. ISBN: 978-3-319-22482-4.

[215] O. Plchot, L. Burget, H. Aronowitz, and P. Matejka. "Audio enhancing with DNN autoencoder for speaker recognition". In: *Proc. of ICASSP*. Mar. 2016, pp. 5090–5094.

[216] Kerlos A. Abdalmalak and Ascensión Gallardo-Antolín. "Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers". In: *Neural Computing and Applications* 29.3 (2018), pp. 637–651. ISSN: 1433-3058.

[217] Carlos Busso and Shrikanth Narayanan. "Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Jan. 2008, pp. 1670–1673.

[218] Dongdong Li, Yubo Yuan, and Zhaohui Wu. "Affect-insensitive speaker recognition systems via emotional speech clustering using prosodic features". In: *Neural Computing and Applications* 26.2 (2015), pp. 473–484. ISSN: 1433-3058.

[219] M. Abdelwahab and C. Busso. "Domain Adversarial for Acoustic Emotion Recognition". In: *IEEE T AUDIO SPEECH* 26.12 (Dec. 2018), pp. 2423–2435. ISSN: 2329-9304.

[220] Ismail Shahin, Ali B. Nassif, and Shibani Hamsa. "Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments". In: *Neural Computing and Applications* 32.7 (2020), pp. 2575–2587. ISSN: 1433-3058.

[221] Didier Meuwly and Andrzej Drygajlo. "Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM)." In: *A Speaker Odyssey - The Speaker Recognition Workshop*. 2001, pp. 145–150.

[222] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. "Support vector machines using GMM supervectors for speaker verification". In: *IEEE Signal Processing Letters* 13.5 (May 2006), pp. 308–311. ISSN: 1070-9908. DOI: 10.1109/LSP.2006.870086.

[223] K. A. Abdalmalak and A. Gallardo-Antolín. "Enhancement of a text-independent speaker verification system by using feature combination and parallel structure classifiers". In: *Neural Computing and Applications* 29.3 (Feb. 2018), pp. 637–651. DOI: 10.1007/s00521-016-2470-x.

[224] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren. "A novel scheme for speaker recognition using a phonetically-aware deep neural network". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2014, pp. 1695–1699. DOI: 10.1109/ICASSP.2014.6853887.

[225] Dong Yu, Michael L. Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide. "Feature Learning in Deep Neural Networks - A Study on Speech Recognition Tasks". In: *International Conference on Learning Representations*. 2013.

[226] Zhaofeng Zhang, Longbiao Wang, Atsuhiko Kai, Takanori Yamada, Weifeng Li, and Masahiro Iwahashi. "Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), pp. 1–13.

[227] Y. Zhao, Z. Wang, and D. Wang. "A two-stage algorithm for noisy and reverberant speech enhancement". In: *Proc. of ICASSP*. 2017, pp. 5580–5584.

[228] M. Kolbœk, Z. Tan, and J. Jensen. "Speech enhancement using Long Short-Term Memory based recurrent Neural Networks for noise robust Speaker Verification". In: *IEEE Spoken Language Technology Workshop (SLT)*. 2016, pp. 305–311.

[229] P. S. Nidadavolu, S. Kataria, J. Villalba, P. García-Perera, and N. Dehak. "Unsupervised Feature Enhancement for Speaker Verification". In: *Proc. of ICASSP*. 2020, pp. 7599–7603.

[230] X. Ji, M. Yu, C. Zhang, D. Su, T. Yu, X. Liu, and D. Yu. "Speaker-Aware Target Speaker Enhancement by Jointly Learning with Speaker Embedding Extraction". In: *Proc. of ICASSP*. 2020, pp. 7294–7298.

[231] Lara Lynn Stoll. "Finding Difficult Speakers in Automatic Speaker Recognition". PhD thesis. EECS Dept., Univ. of California, Berkeley, Dec. 2011.

[232] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011". In: *Artificial Intelligence Review* 43.2 (Feb. 2015), pp. 155–177. ISSN: 1573-7462. DOI: 10.1007/s10462-012-9368-5. URL: https://doi.org/10.1007/s10462-012-9368-5.

[233] Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, Abhimanyu Singh Garhwal, and Ruili Wang. "Speaker identification features extraction methods: A systematic review". In: *Expert Systems with Applications* 90 (2017), pp. 250 –271. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2017.08.015. URL: http://www.sciencedirect.com/science/article/pii/S0957417417305535.

[234] A. Zhang. *Speech Recognition Library for Python (Version 3.8) [Software]*. Accessed on 4 June 2019. URL: https://github.com/Uberi/speech_recognition.

[235] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. "The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings". In: *J ACOUST SOC AM* 133 (May 2013), p. 3591.

[236] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio". In: *Proc. of the Detection & Classification of Acoustic Scenes & Events Workshop (DCASE2017)*. Nov. 2017.

[237] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks". In: *J MACH LEARN RES* 18 (Dec. 2017).

[238] Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. "VESUS: A Crowd-Annotated Database to Study Emotion Production and Perception in Spoken English". In: *Proc. Interspeech 2019*. 2019, pp. 316–320.

[239] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. "Voxceleb: Large-scale speaker verification in the wild". In: *Computer Speech & Language* 60 (2020), p. 101027. ISSN: 0885-2308.

[240] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak. "X-Vectors Meet Emotions: A Study On Dependencies Between Emotion and Speaker Recognition". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 7169–7173. DOI: 10.1109/ICASSP40776.2020.9054317.

[241] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. "librosa, Audio and Music Signal Analysis in Python". In: *Proceedings of the 14th python in science conference*. Jan. 2015, pp. 18–24. DOI: 10.25080/Majora-7b98e3ed-003.

[242] M. Plakal and D. Ellis. *YAMNet*. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet. Accessed: 2020-12-30.

[243] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[244] Francois Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[245] Mehmet Berkehan Akçay and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers". In: *Speech Communication* 116 (2020), pp. 56 –76. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2019.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0167639319302262.

[246] C. Wu, J. Lin, W. Wei, and K. Cheng. "Emotion recognition from multi-modal information". In: *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. Oct. 2013, pp. 1–8. DOI: 10.1109/APSIPA.2013.6694347.

[247] Rosalind W. Picard. "Affective Computing for HCI". In: *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I - Volume I*. USA: L. Erlbaum Associates Inc., 1999, 829–833. ISBN: 0805833919.

[248] Jianhua Tao and Tieniu Tan. "Affective Computing: A Review". In: *Affective Computing and Intelligent Interaction*. Ed. by Jianhua Tao, Tieniu Tan, and Rosalind W. Picard. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995. ISBN: 978-3-540-32273-3.

[249] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases". In: *Pattern Recognition* 44.3 (2011), pp. 572–587. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2010.09.020. URL: https://www.sciencedirect.com/science/article/pii/S0031320310004619.

[250] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. "Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information". In: *Proceedings of the 6th International Conference on Multimodal Interfaces*. ICMI '04. State College, PA, USA: Association for Computing Machinery, 2004, 205–211. ISBN: 1581139950. DOI: 10.1145/1027933.1027968. URL: https://doi.org/10.1145/1027933.1027968.

[251] Björn W. Schuller. "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". In: *Communications of the ACM* 61 (Apr. 2018), pp. 90–99. DOI: 10.1145/3129340.

[252] Mehmet Berkehan Akçay and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers". In: *Speech Communication* 116 (2020), pp. 56 –76. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2019.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0167639319302262.

[253] Fatemeh Noroozi, Dorota KamiÅ„ska, Tomasz SapiÅ„ski, and Gholamreza Anbarjafari. "Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests, and Adaboost". In: *Journal of the Audio Engineering Society* 65 (Aug. 2017), pp. 562–572. DOI: 10.17743/jaes.2017.0022.

[254] P. Jackson & S. Haq. *Surrey Audio-Visual Expressed Emotion (SAVEE) Database, University of Surrey*. 2014. URL: http://kahlan.eps.surrey.ac.uk/savee/Download.html.

[255] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos Dimoulas, and George Kalliris. "Speech Emotion Recognition for Performance Interaction". In: *Journal of the Audio Engineering Society* 66 (June 2018), pp. 457–467. DOI: 10.17743/jaes.2018.0036.

[256] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech Emotion Recognition Using Deep Learning Techniques: A Review". In: *IEEE Access* 7 (2019), pp. 117327–117345.

[257] L. Devillers and L. Vidrascu. "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs". In: *INTERSPEECH*. 2006.

[258] Albert F Ax. "The physiological differentiation between fear and anger in humans". In: *Psychosomatic medicine* 15.5 (1953), pp. 433–442.

[259] Oana Bălan, Gabriela Moise, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. "Fear Level Classification Based on Emotional Dimensions and Machine Learning Techniques". In: *Sensors* 19.7 (2019). ISSN: 1424-8220.

[260] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. "DEAP: A Database for Emotion Analysis ;Using Physiological Signals". In: *IEEE Transactions on Affective Computing* 3 (2012), pp. 18–31.

[261] Jose A. Miranda et al. "Fear Recognition for Women Using a Reduced Set of Physiological Signals". In: *Sensors* 21.5 (2021). ISSN: 1424-8220. DOI: 10.3390/s21051587. URL: https://www.mdpi.com/1424-8220/21/5/1587.

[262] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. "A Multimodal Database for Affect Recognition and Implicit Tagging". In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 42–55.

[263] Abdulkadir Celik, Khaled N. Salama, and Ahmed M. Eltawil. "The Internet of Bodies: A Systematic Survey on Propagation Characterization and Channel Modeling". In: *IEEE Internet of Things Journal* 9.1 (2022), pp. 321–345. DOI: 10.1109/JIOT.2021.3098028.

[264] Yong Zhang, Yang Chen, Yujie Wang, Qingqing Liu, and Andong Cheng. "CSI-Based Human Activity Recognition With Graph Few-Shot Learning". In: *IEEE Internet of Things Journal* 9.6 (2022), pp. 4139–4151. DOI: 10.1109/JIOT.2021.3103073.

[265] Tianming Zhao, Yan Wang, Jian Liu, Jerry Cheng, Yingying Chen, and Jiadi Yu. "Robust Continuous Authentication Using Cardiac Biometrics from Wrist-worn Wearables". In: *IEEE Internet of Things Journal* (2021), pp. 1–1. DOI: 10.1109/JIOT.2021.3128290.

[266] Tao Zhang, Minjie Liu, Tian Yuan, and Najla Al-Nabhan. "Emotion-Aware and Intelligent Internet of Medical Things Toward Emotion Recognition During COVID-19 Pandemic". In: *IEEE Internet of Things Journal* 8.21 (2021), pp. 16002–16013. DOI: 10.1109/JIOT.2020.3038631.

[267] Tong Wang, Yang Shen, Lin Gao, Yufei Jiang, Xu Zhu, and Fu-Chun Zheng. "Long-Term Energy Consumption and Transmission Delay Tradeoff in Wireless-Powered Body Area Networks". In: *IEEE Internet of Things Journal* 9.6 (2022), pp. 4051–4064. DOI: 10.1109/JIOT.2021.3102950.

[268] Arlene John, Stephen Redmond, Barry Cardiff, and Deepu John. "A Multimodal Data Fusion Technique for Heartbeat Detection in Wearable IoT Sensors". In: *IEEE Internet of Things Journal* PP (June 2021), pp. 1–1. DOI: 10.1109/JIOT.2021.3093112.

[269] Sen Qiu, Zhengdong Hao, Zhelong Wang, Long Liu, Jiayi Liu, Hongyu Zhao, and Giancarlo Fortino. "Sensor Combination Selection Strategy for Kayak Cycle Phase Segmentation Based on Body Sensor Networks". In: *IEEE Internet of Things Journal* 9.6 (2022), pp. 4190–4201. DOI: 10.1109/JIOT.2021.3102856.

[270] Yang Bai, Lixing Chen, Mohamed Abdel-Mottaleb, and Jie Xu. "Automated Ensemble for Deep Learning Inference on Edge Computing Platforms". In: *IEEE Internet of Things Journal* 9.6 (2022), pp. 4202–4213. DOI: 10.1109/JIOT.2021.3102945.

[271] Haozhao Wang, Zhihao Qu, Qihua Zhou, Haobo Zhang, Boyuan Luo, Wenchao Xu, Song Guo, and Ruixuan Li. "A Comprehensive Survey on Training Acceleration for Large Machine Learning Models in IoT". In: *IEEE Internet of Things Journal* 9.2 (2022), pp. 939–963. DOI: 10.1109/JIOT.2021.3111624.

[272] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas. "Speech Emotion Recognition Adapted to Multimodal Semantic Repositories". In: *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. Sept. 2018, pp. 31–35. DOI: 10.1109/SMAP.2018.8501881.

[273] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Çaglar Gülçehre, Vincent Michalski, Kishore Reddy Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron C. Courville, Pascal Vincent, Roland Memisevic, Christopher J. Pal, and Yoshua Bengio. "EmoNets: Multimodal deep learning approaches for emotion recognition in video". In: *CoRR* abs/1503.01800 (2015). arXiv: 1503.01800. URL: http://arxiv.org/abs/1503.01800.

[274] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information Fusion* 37 (2017), pp. 98–125. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.02.003. URL: https://www.sciencedirect.com/science/article/pii/S1566253517300738.

[275] Jianhua Zhang et al. "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review". In: *Information Fusion* 59 (2020), pp. 103–126.

[276] Yucel Cimtay et al. "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion". In: *IEEE Access* 8 (2020), pp. 168865–168878.

[277] Yongrui Huang et al. "Combining facial expressions and electroencephalography to enhance emotion recognition". In: *Future Internet* 11.5 (2019), p. 105.

[278] Amir Muaremi, Agon Bexheti, Franz Gravenhorst, Bert Arnrich, and Gerhard Tröster. "Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors". In: *IEEE-EMBS international conference on biomedical and health informatics (BHI)*. IEEE. 2014, pp. 185–188.

[279] Eiman Kanjo, Eman MG Younis, and Nasser Sherkat. "Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach". In: *Information Fusion* 40 (2018), pp. 18–31.

[280] Jonghwa Kim and Elisabeth Andre. "Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation". In: June 2006, pp. 53–64.

[281] Sylvia D Kreibig. "Autonomic nervous system activity in emotion: A review". In: *Biological psychology* 84.3 (2010), pp. 394–421.

[282] Miguel A. Campos-Gaviño and David Larrabeiti. "Toward court-admissible sensor systems to fight domestic violence". In: *IEEE International Conference on Multimedia Communications, Services & Security, MCSS2020*. 2020 (submitted).

[283] J. A. Miranda Calero, R. Marino, J. M. Lanza-Gutierrez, T. Riesgo, M. Garcia-Valderas, and C. Lopez-Ongil. "Embedded Emotion Recognition within Cyber-Physical Systems using Physiological Signals". In: *2018 Conference on Design of Circuits and Integrated Systems (DCIS)*. 2018, pp. 1–6. DOI: 10.1109/DCIS.2018.8681496.

[284] Javier Ramirez, Jose C. Segura, Carmen Benitez, Angel de la Torre, and Antonio Rubio. "Efficient voice activity detection algorithms using long-term speech information". In: *Speech Communication* 42.3 (2004), pp. 271–287. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2003.10.002. URL: https://www.sciencedirect.com/science/article/pii/S0167639303001201.

[285] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano. "On the Personalization of Classification Models for Human Activity Recognition". In: *IEEE Access* 8 (2020), pp. 32066–32079. DOI: 10.1109/ACCESS.2020.2973425.

[286] Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. "100% Classification Accuracy Considered Harmful: The Normalized Information Transfer Factor Explains the Accuracy Paradox". In: *PLOS ONE* 9.1 (Jan. 2014), pp. 1–10. DOI: 10.1371/journal.pone.0084217. URL: https://doi.org/10.1371/journal.pone.0084217.

[287] Peter Mell and Tim Grance. "The NIST definition of cloud computing". In: *National Institute of Standards and Technology* (2011). URL: https://doi.org/10.6028/NIST.SP.800-145.

[288] Weisong Shi et al. "Edge computing: Vision and challenges". In: *IEEE internet of things journal* 3.5 (2016), pp. 637–646.

[289] M. Iorga, L. Feldman, R. Barton, M. Martin, N. Goren, and C. Mahmoudi. "Fog Computing Conceptual Model". In: *Special Publication (NIST SP), National Institute of Standards and Technology* (2018). URL: https://doi.org/10.6028/NIST.SP.500-325.

[290] Gopika Premsankar et al. "Edge computing for the Internet of Things: A case study". In: *IEEE Internet of Things Journal* 5.2 (2018), pp. 1275–1284.

[291] Jorge Portilla et al. "The extreme edge at the bottom of the Internet of Things: A review". In: *IEEE Sensors Journal* 19.9 (2019), pp. 3179–3190.

[292] Farshad Firouzi et al. "The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT)". In: *Information Systems* (2021), p. 101840.

[293] Andrew G. Howard et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *arXiv preprint arXiv:1704.04861* (Apr. 2017).

[294] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. "CNN architectures for large-scale audio classification". In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 131–135. DOI: 10.1109/ICASSP.2017.7952132.

[295] Beáta T. Szabó, Susan L. Denham, and István Winkler. "Computational Models of Auditory Scene Analysis: A Review". In: *Front. Neurosci.* 10 (Nov. 2016). ISSN: 1662-453X. DOI: 10.3389/fnins.2016.00524.

[296] André Fiebig, Pamela Jordan, and Cleopatra Christina Moshona. "Assessments of Acoustic Environments by Emotions – The Application of Emotion Theory in Soundscape". In: *Frontiers in Psychology* 11 (2020). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.573041. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.573041.

[297] Faranak Abri, Luis Felipe Gutiérrez, Prerit Datta, David R. W. Sears, Akbar Siami Namin, and Keith S. Jones. "A Comparative Analysis of Modeling and Predicting Perceived and Induced Emotions in Sonification". In: *Electronics* 10.20 (2021). ISSN: 2079-9292. DOI: 10.3390/electronics10202519. URL: https://www.mdpi.com/2079-9292/10/20/2519.

[298] Thomas Goerne. "The Emotional Impact of Sound: A Short Theory of Film Sound Design". In: Jan. 2019. DOI: 10.29007/jk8h.

[299] Daniel Västfjäll. "The Subjective Sense of Presence, Emotion Recognition, and Experienced Emotions in Auditory Virtual Environments". In: *CyberPsychology & Behavior* 6.2 (2003). PMID: 12804030, pp. 181–188. DOI: 10.1089/109493103321640374. eprint: https://doi.org/10.1089/109493103321640374. URL: https://doi.org/10.1089/109493103321640374.

[300] Felix Weninger, Florian Eyben, Björn W. Schuller, Marcello Mortillaro, and Klaus R. Scherer. "On the Acoustics of Emotion in Audio: What Speech, Music, and Sound Have in Common". In: *Front. Psychol.* 4 (2013). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00292.

[301] Bjorn Schuller, Simone Hantke, Felix Weninger, Wenjing Han, Zixing Zhang, and Shrikanth Narayanan. "Automatic Recognition of Emotion Evoked by General Sound Events". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto, Japan: IEEE, Mar. 2012, pp. 341–344. ISBN: 978-1-4673-0046-9 978-1-4673-0045-2 978-1-4673-0044-5. DOI: 10.1109/ICASSP.2012.6287886.

[302] Weiyi Ma and William Forde Thompson. "Human Emotions Track Changes in the Acoustic Environment". In: *Proc. Natl. Acad. Sci. U.S.A.* 112.47 (Nov. 2015), pp. 14563–14568. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1515087112.

[303] Tom Garner and Mark Grimshaw. "A Climate of Fear: Considerations for Designing a Virtual Acoustic Ecology of Fear". In: *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction with Sound*. AM '11. Coimbra, Portugal: Association for Computing Machinery, 2011, 31–38. ISBN: 9781450310819. DOI: 10.1145/2095667.2095672. URL: https://doi.org/10.1145/2095667.2095672.

[304] Karen Sparck Jones. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". In: *Journal of Documentation* 28.1 (Jan. 1972), pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526.

[305] Akiko Aizawa. "An Information-Theoretic Perspective of Tf–Idf Measures". In: *Information Processing & Management* 39.1 (Jan. 2003), pp. 45–65. ISSN: 03064573. DOI: 10.1016/S0306-4573(02)00021-3.

[306] Nerys Williams. "The Borg rating of perceived exertion (RPE) scale". In: *Occupational Medicine* 67.5 (2017), pp. 404–405.

[307] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2880–2894.

[308] Andreas Triantafyllopoulos, Manuel Milling, Konstantinos Drossos, and Björn W. Schuller. "Fairness and underspecification in acoustic scene classification: The case for disaggregated evaluations". In: *Proc. DCASE*. online, 2021.

[309] Emma Reyner Fuentes. "Studying the existence of a measurable difference in voice emotion expression after suffering from gender-based violence". Master Thesis. Université de Paris, 2022.

[310] I Camey, Laura Sabater, Cate Owren, A Boyer, and Jamie Wen. "Gender-based violence and environment linkages". In: *The Violence of Inequality; Wen, J., Ed.; IUCN: Gland, Switzerland* (2020).

[311] Hamida Khatri and Iheb Abdellatif. "A Multi-Modal Approach for Gender-Based Violence Detection". In: *2020 IEEE Cloud Summit*. 2020, pp. 144–149. DOI: 10.1109/IEEECloudSummit48914.2020.00028.

[312] Paul Fleming, Sofia Gruskin, Florencia Rojo, and Shari Dworkin. "Men's violence against women and men are inter-related: Recommendations for simultaneous intervention". In: *Social Science & Medicine* 146 (Oct. 2015). DOI: 10.1016/j.socscimed.2015.10.021.