Computación Afectiva Multimodal en Dispositivos wearable con Aplicaciones en la Detección de la Violencia de Género

TESIS DOCTORAL PRESENTADA POR

ESTHER RITUERTO GONZÁLEZ

17. COLECCIÓN:
Premio de la Delegación del Gobierno
contra la Violencia de Género
a Tesis Doctorales sobre Violencia contra la Mujer

Segundo premio - Año 2023

GOBIERNO DE ESPAÑA
MINISTERIO DE IGUALDAD

Centro de Publicaciones
C/ Alcalá, 37 - 28071 Madrid

NIPO en línea: 048-24-018-9

Correo electrónico: dgviolenciagenero@igualdad.gob.es

http://www.publicacionesoficiales.boe.es

Dedicada a todas las mujeres que trabajan para conseguir un mundo mejor.

Agradecimientos

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 tesis 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 tesis; 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 esta tesis haya salido adelante, por creer siempre en mí y ayudarme en mis momentos más difíciles. Además, quiero dar especialmente las gracias a mi familia, 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.

Agradecimientos por la financiación

Me gustaría agradecer a las siguientes instituciones el apoyo financiero recibido para completar esta tesis:

Consejería de Investigación e Innovación de la Comunidad de Madrid, por el proyecto de investigación EMPATIA-CM (referencia Y2018/TCS-5046).

Agencia Estatal de Investigación, por el Proyecto SAPIENTAE4Bindi (referencia PDC2021-121071-I00) financiado por MCINAEI10.13039/501100011033 y por la Unión Europea "NextGenerationEU/PRTR".

Programa Predoctoral YEI de la Comunidad de Madrid, por la beca de Personal Investigador Predoctoral en Formación (PEJD-2019-PRE/TIC-16295).

Ministerio de Universidades de España, para la beca FPU19/00448 de "Formación de Personal Universitario (FPU)" y la "Ayuda complementarias de movilidad Estancias Breves 2020 destinadas a beneficiarios del programa de Formación del Profesorado Universitario (FPU)"".

Servicio Alemán de Intercambio Académico (DAAD), para la beca de corta duración (Short-Term Grant) 2020.

Contenido Publicado y Enviado

A continuación, se enumeran los trabajos de contenido publicado y presentado. Algunas partes de las siguientes publicaciones se incluyen total o parcialmente en esta tesis:

Artículos de Revista

[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* (Bindi: Internet de las Cosas Affectivo para combater la Violencia de Género). En *IEEE Internet of Things*. (JCR Q1). Vol. 9, nº 21, pp. 21174-21193. DOI:10.1109/JIOT.2022.3177256

Esta publicación se incluye parcialmente en el Resumen, el Resumen ampliado, el Capítulo 1 (apartados 1.1.3 y 1.2.2), el Capítulo 5 (todos los apartados excepto 5.4.1, 5.4.2 y 5.6.1), el Capítulo 6 (apartado 6.1.1) y el Capítulo 7 (apartado 7.1). El material procedente de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

[2] Esther Rituerto-González, Carmen Peláez-Moreno. (2021) *End-to-end Recurrent Denoising Autoencoder Embeddings for Speaker Identification (*Representaciones de un Autocodificador Recurrente Eliminador de Ruido Extremo a Extremo para la Identificación de Hablantes). En *Neural Computing and Applications*, Springer. (JCR Q1). Vol. 33, no. 21, pp. 14429-14439. DOI:10.1007/s00521-021-06083-7

Esta publicación se incluye íntegramente en el Resumen ampliado y en el Capítulo 4 (Secs. 4.2, 4.4 y 4.6). El material procedente de esta fuente incluido en esta tesis no se destaca con medios tipográficos.

[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* (Aumento de Datos para la Identificación de Hablantes en Condiciones de Estrés para Combatir la Violencia de Género). En *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

Esta publicación se incluye íntegramente en el Resumen, Resumen ampliado, Capítulo 3 (Sec. 3.2.1) y Capítulo 4 (Secs. 4.1 - 4.3). El material de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

Artículos de Conferencias

[4] Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez Moreno. *Affective Acoustic Scene Analysis* (Análisis Acústico Afectivo de Escenas Sonoras). En la Conferencia

TECNIACUSTICA 2022. Disponible en: http://www.sea-acustica.es/fileadmin/Elche22/ID-65.pdf

Esta publicación se incluye parcialmente en el Capítulo 6 (Sec. 6.1.2). El material procedente de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

[5] Clara Luis-Mingueza, Esther Rituerto-González, Carmen Peláez Moreno. *Bridging the Semantic Gap with Affective Acoustic Scene Classification: an Information Retrieval-based Approach* (Salvando la Brecha Semántica con la Clasificación Acústica Afectiva de Escenas Sonoras: un enfoque basado en Métodos de Recuperación de Información). En la Conferencia IBERSPEECH 2022 DOI:10.21437/IberSPEECH.2022-19

Esta publicación se incluye parcialmente en el capítulo 6 (Sec. 6.1.2). El material procedente de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

[6] Emma Reyner-Fuentes, Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez Moreno. *Detecting Gender-based Violence Aftereffects from Emotional Speech Paralinguistic Features* (Detección de las Secuelas de la Violencia de Género a partir de Características Paralingüísticas del Habla Emocional). En la Conferencia IBERSPEECH 2022. DOI:10.21437/IberSPEECH.2022-20

Esta publicación se incluye parcialmente en el capítulo 6 (Sec. 6.3). El material procedente de esta fuente incluido en esta tesis no se destaca con medios tipográficos.

[7] Andreas Triantafyllopoulos, Sandra Ottl, Alexander Gebhard, Esther Rituerto-González, Mirko Jaumann, Steffen Hüttner, Valerie Dieter, Patrick Schneeweiß, Inga Krauß, Maurice Gerczuk, Shahin Amiriparian y Björn W. Schuller. *Fatigue Prediction in Outdoor Running Conditions using Audio Data* (Predicción de la Fatiga en Condiciones de Carrera al Aire Libre utilizando Datos de Audio). En *la International Conference on Engineering in Medicine and Biology* (Conferencia Internacional sobre Ingeniería en Medicina y Biología) (EMBC) 2022. DOI:10.1109/EMBC48229.2022.9871225. Disponible en: [Deep AI Online]

Esta publicación se incluye parcialmente en el capítulo 6 (Sec. 6.2). El material procedente de esta fuente incluido en esta tesis no se destaca con medios tipográficos.

[8] Esther Rituerto-González, Clara Luis-Mingueza, Carmen Peláez-Moreno. (2020). *Using Audio Events to Extend a Multimodal Public Speaking Database with Reinterpreted Emotional Annotations* (Uso de Eventos Sonoros para Ampliar una Base de Datos Multimodal con Anotaciones Emocionales Reinterpretadas). En la Conferencia IBERSPEECH 2020. DOI:10.21437/IberSPEECH.2021-13

Esta publicación se incluye parcialmente en el capítulo 3 (apartados 3.2.2 y 3.2.3), el capítulo 4 (apartado 4.5), el capítulo 5 (apartados 5.6.1 y 5.7) y el capítulo 6 (apartado 6.1.1). El material de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

[9] Esther Rituerto-González, José A. Miranda; Manuel F. Canabal; José M. Lanza-Gutiérrez; Carmen Peláez-Moreno. (2020). *A Hybrid Data Fusion Architecture for BINDI: a Wearable Solution to Combat Gender-based Violence* (Una Arquitectura de Fusión de Datos Híbrida para BINDI: una Solución *Wearable* para Combatir la Violencia de Género). En la *International Conference on Multimedia Communications, Services and Security* (Conferencia Internacional sobre Comunicaciones Multimedia, Servicios y Seguridad) (MCSS) 2020. DOI:10.1007/978-3-030-59000-0_17

Esta publicación se incluye parcialmente en el resumen, el resumen ampliado, el capítulo 3 (apartado 3.2.1), el capítulo 4 (apartado 4.2.1) y el capítulo 5 (apartados 5.1, 5.2.1, 5.2.4, 5.4 y 5.7). El material procedente de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

[10] Esther Rituerto-González, Ascensión Gallardo-Antolín, Carmen Peláez-Moreno. (2018). *Speaker Recognition under Stress Conditions* (Reconocimiento de Hablantes en Condiciones de Estrés). En la Conferencia IBERSPEECH 2018. DOI:10.21437/IberSPEECH.2018-4
Esta publicación se incluye parcialmente en el capítulo 3 (Sec. 3.2.1) y en el capítulo 4 (Secs. 4.1 - 4.3). El material procedente de esta fuente incluido en esta tesis no está singularizado con medios tipográficos.

Bases de Datos

[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 Multimodal Affective Computing dataset* (WEMAC: Base de datos de Computación Afectiva Multimodal de Mujeres y Emociones). Preprint en arXiv. DOI:10.48550/arXiv.2203.00456

Esta publicación se incluye parcialmente en el capítulo 3 (Sec. 3.3). El material procedente de esta fuente incluido en esta tesis no se destaca con medios tipográficos.

[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, y C. López Ongil. "Base de datos UC3M4Safety - Lista de estímulos audiovisuales", 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, y C. López Ongil. "Base de datos UC3M4Safety - Lista de estímulos audiovisuales (vídeo)", 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, y C. López Ongil. "Base de datos UC3M4Safety WEMAC: Cuestionario biopsicosocial y consentimiento informado", 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 y C. López Ongil. "Base de datos UC3M4Seguridad - WEMAC: Señales fisiológicas", 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, y C. López Ongil. "Base de datos UC3M4Safety - WEMAC: Características de Audio", 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, y C. López Ongil, "Base de datos UC3M4Safety - WEMAC: Etiquetado emocional", 2022. DOI:10.21950/RYUCLV

Archivo de Libre Acceso

[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* (Cambio Climático y Audición por Ordenador: Una llamada a la Acción y una Visión General sobre la Inteligencia Artificial de Audio para Ayudar a Salvar el Planeta). *Preprint* en arXiv. DOI:10.48550/arXiv.2203.06064

Esta publicación se incluye parcialmente en el capítulo 6 (Sec. 6.4). El material procedente de esta fuente incluido en esta tesis no se destaca con medios tipográficos.

Otros méritos de la investigación

Las siguientes contribuciones formaron parte de la investigación realizada durante este doctorado, pero no se incluyen en esta tesis:

Comunicaciones

[13] Esther Rituerto-González. "Análisis de Sesgos en Inteligencia Artificial que reflejan Estereotipos de Género y Patrones Sociales de Desigualdad Estructural". VII Congreso Internacional de Jóvenes Investigadores con Perspectiva de Género en la Universidad Rey Juan Carlos (URJC), Madrid. Junio 2022. [Programa del Congreso]

[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. "Descripción de la base de datos UC3M4Safety", 2021. Disponible en: http://hdl.handle.net/10016/32481

[15] Esther Rituerto-González, Carmen Peláez-Moreno. "Caracterización del Miedo a través de la Clasificación Acústica de Escenas Sonoras". II Congreso Internacional I+D+I Tecnología para la Igualdad de Género: Soluciones, Perspectivas y Retos, en UC3M, Getafe. Mayo de 2022. [Programa de la Conferencia]

[16] Esther Rituerto-González. "Ciencia Abierta: Un Concepto Revolucionario". (2021, 28 junio). [Entrada en el blog]

[17] Esther Rituerto-González. "La Influencia del Sexo y el Género en la Inteligencia Artificial". VI Congreso Internacional de Jóvenes Investigadores con Perspectiva de Género en la UC3M, Getafe. Junio 2021. [Programa de la Conferencia]

[18] Esther Rituerto-González, Carmen Peláez-Moreno. "Detección del Miedo en el Habla". I Congreso Internacional I+D+I Tecnología para la Igualdad de Género: Soluciones, Perspectivas y Retos, en UC3M, Getafe. Abril 2021. [Programa del Congreso]

[19] Esther Rituerto-González. "Reconociendo la Voz de la Víctima en Situaciones de Violencia de Género: Un enfoque de Aprendizaje Automático". V Congreso Internacional de Jóvenes Investigadores con Perspectiva de Género en la UC3M, Getafe. Junio 2020. [Programa de la Conferencia]

[20] Esther Rituerto-González. "Dispositivo *wearable* para combatir la Violencia de Género". IV Congreso Internacional de Jóvenes Investigadores con Perspectiva de Género en la UC3M, Getafe. Junio de 2019. [Programa de la Conferencia]

Resumen Ampliado

MOTIVACIÓN

Según la Organización Mundial de la Salud (OMS), 1 de cada 3 mujeres sufre violencia física o sexual en algún momento de su vida, lo que refleja el efecto de la Violencia de Género (VG) en el mundo. En concreto en España, más de 1,100 mujeres han sido asesinadas desde 2003 hasta 2022, víctimas de la violencia de género.

La violencia de género, en todas sus formas, provoca traumas psicológicos que tienen consecuencias físicas y de comportamiento en las mujeres; las supervivientes pueden sufrir depresión y tienen un mayor riesgo de suicidio. Por lo tanto, es urgente encontrar soluciones a corto y medio plazo a este problema imperante y generalizado en nuestra sociedad, siendo este el objetivo de los proyectos en los que se enmarca esta tesis.

Las soluciones a la violencia de género pueden pasar por una inversión adecuada en investigación tecnológica, además de en esfuerzos legislativos, educativos y económicos. Pero a pesar de los esfuerzos tecnológicos, varios expertos en violencia de género cuestionan las soluciones existentes hasta la fecha y las consideran anticuadas. Estos expertos exigen una investigación más avanzada en tecnología para las soluciones a la violencia de género. Y a pesar de los impresionantes avances de la Inteligencia Artificial (IA), aún no existen soluciones tecnológicas para la detección automática de situaciones de riesgo para la vida de las mujeres que incorporen dicha inteligencia.

Una solución impulsada por la IA que realizara un análisis exhaustivo de aspectos como el estado emocional de la persona, además de un análisis del contexto o de la situación externa (por ejemplo, las circunstancias, la ubicación) y, por tanto, detectara automáticamente si la integridad de una mujer está en peligro, podría garantizar su seguridad.

El sistema Bindi descrito en esta tesis se concibe como una solución impulsada por la IA, discreto y *wearable*, enfocado a la detección automática de situaciones de violencia de género. Consta de dos dispositivos inteligentes ocultos en abalorios que monitorizan variables fisiológicas y graban la escena acústica, incluida la voz. Están conectados a un teléfono inteligente con una aplicación con un núcleo impulsado por IA que puede producir diferentes tipos de alertas, así como encriptar y enviar la información a un servidor protegido. Bindi es una tecnología de vanguardia que combina la Computación Afectiva (CA) inteligente con la adquisición y fusión de señales multisensoriales físicas y fisiológicas, y una infraestructura de servidor segura para detectar de forma autónoma las situaciones de riesgo, registrando los datos para posteriores acciones legales.

Así, esta tesis doctoral se basa en la detección de situaciones de riesgo de violencia de género para las mujeres, abordando el problema desde un punto de vista multidisciplinar al aunar las tecnologías de Inteligencia Artificial (IA) y la perspectiva de género. Más concretamente,

dirigimos el foco a la modalidad auditiva, que captamos con Bindi, analizando los datos del habla producidos por la usuaria, ya que la voz puede ser grabada de forma discreta y puede ser utilizada como identificador personal e indicador del estado afectivo reflejado en ella.

BASES DE DATOS DE EMOCIONES

En esta tesis también damos una visión general de las bases del afecto, el estado de ánimo y las emociones. Además, describimos de dónde surgen las diferentes teorías de la emoción en las ciencias cognitivas, sus aplicaciones actuales y algunas consideraciones éticas importantes. Explicamos también su relación con el campo de la Computación Afectiva, que es el encargado de estudiar y desarrollar dispositivos y sistemas tecnológicos que puedan reconocer, procesar, simular e interpretar las emociones y los afectos humanos.

Cuando un ser humano que se ve en una situación potencialmente peligrosa, la respuesta inmediata del cuerpo es la respuesta de *fight-flight-freeze* (lucha-huida-congelación, LHC o *FFF* por sus siglas en inglés). Esta respuesta desencadena automáticamente una reacción fisiológica que se produce cuando se reconoce un acontecimiento como amenazante para la vida. Es una respuesta de defensa activa en la que la persona o bien lucha contra la amenaza, o bien huye de ella, o bien se paraliza o bloquea. El cuerpo se ve afectado por cambios fisiológicos involuntarios con el fin de preparar a la persona para actuar de forma adecuada y rápida.

En respuesta al peligro percibido, el Sistema Nervioso Autónomo inicia una reacción en cadena que implica toda una serie de cambios en el ritmo cardiaco, la respiración y la activación muscular, incluido el aparato de producción del habla, para hacer frente al desafío del momento. La tensión muscular puede llevar a tener la garganta y las cuerdas vocales constreñidas y dar lugar a que la voz de una persona se vuelva aguda, grave o incluso a la ausencia total de voz. La constricción muscular también puede provocar un aumento de la velocidad del habla, tensión en la mandíbula y la lengua, lo que dificulta la inteligibilidad, y el cierre de la salivación hace que la boca se sienta seca y puede producir una voz ronca.

Debido a todos estos cambios físicos y fisiológicos y a su carácter involuntario -la persona no tiene control sobre ellos- que se producen como consecuencia de encontrarse en una situación de riesgo, nos planteamos basarnos en señales fisiológicas como el pulso, la transpiración, la respiración, y también el habla, para detectar el estado emocional de una persona con Bindi, con la intención de reconocer la emoción del *miedo*, que podría ser consecuencia de encontrarse en una situación peligrosa. Un ejemplo de situación de amenaza para la vida que podría desencadenar las respuestas de lucha-huida-congelación en las mujeres son aquellas en las que una mujer sufre una agresión física o sexual, que son situaciones de violencia de género.

Sin embargo, todavía no existe un consenso científico sobre una única teoría válida de la naturaleza de las emociones. Esto se debe a que las emociones son muy subjetivas y no existe

una forma objetiva de categorizarlas y cuantificarlas. Por tanto, su inferencia es todo un reto. El etiquetado o identificación de las emociones -depende de quién lo haga- depende, en primer lugar, de la dificultad intrínseca de interpretar los sentimientos más íntimos de uno mismo. Segundo, de cuánto exteriorizamos una emoción. Y tercero, de cómo cada persona interpreta una determinada situación, lo cual pueden dar lugar a diferentes emociones. Otro reto a tener en cuenta es el de la personalización de género. Parece que existen claras diferencias en la expresión de las emociones según el sexo. Se ha descubierto que las expresiones estereotipadas según el género, -derivadas de la socialización de género-, se manifiestan de forma diferente en hombres y mujeres, ya que los hombres expresan con más frecuencia la ira y el desprecio que las mujeres, que expresan con más frecuencia el *miedo* que los hombres.

CARACTERIZACIÓN DE DATOS

Idealmente, para construir nuestro sistema de IA de Aprendizaje Automático (*Machine Learning, ML*) o Aprendizaje Profundo (*Deep Learning, DL*) para la detección automática de situaciones de riesgo a partir de datos de audio, nos gustaría contar con datos de habla grabada en condiciones reales, propias de la usuaria y escenario objetivo.

En nuestros primeros pasos, nos encontramos con la dificultad de la falta de datos adecuados disponibles, ya que los conjuntos de datos de habla de *miedo real* (no actuado) no estaban disponibles o no existían en el estado del arte. El habla real, espontánea, en situaciones de la vida real y bajo condiciones emocionales, son las categorías ideales que necesitábamos para nuestra aplicación. Pero al no encontrar tales datos, decidimos elegir el *estrés* como la emoción más cercana al escenario objetivo posible, con el fin de poder dar cuerpo a los algoritmos de IA. Así, describimos y justificamos el uso de conjuntos de datos que contienen dicha emoción como punto de partida de nuestra investigación.

Adicionalmente, y como consecuencia del problema anterior, describimos una de las principales aportaciones del equipo UC3M4Safety que es la creación de nuestra propia base de datos para cubrir dicho nicho bibliográfico. Con ello teníamos la intención de recoger variables humanas ante estímulos emocionales que puedan servir en sistemas de IA o ML/DL para distinguir emociones de forma automática y en tiempo real, especialmente las emociones de *miedo* en las mujeres.

En primer lugar, nuestro equipo capturó la "*UC3M4Safety Audiovisual Stimuli Database*" (Base de datos de estímulos audiovisuales UC3M4Safety). Se trata de un conjunto de datos de alta calidad de estímulos audiovisuales, para elicitar hasta 12 emociones diferentes durante su visualización -incluido *el miedo*- en un escenario controlado. Contiene un conjunto de datos de 42 estímulos audiovisuales validados con una categorización emocional

discreta y continua por más de 50 anotadores cada uno en un entorno de *crowdsourcing* con un alto nivel de acuerdo entre ellos.

En segundo lugar, contribuimos a la comunidad con la recopilación de "WEMAC", una base de datos multimodal que comprende un experimento de laboratorio realizado con mujeres voluntarias que visualizan la base de datos de estímulos audiovisuales UC3M4Safety. Su objetivo es inducir emociones reales utilizando un *headset* (casco con auriculares y gafas) de realidad virtual mientras capturan variables fisiológicas, señales del habla y auto-evaluaciones con respuestas emocionales de las usuarias. La realidad virtual se utiliza para maximizar la experiencia inmersiva y, en consecuencia, lograr una mejor elicitación de las emociones.

La base de datos está formada por 101 mujeres voluntarias que nunca sufrieron violencia de género y 43 mujeres voluntarias víctimas de violencia de género. Este último grupo realizó el experimento bajo la supervisión de una psicóloga. Se seleccionaron 28 estímulos audiovisuales de la base de datos de estímulos audiovisuales UC3M4Safety para ser visualizados, algunos de ellos siendo vídeos estereoscópicos de 360º. Justo después de cada visualización de cada videoclip emocional, se pide a las voluntarias que respondan en voz alta a dos preguntas sobre cada vídeo, para hacer que las voluntarias revivan las emociones que han sentido durante la visualización del vídeo, con el objetivo de captar rastros de dicha emoción en su voz.

Además de la voz, las voluntarias etiquetan sus reacciones emocionales tras la visualización de cada vídeo con un joystick. Utilizan los "Maniquíes de autoevaluación modificados (SAM)" para anotar etiquetas continuas de emoción (Valencia/Placer, Excitación y Dominancia) y una etiqueta de emoción discreta de un total de 12 categorías emocionales.

Publicamos la primera versión de la base de datos WEMAC con el objetivo de compartirla con la comunidad investigadora, fomentar la mejora de los resultados de referencia obtenidos y avanzar en la investigación del análisis multimodal de las emociones en general y, en la igualdad de género, en particular. Sin embargo, no podemos publicar o divulgar las señales de voz en bruto por cuestiones éticas y de privacidad, por lo que hemos procesado las señales de voz y extraído características o descriptores de bajo y alto nivel ampliamente utilizados en la literatura de las tecnologías de habla para que la comunidad investigadora pueda analizarlos y trabajar con ellos.

WEMAC aún está lejos de las condiciones de la vida real, ya que grabar el habla en condiciones de *miedo* que sea realista y espontánea es muy difícil, si no imposible. Para acercarnos lo más posible a estas condiciones y tal vez registrar el habla con *miedo*, el equipo UC3M4Safety creó la base de datos "WE-LIVE". Con ella hemos apturado señales fisiológicas, auditivas y contextuales de mujeres en un entorno no controlado (condiciones del mundo real), así como el etiquetado de sus reacciones emocionales ante acontecimientos de su vida cotidiana, utilizando el actual sistema Bindi (pulsera, colgante, aplicación móvil y servidor). Esta base de datos está compuesta por 13 mujeres voluntarias, algunas de ellas víctimas de la violencia de género.

IDENTIFICACIÓN DEL HABLANTE

Para nuestro objetivo de detectar situaciones de riesgo de violencia de género a través del habla, primero necesitamos detectar la voz de la usuaria concreto que nos interesa, una tarea de identificación o reconocimiento del hablante, entre toda la información contenida en la señal de audio. Así, pretendemos rastrear la voz de la usuaria, separada del resto de los hablantes del escenario acústico, intentando evitar la influencia de las emociones o del ruido ambiente en la identificación del hablante, ya que estos factores podrían perjudicarla. Pero el rendimiento de los modelos ML para la detección de hablantes a través de la voz desciende mucho cuando se encuentran en condiciones emocionales. Así pues, el hecho de que la voz de una hablante pueda verse influida por su estado emocional constituye un reto para un sistema de identificación de hablantes.

Estudiamos los sistemas de identificación de hablantes en dos condiciones de variabilidad, 1) en condiciones de estrés, para ver en qué medida estas condiciones de estrés afectan a los sistemas de reconocimiento de hablante -en ausencia de bases de datos del habla en condiciones de *miedo* realistas en ese momento en la literatura-, y 2) en condiciones de ruido real, aislando la identidad de la hablante, entre toda la información adicional contenida en la señal de audio.

Con nuestros estudios, comprobamos que el habla estresada afecta negativamente cuando los sistemas de reconocimiento o identificación del hablante solo han sido entrenados o configurados para habla neutra. En cuanto al caso de la configuración mixta -utilizando los enunciados originales neutros y estresados tanto para el entrenamiento como para las pruebas- el sistema alcanza una tasa muy satisfactoria de reconocimiento de hablante, lo que demuestra que el conjunto de características de habla elegidas para la tarea es adecuado.

En cuanto a nuestros experimentos en los que aumentamos los datos mediante la generación sintética de habla estresada, podemos concluir que dicha generación mediante modificaciones en el tono y la velocidad, añadidos a la base de datos utilizada, amplía significativamente las instancias con las que trabajar, mejorando sustancialmente los resultados de precisión obtenidos por el sistema de identificación.

En la línea del campo de la identificación de hablantees en condiciones reales, estudiamos cómo el habla grabada en condiciones reales, incluyendo ruido ambiental, es perjudicial para los sistemas de reconocimiento de hablante, por lo que exploramos cómo eliminarlo con métodos eficaces de eliminación de ruido para lograr las mejores prestaciones.

Utilizamos *embeddings* (representaciones) de la identidad de los hablantes extraídos de un autocodificador recurrente que elimina el ruido de las señales de audio combinado con una red neuronal superficial que actúa como clasificador para la tarea de identificación de hablantes.

La arquitectura de extremo a extremo propuesta utilizó un bucle de retroalimentación para codificar la información relativa al hablante en representaciones de baja dimensión extraídas por un autocodificador de espectrogramas. Empleamos técnicas de aumento de datos corrompiendo aditivamente el habla limpia con ruido ambiental de la vida real en una base de datos que contenía habla real estresada para mejorar las prestaciones.

Nuestra arquitectura propuesta consigue resultados fiables para toda la gama de señales contaminadas a diferentes niveles Señal a Ruido (*signal-to-noise ratio, SNR*), siendo un enfoque más robusto que el resto de las arquitecturas probadas, especialmente en SNR más bajas (cuando el ruido es más alto). En las tablas de resultados, se observaron tasas de identificación del hablante más bajas al realizar las pruebas con señales de habla estresadas, lo que muestra las dificultades de la tarea inducidas por el estrés. Esto sugiere la necesidad de tener en cuenta específicamente las distorsiones causadas por el habla emocional para las tareas de identificación de hablantes.

DETECCIÓN DE EMOCIONES

En esta tesis, también nos sumergimos en el desarrollo del sistema Bindi para el reconocimiento de emociones *relacionadas con el miedo*, su detección y clasificación en un enfoque altamente multidisciplinar ya que hay muchas contribuciones apoyadas por otros miembros del equipo UC3M4Safety.

En primer lugar, describimos las arquitecturas de las versiones 1.0 y 2.0 de Bindi, la evolución de una a otra, junto con su implementación. Explicamos el enfoque seguido para el diseño de un sistema multimodal en cascada para Bindi 1.0, y también el despliegue de un sistema completo del Internet de las Cosas con componentes de computación en el *edge*, *fod* y *cloud*, para Bindi 2.0; detallando específicamente cómo diseñamos las arquitecturas de inteligencia en los dispositivos Bindi para la detección *del miedo* en la usuaria. En un estudio adicional con datos de la base de datos *Biospeech*, demostramos que ampliar nuestra base de datos con eventos acústicos estresantes es incluso beneficioso para el reconocimiento del estrés en el habla y el audio.

A continuación, describimos nuestra experimentación monomodal con el habla para la detección de emociones *relacionadas con el miedo*, centrándonos primero en la detección del estrés real (no actuado). Posteriormente, como núcleo de la experimentación, trabajamos con WEMAC para la tarea de detección *del miedo* con estrategias de fusión de datos. Hay un fuerte componente multimodal, ya que trabajamos en la parte de reconocimiento de emociones a partir del habla junto con datos de señales fisiológicas, del mismo modo que lo harían los dos dispositivos *wearable* de Bindi. Utilizamos tres estrategias de fusión de datos multimodales que se evalúan y validan. Los resultados experimentales muestran una precisión media del reconocimiento del *miedo* del 63,61% con el método *Leave-hAlf-Subject-Out* (LASO), que es una estrategia de clasificación de entrenamiento dependiente de la persona y adaptada al hablante.

Por lo que sabe el equipo de UC3M4Safety, es la primera vez que se presenta una fusión multimodal de datos fisiológicos y del habla para el reconocimiento del *miedo* en este contexto de violencia de género. Además, es la primera vez que se presenta un modelo LASO que tiene en cuenta el reconocimiento del *miedo*, la fusión de señales multisensoriales y los estímulos de realidad virtual.

LÍNEAS DE INVESTIGACIÓN ADICIONALES SOBRE EL AUDIO Y LA VIOLENCIA DE GÉNERO

Finalizamos esta tesis con algunas líneas de investigación paralelas y complementarias que se abrieron en colaboración con otros miembros del grupo de investigación y que podrían ser de ayuda en la prevención de la violencia de género. Explicamos el trabajo realizado en el campo del "análisis acústico de escenas" y la importancia del análisis de eventos acústicos para la detección de situaciones de riesgo. Definimos el término de "Análisis Acústico *de Escenas Afectivas*" y con él la necesidad de unificar los trabajos realizados sobre escenas acústicas y las emociones que éstas inducen bajo un mismo título, con el fin de elevar la investigación en esta línea y hacerla avanzar. También estudiamos *embeddings* acústicos robustos e interpretables que caracterizan las emociones en la base de datos *UC3M4Safety Audiovisual Stimuli*.

Además, realizamos un breve estudio de la expresión de fatiga en el habla y la voz, observando los resultados por género, y en futuras investigaciones sería interesante caracterizarla para ver las diferencias entre el estrés, el *miedo* y la fatiga sobre las variables fisiológicas y sus efectos en la voz. Asimismo, exploramos cómo la condición de víctima de violencia de género podría detectarse únicamente mediante la información paralingüística del habla. Por último, y siguiendo otro objetivo alineado con el bien social, exploramos brevemente la relación entre la violencia de género y el cambio climático.

CONCLUSIONES

Gran parte del trabajo técnico de esta tesis -al igual que el del resto de los miembros del equipo UC3M4Safety- se lleva a cabo teniendo en cuenta la perspectiva de género. Esto, hasta donde el equipo sabe, es la primera vez que se hace en investigación tecnológica con inteligencia artificial, por lo que podemos considerar que esta investigación se encuentra en una fase preliminar en la que estamos sentando las bases, y en la que pretendemos seguir trabajando en el futuro. En general, esta tesis explora el uso de la tecnología y la inteligencia artificial para prevenir y combatir la violencia de género. Esperamos haber abierto el camino en la comunidad investigadora y más allá, y que nuestra experimentación, resultados y conclusiones puedan ayudar en futuras investigaciones.

El objetivo último de este trabajo es despertar el interés de la comunidad por desarrollar soluciones al problema tan difícil de la violencia de género.

Para concluir la tesis consideramos algunas opciones para trabajos futuros, como el uso de WEMAC y WE-LIVE en arquitecturas de aprendizaje profundo más complejas para desentrañar la identidad del hablante y la información emocional con (por ejemplo, un modelo adversarial), para realizar la detección del hablante y la emoción de forma conjunta.

En los términos generales del desarrollo del Bindi, también debemos tener en cuenta que muchas mujeres permanecen en estado de shock cuando son agredidas o se convierten en víctimas de una agresión, en lugar de producir habla. Debemos tener en cuenta este hecho para futuros desarrollos del sistema Bindi, analizando la aparición de silencios en el audio, junto con las demás variables que ya se han explorado.

En cuanto al análisis de los eventos acústicos y el contexto acústico dentro de Bindi, también es de especial interés la detección de eventos vocales como gritos, gruñidos, respiraciones pesadas, chillidos, y también eventos acústicos como golpes, choques o impactos, que probablemente denotarían que se está produciendo una situación peligrosa.

El análisis de las emociones, en particular del miedo, y de la condición de la violencia de género que se discute en esta tesis también podría ayudar a la investigación de la IA de audio orientada a la salud, en particular con aplicaciones en la atención a la salud mental y la psicoterapia.

Índice

Lista de Figuras

Lista de Tablas

Lista de Abreviaturas

Nota: Durante todo el documento se mantendrán los acrónimos en inglés como criterio, debido a que la mayoría de ellos corresponden a términos técnicos reconocidos en la literatura y estado del arte internacionalmente en el ámbito de la tecnología en este idioma.

Abreviatura en Inglés	Descripción en Inglés	Descripción en Español
AC	Affective Computing	Computación Afectiva
AED/C	Acoustic Event Detection / Classification	Detección / clasificación de eventos acústicos
AI	Artificial Intelligence	Inteligencia artificial
ANS	Autonomic Nervous System	Sistema nervioso autónomo
ASD	Acute Stress Disorder	Trastorno por estrés agudo
ASR	Automatic Speech Recognition	Reconocimiento automático del habla
BT	Bluetooth	Bluetooth
BVP	Blood Volume Pulse	Pulso del volumen sanguíneo
CNN	Convolutional Neural Network	Red neuronal convolucional
CP	Computational Paralinguistics	Paralingüística computacional
CVAWG	Cyberviolence against Women and Girls	Ciberviolencia contra mujeres y niñas
DAE	Denoising Auto-encoder	Autocodificador de eliminación de ruido
DNN	Deep Neural Network	Red neuronal profunda
EEG	Electroencephalogram	Electroencefalograma
EIGE	European Institute of Gender Studies	Instituto Europeo de Estudios de Género
EMG	Electromyography	Electromiografía
EU	European Union	Unión Europea
FFF	Fight-Flight-Freeze	Lucha-Huida-Congelación
FFT	Fast Fourier Transform	Transformada rápida de Fourier
GBV	Gender-based Violence	Violencia de género
GBVV	Gender-based Violence Victim	Víctima(s) de violencia de género
GRU	Gated Recurrent Unit	Unidad recurrente con puerta
GSR	Galvanic Skin Response	Respuesta galvánica de la piel
HR	Heart Rate	Frecuencia cardiaca
IUCN	International Union for the Conservation of Nature	Unión Internacional para la Conservación de la Naturaleza
IoB	Internet of Bodies	Internet del cuerpo cuerpos
IoT	Internet of Things	Internet de las cosas
IPV	Intimate Partnert Violence	Violencia de pareja
LEAs	Law Enforcement Agencies	Agencias policiales
LSTM	Long Short-term Memory	Memoria a corto plazo

MAE	Mean Absolute Error	Error absoluto medio
MFCC	Mel Frequency Cepstral Coefficients	Coeficientes cepstrales de frecuencia mel
ML	Machine Learning	Aprendizaje automático
MSE	Mean Squared Error	Error cuadrático medio
NLP	Natural Language Processing	Procesamiento del lenguaje natural
PAD	Pleasure Arousal Dominance	Placer Excitación Dominancia
PNS	Parasympathetic Nervous System	Sistema nervioso parasimpático
PTSD	Post-traumatic Stress Disorder	Trastorno de estrés postraumático
RDAE	Recurrent Denoising Auto-encoder	Autocodificador recurrente de eliminación de ruido
RMS	Root Mean Square	Raíz cuadrática media
SD	Speaker Dependant	Dependiente del hablante
SDG	Sustainable Development Goals	Objetivos de Desarrollo Sostenible
SER	Speech Emotion Recognition	Reconocimiento de las emociones del habla
SI	Speaker Independent	Independiente del hablante
SI	Speaker Identification	Identificación del hablante
SKT	Skin Temperature	Temperatura de la piel
SNN	Shallow Neural Network	Red neuronal superficial
SNR	Signal to Noise Ratio	Relación Señal a Ruido
SNS	Sympathetic Nervous System	Sistema Nervioso Simpático
SR	Speaker Recognition	Reconocimiento de hablante
SV	Speaker Verification	Verificación de hablante
TRL	Technology Readiness Level	Nivel de Madurez Tecnoloógica
TTS	Text to Speech	Texto a Habla
UN	United Nations	Naciones Unidas
UTC	Universal Time Coordinated	Tiempo Universal Coordinado
WHO	World Health Organization	Organización Mundial de la Salud

Capítulo 1: Introducción a la Violencia de Género

Este capítulo presenta la motivación y el contexto de esta Tesis Doctoral. La primera parte da respuesta a preguntas sobre el problema que queremos resolver y la justificación de la relevancia de este trabajo. La segunda parte abarca aspectos como los retos a los que nos enfrentamos tanto desde el punto de vista técnico como social.

1.1. Motivación

Según la Organización Mundial de la Salud (OMS), 1 de cada 3 mujeres sufre violencia sexual o física en algún momento de su vida, lo que refleja el efecto de la Violencia de Género (GBV) en el mundo [21]. Concretamente en España, más de 1.100 mujeres han sido asesinadas desde 2003 hasta 2022, víctimas de la violencia de género [22]. La discriminación de género y su manifestación de esta violencia, son un problema omnipresente en nuestra sociedad que afecta al 50% de la población mundial.

Según el Instituto Europeo de la Igualdad de Género (EIGE) el término de violencia de género se define como "la violencia dirigida hacia una persona por razón de su sexo". También afirman que "tanto las mujeres como los hombres sufren violencia de género, pero la gran mayoría de las víctimas son mujeres y niñas, y la mayoría de los agresores son hombres" [23]. Por lo tanto, a lo largo de esta tesis, utilizaremos indistintamente los términos violencia de género y violencia contra las mujeres, ya que creemos que utilizar el término "basada en el género" pone el foco en las desigualdades de poder existentes entre hombres y mujeres, que es el origen de la violencia de género.

La violencia de género se presenta bajo muchas manifestaciones distintas no excluyentes entre sí, varias incidencias de violencia pueden ocurrir al mismo tiempo y reforzarse mutuamente. Los actos de violencia pueden dirigirse contra personas que sufren desigualdades, como las relacionadas con su edad, raza, discapacidad, religión, clase social o sexualidad. Así pues, la violencia y la discriminación a las que se enfrentan las mujeres no sólo se basan en el género, sino que también experimentan formas de violencia diversas e interrelacionadas [24].

La violencia contra las mujeres puede englobarse en cuatro formas clave de violencia, lo que favorece una comprensión exhaustiva de lo que se considera violencia de género. Estas formas son: física, sexual, psicológica y económica [25].

- Violencia física: Toda fuerza física ilícita que provoque cualquier tipo de daño físico. La violencia física puede manifestarse como agresiones graves y leves, limitación de la libertad y, en última instancia, homicidio, entre otras.
- Violencia sexual: Cuando se realiza cualquier acto sexual sobre una persona en contra de su voluntad, ya sea cuando no puede dar su consentimiento o cuando no lo da

explícitamente -ya sea porque la persona tiene una discapacidad mental, es una niña o está inconsciente o intoxicada como consecuencia de las drogas o el alcohol- [26]. Puede adoptar la forma de agresión sexual o violación.

- Violencia psicológica o emocional: Cualquier acción o comportamiento que cause daño psicológico a una persona, provocando *miedo* mediante la intimidación o saboteando el sentido de autoestima de una persona mediante la crítica continua. La violencia psicológica puede manifestarse, por ejemplo, como intimidación, difamación, acoso, humillación o insulto verbal.
- Violencia económica: Todas las acciones o comportamientos que causan un perjuicio económico a una persona, haciéndola dependiente económicamente, manteniendo un control parcial o total sobre sus recursos financieros. La violencia económica puede manifestarse como restricción del acceso a la educación, a los recursos financieros o al mercado laboral; daños a la propiedad, o incumplimiento de las responsabilidades económicas -como la pensión alimenticia- [27] entre otras.

En esta era del espacio digital han surgido recientemente nuevos tipos de discriminación y violencia contra las mujeres, que actualmente se definen como *ciberviolencia* contra las mujeres y las niñas (CVAWG). Este tipo de *ciberviolencia* incluye acciones como la pornografía no consentida (*revenge porn*), el ciberacoso, el "*slut-shaming*", el "*doxing*", la pornografía no solicitada, las amenazas de violación y de muerte, las difamaciones y el acoso por razón de sexo y la "sextorsión" [28]. La CVAWG es una continuación de la violencia que se produce *offline*. Por ejemplo, el ciberacoso por parte de una expareja o pareja tiene consecuencias similares al acoso *offline* y se considera el mismo tipo de violencia de pareja. La única diferencia es que se ve facilitado por la tecnología. Así pues, la CVAWG puede manifestarse como múltiples formas de violencia, incluida la psicológica y la sexual. Las nuevas tendencias también señalan que está aumentando la violencia económica, que se produce, por ejemplo, cuando la situación laboral (o el futuro empleo) de la víctima se ve en peligro por la información que se difunde en Internet.

También debe tenerse en cuenta la importancia de que la violencia en el ciberespacio se manifieste también psíquicamente [29]. En los últimos años, la pandemia del COVID-19 ha agravado los riesgos de ciberviolencia contra niñas y mujeres. Según [30], "el uso de Internet ha aumentado entre un 50% y un 70% con respecto a los niveles en que se utilizaba antes de la pandemia, y esta mayor vulnerabilidad ha provocado una pandemia fantasma de violencia de género en línea".

Las desigualdades estructurales son una de las causas de que la violencia de género se normalice y se reproduzca. Éstas son las normas, actitudes y estereotipos sociales en torno al género en la sociedad. Por lo tanto, es importante reconocer la violencia estructural e institucional cuando se intenta explicar la omnipresencia de la violencia de género en nuestra sociedad. Ésta se define como "la subordinación de la mujer en la vida económica, social y política" [24].

Gran parte de la violencia contra las mujeres es invisible, apenas se denuncia debido a la vergüenza y la estigmatización que sufren las víctimas y a la impunidad de la que gozan los agresores. La violencia de género no es un problema individual, sino un fenómeno social que se entrecruza en muchos ámbitos diferentes de la vida, en el que la violencia invisible es la base sobre la que se sustentan las formas de violencia más mortíferas. La desigualdad de género es lo que se esconde bajo la superficie, por lo que es esencial que la sociedad reconozca y admita los tipos de violencia visible e invisible, representados en el iceberg de la violencia de la Fig. 1.1, para desarmar el marco social y cultural que perpetra dicha violencia.

Figura 1. 1: Metáfora del iceberg de las formas visibles e invisibles de la violencia de género. Ilustración basada en [31].

La violencia de género, en todas sus formas, provoca traumas psicológicos que tienen consecuencias físicas y de comportamiento; las supervivientes pueden padecer depresión y corren un mayor riesgo de suicidio. Por lo tanto, urge encontrar soluciones a corto y medio plazo a este problema imperante y generalizado en nuestra sociedad, siendo este último el propósito de los proyectos en los que se enmarca esta tesis.

1.1.1. Marco Económico de la Violencia de Género en Europa

Según el Instituto Europeo de la Igualdad de Género (EIGE), la Unión Europea (EU) gasta 366.000 millones de euros al año en las consecuencias de la violencia de género [32]. Este estudio actualizado basado en el informe de 2014 titulado *Estimación de los costes de la violencia de género en la Unión Europea* [33] ofrece "estimaciones revisadas de los costes de la violencia de género y de pareja en la EU".

Extrapolando los resultados del estudio del caso del Reino Unido a la Unión Europea, "el coste estimado de la violencia de género contra las mujeres en la EU-27 fue de más de 290.000 millones de euros, lo que representa el 79% de todos los costes de la violencia de género tanto contra las mujeres como contra los hombres". Además, "el coste estimado de la violencia de pareja contra las mujeres en la EU-27 fue de casi 152.000 millones de euros, lo que representa el 87% de todos los costes de la violencia de pareja tanto contra las mujeres como contra los hombres" [33].

Los diferentes costes de la violencia de género en el informe se desglosan de la siguiente manera, siendo el mayor coste el impacto emocional y físico en las víctimas (56%) entendido como el daño que sufrieron como consecuencia del delito, seguido de los servicios de justicia penal (21%) (sistema judicial y policía) y la pérdida de producción económica (14%) explicada como un aumento general de los datos de incidencia tanto para mujeres como para hombres. Otros costes a tener en cuenta pueden incluir los servicios de justicia civil (por ejemplo, los procedimientos de divorcio y custodia de los hijos) y el apoyo financiero a la vivienda y los servicios de protección de la infancia. En concreto, durante la pandemia de Covid-19 y como una de las consecuencias de las restricciones de cierre, "la violencia de pareja se disparó, representando casi la mitad (48%, 174.000 millones de euros) del coste de la violencia de género". De ellos, "la violencia de pareja contra las mujeres representa el 87% de esta suma (151.000 millones de euros)".

Incluso cuando las cantidades que se tienen en cuenta son del orden de miles de millones de euros, el dinero que se destina a apoyar a las víctimas de la violencia de género (GBVV) no es suficiente, ya que servicios como los refugios para mujeres en situación de violencia sólo representan el 0,4% del coste de la GBV.

La Organización de las Naciones Unidas (UN) proclamó en 1993 la "Declaración sobre la eliminación de la violencia contra la mujer" [34] y, desde entonces, "la violencia contra la mujer y la violencia doméstica se consideran formas de discriminación, asuntos de derecho penal y violaciones de los derechos humanos". La Agenda 2030 para el Desarrollo Sostenible [35] fue adoptada por los Estados miembros de la UN en 2015 y proporciona un "plan compartido de prosperidad y paz para el planeta y las personas, en el presente y en el futuro"[1] . En él proclaman los 17 Objetivos de Desarrollo Sostenible (SDG), entre los que reconocen "el objetivo de lograr la igualdad de género y empoderar a todas las mujeres y niñas" (SDG 5). Los SDG presentan un llamamiento urgente a la acción en una cooperación global de todos los países.

Además, la lucha contra la violencia de género forma parte de las actividades de la Comisión Europea para proteger los principales valores de la EU y garantizar el mantenimiento de la Carta de los Derechos Fundamentales de la EU. En marzo de 2022, la Comisión Europea propuso nuevas normas aplicables a toda la EU para combatir la violencia de género y la violencia doméstica [36], que incluyen la tipificación como delito de la mutilación genital femenina, la

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

violación basada en la falta de consentimiento y la ciberviolencia, así como reforzar el acceso de las víctimas a la justicia, entre otros.

Aunque es imposible que el dolor y el sufrimiento humanos, incluso una vida humana, tengan un "precio" [33], ser conscientes del coste de la violencia puede orientar a los países a dirigir el dinero hacia donde sea realmente necesario; hacia donde sea más eficaz y rentable para salvar vidas, lo que constituye tanto un imperativo moral como un uso inteligente de la economía.

1.1.2. Erradicación de la violencia de género

La causa subyacente de la violencia de género está relacionada con la desigualdad estructural de género, basada en el sistema patriarcal de la sociedad y en el desequilibrio de poder entre hombres y mujeres. El concepto de *interseccionalidad* reconoce que las desigualdades sistémicas están conformadas por la superposición de diferentes factores sociales, como la orientación sexual, la identidad de género, la etnia, la raza, la discapacidad y la clase económica, entre otros factores de discriminación. Todos ellos se entrecruzan para crear dinámicas y efectos particulares de discriminación cruzada. Todas las formas de desigualdad se refuerzan mutuamente y deben analizarse y abordarse al mismo tiempo para evitar que las desigualdades se refuercen unas a otras [37]. Por lo tanto, en particular hay ciertos grupos de mujeres que son más vulnerables a la violencia, que tienen más dificultades para resistir cuando se enfrentan a este fenómeno amenazador, porque además de sufrir discriminación de género, sufren de otros tipos de discriminación. Entre estos grupos se encuentran las mujeres jóvenes y mayores, las niñas, las discapacitadas, las de raza, etnia, migrantes o indígenas, las percibidas como LGBTIQ+ (definidas como "personas que se han identificado como lesbianas, gays, bisexuales, transexuales, intersexuales o *queers*"), así como las que abusan de sustancias y las que tienen dificultades familiares o económicas. Todas ellas sufren de mayor riesgo de sufrir violencia de género.

Esto lleva a la conclusión de que el perfil de una GBVV no es homogéneo, y para garantizar una ayuda y un apoyo eficaces a las víctimas, se necesitan diversas estrategias de intervención y programas de prevención, educación y terapia. Las víctimas de la violencia de género suelen tener necesidades de protección individuales/especiales (algunas son especialmente vulnerables a la re-victimización) basadas sobre todo en aspectos psicosociales y culturales que deben tenerse en cuenta a la hora de orientar las soluciones a este problema predominante.

Las soluciones a la violencia de género pueden implicar la correcta asignación de recursos económicos para medios legales y sociales (por ejemplo, para la protección de las víctimas y sus hijos), para la educación y, además, para la inversión en investigación tecnológica.

A nivel de la población en general, es necesario extender la prevención desde una perspectiva integral como medidas de sensibilización, basadas en el respeto a los derechos humanos,

enseñando el rechazo a todo tipo de violencia e incluyendo acciones específicas contra la violencia de género.

En el ámbito educativo es fundamental ir más allá de la elaboración de materiales y programas puntuales para que prevalezcan las medidas educativas. A través de experiencias de colaboración entre chicas y chicos en el aula, basadas en el respeto mutuo, se podría avanzar mucho en la superación de dos de las principales condiciones que subyacen a la violencia de género: la resistencia al cambio que esta situación produce y la desigual distribución del poder en la sociedad [38]. Además, instituciones como los colegios y los institutos, deberían desarrollar protocolos sobre cómo actuar en caso de que se tenga conocimiento de violencia entre los alumnos o sus familias, mediante la intervención educativa. Porque no basta con informar, sino que es necesario construir la igualdad desde la práctica a partir de la concienciación y educación de las nuevas generaciones.

En el ámbito legislativo, existen normas estatales en España [39], Europa [40] e Internacionales [41] que se ocupan de regular aspectos como el judicial, o el laboral en materia de violencia de género. En España, la Ley 1/2004 de Medidas de Protección Integral contra la Violencia de Género [42] establece "los mecanismos judiciales imprescindibles para evitar una doble victimización de las mujeres, suponiendo la unificación de un marco de asistencia y protección para todas las mujeres, cualquiera que sea su situación personal". Esta ley también establece "medidas de protección integral, cuya finalidad es prevenir, erradicar y sancionar la violencia de género y prestar asistencia a las mujeres y menores bajo su tutela, víctimas también de esta violencia".

En la Unión Europea, todos los Estados miembros se han comprometido con los "principales mecanismos de derechos humanos", que les obligan a "combatir la violencia contra las mujeres por considerarla una violación de los derechos humanos, y una forma específica de violencia de género vinculada a la discriminación contra las mujeres". De este modo, los Estados miembros están obligados a poner fin a la impunidad y a prohibir todo tipo de violencia, a proporcionar una protección adecuada a las supervivientes, a tomar medidas para prevenirla y a garantizar la ayuda [40]. Ejemplos del interés por la prevención de la violencia y la igualdad de género son la creación en 2006 del EIGE[2] -ya mencionado en el apartado 1.1.1- que se encarga de la recopilación, el análisis y la difusión de información sobre igualdad y violencia de género; así como el establecimiento del Convenio de Estambul [43] por parte del Consejo de Europa en 2011 sobre la lucha y la prevención de la violencia doméstica y la violencia de género.

Desde el ámbito tecnológico y de la investigación, la lucha contra la violencia en la Unión Europea ha hecho que la EU financie proyectos de investigación e innovación para combatir la

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

Violencia de Género desde hace más de dos décadas[3]. Y los resultados de sus investigaciones se han traducido en recomendaciones adaptadas a los diferentes sectores implicados en la protección de las víctimas, es decir, la policía, la sanidad y los sectores sociales, que necesitan de cooperación entre varios organismos.

1.1.3. Soluciones Tecnológicas para Combatir la Violencia de Género

La tecnología ha logrado muchos avances médicos y científicos, pero también ha dado lugar a nuevos tipos de discriminación en línea (*online*), discursos de odio y violaciones virtuales de los derechos humanos, incluida la violencia de género en línea o ciberviolencia.

La tecnología puede ser una herramienta de empoderamiento y seguridad para las mujeres, para convertirlas en participantes activas que escapan de las relaciones violentas. Varios grupos destinatarios pueden beneficiarse del uso de la tecnología: no sólo las víctimas que se recuperan, sino también los agentes sociales implicados en la prevención de la violencia de género y la protección contra ella, como los terapeutas, las fuerzas de seguridad y los servicios sanitarios. Las soluciones tecnológicas mejoran la eficacia de los profesionales que intervienen contra la violencia de género. Ello redunda en una mejor calidad de los servicios ofrecidos a los ciudadanos y en una mayor seguridad.

Todas las mujeres corren el riesgo de sufrir violencia de género, por eso las estrategias eficaces para prevenir y reducir la violencia deben apuntar a sus raíces. "El desarrollo de una tecnología segura para abordar la violencia de género requiere el liderazgo de las mujeres y las niñas y la colaboración con ellas", afirma UNICEF [44]. Dado que las causas de la violencia de género difieren significativamente de otros tipos de violencia, debemos sistematizar los conocimientos existentes y utilizarlos como norma para adaptar las herramientas tecnológicas de forma que respondan a ellos. Los derechos, las necesidades y los requisitos de las supervivientes de esta violencia en particular, son clave para su diseño. Porque UNICEF también afirma que "la tecnología y sus herramientas no deben exponer a las niñas y a las mujeres a más daños; las soluciones tienen que construirse con una base amplia y sólida de protocolos éticos y normas de la comunidad de la violencia de género, garantizando al mismo tiempo la seguridad digital y las normas de privacidad -por ejemplo, el anonimato y la protección de datos-, con el fin de evitar la discriminación y la victimización".

En los últimos años, el crecimiento de la tecnología digital ha favorecido el desarrollo de novedosas aplicaciones web y para *smartphones* (teléfonos inteligentes) destinadas a luchar contra la violencia de género. Junto con la llegada del Internet de las cosas (IoT), estas tecnologías han desencadenado el desarrollo de varias soluciones que van desde los organismos encargados

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

de hacer cumplir la ley (LEA) hasta cartografiar la exposición a la violencia sexual en distintos lugares [45].

Las aplicaciones basadas en características de geolocalización pueden aumentar la concienciación y reducir el riesgo de violencia de una usuaria, apoyando la prevención [44]. Por ejemplo, *Ec Shlirë* (Caminar libremente)[4] , es una aplicación desarrollada por *Girls Coding Kosova*, que permite a las usuarias denunciar casos de acoso sexual de forma discreta, que se comparten con las autoridades. Una aplicación similar para *smartphones*, *Safetipin*[5] , recopila y mapea en tiempo real los datos de las usuarias -principalmente mujeres y niñas- para proporcionar puntuaciones de seguridad según su ubicación con el fin de mejorar la seguridad pública.

Otro uso innovador de la tecnología para facilitar el acceso a la información y los servicios sin necesidad de acudir in situ, de forma segura, culturalmente adecuada y con una alta accesibilidad para las usuarias, es el uso de *chatbots* interactivos o *apps* de difusión. Algunos ejemplos son el Proyecto Caretas[6] en Brasil, Maru[7] de la ONG Plan Internacional, *Virtual safe spaces* (VSS)[8] y *Springster*[9] de UNICEF. Estas aplicaciones proporcionan recursos y consejos reales de activistas y expertas, suministrando información sobre autocuidado y empoderamiento, violencia de género y salud reproductiva y sexual para mujeres y niñas. Sin embargo, en el caso de los *chatbots*, algunos se basan en la IA y sólo unos pocos incluyen la interacción humana detrás. Estudios recientes se preocupan por el potencial y la eficacia de estos *chatbots* a la hora de proporcionar un apoyo emocional en línea eficaz a las personas, y concluyen que "las usuarias parecen considerar más fiable el apoyo generado por personas que el automatizado por máquinas" [46]. Por eso, la inclusión de *human-on-the-loop* (intervención de personas dentro del circuito de las máquinas, como retroalimentación) [47] es crucial en este tipo de aplicaciones.

En el caso de las aplicaciones para la violencia de género, la tecnología también ofrece una mejora en la prestación de servicios contra la violencia de género y en la calidad de las reacciones. Primero/GBVIMS+[10] es una solución tecnológica de código abierto para la gestión de casos de violencia de género. El sistema mejora la calidad de la atención a las supervivientes y la colaboración a distancia entre los supervisores y los trabajadores que se ocupan de esos casos.

Otra aplicación, ROSA[11] proporciona educación esencial e intercambio de conocimientos para que el personal de apoyo a las personas que sufren violencia de género. Medicapt[12] recoge

[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/

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

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

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

pruebas forenses -que son admisibles ante los tribunales- de las supervivientes de violencia sexual, y puede transmitir estos datos de forma segura a la policía, los jueces y los abogados. Y, VictimsVoice[13] , es una aplicación que permite a las supervivientes de la violencia de género anotar incidencias de abusos de forma legalmente admisible, segura y protegida [44].

Aún así, estas soluciones carecen de características importantes en lo que se refiere a la protección contra la violencia de género en tiempo real, garantizando la seguridad de las mujeres frente a la violencia de género física, es decir, los ataques y las agresiones. Por ello, otras soluciones tienen como objetivo atajar este tipo de situaciones, como SAFER PRO[14] , que es un dispositivo *wearable* desarrollado por una empresa de Nueva Delhi, India, que contiene un chip integrado en una pulsera que envía alertas a los contactos de emergencia de la usuaria, cuando el dispositivo es activado por ella, informando de una situación de emergencia. Además, India emitió una directiva relacionada con la inclusión obligatoria de un botón de pánico en todos los teléfonos móviles que se vendieran a partir de 2017. Sin embargo, los botones de pánico presentan importantes limitaciones en cuanto a la seguridad de las mujeres, ya que exige que ellas mismas tomen un papel activo en su autoprotección -a veces imposible en algunos tipos de agresión-, su falta de diseño discreto -que puede llevar a estigmas en la violencia de género-, o lo que es peor, la falta de apoyo infraestructural [48].

Particularizando en las soluciones tecnológicas contra la violencia de género desarrolladas hasta la fecha en territorio nacional, España es pionera en tecnología contra la violencia de género. Algunas soluciones institucionales incluyen herramientas tecnológicas de apoyo y protección a la violencia de género como las siguientes:

- VioGén[15] : Es un protocolo que siguen los agentes de policía que toman declaración a las denunciantes de una víctima de violencia de género. Rellenan un cuestionario específico que da como resultado una calificación de riesgo que, si es alta, se activan medidas de protección policial. Éstas pueden ir desde la realización de llamadas de seguimiento hasta la colocación de un coche patrulla las 24 horas del día en la puerta del domicilio de la víctima.
- ATENPRO[16] : Usa un teléfono móvil y dispositivo de telecomunicaciones que permite a las usuarias ponerse en contacto en cualquier momento con un centro de llamadas atendido por personal específicamente formado para dar una respuesta adecuada a su situación de GBV.

[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

- AlertCops [49]: Es un servicio de ayuda a las personas ciudadanas en situaciones de peligro, con el objetivo de enviar avisos, incluyendo datos geolocalizados, con fotografías o audio, a los agentes de policía para advertirles de un testigo o de la presencia de un delito. AlertCops incorpora un botón SOS para reforzar la protección de las víctimas de violencia de género y del personal sanitario[17] .
- Más específicamente para la protección de la violencia de género, el Centro COMETA[18] ofrece los servicios de vigilancia, funcionamiento e instalación de dispositivos que vigilan tanto a la víctima como al agresor en casos de violencia de género. El sistema está diseñado para activar una alarma en los casos en que la GBVV esté en peligro, como en caso de que el agresor se acerque demasiado a ella, o en caso de manipulación de la correa o rotura de la pulsera, entre otros.

Un estudio de investigación realiza un análisis exhaustivo de las soluciones tecnológicas hasta la fecha de su publicación [50] y afirma que "el objetivo debe ser lograr una solución holística, ya que la integración adecuada de diversos enfoques podría conducir a una propuesta multi-estratégica que podría mejorar la seguridad de las mujeres y contribuir al fin de este tipo de violencia".

A pesar de los esfuerzos tecnológicos, las soluciones existentes hasta la fecha presentan diferentes lagunas de investigación cuestionadas por varios expertos en violencia de género que reclaman investigaciones más avanzadas [51] y tecnologicas para algunas soluciones consideradas obsoletas. Y a pesar de los impresionantes avances de la Inteligencia Artificial (AI), no existen soluciones que incorporen inteligencia para la detección automática de una situación de riesgo que pueda poner en peligro la vida de las mujeres. Ya hemos mencionado que los botones de pánico, o los centros de ayuda telemáticos, son soluciones que implican el compromiso de la víctima en su propia seguridad. Y en los casos en los que las mujeres son atacadas por un delincuente o agresor, es posible que no dispongan de los recursos necesarios para llevar a cabo estas acciones.

Están surgiendo nuevos paradigmas en los que se proponen nuevas herramientas de AI, pero no como sustitutas de las ya existentes, sino como complementarias, que incluyen ventajas frente a las tradicionales. El análisis predictivo impulsado por la AI es capaz de recopilar datos de las usuarias, analizarlos y extraer de ellos valiosas conclusiones. Las predicciones relativas a la seguridad de las mujeres pueden estimarse con precisión con un algoritmo de AI correctamente entrenado, analizando los datos (es decir: estado de las usuarias, contexto) con los que ha sido entrenado. La IA proporciona funciones como el estudio analítico del estado y el contexto de la usuaria, predicciones y decisiones personalizadas, resultados precisos basados en datos en tiempo

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

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

real y la eliminación de información irrelevante o redundante, proporcionando una solución integral que garantizaría la seguridad de las mujeres.

Una solución impulsada por la AI que realizara un análisis exhaustivo de aspectos como el estado emocional de la persona, además de un análisis del contexto o de la situación externa (por ejemplo, las circunstancias, la ubicación) y, por lo tanto, detectara automáticamente si la integridad de una mujer está en peligro, evitaría que la persona tuviera que implicarse activamente en su autoprotección -algo ciertamente imposible en algunos tipos de agresión- y podría proporcionar una respuesta automática y más rápida para garantizar su seguridad.

1.1.4. Una solución tecnológica puntera de Inteligencia Artificial para combatir la Violencia de Género: Bindi

En respuesta a los requerimientos comentados y tras un estudio socio-psicológico de las ventajas e inconvenientes de la tecnología empleada actualmente en España, nació en 2016 el equipo multidisciplinar UC3M4Safety -al que pertenecen tanto la autora como la directora de esta tesis- para desarrollar una innovadora solución de inteligencia artificial denominada Bindi.

Este proyecto de esta tesis se enmarcó en EMPATIA-CM (*ProtEcción integral de las víctimas de violencia de género Mediante comPutación AfecTIva multimodAl*), un proyecto de la *Convocatoria 2018 de Proyectos Sinérgicos de I+D en Nuevas y Emergentes Áreas Científicas Comunidad de Madrid* que fue concedido al equipo multidisciplinar UC3M4Safety desde Abril de 2019 hasta Junio de 2022. Actualmente, al grupo se le ha concedido un segundo proyecto de continuación hasta Junio de 2023 denominado S4B (*SistemA ciberfísico Para el seguimIento y prevencIón de cAsos de violencia de género*: SAPIENTAE4Bindi) para incrementar el nivel de madurez tecnológica (TRL) de los resultados de investigación de EMPATIA-CM y esta tesis se enmarca en ambos proyectos, EMPATIA-CM y S4B.

El equipo UC3M4Safety surgió de la necesidad de aunar esfuerzos en la lucha contra la Violencia de Género (GBV) desde múltiples disciplinas. El proyecto cuenta con 42 investigadores del Instituto de Estudios de Género (IEG) y de los departamentos de Tecnología Electrónica, Telemática y Teoría de la Señal y Comunicaciones de la Universidad Carlos III de Madrid (UC3M). Además, el grupo colabora estrechamente con el Centro de Electrónica Industrial (CEI) de la Universidad Politécnica de Madrid (UPM). Y esta multidisciplinariedad del equipo se refiere a la implicación de diversas investigadoras e investigadores que aportan conocimientos de varias disciplinas, cada persona contribuyendo al proyecto desde su propia área.

El equipo UC3M4Safety comenzó su andadura en el concurso Anu and Naveen Jain Women's Safety XPrize[19] en el que quedó semifinalista. Ha registrado una solicitud de patente [52] y ha obtenido varias subvenciones y premios.

El objetivo principal que tenía EMPATIA-CM y que ahora continúa en S4B es mejorar la protección que la sociedad ofrece a las mujeres en situaciones de agresión por violencia de género, generando un protocolo fiable y robusto para detectar, prevenir y resolver estos delitos. La misión de las tecnologías innovadoras propuestas en los proyectos realizados es ayudar a prevenir la GBV mediante: 1) la detección precoz de situaciones de riesgo, 2) la interconexión de víctimas potenciales y agentes protectores y 3) la recogida segura y precisa de pruebas del presunto delito. Todos estos medios nos ayudarán a estudiar el problema de la violencia de género de forma integral y multidisciplinar. Para ello, el equipo propone el uso de sistemas ciberfísicos con Computación Afectiva (AC). En concreto, dispositivos *wearable (*vestibles, que se pueden llevar puestos) con sensores inteligentes que monitoricen a una usuaria en tiempo real, detecten las circunstancias en las que se encuentra y las emociones que experimenta en situaciones de riesgo y conecten con agentes protectores, gubernamentales y/o no gubernamentales, alertando de la situación en tiempo real. El impacto esperado es la mejora notable de la vulnerabilidad de las mujeres al proporcionarles herramientas que mejoren su seguridad y favorezcan así su desarrollo personal y profesional.

Este es el entorno multidisciplinar de compromiso social e investigación puntera en el que se desarrolla este proyecto de tesis. Los recursos humanos y materiales dedicados al proyecto son óptimos gracias a la adecuación de los perfiles de las investigadoras e investigadores implicados, la infraestructura tecnológica disponible y la financiación obtenida en los proyectos EMPATIA-CM y S4B.

Figura 1. 2: Esquema de la operación de Bindi [53]. Reproducido con permiso del propietario del copyright, el equipo UC3M4Safety.

[19] https://portal.uc3m.es/portal/page/portal/inst_estudios_genero/proyectos/Women_Saf ety_XPrize_2018

En cuanto al funcionamiento técnico de Bindi, este sistema se concibe como una solución integral, inteligente, discreta, conectada, de *edge-computing* (de computación periférica) y *wearable* dirigida a la detección automática de situaciones de violencia de género. Como puede observarse en la Fig. 1.2, la GBVV lleva dos dispositivos inteligentes ocultos en joyas o abalorios que monitorizan las variables fisiológicas y el entorno acústico, incluida la voz. Estos están conectados a un teléfono inteligente con una aplicación con un núcleo impulsado por AI que puede producir diferentes tipos de alertas, así como encriptar y enviar la información a un servidor seguro.

Bindi es una tecnología de vanguardia que combina la computación afectiva inteligente y el IoT con la adquisición y fusión de señales físicas y fisiológicas multisensoriales y una infraestructura de servidor segura para detectar de forma autónoma situaciones de riesgo, marcando alarmas y registrando datos para posteriores acciones legales. Más concretamente, Bindi captura metadatos y datos relativos a la usuaria y su contexto (por ejemplo, rutinas personalizadas, geolocalización, variables fisiológicas, habla, eventos acústicos, ...) y determina el estado afectivo de la usuaria teniendo en cuenta las circunstancias en las que se encuentra. Con estos datos, Bindi utiliza su núcleo de AI para evaluar cada situación, y tiene la capacidad de detectar automáticamente cuándo una situación puede poner en peligro la vida de la usuaria, activando sus alarmas, alertando a los servicios de emergencia. En la Fig. 1.3 representamos la evolución de los dispositivos *wearable* desde la versión 1.0 de Bindi hasta la 2.0.

1.2. Contexto: Desafíos Técnicos

Tras haber discutido la motivación de este trabajo, consideramos esencial explicar el contexto en el que se enmarca. En este subapartado hablaremos de los retos técnicos que plantea este trabajo.

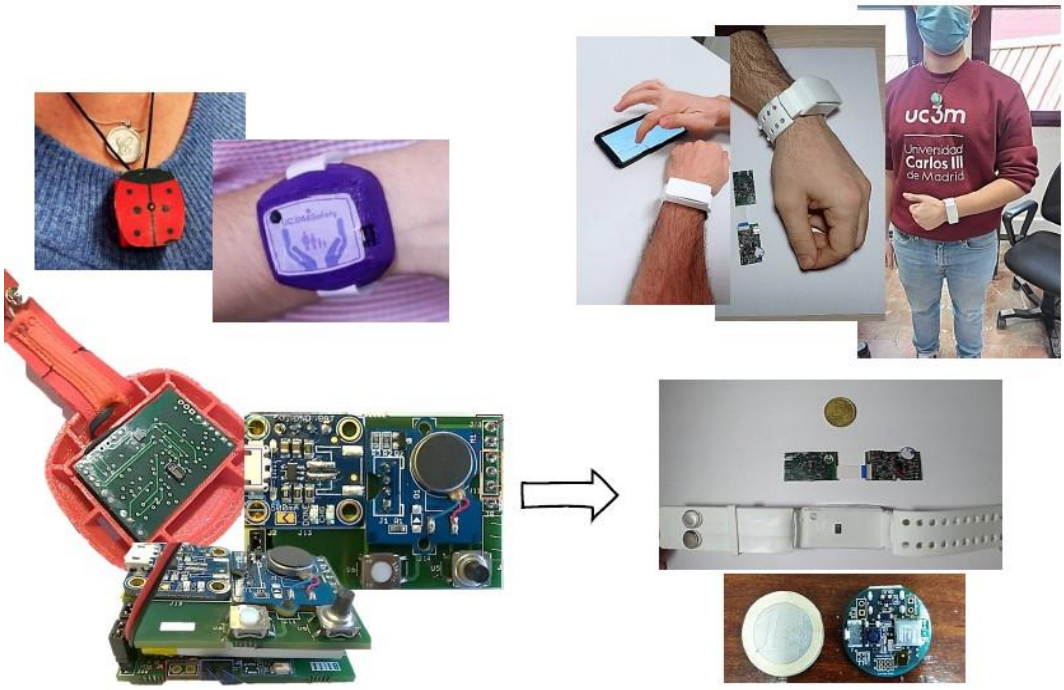


Figura 1. 3: Evolución de los dispositivos portátiles Bindi [53]. Reproducido con permiso del propietario del copyright, el equipo UC3M4Safety.

La tesis se centra en el análisis del estado afectivo de una persona mediante diferentes modalidades de entrada o datos, pero más concretamente del auditivo, y en particular del habla. Así, revisamos algunos de los retos relativos al problema del sesgo en los algoritmos de AI, concretamente en el aspecto del género. Además, esta solución tecnológica está pensada para ser integrada en dispositivos *wearable* y *smartphones*, por lo que exponemos la preocupación sobre los retos computacionales de dichos dispositivos. Por último, nuestro objetivo final con ella es detectar cuándo la vida de una mujer está en peligro debido a una situación de violencia de género, por lo que presentamos brevemente el gran reto que supone poder ofrecer soluciones a todas las mujeres/personas.

1.2.1. Investigación, Datos y Sesgos

Las jerarquías sociales entrecruzadas y superpuestas que se encuentran en el poder, la religión, la raza, la etnia, el género, la edad, la orientación sexual o la clase, dan lugar a la distribución desigual de acceso a los recursos y a los derechos, lo que constituye una desigualdad social. Y desde una perspectiva histórica, los avances en investigación -y más concretamente, en ciencia y tecnología- se han orientado, focalizado y dirigido hacia un perfil concreto de personas debido a la desigualdad social imperante e histórica.

Se ha demostrado que los grupos sociales hacia los que la investigación, la ciencia y la tecnología no se han orientado históricamente tienen dificultades para conseguir que estos avances se apliquen y funcionen para ellas. Un claro ejemplo de tales grupos sociales son las mujeres.

En el ámbito médico, hasta el año 2000 se produjo casi una duplicación de la tasa de mortalidad en las mujeres con respecto a los hombres debido a enfermedades cardiacas [54]. Existían claras lagunas en la investigación sobre el infarto de miocardio, porque se había realizado teniendo en cuenta mayoritariamente a pacientes masculinos. Esto provocó que las mujeres recibieran un diagnóstico y un tratamiento menos adecuados que los hombres, como consecuencia de la educación y la información recibidas por los médicos. Hasta esa fecha, se pensaba que las mujeres experimentaban una mayor variedad de síntomas, denominándose "síntomas atípicos", porque no se correspondían con los experimentados por los hombres y, por tanto, las enfermedades miocárdicas rara vez se identificaban adecuadamente en las mujeres.

En el campo de la tecnología y la AI, existen innumerables ejemplos de este tipo de discriminación, algunos de los más recientes se describen a continuación. Una conocida marca de coches se vio obligada a solicitar la devolución del mercado de uno de sus modelos de coche porque los conductores masculinos de Alemania no se fiaban de la voz femenina que daba las indicaciones en el sistema de navegación del coche [55]. Otro estudio [56] que evaluó la precisión de los subtítulos generados automáticamente por YouTube en los dos géneros estándar mostró grandes diferencias de precisión en ambos géneros, con una precisión significativamente menor en las voces femeninas, lo que pone de manifiesto la necesidad de una validación de los sistemas que esté estratificada sociolingüísticamente. Un estudio adicional [57] evaluó 3 sistemas comerciales de grandes compañías de clasificación facial mostrando discrepancias considerables en "la precisión de la clasificación de las mujeres de piel oscura, las mujeres de piel clara, los hombres de piel oscura y los hombres de piel clara en los sistemas de clasificación de género".

Las mujeres de piel oscura fueron el grupo peor clasificado - "tasas de error de hasta del 34,7%" [58]-, mientras que "la tasa de error máxima para los varones de piel más clara fue del 0,8%".

Podemos encontrar otro ejemplo de la brecha de investigación entre grupos sociales en [59], donde "los dispositivos *wearable* son menos precisos para detectar la frecuencia cardiaca en individuos con tonos de piel más oscuros". Los investigadores concluyeron que "los *wearable* son menos precisos para la detección de la frecuencia cardiaca en dichos participantes, siendo una posible causa la falta de investigación en sujetos de piel más oscura en el momento de desarrollarse dichos dispositivos".

También nos gustaría destacar que, en el campo del reconocimiento del habla, se sabe desde hace tiempo que los modelos orientados al reconocimiento de la voz de los hombres, por un lado, y de las mujeres, por otro, funcionan mejor si se diseñan por separado, debido a las características del habla conocidas y diferentes que tiene cada sexo [60].

Estos son sólo algunos de los muchos ejemplos que evidencian que la AI está generalmente sesgada hacia grupos sociales específicos, discriminando a otros, como es el caso de las mujeres [61], donde históricamente la AI ha carecido claramente de perspectiva de género.

Esto señala que se necesita atención específica en investigación para construir algoritmos de inteligencia artificial absolutamente justos, transparentes y responsables. El papel de la AI en la consecución de los Objetivos de Desarrollo Sostenible de las Naciones Unidas es controvertido: puede permitir la consecución de 134 metas en todos los objetivos, pero también inhibir 59 [62]. Junto con la enorme explosión de avances en AI de las últimas décadas, también ha surgido un encendido activismo por los derechos sociales, ya que los sistemas algorítmicos han sido criticados por perpetuar los prejuicios, la discriminación injusta y contribuir a la desigualdad [63]. A pesar de los avances que han descubierto sesgos en subcampos de la AI como el Procesamiento del Lenguaje Natural (NLP) [64], la Visión por Ordenador [65] y la Medicina y la Sanidad [66], desde el punto de vista de la Computación Afectiva -cuando la tarea requiere reconocer o simular afectos humanos-, hay pocas evidencias en la literatura de los sesgos de los algoritmos de IA (por ejemplo, utilizando rasgos psicobiológicos [67]) cuando los datos utilizados para la tarea son señales del habla. No obstante, esto no significa que no existan sesgos en estos casos, sino que se trata de un campo tan reciente que aún no se ha revisado sistemáticamente la literatura sobre el tema.

Así pues, extrapolando el análisis de otros campos de investigación, ofrecemos un breve repaso de algunos de los posibles sesgos en la Computación Afectiva, y más concretamente en la tarea del Reconocimiento de Emociones del Habla (SER), que dan lugar a posibles soluciones desbalanceadas en función del género, así como posibles formas de mitigarlos:

- Sesgo de muestreo en la creación de conjuntos de datos: Mayor cantidad de datos recogidos de sujetos u oradores masculinos. Existe una tendencia activa, en las bases de datos de reconocimiento del habla de las emociones, a proporcionar la misma cantidad de datos de hablantes masculinos y femeninos, pero en la gran mayoría de las bases de datos que presentan un desequilibrio, éste se debe a un mayor número de datos de hablantes masculinos sobre los femeninos [68].
- Sesgo del conjunto negativo (o falacia del "mundo cerrado"): Se refiere a no disponer de suficientes muestras para hacer una representación fiable de todo el mundo en el conjunto de datos disponible. Los conjuntos de datos pueden estar desequilibrados si queremos representar todas las categorías o tipos de datos que existen en el mundo y el conjunto de datos no los contiene. Se trata de un sesgo muy común y difícil de abordar.
- Sesgo de etiquetado: Las etiquetas que acompañan a los datos suelen ser generadas por humanos (por ejemplo, anotadores expertos, *crowdsourcing*,...) y este sesgo se refiere al hecho de que varios anotadores pueden anotar los mismos datos de forma diferente. Cada anotador se basa en sus antecedentes para etiquetar, según su formación, origen, contexto o trayectoria. Por lo tanto, la anotación es una consecuencia de las percepciones previas y la experiencia de los anotadores. Por ejemplo, en el caso de la anotación de emociones,

cuyo carácter es intrínsecamente subjetivo, la etiqueta puede variar en función de las características del anotador, como su cultura, su inteligencia emocional, etc.

- Sesgo de evaluación humana: Consiste en extraer conclusiones basadas en la experiencia o trayectoria de la persona o personas que realizan la investigación, que tendrán sesgos de percepción y análisis en función de su contexto.

La solución a todos estos sesgos pasa por incluir un número equilibrado, rico y diverso de: datos, anotadores para dichos datos y evaluadores; con el fin de captar fielmente la realidad, caracterizar el contenido y los datos de forma fiable y extraer conclusiones adecuadas teniendo en cuenta dichos sesgos.

Es importante vigilar los sesgos en los datos y los algoritmos de AI y mitigarlos, porque la IA puede exacerbar y reforzar los sesgos sociales, con todas las consecuencias negativas que ello conlleva. Los datos captan la realidad y la realidad está sesgada, si esos sesgos alimentan los algoritmos de AI, acabarán por reforzar y perpetuar la discriminación. Las y los investigadores deben tener en cuenta los sesgos y trabajar para ser justos y lograr la igualdad en la sociedad con sus sistemas, pero todas nosotras debemos ir más allá y atajar el problema de la discriminación de raíz, asumiendo la responsabilidad social contra la desigualdad estructural.

1.2.2. Hardware: Complejidad Computaciónal y Batería

El éxito de los algoritmos de AI depende de muchos factores, como la cantidad o la calidad de los datos utilizados para el entrenamiento, la complejidad de los modelos y la precisión de las etiquetas, entre otros. Los modelos computacionales complejos y profundos han demostrado tener éxito debido a su buena capacidad de generalización, por lo que son más adecuados cuando la fase de entrenamiento contiene muchos datos, en comparación con el uso de conjuntos de datos más pequeños y algoritmos de aprendizaje automático poco profundos [69]. Así, uno de los principales inconvenientes de los modelos de AI de estos últimos años es esta tendencia a crecer hacia arquitecturas y esquemas más grandes [70], con el fin de lograr mejores rendimientos. Pero más grande no siempre es mejor para el aprendizaje automático. Por muy rompedores que sean, las consecuencias de los modelos más grandes son graves tanto en términos de presupuesto -necesidad de más potencia de cálculo y suministro de energía- como para el medio ambiente -a mayor consumo de energía, peor para el medio ambiente en términos de contaminación- [71].

En Bindi, la arquitectura IoT diseñada considera una división en tres capas, es decir, computación de *edge*, de *fog* y en *cloud* (en la "nube") [1]. En el sistema, la capa de computación de *edge* se concibe como una red ciberfísica inteligente compuesta por dos dispositivos (un colgante y una pulsera). La capa de informática de *fog* en Bindi se concibe como una aplicación para *smartphone*. Por último, la información relevante obtenida a través de las capas de *edge* y *fog* se almacena de

forma segura en servicios informáticos específicos en el *cloud*. Y la gestión del consumo de energía es un requisito para el diseño de estos sistemas *wearable*. Si queremos emplear algoritmos de AI en tales dispositivos, es esencial realizar un análisis preciso del consumo de la batería y de la autonomía de los dispositivos de hardware implicados en cada capa para garantizar que el sistema funcione cuándo y cómo sea necesario.

Como parte del trabajo del equipo UC3M4Safety, se estudia el análisis cuantitativo del consumo en los dispositivos *wearable* [72], midiendo las acciones que más energía demandan en la parte de monitorización [1]. El equipo midió el consumo de energía empleado en la comunicación y adquisición de datos de los sensores, ya que son esenciales para el sistema y están intrínsecamente relacionadas con el diseño de hardware específico de los dispositivos [1].

1.2.3. ¿Una solución para todas las mujeres? El reto de la generalización de la Inteligencia Artificial en la Violencia de Género

A pesar de las expectativas actuales sobre el papel de la Inteligencia Artificial en nuestra sociedad y de los impresionantes avances que hemos presenciado en la última década, la solidez de la AI es motivo de gran preocupación. De hecho, la Unión Europea ha desarrollado la Ley de Inteligencia Artificial [73], que proporciona un conjunto de directrices para proteger a las ciudadanas y ciudadanos europeos de posibles usos indebidos y errores, haciendo hincapié en la fiabilidad y en evitar todo tipo de sesgos en relación con las características demográficas de las personas.

Existen diferentes enfoques para mejorar la robustez de la AI en la literatura, pero la mayoría de ellos coinciden en el diagnóstico de la raíz del problema: el desajuste entre los modelos matemáticos obtenidos del entrenamiento con datos de laboratorio (captados en condiciones controladas) y los utilizados después en el mundo real, donde las condiciones son incontroladas. Cuanto más compleja es la realidad, más datos se necesitan para obtener correctamente modelos precisos. Todos los datos recopilados deben captar la diversidad y la complejidad del fenómeno que se quiere modelizar.

Somos conscientes, como ya hemos dicho, de que la vulnerabilidad individual a la violencia de género está relacionada con aspectos psico-socio-culturales. Por lo tanto, es un reto enorme desarrollar una única herramienta de apoyo para todas las mujeres, y más concretamente para la violencia de género. Estas soluciones pueden no ser aceptables, ni apropiadas, y/o no estar disponibles, o incluso ser peligrosas en algunas circunstancias. Esta es la razón por la que consideraremos una dimensión más amplia. Porque esta solución tecnológica es muy importante para salvar vidas y garantizar la seguridad de las mujeres. Somos conscientes de que deberíamos ser capaces de proporcionar soluciones a todas las mujeres/personas, pero sigue siendo un reto grande y complejo.

Uno de los inconvenientes de este tipo de soluciones tecnológicas es su limitada capacidad de generalización. Esto significa que actualmente las tecnologías no son capaces de adaptarse automáticamente a la diversidad de la violencia de género y a sus situaciones cambiantes, por ejemplo, en las rutinas diarias, los hábitos culturales, la diversidad de situaciones familiares, etc. Esto podría repercutir gravemente en el rendimiento y elevar significativamente la tasa de falsas alarmas activadas, lo que haría totalmente inasequible el mantenimiento de los recursos necesarios para atenderlas.

Como hemos mencionado, la complejidad del problema de la violencia de género es dinámica y difícil de medir, ya que las medidas adoptadas para combatirla, junto con la sensibilización de la opinión pública y los esfuerzos educativos, que aplican los diferentes países, modifican su aspecto y prevalencia. Desde el punto de vista de su modelización automática mediante el uso de herramientas tecnológicas y diferentes dispositivos sensoriales, está claro que necesitamos dividir el problema, primero comprendiendo las diferentes realidades socioculturales, y segundo, la situación psicológica de las víctimas, mediante el conocimiento experto y las metodologías cuantitativas de las ciencias sociales.

El equilibrio observado entre los comportamientos colectivos e individuales debe traducirse en una metodología para recopilar y modelar datos, y para articular las relaciones entre los submodelos matemáticos obtenidos por la AI. Pero no sólo el círculo se cierra cuando las tecnologías obtenidas con la ayuda de las ciencias sociales se utilizan para mejorar la comprensión socio-psicológica de la violencia de género, sino cuando a partir de la integración y la agregación inteligentes de los datos capturados, mejoran sustancialmente los métodos cuantitativos.

1.3. Objetivos y Relevancia

Una vez expuesta la motivación y el contexto de los retos que hay que afrontar, en esta sección desglosamos los objetivos de esta tesis y su justificación y relevancia.

Creemos que las soluciones al problema de la GBV desde las ciencias sociales podrían venir de la mano de la tecnología y la Inteligencia Artificial. Creemos que la tecnología es un facilitador para erradicar la violencia de género, pero no la única solución en sí misma. Por ello, esta tesis doctoral apunta a la detección de situaciones de riesgo de violencia de género para las mujeres, abordando el problema desde un punto de vista multidisciplinar al aunar las tecnologías de la AI y la perspectiva de género, necesitando de técnicas de varias disciplinas.

Queremos analizar el estado emocional de la usuaria a través de las modalidades auditiva y fisiológica, -siendo auditiva, del habla y acústica-, pero más concretamente, dirigimos el foco a la modalidad auditiva, analizando el habla producida por la usuaria, para la detección de emociones en la voz y las formas en que se puede combinar con la información fisiológica.

Así, pretendemos investigar sobre la Computación Afectiva, particularizando en las tecnologías del habla y sus aplicaciones, desde una perspectiva de género, para dar una solución tecnológica que pueda proteger a las mujeres de las situaciones de riesgo de violencia de género. La Fig. 1.4 presenta esta definición en un mapa conceptual de esta tesis.

Figura 1. 4: Mapa conceptual que enmarca esta tesis doctoral.

Hay varios aspectos clave por los que se seleccionó la voz como variable a registrar con Bindi para la protección de las usuarias. En primer lugar, para captar la voz no es necesario llevar dispositivos pesados o complejos como un casco, una cinta para la cabeza o una banda pectoral. La voz es una de las señales que pueden grabarse de forma discreta, simplemente utilizando un *smartphone* o un micrófono de solapa; por lo tanto, es fácil capturar y utilizar los datos de la voz, y además las usuarias pueden acceder fácilmente a la tecnología que puede grabarla.

En segundo lugar, la voz es un identificador único de una persona. Cualquier otra señal corporal, como la temperatura de la piel, no incluye información relativa a la persona, sin poder establecer un vínculo directo entre la persona y la señal. En cambio, al capturar la voz, captamos mucha información relevante sobre la persona que habla [74], la voz puede utilizarse como identificador personal, por ejemplo, puede utilizarse como prueba judicial si se almacena correctamente y se protege para evitar modificaciones y hackeos. Y en tercer lugar, porque las emociones se reflejan en la voz. Esto lo veremos más adelante en el capítulo 2, especialmente *el miedo*, para el que la ausencia de voz también es importante (en la sección 2.2.1 se profundiza en la ayuda de las variables fisiológicas).

Como se ha indicado anteriormente, esta tesis nace enmarcada dentro del proyecto EMPATIA-CM, más concretamente bajo el Objetivo 2: *Investigar, diseñar y verificar algoritmos para detectar automáticamente situaciones de riesgo en víctimas de Violencia de Género*. Hacia el final

de la tesis, también ha tomado forma desde SAPIENTAE4Bindi, bajo el Objetivo 1: *Alcanzar un nivel de madurez del sistema Bindi equivalente a un TRL (Technology Readiness Level) 7-8 para poder implementarlo con garantías en los dispositivos de protección y atención a las Víctimas de Violencia de Género en las administraciones públicas*, continuando con el análisis de las modalidades de entrada de los dispositivos portátiles Bindi y su procesamiento para la detección automática de situaciones de riesgo.

Para definir el enfoque de esta tesis, describimos a continuación los objetivos generales y específicos:

Objetivo general (OG):

(**OG.1):** Comprender las reacciones de las mujeres -incluidas las víctimas de la violencia de género- ante situaciones de riesgo o peligro, hasta el punto de poder generar mecanismos automáticos de detección de estas situaciones a partir de la modalidad auditiva en particular.

Objetivos específicos (OE):

(OE.1): **Identificar la voz de la hablante** entre toda la información contenida en la señal de audio, haciendo frente a la influencia de las emociones o del ruido ambiente. Esto es necesario en primer lugar para realizar después el reconocimiento de emociones en el habla.

(OE.2): Desarrollar modelos computacionales robustos de aprendizaje automático o profundo (ML o DL) basados en la **señal del habla** para **detectar el *miedo*** o el *pánico* -o en su ausencia, su pariente cercano: *el* estrés- en la voz de la hablante objetivo, reflejando el estado emocional de la usuaria.

(OE.3): Investigar y desarrollar modelos computacionales multimodales robustos con la misma función que OE.2 en los que se combinen diferentes modalidades de forma inteligente según las limitaciones de los dispositivos *wearable* de Bindi. Este proyecto de tesis explorará métodos de fusión de los modelos OE.2 con los de otras modalidades aportados por otros miembros del equipo UC3M4Safety. Este objetivo es interdisciplinar, ya que requiere de **técnicas de fusión de datos**, realizando un trabajo muy minucioso, concienzudo y preciso.

(OE.4): Investigar y desarrollar métodos para adaptar los modelos OE.2 y OE.3 a la usuaria individual, en particular en lo que respecta a la modalidad auditiva. Este paso de **personalización** es crucial para aumentar el rendimiento de los modelos genéricos.

(OE.5): **Despertar el interés de la comunidad investigadora** en el desarrollo de soluciones para la prevención del complejo y difícil problema de **la violencia de género**

1.4. Contribuciones y Estructura de la Tesis

En esta sección presentamos las contribuciones de esta tesis y la estructura que se seguirá a lo largo de los capítulos. Las principales contribuciones científicas de esta tesis son las siguientes:

- Un análisis exhaustivo del omnipresente problema de la violencia de género y el uso de la AI como solución tecnológica, incluidos los retos, las estrategias y soluciones existentes y las limitaciones. En particular, las consideraciones éticas y los retos del uso de la Computación Afectiva desde la perspectiva de género.
- Un estudio de las bases de datos de habla estresada realista existentes en la literatura que serían adecuadas para nuestro objetivo de reconocimiento del *miedo* a través del habla y sus limitaciones, incluido un método robusto y estable para etiquetar emociones cuando son anotadas de manera continua por una cantidad limitada de calificadores expertos.
- Una novedosa base de datos multimodal grabada en condiciones de laboratorio para la elicitación de emociones reales -incluido *el miedo*- mediante realidad virtual denominada WEMAC, capturada junto con otros miembros del equipo UC3M4Safety. Incluye las variables fisiológicas y del habla de las participantes, junto con diferentes anotaciones emocionales.
- Una novedosa base de datos multimodal en condiciones de la vida real capturada con los dispositivos de Bindi llamada WE-LIVE, que incluye variables fisiológicas y del habla, geolocalización y acelerómetros, junto con diferentes anotaciones emocionales, también grabada en estrecha colaboración con más personas del equipo de UC3M4Safety.
- Una estrategia de aumento de datos para hacer frente a los efectos negativos de las condiciones de estrés en el habla para la tarea de reconocimiento de hablantes mediante habla estresada generada sintéticamente. Esta técnica podría extrapolarse al problema de la detección *del miedo a* través del habla cuando los datos son limitados.
- Un modelo robusto de aprendizaje automático para la tarea de reconocimiento del hablante en condiciones de ruido que elimina el ruido del habla en el momento en que identifica al hablante, incluido el habla en condiciones de estrés. Demostramos que el modelo es más robusto y estable que otros métodos para la detección del hablante en condiciones de mucho ruido.
- El diseño de un sistema multimodal en cascada para Bindi 1.0 y su consiguiente evolución a un sistema híbrido asíncrono de fusión de las modalidades fisiológica y del habla para la detección *del miedo*.
- Un diseño de un sistema global de Internet de las Cosas con componentes de computación de *edge*, *fog* y *cloud* para de BINDI 2.0 de nuevo junto al equipo de UC3M4Safety. Específicamente detallando cómo diseñamos las arquitecturas de inteligencia en los dispositivos Bindi para la detección del *miedo* en la usuaria y la validación experimental de dicha metodología de procesado de datos.
- Un diseño de modelos de baja complejidad computacional para la detección del estrés real a través del habla.
- Diseño y validación de un sistema para la detección del *miedo* realista utilizando sistemas de datos monomodales -voz- y multimodales -voz y señales fisiológicas-, emulando el

funcionamiento en tiempo real de los dos dispositivos wearables de Bindi, utilizando diferentes enfoques de fusión de datos y una estrategia de adaptación a la hablante.

- Una metodología y un caso de uso en una investigación preliminar en el campo del *Análisis Acústico Afectivo de Escenas*, para estudiar la relación entre unas escenas acústicas y las emociones que pueden provocar en las personas inmersas en ellas, en colaboración con más personas del equipo de UC3M4Safety.
- Un estudio preliminar sobre la detección de condiciones de violencia de género a través del habla y de claves paralingüísticas, también junto con miembros del equipo de UC3M4Safety.

Como hemos introducido antes, esta tesis es un estudio exhaustivo de cómo podemos utilizar la modalidad auditiva desde la perspectiva de género para la protección de las mujeres contra la violencia de género. A continuación, presentamos brevemente la estructura del resto de la tesis.

En el capítulo 1 describimos qué es la violencia de género y sus consecuencias, sirviendo de motivación y antecedentes para esta tesis.

El capítulo 2 presenta una introducción al afecto, las emociones y cómo surgen y sus efectos en el cuerpo humano. También introduce al tema de la Computación Afectiva, el campo de investigación de la AI que pretende dotar a las máquinas de la capacidad de la inteligencia emocional, incluida la de simular la empatía.

El capítulo 3 describe y justifica el uso de conjuntos de datos que contienen estrés como punto de partida de nuestra investigación. Además, y como consecuencia de la falta de bases de datos realistas de habla con *miedo* en la literatura, describimos una de las principales contribuciones de nuestro equipo UC3M4Safety que es la creación de nuestro propio conjunto de conjuntos de datos para cubrir dicho nicho en la literatura: UC3M4Safety Audiovisual Stimuli Dataset, WEMAC y WE-LIVE.

Los capítulos 4 y 5 están centrados en tareas y son experimentales, cada uno de ellos introduce los temas del reconocimiento de hablantes -estudiando temas como la eliminación de ruido del habla y la identificación de hablantes en condiciones de estrés-, y el reconocimiento de emociones -particularmente centrado en las negativas, por ejemplo, el *estrés*, el *miedo*-, respectivamente, incluyendo los trabajos realizados en esta tesis para cada campo. En el capítulo 5 también se ofrece una visión general y un debate sobre el funcionamiento y la importancia de Bindi.

En el capítulo 6 abordamos otros trabajos de investigación complementarios a esta tesis, como el *Análisis Acústico Afectivo de Escenas* o la detección de la condición de víctima de violencia de género a partir del habla.

Por último, el capítulo 7 presenta las conclusiones de los trabajos de investigación realizados en esta tesis multimodal y multidisciplinar y lo que pretendemos seguir haciendo después de ella como trabajo futuro.

Capítulo 2: Una Perspectiva Multidisciplinar de la Computación Afectiva

El presente capítulo consiste en una definición del campo de la Computación Afectiva. Abarca la base del afecto, el estado de ánimo y las emociones y de dónde surgen, las diferentes teorías de la emoción en las ciencias afectivas, sus aplicaciones actuales y algunas consideraciones éticas importantes.

2.1. Afecto, Emociones y Estado de Ánimo

El campo de investigación de la Computación Afectiva (AC) comprende "el estudio y el desarrollo de sistemas que pueden reconocer, interpretar, procesar y simular los afectos y las emociones humanas" [75]. Se trata de un campo multidisciplinar en el que intervienen los campos de la informática, la psicología y las ciencias cognitivas, centrado en permitir que los robots y los ordenadores respondan de forma inteligente a la retroalimentación emocional humana natural.
Afecto es el término unificado para describir los estados de ánimo, como las emociones. En [76] los autores definen que "los estados afectivos varían de varias maneras, incluyendo su intensidad, duración y niveles de excitación y agradabilidad" - que describiremos con más detalle en la Sec. 2.3.2 -. El estudio también declara que "las emociones desempeñan un papel importante en la regulación de la cognición, el comportamiento y las interacciones sociales, y el afecto se considera el estado experimental de los estados de ánimo". Aunque en el lenguaje cotidiano a menudo se utilizan indistintamente términos como afecto, emoción y estado de ánimo, el afecto se concibe como la categoría superior a la que pertenecen las emociones y los estados de ánimo" [76].
Los estados de ánimo y las emociones se diferencian sobre todo por su duración en el tiempo y las causas que los desencadenan. Las emociones son experiencias bastante intensas y efímeras que pueden producirse por dos motivos. Por un lado, pueden desencadenarse en respuesta a un estímulo externo concreto (por ejemplo, acontecimientos, acciones u objetos) y pueden surgir de forma algo inconsciente. Por otro lado, pueden seguir a un juicio cognitivo de un estímulo que ocurre en el momento (por ejemplo, contestando a ¿qué relevancia personal tiene este estímulo para mí?, ¿tiene este estímulo alguna relación con mis objetivos?) [77].
Por otro lado, los estados de ánimo tienen una mayor duración en el tiempo que las emociones y su naturaleza es más diversa. Por ejemplo, un sentimiento generalizado de tristeza sin un origen definido podría entenderse o interpretarse como un estado de ánimo. Estas experiencias de afecto que se producen de forma recurrente durante un periodo de tiempo prolongado pueden denotar el bienestar subjetivo de las personas, por ejemplo, su satisfacción global con la vida, o ser un signo de depresión.

El afecto tiene funciones cognitivas esenciales, lo utilizamos como suministro de información al hacer deducciones sobre objetos o personas, imprimando recuerdos agradables e influyendo en el procesamiento de la información y la toma de decisiones [77].

En esta tesis nos centramos en la detección de emociones más que de estados de ánimo. Las emociones pueden describirse como "las mejores conjeturas del cerebro sobre el significado de las sensaciones corporales, guiadas por la experiencia pasada" [78].

2.2. Bases Neurofisiológicas del Afecto y las Emociones

El sistema límbico del cerebro comprende un grupo de estructuras encargadas de regular las emociones y el comportamiento. Situado en las profundidades del cerebro, es la parte responsable de las respuestas conductuales y emocionales [79]. Algunas de las estructuras que están implicadas en las acciones del sistema límbico se encuentran debajo de la corteza cerebral y sobre el tronco encefálico [80]. Algunas de ellas son el tálamo, el hipotálamo que se encarga de la producción de las principales hormonas y de controlar la sed, el hambre y los estados de ánimo entre otros; y los ganglios basales, que recompensan el procesamiento, la formación de hábitos, la inclinación y el movimiento. Pero las dos estructuras principales y más importantes para el procesamiento emocional son el hipocampo y la amígdala [81].

– **Hipocampo**: Es en esencia el centro de la memoria del cerebro. Es donde se forman los recuerdos episódicos, se catalogan y se archivan en el almacenamiento a largo plazo a través de varias partes de la corteza cerebral. Las conexiones creadas en el hipocampo ayudan a asociar los sentidos con los recuerdos. La orientación espacial también tiene cierto origen en el hipocampo [80].

– **Amígdala**: Es clave para la generación de respuestas emocionales y está situada justo al lado del hipocampo, es especialmente responsable de sentimientos como el *placer*, el *miedo*, *la ansiedad* y *la ira*. El contenido emocional ligado a los recuerdos se debe a la amígdala. La amígdala modifica la intensidad y el contenido emocional de los recuerdos y desempeña un papel crucial en la creación de nuevos recuerdos, especialmente los relacionados con el *miedo*. Los recuerdos de *miedo* sólo se crean tras unas pocas repeticiones, lo que convierte al "aprendizaje del *miedo*" en un método muy conocido para investigar la formación y consolidación de los recuerdos [80].

– **Hipotálamo**: El hipotálamo es una parte del cerebro que se encarga del crecimiento, el metabolismo, la diferenciación sexual, las respuestas emocionales y los deseos e impulsos necesarios para que un individuo pueda sobrevivir [82]. El hipotálamo, junto con la glándula pituitaria, administra la presión sanguínea, la emisión de hormonas, la fuerza y el ritmo de los latidos del corazón, la temperatura corporal y los niveles de electrolitos y agua. El hipotálamo es también el núcleo para la administración de la actividad de las dos partes del Sistema Nervioso

Autónomo (ANS), los sistemas nerviosos simpático y parasimpático. La expresión emocional depende en gran medida del sistema nervioso simpático y está controlada por regiones de los hemisferios cerebrales situadas por encima del hipotálamo y por el mesencéfalo, situado por debajo.

El sistema límbico, en particular la amígdala, es clave para controlar diferentes comportamientos emocionales, como la ansiedad, la rabia y el *miedo* [83]. Y desde un punto de vista biológico, el *miedo* es una emoción clave, ya que ayuda al organismo a responder en consecuencia ante situaciones amenazantes que podrían ser perjudiciales para un individuo. La respuesta del *miedo* está provocada por la estimulación de la amígdala. Inicialmente, la amígdala activa el hipotálamo, que inicia la respuesta de lucha o huida.

2.2.1. Respuesta de lucha-huida-congelación (LHC o FFF)

La respuesta de lucha-huida-congelación -también conocida como respuesta de lucha o huida, o en inglés *fight-flight-freeze*, FFF- se produce como consecuencia de un acontecimiento que se reconoce como amenazante, es una reacción fisiológica automática del organismo [84]. Ejemplos de ello pueden ser ver que un vehículo que se aproxima se interpone rápidamente en nuestro camino, o ver que alguien nos está siguiendo mientras se camina por una calle. La respuesta de lucha o huida, desde una perspectiva evolutiva, se considera un instinto adaptativo que los humanos desarrollaron cuando los estímulos ambientales o los depredadores ponían en peligro la supervivencia de los seres humanos.

En concreto, la lucha o huida es una respuesta de defensa activa en la que la persona lucha o huye. El cuerpo se ve afectado por cambios fisiológicos para que la persona esté preparada para actuar adecuada y rápidamente. La congelación puede producirse antes de que el cerebro decida luchar o huir, quedando la lucha o huida en suspenso momentáneamente. También se llama inmovilidad reactiva. También ocurre en el momento en que la mente y el cuerpo son conscientes, mediante un proceso llamado neurocepción, de que luchar o huir ya no son alternativas para hacer frente a la amenaza percibida, es entonces cuando la respuesta cambia a la opción de permanecer inmóvil durante toda la situación amenazante como última alternativa para salvarse [85]. La lucha-huida-congelación es una reacción automática -no consciente- del cerebro y el cuerpo, que no se puede desencadenar ni controlar.

La respuesta de lucha-huida-congelación no sólo puede desencadenarse por un acontecimiento, sino también por un *miedo* psicológico. El cerebro asocia experiencias negativas a una situación concreta, lo que significa que el *miedo* está condicionado. Lo que puede causar *miedo* se denomina una amenaza percibida, o algo que el cerebro considera peligroso, que pueden ser diferentes para cada persona. Cuando se enfrenta a una amenaza percibida, el cerebro piensa que la persona está

en peligro, ya que reconoce la situación como una amenaza para su vida. Así, el cuerpo reacciona inconscientemente con la respuesta de lucha-huida-congelación para preservar la propia vida [86]. En estos casos, se dice que la respuesta de lucha-huida-congelación es hiperactiva. Lo que significa que se produce cuando situaciones de la vida normal que no son realmente amenazantes desencadenan la reacción. Estas respuestas hiperactivas son frecuentes en personas que han vivido acontecimientos traumáticos o padecen un trastorno de ansiedad. El ejemplo de ver a alguien caminando detrás de una misma por la calle no tiene por qué ser peligroso per se, pero puede desencadenar la respuesta de lucha o huida si usted es una mujer, la persona es un hombre corpulento y es más de medianoche.

A diferencia de los hombres, que experimentan mayoritariamente la respuesta de lucha o huida ante una situación amenazante -propuesta por primera vez por Bradford Cannon en 1915-, las mujeres parecen tener dos respuestas igualmente probables ante las condiciones estresantes, la de lucha o huida y la de "tiende y hazte amiga" (en inglés *tend-and-befriend*), presentada por Shelley Taylor en 2000 [87]. Esta respuesta de *tend-and-befriend* produce en el organismo cambios bioquímicos similares a los de la respuesta de lucha o huida.

Debido a la exclusión de las mujeres de los ensayos clínicos en la investigación -como ya describimos en la subsección 1.2.1-, hace tan sólo dos décadas se descubrió la teoría de *tend-and-befriend*. En ella se afirma que, ante una amenaza percibida, las mujeres tenderán a la protección de sus crías (*tend*) y a buscar a su grupo social para conseguir defensa mutua (*befriend*) [88]. Se cree que, debido a la selección natural, los humanos tienen un sistema biológico que gestiona las interacciones sociales del mismo modo que regula necesidades básicas como la sed o el hambre. Y parece tener sus raíces en esta necesidad instintiva de proteger a los niños y afiliarse con otros para mayor seguridad. Además de las necesidades básicas de bienestar físico, los humanos son criaturas sociales que se basan en sus instintos para interactuar con los demás. Las mujeres suelen inclinarse por proteger y cuidar a las crías.

Consecuencias en fisiología

Ante una situación peligrosa o amenazante, el estado emocional de la persona puede ser *miedo*, pánico, estrés, nerviosismo, shock, inseguridad, preocupación, ... Una serie de reacciones desencadenadas por el mecanismo de lucha-huida-congelación provocan cambios físicos y fisiológicos en el cuerpo humano.

La primera reacción ante una situación amenazante la genera la amígdala en el cerebro, más concretamente en el sistema límbico [77]. La amígdala es responsable de regular la respuesta de lucha o huida y desempeña un papel clave en el procesamiento del *miedo*. En el momento en que se percibe el peligro, la amígdala envía una señal al hipotálamo, responsable de la liberación de hormonas, y conecta los sistemas endocrino y nervioso. Este último informa entonces al resto del

organismo a través del sistema nervioso autónomo, encargado de controlar los mecanismos involuntarios del cuerpo, y estimula el sistema nervioso simpático (SNS).

Cuando la amígdala desencadena una señal de emergencia, el hipotálamo activa el sistema nervioso simpático transmitiendo señales a las glándulas suprarrenales [89]. Estas glándulas reaccionan inyectando adrenalina a través del torrente sanguíneo.

Esto provoca una serie de cambios fisiológicos que reconducen la energía de las zonas del cuerpo que están asociadas a procesos de reposo o pasivos - como el sistema digestivo - a las zonas del cuerpo que ayudan a la persona a estar preparada para la acción para que pueda evitar daños [90]. El corazón empieza entonces a latir más rápido que en estado normal, proporcionando sangre a los músculos, al corazón y otros órganos vitales. La frecuencia cardiaca y la presión arterial también aumentan. Los vasos sanguíneos de los músculos se dilatan y la tensión muscular aumenta para dotar al cuerpo de mayor velocidad y fuerza [77]. La frecuencia respiratoria aumenta y las vías respiratorias de los pulmones se abren para que pueda entrar la máxima cantidad de oxígeno en cada respiración. Todo el oxígeno extra disponible se envía al cerebro para que aumente el estado de alerta. Los sentidos como el oído y la vista también se agudizan [89] [91].

Durante la respuesta de lucha o huida, además de provocar vasodilatación en los músculos para conseguir velocidad y fuerza, el SNS también provoca vasoconstricción en la piel, para permitir que la sangre llegue a los órganos principales, dejando la piel con un aspecto más pálido. Al estrecharse los vasos sanguíneos, el cuerpo puede calentarse muy rápidamente, por lo que la respuesta de lucha o huida también aumenta la transpiración (sudoración). El cuerpo se enfría para evitar el sobrecalentamiento mediante la evaporación del sudor, y permite así seguir huyendo o luchando contra el daño sin sentirse agotado por el calor.

Además, la epinefrina -también conocida como adrenalina- desencadena la liberación de grasas y azúcar en la sangre, ya que son estos nutrientes los que inundan el torrente sanguíneo, proporcionando energía a todas las partes del cuerpo. Mientras persista la amenaza, el hipotálamo seguirá enviando señales al SNS para que siga segregando adrenalina y cortisol a fin de mantener la activación del organismo [92]. La liberación excesiva de adrenalina puede provocar un trastorno simpático por estrés agudo (ASD), también conocido como "estado de shock".

En concreto, el ANS -compuesto por los sistemas simpático y parasimpático- es responsable de regular los procesos fisiológicos involuntarios. Dentro del sistema límbico, el hipocampo -responsable de la creación de recuerdos con contexto emocional- forma representaciones emocionalmente significativas e interpreta los acontecimientos. La amígdala -responsable de asociar el *miedo* con las situaciones amenazantes- también es especialmente importante en el procesamiento de las emociones relacionadas con el *miedo*. Esta conexión entre la amígdala, el hipotálamo y el ANS está estrechamente relacionada con los reflejos y las respuestas de lucha o

huida, las expresiones de *miedo* en el cuerpo y la activación de neurotransmisores como la adrenalina y el cortisol, vinculados a las respuestas de estrés.

Cuando la respuesta no es ni de lucha ni de huida, sino de congelación o parálisis, el sistema nervioso parasimpático (PNS) toma el mando y lleva al extremo su papel de relajación activando el nervio vago, el nervio principal del PNS [93]. Esta red de nervios funciona como un freno al corazón, ya que lo ralentiza a un latido inferior al del estado de reposo, también adormece los sentidos y los músculos dispensando sustancias químicas en el torrente sanguíneo [85].

A partir de las respuestas de lucha-huida-congelación, el resultado depende de cómo el organismo haya aprendido a través de la experiencia a enfrentarse a cada tipo de amenaza, junto con el plan innato de lucha-huida en el cerebro, que determina la reacción más favorable para superar la amenaza [94] [95].

Consecuencias en la producción del habla

Como reacción a un peligro percibido o real, el ANS realiza diferentes cambios en el cuerpo, en relación con el ritmo cardiaco, la activación muscular y la vocalización, la respiración, para poder hacer frente a la situación amenazante. Dependiendo de la situación y de cada persona, hay diferentes variaciones de las respuestas de huida, lucha y congelación [96].

Uno de los principales responsables de estos cambios es el 10º nervio craneal o nervio vago. Es el nervio más largo del ANS y atraviesa la boca, la lengua, la laringe, el corazón, los pulmones y el aparato digestivo [97]. El tracto vocal específicamente -formado por las cuerdas vocales, la laringe y la faringe- tiene un complejo sistema nervioso parte del SNS.

El estrés, la ansiedad y el *miedo* pueden afectar profundamente al rendimiento vocal al desencadenar la respuesta de lucha o huida. La tensión muscular puede provocar la constricción de la garganta y las cuerdas vocales y hacer que la voz de la persona se vuelva más aguda, que la voz se apague o incluso que la pierda por completo [96]. La constricción muscular también puede provocar un aumento de la velocidad del habla, tensión en la mandíbula y la lengua, que dificulta la inteligibilidad, y parar la salivación, lo hace que la boca se sienta seca y convierte la voz más áspera. El aumento de la frecuencia respiratoria también puede provocar quedarse sin aire al hablar [96].

Al tener una respuesta de congelación o parálisis, el corazón se ralentiza a un ritmo inferior al de reposo y los músculos se sienten entumecidos debido a la liberación de sustancias químicas en el torrente sanguíneo. Esta reacción también bloquea las cuerdas vocales para mantener el flujo de oxígeno hacia los pulmones. Esta es la razón por la que en el modo de congelación una persona puede sentir que es físicamente imposible hablar, gritar o pedir ayuda [85].

Debido a todos estos cambios físicos y fisiológicos y a su naturaleza involuntaria -la persona no tiene control sobre ellos- que se producen en una persona como consecuencia de encontrarse en

una situación de riesgo, nos planteamos basarnos en señales fisiológicas como el pulso, la transpiración, la respiración, y también el habla, para detectar el estado emocional de una persona -con la intención de reconocer el *miedo*-, que podría ser consecuencia de encontrarse en una situación peligrosa. Un ejemplo de situación de amenaza para la vida que podría desencadenar las respuestas de lucha-huida-congelación en las mujeres son las situaciones de violencia de género, aquellas en las que una mujer sufre una agresión física o sexual. Encontrarse en una situación de peligro potencial puede conllevar los mencionados cambios físicos y fisiológicos en el organismo.

2.3. Teorías de la Emoción en la Ciencia

Tras examinar de dónde surge la respuesta emocional en el cerebro y cómo afecta al cuerpo, -especialmente el *miedo*-, ahora presentamos las dos diferentes perspectivas principales de la teoría emocional en el campo de la percepción de las emociones. Si bien es la disciplina de la neurociencia afectiva la que pretende desarrollar una comprensión profunda de las emociones, los estados de ánimo y los sentimientos y de cómo se integran en el cerebro, todavía no existe un consenso científico sobre que haya una única teoría válida de la naturaleza fundamental de la emoción. De ahí surgen las dos teorías aparentemente opuestas que rigen el campo de la percepción de las emociones: la teoría categórica y la teoría dimensional.

2.3.1. Las Emociones como Categorías Discretas

La teoría categórica de las emociones postula la existencia de seis emociones universales básicas bien definidas: "*felicidad*, *ira*, *tristeza*, *sorpresa*, *asco* y *miedo*" [98].

Estas emociones son básicas para los humanos, ya que estamos dotados de instrumentos biológicos para reaccionar ante situaciones vitales universales -como los éxitos o la pérdida-, y cada emoción orienta hacia una reacción que, durante la evolución, funcionó más eficazmente que otras soluciones en circunstancias relevantes similares para la supervivencia humana [99]. Cada una de las emociones básicas no es un estado fisiológico o afectivo individual, sino algo así como una familia de estados relacionados que las personas de todas las culturas pueden haber experimentado debido a problemas adaptativos similares, por lo que estas emociones se describen como universales.

Aunque la perspectiva categórica de la emoción humana no necesita esencialmente una explicación evolutiva de sus orígenes, los humanos y los animales experimentan categorías discretas de cada emoción, ya que se cree que cada una surgió de una adaptación que se desarrolló para resolver un problema adaptativo parecido [99]. Como ejemplo, se cree que la emoción discreta del *miedo* se desarrolló como un mecanismo para mejorar la supervivencia de los individuos evitando peligros a lo largo del tiempo a través de la evolución [100]. La tabla 2.1

proporciona las seis emociones básicas categóricas y, junto a cada una, la dificultad adaptativa para cuya resolución podría haber evolucionado la emoción [101].

Emoción discreta	Problema de adaptación
Felicidad	Buscar compañeros valiosos
Ira	Lidiar con una amenaza física en el entorno
Tristeza	Reforzar los vínculos sociales induciendo a la compasión [102]
Sorpresa	Darse cuenta de una anomalía [103]
Asco	Evitar o expulsar alimentos tóxicos
Miedo	Evitar el peligro

TABLA 2.1: Problemas adaptativos resueltos por las 6 categorías básicas de emociones, desde una perspectiva evolutiva [101].

La teoría básica de las emociones, aplicando la teoría *darwiniana*, sugiere que la característica de adaptabilidad de las emociones aumenta nuestra supervivencia genética mejorando las opciones reproductivas (por ejemplo, la alegría anima a la gente a explorar y conocer nuevas posibles parejas) o haciendo frente a las amenazas para la reproducción (por ejemplo, el asco nos ayuda a evitar la muerte) [101] [100]. Sin embargo, algunas investigaciones aplicadas a la AI están investigando las emociones más allá de las seis básicas [104], y la teoría definitiva de las emociones aún no ha sido consensuada en la comunidad investigadora.

2.3. Espacio Dimensional de las Emociones

Las teorías dimensionales se oponen a las teorías de las emociones discretas y básicas debido a que estas últimas no explican plenamente algunas observaciones de los estudios empíricos de neurociencia afectiva. El modelo del mapa del afecto [105] es una teoría dimensional que sugiere que "todos los estados afectivos derivan de percepciones cognitivas de impresiones neuronales que son el producto de un mínimo de dos sistemas neurofisiológicos independientes: uno relacionado con la excitación o el estado de alerta (también llamado *arousal)*, y otro relacionado con la valencia -una dimensión continua de placer-desagrado-" [106].

Estos modelos basados en dimensiones continuas -modelos dimensionales- piensan en las experiencias afectivas como una gama continua de estados bien interconectados e indefinidos [105]. Al final, las emociones se ven como "el producto de una comunicación intrincada entre las cogniciones, probable de ocurrir inicialmente en las estructuras neocorticales, y los cambios neurofisiológicos relacionados con estos sistemas de valencia y excitación" [105]. Existe un sistema asociado generalmente al placer y la recompensa, el sistema mesolímbico dopaminérgico, y podría representar la base neuronal para la dimensión de la valencia [105]. Además, se cree que

la formación reticular ajusta el equilibrio de la excitación del sistema nervioso central a través de sus conexiones con el sistema límbico, el tálamo y la amígdala [106].

Sin embargo, desde 1974, hay psicólogos que discuten en torno a esta teoría, sobre la interpretación específica de las dimensiones relacionadas con el afecto y la cognición. Y desde el punto de vista de la categorización de las emociones, utilizando la representación o espacio bidimensional *arousal*-valencia, emociones como el *miedo* o la *ira* estarían muy próximas entre sí, cuando en realidad tienen consecuencias fisiológicas diferentes y la sensación de cada una es distinta [107].

La nueva teoría dimensional del espacio PAD añade un eje al espacio *arousal*-valencia [108]. El espacio PAD (**p**lacer (valencia), ***a**rousal* (excitación) y **d**ominancia), está formado por tres dimensiones emocionales independientes que se cree que describen las emociones humanas [109]. Se cree que el placer -que es la valencia- es como un continuo que va de la felicidad intensa a la infelicidad extrema o el dolor; comprende extremos como felicidad y el enfado, satisfacción-insatisfacción y agrado-desagrado, para determinar el nivel de placer de una persona. Se sabe que la excitación (*arousal*) es la cantidad de actividad mental a lo largo de una única dimensión, que va desde el sueño hasta la excitación extrema. En cada extremo, las palabras que describen el *arousal* son estimulación-relajación, excitación-tranquilidad y despertar-somnolencia. Asimismo, se piensa que la dominancia esta relacionada con los sentimientos de control y restricción, expresando hasta qué punto una persona domina la emoción a partir de su comportamiento. El grado de dominancia se sitúa en un continuo que va desde la dominancia total a la sumisión, con descriptores como autonomía, influencia y control [109].

Las bases de datos emocionales pueden etiquetarse con emociones discretas o con emociones continuas, por lo que trabajos recientes proponen un mapeo discreto-continuo en la investigación sobre AC [110], [111], [112] y algunos incluso sugieren la necesidad de un cuarto eje para representar con precisión las emociones discretas en un espacio continuo 4D [113].

Centrándonos en la dominancia, ésta se refiere a la sensación de influencia y control sobre otras personas o/y el entorno o ambiente, frente a sentirse controlado o influenciado por otros o por la situación (por ejemplo, *ira*, *poder*, frente a *ansiedad* y *miedo*) [109]. Y puesto que en esta tesis nos centramos en detectar el *miedo* provocado por una situación de violencia de género, el eje de dominancia es de especial relevancia para el etiquetado de las emociones, explicando claramente cuándo una persona se siente completamente abrumada por la emoción y controlada en dicha situación.

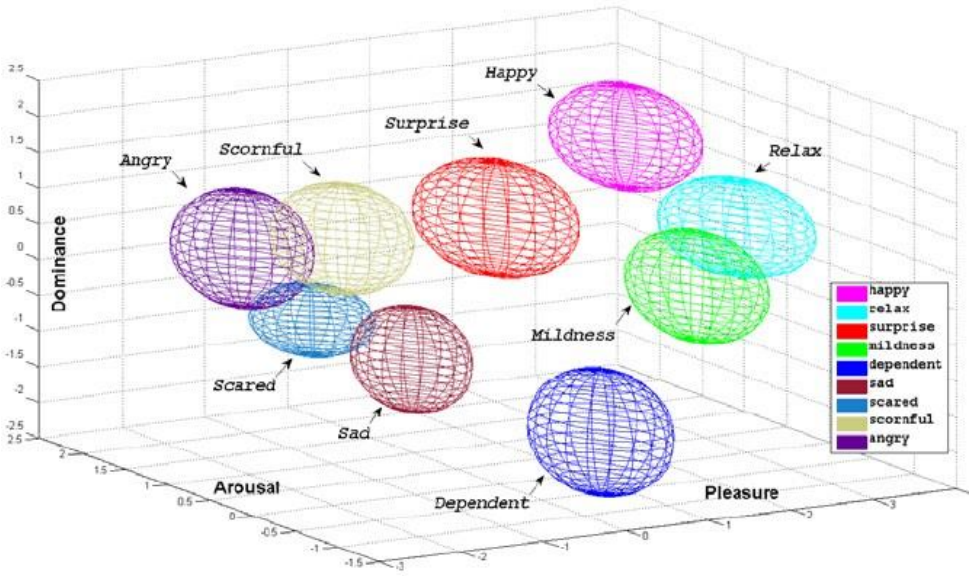


Figura 2. 1: Mapeo emocional del espacio de las emociones PAD abreviado de discreto a continuo [112]. Reproducido con permiso del propietario del copyright, Springer Nature.

2.4. Interpretación y Comprensión en la Computación Afectiva

Tras describir cómo surgen las emociones y qué teorías existen sobre cómo clasificarlas, en esta sección explicamos más a fondo la comprensión de esta rama de la inteligencia artificial que se ocupa de las tareas relacionadas con las emociones.

La Computación Afectiva (AC) surge a partir de las emociones o incide en ellas [114]. Se trata de un campo multidisciplinar en continuo crecimiento. Éste investiga cómo las máquinas pueden llegar a interpretar el afecto, y cómo la comunicación entre humanos y máquinas puede estar integrada por el afecto, cómo podemos diseñar sistemas con afecto de modo que se mejoren sus capacidades, y cómo la interacción con las máquinas puede transformarse mediante la detección y las estrategias afectivas [115]. Abarca varias disciplinas como la psicología, la ingeniería, la ciencia cognitiva, la educación y la sociología, entre otras.

La Computación Afectiva se basa en el Aprendizaje Automático y el Aprendizaje Profundo. Según [116], "el Aprendizaje Automático es el subcampo de la informática y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender". Se considera que un agente aprende cuando su eficacia mejora gracias a la experiencia con el uso de datos. En el proceso de aprendizaje automático, primero se alimenta a un programa informático con datos y éste los observa, luego construye un modelo de predicción basado en dichos datos, y utiliza el modelo de dos maneras, primero como una hipótesis sobre el mundo y segundo como un método para resolver problemas y hacer inferencias y predicciones sobre nuevos datos.

Por otro lado, el aprendizaje profundo es un conjunto de algoritmos de aprendizaje automático con el mismo propósito, pero además pretenden modelar representaciones de alto nivel en los datos utilizando arquitecturas computacionales complejas que soportan transformaciones no lineales de los datos [117]. Así, éstas tienen la capacidad más flexible para modelar problemas del mundo real con mayor generalizabilidad (posibilidades de generalización). El aprendizaje profundo forma parte de un grupo más amplio de métodos de aprendizaje automático que se basan en la comprensión de las representaciones de los datos. La investigación en este ámbito intenta definir qué representaciones son óptimas para ser comprendidas y cómo crear modelos para reconocer e interpretar estas representaciones.

Las disciplinas de la Inteligencia Artificial pueden clasificarse según el tipo de datos que emplean, por ejemplo, el campo de la Visión por Computador (CV) analiza imágenes y vídeos, las Tecnologías del Habla (ST) trabajan con el procesamiento de datos del habla, el Procesamiento del Lenguaje Natural (NLP) utiliza datos textuales, ... Pero la Computación Afectiva es una disciplina de la IA definida por la tarea a realizar, no restringida por el tipo de datos utilizados.

Una categorización que se ha hecho en este campo [118] establece cuatro áreas distinguidas sobre el mismo, que son: "1. el análisis y la caracterización de los estados afectivos -identificación o detección-, 2. el reconocimiento automático del estado afectivo -reconocimiento-, 3. la expresión de los estados afectivos -generación y elicitación-, 4. y la adaptación de la respuesta al estado afectivo de la usuaria".

Así, dentro de las tareas objetivo que implican emociones en la AC, ayudándonos por las definiciones del diccionario [119] (descritas a continuación con letra cursiva), podemos definir a continuación las siguientes tareas:

Detección - "*darse cuenta de algo que está parcialmente oculto o no está claro o descubrir algo, especialmente utilizando un método especial*". En la AC, también puede conocerse como identificación, y se aplica a encontrar una alteración del estado emocional de una persona, por un descubrimiento de algo que está sucediendo y que está alterando el estado basal neutro en el que la persona se encontraba inercialmente. Entre las aplicaciones de esta área se encuentra la identificación de emociones o desencadenantes que las provocan.

Reconocimiento: "*conocer a alguien o algo porque ya se ha visto, oído o experimentado antes*". Va un paso más allá que la detección en AC, ya que analiza el nuevo estado detectado y lo categoriza, en alguna categoría -ya sea continua o discreta- de estados emocionales ya conocidos. Esta área es una consecuencia de la anterior, ya que además de identificar el momento en el que se produce una emoción, también se encarga de clasificarla en un tipo conocido.

Generación - "*la producción o creación de algo*". A este ámbito pertenece la capacidad de crear y sintetizar emociones mediante máquinas. Ejemplos como la generación de emociones en el habla -*Emotional Text-to-Speech* (TTS)- o la generación de texto con contenido emocional

Elicitación - "*producir algo, especialmente una reacción*". Esta área se encarga de inducir, desencadenar o evocar un estado emocional en una persona, que en última instancia puede influir en sus acciones o reacciones. Con aplicaciones como la inducción al confort en la atención al cliente telefónica.

La convergencia de las cuatro áreas mencionadas da como resultado la capacidad de **adaptación** de las máquinas, reflejo de la *empatía*. La adaptación se define como "*algo producido para ajustarse a diferentes condiciones o usos, o para hacer frente a diferentes situaciones*". Entonces puede decirse que simulan la empatía, cuando son capaces de identificar un estado emocional en una persona, y reconocerlo, para luego generar una respuesta afectiva sintética, y ser capaces de provocar otras reacciones emocionales en las personas, imitando el mecanismo de la *empatía* humana.

Pero esta llamada empatía es un proceso holístico para el que las máquinas necesitan comprender no sólo las emociones, sino también el contexto y la situación que llevaron a la persona a ese estado afectivo. En nuestro caso concreto, pretendemos detectar situaciones de riesgo para las mujeres, y esto no sólo se consigue reconociendo el estado afectivo de una mujer mediante sus datos de usuaria -es decir, variables fisiológicas y/o habla- sino también junto con los datos contextuales -es decir, ubicación GPS, sonidos ambientales, hora del día,… -. Todos estos datos juntos proporcionan información contextual para realizar una interpretación y comprensión holísticas de la situación, lo que podría, en última instancia, determinar cuándo está en peligro la vida de una persona.

La interpretación y la comprensión son clásicas en otros campos de la IA, por ejemplo, en el reconocimiento del habla, se identifican los fonemas del habla hablada -decisión sobre si hay silencio o voz-, y se reconocen -vocales y consonantes- qué palabras y frases tienen un significado, interpretado por nuestro entendimiento cognitivo superior. En este trabajo perseguimos un objetivo similar, hacer que sea posible identificar, por ejemplo, un aumento repentino de la frecuencia cardiaca, junto con el reconocimiento de pasos apresurados y jadeos a partir de una señal de audio, y que podamos interpretar conjuntamente esos acontecimientos -a primera vista, aislados-, realizando una comprensión holística de la situación, determinando su nivel de riesgo para la usuaria.

Dicho esto, con toda esta información pretendemos dar un paso adelante con Bindi en la Computación Afectiva, hacia un nivel cognitivo superior. Pretendemos no sólo analizar los datos de una usuaria para detectar y reconocer emociones aisladas desde un nivel computacional básico, sino ir más allá hasta *interpretar* y *comprender* dichos estados afectivos, junto con la información situacional. Esto nos llevará a comprender su contexto y circunstancias, para finalmente poder detectar una situación de riesgo, amenaza o peligro para una mujer.

2.5. Desafíos: Subjetividad, Anotaciones y Género

Rosalind Picard, quien acuñó el término Computación Afectiva [75], describe el término *emoción* como "las relaciones entre los incentivos externos, los pensamientos y los cambios en los sentimientos internos; al igual que el tiempo atmosférico es un término superordinado para las relaciones cambiantes entre la velocidad del viento, la humedad, la temperatura, la presión barométrica y la forma de las precipitaciones" [120]. Define una metáfora meteorológica afirmando que "una combinación única de cualidades meteorológicas crea una tormenta, un tornado, una ventisca o un huracán, acontecimientos que son análogos a las emociones del *miedo*, la *alegría*, la *excitación*, el *disgusto* o *la ira*. Pero el viento, la temperatura y la humedad varían continuamente y no necesariamente producen combinaciones tan extremas. Así pues, los meteorólogos no se preguntan qué significa *el tiempo*, sino que determinan las relaciones entre las cualidades medibles y más tarde dan nombre a las coherencias que descubren" [120]. Al final, Rosalind Picard afirma que es difícil esperar que los investigadores tengan éxito a la hora de hacer coincidir etiquetas humanas cuando esas etiquetas pueden no existir específicamente, comparando el problema con no tener términos específicos para la mayoría de los estados del tiempo, sino sólo nombres para sus estados extremos, y lo mismo se aplica a las emociones.

Esta metáfora deja claro que las emociones no son objetivas, no son dígitos que puedan reconocerse y diferenciarse claramente. Las emociones están impregnadas de subjetividad, y esta particularidad es bidireccional, ya que tiene dos direcciones.

En primer lugar, existe una dificultad intrínseca para etiquetar o categorizar los sentimientos más íntimos de una misma, a pesar de que tenemos mejor acceso a ellos que cualquier otra persona. Todavía muchas personas no saben cómo conectar con su propio estado de sentimientos y reconocerlos, aunque la gente tiene sentimientos permanentemente [121]. Muchas bases de datos de Computación Afectiva se etiquetan mediante auto-anotaciones de los sujetos participantes; a veces con respecto a categorías discretas de emociones y a veces refiriéndose a ejes continuos -espacio PAD-. También depende de la formación emocional de cada persona, es decir, si a la persona no se le ha enseñado a identificar sus propias emociones o no tiene experiencia en reconocerlas - falta de inteligencia emocional, aún asignatura pendiente en muchas escuelas [122]- sus propias emociones, puede tener percepciones diferentes de lo que significan las escalas *Likert* de *avrousal* o valencia. Esto puede variar mucho en función de los antecedentes y la cultura de la persona. Así como 2mm de precipitación por hora es una lluvia débil en España, en las islas Filipinas puede que ni siquiera se considere lluvia.

En segundo lugar, otras bases de datos son etiquetadas por anotadores externos, independientes de la persona que experimenta la emoción. Y en el camino de una emoción desde su generación hasta su exteriorización, la persona puede tener cierto grado de control de dicha exteriorización - ya explicamos algunas de las consecuencias incontrolables del *miedo* desde el sistema nervioso

autónomo, pero no toda la exteriorización es automática, directa o inevitable -, lo que dificulta que el anotador determine correctamente la emoción que presenta la persona si ésta no la exterioriza abiertamente.

A estas dos se añade una tercera subjetividad. En el caso particular en el que se trata de anotar las emociones elicitadas en una persona mediante la visualización de un estímulo audiovisual -u otro sensorial (por ejemplo, olfativo, gustativo o táctil)-, o de una experiencia de la vida real de un tipo específico, con el objetivo de lograr una emoción, la situación puede ser percibida de forma diferente por una persona que por otra. Para una persona la visualización -o la experiencia- de la vida nocturna en una ciudad abarrotada puede ser excitante, atractiva, emocionante, pero para otras puede ser estresante, perturbadora o tensa. Esto significaría que incluso cuando se intenta suscitar una emoción determinada en los espectadores -por ejemplo, para la generación de una base de datos- puede que no se consiga suscitar la emoción objetivo porque no todas las personas reaccionan de la misma manera ante los mismos estímulos.

Toda esta subjetividad de las emociones significa que, en el campo de la Computación Afectiva, donde los modelos de AI se entrenan con datos y etiquetas, no disponen de etiquetas objetivas "en blanco y negro" como en otras áreas de la AI. Entonces, las etiquetas emocionales auto-anotadas o anotadas externamente no deben tomarse como etiquetas estándar absolutas. Las personas pueden reaccionar de forma diversa ante los mismos estímulos, incluso en distintos momentos, en función de muchas variables, como el estado de ánimo, las experiencias pasadas, la cultura y los antecedentes. Eso hace de la Computación Afectiva, un campo de la Inteligencia Artificial subjetivo y ligeramente escurridizo, en el que deben tenerse en cuenta todos estos matices.

En consonancia con la interpretación y comprensión de las situaciones en su conjunto, no existen hasta la fecha -que sepamos- bases de datos que sirvan específicamente para identificar y comprender las situaciones emocionales. Y como se describe en el apartado 2.4, hay que alejarse del marco teórico en el que se analizan y procesan las emociones aisladas, para entender las situaciones emocionales o afectivas de forma holística, teniendo en cuenta el contexto, para comprender plenamente por qué una situación suscita una emoción determinada en una persona. Esto es crucial en la detección de situaciones de riesgo.

En otro orden de cosas, en paralelo a la sección 1.2.3, y en línea con la subjetividad de las emociones que dependen de la persona, otro reto que se plantea es el de la personalización en función del sexo. Parece haber claras diferencias en la expresión de las emociones según el sexo [123]. Se ha descubierto que los hombres y las mujeres muestran con mayor precisión las expresiones estereotipadas de género, -derivadas de la socialización de género-, ya que los hombres expresan más exactamente emociones como la *ira* y el *desprecio*, mientras que las mujeres expresan con mayor exactitud el *miedo* y la *felicidad* [124] [125]. En lo que respecta específicamente a la elicitación de emociones, al visualizar vídeos de violencia de género, la

identificación de los espectadores con el o la protagonista del vídeo afecta directamente al etiquetado, ya que las mujeres prefieren etiquetar principalmente el *miedo*, mientras que los hombres etiquetan la emoción como *ira* o *tristeza* [126]. Podría ser que las restricciones sociales a la expresión emocional en los hombres fueran una de las razones de los elevados índices de violencia contra las mujeres perpetrada por hombres. Estos ideales masculinos, como la presión para cumplir las expectativas de dominación que impone la sociedad, podrían aumentar el potencial de los chicos para implicarse en actos de violencia general, como asaltos, acoso escolar y/o agresiones físicas y verbales [127].

En cuanto a esas diferencias de género en los datos utilizados para entrenar modelos de aprendizaje automático (*Machine Learning,* ML), algunos estudios muestran que tener en cuenta la edad y el género en la AC puede suponer una mejora en la precisión de las tareas de reconocimiento de emociones [67], y afirman que las variables específicas de la persona no deben supervisarse en el análisis de la AC, como el género, la personalidad y la edad. Es bien sabido que los sistemas de reconocimiento de emociones dependientes del género rinden más que los independientes del género, por lo que algunos estudios mejoran la calidad de discriminación de los rasgos dependientes del género [128], o modelan la información de género para una representación emocional más robusta [129], con el fin de lograr mejores precisiones. Se describen resultados sobre este tema en la sección 3.3.1.

2.6. Consideraciones Éticas, Prácticas y Legales

La Computación Afectiva ha tenido un gran impacto social desde su aparición. Algunas preocupaciones éticas que deben debatirse en la AC están relacionadas, de forma genérica, con la discriminación y los prejuicios, el abuso de influencia y la manipulación, la salud mental y la seguridad, y la privacidad de los datos sensibles.

Sin embargo, no existe una guía de principios para los protocolos y normas éticas de investigación contemporáneos a fecha de la escritura de esta tesis, pero algunos estudios pretenden recoger los más comunes, como [130], [131]: "1. consentimiento informado, que implica evitar la observación encubierta o secreta de los participantes 2. privacidad de los participantes (confidencialidad y anonimato) 3. evitar el daño (incluido el efecto psicológico) y hacer el bien 4. conocimiento de los grupos vulnerables 5. derecho de los participantes a retirarse o darse de baja 6. uso restringido de los datos 7. debido cuidado en el almacenamiento de los datos 8. evitar los conflictos de intereses".

Rosalind Picard [132] plantea la preocupación de que "un ordenador que pueda expresarse emocionalmente actuará emocionalmente algún día". En el caso de la mencionada *adaptación* de las máquinas, el inconveniente es la capacidad de éstas de manipular a los humanos. Por ejemplo,

en el caso de las empresas que comprenden mejor las necesidades y los deseos de sus clientes, lo que hace posible crear un nuevo tipo de marketing, dirigido al apego emocional y al control.

En [133], hay un amplio debate sobre el tema de la preservación de la privacidad y las tecnologías en las tareas de caracterización del hablante y del habla. En [134], ofrecen una visión general de los fenómenos paralingüísticos que pueden utilizarse o incluso se utilizan para obtener información personal mediante las señales del habla. En [131], los autores sugieren directrices de buenas prácticas en paralingüística computacional (CP) y AC, como elegir la métrica de rendimiento adecuada y tener en cuenta la interpretabilidad y la representatividad [135].

Desde el punto de vista de **nuestra aplicación** -desarrollar un dispositivo *wearable* capaz de detectar situaciones de riesgo para la usuaria y alertar automáticamente a los servicios de emergencia en caso necesario-, hay algunas preocupaciones éticas clave a tener en cuenta.

Es posible que un sistema de este tipo -especialmente en sus primeras fases de desarrollo- pueda cometer errores. Podría disparar falsas alarmas e incluso pasar por alto situaciones de riesgo. Y esta posibilidad de fallo podría tener consecuencias peligrosas. Es necesario encontrar un equilibrio entre tener "falsos negativos" (que el sistema prediga que no existe riesgo cuando en realidad sí lo hay) y no tener "falsos positivos" (que el sistema prediga que existe un riesgo que en realidad no existe), siendo preferible no tener nunca ninguno de los primeros a costa de tener alguno de los segundos. Un enfoque denominado aprendizaje *pasivo-agresivo se* encarga de arreglar los falsos positivos y reducir las alertas, en los modelos de aprendizaje automático [136].

Ya hemos hablado anteriormente de los sesgos en ML en la sección 1.2.1. En muchas ocasiones, el problema de los sesgos en el ML proviene del hecho de que la mayoría de los algoritmos se consideran cajas negras, es decir, proporcionan las salidas deseadas en respuesta a las entradas que se introducen, pero no son capaces de explicar cómo han llegado a esa conclusión. Cuando dejamos que estos algoritmos tomen decisiones que tienen gran importancia para las personas que toman esas decisiones, como en el caso de la detección de situaciones de riesgo, deberíamos ser capaces de explicar las razones de esas decisiones. Tanto para el personal técnico de supervisión como para la usuaria que utiliza Bindi, disponer de un sistema altamente explicable proporcionaría a todos confianza en su uso, además de poder ver con mayor claridad cómo los cambios que los desarrolladores introducen en el sistema afectan a las decisiones que toma Bindi.

Un par de puntos a tener en cuenta, también mencionados en la Sec. 1.2.3 son el PTSD y la diversidad.

El hecho de que entre las usuarias que utilicen Bindi se encuentren necesariamente víctimas de violencia de género, requiere una atención especial porque la violencia que sufrieron tiene consecuencias de estrés postraumático en ellas. Bindi debe tener en cuenta las necesidades específicas que pueden necesitar este tipo de usuarias respecto a las mujeres que nunca la han sufrido, teniendo en cuenta de alguna manera una evaluación de la sintomatología postraumática.

Existe una gran diversidad cultural en España, y aún más en Europa y en el mundo. La cultura es un sistema de orientación para una nación, sociedad, organización o grupo. Ésta es también un sistema de valores y normas que influye en la acción subconsciente. Ytodos estos aspectos afectan a la forma en que interactuamos con el mundo, incluida la forma en que expresamos las emociones. Bindi debe tener en cuenta el grupo destinatario al que se orienta y que utilizará en cada momento, para poder ofrecer un sistema que pueda proteger a todas las mujeres, teniendo en cuenta sus diferencias individuales.

En el aspecto jurídico, el marco normativo europeo y nacional hace hincapié en los retos de la AI ligados a la necesidad de procesar datos reales, incluido el compromiso entre privacidad y protección de datos y las tensiones entre explicación y predicción. Abordar estos retos es una tarea que requiere la existencia de una legislación adecuada. Ya existe legislación europea relacionada con la protección de datos, el Reglamento General de Protección de Datos (RGPD) [137], pero además se está trabajando para regular específicamente la AI: la Ley de Inteligencia Artificial de la Comisión Europea [73] es un borrador que se espera aprobar en breve. La *Secretaría de Estado de Digitalización e Inteligencia Artificial (SEDIA)* propone probar en España el nuevo Reglamento Europeo sobre Inteligencia Artificial a través de un proyecto piloto que pondrá a prueba una nueva agencia: la *Agencia de Supervisión de Algoritmos,* cuya creación está prevista para finales de 2022. El nuevo reglamento europeo entrará en vigor el 1 de enero de 2024.

2.7. Revisión de la Literatura sobre la Computación Afectiva y la Violencia de Género

La Computación Afectiva hace uso de modelos de IA - Machine Learning y Deep Learning - para las tareas mencionadas: detección, reconocimiento, generación y elicitación, de emociones. Y no hay mucha literatura sobre AC y GBV juntas debido a la falta de datos de víctimas de violencia de género, como se explica en la sección 1.2.1. Pero se trata de un campo emergente, ya que el mundo y la comunidad investigadora lo están viendo como la amenaza a la vida humana y a los derechos humanos que es, y la investigación sobre la fusión de estos dos campos está creciendo. En el campo de la NLP, unos investigadores [138] capturaron un conjunto de datos para la identificación de feminicidios (el asesinato de mujeres por el hecho de ser mujeres) a partir de 400 informes de medios de comunicación escritos, junto con sus etiquetas. También entrenaron un modelo de aprendizaje automático utilizando estos datos, logrando una precisión en el conjunto de datos de prueba del 81,1%, con 400 muestras (artículos escritos).

En [139], los autores examinan los fundamentos de los esfuerzos de los activistas y de la sociedad civil para recopilar contra-datos sobre feminicidios y asesinatos relacionados con el género, revisando los esfuerzos de los activistas para vigilar y cuestionar la violencia de género.

En [140], utilizan una base de datos recopilados a lo largo de dos décadas sobre la violencia de género en España. Utilizan la selección de características, se aplican algoritmos predictivos y se comparan para predecir con bastante éxito el número de denuncias por violencia de género que se presentarán ante un tribunal en los próximos seis meses en el país. El mismo equipo [141] tiene un estudio sobre una solución de vigilancia basada en biosensores para la protección de la violencia de género, similar a nuestra contribución [1], que sirve como constatación de que la tecnología es cada vez más aceptada y utilizada como solución para combatir y mitigar la violencia de género.

Con el auge de las redes sociales en la última década, y junto con el ciberactivismo y el ciberacoso, algunos trabajos [142], [143] exploran modelos de redes neuronales para identificar la violencia de género en mensajes de Twitter en lengua española con base en México, y discriminar entre tuits etiquetados manualmente con intención de violencia de género.

Desde el punto de vista de la otra parte de la violencia, es decir, del perpetrador, en [144] analizan las variables más influyentes y también predicen la probabilidad de perpetración de violencia de género utilizando datos de cuestionarios de jóvenes sin hogar de Los Ángeles. Se utilizan varios algoritmos de aprendizaje automático supervisado para construir una herramienta de triaje de la perpetración de violencia en la pareja íntima (IPV) con el fin de detectar qué jóvenes corren un alto riesgo de participar en la perpetración de actos violentos.

En lo que respecta a la salud mental, la violencia de género provoca trastornos traumáticos como el Trastorno por Estrés Agudo (ASD) y el Trastorno por Estrés Postraumático (PTSD), y existe literatura actual sobre el uso de métodos de aprendizaje automático en la estimación de sujetos con ASD y PTSD [145] en la que se utilizan múltiples niveles de datos biológicos -clínicos, neuroendocrinos, psicofisiológicos- u otras fuentes de datos -por ejemplo, información demográfica- para predecir síntomas tempranos o identificar factores de riesgo relacionados con el PTSD o el ASD.

También existe un interés en la comunidad investigadora por generar voces neutras desde el punto de vista del género para los asistentes de voz y eliminar los prejuicios sexistas [146]. Pero, en general, en el campo de las tecnologías del habla hay poco o ningún trabajo para combatir o prevenir la violencia de género. Así pues, esta tesis pretende llenar ese nicho y explorar e investigar el uso de las tecnologías del habla para la prevención de la violencia de género, despertando también el interés de la comunidad investigadora en el desarrollo de soluciones para la prevención del tan complejo problema de la violencia de género.

Capítulo 3: Caracterización de Datos para la Detección de Situaciones de Violencia de Género

La forma en que se capturan los datos influye en la metodología que puede aplicarse a los mismos, por lo que la metodología tiene que ir de la mano del proceso de captura de datos. En este capítulo queremos reunir el esfuerzo metodológico desde el punto de vista de las bases de datos, explicando las decisiones tomadas con respecto a los datos utilizados en orden cronológico para la investigación de esta tesis. Detallamos las dificultades encontradas para alcanzar nuestros objetivos debido a la falta de datos adecuados disponibles, ya que los conjuntos de datos de habla de *miedo* real (no actuado) no están disponibles o no existen en la literatura. La emoción realista más cercana al *miedo* es *el estrés*, por lo que en este capítulo describimos y justificamos el uso de conjuntos de datos que contienen dicha emoción como punto de partida de nuestra investigación. Además, y como consecuencia del problema anterior, describimos una de las principales contribuciones del equipo UC3M4Safety, que es la creación de nuestro propio conjunto de bases de datos para cubrir ese nicho de la literatura.

El diseño y la recopilación de los conjuntos de datos descritos en los apartados 3.3 y 3.4 ha supuesto un enorme esfuerzo de los miembros del equipo UC3M4Safety que participan en el proyecto EMPATIA-CM. Como parte de este esfuerzo, se han realizado las siguientes contribuciones en esta tesis: el diseño de la recopilación de datos de habla y audio, la asistencia técnica y el apoyo al protocolo de captura de datos, el procesamiento de la canalización de datos de habla, así como la asistencia en su captura y el seguimiento de la usuaria tanto durante WEMAC como en WE-LIVE.

En línea con una conceptualización moderna de la AI centrada en los datos (*data-centric AI*) [147], queremos centrarnos en el uso de datos apropiados para nuestra tarea, dando prioridad a la importancia de los datos más adecuados. Esta técnica única y reciente consiste en construir sistemas de AI con datos de calidad, haciendo hincapié en que los datos expresen claramente lo que la AI debe aprender, en lugar de centrarse en escribir código. Esta conceptualización surgió debido a las primeras soluciones de AI que eran más costosas adoptadas para mejorar los modelos de AI a lo largo de los años -en términos de recursos y económicos-, y este enfoque apuesta por un cambio fundamental necesario para liberar realmente todo el potencial de la AI, proporcionando un método sistemático para mejorar los datos, llegar a un consenso sobre los mismos y limpiar los datos incoherentes.

3.1. Desafíos de los Datos de Audio para la Detección de la Violencia de Género

En nuestra aplicación queremos detectar situaciones de riesgo de violencia de género a través de la modalidad auditiva. Definimos en la Fig. 3.1 un esquema de los datos auditivos, las tareas y

las condiciones en las que pueden registrarse, para utilizarlos en la detección de situaciones de violencia de género. Hay 3 componentes principales en los que se pueden dividir las tareas que pueden utilizar datos auditivos: *habla* -en la que se pueden realizar tareas relacionadas con el hablante y la emoción-, *audio (*que no es habla*)* -para la detección de eventos acústicos y la clasificación de sonidos-, entre otros-, y *el ruido de fondo* -que podría añadir información beneficiosa o perjudicial, dependiendo de la tarea-.

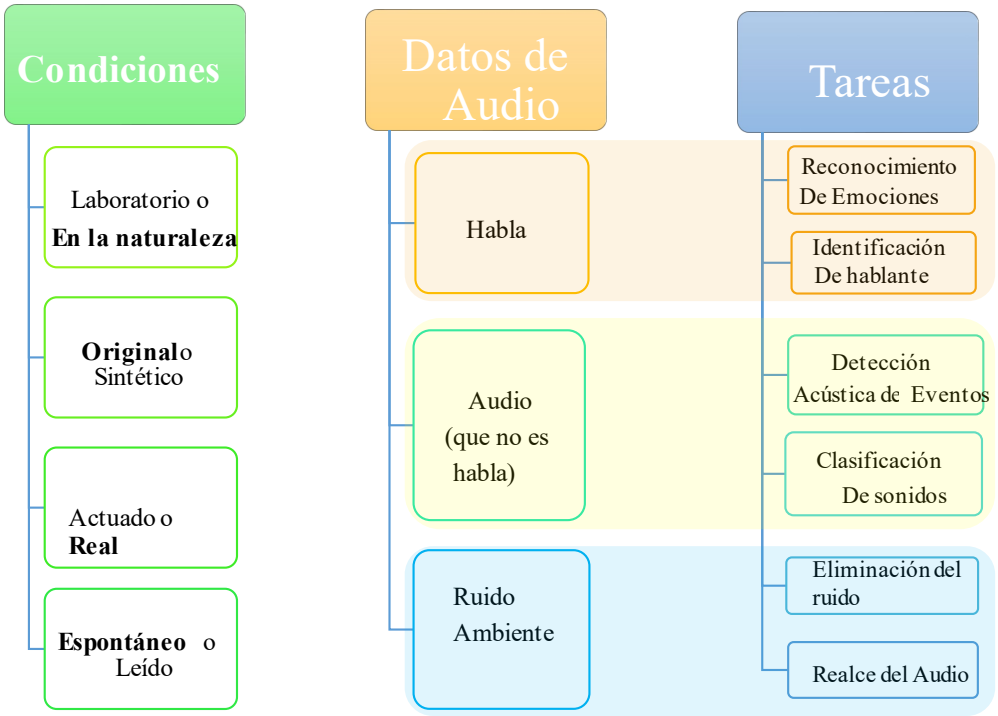


Figura 3. 1: Esquema de los datos auditivos que se utilizarán en la detección de situaciones de violencia de género.

Idealmente, para construir nuestro sistema de AI, ML o DL para la detección automática de situaciones de riesgo a partir de datos auditivos, nos gustaría contar con el habla propia de la usuaria en cuestión, más el audio de fondo que podría darnos el contexto acústico en el que se encuentra la usuaria. Lo ideal sería que el primero fuera habla limpia (no ruidosa) propia y espontánea, incluyendo el habla neutra y emocional -y con *miedo*, a ser posible-. El segundo debería incluir eventos acústicos y ruido de fondo -que no siempre es perjudicial pero puede añadir información contextual- para que la situación pueda comprenderse totalmente. También serían deseables fuentes de información adicionales, como biosensores que midan las señales fisiológicas de la usuaria, ya que proporcionarían una visión más fiable de las circunstancias y así analizar de manera exhaustiva la situación. Así el uso de estos datos conduciría al desarrollo de modelos ML eficaces para la detección de situaciones de riesgo.

En esta tesis seguimos un enfoque ascendente (*bottom-up*), centrándonos primero en identificar a la hablante y el contenido emocional de su voz, y luego complementamos las conclusiones obtenidas con la información auditiva adicional, para obtener una visión completa de la situación. Como primer paso, necesitamos reconocer a la usuaria (identificación de la hablante) y después detectar la aparición del *miedo* en el habla (reconocimiento de la emoción). Para ello hacemos un análisis de las bases de datos disponibles en la literatura que podemos utilizar y después proporcionamos una descripción completa de nuestra iniciativa por crear una base de datos adecuada para nuestros fines.

Las bases de datos de audio suelen grabarse en condiciones de laboratorio, pero cada vez hay más que se graban *in-the-wild* (en la naturaleza, o en condiciones de la vida real). Esta metáfora, parecida a la definición del diccionario de *in-the-wild* -seres que viven libres y en estado natural[20] -, invoca condiciones reales, de la vida real, y con características naturales. Sin embargo, el rendimiento de los modelos de AI evaluados en la naturaleza sigue siendo poco fiable, en parte debido a la diversidad de los datos y a los factores contextuales, a menudo desconocidos, como por ejemplo las condiciones de captura.

Los datos en condiciones de laboratorio suelen tener la ventaja de proceder de un trabajo de estudio en el que hay un objetivo al grabar, estableciendo exactamente qué se pretende captar, lo que resulta de gran ayuda. La desventaja es que estas condiciones de laboratorio suelen estar lejos de las condiciones de la vida real porque no incluyen todo lo que puede aparecer al grabar en la naturaleza. Las grabaciones de audio en laboratorio suelen ser claras y limpias -sin ruido o con muy poco ruido-, mientras que en la naturaleza esas condiciones no se cumplen (las grabaciones suelen contener ruidos, como ruido de vehículos, sonidos domésticos, exteriores e interiores, incluso sonidos de golpes en el micrófono o roces si se lleva puesto, etc.). Estas dos configuraciones son muy diferentes, los datos grabados en la naturaleza se ven afectado por una gran variabilidad que puede empeorar el rendimiento de los modelos.

Los datos de audio generados sintéticamente no pueden emular a la perfección los datos de audio reales, pero esta generación artificial es una forma de probar cómo funcionan los modelos en condiciones específicas utilizando sólo unos pocos recursos -sin emplear la gran cantidad de recursos necesarios (personal, material, tiempo, …) para la grabación de una base de datos-. Cuando nos referimos a datos generados *sintéticamente* no estamos describiendo datos *sintetizados* que son resultado de usar modelos que convierten texto en voz, hablantees o generadores de voz; en su lugar nos referimos a datos de audio *aumentados* (que han sido modificados artificialmente en factores como la velocidad, el tono, la combinación o suma de dos señales, etc.). Estos datos generados sintéticamente podrían servir como forma preliminar de

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

probar modelos de ML y obtener resultados similares a los que podríamos esperar al usar datos grabados en esas condiciones específicas.

Las bases de datos de habla emocional y neutra pueden ser grabadas por actores que simulan hablar bajo esas emociones, o por personas con emociones reales previamente inducidas. Debido a las características de nuestra tarea, preferimos priorizar el uso de bases de datos emocionales reales en lugar de las actuadas, ya que las bases de datos de actores pueden estar sobreactuadas, lo que pone en duda el valor de utilizar actores para investigar las emociones reales [148]. La diferencia entre habla espontánea o leída da lugar a otro tipo de categorización. Para nuestro caso de uso, buscamos un habla que sea preferiblemente espontánea, de nuevo, más parecida al habla de la vida real.

Un modelo ML que pudiera entrenarse con este tipo de datos sería bueno para la inferencia y la capacidad de predicción en situaciones del mundo real con habla emocional espontánea y real. Pero en la literatura hay muy pocas bases de datos originales, reales y espontáneas que incluyan habla emocional -y lo que es más importante, que incluyan el *miedo*- de diferentes hablantes, especialmente mujeres. Así pues, teniendo en cuenta estas tres limitaciones, en esta tesis trabajamos inicialmente con las bases de datos más adecuadas encontradas en la literatura con habla real y espontánea, que incluyen habla con *estrés*, un pariente cercano del *miedo*.

Por su importancia a lo largo de la tesis, creemos necesario desarrollar más aún el tema del *estrés*. A pesar de que *el estrés* no se considera una emoción reconocida, la *ansiedad* y el *nerviosismo están* estrechamente ligados a él [3]. Se describe como una condición de *tensión provocada* por una situación desafiante o desfavorable. Esto puede ocurrir tanto por variables internas como las externas, como la carga de trabajo, los ruidos, las vibraciones, la falta de sueño, la fatiga, etc., todas pueden provocar *estrés*. Este trabajo bibliográfico [149], ofrece una visión muy completa del sistema de detección del estrés, incluyendo el papel del aprendizaje automático en los sistemas de detección de emociones, los métodos de selección de características, las diferentes medidas de evaluación, las tareas y las aplicaciones. Además estudia la conexión entre la naturaleza biológica de las emociones de las personas y el *estrés* mental.

El estrés tiene varias consecuencias fisiológicas, como cambios respiratorios (respiración más rápida), aumento de la frecuencia cardiaca, más sudoración (transpiración de la piel), incluso un aumento de la tensión muscular que también se refleja en las cuerdas vocales y el tracto vocal, lo que afecta a la producción del habla. Todos estos factores pueden, directa o indirectamente, afectar negativamente a la calidad del habla [150] y nos ayudan a discriminar entre habla estresada o neutra cuando utilizamos algoritmos de aprendizaje automático [3].

3.2. Bases de Datos de Habla Compatibles y Disponibles en la Literatura

Si nos remitimos a la categorización de los datos del habla en la Fig. 3.1, las categorías en negrita son idealmente las que necesitamos para nuestros datos del habla con *miedo* y nuestra aplicación, pero debido a la falta de disponibilidad de este tipo de bases de datos de manera abierta y accesible en la literatura, intentamos buscar otras más adecuadas para nuestros objetivos, y priorizamos al máximo el uso del habla real sobre la actuada. La alternativa más cercana que existe al *miedo real* son los conjuntos de datos grabados por actores [151], [152].

Entre ellas se incluyen bases de datos con fragmentos de películas, como la base de datos SAFE [153], centrada en la emoción del *miedo*. A su vez, la grabación de *estrés* real y espontáneo es difícil de encontrar en la literatura, ya que hay muy pocos conjuntos de datos en los que el habla estresada sea simulada o grabada en condiciones reales. Algunos ejemplos son la base de datos SUSAS [154], una recopilación de datos del habla para el análisis del reconocimiento del habla y el diseño de algoritmos robustos al *ruido* y al *estrés*; o UT-Scope [155], que proporciona una estimación automática del habla *lombarda* a partir del reconocimiento del hablante -el efecto lombardo es la tendencia inconsciente de los hablantes a aumentar el volumen de su voz para mejorar la inteligibilidad cuando hablan en un entorno ruidoso-; o el Corpus VOCE [156] -una base de datos en condiciones neutras y de estrés en el habla realista, leída y espontánea-. Otra base de datos de estrés leído que utilizamos es Biospeech [157], a la que sus autores nos dieron acceso. Algunos de los trabajos re para el *estrés* o *el miedo* real utilizan bases de datos propias que no se han hecho públicas. Para el *miedo* encontramos algunos como [158], donde presentan grabaciones de datos de habla en servicios de emergencias (situaciones reales de urgencia y *miedo*) de un centro de llamadas de emergencia; o [159], donde se graba el habla de usuarios con agorafobia inducida por el *miedo*. En cuanto al *estrés*, algunas bases de datos realistas del estado de la técnica no están del todo disponibles para su uso, como [160], que incluye grabaciones de voz en ruso (palabras, frases y oraciones) grabadas por testigos de en sucesos adversos que experimentan estrés; o [161, 142], donde utilizan entornos virtuales para inducir *estrés* en los participantes; o tres corpus alemanes -el FAU, el Ulm- y el Reg-TSST- que se recopilaron siguiendo el conocido protocolo de la Prueba de Estrés Social de Trier (TSST) [163]. Aunque es positivo que haya trabajos realizados y descritos en este campo, no podemos beneficiarnos plenamente de ellos, ya que no están disponibles abiertamente para la investigación.

3.2.1. Corpus VOCE

Dado que Bindi se utilizará en la vida real para detectar situaciones de peligro es necesario 1. trabajar con bases de datos que contengan habla en condiciones reales y 2. que éstas incluyan

sentimientos reales de *miedo*, *pánico* o *ansiedad*, que podrían evocarse en el tipo de situaciones que se detectarán en el caso de uso. En los trabajos [2], [3] y [10] se utilizó la base de datos VOCE.

La primera condición es relativamente fácil de obtener en la bibliografía, pero no la segunda. Por ello, optamos por seleccionar un conjunto de datos generados en condiciones reales pero que estudiaran una sensación relativamente cercana al *miedo*, como es el *estrés*. En concreto, seleccionamos el Corpus VOCE [156] por tres razones principales, 1) incluye datos capturados en condiciones reales de *estrés,* 2) algunos sensores utilizados durante la fase de captura son similares a los presentes en la pulsera de Bindi para obtener mediciones de la frecuencia cardiaca, y 3) debido a la existencia de estudios previos [164] que confirman la viabilidad de relacionar las métricas de la frecuencia cardiaca con el *estrés* en el habla.

VOCE [156] comprende grabaciones de 45 hablantes en condiciones neutras y de *estrés* en el habla, siendo: realista, leída y espontánea [3]. La última versión actualizada de este conjunto de datos incluye un total de 135 grabaciones de voz procedentes de un conjunto de 45 estudiantes (21 hombres, 17 mujeres y 7 sin identificar) de la Universidad de Oporto, con edades comprendidas entre los 19 y los 49 años. Para cada usuario, el habla se grabó en tres escenarios diferentes: *pre-baseline*, *baseline* y *recording*, que se adquirieron: mientras la persona voluntaria está leyendo un artículo 24 horas antes de hablar en público, mientras la persona lee el mismo artículo sólo 30 minutos antes de hablar en público, y en un escenario en el que la persona se encuentra hablando en público y en condiciones de estrés; respectivamente. La frecuencia cardiaca (HR) también se adquirió cada segundo para las tres grabaciones.

Junto con estos archivos de audio, se proporcionan 117 archivos que contienen 2 variables fisiológicas medidas y utilizadas para estimar la frecuencia cardiaca (FC). Estas mediciones, realizadas con un dispositivo Zephyr HxM BT2, son 1. (i) Z_{ecg} que representa un valor de HR promediado y filtrado con un periodo de muestreo de 1s; y 2. (ii) Z_{ts} valores que se refieren a los instantes de tiempo en los que se producen picos de *R* en el electrocardiograma obtenido con el dispositivo, medidos con un reloj interno de 16 bits. Cada uno de estos valores va acompañado del instante de tiempo universal coordinado (UTC) correspondiente. Además, la base de datos contiene un archivo de metadatos que incluye sexo, edad, información sanitaria, experiencia en hablar en público, puntuaciones de la prueba STAI (*State-Trait Anxiety Inventory*) [165] y otra información sobre la calidad de las grabaciones de audio (nivel de energía, saturación...). Desgraciadamente, esto sólo se proporciona para 38 de las 45 personas voluntarias de la base de datos y ésta sólo recoge información completa (los 3 archivos de audio y sus correspondientes valores de HR) de 21 voluntarios y voluntarias.

Dividimos a estos 21 hablantes en dos conjuntos, el Set 1 estaba compuesto por 10 oradores cuyo HR era coherente con las grabaciones -en el sentido de que, cuando un orador estaba leyendo, la frecuencia cardiaca permanecía estable, pero en el escenario de hablar en público el HR aumentaba-. El Set 2 estaba formado por los otros 11 hablantes restantes. En la tabla 3.1 se especifica el número de muestras por Set, cada muestra representa una señal de audio de *1* s.

Muestras	**Neutro**	**Estrés**	**Total**
Set 1	1.389	3.989	5.378
Set 2	1.716	4.858	6.574
Total	3.105	8.847	11.952

TABLA 3.1: Número de muestras de habla de la base de datos preprocesada VOCE Corpus [10].

Preprocesamiento de datos

Para su uso en esta tesis, procesamos tanto los datos del habla como las señales de HR. Para simplificar, comenzamos con una conversión de estéreo a mono de las grabaciones de audio, seguida de un *downsampling* de la frecuencia de muestreo de 44,1 kHz a 16 kHz para reducir el coste computacional sin perder demasiada información. Después continuamos realizando una normalización z-score (las señales tienen media 0 y desviación estándar 1), estandarizada en la literatura. Por último, las señales pasan por un detector de actividad vocal (VAD) [166] que elimina las muestras de audio en silencio, ya que no incluyen información sobre las características de la voz, que es la información que querríamos usar para nuestra tarea. El algoritmo VAD específico elegido está diseñado para mejorar la robustez de la detección del habla en entornos ruidosos, eliminando partes de un segundo de duración de audio sin habla en las que no se puede tomar ninguna decisión sobre si hay *estrés* o quién es el hablante. En cuanto a las medidas HR recogidas en la base de datos, los valores Z_{ecg} originales con signo se convirtieron en otros sin signo de 0 a 255. Las secuencias Z_{ts} se descartaron por considerarse demasiado ruidosas y Z_{ecg} ya proporcionaba la información de HR necesaria con una resolución temporal razonable.

Etiquetado

Etiquetar una señal de audio para determinar la presencia de *estrés* es un asunto delicado, ya que no existe una forma prescrita de hacerlo dado que *el estrés* no es binario y es muy subjetivo. Adoptando una perspectiva pragmática, una vez más nos basamos en un trabajo anterior [164] en el que las grabaciones de este corpus se etiquetaron en función de la frecuencia cardiaca de cada hablante. En lugar de las etiquetas incluidas en el corpus VOCE original para cada situación de grabación (0 para las secuencias completas *pre-baseline* o *baseline* y 1 para la *recording*) generamos las etiquetas a partir de las secuencias de HR. Cada audio de *1s* se etiqueta como estresada o neutra utilizando un umbral de HR dependiente del hablante establecido para cada uno de los hablantes utilizando sus respectivas grabaciones *pre-baseline*. Se compararon dos

umbrales de HR diferentes: la media de la HR *pre-baseline* más la desviación estándar y el percentil del 75% del valor de la FC, y finalmente se descartó el primero.

Equilibrio y aumento de datos

El hecho de que las muestras de datos no estuvieran equilibradas -es decir, hay hablantes con un número de muestras significativamente mayor que otros, y la cantidad de muestras de habla estresadas y de habla neutra tampoco son idénticas- nos llevó a realizar un ajuste para cada conjunto (Set) y condición con el fin de obtener buenas estimaciones con nuestros modelos. Entonces, todas las clases -en este caso, los hablantes- deben considerarse igual de importantes desde el punto de vista de un clasificador de reconocimiento de hablantes para optimizar su entrenamiento. Sin embargo, el uso de una técnica de sobremuestreo (*oversampling*) puramente estadística tendría un gran inconveniente en nuestro caso, ya que el desequilibrio es muy grande y la cantidad de datos artificiales creados sería demasiado grande. Para hacer frente a este problema, primero submuestreamos (*undersampling*) el conjunto de datos neutros admitiendo un máximo de 120 muestras por hablante en ambos conjuntos (1 y 2), así como el conjunto estresado, utilizando un umbral de 300 muestras. La aplicación de una técnica de sobremuestreo (en concreto, SMOTE [167]) a los datos submuestreados dio como resultado un número suficiente de muestras nuevas, consiguiendo un conjunto de datos equilibrado pero sin incluir una cantidad desproporcionada de datos artificiales.

Además, experimentamos aplicando modificaciones en las señales de habla a la velocidad de locución y el tono de la base de datos original, para producir muestras de habla estresada generadas sintéticamente, y medir su efecto con clasificadores ML. Este proceso se detalla en la sección 4.3.1.

3.2.2. BioSpeech

Biospeech (BioS-DB) [157] es una base de datos multimodal de habla pública que incluye anotaciones emocionales de manera continua en el tiempo. Consta de 55 hablantes que leen dos textos, uno en alemán y otro en inglés, mientras se registran sus variables fisiológicas -el pulso del volumen sanguíneo (BVP), la conductancia de la piel (SKT)- y el habla [8].

Esta base de datos responde a la idea de que la *performance anxiety (*en español sería el *miedo* escénico*)* puede producirse al hablar en voz alta delante de un público, y puede reflejarse en las variables fisiológicas y en el habla. Tres anotadores con formación previa utilizan un *joystick* para etiquetar la emoción que presenta el hablante de manera continua en el tiempo en un espacio 2D, cuyos ejes representan el *arousal* y la valencia descritos en la sección 2.3.2. Biospeech se utilizó en [8].

El objetivo de utilizar estos datos para la tesis es doble, 1. detectar *el estrés* en el habla y 2. reconocer al hablante incluso cuando el habla está bajo condiciones de *estrés*. Para ello, realizamos una clasificación con los datos, en lugar de una regresión. Tanto las tareas de regresión como las de clasificación realizan predicciones sobre los datos, pero la diferencia reside en que la regresión pretende predecir valores continuos, y la clasificación predice valores discretos entre un número limitado de clases.

Para crear unas etiquetas *ground truth* (en español sería etiquetas de 'verdad absoluta', verdaderas u originales) para las etiquetas emocionales a partir de las tres anotaciones individuales en tiempo continuo, los autores de BioS-DB utilizaron la métrica de estimación ponderada por evaluador (EWE) [168]. La EWE es fiable cuando el número de anotadores es bastante grande, pero en este caso sólo contamos con 3 evaluadores, lo que hace que la posibilidad de disparidad en las calificaciones sea muy alta.

Los antecedentes de cada anotador afectan a sus valoraciones, además del sesgo de las posibles comparaciones entre hablantes consecutivos. Estos factores pueden inducir variabilidad y discrepancias en las calificaciones, y una combinación ponderada de las etiquetas de cada anotador puede no ser el método de fusión óptimo. Pensamos que esto podría ser perjudicial para nuestros fines de clasificación, que son diferentes de los de los creadores del conjunto de datos, que fue la regresión.

Reinterpretación de etiquetas para un enfoque de clasificación

Así, en el marco de esta tesis, proponemos un reetiquetado de los valores BioS-DB de *arousal* y valencia cuantificándolos en 4 cuadrantes categóricos [169]. Esto es crucial para definir una tarea de clasificación en lugar de una de regresión. Estos cuatro cuadrantes son

- Alta valencia, alta excitación (HVHA): Q1
- Baja valencia, alta excitación (LVHA): Q2
- Baja valencia, baja excitación (LVLA): Q3
- Alta valencia, baja excitación (HVLA): Q4

También creemos que, aunque BioS-DB tiene una resolución temporal muy precisa en el etiquetado, una resolución temporal más gruesa para captar las emociones subyacentes en el habla es más adecuada en tareas de clasificación como la nuestra. En concreto, las anotaciones en bruto en BioS-DB de cada anotador se muestrearon originalmente a *2* Hz y su rango era [-1000, 1000]. Por lo tanto, para nuestros propósitos, reducimos el muestreo de las señales a *1Hz* para obtener una etiqueta por segundo, que será nuestra frecuencia de trabajo de referencia para futuros esquemas de fusión de datos. Para calcular una etiqueta final combinada para cada segundo, elegimos los dos anotadores que habían etiquetado más cerca en el espacio 2D, y basándonos en

el signo de los valores de *arousal (*excitación*)* y valencia, los convertimos en una etiqueta categórica en cada uno de los cuatro cuadrantes. Si el cuadrante de las dos etiquetas coincide, se elige como etiqueta agregada, en caso contrario, asignamos un valor indeterminado provisional, *x*.

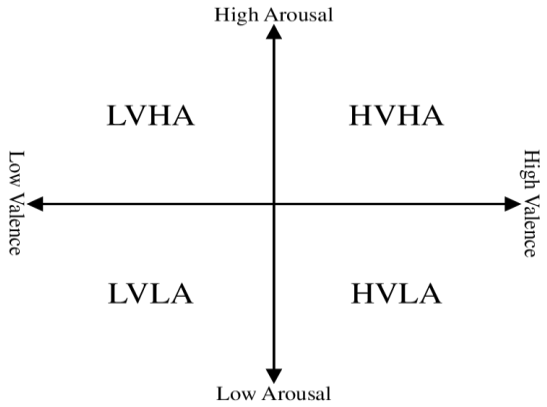


Figura 3. 2: Cuatro cuadrantes del espacio valencia-arousal [169]. Reproducido con permiso del propietario del copyright © 2012 IEEE.

A continuación, analizamos varios casos para las etiquetas indeterminadas, como muestra la Fig. 3.3. Si *x* se debe a una transición entre cuadrantes (un anotador ha cruzado el límite pero el otro aún no), elegimos al azar cualquiera de los dos cuadrantes. En caso contrario, consideramos la etiqueta si dos anotadores caen en el mismo cuadrante aunque no sean los más cercanos en el espacio 2D. Si es así, la etiqueta agregada es la correspondiente a ese cuadrante. Este proceso resuelve una gran cantidad de etiquetas indeterminadas. Para el resto y para los casos en los que encontramos varias indeterminaciones seguidas, utilizamos una ventana de 5 segundos y sustituimos las etiquetas desconocidas mediante *majority voting* (votación por mayoría). Nuestro proceso tiene en cuenta la proximidad de las etiquetas de los anotadores, lo que proporciona confianza sobre la etiqueta resultante, ya que los anotadores interpretan el espacio 2D en términos del significado de los cuadrantes.

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

Figura 3. 3: Procedimiento propuesto para determinar la nueva etiqueta de cuadrante combinado para Biospeech.

Las transiciones entre cuadrantes se consideran cuidadosamente, ya que las personas no saltan de un estado emocional a otro de forma repentina. La ventana de suavizado proporciona una señal de etiqueta suave al evitar los cambios bruscos entre cuadrantes. Por último, para nuestra tarea de detección automática de situaciones de violencia de género, se elegirá como objetivo el segundo cuadrante *Q2*, donde se encuentran las emociones relacionadas con *el estrés*, *la ansiedad* y el *miedo*.

Así, consideramos dos tipos de etiquetado para nuestras tareas: 4 cuadrantes y binario (considerando *Q1*, *Q3*, *Q4* como la etiqueta negativa o 0, y *Q2* como la positiva, o 1, o de interés), y el resultado del reetiquetado puede observarse en la Tabla 3.2.

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

TABLA 3.2: Porcentaje (%) de etiquetas en cada cuadrante PAD para el reetiquetado de Biospeech [8]. Reproducido con permiso del propietario del copyright, ISCA.

3.2.3. BioSpeech+

Como se ha indicado en secciones anteriores, nuestro objetivo final es desarrollar una herramienta autónoma para detectar situaciones de riesgo de violencia de género. En lo que respecta al habla y al audio, pretendemos rastrear e identificar la voz de la usuaria y utilizarla para detectar el *miedo* o el *pánico* -o su pariente cercano, *el* estrés-. Para mejorar la precisión del sistema, pretendemos contextualizarlo -en línea con la comprensión e interpretación de la

situación 2.4 - mediante el análisis de la escena acústica (sonidos y ruidos de fondo) utilizando un sistema de detección y clasificación de eventos acústicos (AED/C).

BioS-DB se utiliza como aproximación a este problema. Sin embargo, para nuestros fines específicos es clave complementar la información hablada con conocimientos sobre los eventos presentes en la escena acústica: en muchos casos, *el pánico* puede hacer que una víctima de VG permanezca en silencio. Por ello, los sonidos ambientales, es decir, la caracterización de la escena acústica, pueden proporcionar información útil para el sistema de detección. Junto con otros miembros del equipo UC3M4Safety, introdujimos un procedimiento preliminar para ampliar BioSpeech y convertirlo en Biospeech+, compuesto por los archivos de voz originales enriquecidos sintéticamente con sonidos ambientales [8]. En él, hacemos uso de AudioSet [170], una colección a gran escala de clips de sonido de 10 segundos etiquetados por humanos y capturados de YouTube. Audioset proporciona 2,084,320 muestras que contienen 527 etiquetas a nivel de clip de eventos acústicos. Hemos seleccionado un subconjunto de 2,108 muestras de Audioset, pertenecientes a 83 clases, para ampliar la BioS-DB original. Para elegir las clases relacionadas con eventos que inducen *miedo*, seleccionamos eventos violentos y empleamos la colección de estímulos audiovisuales [126], [11.1], seleccionada para el desarrollo del conjunto de datos WEMAC [11]. La selección inicial fue realizada por expertos en VG y posteriormente validada por más de 1300 voluntarios [126].

En la fase de preprocesamiento, la señal de audio se normaliza y se convierte a 16 *kHz y* canal mono. A continuación, se calcula un espectrograma *log-mel* de 64 *bins* para extraer una representación tiempo-frecuencia de la señal de audio como una imagen.

En cuanto a la mezcla sintética, el proceso se basa en el aumento de datos seguido en la tarea 4 del desafío Detección y clasificación de escenas y eventos acústicos (DCASE) 2019 [171]. El algoritmo *scaper* [172] nos permite definir distribuciones de probabilidad para la aparición y duración de los eventos sonoros. Así, el sistema genera tantas mezclas sintéticas como se desee a partir de audio previamente clasificado como de primer plano o de fondo. En nuestro caso particular, los eventos en primer plano son las muestras originales de BioS-DB y los eventos de fondo son las muestras del subconjunto Audioset.

```
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
```

Figura 3. 4: Procedimiento de generación de Biospeech+, mezclado de muestras de BioSpeech y Audioset con Scaper [8]. Reproducido con permiso del propietario del copyright, ISCA.

El número de mezclas generadas se ha fijado en 110: generamos una mezcla por cada archivo BioS-DB, considerando las grabaciones captadas por el micrófono de solapa, es decir, 55 grabaciones de audio en alemán y 55 en inglés.

El algoritmo que detalla el procedimiento de mezcla, teniendo en cuenta las nuevas etiquetas binarizadas explicadas en la Sec. 3.2.2, se presenta en formato de pseudocódigo en la Fig. 3.4. El fundamento de esta metodología para el aumento del conjunto de datos es proporcionar una relación no determinista entre los sonidos estresantes o potencialmente aterradores y la aparición de *estrés* en el hablante. Además de gestionar las distribuciones de probabilidad y la temporización de los eventos, Scaper permite realizar operaciones de desplazamiento del tono y de ralentización en el tiempo sobre las muestras en primer plano, y ambas podrían utilizarse para aumentar aún más el conjunto de datos.

3.3. WEMAC: Conjunto de Datos de Computación Afectiva Multimodal de Mujeres y Emociones

Hasta ahora hemos hablado de la falta de bases de datos etiquetadas adecuadas para nuestro propósito en la literatura -hasta el momento del trabajo aquí presentado- sobre el habla real en condiciones de *miedo*. Parecía claro que el siguiente paso era contribuir con la recopilación de una base de datos que sirviera exactamente a nuestro objetivo de detectar situaciones de riesgo a través de la voz. Como dijimos en el apartado 1.1.4, esta tesis forma parte del proyecto EMPATIA-CM que pretende desarrollar un dispositivo multimodal vestible para la detección automática y discreta de estas situaciones, por lo que las bases de datos recopiladas y explicadas a continuación también son multimodales.

WEMAC es un conjunto de datos multimodal que consiste en experimentos de laboratorio con voluntarias expuestas a estímulos audiovisuales validados para evocar emociones reales mediante un *headset* (casco y gafas) de realidad virtual, capturando y recopilando variables fisiológicas, el habla y etiquetas emocionales. Para su recopilación hemos utilizado la colección

de estímulos audiovisuales validados, que forma parte del conjunto de datos de estímulos audiovisuales UC3M4Safety, pero su creación se debe a los demás miembros del equipo UC3M4Safety. Surge de la necesidad de elicitar emociones realistas, especialmente *el miedo*, que es clave para la detección de situaciones de riesgo de violencia de género. Creemos que esta base de datos servirá y ayudará a la investigación sobre Computación Afectiva multimodal utilizando información fisiológica y del habla, y que será especialmente eficaz para la tarea de detección de situaciones de riesgo de violencia de género.

3.3.1. Base de datos UC3M4Safety Audiovisual Stimuli

Otros miembros del equipo UC3M4Safety realizaron el estudio [126] para obtener un conjunto de alta calidad de estímulos audiovisuales para provocar emociones en escenarios controlados. Este conjunto de estímulos está diseñado para recopilar señales y respuestas adicionales (variables fisiológicas y habla) que puedan ser utilizadas por sistemas de AI ML/DL destinados a la identificación automática y en tiempo real de emociones. Aunque el objetivo principal es reconocer el *miedo* o el *pánico*, utilizaremos videoclips cuidadosamente seleccionados y un completo sistema de etiquetado de 12 emociones categóricas.

El trabajo presenta la identificación de las emociones elicitadas tras la visualización de los estímulos audiovisuales. Además, los autores realizaron un estudio estadístico de las diferencias de género en las respuestas emocionales de 1.332 personas voluntarias (811 mujeres y 521 hombres). El estudio de investigación produjo un conjunto de datos de 42 estímulos audiovisuales -denominado Base de Datos de Estímulos Audiovisuales UC3M4Safety [11.1] [126]- que desencadenan una gama de 12 emociones en los espectadores. Cada estímulo tiene un alto nivel de acuerdo y una categorización emocional discreta, así como una categorización emocional continua en el Espacio Afectivo Placer-Arousal-Dominancia (PAD).

La selección de la serie de estímulos audiovisuales se realizó en cinco pasos, como se muestra en la Fig. 3.5. Cada recuadro de color azul refleja el proceso paso a paso y los criterios utilizados para la selección de los clips, mientras que los recuadros blancos denotan a los supervisores que participaron en el proceso.

Inicialmente, cinco investigadores recogieron muestras de contenido emocional de películas comerciales, series de televisión, documentales, cortometrajes, anuncios y vídeos de Internet. Estos clips fueron etiquetados originalmente con una emoción objetivo por otros miembros del equipo UC3M4Safety, con el asesoramiento de un panel de expertos. Las emociones discretas contenidas en los estímulos audiovisuales buscados por los investigadores fueron *alegría, tristeza, sorpresa, desprecio, esperanza, miedo, atracción, asco, ternura, ira, calma* y *tedio*.

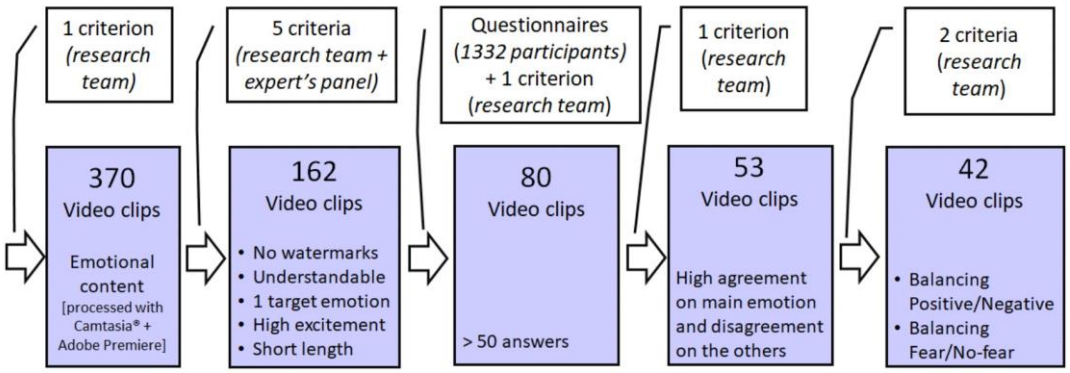


Figura 3. 5: Procesamiento de videoclips en la creación de la base de datos de estímulos audiovisuales de la UC3M. Reproducido con permiso del propietario del copyright, los autores de [126] a través de la licencia Creative Commons CC-BY 4.0 de MDPI.

En segundo lugar, de los 370 videos obtenidos en la Etapa 1, se seleccionaron 162 *clips* para su posterior evaluación basándose en criterios de selección (véase la Etapa 2, Fig. 3.5). Se tuvo en cuenta un criterio adicional para las películas sobre violencia de género: la protagonista de las películas debe ser una mujer víctima de algún tipo de violencia (sexual, física, psicológica, etc.). En tercer lugar, la lista de 162 estímulos fue etiquetada con categorías discretas de emoción en un entorno de *crowdsourcing* por un amplio conjunto de voluntarios. Cada clip se etiqueta con la emoción experimentada tras su visualización. Las emociones notificadas por las voluntarias no siempre coincidían con las que el equipo y los expertos esperaban que suscitara el clip en un principio. Sólo 80 videoclips obtuvieron suficientes respuestas para ser considerados para un análisis posterior (cada uno fue visualizado por más de 50 personas voluntarias que lo etiquetaron con etiquetas de emociones).

Para la selección final de los estímulos audiovisuales debían cumplirse dos condiciones. La primera condición buscaba el mayor porcentaje de consentimiento entre los participantes, es decir, al menos el 50% de las voluntarias considerando ambos géneros o al menos el 50% de un género individualmente, que visualizaban cada estímulo, debían etiquetarlo con la misma emoción categórica. En la segunda condición también se comprobó la exclusividad de esta etiqueta comprobando que todas las demás emociones posibles coincidían como máximo el 30% de las veces. Por último, se eliminaron algunos vídeos para equilibrar la distribución entre las emociones objetivo consideradas como *miedo* y *no miedo*, produciendo una selección de 42 clips. Se obtuvo un porcentaje del 44,44% de estímulos para la elicitación del *miedo* y del 55,55% para el resto de emociones, como figuran en la tabla 3.3.

Miedo	44.44%	**Tedio**	2.22%
Felicidad	8.89%	**Ternura**	6.67%
Esperanza	2.22%	**Calma**	11.11%
Sorpresa	4.44%	**Asco**	8.89%
Ira	4.44%	**Tristeza**	6.67%

TABLA 3.3: Porcentajes de emociones categóricas suscitadas por el conjunto de estímulos audiovisuales de UC3M4Safety Audiovisual Stimuli para la muestra final de 42 clips [126].

Diferencias de género en las anotaciones emocionales

Los resultados obtenidos por el equipo de [126] muestran unas emociones positivas notificadas similares en valores discretos (y también en el espacio PAD) para ambos géneros, mientras que las emociones negativas (especialmente el *miedo* y el *desprecio*) notificadas para ambos géneros son más diferentes. La memoria autobiográfica puede influir en la percepción del *miedo* en aquellos videoclips relacionados con la violencia de género. La violencia de género es difícil de etiquetar, y el hecho de que la persona espectadora se identifique con los protagonistas del videoclip puede estar teniendo un gran impacto en su estado emocional. En los clips en los que las mujeres etiquetan mayoritariamente el *miedo,* los hombres etiquetan la *ira* y la *tristeza*.

Etiquetar correctamente el *miedo* de las mujeres beneficiará nuestro objetivo principal de desarrollar un sistema automático para protegerlas de las agresiones violentas. Teniendo en cuenta los resultados observados, la variable de género debería tenerse en cuenta tanto en la fase de selección de estímulos de la base de datos como en la fase de entrenamiento de los algoritmos de aprendizaje automático. Aunque en el estudio no se observaron diferencias significativas entre unas emociones y otras (especialmente *el miedo* y *la esperanza*), deberían tenerse en cuenta las diferencias de género en las emociones declaradas para mejorar la clasificación de las emociones. Esto es especialmente importante en este trabajo, porque el objetivo principal es identificar las condiciones que causan *miedo* a las mujeres, incluyendo (y especialmente) a las víctimas de la violencia de género. En este caso, el género debe tenerse en cuenta porque la emoción percibida al ver un vídeo de *miedo* difiere según el género. Incluso es clave para el caso particular de los estímulos que reproducen situaciones de violencia de género, donde existe una gran diferencia en el etiquetado entre mujeres y hombres (*miedo* frente a *ira* y *tristeza*).

Estos resultados corroboran los de otros estudios [173], en los que los autores concluyen que las mujeres declaran sentir más *miedo* que los hombres; y que las emociones de *tristeza*, *compasión* y *miedo* las sienten más las mujeres que los hombres, lo que podría deberse a que los rasgos empáticos más fuertes y el *caretaker syndrome* (síndrome del cuidador) se dan sobre todo en las mujeres.

3.3.2. Colección de Bases de Datos WEMAC

En la base de datos denominada propiamente como WEMAC [11], nosotros -el equipo UC3M4Safety- utilizamos un entorno de realidad virtual -un casco con gafas de realidad virtual y un joystick- para presentar a mujeres voluntarias los estímulos audiovisuales de manera inmersiva (es decir, videoclips) con el fin de provocar, etiquetar y medir reacciones emocionales reales ante ellos.

Las participantes son mujeres voluntarias, incluidas mujeres que han sufrido violencia de género -por tanto, voluntarias con y sin violencia de género-. Están divididas en grupos de edad equilibrados definidos por intervalos de 10 años: G1 (18 - 24), G2 (25 - 34), G3 (35 - 44), G4 (45 - 54) y G5 (55 en adelante). La base de datos está formada por 104 mujeres voluntarias que nunca sufrieron violencia de género (47 en la primera versión y 57, en la segunda) y 43 mujeres voluntarias víctimas de violencia de género. Este último grupo realizó el experimento bajo la supervisión de una psicóloga. La Fig. 3.6 muestra un diagrama simplificado de la metodología específica seguida durante la experimentación para cada voluntaria y estímulo.

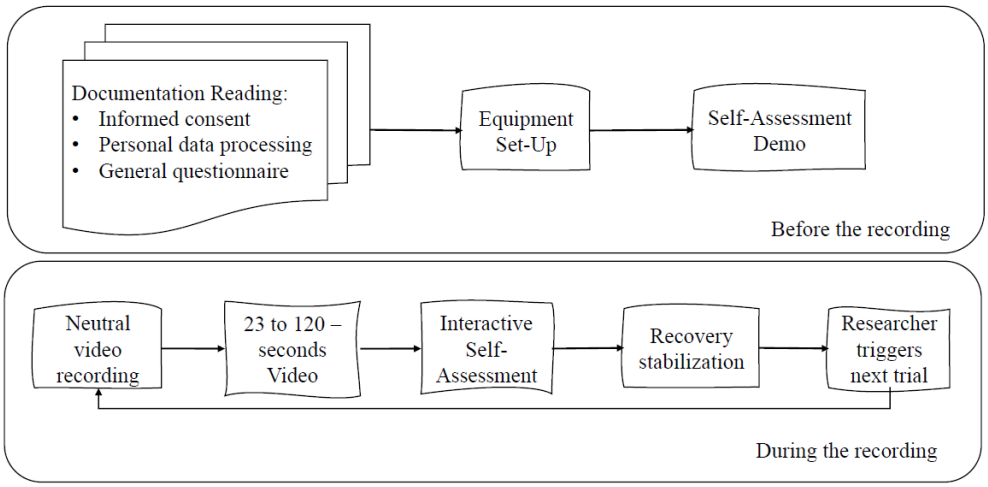


Figura 3. 6: Metodología para la captura del dataset WEMAC, antes y durante las visualizaciones [11].

Las voluntarias fueron reclutadas a través de anuncios en las redes sociales y de los canales de comunicación interna de la Universidad Carlos III. Previamente al experimento, se explica a las voluntarias reclutadas las diferentes fases a seguir, incluyendo un conjunto de documentos que deben rellenar: un consentimiento informado y un cuestionario genérico inicial. El primero es necesario para el tratamiento de los datos personales y la normativa de protección de datos. El segundo recoge información como rasgos de personalidad, sexo, edad, si realizaron alguna actividad física reciente o si estaban tomando medicación -el uso de medicamentos podría modificar la respuestas fisiológicas de cada usuaria-, sobrecarga emocional debido a situaciones laborales, económicas y personales, sesgos del estado de ánimo (miedos, fobias, experiencias traumáticas), entre otros. Esta información podría ser relevante e informativa de las reacciones

emocionales de las usuarias captados en el experimento, afectando a su percepción, evaluación y atención.

En esta recopilación de datos, el equipo UC3M4Safety siguió la metodología presentada también en [174], que es un estudio para la detección del miedo utilizando la concentración de la hormona catecolamina en sangre. El último paso en la fase de preparación del experimento es una demostración introductoria en la que las voluntarias se acostumbran al entorno de realidad virtual -casco, gafas y joystick- y se familiarizan con las particularidades del etiquetado. Este entorno se utiliza para presentar los clips a las usuarias, y también para anotar los clips según diferentes categorías a través de pantallas interactivas.

Todo el proceso de lectura de la documentación, montaje de los equipos, demostración del entorno virtual, junto con la visualización y etiquetado de los vídeos, suele durar entre 60 y 100 minutos por participante.

Visualización de estímulos audiovisuales

Utilizamos un casco de realidad virtual Oculus Rift-S [21] para presentar los estímulos audiovisuales. La realidad virtual se utiliza para maximizar la experiencia inmersiva y, en consecuencia, lograr una mejor elicitación de la emoción. Durante el experimento de grabación, cada voluntaria visualiza un total de 14 estímulos audiovisuales emocionales, algunos de los cuales presentan una experiencia de 360º. Estos estímulos se seleccionaron de un conjunto de 28 estímulos audiovisuales, lo que dio lugar a dos *batches* (grupos) de 14 vídeos cada uno, de los 42 vídeos finales de los estímulos audiovisuales UC3M4Safety [126], como se ve en la Fig. 3.7. Los criterios aplicados para la selección fueron los siguientes 1) un alto porcentaje de acuerdo en la etiqueta emocional discreta observada en las anotadoras mujeres durante el experimento de preetiquetado [126], 2) una duración adecuada del experimento de laboratorio y 3) una distribución equilibrada de clips de *miedo/no-miedo* en cada *batch* [53].

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

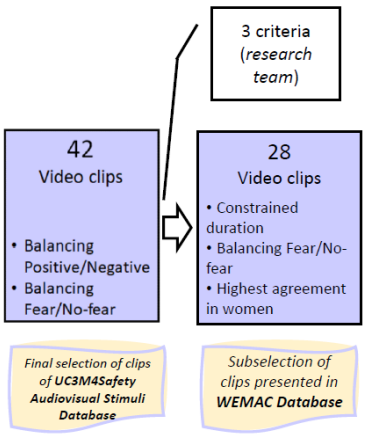


Figura 3. 7: Esquema de la subselección de clips del conjunto de datos UC3M4Safety Audiovisual Stimuli utilizados en la base de datos WEMAC.

La duración media de cada estímulo audiovisual es de 100 segundos. Ambos *batches* tienen 8 estímulos cuya etiqueta pertenece al segundo cuadrante de *arousal*-valencia de los 4 cuadrantes de emociones categóricas para mantener un equilibrio adecuado entre las emociones de *miedo* y las que no lo son. Nótese que la premisa de equilibrio se evalúa considerando el modelo valencia-*arousal* -o placer-*arousal*, PA-, en lugar del espacio placer-*arousal*-dominancia (PAD), por simplicidad. Debido a este hecho, los estímulos pre-etiquetados como *ira* o *miedo* se consideran dentro del segundo cuadrante, estando entonces dentro de la clase positiva (o de interés) para el etiquetado binario que consideramos.

Antes de la presentación de cada uno de los estímulos, se muestra un videoclip neutro para situar a la participante en un estado emocional neutro. Estos videoclips neutros se han seleccionado del amplio grupo de videos proporcionado por el Laboratorio de Psicofisiología de Stanford [175]. Del mismo modo, también se muestran a las voluntarias escenas estáticas 3D de paisajes para inducir a la recuperación del estado emocional, después del proceso de etiquetado interactivo de emociones. Estas escenas 3D se seleccionaron por consenso unánime del equipo de investigación. La principal diferencia entre los clips neutros y los de recuperación es que mientras que durante la visualización de los primeros no se realiza ninguna acción -es decir, no hay seguimiento de la recuperación-, en el caso de los segundos hay un seguimiento fisiológico a través de la pulsera Bindi para garantizar la estabilización fisiológica de la voluntaria.

Señales fisiológicas grabadas

Durante la presentación de los estímulos audiovisuales, se capturan las señales fisiológicas de las participantes[22] . El equipo utilizado para este fin incluye los siguientes dispositivos y sensores:

[22] Procesamiento disponible en: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/

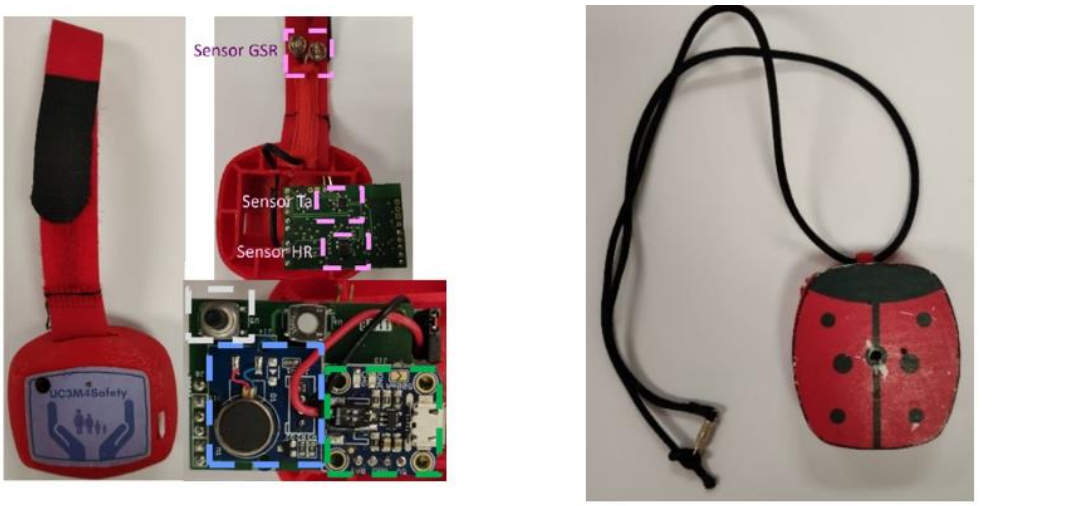
- El sistema de herramientas de investigación BioSignalPlux [23] , que se utiliza habitualmente para adquirir diferentes señales fisiológicas, en particular: el pulso del volumen sanguíneo del dedo (BVP), la respuesta galvánica de la piel de la muñeca ventral (GSR), la temperatura de la piel del antebrazo (SKT), la electromiografía trapezoidal (EMG), la respiración torácica y el movimiento inercial de la muñeca a través de un acelerómetro.
- La pulsera de Bindi, representado en la Fig. 3.8a, que mide el BVP dorsal de la muñeca, el GSR ventral de la muñeca y el SKT. Las particularidades de hardware y software de este elemento se detallan en publicaciones anteriores del equipo [176], [177], [178].
- Un sensor GSR adicional que se integrará en la próxima versión de la pulsera Bindi. Sus particularidades de hardware y software se detallan en [179].

El conjunto de herramientas BioSignalPlux se emplea para proporcionar medidas *ground truth* (verdaderas y fiables, con las que contrastar), con que se compararán con los sensores incluidos en la pulsera de Bindi. De hecho, las señales BVP y GSR obtenidas de BioSignalPlux y Bindi se compararon y correlacionaron con éxito en [176] y [178]. La sincronización de la adquisición de todos los sensores junto con las etapas del experimento se ejecuta en un ordenador portátil a través de un programa basado en el framework Unity[24] . En este sentido, la frecuencia de muestreo de los dispositivos que captan la información fisiológica es de 200 Hz.

Proceso de etiquetado: Señales verbales y auto-anotaciones emocionales

Tras la visualización de cada videoclip emocional, las voluntarias ven un conjunto de pantallas interactivas dentro del entorno de realidad virtual, desarrollado con el software Unity [180].

[23] https://biosignalsplux.com/products/kits/researcher.html https://unity.com/es[24]



(A) Bindi 1.0 brazalete para la captura de señales fisiológicas

(B) Colgante Bindi 1.0 para la captura de señales auditivas

Figura 3. 8: Dispositivos portátiles Bindi 1.0. Reproducido con permiso del propietario del copyright, el equipo UC3M4Safety.

En estas pantallas, a las voluntarias se les pide que etiqueten sus reacciones emocionales. La anotación se realiza en el siguiente orden:

1. Se presentan dos preguntas por pantalla a las voluntarias después de visualizar el estímulo audiovisual -con la intención de captar al menos una señal de voz de un minuto de duración-, como resultado, el micrófono integrado en el casco Oculus Rift S capta una señal de audio que contiene el habla emocional de la usuaria. Estas preguntas se diseñaron para hacer revivir a las voluntarias las emociones sentidas durante la visualización del vídeo, con el objetivo de capturar los últimos rastros de emoción en su voz. La Tabla 3.4 presenta el conjunto de preguntas.
2. Los maniquíes de autoevaluación modificados (SAM) se utilizan para anotar los valores de *valencia/placer, arousal/excitación y dominancia mediante* una escala Likert de 9 puntos. Estos SAM modificados aparecen en la Fig. 3.9, y son resultado de un proceso de rediseño y evaluación que se detalla en [181].
3. También se anota la *familiaridad* con la emoción sentida y la situación mostrada en el videoclip. Ambas se responden utilizando la misma escala Likert de 9 puntos que para las SAM.
4. *El agrado/interés (*en inglés *liking)* que viene a indicar si les ha gustado el vídeo se anota mediante una pregunta binaria sí-no.
5. La selección de una *emoción discreta de* un total de 12, ya descritas en la Sec.3.3.1 [126].

Primera pregunta	Segunda pregunta "Cierra los ojos y piensa en la situación que has visto..."
Describe lo que acaba de ocurrir con tus propias palabras	¿Qué detalles puedes describir?
Explica la situación que has visto con tus propias palabras	¿Qué detalles recuerdas?
Describe lo que has visto con tus propias palabras	¿Qué es lo que más te ha impactado?
Describe lo que has oído con tus propias palabras	¿Qué ocurrió al principio?
Describe dónde y cuándo se ha producido la situación	¿Qué habrías hecho si hubieras estado allí?
	¿Qué habrías hecho si hubieras estado en esa situación?

TABLA 3.4: Preguntas formuladas en la fase de anotación del WEMAC. Se formularon dos preguntas a cada participante, elegidas al azar después de cada visualización de vídeo.

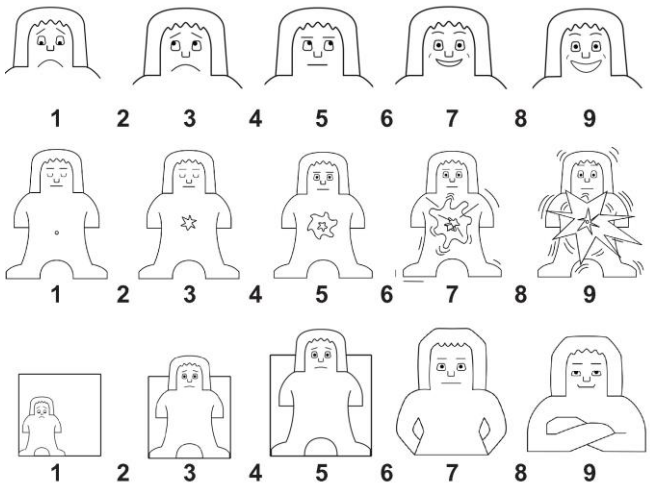


Figura 3. 9: SAM modificados por el Equipo UC3M4Safety [11]. Reproducido con permiso del propietario del copyright, los autores de [181] a través de la licencia Creative Commons CC-BY 4.0 de Frontiers.

Extracción de Características de Audio y *Embeddings* (representaciones)

Dado que no podemos divulgar las señales de voz en bruto por cuestiones éticas y de privacidad -ya que esto podría identificar a las usuarias y relacionarlas con si han sufrido o no violencia de género-, hemos procesado las señales de voz y extraído características de bajo y alto nivel para que la comunidad investigadora pueda analizarlas y trabajar con ellas[24]. Utilizamos distintos conjuntos de herramientas del lenguaje de programación Python para extraer información con un tamaño de ventana de *1 segundo* y un tamaño de salto de *1 segundo* por archivo de audio. Seguimos un enfoque similar al seguido en el *MuSe Challenge* 2021 [182] para la extracción de características y *embeddings* (representaciones numéricas abstractas, ampliamente utilizadas en el campo del Machine y Deep Learning) de las señales de audio:

[24] Disponible en: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing

1. *librosa* [183]: extraemos la media y la desviación estándar de una colección de características de habla calculadas con un tamaño de ventana de *20* ms y un tamaño de salto de *10 ms* mediante el conjunto de herramientas *librosa*. Las 38 características extraídas son 13 coeficientes cepstrales de frecuencias mel (*MFCC*), la raíz cuadrada media (*RMS*) o energía, la tasa de cruce por cero (*ZCR*), el centroide espectral (*spectral centroid*), el balanceo espectral (*spectral rolloff*), la planitud espectral (*spectral flatness*) y la frecuencia fundamental (*pitch*).
2. *eGeMAPS* [184]: calculamos 88 características relacionadas con el habla y el audio mediante la librería *openSMILE* de Python [185] en su configuración por defecto, es decir, un tamaño de ventana de *25* ms y un tamaño de salto de *10* ms.
3. ComParE: extraemos las 6,373 características utilizadas en el *ComParE Challenge* 2016 [186] utilizando de nuevo *openSMILE* de Python.
4. *DeepSpectrum* [187]: extraemos *embeddings* de 6,144 dimensiones mediante este kit de herramientas para la extracción de *embeddings* de audio basado en diferentes arquitecturas de redes neuronales profundas (DNN) entrenadas con ImageNet [188]. En concreto, se consideraron dos configuraciones diferentes, la red *ResNet50* y la salida de la última capa *Average Pooling* (*avg_pool),* que dio como resultado *embeddings* de 2,048 dimensiones, y la red VGG-19 y la última capa *Fully Connected* (*fc2*), que dio como resultado *embeddings* de 4,096 dimensiones.
5. *VGGish*: extraemos *embeddings* de 128 dimensiones de la capa de salida de la red VGG-19 entrenada para AudioSet [170].

Publicamos -Tabla 3.5- la primera versión de la base de datos WEMAC con el objetivo de compartirla con la comunidad investigadora, fomentar la mejora de los resultados de referencia -presentados en [1]- mediante el uso de métodos de fusión, modelos de atención, aprendizaje por transferencia, estrategias semi-supervisadas o de auto-aprendizaje o cualquier otro que la comunidad investigadora considere adecuado; y avanzar en la investigación del análisis multimodal de las emociones en general, y en la igualdad de género en particular.

Base de datos	Conjuntos de datos	Condiciones	Participantes
Base de datos UC3M4Safety [14]	Estímulos audiovisuales: vídeos [11.2] Estímulos audiovisuales: valoraciones emocionales [11.1]	*Crowdsourcing*	Público general y jueces expertos
	WEMAC: Cuestionario biopsicosocial [11.3] WEMAC: Señales fisiológicas [11.4] WEMAC: Características de audio [11.5] WEMAC: Auto-anotaciones emocionales [11.6]	Laboratorio	Víctimas y no víctimas de violencia de género

TABLA 3.5: Jerarquía, subdivisiones y referencias de los conjuntos de datos de la base de datos de seguridad UC3M4 [126] [11].

3.4. Conjunto de datos WE-LIVE: Mujeres y Emociones en la Vida Real

WEMAC es una base de datos de laboratorio para detectar emociones reales desde un punto de vista multimodal en mujeres, pero aún está lejos de las condiciones de la vida real. Grabar habla emocional en condiciones de *miedo* que sea realista y espontánea es muy difícil, si no imposible. Para acercarnos lo más posible a estas condiciones y, tal vez, registrar el habla con *miedo*, el equipo UC3M4Safety creamos el conjunto de datos "*Women and Emotion in real LIfe affectiVE computing dataset: WE-LIVE*".

El objetivo con WE-LIVE es recoger señales fisiológicas, de audio y contextuales de las mujeres voluntarias en un entorno relevante y no controlado, así como recoger el etiquetado de sus reacciones emocionales ante los acontecimientos cotidianos de su vida, utilizando el actual sistema Bindi (pulsera, colgante, aplicación móvil y servidor). Através de la conexión Bluetooth® con el teléfono móvil, los datos captados por Bindi se envían a un servidor protegido y encriptado. Por 'entorno relevante' se entiende la actividad cotidiana dentro de sus rutinas habituales. Los dispositivos sólo realizarán la recogida de datos y la adquisición de señales se realizará de forma simultánea: las señales fisiológicas, de geolocalización, de audio y de voz se contextualizan temporalmente.

La base de datos está compuesta por 13 mujeres voluntarias, entre las que se encuentran víctimas de violencia de género (GBVV). Algunas de ellas también participaron en la recopilación de WEMAC. Como en el caso de este último, las voluntarias fueron reclutadas a través de anuncios en las redes sociales y de actividades de divulgación dirigidas a estudiantes e investigadoras de la universidad.

En primer lugar, para capturar la base de datos, se explica a cada participante la guía informativa que incluye en qué consistirá el experimento. Durante esa sesión se entregan los dispositivos a la usuaria y se instala la aplicación en su *smartphone*. También se explica detalladamente su funcionamiento y uso cotidiano, así como ciertas particularidades asociadas a ellos. También se pide a las voluntarias que rellenen un cuestionario rutinario para recoger sus actividades habituales, que luego serán clasificadas por una psicóloga según la relevancia de la actividad. En cualquier momento del experimento, la voluntaria puede decidir no continuar con él.

3.4.1. Datos Capturados

Cada voluntaria recibe los dispositivos durante 7 días, prorrogables a 10, si ella lo desea. Hay una persona específica -del personal técnico del equipo UC3M4Safety- responsable de cada voluntaria con la que puede ponerse en contacto en cualquier momento. La usuaria lleva una vida normal y los dispositivos captan los datos pertinentes. Los dispositivos,

colgante y pulsera de Bindi, capturan GSR, BVP, SKT, audio, geolocalización y también señales de acelerómetros, en diferentes patrones de grabación según la rutina en la que se encuentre la usuaria en cada momento. Los diferentes patrones incluyen grabaciones más largas durante las franjas de actividad más relevantes del día y breves durante las franjas horarias con poca actividad. En la Fig. 3.10 representamos la versión 2.0 de los dispositivos portátiles de Bindi.

La parte del habla es la más relevante para esta tesis, y es grabada por el micrófono situado en el colgante. El colgante incluye un micrófono omnidireccional MP34DT06 [25] . Según la normativa española[26] , no se requiere el consentimiento de terceros cuando la finalidad del tratamiento de datos es proteger un interés vital del interesado. Además, el audio capturado no está destinado a ser escuchado por ninguna persona, sólo analizado por algoritmos ML, por lo que la privacidad de las usuarias permanece intacta.

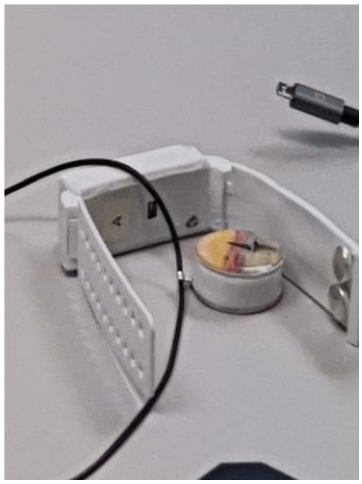


Figura 3. 10: Dispositivos wearable Bindi 2.0. Reproducido con permiso del propietario del copyright, el equipo UC3M4Safety.

La app para smartphone permite controlar la conexión de los dispositivos con el teléfono móvil de la usuaria, el etiquetado de las situaciones y la desactivación de los dispositivos. Esta app también permite el seguimiento de la usuaria por parte del equipo técnico. También dispone de un modo de *reposo* que permite a la usuaria desactivar los dispositivos a voluntad en caso de que desee desactivarlos durante un periodo de tiempo, lo que suele hacerse durante la noche cuando la usuaria duerme y los dispositivos se están cargando. También tiene un modo de *deporte* manual, que etiqueta el periodo de tiempo mientras está activado como de alta actividad física, debido a que la usuaria hace deporte.

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

[26] 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

3.4.2. Etiquetado

Los datos capturados nos proporcionan información detallada sobre la actividad de la usuaria, pero también necesitamos un etiquetado emocional por su parte para que nuestros modelos ML puedan trabajar con los datos y hacer inferencias y predicciones.

Se pide a las usuarias que etiqueten el momento en que sienten una emoción, definida para ellas como *"una reacción breve e intensa a un estímulo concreto (interno, por ejemplo, un recuerdo; o externo, por ejemplo, el sonido de un choque), que provoca cambios corporales (pulso, tensión muscular, expresión facial, etc.) e influye en nuestro comportamiento y pensamiento"*. *Se* pide a las voluntarias que caractericen y definan los acontecimientos emocionales que se han producido en un periodo determinado, así como el contexto de dichos acontecimientos a través de la pantalla de etiquetado de la aplicación móvil de la participante. Para facilitar la anotación puntual de las usuarias, existen dos **modos de etiquetado**:

- **Desencadenado por avisos**: Este tipo de etiquetado se solicita periódicamente después de cada una de las franjas horarias de actividad rutinaria correspondientes y recuerda a la usuaria que es conveniente que etiquete las emociones que experimenta en cada una de ellas al menos una vez.
- **A la carta**: Este tipo de etiquetado se activa cuando la usuaria lo requiere a través de la aplicación del smartphone directamente.

Además, se pide a la usuaria que rellene algunos otros campos de etiquetado para cada etiqueta emocional con el objetivo de que sean más informativos, precisos y proporcionen el mayor contexto posible:

1. ***Arousal*/excitación:** En una escala de calma-activación de 9 puntos utilizando el SAM modificado [181].
2. **Valencia:** En una escala de negatividad-positividad de 9 puntos.
3. **Dominancia:** En una escala de dominancia-sumisión de 9 puntos.
4. **Emoción categorica:** Varias etiquetas de categorías de emociones discretas para seleccionar una entre *aburrimiento, asco, alegría, calma, atracción, sorpresa, esperanza, desprecio, gratitud, ira, miedo* y *tristeza*.
5. **Intensidad de la emoción:** De baja a alta en una escala de 9 puntos.
6. **Experiencia pasada:** Un interruptor booleano (sí o no) para marcar si la usuaria cree que la reacción emocional está relacionada con experiencias traumáticas y/o impactantes del pasado.

7. **Contexto:** Categorías discretas entre las que seleccionar tantas como desee la usuaria, que aportan información contextual. Algunas son: *casa, escuela, trabajo, gimnasio, restaurante, hospital, transporte, café, fiesta, bebidas.*
8. **Nota de audio:** Se trata de una respuesta opcional, en la que la usuaria puede grabar una señal de audio para describir la situación, el estímulo o los motivos que desencadenaron la reacción emocional, los sentimientos, a qué lo atribuye, etc.

Una vez completados todos los campos, la etiqueta emocional se guarda en el sistema. Para cada una de las etiquetas o anotaciones creadas en la aplicación, existen 3 tipos de **marcas de tiempo** asociadas a ella:

- **Creada**: Manual o automática, cada vez que cambia la franja horaria de las rutinas, o justo cuando la usuaria crea la etiqueta.
- **Ocurrida**: Manual, la usuaria selecciona el momento de tiempo en el que sucedieron los estímulos que provocaron la emoción.
- **Enviado**: Manual, denota cuando la etiqueta está completa y las usuarias hacen clic en "Enviar" o "Guardar".

Durante los 7 a 10 días que duró cada experimento, se realizó una llamada telefónica de seguimiento del equipo técnico a la usuaria en dos ocasiones para comprobar que todo funcionaba correctamente y recoger opiniones sobre la comodidad de utilizar los dispositivos y la aplicación, aportando factores como la facilidad de uso.

Todos los datos personales, así como los datos registrados durante el experimento, son anónimos. Eso significa que generamos un código para cada uno de ellos, de modo que es imposible asociar los datos recogidos con la persona que se ofreció voluntaria para el experimento. Los datos anonimizados se conservan durante un máximo de 5 años. Los almacenamos en un servidor seguro y encriptado.

3.5. Conclusiones, Percepciones y Mejoras

En este capítulo hemos explicado qué bases de datos disponibles en la bibliografía se ajustaban más a nuestras necesidades. Hemos detallado las modificaciones que les hicimos y explicado nuestros propósitos al hacerlo. También hemos dejado claro que para nuestros fines específicos -entrenar un detector de situaciones de riesgo de violencia de género- esto no era suficiente. Por ello, hemos creado nuevos conjuntos de datos en consecuencia, describiendo el esfuerzo que nos ha supuesto e ilustrando por qué creemos que se trata de una contribución importante, ya que constituyen un primer paso hacia nuestro objetivo. En esta sección proporcionamos algunos breves mensajes aprendidos de la recopilación de bases de datos y la forma en que vamos a continuar con el uso de estas bases de datos para el desarrollo de mejores modelos ML para la detección de situaciones de riesgo de violencia de género.

El Corpus VOCE nos sirvió como primera aproximación para evaluar las diferencias entre el habla estresada y la neutra, así como para observar la relación entre el habla y los cambios en la frecuencia cardiaca. En la misma línea de investigación, también utilizamos Biospeech. El uso de esta última nos permitió pasar de una configuración binaria estresada a un estado emocional más específico que incluía 4 combinaciones de *arousal*-valencia. Estas bases de datos también permitieron trabajar para mejorar el reconocimiento de hablantes gracias a la diversidad de hablantes disponibles. Además, la reinterpretación de las etiquetas de Biospeech y la adición de eventos acústicos que dieron lugar a Biospeech+ nos permitieron estudiar la influencia de dichos eventos en la detección del estrés en el habla.

Como hemos señalado anteriormente, somos conscientes de que el habla estresada no es habla con *miedo*, pero el uso de la primera nos ha ayudado a analizar el habla en condiciones emocionales similares, sus características y a buscar modelos adecuados a nuestros retos y necesidades, mientras perseguimos el objetivo de la detección de situaciones de riesgo de violencia de género.

Algunos de los retos que plantea la creación de bases de datos para nuestra aplicación específica ya se han descrito en la sección 1.2. En nuestro caso, recopilamos dos bases de datos sólo con datos de usuarias, sin necesidad de realizar un balance de género debido al uso que se le va a dar en el ámbito de la igualdad de género, pero los datos estaban realmente equilibrados en cuanto a la edad: había 5 grupos equilibrados por edad, tanto en WEMAC como en WE-LIVE. En cuanto al trasfondo cultural, todas los participantes eran residentes en España, el país en el que se va a poner en práctica Bindi y para el que se utilizarían los datos registrados.

En lo que respecta al proceso de etiquetado, explicamos minuciosamente el proceso de anotación tanto en WEMAC como en WE-LIVE con el fin de obtener las etiquetas más fiables y precisas posibles, pero es realmente posible que exista un sesgo de fondo en los datos, tanto por parte de los anotadores durante el *crowdsourcing* en el conjunto de datos de estímulos audiovisuales, como en WEMAC y WE-LIVE. Por eso debemos manejar estas etiquetas con cuidado, especialmente las que sólo tienen auto-anotaciones. En lugar de utilizarlas como etiquetas *ground-truth* (verdad absoluta), podríamos estudiar nuevas formas de utilizarlas en el futuro, por ejemplo, teniendo en cuenta otros aspectos de los antecedentes de cada usuaria que puedan influir en el etiquetado, o utilizando una etiqueta de emoción agregada para tener una etiqueta más fiable para cada vídeo en el caso de WEMAC.

Centrándonos en el habla, esperamos que el habla capturada en WEMAC contenga rastros de la emoción, pero no podemos afirmarlo ya que la grabación se realiza justo después de la visualización de los estímulos que provocan la emoción y no durante. Durante la recopilación de la base de datos, hemos observado que algunas usuarias pedían hacer una

breve pausa entre la visualización de los estímulos y la captura de la señal del habla, para recuperarse de los estímulos que habían visto, lo cual es normal, ya que había vídeos que pretendían provocar *miedo* real. Sin embargo, esto significa que el contenido emocional de las señales del habla puede ser variable, y eso es algo que también debemos tener en cuenta al trabajar con esos datos.

En cuanto a la base de datos WE-LIVE, su grabación finalizó en julio de 2022, y el procesamiento y análisis de los datos capturados para su uso posterior sigue siendo un trabajo en curso. Es necesario realizar un trabajo exhaustivo en este sentido, en primer lugar, con la limpieza de datos, ya que los dispositivos han registrado una enorme cantidad de datos que no siempre contienen información relevante para la tarea; después en cuanto a la coordinación y alineación de los datos, ya que es necesario sincronizar las diferentes modalidades y resolver el problema de los datos que faltan en cada modalidad en particular. También en el caso de WE-LIVE, hubo algunos problemas técnicos durante el momento de la captura, la mayoría de los fallos se debieron a desconexiones de los dispositivos con el smartphone -vía Bluetooth (BT)-, y al inestable funcionamiento de la pulsera cuando entraba el sudor. El equipo está abordando esta cuestión y se espera diseñar una versión mejorada en el próximo prototipo de Bindi: Bindi 3.0.

Capítulo 4: Reconocimiento del Hablante en Condiciones de Variabilidad

Para nuestro objetivo de detectar situaciones de riesgo de violencia de género a través del habla, el primer paso a dar parece ser detectar la voz de la usuaria concreta que nos interesa, de entre toda la información contenida en la señal de audio. Para detectar de forma coherente la emoción, especialmente *el miedo*, en el habla de una usuaria, primero hay que aislar su voz de la de otros hablantes en un escenario acústico. Esta práctica también abre interesantes oportunidades en situaciones en las que es necesario identificar a todos los hablantes de una escena, por ejemplo, en las pruebas forenses. Así, dedicamos este capítulo a la investigación en el campo del reconocimiento de hablantes (*speaker recognition*, SR) porque después de detectar a la usuaria podemos hacer un análisis de las emociones que experimenta.

Este campo es ligeramente esquivo en nuestro caso, porque necesitamos detectar la voz de la usuaria para identificarla, pero el rendimiento de los modelos ML para detectar hablantes a través de la voz desciende cuando se encuentran en condiciones emocionales. Así pues, el hecho de que la voz de una usuaria pueda verse influida por su estado emocional constituye un reto para un sistema de identificación de hablantes. Nos interesa conseguir buenos rendimientos para el reconocimiento del hablante incluso cuando la voz expresa estrés o *miedo*.

A continuación, las aportaciones de este capítulo se basan en nuestro estudio de los sistemas de reconocimiento de hablantes en condiciones de variabilidad, 1) la identificación de hablantes en condiciones de estrés, para ver en qué medida estas condiciones de estrés afectan a los sistemas de SR[27] y 2) el reconocimiento de hablantes en condiciones de ruido real, que es donde funcionará en última instancia nuestra aplicación, aislando la voz del hablante, entre todos los ruidos ambientales adicionales.

4.1. Introducción

El reconocimiento del hablante (SR) se refiere a la tarea de detección automática de una persona a partir de las características de su voz, también conocida como biometría vocal [189]. En él podemos distinguir dos subtareas, la identificación del hablante (SI) y la verificación del hablante (SV). La primera se refiere al reconocimiento de una usuaria concreto entre un número conocido de usuarios (un escenario multiclase), y la segunda pretende identificar a una usuaria frente al resto (escenario binario). Es en la identificación del hablante donde nos centramos a lo largo de este capítulo, en la capacidad de detectar a

[27] A falta de bases de datos sobre el habla en condiciones de miedo realista en la literatura - lo que se explica detalladamente en la Sec. 3.1 - en el momento en que se realizó esta parte de esta tesis

quién pertenece la voz, incluso en condiciones emocionales. Los efectos de las emociones en SI [190] se han estudiado en la literatura, pero son escasas las investigaciones sobre la influencia del estrés específicamente. El desarrollo tecnológico de Reconocimiento de Emociones del Habla (SER) es abundante, pero la tarea del SR en condiciones emocionales es todavía un campo científico en fase inicial.

En cuanto al sistema SI a diseñar, debe adaptarse a lo que esperamos que Bindi encuentre en una situación del mundo real: el objetivo de nuestro sistema es detectar a las personas que hablan incluso cuando sus voces presentan condiciones de *miedo* o estrés. Por este motivo, podríamos enfrentarnos a un problema de aprendizaje desajustado en el que sólo dispondríamos de señales de voz neutros para el entrenamiento de nuestros modelos, recogidos en una configuración inicial de Bindi -dado que la posibilidad de forzar a la usuaria a hablar mientras se encuentra en condiciones de *miedo* o estrés es difícil-, mientras que las condiciones de funcionamiento del entorno real contendrían muestras tanto neutras como estresadas o con *miedo*.

Desde un punto de vista más técnico, en este capítulo estudiamos y diseñamos sistemas de identificación del hablante robustos a las condiciones de variabilidad que podría inducir un micrófono embebido en un dispositivo *wearable* que trabaja en un entorno real, como son las emociones -el estrés y el *miedo*, en particular- y las condiciones de ruido ambiental en el habla. Rastreamos estos problemas mediante distintas técnicas, como el aumento de datos (*data augmentation*, DA) o la generación de datos sintéticos, que tienen en cuenta las restricciones computacionales de Bindi y las características de la entrada de audio.

4.2. Trabajos Relacionados

Ya hemos mencionado las dificultades a las que nos enfrentamos a la hora de buscar bases de datos adecuadas para desarrollar modelos de aprendizaje automático apropiados para nuestra tarea en la sección 3.1. Existen muy pocas bases de datos en las que se registre el habla estresada en condiciones reales -y no hay ninguna en el caso del *miedo* real-, además del reto que supone el proceso de etiquetado.

Características extraídas manualmente (*handcrafted features*)

En la literatura se utilizan diversas estrategias de características elaboradas o seleccionadas a mano para aplicaciones relacionadas con el habla, [191], [192], [193]. Los sistemas de identificación de hablantes trabajan con señales del habla e intentan utilizar características acústicas que difieran entre personas para poder discriminar entre ellas. Algunas de las características que muestran un buen rendimiento cuando se utilizan en condiciones neutras o sin emoción en los sistemas basados en el habla son los coeficientes cepstrales de frecuencia mel (MFCC) [194]. Los MFCC modelan el sistema auditivo humano para captar las

características fonéticamente importantes del habla, distribuyendo los coeficientes de frecuencia mel de forma casi lineal en las frecuencias bajas y logarítmicamente en las altas. Aunque pueden calcularse muchos de los coeficientes, normalmente se calculan los 12 o 13 primeros. Las características prosódicas también se utilizan mucho.

Los rasgos prosódicos son características suprasegmentales que aparecen cuando los sonidos se juntan en el habla conectada. Algunos de los ámbitos o fenómenos para los que se aplican estas carácterísticas son la entonación, las sílabas acentuadas y el ritmo. Además, también se utilizan rasgos fonéticos. Éstas también modelan la acústica captando la variación de pronunciación entre hablantes y en el espacio acústico, lo que permite modelar patrones a más largo plazo como la detección de los fonemas y sus estadísticas. En la misma línea, la predicción lineal (LP) se utiliza en el procesamiento del audio y del habla para definir la llamada envolvente espectral de las señales del habla de forma comprimida, utilizando los coeficientes optimizados de un modelo predictivo lineal. También es una potente técnica de análisis del habla para proporcionar estimaciones de parámetros del habla como el tono, la duración y la energía [195].

Aunque para la identificación de hablantees en condiciones de estrés apenas hay trabajos previos, se utilizan los MFCC [196] junto con características prosódicas como el tono, la energía y la duración de las palabras [197] y consiguen buenos resultados [198]. Sin embargo, su capacidad de generalización y su robustez frente a la variabilidad son limitadas.

Características automáticas (*automatic features* o *embeddings*)

Más allá de las características manuales mencionadas anteriormente, los *embeddings* son representaciones numéricas abstractas obtenidas automáticamente a partir de meter los datos en bruto en DNN (redes neuronales profundas). Esta es una tendencia novedosa que logra resultados muy innovadores [199], [200]. En la última década, se ha descubierto que, cuando se dispone de datos suficientes, las representaciones de características aprendidas automáticamente o *embeddings* basados en DNN suelen ser más eficaces que las características elaboradas a mano o diseñadas manualmente, lo que permite desarrollar modelos predictivos mejores y más rápidos [201], [202]. Y lo que es más importante, las representaciones/*embeddings* aprendidas automáticamente suelen ser más flexibles y potentes. El aprendizaje a partir de *embeddings* consiste en obtener características abstractas y útiles normalmente a partir de la forma de onda de la señal de audio directamente, o a partir de representaciones de baja dimensión relativamente complejas, mediante el uso de autocodificadores y otras arquitecturas de aprendizaje profundo que a menudo generalizan mejor a tipos de datos con los que no han sido entrenados [203], [204]. No obstante, en nuestro caso, el uso de enfoques DNN complejos debe manejarse con cuidado debido a su elevada carga computacional, el tiempo que tarda y la disponibilidad de conjuntos de datos de entrenamiento

suficientemente grandes, que son tres limitaciones muy importantes dentro del sistema Bindi, detalladas en la sección 1.2.2. Además, las características automáticas pueden ser difíciles de entender o interpretar por los humanos, al tener poca "explicabilidad", lo que va en detrimento de la transparencia de los modelos de aprendizaje automático.

Debido a la naturaleza secuencial de las señales del habla, su contexto temporal es de gran relevancia para las tareas de clasificación y predicción [205]. Además, el carácter secuencial de sus contenidos frecuenciales conlleva información muy relevante del habla [206]. Las redes neuronales recurrentes (RNN) son potentes herramientas para modelar datos secuenciales [207], habiéndose convertido en el estado del arte debido a su mayor rendimiento y capacidad de generalización. Sin embargo, la disponibilidad de bases de datos grandes es, de nuevo, de vital importancia para entrenar dichas redes. Desafortunadamente, éste no es el caso de los datos de situaciones de estrés reales en particular, como las que nos ocupan.

Las redes neuronales profundas son incluso capaces de condensar eficazmente la información relacionada con la identidad del hablante, pudiendo excluir el resto de información irrelevante para tareas de SR, en lo que se denominan *embeddings* de hablante. En un sentido amplio, todos los *embeddings* neuronales que incluyen alguna forma de agrupación temporal global y se entrenan para identificar a los hablantes en un conjunto de grabaciones de entrenamiento se unifican bajo el término *x-vectors* según [208], [209]. Las variantes de los sistemas *x-vector* se caracterizan por diferentes arquitecturas del codificador; métodos de *pooling* y objetivos de entrenamiento [210] y, en este sentido, todos los *embeddings* probadas en esta sección podrían considerarse como tales.

Aumento de datos (*data augmentation*, DA)

El DA es un ingrediente clave de las tecnologías del habla más avanzadas, ya que es una estrategia común adoptada para aumentar la cantidad de datos de entrenamiento. También puede actuar como regularizador para evitar el sobreajuste [211] y mejorar el rendimiento en problemas de clases desbalanceadas [167]. Esto hace que todo el proceso sea más robusto y se consiga un mejor rendimiento. Debido a la escasez de datos que mencionamos en la Sec. 3.1 y aunque los datos resultantes de esta técnica no sean totalmente realistas, es una buena opción para nuestro caso, ya que las bases de datos que utilizamos pueden ser bastante pequeñas y presentar desequilibrios de datos. Utilizando DA, podemos aumentar la cantidad de datos disponibles y hacer frente a la falta de equilibrio entre clases [212], aproximándonos a cómo funcionaría el sistema en un caso real para el que aún no dispondríamos de datos.

Junto con la gran cantidad de investigaciones para hacer frente al problema generalizado de la variabilidad de las señales de voz, el DA es una técnica muy aplicada para ampliar las bases de datos [213], por ejemplo, añadiendo ruido o aplicando transformaciones a las señales de voz, similares a las que introducen los canales de transmisión.

Las técnicas de realce del habla (*speech enhancement,* SE) también se utilizan para mejorar la calidad perceptiva general del habla, concretamente la inteligibilidad [214], [215], [216]. Notablemente, estas técnicas pueden modificarse hacia un objetivo de reconocimiento del hablante, en lugar de simplemente mejorar la calidad del audio [210].

Además, para paliar el desajuste intrínseco de la variación y, en concreto, el provocado por las emociones en las tareas de identificación de hablantes, la literatura considera varias soluciones, como elicitar emociones en los hablantes de forma que se consigan efectos similares a los espontáneos [217] debido a las dificultades de registrar emociones auténticas -tanto en términos de privacidad como de etiquetado-. Asimismo, también se utilizan estimaciones estadísticas y métodos de adaptación al dominio [218], [219], [220]. Esta falta de conjuntos de datos que contengan emociones negativas reales y naturales -no actuadas- en el habla, como las que podría experimentar una usuaria en una situación de riesgo o violencia, es sin duda un reto.

Modelos de clasificación

En cuanto a los modelos que se utilizan para este tipo de tareas, algoritmos como los modelos de mezclas gaussianas (GMM) se emplearon generalmente para el reconocimiento de hablantes [221] y las máquinas de vectores de soporte (*support vector machine*, SVM) también se aplican ampliamente [222],[223]. Otros estudios sugieren el uso de DNN para el reconocimiento de hablantees [224], [225] La mayor precisión alcanzada por las DNN, en comparación con otros sistemas del estado del arte, es el resultado de su capacidad para extraer representaciones (*embeddings*) discriminantes de los datos que son robustas a la variabilidad, especialmente en las señales del habla. En los últimos años, los algoritmos de aprendizaje profundo se han disparado en muchos campos científicos gracias a la disponibilidad de grandes cantidades de datos.

Sin embargo, en la investigación llevada a cabo en este capítulo, pretendemos mantener un equilibrio entre la complejidad computacional y la precisión debido a las limitaciones de hardware del dispositivo, donde el consumo de batería es crítico -que impone el *hardware* de nuestros dispositivos de Bindi- y la escasez de datos de entrenamiento disponibles en un principio (véanse los retos en la sección 1.2).

Variabilidad del ruido

Recientemente, para hacer frente a los problemas del ruido ambiental que aparece en situaciones de la vida real en las señales de audio, el aumento de datos (*data augmentation*, DA) con ruido aditivo y convolucional con redes neuronales se ha alzado como uno de los mejores enfoques en SR. Después, el uso de modelos para eliminar el ruido (*denoising*) -o

dereverberar- eficazmente muestras de habla manteniendo la información específica del hablante mediante DNN es un campo con trabajos emergentes [226]. La investigación actual incluye modelos en dos etapas que muestran una mejora de la inteligibilidad del hablante [227], arquitecturas de memoria a largo plazo (LSTM) que explotan las características secuenciales del habla [228], módulos de mejora de características no supervisados y robustos en condiciones sin restricciones de ruido [229], y módulos de mejora del habla especialmente dirigidos con la optimización conjunta de los módulos de identificación del hablante y de extracción de características [230],[204],[215].

4.2.1. Desafíos de la Variabilidad en el Reconocimiento del Hablante

El habla en la vida real suele contener ruido y se encuentra en condiciones no restringidas que son difíciles de predecir, esto complica su reconocimiento y comprensión. Los sistemas de reconocimiento del habla (SR) necesitan un alto rendimiento en estas condiciones del "mundo real". Esto es extremadamente difícil de conseguir debido a las variaciones tanto extrínsecas como intrínsecas y se conoce comúnmente como reconocimiento del hablante *en la naturaleza (speaker recognition in-the-wild)*. Normalmente, este problema de variabilidad afecta a los modelos del habla porque estos suelen haberse entrenado con datos de habla en condiciones de laboratorio (es decir, limpios de ruido) y su rendimiento baja cuando en la fase de *test,* predicción o inferencia se encuentra con datos con ruido. Esto significa que es necesario desarrollar sistemas robustos que puedan manejar la variabilidad sin una degradación del rendimiento. Las variaciones extrínsecas abarcan la conversación y la música de fondo, el ruido ambiental, la reverberación, los efectos del canal de transmisión y del micrófono, etc. Por otro lado, las variaciones intrínsecas son los factores presentes en el habla inherentes de los propios hablantes, como la edad, el acento, la emoción, la entonación o la velocidad del habla [231].
Los sistemas de reconocimiento automático del habla (ASR) pretenden extraer la información lingüística del habla a pesar de las variaciones intrínsecas y extrínsecas [210]. Sin embargo, el reconocimiento del hablante (SR) aprovecha las variaciones intrínsecas o idiosincrásicas para averiguar la identidad de cada hablante. Además de la variabilidad intra-hablante (emoción, salud, edad), la identidad del hablante resulta de una compleja combinación de aspectos fisiológicos y culturales. Aun así, el papel del habla emocional no se ha explorado en profundidad en SR. Este podría considerarse un rasgo intrínseco, plantea un reto debido a las distorsiones que produce en la señal del habla para otras tareas. Influye significativamente en el espectro del habla, lo que tiene un impacto considerable en las características frecuenciales que se extraen de ella y por eso deteriora el rendimiento de los sistemas de reconocimiento de hablante.

El problema del desajuste entre las características estadísticas de los datos de entrenamiento de los modelos y de los datos en condiciones de la vida real puede dar lugar a características muy diferentes en la voz del hablante, lo que hace que los modelos de reconocimiento de hablantes pierdan parte de su precisión y poder predictivo. Las variaciones extrínsecas constituyen desde hace tiempo un reto que afecta a la base de todas las tecnologías del habla. Las redes neuronales profundas han dado lugar a mejoras sustanciales gracias a su capacidad para tratar con conjuntos de datos ruidosos del mundo real sin necesidad de características diseñadas específicamente para ser robustas. Sin embargo, uno de los ingredientes más importantes para el éxito de estos métodos es la disponibilidad de conjuntos de datos de entrenamiento muy amplios y diversos.

4.3. Efectos del Estrés en las Tasas de Reconocimiento de Hablantes

En esta sección detallamos los experimentos realizados en nuestras propias contribuciones publicadas en [3] y [10]. En ellos, pretendemos analizar cómo afecta el estrés en el habla a las tasas de reconocimiento de hablantes. Pretendemos encontrar técnicas para reforzar los sistemas de reconocimiento de hablantes, ya sea neutralizando los efectos del estrés -y, en última instancia, del *miedo*- o siendo capaces de sintetizarlo a partir del habla neutra, para reforzar el entrenamiento de los modelos de inferencia del aprendizaje automático. Proponemos el uso de técnicas de DA utilizando habla generada sintéticamente en condiciones de estrés, modificando el tono y la velocidad del habla. Además también proponemos un análisis de los mejores métodos de extracción de características para crear modelos de inferencia de SR robustos a la variabilidad de las emociones adaptando los datos y manteniendo al mismo tiempo una arquitectura ligera.

El diagrama de bloques de la metodología seguida se representa en la Fig. 4.1. A continuación se describen en detalle las características de la base de datos empleada, el proceso de etiquetado automático basado en la medición de la frecuencia cardiaca, la extracción de características en dos etapas y las técnicas de aumento y normalización de datos.

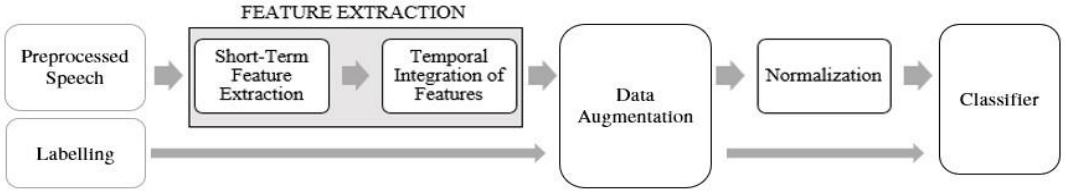


Figura 4. 1: Diagrama de bloques de la metodología de reconocimiento de hablantes en condiciones de estrés con el corpus VOCE. Reproducido con permiso del propietario del copyright, los autores de [3] a través de la licencia Creative Commons CC-BY 4.0 de MDPI.

En esta parte de la tesis utilizamos la base de datos del corpus VOCE [156] que ya describimos en detalle, junto con el preprocesamiento y las técnicas de aumento de datos utilizadas en la sección 3.2.1. En cuanto al etiquetado, trabajamos con dos tipos de etiquetas para cada muestra de audio: etiquetas booleanas de *estrés* que indican la presencia o ausencia de estrés, y etiquetas de hablante, que toman valores de 1 a *n*, representando el identificador de hablante de cada la muestra de audio, donde *n* es igual al número total de hablante. Esta *n* es, diferente en cada conjunto o set; en el Set 1, *n* es igual a 10, y en el Set 2, *n es* igual a 11, con un total de 21 hablantes - véase la Sec. 3.2.1 -.

Tras el bloque de pre-procesamiento del habla, extraemos de ella características acústicas extraídas manualmente (*handcrafted*). Éstas deben reflejar la anatomía del sistema de producción del habla (por ejemplo, el tamaño y la forma de la laringe y la boca) y los patrones de comportamiento aprendidos que conforman los hábitos de habla (por ejemplo, el tono de voz, el estilo de habla, etc). Para la extracción de características utilizamos una ventana de 20 ms con un 50% de solapamiento -valores muy comunes utilizados para analizar la evolución temporal de las señales en la literatura de las tecnologías del habla [232]-. Para convertir los vectores de características a la resolución de 1s y alinearlos con las etiquetas de estrés, realizamos la media y la desviación estándar de las características acústicas sobre segmentos de 1 segundo. Así obtenemos un vector de características por cada segundo de señal de audio. Estas características suman un total de 34 y son la media y la desviación estándar de MFCC 1-13, formantes de 1 a 3 y frecuencia fundamental de la voz. Se seleccionaron de acuerdo con la bibliografía sobre las tareas de reconocimiento de emociones y de hablantes [232, 213].

4.3.1. Muestras de Habla Estresadas Generadas Sintéticamente

El tono y la velocidad de elocución fueron dos variables que observamos informalmente que cambiaban entre las señales de habla neutras y las estresadas. En consecuencia, realizamos un análisis para medir las diferencias entre el tono medio de las locuciones neutras y las estresadas de cada hablante utilizando la librería de VoiceBox [166]. La estimación de la velocidad media de elocución de cada usuaria se calculó computando el número medio de palabras por segundo de cada hablante, obteniendo una transcripción automática de cada una de las grabaciones mediante Google Speech Recognition [234], y dividiéndola entre la longitud de las señales de audio tras haber eliminado las partes de la señal de audio en silencio con un módulo VAD.

Las diferencias de tono del habla neutra a la estresada se situaron en un rango entre un porcentaje relativo de -2% y +7% para todos los hablantes, aumentando una media de 2,2% Hz. En cuanto a la velocidad de elocución, subjetivamente, parecía aumentar en los enunciados de habla estresada, sin embargo nuestro análisis nos dio la conclusión contraria.

El número de palabras por segundo era mayor cuando el hablante leía un texto, 2,2 palabras/s de media, en comparación con cuando el orador realizaba una presentación oral, 1,85 palabras/s. Al escuchar las señales, determinamos que las palabras se pronunciaban más rápido durante la exposición oral, pero había muchas pausas cortas y palabras de relleno - palabras como '*ehm*', '*um*', '*ah*' - entre ellas, que no contaban como palabras para la transcripción pero que tampoco eran eliminadas por el módulo VAD. Estas causas provocan una velocidad de elocución inferior en general.

Así, aplicamos modificaciones en la velocidad de locución y la frecuencia fundamental de la voz sobre la base de datos original, para producir muestras de habla estresada generadas sintéticamente. La frecuencia se modificó en los siguientes porcentajes relativos [-6%, -3%, +3%, +6%] y las señales de habla se ralentizaron -con el objetivo de alargar la duración- en los siguientes porcentajes [-20%, -15%, -10%, -5%]. Todas estas modificaciones se aplicaron a las señales de audio originales y dieron como resultado lo que denominamos un nuevo conjunto de datos de habla estresada generado sintéticamente. De este modo, aumentamos nuestros datos en un factor de 9, el conjunto de datos original más 8 modificaciones.

4.3.2. Configuración Experimental y Resultados

Originalmente, para la configuración experimental inicial utilizamos los datos disponibles para los Sets 1 y 2 juntos (21 hablantes, detallados en la Sec. 3.2.1, el número de muestras puede observarse en la Fig. 4.1). Este experimento preliminar se realiza para observar el comportamiento de la tasa de identificación de hablantes en condiciones de *desajuste*. En primer lugar, dividimos los datos en muestras neutras (N) y estresadas (S) y experimentamos entrenando un modelo de aprendizaje máquina con un tipo de dato, y haciendo la inferencia con el otro y luego mezclando ambos tipos, utilizando un perceptrón multicapa (MLP) convencional como modelo de inferencia. Los **resultados** en términos de precisión -el porcentaje de segmentos de audio clasificados correctamente- pueden consultarse en la tabla 4.2. Para obtener resultados fiables, estos experimentos se repitieron 50 veces, y en cada repetición los datos utilizados para las pruebas (30%) se eligieron al azar.

Muestras	**Neutro**	**Estresado**	**Total**
Set 1	1389	3989	5378
Set 2	1716	4858	6574
Total	3105	8847	11952

TABLA 4.1: Número de muestras de VOCE utilizadas [3].

Como cabía esperar en un principio, la configuración en la que se utilizan el mismo tipo de datos para entrenamiento e inferencia (condiciones de *match*) versus la configuración en la que se

utilizan tipos de datos distintos para entenamiento e inferencia (condiciones de *mismatch*), se obtienen mejores resultados en la primera configuración que en la segunda. Cuando se entrena con habla neutra y se hace inferencia con datos de habla estresada, la precisión disminuye más de un 15% con respecto a las condiciones de *match*, por lo que parece que el habla estresada tiene características diferentes en comparación con el habla neutra que afectan al modelo de SI. Por el contrario, cuando se entrena con datos de habla estresada y se realizan pruebas con datos de habla neutral, la disminución de la precisión con respecto a la configuración *match* no es tan importante (5% absoluto) en comparación con el caso inverso. Esto puede indicar que el habla estresada podría tener una distribución de datos menos homogénea en el que podría estar contenido el habla neutra, pero no a la inversa. En cuanto a los experimentos con condiciones mixtas, la precisión alcanzó un 96,05%, logrando un resultado positivo para esta tarea en particular.

En la siguiente configuración experimental, pretendemos medir la precisión alcanzada por el sistema al entrenarse con los distintos conjuntos de datos estresados generados sintéticamente. Realizamos modificaciones de frecuencia y velocidad para estresar artificialmente los datos del set 1 de hablantes, y haremos inferencia con el habla estresada original. Esperamos que los resultados obtenidos en estos experimentos reflejen qué modificación imita mejor el habla estresada original.

Conjunto de entrenamiento	**Conjunto de inferencia**	**Media (%)**	**Std (%)**
Neutro	Neutro	96.73	0.33
	Estresado	79.21	0.90
Estresado	Estresado	95.87	0.28
	Neutro	90.89	0.49
Mixto	Mixto	96.05	0.12

TABLA 4.2: Resultados de precisión para el reconocimiento de hablantes en condiciones de estrés con VOCE en configuración de match y mismatch [3].

Mantuvimos fijo el conjunto de test o inferencia para estos experimentos, que fue un 30% de las muestras del habla estresada original. Además, se eliminó el mismo 30% en cada conjunto estresado generado sintéticamente para lograr una comparación más precisa entre experimentos y garantizar que las muestras de inferencia nunca estuvieran presentes en el conjunto de entrenamiento, aunque hubieran sido modificadas por nuestro procedimiento de aumento de datos.

Los grupos de los conjuntos de la Fig. 4.2 se agruparon formando distintas combinaciones para reconocer las diferencias de precisión del modelo para cada configuración concreta. En el lado izquierdo, representamos el conjunto de datos original, compuesto por muestras de habla

neutras y estresadas. En este caso, utilizamos el 30% de los datos de la colección estresada como conjunto de test para la inferencia para los experimentos posteriores. A la derecha, representamos un esquema de uno de los conjuntos estresados generados sintéticamente (el habla neutra original se convierte en "habla estresada sintética" [*synthetic stress, SS]* y el habla estresada original se convierte en "habla *super estresada* sintética" [*synthetic super stressed*, SSS]). El 30% de los datos utilizados antes como conjunto de test se eliminó para obtener resultados más fiables. Hay varios conjuntos de datos generados sintéticamente, uno por cada modificación aplicada.

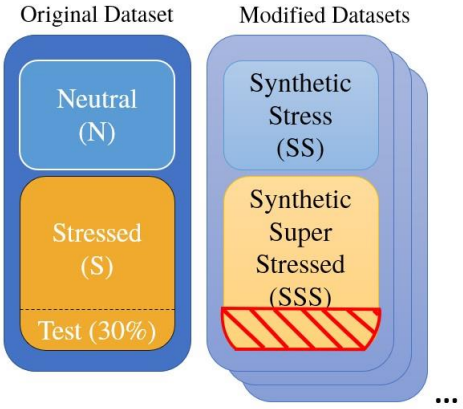


Figura 4. 2: Esquema de los conjuntos de datos original y modificado de VOCE. La parte roja se refiere al equivalente a las muestras de prueba del bloque de la izquierda, lo que significa que se eliminaron correctamente cuando se utilizó SSS para el entrenamiento.

En las Figs. 4.3 y 4.4 presentamos los **resultados obtenidos**, enumeramos los datos utilizados para la fase de entrenamiento en el eje *X*, el eje *Y* representa la precisión alcanzada, y cada barra de color indica el conjunto modificado utilizado para el entrenamiento. En ambas se utilizan solo los datos del conjunto o Set 1, y el conjunto de test se mantiene siempre igual por fiabilidad. En la Fig. 4.3 podemos observar que las modificaciones que obtienen las mayores precisiones son *Pitch +3%* y *Pitch -3%* (el *pitch* equivale a la frecuencia fundamental de la voz). En cuanto a la Fig. 4.4, aunque los resultados de velocidad son muy similares, la modificación que funciona peor es *Velocidad -20%*. En cuanto a los conjuntos de entrenamiento utilizados, el conjunto SSS funciona mejor que el SS en ambos casos.

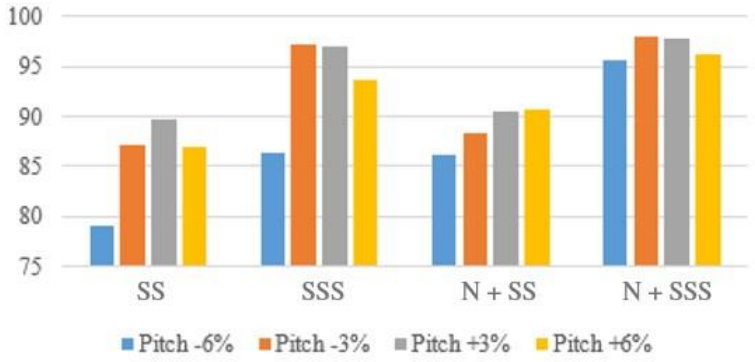


Figura 4. 3: Resultados de precisión entrenando el modelo con datos estresados generados sintéticamente con modificaciones de tono. Reproducido con permiso del propietario del copyright, los autores de [3] a través de la licencia Creative Commons CC-BY 4.0 de MDPI.

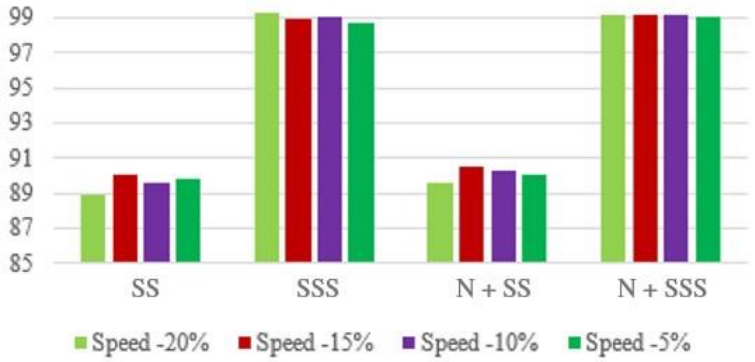


Figura 4. 4: Resultados de precisión entrenando el modelo con datos estresados generados sintéticamente con modificaciones de la velocidad. Reproducido con permiso del propietario del copyright, los autores de [3] a través de la licencia CC-BY 4.0 de MDPI.

Para la siguiente serie de experimentos decidimos realizar las modificaciones en las grabaciones de audio del set o conjunto 2 que han logrado mayores índices de precisión en el set 1. Estas fueron el *pitch* [-3%, +3%] y la velocidad [-15%, -10%, -5%], como ya se ha mencionado. Unimos los conjuntos 1 y 2, transformando el problema en una tarea SI de 21 hablantes y combinamos todos los datos de estrés generados sintéticamente en un conjunto único de datos, aumentando en un factor de 6 el tamaño de los datos originales: 5 modificaciones más el conjunto de datos original. Se realizaron los mismos análisis para el conjunto 1 y los conjuntos 1 + 2.

En la tabla 4.3 observamos dos tipos de experimentos, unos en los que sustituimos los datos y otros en los que los aumentamos los datos de la fase de entrenamiento. Estos experimentos los repetimos 20 veces para mayor fiabilidad. En cuanto a la sustitución del conjunto original por uno estresado generado sintéticamente, tenemos los experimentos número 6 y 7 para compararlos con los experimentos número 1 y 2 respectivamente. La sustitución de los datos logra resultados similares a los obtenidos con los datos originales cuando se utilizan datos generados sintéticamente de habla neutra para el entrenamiento (caso 1 frente al caso 6), así

como mejores tasas de identificación cuando se utilizan datos generados sintéticamente y obtenidos a partir del habla estresada (caso 2 frente al caso 7).

Exp. Num.	Caso	Set 1, *Media*	Set 1, *Std*	Sets 1 + 2, *Media*	Sets 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

TABLA 4.3: Resultados de precisión para el reconocimiento de hablantes en condiciones de estrés con VOCE con habla generada sintéticamente utilizando diferentes combinaciones [3].

Los experimentos de aumento de datos son el 3, 4, 5, 8, 9 y 10. El resultado es realmente positivo, los mejores resultados se obtienen en el experimento 8 con un 99,45% de precisión para los conjuntos 1 + 2. Estos resultados nos muestran que aumentar los datos con habla estresada generadas sintéticamente aumenta la tasa de SI.

Uno de los objetivos de estos experimentos era determinar si el experimento 4 podía superar al experimento 2. Esto significaría que habíamos cumplido la tarea de generar un habla estresada sintéticamente adecuada a partir del habla neutra. Sin embargo, vemos que el procedimiento que empleamos no era suficiente para ser utilizado como sustituto. No obstante, en la tabla 4.3 observamos que el caso 4 obtiene mejores resultados que el caso 6, que a su vez supera al caso 1. Esto demuestra que el habla estresada generada sintéticamente y utilizada como datos de entrenamiento junto con los datos estresados originales aumenta el rendimiento del sistema SI.

4.3.3. Discusión

Nuestro objetivo en esta sección era analizar cómo influía el habla estresada en el rendimiento de los sistemas de identificación de hablantes. Hemos identificado un problema, el habla estresada en la fase de test afecta negativamente cuando los sistemas de SI se entrenan sólo con habla neutra.

En cuanto al caso de las condiciones de *match* y *mismatch*, en la configuración mixta -utilizando habla original neutra y estresada tanto para el entrenamiento como para el test- el sistema SI alcanza un 96,05% de precisión, una tasa muy satisfactoria para este tipo de tareas, lo que demuestra que el conjunto de características elegido para la tarea es adecuado.

En los experimentos preliminares de sustitución de datos, dependiendo de la diferencia entre los datos generados sintéticamente y los originales utilizados para el entrenamiento, algunas sustituciones superan los resultados obtenidos con los datos originales. Además, las modificaciones sobre la frecuencia fundamental de la voz del hablante funcionan mejor cuando incluimos muestras estresadas generadas sintéticamente para el entrenamiento, que cuando incluimos las modificaciones en la velocidad del habla. Sin embargo, cuando utilizamos muestras acentuadas *superestresadas* generadas sintéticamente para el entrenamiento, los conjuntos modificados por cambios en la velocidad obtienen mejores resultados.

En cuanto a los experimentos para aumentar la base de datos con estrés artificial, podemos concluir que la generación de diferentes tipos de habla estresada generada sintéticamente mediante modificaciones en la frecuencia y la velocidad, y su adición a la base de datos, amplía significativamente las muestras con las que trabajar, mejorando sustancialmente los resultados obtenidos por el sistema de identificación de hablantes con un 99,45% de precisión.

Varios experimentos y métodos quedaron sin explorar y se dejan para futuros trabajos, como cambiar el paradigma a un entorno de verificación de hablantes, lo que podría reducir las condiciones del problema y hacerlo más conveniente para nuestra tarea. Además, utilizar el habla en condiciones de *miedo* real haría que las condiciones de entrenamiento y prueba coincidieran por completo. Dado que Bindi trabaja en entornos reales, sería oportuno reforzar el sistema entrenándolo con audios modificados como si hubieran sido grabados en un entorno real, por ejemplo, añadiendo ruido a la base de datos utilizada y analizar su efecto. También sería interesante seguir analizando las diferencias entre el habla neutra y el habla estresada para encontrar nuevas modificaciones que aplicar al habla neutra para transformarla en un habla acentuada adecuada generada sintéticamente.

4.4. *Embeddings* de Hablantes a partir de un *Auto-encoder* Recurrente con Eliminación de Ruido de Extremo a Extremo

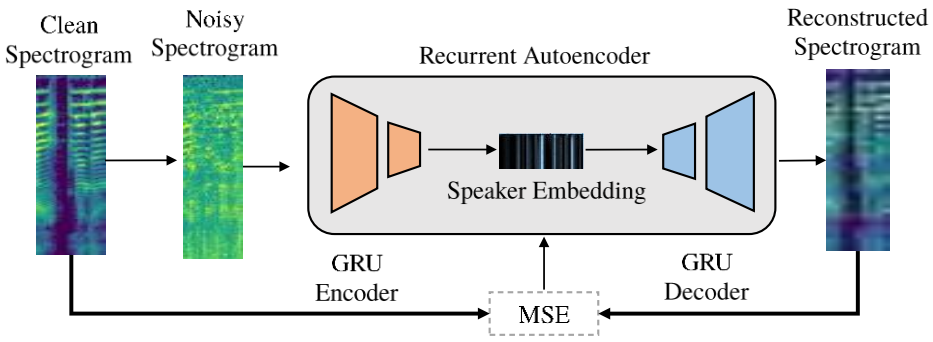
La variabilidad que las condiciones de la vida real inducen en el habla son un hándicap para los sistemas de reconocimiento de hablantes, por ejemplo, el estado emocional del hablante, el ruido ambiental, ... Éste habla se denomina "*in-the-wild*". Mediante los principios del aprendizaje máquina con *embeddings*, en esta sección pretendemos detallar nuestra propia contribución publicada en [2], sobre el diseño de un *autoencoder* recurrente capaz de eliminar el ruido que extrae representaciones o *embeddings* robustas del hablante a partir de espectrogramas con ruido de señales de habla para realizar la identificación del hablante.

Abordamos el problema combinado de la falta de robustez frente al ruido ambiental de los sistemas de SR y los efectos del habla emocional negativa en su rendimiento. Nuestra contribución capitaliza el uso de *embeddings* robustos orientados al reconocimiento de hablantes extraídas de un *auto-*

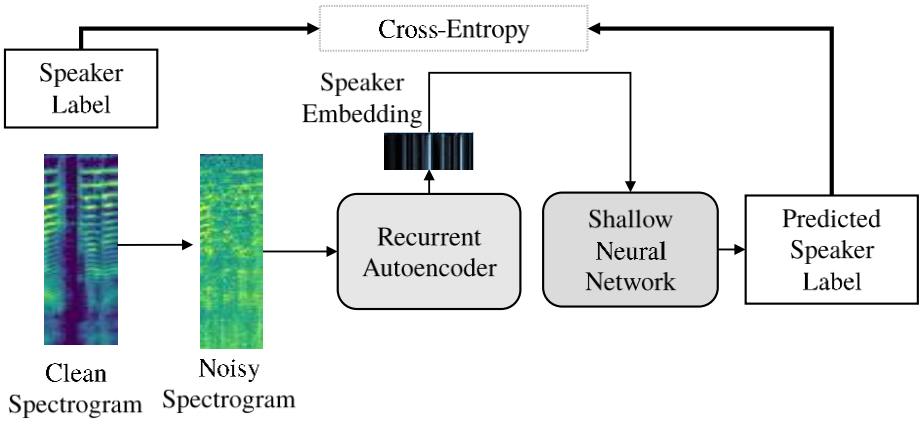
encoder recurrente capaz de eliminar el ruido, combinado con una red neuronal superficial -una red neuronal *feed-forward*, equivalente a un perceptrón multicapa (MLP)- que actúa como clasificador *back-end* para la tarea de identificación del hablante, como se detalla en la Fig. 4.5. Esta arquitectura de extremo a extremo está diseñada para trabajar en condiciones adversas, tanto desde el punto de vista de la distorsión del habla debida a situaciones de estrés, como del ruido ambiental.

Elegimos la base de datos VOCE Corpus [156] porque incluye habla grabada en condiciones de estrés espontáneo y debido a su naturaleza real. Además, aumentamos nuestra base de datos con señales ruidosas sintéticas contaminando aditivamente el conjunto de datos con ruido ambiental para emular el habla grabada en entornos reales.

Analizamos una arquitectura recurrente de *auto-encoder* capaz de eliminar el ruido basada en neuronas *gated recurrent units* (GRU), en la que la arquitectura recurrente tiene como objetivo modelar el contexto temporal de los enunciados del habla. El *encoder* extrae *embeddings* a nivel de muestra de los espectrogramas del habla y se optimiza conjuntamente con una red *feed forward* cuya capa de salida calcula los *posteriors* de la clase de hablante. Con la ayuda del módulo de eliminación de ruido, que intenta eliminar la información sobre el ruido ambiental, y de la SNN, cuyo objetivo es reconocer al hablante, toda la información que no se emplea directamente para la identificación de los hablantes se descarta. En concreto, la función de pérdida asociada a esta última red densa también se introduce en el *autoencoder* de eliminación de ruido para guiar sus esfuerzos hacia la tarea de SR.



(A) Limpieza del ruido del espectrograma



(b) Identificación del hablante

Figura 4. 5: Componentes de la arquitectura propuesta: Recurrent Denoising Autoencoder y red neuronal superficial [2]. Reproducido con permiso del propietario del copyright, Springer Nature.

Por último, observamos que estas representaciones orientadas a la discriminación de hablantes son más robustas al ruido y a la variabilidad que las optimizadas por separado. También comparamos los efectos de los *embeddings* extraídos automáticamente por esta arquitectura conectada en dos etapas frente a los dos módulos por separado, junto a las características extraídas manualmente que anteriormente demostraron ser adecuadas para este problema y una alternativa recurrente en frecuencia obtenida transponiendo las entradas al autocodificador GRU.

La principal diferencia con respecto a trabajos similares como los mencionados en la sección 4.2 -en particular [204]- consiste, en primer lugar, en el enfoque poco profundo del *back-end* orientado a disponer de un sistema de ejecución rápida y en tiempo real en un dispositivo *wearable*, buscando un equilibrio entre complejidad computacional y rendimiento; y, en segundo lugar, en el uso de un sistema de extremo a extremo para extraer *embeddings* que contengan únicamente información relevante sobre el hablante para la tarea de identificación.

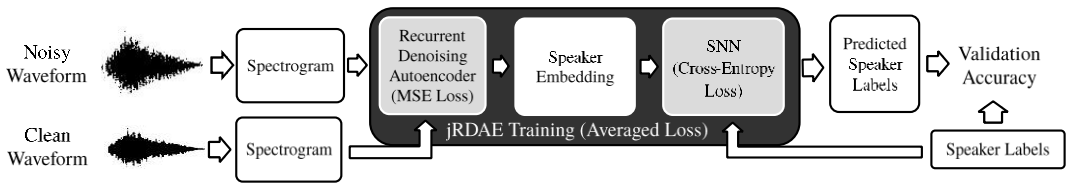
4.4.1. Arquitectura del Modelo

La arquitectura propuesta es la combinación de un *auto-encoder* recurrente de eliminación de ruido (RDAE) y una red neuronal superficial totalmente conectada (SNN) -que equivale a un MLP- en un sistema de extremo a extremo.

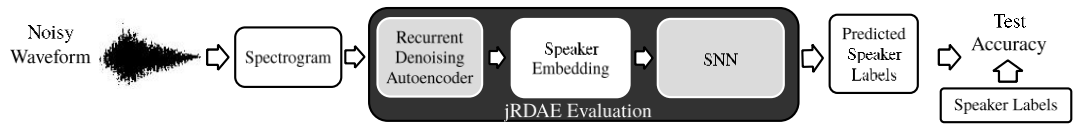
Los *auto-encoder* son, por lo general, algoritmos de aprendizaje automático sin supervisión entrenados para reconstruir sus entradas a través de una serie de capas. Los *auto-encoder* de eliminación de ruido (DAE) toman una versión ruidosa de los datos como entrada y una versión limpia como salida deseada e intentan reconstruir la segunda a partir de la primera. Nuestro RDAE propuesto se compone de un codificador de dos capas y un decodificador simétrico basado en GRUs. La SNN incluye una capa densa y una capa de salida.

Un *auto-encoder* es un modelo matemático entrenado en datos no etiquetados y utilizado para convertir los datos de entrada en una representación de características comprimida (también llamadas representaciones o *embeddings*) en el llamado cuello de botella o *bottleneck*, y luego convertir esa representación de características, de nuevo a la dimensión de los datos de entrada. En nuestro caso, el codificador toma como entrada un mel-espectrograma de un segundo en escala logarítmica y lo codifica en una representación de baja dimensión. Aunque los sistemas de SI tienden a utilizar ventanas temporales más largas para asegurar sus decisiones, Bindi necesitaba un resultado más rápido y en tiempo real, lo que ha motivado esta arquitectura de identificación de hablantes *de corta duración*.

Tras su extracción, el embedding alimenta simultáneamente al descodificador y a la SNN (véase la Fig. 4.6). En primer lugar, el descodificador intenta reconstruir un espectrograma limpio a partir de esta representación extraída de un espectrograma ruidoso obteniendo el error cuadrático medio (MSE) entre el espectrograma reconstruido y el limpio. En segundo lugar, la SNN encargada de identificar al hablante al que pertenece esa muestra de habla calcula la entropía cruzada (*cross-entropy*) de las etiquetas del hablante predicho y del hablante real.



(A) Fase de entrenamiento y optimización conjunta



(B) Fase de prueba/inferencia/*test*

Figura 4. 6: Esquema de fases de entrenamiento y test en la arquitectura propuesta en [2]. Reproducido con permiso del propietario del copyright, Springer Nature.

Las ecuaciones 4.1 y 4.2 representan las funciones de pérdida, L_d y L_s, del RDAE (error cuadrático medio) y del SNN (entropía cruzada) respectivamente

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

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

donde s es el espectrograma limpio, $\hat{s}$ el espectrograma reconstruido a partir del ruidoso, e y e $\hat{y}$ son las etiquetas de hablante original y predicha. N representa el número total de muestras de habla. Por último, en lugar de entrenar secuencialmente el RDAE y el SNN, toda la arquitectura

se optimiza conjuntamente utilizando una función de costes ponderada que combina linealmente las dos métricas anteriores en la ecuación 4.3.

$$L_T = \lambda * L_d + (1 - \lambda) * L_s \tag{4.3}$$

Hemos observado empíricamente que la normalización de los espectrogramas da como resultado una pérdida MSE normalizada que cae aproximadamente dentro del mismo rango dinámico que la pérdida de entropía cruzada. Dado que no teníamos ninguna razón *a priori* para pensar que una de las tareas pudiera influir en el resultado más que la otra, fijamos $\lambda = 0{,}5$. Esto dio buenos resultados en nuestra prueba, pero habría que seguir explorando este parámetro como trabajo futuro.

4.4.2. Aumento de Datos

En esta experimentación se utiliza el corpus VOCE [156] ya que, en primer lugar, contiene datos de habla en condiciones reales de estrés y, en segundo lugar, ofrece datos de sensores similares a los presentes en Bindi. Para simular entornos reales, las señales de voz se contaminaron aditivamente con 5 ruidos diferentes de *-5dB* a *20dB* en pasos de 5dB de relación señal/ruido (SNR). Las señales de ruido se eligieron del conjunto de datos DEMAND [235]: *DWASHING*, *OHALLWAY*, *PRESTO*, *TBUS*, *SPSQUARE* y *SCAFE*. Los ruidos se eligieron para emular condiciones de la vida cotidiana similares a las previstas para el funcionamiento de Bindi. Los ruidos se filtraron con un filtro de frecuencia *paso alto* para eliminar las frecuencias inferiores a *60* Hz y eliminar así las interferencias de la línea eléctrica, especialmente notables en el ruido *DWASHING*.

Utilizamos una ventana FFT de 70 *ms*, un solapamiento del 50% y 140 bandas de frecuencias mel y extrajimos los espectrogramas de las señales de voz para cada segundo de audio utilizando el módulo de extracción de espectrogramas de [236, 217], con lo que obtuvimos 27 pasos temporales y 140 mel-espectrogramas de bandas de frecuencias mel. Estas elecciones resultaron razonables durante una evaluación preliminar. Nuestra elección de un mayor número de bandas de frecuencias mel y de ventanas temporales más largas que las elegidas habitualmente en la extracción manual de características permite un equilibrio de la resolución de frecuencia y tiempo más adecuado para las redes recurrentes. Aunque las elecciones clásicas de estos valores se inspiran en el sistema auditivo humano, nuestra hipótesis es que las máquinas podrían aprovechar su potencia computacional al analizar datos más allá de lo que los humanos pueden oír y, por tanto, podrían superar la tasa de error humana si se aportan datos suficientes.

4.4.3. Configuración Experimental y Resultados

Para medir la robustez del sistema diseñamos un escenario de *multi-condicionamiento (multiconditioning)* en el que se combinan todas las señales de habla contaminadas a diferentes SNR, así como las señales de habla limpias. Se trata de un escenario más realista en el que la SNR específica no se fija *a priori* para cada entrenamiento. Se prestó especial atención a que todas las muestras pertenecientes a la misma configuración de ruido pero contaminadas SNR diferentes se agruparan en el mismo conjunto de datos de validación, de modo que ninguna de las distintas versiones de las muestras del subconjunto de validación apareciera en el conjunto de entrenamiento.

Se utilizó la validación cruzada anidada (*nested cross-validation*) para optimizar los hiperparámetros del *auto-encoder* y del SNN como clasificador de hablantes. En la validación cruzada anidada, se utiliza un bucle exterior del 33% de datos no vistos en la fase de entrenamiento para obtener los resultados finales del test. Se utiliza un bucle interior (3 conjuntos de validación) para encontrar los hiperparámetros óptimos para el modelo mediante la búsqueda en cuadrícula (*grid search*). El conjunto de test no es visto por el modelo. En la Fig. 4.6 se detalla un diagrama de bloques de las fases de entrenamiento y test.

Los espectrogramas se reducen en el eje de frecuencias de dimensiones 27×140 a 27×40. Esta imagen de baja dimensión se aplana, obteniendo un *embedding* de hablante unidimensional de 1080. Los tamaños de las capas de la arquitectura se muestran en la tabla 4.4.

El número de neuronas ocultas de la capa densa de la SNN se fijó en 1.000, el porcentaje de test en 30% y el parámetro de regularización L2 en 0,01. Entrenamos durante 15 *epochs* con un tamaño de *batch* de 128 y una probabilidad de aprendizaje *(learning rate)* de 0,001. También añadimos un retardo al criterio de parada, una paciencia de 5 iteraciones, tras las cuales, si no se observan mejoras, se detiene el entrenamiento. El modelo con menor pérdida de validación durante el entrenamiento se selecciona como el óptimo. Los espectrogramas se normalizaron con respecto a la media y la desviación estándar de su conjunto de entrenamiento. Cada espectrograma del conjunto de validación se normalizó con respecto a la media y la desviación estándar obtenidas de su correspondiente conjunto de entrenamiento.

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)

TABLA 4.4: Dimensiones de salida de las capas del autocodificador y arquitectura SNN back-end. Codificador (izquierda), decodificador (centro) y SNN (derecha) [2].

Comparamos el rendimiento de nuestro método optimizado *conjuntamente* propuesto (jRDAE) frente a tres arquitecturas diferentes. En primer lugar, el mismo sistema que el nuestro en el que el RDAE y el SNN back-end se han optimizado *de forma independiente* (iRDAE). En segundo lugar, un *auto-encoder denoising* recurrente transpuesto que difiere de nuestro enfoque en que los espectrogramas utilizados como entrada están *transpuestos*, así como las capas GRU, y es el eje temporal el que se reduce en dimensionalidad. Con ello se pretende modelizar de forma recurrente el dominio de la frecuencia. Por último, un sistema en el que las características extraídas manuamente, como la frecuencia fundamental, los formantes, los MFCC y la energía, se eligieron basándose en la bibliografía [3], se introducen directamente en el componente back-end SI, siendo un único módulo que hay que entrenar.

Nuestros resultados se muestran en la Fig. 4.7, donde también se representan los intervalos de confianza para cada uno de los resultados tomados como una desviación estándar en la validación triple. Como métrica para comparar los algoritmos, elegimos la precisión (*accuracy*) en términos de identificación del hablante, ya que las clases estaban bastante equilibradas. Para cada experimento, el intervalo de confianza se muestra como un pequeño *diagrama de bigotes* que representa la desviación estándar de los experimentos de validación cruzada realizados para indicar su importancia estadística. Nuestro objetivo es lograr robustez y, por tanto, obtener un rendimiento mejor cuando la SNR es baja.

La arquitectura en cascada optimizada de forma independiente (iRDAE) es el algoritmo que obtiene los resultados más bajos en todas las SNR (con la excepción del ruido *OHALLWAY* en SNR inferiores a *10* dB, donde es el segundo peor). Podemos concluir que la optimización del RDAE únicamente, hacia la minimización del MSE no es coherente con las necesidades de la SI. La arquitectura transpuesta es el resultado de tomar el eje de los espectrogramas transpuesto y, por tanto, de reducir el eje temporal en el *auto-encoder* en lugar del frecuencial. Como puede verse en los gráficos, esto da lugar a una detección inexacta del hablante. Creemos que reducir las características temporales secuenciales de los espectrogramas es una desventaja para el sistema SI.

Las características extraídas a mano (*handcrafted, HC*), por su parte, consiguen buenos resultados para SNR altas, ya que las características se eligieron específicamente para la tarea. Las HC funcionan aceptablemente bien cuando se dispone de una pequeña cantidad de datos, pero su rendimiento empeora muy rápidamente cuando disminuye la SNR.

Para la mayoría de los ruidos, la arquitectura propuesta (jRDAE) consigue los mejores resultados para las SNR más bajas y tasas estables para las más altas. jRDAE consigue resultados fiables para toda la gama de SNR, siendo un enfoque más robusto que el resto de las arquitecturas. La excepción es el ruido *PRESTO*, en el que un examen más detallado reveló que los espectrogramas que se habían reconstruido estaban bastante lejos de los limpios.

Además, dividimos los resultados para el sistema jRDAE propuesto (tabla 4.5) para observar las diferencias de rendimiento para las muestras *neutras* (N) y *estresadas* (S), donde en las dos últimas columnas de la tabla 4.5 se ofrecen los valores *medios* y *std* a modo de resumen. Claramente, se observaron índices de SI más bajos en el habla estresada, lo que muestra las dificultades inducidas por el estrés, siendo *PRESTO* y *SCAFE* los más afectados. Esto sugiere la necesidad de atender específicamente a las distorsiones causadas por el habla emocional.

Ruido \ SNR		-5	0	5	10	15	20	Limpio	**Media**	**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	-	-

TABLA 4.5: Resultados de precisión detallados por ruido aditivo y SNR, estratificados por muestras Estresadas (S) y Neutras (N) para el jRDAE propuesto [2].

En la Fig. 4.8 mostramos un desglose de los resultados en términos de habla neutra y estresada, comparando el enfoque HC y el jRDAE propuesto. En ella podemos observar un deterioro similar en los casos estresados para todos los ruidos aditivos. Concretamente para el modelo HC, que sigue una tendencia similar a la de los resultados de la tabla 4.5. También se representan los intervalos de confianza para cada uno de los resultados tomados como una desviación estándar en la validación triple. Los valores *std* denotan la media de los valores *std* del proceso de validación triple para los 6 SNR. Los resultados de precisión en estrés son ligeramente peores que los neutros, y para SNRs más bajos, los resultados son notablemente peores que para jRDAE.

Salvo algunas excepciones no significativas, observamos mejores resultados para el habla neutra, mientras que para el habla estresada la SR logra tasas de precisión más bajas, para ambos enfoques - HC y jRDAE -. Esto pone de manifiesto que el estrés afecta al habla y deteriora las tasas de reconocimiento del hablante a pesar de haber incluido esta degradación concreta en el conjunto de entrenamiento. En esta sección, no utilizamos las etiquetas de estrés frente a neutro para trabajar activamente en la lucha contra el estrés o para reducir sus efectos, por lo que creemos que aún hay margen de mejora.

4.4.4. Discusión

En esta sección evaluamos el rendimiento de los *embeddings* orientados al hablante extraídos con una arquitectura de extremo a extremo compuesta por un *auto-encoder* de eliminación de ruido recurrente para una tarea de SR utilizando una red neuronal superficial. Con este enfoque, pretendíamos mitigar los efectos en los sistemas de SR causados por la variabilidad inducida por el ruido ambiental, tanto para el habla neutra como para la estresada.

La arquitectura de extremo a extremo propuesta utilizó un bucle de realimentación para codificar la información relativa al hablante en *embeddings* de baja dimensión extraídas por un *auto-encoder* de eliminación de ruido utilizando espectrogramas. Empleamos técnicas de aumento de datos corrompiendo aditivamente el habla limpia con ruido ambiental real en una base de datos que contenía habla real estresada. Nuestro estudio demostró que la optimización conjunta de los módulos de *denoising* e identificación del hablante superaba -especialmente en SNR más bajas- a la optimización independiente de ambos componentes bajo distorsiones de estrés y ruido, así como al uso de características extraídas a mano.

Nuestra arquitectura jRDAE propuesta consigue resultados estables para toda la gama de señales SNR contaminadas, siendo un enfoque más robusto que el resto de arquitecturas probadas. En las tablas resultantes, se observaron tasas de SI más bajas al realizar la inferencia en habla estresada, lo que demuestra las dificultades inducidas por el estrés. Esto sugiere la necesidad de tener en cuenta específicamente las distorsiones causadas por el habla emocional.

En cuanto al coste computacional del sistema, hay que tener en cuenta que se espera que este módulo de identificación de hablantes esté integrado en un dispositivo con limitaciones computacionales, y para ello se prefieren los sistemas ligeros, con el fin de aumentar la duración de la batería. La decisión de utilizar capas GRU en lugar de capas LSTM se basó en que el número de parámetros es significativamente menor y, por tanto, las GRU son rápidas y menos costosas computacionalmente que las LSTM. Con esta decisión, el principal cuello de botella de velocidad es ahora la SNN, con 1,1 millones de parámetros. En el futuro, pretendemos reducir el número de parámetros de este modelo para desarrollar un algoritmo inteligente ligero que se pueda integrar en el sistema Bindi.

Para analizar más a fondo la robustez de estos *embeddings* orientados al hablante y de la arquitectura de extremo a extremo, podría probarse en una estrategia *adversarial* utilizando un clasificador de emociones -o estrés- como módulo *adversarial* de dominio. También tenemos la intención de utilizar conjuntos de datos más ricos que contengan habla de la vida real, concretamente WEMAC [11]. Para hacer frente al problema de la escasez de datos, se podrían utilizar modelos de transferencia y adaptar al habla emocional otros conjuntos de datos a gran escala para la identificación de hablantes, como VESUS [238] y VOXCeleb [239].

4.5. Respuesta de los Modelos de Reconocimiento de Hablante frente a Eventos Acústicos

En la línea del reconocimiento de hablantes en condiciones de estrés, realizamos adicionalmente algunos experimentos de SR con otra base de datos de habla que incluye condiciones de estrés reales. El trabajo que se detalla en esta sección está publicado en [8], de cuya contribución somos responsables, con la ayuda de otros miembros del equipo UC3M4Safety.

Determinamos que la base de datos BioSpeech (BioS-DB) [157] se ajustaba bien a nuestros intereses entre otras bases de datos adecuadas, ya que incluye etiquetas en tiempo continuo en el espacio *arousal*-valencia para el habla estresada no actuada realista debido a sus condiciones de habla en público. También incorpora datos fisiológicos (pulso de volumen sanguíneo, BVP, y Conductancia de la piel, SC) iguales que los que captura Bindi, y disponer de ellos podría ser de gran utilidad para los modelos multimodales en el futuro. Nótese, sin embargo, que el objetivo de esta base de datos es muy diferente al nuestro, ya que sus autores pretendían predecir bioseñales a partir de la señal del habla.

En general, la principal dificultad de los datos etiquetados emocionalmente radica en el proceso de etiquetado adecuado (véase la sección 2.5). No existe un acuerdo universal sobre cómo categorizar o medir las emociones. Las etiquetas anotadas por la propia persona (autoevaluaciones o autoetiquetas) pueden ser diferentes de las etiquetas anotadas por evaluadores externos que observen a dicha persona. Además, en la sección 3.2.2, ya introdujimos nuestra propia reinterpretación del etiquetado de BioS-DB, más adecuada para una tarea de clasificación del habla estresada (más información en la sección 5.6.1).

Por otra parte, el estado emocional de la persona podría influir negativamente en el rendimiento de cualquier tecnología del habla y, en particular, en la identificación del hablante [240]. Identificar la voz de la usuaria objetivo, separándola del resto de los hablantes, abre una posibilidad interesante para situaciones en las que sería deseable identificar a todos los hablantes implicados en la escena, por ejemplo, en caso de que se requieran pruebas legales.

La creación de Biospeech+ (véase la sección 3.2.3) surge de la necesidad de averiguar si podíamos utilizar la detección o clasificación de eventos acústicos (AED/C) de los sonidos o ruido de fondo para ayudar a las tareas de reconocimiento del hablante (SR) y de las emociones en el habla (SER). Queremos investigar si los eventos acústicos podrían causar con mayor probabilidad una reacción estresante ligeramente sincronizada con los instantes

de tiempo en los que las etiquetas emocionales denotan una aparición de estrés agudo. Para Biospeech+ los combinamos en diferentes relaciones SNR[28] (-5, 5 y 15 dB).

4.5.1. Configuración Experimental y Resultados

Extrajimos características de habla de librerías muy conocidas utilizadas para SR, SER y AED/C, respectivamente: librosa [241], eGeMAPS [184] del conjunto de herramientas openSMILE [185] y *embeddings* de YAMNet [242], como explicaremos más adelante. El tamaño de nuestra ventana de trabajo es de 1 segundo. Se trata de un compromiso entre la complejidad computacional y la velocidad, como requisitos para Bindi. Así, de librosa extrajimos 19 características con un tamaño de ventana de *20* ms y un solapamiento de *10* ms y luego su media y sus desviaciones estándar cada segundo, lo que dio como resultado 38 características por segundo. Con openSMILE extrajimos el conjunto de características eGeMAPS con 88 características *low-level*. Para extraer las características adecuadas para los eventos de audio utilizamos los *embeddings* de 1024 dimensiones correspondientes a las activaciones de la capa convolucional superior de YAMNet. Utilizamos un método de selección de características en el que se utilizó la correlación de la concatenación de los tres conjuntos de características para eliminar las características con una correlación superior al 95%. El resultado fue una reducción del 68% de las características. Examinando las matrices de correlación confirmamos que la mayoría de las características de YAMNet estaban muy correlacionadas entre sí. Todas las características se estandarizaron utilizando la normalización z-score.

Con el tamaño de ventana elegido, BioS-DB contiene aproximadamente 5000 muestras. Se trata de un conjunto de datos de tamaño pequeño para el uso de redes neuronales profundas, por lo que probamos una red sencilla de perceptrón multicapa (MLP) implementado con scikit-learn [243] y dos arquitecturas de red poco profundas implementadas con Keras [244], que trabajan para mantener una complejidad computacional baja. La primera de ellas consta de dos capas ocultas totalmente conectadas con 50 y 20 neuronas, respectivamente. La segunda es una combinación de una capa convolucional 1D, una capa bidireccional Gated Recurrent Unit (GRU) y una capa *fully connected*. Este modelo responde a la idea de que es importante extraer información de la distribución del contexto temporal de las características. Los modelos en Keras se compilaron utilizando un optimizador Adam y una tasa de aprendizaje de 0,001, además de usar la entropía cruzada categórica como función de pérdida. Todos los modelos utilizaron la F1-score como métrica de evaluación del rendimiento, ya que esta métrica tiene en cuenta los desequilibrios del

[28] Para la medida SNR consideramos el habla en primer plano de Biospeech como la "señal" y los eventos de audio como "ruido".

conjunto de datos. Para todos los experimentos utilizamos una validación cruzada de 5 *folds* (particiones).

Modelo	librosa	*p*	eGeMAPS	*p*	yamNET	*p*	L+E+Y	*p*	feat sel	*p*
Reconocimiento de hablantes										
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*

TABLA 4.6: Resultados de F1 para el reconocimiento de hablantes en BioSpeech limpio [8]. Se muestran los resultados de la media y la desviación estándar para una validación quíntuple (5-fold).

Los resultados para Biospeech sin los eventos de audio se muestran en la tabla 4.6. En ella, MLP se refiere al Perceptrón Multicapa, K2D se refiere al modelo de 2 capas densas en Keras y KCGD se refiere al modelo de Keras compuesto por una GRU bidireccional, una capa convolucional 1D y una capa densa. Al utilizar la base de datos Biospeech, las muestras de hablantes no están equilibradas por igual –es decir, que cada hablante tiene un número distinto de muestras -- y pretendemos utilizar la métrica de F1-score, ya que es muy utilizada habitualmente en modelos de inferencia para datos desequilibrados. Esta métrica la utilizaremos en lugar de la métrica de precisión (*accuracy*) como hacemos cuando utilizamos la base de datos VOCE. En la tabla 4.6, *p* representa el número de parámetros de cada modelo. Para las tres tareas consideradas, MLP con librosa consigue el mejor rendimiento. Cabe destacar que las características de librosa alcanzan la máxima puntuación en la tarea SR.

Las diferencias de rendimiento entre las características pueden deberse a múltiples razones, por ejemplo, su naturaleza. Las características de librosa y eGeMAPS se extraen manualmente, mientras que las de YAMNet se extraen automáticamente de una red de detección de eventos sonoros pre-entrenada. También, su número -38, 88 y 1024 respectivamente-, y además, su potencial específico para representar emociones o información del hablante. Cabe destacar los resultados para eGeMAPS con el modelo K2D, que es el más ligero de los 3 modelos utilizados, da mejores resultados que el KCGD con más parámetros; y logrando además un rendimiento sólo inferior en un 10% al MLP, teniendo 1/6 de sus parámetros.

La Fig. 4.9 ofrece los resultados de SR para Biospeech+ para diferentes SNR. Podemos observar de nuevo un rendimiento casi perfecto para las características de librosa, y buenos rendimientos para eGeMAPS, l+e+y (*early fusion* de características de librosa, eGeMAPS y YAMNet) y selección de características, pero una disminución considerable de la eficacia sólo para los *embeddings* de YAMNet. Esto significa que los eventos acústicos no afectan a la tarea de SR. Además, los embeddings de YAMNet no parecen captar información relevante sobre la información acústica del habla que podrían ayudar a distinguir entre hablantes.

4.5.2. Discusión

En esta sección nos propusimos evaluar si añadir eventos acústicos estadísticamente poco relacionados con la aparición de habla estresada podría mejorar el rendimiento de un sistema de SR. Partimos de la premisa de que detectar acústicamente situaciones de riesgo de violencia de género implica tener en cuenta el habla y los contextos acústicos, ya que podrían estar correlacionados. Sin embargo, no existen conjuntos de datos no actuados que contengan esta relación para poder estudiarla, por lo que generamos Biospeech+, aumentando BioS-DB con eventos acústicos.

Además, en la sección 3.2.2 reinterpretamos las etiquetas BioS-DB, incluyendo que las muestras etiquetadas como *Q2* se interpretaron como las relacionadas con el *miedo*, la ansiedad o el estrés. En este caso debíamos tener en cuenta que sin la dimensión de *dominancia*, emociones como la ira o la rabia podrían situarse también en ese cuadrante, generando una interpretación errónea de ambas.

En este estudio preliminar nos centramos en la relación entre el hablante, el estrés y los eventos acústicos. Tanto los conjuntos de características como los algoritmos se utilizaron con el objetivo de mantener baja la complejidad computacional y teniendo en cuenta el número de muestras de la base de datos utilizada. Y como conclusión, los eventos acústicos estresantes con una correlación no determinista con las muestras estresadas del habla demostraron no afectar en el reconocimiento del hablante.

En cuanto a los distintos métodos de extracción de características, las librerías ampliamente utilizadas para la extracción de características en tareas de reconocimiento de hablantes - librosa y eGeMAPS - funcionan mucho mejor que las características de YAMNet, que se utilizan para la detección de eventos acústicos. El hecho de que el conjunto de características de librosa alcanzara una F1-score del 100% se comprobó minuciosamente, ya que éramos conscientes de que una precisión tan alta podría denotar un problema de entrenamiento, pero no se encontró ningún error. Las posibles razones de esta puntuación del 100% podrían deberse a que se trata de un conjunto de características que funciona muy bien para la tarea de reconocimiento de hablantes, haciendo que el modelo aprenda muy bien los patrones de identidad de la voz, o al hecho de que la cantidad de datos del conjunto de test es limitada. Por último, en la sección 5.6.1 volveremos a utilizar esta base de datos para medir el efecto de los eventos acústicos en la tarea de reconocimiento del estrés, y avanzamos que, para dicha tarea, la presencia de eventos acústicos es beneficiosa.

4.6. Conclusiones y Trabajo Futuro en el Reconocimiento de Hablantes

En este capítulo hemos abordado la tarea de reconocimiento del hablante con la intención de identificar primero a la usuaria, para después detectar emociones en su voz que puedan indicar

una situación de riesgo. Nos centramos en dos aspectos de variabilidad, en primer lugar, en el reconocimiento del hablante en condiciones de estrés, para comprender cómo y cuánto afectan estas condiciones de estrés a la tarea de reconocimiento del hablante, y en segundo lugar en el reconocimiento del hablante en entornos reales (ruidosos), que es donde funcionará finalmente nuestra aplicación de Bindi.

En la sección 4.3 analizamos cómo influye el habla estresada en el rendimiento de los sistemas de SI e identificamos que sí repercute negativamente en las tasas de reconocimiento de hablantes cuando los sistemas de SI utilizan modelos ML entrenados sólo con habla neutra. Para ello, trabajamos con configuraciones de *match* y *mismatch*, y creamos datos de habla estresada generados sintéticamente para sustituir los datos del habla neutra, lo que amplía significativamente las muestras con las que trabajar, mejorando sustancialmente los resultados obtenidos por los sistemas. Por lo tanto, en ausencia de habla estresada emocional real -y, en última instancia, habla en condiciones de *miedo*- podemos aumentar los datos con mayores modificaciones del tono y ralentizaciones de la velocidad del habla para conseguir datos que se asemejen al estrés real y puedan ayudar a mantener una tasa de reconocimiento de hablante estable en los sistemas de SR.

En la sección 4.4 empleamos técnicas de aumento de datos mediante la corrupción aditiva del habla limpia con ruido ambiental de la vida real en una base de datos que contenía habla estresada real, para estudiar la relación entre estos 3 factores -hablante, ruido y emociones-. Partíamos de la premisa de que el habla 'en la naturaleza' (*in-the-wild*) es una desventaja para los sistemas de reconocimiento de hablantes debido a la variabilidad inducida por las condiciones de la vida real, como el ruido ambiental y el estado emocional del hablante. El diseño de un auto-encoder recurrente de extremo a extremo eliminador de ruido que extrae *embeddings* robustos sobre la identidad del hablante hablante a partir de espectrogramas del habla que contienen ruido, aborda el problema combinado de la falta de robustez frente al ruido ambiental de los sistemas de RS, incluso cuando incluyen el habla estresada para su entrenamiento. Este aprendizaje basado en *embeddings* o representaciones o *features* automáticas aprovecha la optimización conjunta de un *denoiser* y un bloque SI con una función de pérdida combinada, que se demuestra que funciona mejor que un *denoiser* aislado. Esta arquitectura consigue resultados estables para las señales contaminadas en todo el rango de SNR, siendo un enfoque más robusto que el resto de las arquitecturas probadas. Las dificultades de rendimiento de la arquitectura en el habla estresada se observan cuando se consiguen tasas de SI más bajas para las muestras de habla estresada que para las neutras. Esto sugiere la necesidad de seguir abordando específicamente las distorsiones causadas por el habla emocional.

En la sección 4.5 aumentamos los datos del habla con eventos acústicos aditivos, partiendo de la premisa de que detectar situaciones de riesgo implica tener en cuenta el habla y el contexto acústico, ya que podrían estar correlacionados. Pero los eventos acústicos estresantes con una correlación no determinista con las muestras de habla estresada demostraron no afectar negativamente en el reconocimiento del hablante. Así pues, la relación entre la detección del estrés y este tipo de eventos acústicos queda pendiente de estudio en el capítulo 5.Para terminar, la falta de datos de habla en condiciones emocionales reales es sin duda un inconveniente para los sistemas de SR, ya que sin ellos es difícil que los modelos ML alcancen las mejores tasas de reconocimiento -en condiciones de *match*-. También es muy importante tener en cuenta que el habla grabada en condiciones reales incluye ruido ambiental que también es perjudicial para los sistemas de SR, por lo que debe eliminarse con métodos correctos de eliminación de ruido para lograr las mejores prestaciones de SR. A falta de habla estresada real y de condiciones ruidosas con las que entrenar y hacer inferencia en nuestros modelos de SR, hemos descubierto que aumentar los datos estresándolos sintéticamente y añadiendo ruido ambiental nos permite estudiar y diseñar modelos de SR más robustos al estrés y al ruido.

WEMAC y WE-LIVE se crearon para dar respuesta a este problema, con el que pretendemos seguir trabajando y aplicar todos los conocimientos obtenidos en los estudios detallados anteriormente realizados en el campo de la SR para nuestra aplicación de detección de situaciones de riesgo de violencia de género. Para seguir analizando la robustez de los *embeddings* orientados al hablante, se podría probar un modelo de SR de con un entrenamiento *adversarial* utilizando un clasificador de emociones -o estrés.

El uso de técnicas de aumento de datos, junto con muestras de datos reales como las de nuestros conjuntos de datos -WEMAC y WE-LIVE- allana el camino para utilizar redes neuronales profundas más complejas en nuestro problema. Como trabajo futuro, el siguiente paso podría ser utilizar una red neuronal pre-entrenada (*pre-trained*) y ajustarla (*fine-tuning*) con los datos que tenemos disponibles en WEMAC. Para trabajar en la difícil tarea de la detección de situaciones de riesgo de violencia de género utilizando un dispositivo discreto como el Bindi, debemos tener siempre presentes sus limitaciones, ya descritas en los apartados 1.1.4 y 1.2.2. Sin embargo, conviene tener en cuenta que, con la rápida evolución de los dispositivos, es posible que en futuras versiones de Bindi podamos introducir redes neuronales más complejas y profundas, con mayor poder de inferencia y capacidad computacional.

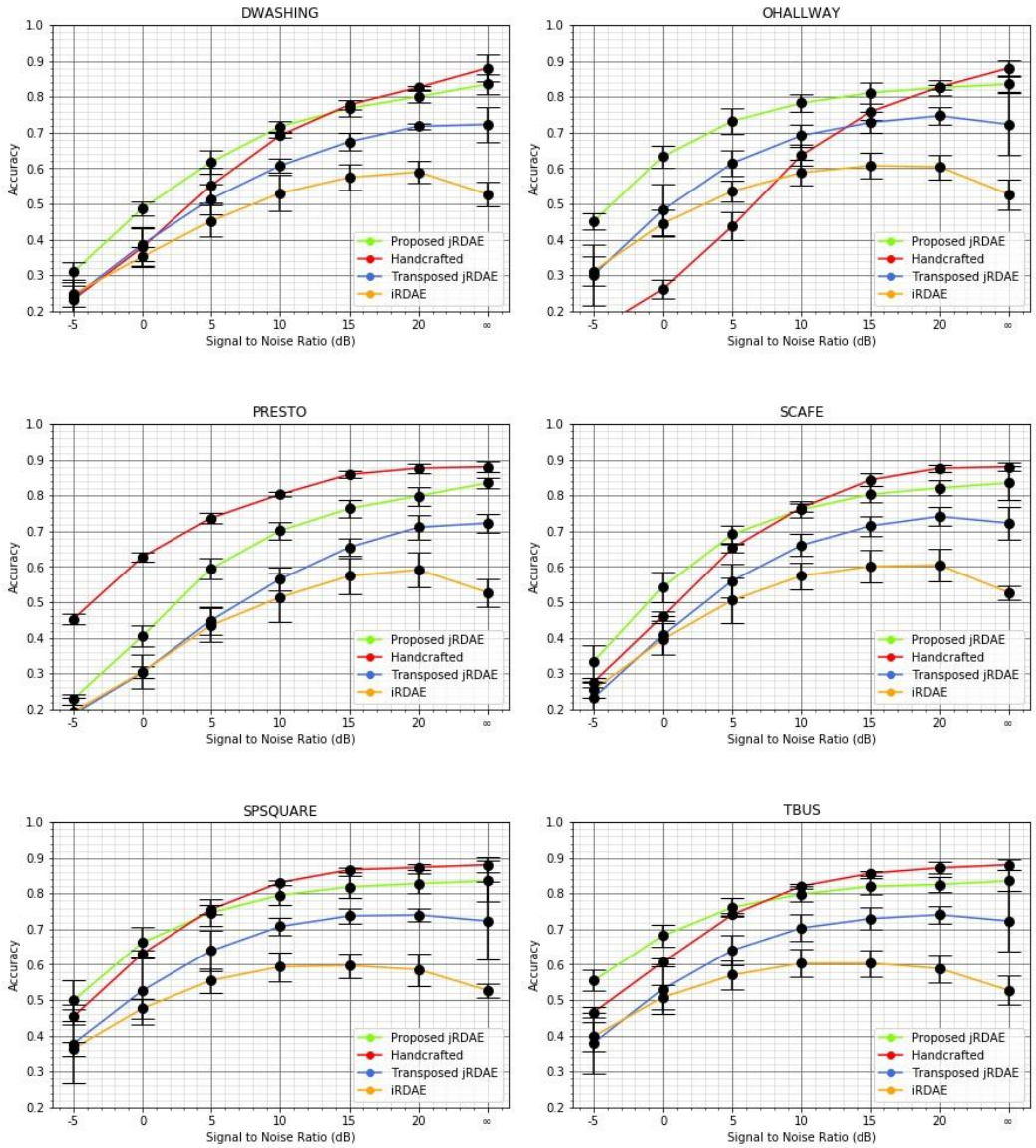
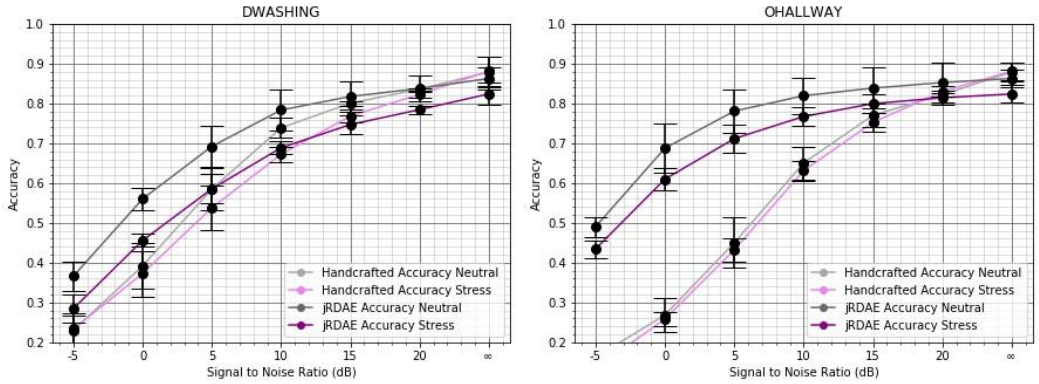


Figura 4. 7: Resultados detallados por ruido aditivo y SNR en términos de precisión para diferentes arquitecturas [2]. Reproducido con permiso del propietario del copyright, Springer Nature.



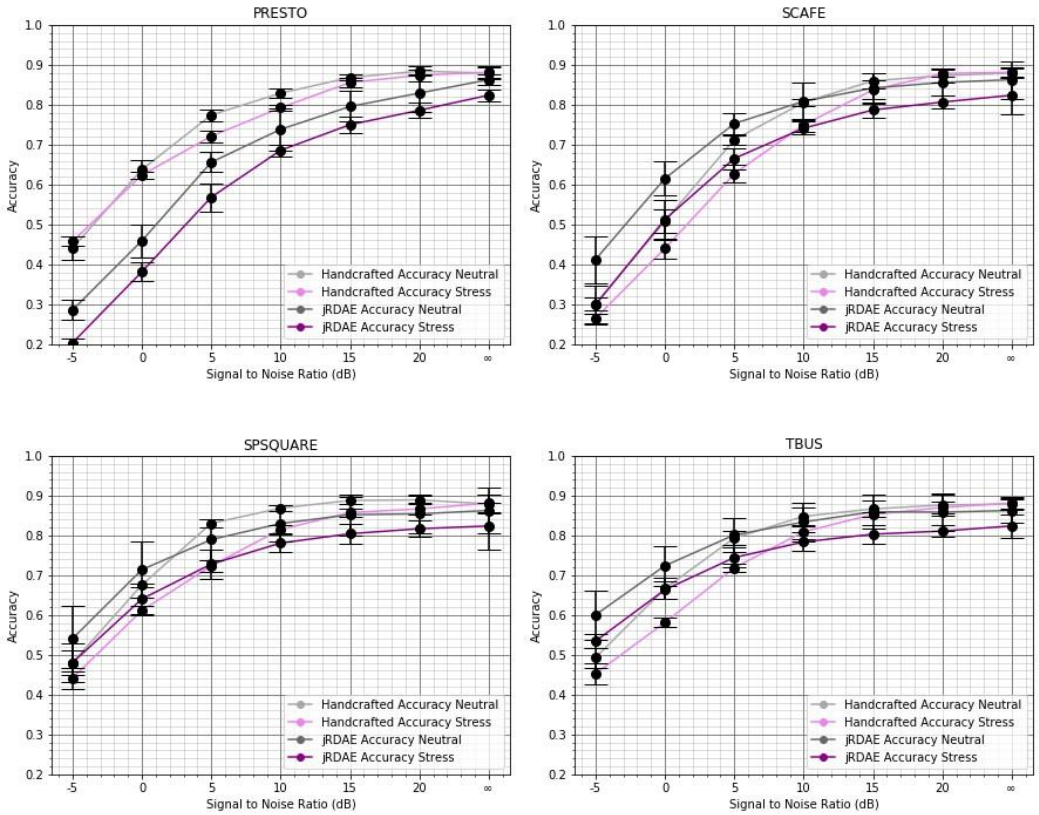


Figura 4. 8: Resultados detallados por ruido aditivo y SNR en términos de precisión para muestras de tensión y neutras para las configuraciones Handcrafted y jRDAE [2]. Reproducido con permiso del propietario del copyright, Springer Nature.

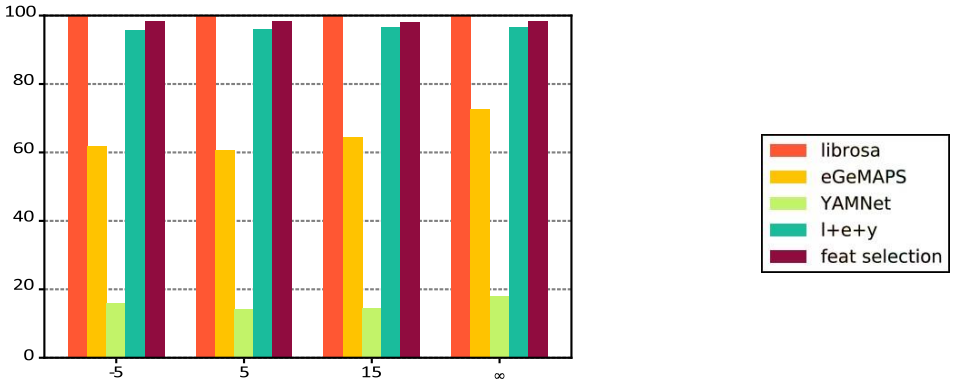


Figura 4. 9: Resultados de la F1-score del reconocimiento del hablante con un Perceptrón Multicapa (MLP) en Biospeech+ [8]. Reproducido con permiso del propietario del copyright, ISCA.

Capítulo 5: Sistema Multimodal de Reconocimiento de la Emoción del Miedo para Bindi

En este capítulo nos sumergimos en el desarrollo del sistema Bindi para el reconocimiento de emociones relacionadas con el *miedo*. Este capítulo es altamente multidisciplinar, ya que cuenta con numerosas contribuciones apoyadas por otras personasmbros del equipo UC3M4Safety. Este capítulo retoma y reproduce total o parcialmente el contenido de los artículos publicados en [9] y [1].

En primer lugar, describimos la arquitectura de hardware del sistema Bindi, desarrollando los componentes de los sistemas de computación de *edge*, *fog* y *cloud*. A continuación, explicamos el enfoque seguido para el diseño de las estrategias de fusión multimodal para el sistema de alerta automática para Bindi, primero un sistema multimodal en cascada para Bindi 1.0, y también, el despliegue de un sistema completo del Internet de las Cosas (*Internet of Things*, IoT) con componentes de computación de *edge*, *fog* y *cloud*, para Bindi 2.0. En este último, detallamos específicamente cómo diseñamos las arquitecturas de inteligencia en los dispositivos Bindi para la detección del *miedo* en la usuaria. Estas contribuciones se realizaron en colaboración con otros miembros del equipo.

Además, describimos los diferentes *pipelines* de procesamiento de datos (fisiológicos y del habla) y la experimentación monomodal con el habla para la detección de emociones relacionadas con el *miedo*, centrándonos primero en la detección del estrés realista -ya que los datos estaban en el momento disponibles en fácilmente en la literatura- [1].

Posteriormente, como núcleo de experimentación, conjuntamente con otros miembros del equipo UC3M4Safety, trabajamos con nuestra propia base de datos diseñada y capturada por el equipo -WEMAC [11]- para la tarea de detección del *miedo*. En este capítulo hay un fuerte componente multimodal, ya que trabajamos en la parte de reconocimiento de emociones a partir del habla junto con datos de señales fisiológicas, del mismo modo que lo harían los dos dispositivos *wearable* de Bindi.

Por último, en las últimas secciones se ofrece una discusión sobre la arquitectura de Bindi, los resultados y su discusión, seguidos de las conclusiones y las futuras líneas de investigación.

La principal aportación de esta tesis se basa en el trabajo en la modalidad del habla, pero también aborda la multimodalidad y la fusión de modalidades. La tesis doctoral complementaria donde se encuentra con gran detalle todo el trabajo realizado sobre la modalidad fisiológica -y en particular sobre los componentes hardware de Bindi, concretamente la pulsera- fue desarrollada por el miembro del equipo UC3M4Safety José Miranda Calero, y puede encontrarse en [53].

5.1. Introducción

Como explicamos en el apartado 1.1.4, la solución tecnológica de Bindi mejora a los botones de pánico para detectar situaciones de riesgo, ya que estos últimos pueden causar dificultades porque

las víctimas tienen que utilizarlos activa y manualmente incluso en condiciones en las que pueden no ser capaces (estar en estado de shock, etc). En su lugar, Bindi, nuestro sistema multimodal totalmente autónomo, se basa en técnicas de inteligencia artificial que detectan automáticamente situaciones de riesgo, basándose en la detección de emociones relacionadas con el *miedo*, e inicia un protocolo de protección cuando es necesario. Para ello, "Bindi integra tecnologías modernas de vanguardia, como el *Internet of Things*, la computación afectiva y los sistemas ciberfísicos, reuniendo i) el IoT afectivo con sensores inteligentes comerciales auditivos y fisiológicos incrustados en dispositivos *wearable*, ii) la fusión jerárquica de información multisensorial, y iii) la arquitectura IoT *edge-fog-cloud*" [1].

Para la detección de situaciones de riesgo en el contexto de la violencia de género, nos basaremos en el reconocimiento de las emociones relacionadas con el *miedo* a partir del habla y las variables fisiológicas de una usuaria, entrando directamente en el campo de la SER.

El reconocimiento de las emociones del habla, abreviado como SER, es "la tarea de reconocer las emociones y los estados afectivos humanos a partir de su habla". Se basa en el hecho de que la voz suele reflejar las emociones que se sienten, y esto se hace a través de las características que se extraen de ella. Para el desarrollo de Bindi, nuestro objetivo es detectar emociones de *miedo* en la usuaria que sean consecuencia de una situación de riesgo para ella. En el capítulo anterior dimos el primer paso para reconocer emociones, que es el reconocimiento del hablante del que se desea reconocer la emoción. Y en este capítulo, nos centramos en la clasificación de las emociones utilizando variables del habla y fisiológicas -centrándonos en las primeras-, concretamente las relacionadas con el *miedo*. No existen en la literatura bases de datos adecuadas para esta tarea, que incluyan el habla en condiciones reales de *miedo*, por lo que primero trabajamos con el habla estresada, por ser un pariente cercano del *miedo,* y luego trabajamos con nuestra base de datos WEMAC, que se desarrolló para cubrir ese nicho. Con esos datos, pretendemos entrenar modelos de aprendizaje automático para detectar automáticamente el estado emocional de una usuaria, en particular, *el miedo*.

Hacemos uso tanto del habla como de las señales fisiológicas, ya que Bindi pretende ser un dispositivo discreto -para el reconocimiento de emociones- y dichas modalidades no son invasivas y pueden capturar datos en un entorno cotidiano de la vida real, que es donde pretendemos que funcione Bindi. Este tipo de dispositivos *wearable* interconectados pertenecen al campo del *Internet of Bodies* (IoB), un subcampo del *Internet of Things* (IoT). Con el uso de estas dos modalidades desarrollamos sistemas de reconocimiento de la emoción del *miedo* unimodales y multimodales.

Tipos de fusión de modalidades en el aprendizaje automático

La multimodalidad es un concepto natural para los seres vivos como medio de interacción con el mundo que nos rodea. En todos los individuos, la información adquirida procede de sensores internos

y externos. Esta información se combina y fusiona para proporcionar respuestas rápidas al entorno externo en constante cambio.

Centrándonos en el campo de la computación afectiva, manejar más de una modalidad supone un reto porque los datos difieren en diferentes aspectos, como puede ser el origen, la estructura y la relevancia. Sin embargo, la diversidad dentro de un sistema multimodal de reconocimiento de emociones (por ejemplo, la combinación de señales fisiológicas y auditivas) suele permitir mejorar los conocimientos de una forma que no puede lograrse con una sola modalidad [245]. En la literatura, existen cuatro técnicas principales para la fusión de datos: fusión a nivel de características, a nivel de decisiones, a nivel de modelos y fusión híbrida [246].

En la fusión a nivel de características o fusión temprana (también llamada *early fusion*), las diferentes características obtenidas de cada sensor de entrada se combinan en otro vector de características antes de la clasificación con un modelo de aprendizaje automático. El principal inconveniente de este método es la elevada dimensionalidad del vector de características resultante, que podría dar lugar a la conocida 'maldición de la dimensionalidad' *(curse of dimensionality)*.

A diferencia de la fusión temprana, la fusión a nivel de decisión o fusión tardía (*late fusion*) requiere múltiples etapas de entrenamiento, una por modalidad (por ejemplo, una etapa de entrenamiento sólo para señales fisiológicas y otra sólo para señales auditivas). Este mecanismo de fusión se basa en la combinación tardía de los resultados del reconocimiento unimodal mediante algún criterio. En este caso, cada una de las modalidades puede ser modelada con mayor precisión por sus clasificadores, pero el sistema no maneja en modo alguno las interacciones o correlaciones entre modalidades. La versión inicial de Bindi -Bindi 1.0- consideraba una técnica de fusión a nivel de decisión según los resultados de inferencia unimodal basados en datos fisiológicos y del habla.

Se pueden aplicar otras dos metodologías de fusión para abordar el problema de la interacción: los enfoques híbridos (*hybrid fusion*) y de fusión a nivel de modelo. Ambos combinan aspectos de las dos técnicas ya comentadas (fusión temprana y tardía). La fusión a nivel de modelo se basa en la correlación mutua entre los distintos flujos de datos procedentes de las modalidades del sistema. Normalmente se considera que explora la correlación temporal entre esos flujos [246]. La fusión híbrida implementa más de un nivel de fusión dentro del mismo sistema (por ejemplo, combinando enfoques a nivel de característica y a nivel de decisión), lo que suele proporcionar mejores resultados de reconocimiento que aplicando únicamente una técnica de fusión.

5.2. Trabajos Relacionados

El campo de investigación multidisciplinar destinado a reconocer las emociones humanas es la ya mencionada computación afectiva [75], [247]. Entre las aplicaciones de la AC podemos encontrar proporcionar mejores condiciones de trabajo, entretenimiento o servicios a las personas. No sólo se basa en sensores inteligentes y en el procesamiento digital de señales, sino también en técnicas de

IA, como el aprendizaje automático y profundo (ML y DL). La investigación colaborativa entre los campos de la psicología, la informática, los sensores inteligentes y las ciencias cognitivas [248] permite detectar diferentes estados emocionales mediante la monitorización de señales humanas, como las fisiológicas y las físicas. Algunos ejemplos de señales físicas incluyen el audio, el habla o la voz, las señales de imagen o vídeo, el seguimiento del fondo de la escena o de la usuaria. Algunos ejemplos de variables fisiológicas incluyen la frecuencia cardiaca (FC), la respuesta galvánica de la piel (GSR), la temperatura corporal (SKT), el electromiograma (EMG) y el electroencefalograma (EEG).

5.2.1. Perspectiva del Habla: Reconocimiento de las Emociones en el Habla (SER)

La detección de emociones se ha descrito ampliamente en la literatura con el uso de señales del habla [249], [250]. En los últimos años, el interés por detectar e interpretar las emociones en el habla es muy amplio [251], [252]. El reconocimiento de las emociones del habla (SER) consiste en la identificación del contenido emocional de las señales del habla, la tarea de reconocer las emociones humanas y los estados afectivos a partir del habla. En este campo, hay tres aspectos importantes que se estudian y debaten en la comunidad y la literatura del aprendizaje automático: i) la elección de características acústicas adecuadas [232], ii) el diseño de un clasificador apropiado [253] y, iii) la creación de bases de datos de habla emocional [254], [255].

En los apartados 3.1 y 3.2 se hace un repaso de las bases de datos existentes en la literatura que pueden ser adecuadas para esta tesis, así como de sus retos.

El reconocimiento de las emociones por voz tiene aplicaciones en la interacción persona-ordenador (*human-computer interaction,* HCI), así como en robots, servicios móviles, juegos de ordenador y evaluaciones psicológicas, entre otros. A pesar de sus numerosas aplicaciones y de los avances sustanciales debidos a la llegada de las técnicas de aprendizaje profundo [256], el reconocimiento de emociones sigue siendo una tarea difícil, sobre todo por la subjetividad que entrañan las emociones (véase la sección 2.5).

La falta de corpus de habla existentes con un *miedo* elicitado en situaciones reales es un problema particular de nuestra investigación concreta (véase la sección 3.1). Sin embargo, algunos estudios han conseguido resultados en este sentido. Por ejemplo, Clavel et al. [153] desarrollaron un sistema de detección de situaciones de emergencia o anómalas basado en audio para clips de películas. Sus resultados alcanzaron una precisión de hasta el 70,3% para la detección del *miedo* mediante una estrategia Leave One Trial Out (LOTO) para 30 películas. En [257], realizaron una detección de emociones con características paralingüísticas en un corpus de diálogos que contenía grabaciones reales agente-cliente obtenidas de un centro de llamadas de urgencias médicas. Como resultado, lograron una tasa de reconocimiento con una precisión de hasta el 64% para el reconocimiento del *miedo*.

5.2.2. Reconocimiento de Emociones mediante Señales Fisiológicas

Uno de los dispositivos Bindi es una pulsera inteligente (*smartband*) que puede capturar señales fisiológicas, ya que, para el diseño de un sistema de reconocimiento de emociones, la perspectiva fisiológica es extremadamente informativa. En la investigación sobre fisiología y emociones, el reconocimiento del *miedo* -entre otras emociones- no es nuevo [258].

Sin embargo, hasta donde sabemos en el momento de la publicación de esta tesis, sólo existen dos sistemas de reconocimiento del *miedo* basados únicamente en información fisiológica y etiquetas auto-anotadas.

Por un lado, los autores de [259] utilizaron todas las señales disponibles en la Base de Datos para el Análisis de Emociones mediante Señales Fisiológicas (*Database for Emotion Analysis using Physiological Signals*, DEAP) [260] para proporcionar un sistema especializado de reconocimiento del *miedo*. Consiguieron una tasa de detección precisa del *miedo* inferior al 90%, aunque también tuvieron en cuenta el EEG, que actualmente no es factible como dispositivo *wearable* discreto. Por otro lado, en una investigación anterior de otros miembros del equipo UC3M4Safety [261], sólo se utilizaron tres variables fisiológicas disponibles en el conjunto de datos *Multimodal Analysis of Human Nonverbal Behaviour in real-world settings* (MAHNOB) [262], obteniendo una tasa de precisión en el reconocimiento del *miedo* de hasta el 76,67% para un enfoque independiente de la persona (*speaker-independent*) utilizando datos de 12 mujeres voluntarias. En este último concluyeron que era necesario capturar un conjunto de datos novedoso centrado en la detección del *miedo*, que incluyera el uso de tecnología inmersiva, la consideración de la perspectiva de género, el logro de una distribución de estímulos equilibrada y adecuada con respecto a las emociones objetivo y un mayor número de participantes.

5.2.3. *Internet of Bodies*

El crecimiento de la investigación sobre dispositivos que monitorizan señales del cuerpo humano durante los últimos años implica una inminente ampliación del dominio de la Internet de las Cosas (IoT). Esta tendencia surge en relación con los dispositivos interconectados (por ejemplo, *wearable*, implantados, incrustados e ingeridos) situados dentro y alrededor del cuerpo humano formando una red, que actualmente se denomina *Internet of Bodies* (IoB) [263]. Este novedoso campo tiene muchas aplicaciones, como el reconocimiento de la actividad humana [264], la autenticación de usuarios [265], e incluso reconocimiento de emociones [266]. Este campo también abarca estudios esenciales sobre las limitaciones de dichos sensores, como los problemas de retardo temporal y consumo de energía [267]. Así, estos sensores en el cuerpo pueden adquirir diferentes tipos de información fisiológica al mismo tiempo, lo que deriva en estudios relacionados con el uso de técnicas de fusión de datos multimodales [268], [269].

Esta proliferación de IoB va acompañada de avances en las tecnologías de aprendizaje automático y aprendizaje profundo, lo que da lugar a una explosión de inteligencia móvil y plantea crecientes demandas de recursos informáticos que los dispositivos móviles *edge* no pueden satisfacer. En consecuencia, se están potenciando y explorando las capacidades de la computación de *edge* para ofrecer mejores servicios de inferencia de motores de inteligencia a las usuarias finales [270]. Por ejemplo, en [271] trabajaron en la aceleración del proceso de entrenamiento de grandes modelos de aprendizaje automático en IoT para hacer frente a las limitaciones del hardware.

Dentro de este contexto de IoB, los trabajos que se explican en los siguientes apartados pretenden aportar y fomentar la generación de técnicas novedosas y ligeras de fusión de datos multimodales alimentados por la monitorización del cuerpo humano hacia su aplicabilidad a los actuales dispositivos de *edge-computing*, como los de Bindi [1].

5.2.4. Técnicas de Fusión Multimodal

Como Bindi es un sistema multimodal -utiliza señales fisiológicas y del habla para detectar las emociones de la usuaria-, en esta sección hacemos un breve repaso de la teoría y las arquitecturas multimodales.

Algunos trabajos han propuesto enfoques multimodales que combinan datos visuales y del habla para mejorar y reforzar el reconocimiento de las emociones [272] [273]. Esta conceptualización no es posible en Bindi porque no hay componente visual. Por tanto, la información adicional procederá de variables fisiológicas. Dado que Bindi es un sistema multimodal que consta de una pulsera que capta señales fisiológicas (SKT, GSR, BVP) y un colgante que incluye un micrófono que capta señales de audio (eventos acústicos, habla) (más detalles en la sección 1.1.4), estas dos modalidades -fisiológica y auditiva- pueden trabajar conjuntamente para la detección del *miedo* y, en consecuencia, la detección de situaciones de riesgo.

Cuando se trata del reconocimiento de emociones combinando diferentes modalidades de datos, se pueden encontrar algunas revisiones exhaustivas que presentan el estado actual de las técnicas de fusión de datos en la literatura [274, 255]. Estos trabajos plantean la necesidad de 1) nuevos enfoques para avanzar en la comprensión de la casuística multimodal por parte de la comunidad, y 2) modelos de reconocimiento de emociones independientes de la persona (*speaker-independent*) para facilitar los despliegues posteriores en condiciones reales. También coinciden en las posibles mejoras de rendimiento con los enfoques multimodales en comparación con los unimodales.

De hecho, recientemente, la investigación en sistemas multimodales está en auge. Por ejemplo, los autores de [276] propusieron un sistema híbrido multimodal de fusión de reconocimiento de emociones que incluía expresiones faciales, GSR y EEG. Sus resultados arrojaron una precisión máxima por persona del 91,50% y una precisión media del 53,80% utilizando una estrategia de exclusión de un sujeto (*Leave One Speaker Out*, LOSO) y una base de datos de acceso público

(DEAP) para diferentes casos de uso de detección de emociones, como enfado, asco, *miedo*, felicidad, neutralidad, tristeza y sorpresa. Además, crearon su propio conjunto de datos con el que lograron una precisión máxima de la persona del 81,2% y una precisión media del 74,2% utilizando una estrategia LOSO para tres clases de emoción, que fueron triste, neutral y feliz. En [277], se aplicó una estrategia de fusión ponderada acompañada de técnicas de aprendizaje por transferencia para el reconocimiento multimodal de emociones mediante EEG y la detección de expresiones espaciales espontáneas. El trabajo empleó una configuración dependiente de la persona Leave-One-Trial-Out (LOTO) e informó de una precisión media de hasta el 69,75% y el 70,00% para la clasificación de la valencia y la *arousal*, respectivamente. Además de estos trabajos, se pueden encontrar más investigaciones sobre la fusión de datos multimodales para casos de uso relacionados con el estrés en [278], [279].

Analizando estos trabajos relacionados, la mayoría de los sistemas de reconocimiento de emociones no se dirigen a la fusión de las modalidades fisiológica y auditiva ni tienen en cuenta a los grupos vulnerables, como las GBVV. En lo que respecta específicamente a dicha fusión bimodal de información fisiológica y vocal, uno de los pocos trabajos que destaca es [280], hasta donde sabemos. Este trabajo consideró diferentes esquemas de fusión de datos y logró una precisión media de hasta el 55,00% para una estrategia independiente de la persona que utilizaba una fusión de características cuando se dirigía a una clasificación binaria de valencia y excitación. En consecuencia, existe una necesidad actual de investigación sobre estos temas, en la que este capítulo pretende profundizar.

5.3. Arquitectura Hardware del Sistema Bindi

Al principio de la sección 1.1.4 describimos Bindi, en esta sección profundizamos en el diseño de su arquitectura. En la Fig. 5.1 se presenta una arquitectura de sistema simplificada de Bindi. Las siguientes subsecciones ofrecen una visión técnica general de cada componente del sistema. Los dispositivos de borde de la arquitectura Bindi son la pulsera y el colgante.

Edge computing: Pulsera

Este dispositivo ejecuta un motor de inteligencia integrado para la detección del *miedo* basado en información fisiológica. La fig. 5.2 muestra los componentes de hardware integrados en este dispositivo, que pueden clasificarse en cuatro grupos: sensores fisiológicos, actuadores, elementos gestores de la energía y la unidad del microprocesador. Para más detalles sobre ellos, consulte [1]. Hay que tener en cuenta que el módulo de radiofrecuencia mediante comunicación Bluetooth Low Energy® también está integrado en esta unidad.

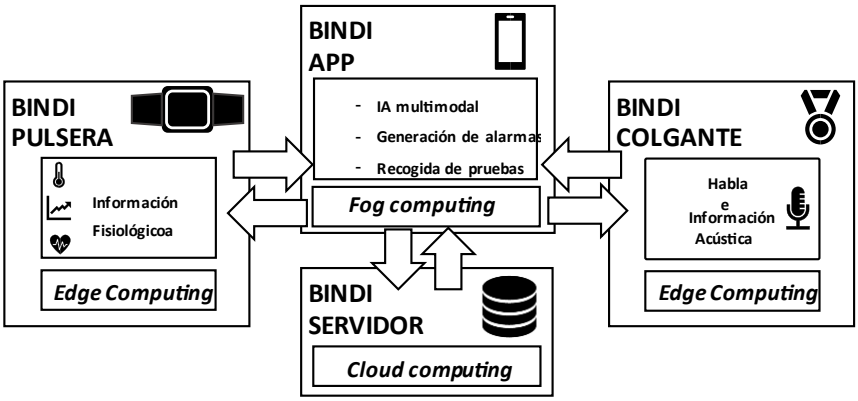


Figura 5. 1: Arquitectura del Hardware de Bindi simplificada [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

La pulsera está equipada con un botón electromecánico convencional para la activación manual por parte de la usuaria, que actúa como botón del pánico. Los sensores fisiológicos captan las siguientes variables:

- HR: Se basa en un sensor de fotopletismografía que detecta los cambios del pulso del volumen sanguíneo (BVP) midiendo la absorción de la luz emitida a través de la piel.
- GSR: Este sensor utiliza dos electrodos para medir la conductividad de la piel a través de una medición exosomática de corriente continua.
- SKT: Este circuito integrado se define como un sensor de grado clínico para aplicaciones *wearable*, que proporciona una precisión de ±0,1 °C en un rango de temperatura de 30 °C a 50 °C.

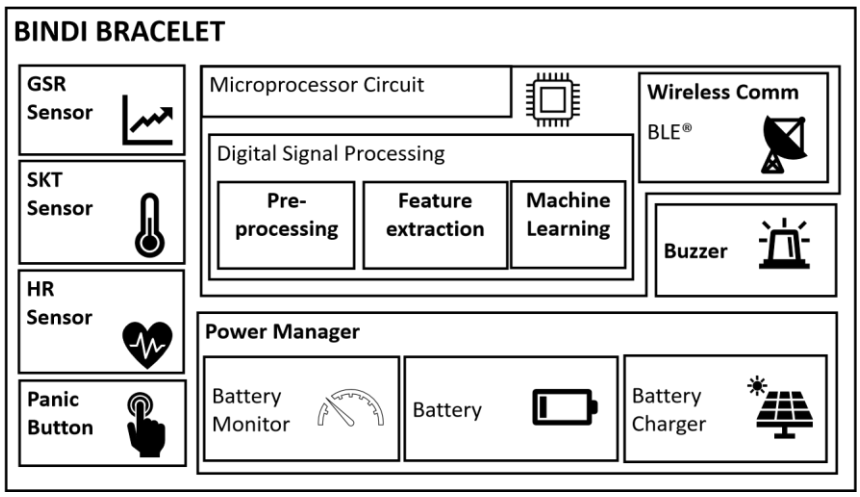


Figura 5. 2: Arquitectura simplificada de la pulsera de Bindi [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Las variables fisiológicas comentadas anteriormente se eligieron debido a su demostrada fuerte relación con el reconocimiento de emociones [281] y a su facilidad de implementación en dispositivos *wearable*. Este último punto es especialmente relevante y nos llevó a descartar otros

sensores fisiológicos típicos utilizados en este campo (como el EEG), que no cumplen el requisito de discreción. La cadena de procesamiento digital de señales dentro de la pulsera implica tanto la adquisición y el filtrado de las señales fisiológicas como las etapas de extracción e inferencia de características.

Edge computing: Colgante

Este dispositivo capta información de audio y voz, que se transmite a un motor inteligente para la detección del *miedo*. El colgante tiene la misma arquitectura de hardware que la pulsera, pero integra un micrófono en lugar de sensores fisiológicos. Su arquitectura se muestra en la Fig. 5.3. El micrófono se basa en un sistema microelectromecánico con un sensor de audio omnidireccional. Esta pieza incluye un elemento sensor capacitivo y una interfaz de circuito integrado que permite obtener directamente una señal digital. La cadena de procesamiento de la señal digital dentro del colgante comprende tanto la recepción y el filtrado de las señales auditivas (audio y voz) como la transmisión inalámbrica al teléfono inteligente. Hay que tener en cuenta que, debido al ancho de banda limitado de la comunicación inalámbrica, el audio se comprime antes de ser transmitido.

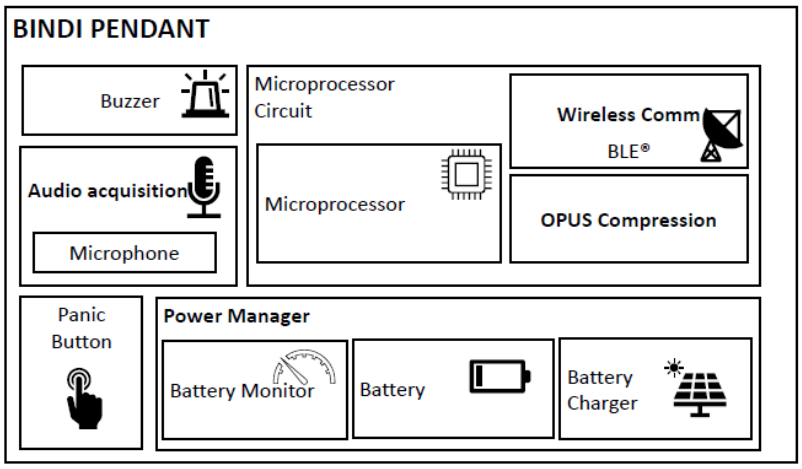


Figura 5. 3: Arquitectura simplificada del colgante de Bindi [1]. Reproducida con permiso del propietario del copyright, © 2022 IEEE.

Fog computing

La *fog computing* dentro de Bindi está representada por la aplicación Bindi App que se ejecuta en un teléfono inteligente (*smartwatch*). Proporciona una interfaz gráfica al usuario final y realiza las siguientes funcionalidades técnicas:

- Solicita datos fisiológicos y auditivos a la pulsera y al colgante respectivamente, de acuerdo con los *pipelines* de procesamiento de datos implementados.
- Gestiona los disparos de alarma (alertas por SMS/unidad de protección o servicios de emergencia) y los registra en el servidor en función de la respuesta inteligente del motor o del botón de pánico manual.
- Realiza un seguimiento de la ubicación de cada usuaria mediante GPS.

- Gestiona las comunicaciones seguras con el servidor adaptándose al estado actual de la batería del *smartphone*.
- Recoge y sube a la nube (*cloud*) datos auditivos y fisiológicos cifrados como prueba de un presunto delito si se dispara la alarma.
- Realiza los procesos de extracción de características y de inferencia para el sistema monomodal auditivo. Además, maneja diferentes estrategias de fusión de datos.

Computación en la nube (*Cloud Computing*)

La parte de computación en nube es donde entra en funcionamiento el Servidor de Bindi. La implementación del Servidor Bindi consiste en una base de datos MongoDB[29] y un servidor de aplicaciones web NodeJS[30]. Este servidor almacena la información capturada en el *edge* con tres objetivos principales. En primer lugar, sirve como monitor de actividad, indicando posibles situaciones problemáticas sobre la evolución afectiva a largo plazo de las víctimas de violencia de género para las personas que supervisan el bienestar de las usuarias. Segundo, almacena datos encriptados, sirviendo como prueba digital en un eventual juicio. En tercer lugar, toma decisiones tras la activación de las alarmas siguiendo procedimientos de seguridad predeterminados.

5.4. Estrategias de Fusión Multimodal para Bindi

La fusión de datos es una forma poderosa de mejorar la robustez del motor de inteligencia multimodal de Bindi. Así, en esta sección se proponen las arquitecturas de fusión de datos consideradas para reforzar la fiabilidad y robustez de Bindi, de modo que los datos fisiológicos y de audio puedan considerarse conjuntamente en la decisión de activar la alarma.

Presentamos un análisis de una estrategia de fusión tardía (*late fusion*) multimodal para combinar los *pipeline* de procesamiento de datos fisiológicos y del habla con el fin de determinar la mejor estrategia para el motor de inteligencia para Bindi. Nuestro objetivo es analizar y comprender mejor las respuestas de las mujeres a la emoción del *miedo* en situaciones de riesgo.

5.4.1. *Late Fusion* Inicial en Cascada: Bindi 1.0

En una primera aproximación en Bindi, se fusionaron los sistemas de detección de alertas del habla y de las señales fisiológicas siguiendo un enfoque a nivel de decisión, también denominado fusión tardía (*late fusion*). Para ello, los autores consideraron un enfoque en cascada en el que los dos sistemas se ejecutan uno tras otro.

[29] https://www.mongodb.com

[30] https://nodejs.org/es/

El sistema comienza ejecutando el sistema de señales fisiológicas, que analiza los datos captados por la pulsera y decide si la usuaria se encuentra en una situación peligrosa o no. Este sistema de señales fisiológicas se basa en un algoritmo de clasificación KNN, que se ejecuta en el procesador del interior dla pulsera. Si el sistema da como resultado una detección positiva, se comunica con el teléfono inteligente, que lanza una petición al sistema de voz para que analice la situación actual. El sistema de voz captura los datos de audio durante un tiempo determinado, que se envían al *smartphone* con un proceso de compresión previo. En el *smartphone*, los datos de audio se analizan y se envían a un modelo MLP que se ejecuta en el microprocesador del smartphone. La predicción realizada por el sistema de voz es entonces la predicción global alcanzada en Bindi.

De este modo, el sistema fisiológico actúa como desencadenante para activar la siguiente etapa de la cascada. Se asumió esta decisión de diseño porque el coste energético de la pulsera que captura esos datos fisiológicos, así como los ligeros algoritmos de aprendizaje automático dentro del procesador, permiten que el dispositivo funcione durante horas (al menos dos días, en el momento de la publicación de [9]). Por el contrario, la captura de datos de audio y la composición de la información para su envío son costosas, por lo que deben reducirse al máximo. Además, ejecutar muchas veces el análisis de los datos de audio también es costoso para el smartphone en términos de batería. Por todas estas razones, se decidió que el sistema de voz estuviera en la segunda etapa de la cascada.

5.4.2. Enfoque de Fusión Híbrida (*Hybrid Fusion*)

El trabajo de esta sección se ha publicado en [9] junto con otros miembros del equipo UC3M4Safety. La arquitectura de fusión inicial en Bindi comentada anteriormente se realiza a nivel de decisión. Esta estrategia es fácil de implementar, pero incluye la desventaja de no considerar las posibles relaciones entre las distintas modalidades del sistema, es decir, las posibles correlaciones entre la información fisiológica y la auditiva. Además, otra desventaja es la heterogeneidad entre las puntuaciones de confianza proporcionadas por los modelos de cada modalidad. Antes de hablar de otras arquitecturas de fusión para Bindi hay que tener en cuenta algunos **aspectos clave**:

- Bindi es un sistema distribuido compuesto por tres dispositivos, un teléfono inteligente y dos dispositivos integrados (una pulsera y un colgante). Esto significa que la comunicación entre ellos es esencialmente necesaria.
- Bindi se encuentra dentro de un sistema ciberfísico restringido, lo que significa que tanto los recursos informáticos como la batería son limitados, especialmente para los dos dispositivos integrados. Centrándonos en la duración de la batería, la transmisión de datos consume más energía que otras tareas habituales, como el procesamiento y la detección. Por tanto, cuantos menos datos se transmitan, más durará la batería de los dispositivos.

- La arquitectura inicial a nivel de decisión implica que las señales de las dos modalidades están desalineadas en el tiempo. Así, las señales fisiológicas que activan la alarma se adquieren antes que la grabación de audio.

  Teniendo en cuenta estos aspectos clave, y a diferencia de otros métodos para fusionar información fisiológica y vocal mediante la fusión a nivel de características que influyeron en este trabajo [280], los autores proponen una arquitectura híbrida de fusión de datos combinando los enfoques a nivel de decisión (*late*) y a nivel de característica (*early*). Por lo que sabemos -en el momento de la publicación de [9]-, este enfoque híbrido nunca se había considerado antes para un sistema multimodal fisiológico-auditivo *wearable*.

  El equipo tomó dos decisiones de diseño principales para esta arquitectura híbrida. En primer lugar, los dos dispositivos integrados no pueden realizar la fusión a nivel de características debido a las limitaciones de capacidad computacional y batería. Por lo tanto, el teléfono inteligente se encargaría de esta tarea. En segundo lugar, no es posible enviar continuamente información fisiológica y auditiva al teléfono inteligente para que realice la fusión a nivel de rasgos y, por tanto, no puede tener lugar en todo momento.

  La Fig. 5.4 muestra la fusión híbrida de datos propuesta para Bindi. Por defecto, el sistema está realizando la fusión tardía ya incluida en el planteamiento inicial de Bindi 1.0 (véase la sección 5.4.1). Esto significa que la pulsera está capturando datos fisiológicos a lo largo del tiempo (paso 1). A continuación, el sistema ML de la pulsera analiza los datos de entrada. En caso de que detecte la emoción objetivo, genera un *trigger* al teléfono inteligente (2). El smartphone solicita al colgante (3) que capture datos de audio (4).

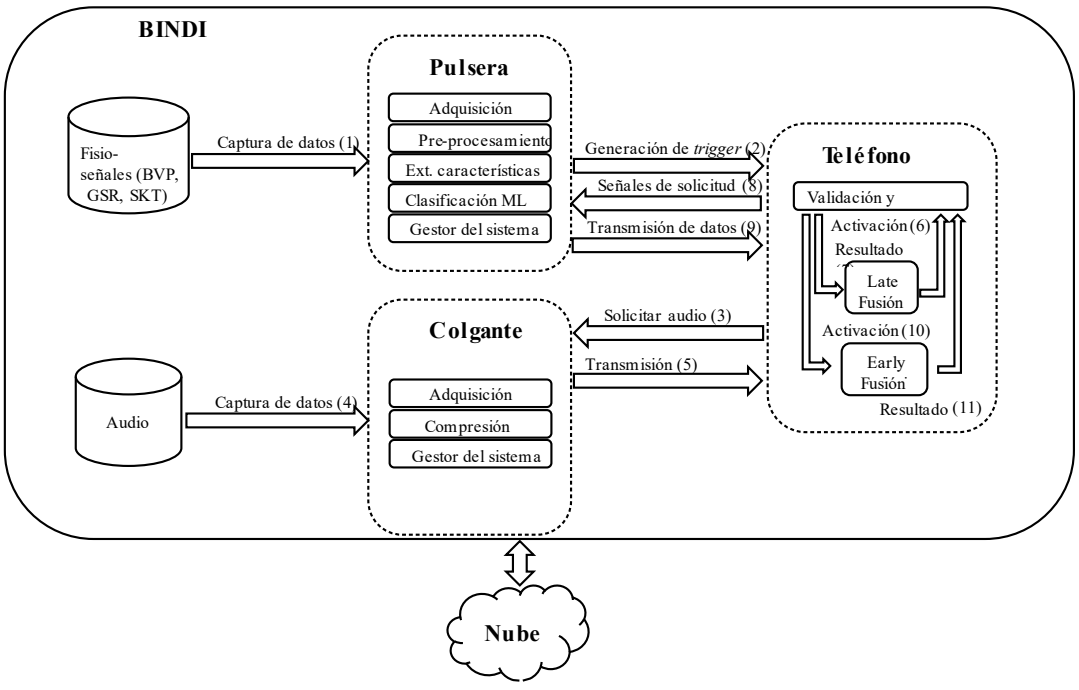


Figura 5. 4: Arquitectura híbrida de fusión de datos para Bindi 2.0 [9]. Reproducido con permiso del propietario del copyright, Springer Nature.

La información de audio se comprime y se envía al smartphone (5). El smartphone ejecuta el modelo basado en MLP (6), obteniendo la respuesta para la arquitectura de fusión tardía (7). En caso de que la fusión tardía dé como resultado una detección positiva de *miedo*, entonces se ejecuta la arquitectura de fusión temprana. En tal caso, el teléfono inteligente solicita los datos fisiológicos de la pulsera (8), que fueron capturados en el pasado (es decir, los datos fisiológicos que generaron el disparo en la fusión tardía y los obtenidos durante el tiempo en que el colgante obtuvo datos de audio). Tras aplicar algunas técnicas de compresión de los sensores para aliviar el uso de la batería, la información solicitada se envía al teléfono inteligente (9) que ejecuta un algoritmo de clasificación que combina tanto los datos fisiológicos como los auditivos (10). El resultado de esta arquitectura de fusión temprana es la salida de todo el sistema Bindi. El procesamiento posterior se realizará en la nube [282].

5.4.3. Estrategias de *Weighted Late Fusion*: Bindi 2.0

Además de la fusión híbrida, el equipo UC3M4Safety también ha sopesado otras arquitecturas de fusión intermedias, y su validación y comparación forma parte del plan a seguir del equipo.
La arquitectura original de Bindi se basa en una estrategia de fusión de datos tardía, que se ejecuta siguiendo una cascada de dos capas, en la que cada capa tiene un modelo de inteligencia asociado a cada modalidad de datos -las señales fisiológicas adquiridas por la pulsera y el audio y el habla captados por el colgante, respectivamente-. El modelo de la primera capa actúa como un interruptor de bajo coste para activar una segunda capa más exigente, relacionada también con una capacidad de detección más potente. Esta estrategia inicial de bajo consumo es útil para decidir cuándo debe llevarse a cabo la captura de audio más potente y costosa en el colgante. Sin embargo, el uso de los datos captados en la pulsera sólo para fines de conmutación podría implicar que el motor de decisión inteligente no está teniendo en cuenta toda la información disponible.
En esta sección proponemos tres estrategias de fusión tardía ponderada (*weigthed*) basadas en la literatura (por ejemplo, [277]) que se consideran un compromiso entre la baja complejidad computacional y la solidez teniendo en cuenta la confianza del sistema en las predicciones [1]. Estas estrategias de fusión tardía se alimentan de las etiquetas binarias proporcionadas por los motores de inteligencia monomodal fisiológica y del habla, como se muestra en la Fig. 5.5.

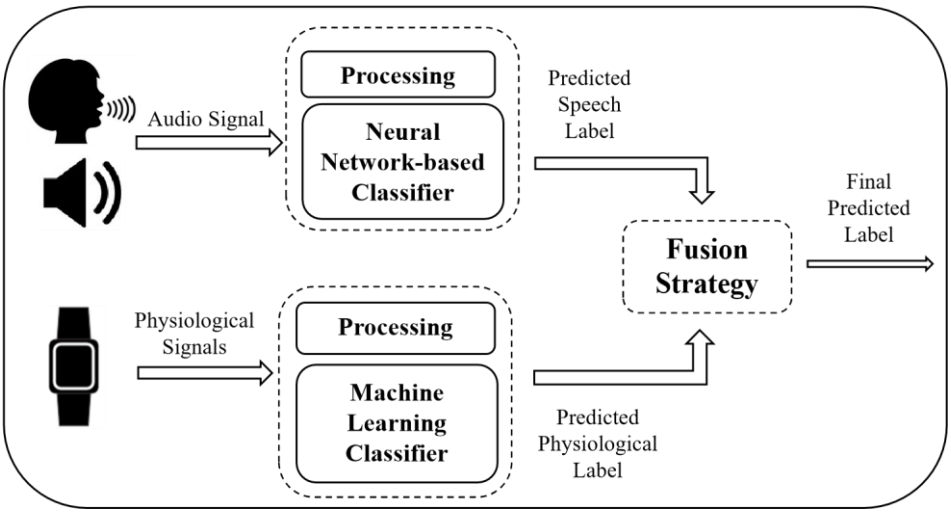


Figura 5. 5: Diagrama de bloques de fusión de datos de Bindi [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Como se ha comentado anteriormente, los subsistemas monomodales fisiológico y del habla estiman una etiqueta binaria, $y^m_k \in \{0, 1\}$, para cada ventana temporal k y modalidad $m \in \{\text{phy}, \text{sp}\}$, refiriéndose phy y sp a los subsistemas fisiológico y del habla, respectivamente. Nótese que cada una de las modalidades utiliza una longitud de ventana temporal diferente, T_m, en segundos, debido a sus peculiaridades específicas. Bindi pretende emitir una respuesta por periodo de tiempo n (cada uno de la misma longitud L), con $n \in 1, 2, \ldots$, en segundos. Así, una estimación de la probabilidad de *miedo* p^m_n para el *n-ésimo* periodo de tiempo y la *m-ésima* modalidad se calcula como

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

donde $K_m = \left\lfloor \frac{L}{T_m} \right\rfloor$, es decir, el número de ventanas temporales que consideramos para cada modalidad para la estimación de probabilidades.

A continuación, una única etiqueta binaria, Y_n^m, basada en p^m_n puede calcularse como

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

es decir, dará como resultado "1" (*miedo*) si p^m_n es superior al umbral predefinido relacionado con la modalidad, $\text{th}_m \in \{0,1\}$, o "0" (*no miedo*) en caso contrario. Tenga en cuenta que los valores th_{phy} y th_{sp} se tratan en la sección 5.6.2.

Como métrica para representar el grado de confianza de cada sistema monomodal en la etiqueta de clase predicha en un periodo determinado, la entropía h^m_n para el *n-ésimo* periodo de tiempo y la *m-ésima* modalidad se calcula como

$$h^m{}_n = -[p_n^m \cdot \log(p_n^m) + (1 - p_n^m) \cdot \log(1 - p_n^m)]. \quad (5.3)$$

Sobre esta base, se estudian tres estrategias de fusión tardía para producir la respuesta fusionada del sistema Y_n^{f} para el periodo de tiempo *n-ésimo*:

- Caso 1, Entropía más baja (*lowest entropy*): La respuesta del sistema corresponde a la etiqueta binaria producida por el sistema monomodal con la entropía más pequeña, es decir, para la que el modelo tiene más confianza. Para ello, la probabilidad de *miedo* fusionada p_n^f para el *n-ésimo* periodo de tiempo se calcula como

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

A continuación, aplicando el mismo razonamiento que en la ecuación (5.2), se obtiene una etiqueta binaria fusionada como

$$Y_n^{\mathrm{f}} = \begin{cases} 0 & \text{if } p_n^{\mathrm{f}} < \mathrm{th_f} \\ 1 & \text{otherwise} \end{cases}, \quad (5.5)$$

donde, por ahora, $\mathrm{th_f}$ es el 0,5 convencional.

- Caso 2, Combinación ponderada de entropía inversa (*Inverse Entropy Weighted Combination*): La probabilidad de *miedo* fusionada p_n^f para el *n-ésimo* periodo de tiempo se calcula como una suma ponderada de probabilidades, tal y como viene dada por:

$$p_n^{\mathrm{f}} = \sum_m w_n^m \cdot p_n^m, \quad (5.6)$$

$$w_n^m = \frac{1/h_n^m}{\sum_m 1/h_n^m}. \quad (5.7)$$

A continuación, se obtiene una etiqueta binaria fusionada según la ecuación (5.5).

- Caso 3, *Logical OR*: La respuesta del sistema corresponde al cálculo OR lógico sobre las etiquetas binarias para cada sistema monomodal. Es decir

$$Y_n^{\mathrm{f}} = Y_n^{\mathrm{phy}} \vee Y_n^{\mathrm{sp}}. \quad (5.8)$$

Al comparar teóricamente las tres estrategias de fusión, el caso 3 facilita la obtención de una predicción de la clase de *miedo* sin comprobar la confianza del subsistema, lo que podría dar lugar a una detección falsa. Sin embargo, la estrategia de entropía más baja (Caso 1) se fía del modelo con más confianza en la inferencia sin tener en cuenta las diferencias en las probabilidades. Por último, la combinación ponderada de entropía inversa (Caso 2) establece un equilibrio entre las probabilidades y las entropías para cada subsistema monomodal. Así, la confianza de esta última estrategia, el Caso 2, puede ser mayor que la de las demás.

5.5. *Pipelines* de Procesamiento de Datos

Uno de los objetivos clave de nuestro trabajo es validar el *pipeline* de procesamiento de datos dentro de Bindi, desde la adquisición de datos hasta la generación de alarmas. Para lograr este objetivo se han aplicado y comparado diferentes disposiciones de los componentes del sistema. Este hecho ha conducido a una exploración del espacio de diseño de diferentes arquitecturas multimodales (información fisiológica y auditiva) del sistema [1]. En concreto, se han evaluado tres disposiciones temporales:

1. La primera versión es Bindi 1.0 [177], que se basa en una estrategia jerárquica o en cascada. En esta versión (descrita en la sección 5.4.1), la información fisiológica es recogida continuamente por la pulsera, que ejecuta un motor ligero de inteligencia fisiológica monomodal. Cuando detecta que la usuaria está experimentando *miedo*, dispara una pre-alarma a la aplicación para *smartphone* de Bindi. Esta acción hace que el colgante comience a grabar audio durante un breve periodo, lo que supone una estrategia de bajo consumo energético para el micrófono. A continuación, la señal de audio se envía a la App Bindi para que realice la detección del *miedo* mediante un motor de inteligencia monomodal basado en el habla. Por último, si este último sistema -el del habla- confirma la detección, la Bindi App inicia un procedimiento de seguridad para ayudar a la usuaria, disparando una alarma al Bindi Server.
2. La versión posterior, Bindi 2.0a, se basa en las mismas dos canalizaciones monomodales de procesamiento de datos de Bindi 1.0, pero en la fase de decisión final aplica una técnica de fusión tardía en lugar de una estrategia jerárquica de confirmación [9]. Hereda la funcionalidad de pre-alarma de Bindi 1.0 para que el micrófono consuma poca energía.
3. Como variación de Bindi 2.0a, el Bindi 2.0b sigue el esquema de fusión tardía introducido en Bindi 2.0a, pero lo basa en la adquisición continua de datos fisiológicos y auditivos, lo que significa que la funcionalidad de pre-alarma no está habilitada.

   En los siguientes subapartados se detallan los *pipeline* de procesamiento de datos fisiológicos y auditivos. La contribución del procesamiento fisiológico ha sido desarrollada por otros miembros del equipo UC3M4Safety, el *pipeline* de datos de audio es una contribución propia como parte de la investigación realizada para esta tesis, y las estrategias de fusión son una contribución conjunta de todo el equipo UC3M4Safety.

   La naturaleza particular de los tipos de datos (fisiológicos, del habla) conlleva retos diferentes. Por ello, los esquemas de procesamiento de datos, los métodos y las técnicas de extracción de características se adaptan a cada señal.

Subsistema de datos fisiológicos

En esta sección haremos un breve recuento del sistema fisiológico en Bindi para ayudar a la comprensión del sistema de fusión. Para más detalles, referirse a [1] y [53].

La primera etapa de procesamiento de los datos fisiológicos es la adquisición de la señal y el enventanado. En nuestro caso, las frecuencias de muestreo seleccionadas son 100, 10 y 5 Hz para la BVP, la GSR y la SKT, respectivamente. Estas frecuencias son adecuadas para captar la dinámica de la señal con la resolución temporal apropiada. Para la segmentación de la señal, se utiliza una estrategia de superposición de longitud fija de ventanas de 20s con un solapamiento de 10s. Esta configuración proporciona una resolución de frecuencia de 0,05 Hz, lo que resulta en una buena compensación entre el almacenamiento de datos y la información fisiológica disponible para ser extraída. Una vez capturadas y segmentadas las señales, la etapa de filtrado elimina el ruido fuera de banda de forma específica para cada señal.

El bloque de extracción de características extrae la información contenida en las señales fisiológicas y es la siguiente etapa en el *pipeline* de procesamiento. En concreto, hay 25 características para el BVP, 17 características para el GSR y seis características para el SKT. En [261] se ofrece una amplia descripción de las características. Para la clasificación, se utiliza un algoritmo ligero de aprendizaje automático supervisado binario *K-Nearest Neighbors* (KNN). Durante la fase de entrenamiento, se aplica el aprendizaje sensible a los costes modificando el coste de clasificación errónea del KNN, lo que aumenta la sensibilidad, es decir, el sistema tendrá menos probabilidades de omitir una situación peligrosa para el caso de uso [283]. Por último, la salida del subsistema de datos fisiológicos es una etiqueta binaria cada 10s. Este *pipeline* fisiológico se ha probado en trabajos anteriores utilizando un conjunto de datos públicos [261].

Subsistema de datos de voz

El procesamiento de los datos del habla incluye los siguientes módulos fundamentales: detección de la actividad vocal (VAD), filtrado en el dominio de la frecuencia, extracción de características, normalización y un clasificador basado en una red neuronal.

Se emplea un módulo VAD ligero básico [284] basado en la energía espectral para detectar y eliminar las partes silenciosas de las señales del habla en las que el extractor de características de voz posterior no extraería ninguna información relevante del habla debido a su ausencia. No obstante, la detección del silencio es crucial para el correcto funcionamiento del dispositivo, ya que las mujeres en situaciones de peligro suelen reaccionar con un sobresalto y permanecer en silencio, por lo que se pretende trabajar más en la caracterización del silencio en el futuro.

En combinación con el módulo VAD, para facilitar la manipulación de las señales y conservar al mismo tiempo toda la información significativa de los datos del habla, es necesario reducir el muestreo de las señales a 16 kHz. A continuación, se aplica un filtro de paso bajo a 100 Hz para

eliminar el ruido de baja frecuencia captado por el micrófono y posiblemente causado por el aire acondicionado y los zumbidos de la red eléctrica, entre otros factores, ya que las bases de datos con las que trabajamos se graban en condiciones de laboratorio. Después, se filtran las señales con un filtro de paso bajo a 8 kHz para conservar la información clave sobre el habla y seguir manteniendo una baja complejidad.

A continuación, el extractor de características del habla computa 38 características del habla dedicadas a la detección de emociones utilizando una ventana de 20 ms con 10 ms de solapamiento, que son valores estándar de la bibliografía. Entre las características consideradas están el tono, Mel coeficientes cepstrales de frecuencia, formantes, energía y características espectrales adicionales, todos ellos calculados mediante la librería de herramientas Python librosa [241]. Las características se agregan por segundo calculando sus estadísticas de media y desviación estándar para normalizarlas posteriormente. Se realizan experimentos preliminares de ablación antes de fijar esta agregación de 1s, variando el contexto temporal de las características del habla agregadas para 1, 5 y 10s.

La normalización de las características se realiza aplicando los valores de la puntuación de media y desviación estándar de las características de referencia extraídas cuando la usuaria se encuentra en estado de reposo o neutro, lo que se denomina normalización del estado basal. Se prueban informalmente otros esquemas de normalización (por ejemplo, por vídeo, por usuaria y z-score tradicional) antes de considerar la normalización del estado basal descrita, pero se seleccionó esta última por mostrar un mejor rendimiento del sistema. Las características agregadas normalizadas se introducen en un clasificador de red neuronal MLP adaptado a la usuaria y entrenado para la detección del *miedo*. Este subsistema genera una etiqueta binaria cada 1 s. Las etiquetas predichas por el subsistema de habla monomodal cada segundo se suavizan en el tiempo utilizando una ventana de 7s para mantener una detección coherente y estable. Nótese que cada una de las modalidades utiliza una longitud de ventana temporal diferente en segundos, debido a sus peculiaridades específicas, que se fusionan utilizando las distintas estrategias de fusión (véase la sección 5.4.3).

5.6. Configuración Experimental y Resultados sobre el Reconocimiento del Estrés y el Miedo

En el ámbito del reconocimiento de emociones basado en el habla, primero realizamos experimentos de detección de estrés basados en la voz para evaluar si los eventos acústicos (que podrían definir el contexto acústico en el que se encuentra la usuaria) podrían ayudar a detectar el estrés en la modalidad auditiva. Después, y para validar las estrategias de fusión propuestas en la sección 5.4, utilizamos WEMAC, una base de datos capturada por nuestro equipo UC3M4Safety cuyo objetivo es la elicitación del *miedo* relacionado con la violencia de género.

5.6.1. Experimentos sobre el Reconocimiento Unimodal de Estrés

En esta sección explicamos brevemente nuestra aportación sobre la experimentación realizada para la clasificación del estrés en el habla, ya que podría entenderse como una emoción relacionada con el *miedo*, en una fase preliminar previa al trabajo con la base de datos WEMAC. En nuestro estudio detallado anteriormente [8] en la sección 4.5 realizamos una tarea de identificación de hablantes en Bios-DB [157] y Biospeech+ (véase la sección 3.2.3), una base de datos aumentada con eventos acústicos basada en Bios-DB. En esta sección queremos detallar la tarea de reconocimiento de emociones realizada sobre los mismos datos y sus resultados. La metodología seguida para esta tarea es exactamente la misma que en la sección 4.5, en lo que respecta a las características extraídas de las señales de voz y a los clasificadores utilizados para la tarea. La principal diferencia radica en las etiquetas utilizadas, que ahora son dos en particular, i) etiquetas binarias referidas a habla estresada y neutra, y ii) las emociones reinterpretadas en los 4 cuadrantes del espacio PAD, tal y como se describe en la Sec. 3.2.2.

Los resultados se presentan en la tabla 5.1, donde *p* representa el número de parámetros de cada modelo. MLP se refiere al perceptrón multicapa, K2D al modelo de 2 capas densas de Keras y KCGD al modelo de Keras compuesto por un GRU convolucional 1D, bidireccional y capas densas. Se muestran los resultados de la media y la desviación estándar para una validación quíntuple. Para las dos tareas consideradas, MLP con librosa consigue el mejor rendimiento.

Modelo	librosa	*p*	eGeMAPS	*p*	yamNET	*p*	L+E+Y	*p*	feat sel	*p*
Reconocimiento binario de tensiones										
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*
Reconocimiento de las emociones del habla (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*

TABLA 5.1: Resultados de F1-score para el reconocimiento del estrés y las emociones en señales de audio limpias [8].

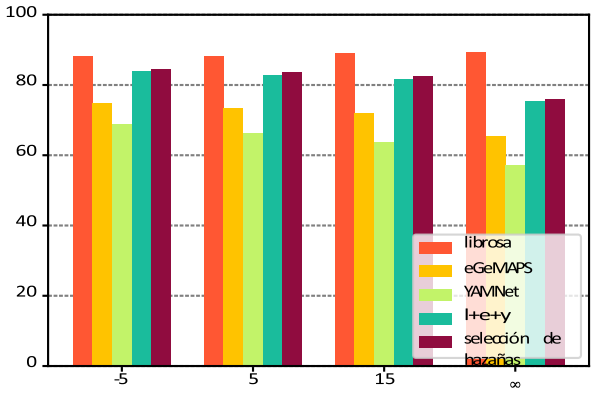


Figura 5. 6: Resultados de F1-score para el reconocimiento binario de estrés con un MLP en Biospeech+ [8]. Reproducido con permiso del propietario del copyright, ISCA.

La Fig. 5.6 muestra los resultados para diferentes SNR (en el eje horizontal) en Biospeech+ (véase la Sec. 3.2.3). En concreto, muestra los resultados para la clasificación binaria de las etiquetas de estrés para el modelo que mejor funcionó (MLP). Todos los conjuntos de características -excepto quizá librosa, que se mantiene estable- muestran una tendencia a mejorar la F1-score a medida que el valor de SNR es más bajo, es decir, cuando los eventos acústicos se superponen a las relaciones del habla[31] (-5 y 5 dB)[32] . Esto demuestra que ampliar nuestra base de datos con eventos acústicos resulta útil para el reconocimiento del estrés en el habla y el audio. Todos los conjuntos de características, en mayor o menor medida, parecen ser capaces de captar información sobre los eventos acústicos que se consideran desencadenantes de estrés.

5.6.2. Experimentos Monomodales y Multimodales de Reconocimiento del Miedo utilizando WEMAC para Bindi

En esta sección, pretendemos validar y evaluar las distintas arquitecturas de fusión de Bindi para la tarea de reconocimiento *del miedo* mediante WEMAC. Este estudio se ha publicado en [1] junto con otros miembros del equipo UC3M4Safety. Este trabajo pretende ser el primer marco multimodal que sirva de referencia para poder seguir trabajando con el *miedo* provocado en la vida real de las mujeres. Hasta donde sabemos, es la primera vez que que se ha dado una fusión multimodal de datos fisiológicos y del habla para el reconocimiento del *miedo* en este contexto de violencia de género.

Comenzamos trabajando en el análisis de las etiquetas. Primero binarizamos las emociones discretas anotadas a cada estímulo audiovisual por cada usuaria, para transformar el problema de modelado en una clasificación binaria, donde "1" (clase positiva o también llamada “de interés”)

[31] Para la medida SNR consideramos el habla en primer plano de Biospeech como la "señal" y los eventos de audio como "ruido".

[32] Observe que el símbolo *"infinito"* denota la línea de base para cuando no se añaden eventos acústicos a la base de datos.

representaba *el miedo* y "0" (clase negativa) cualquier otra emoción, convirtiendo el problema en un problema de clasificación binaria del *miedo*.

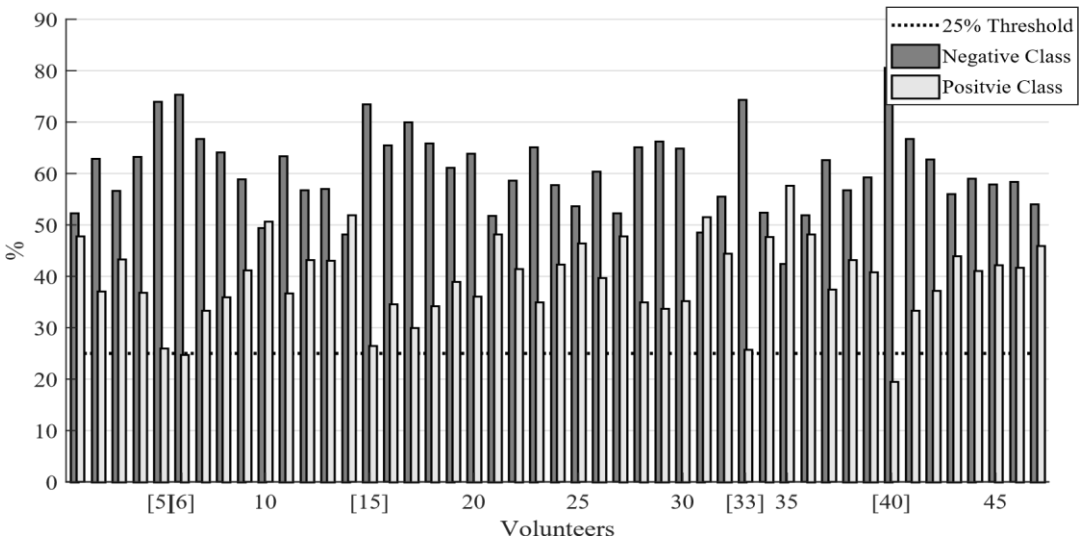


Figura 5. 7: Distribuciones estadísticas de las clases positivas y negativas para las etiquetas de emociones auto-anotadas binarizadas en cuanto a miedo en WEMAC. Las voluntarias entre paréntesis con las excluidas [1]. Reproducido con permiso del propietario de copyright, © 2022 IEEE.

Se observó que algunas voluntarias en particular presentaban una distribución considerablemente desequilibrada en sus etiquetas auto-anotadas, como se muestra en la Fig. 5.7. Por lo tanto, decidimos excluir a las voluntarias 5, 6, 15, 33 y 40 de la evaluación, ya que sólo presentaban alrededor del 25% de etiquetas positivas. En consecuencia, la evaluación debía realizarse sólo con 42 de las 47 voluntarias iniciales. La distribución de clases de estas 42 voluntarias era de alrededor del 60% y el 40% para las clases negativa y positiva, respectivamente. Esta distribución se ajusta a la información presentada en la tabla 3.3 para las distintas emociones.

Obsérvese que los resultados experimentales de esta sección tienen en cuenta el proceso de validación realizado *off-line*, para evaluar la funcionalidad de los *pipeline* de procesamiento de datos y las estrategias de fusión, y posteriormente incorporar dichos módulos en la arquitectura, equilibrando las compensaciones observadas.

Consideraciones para el montaje experimental en las fases de entrenamiento y *test* de los subsistemas monomodales

Hubo que tener en cuenta algunos puntos para diseñar las estrategias de entrenamiento y test de los dos subsistemas monomodales. En primer lugar, según el diseño de la base de datos WEMAC, hay que tener en cuenta que los datos fisiológicos se capturaron durante la visualización del estímulo audiovisual (y también la elicitación de la emoción ocurría en ese momento), mientras que la grabación del habla se registró durante la anotación posterior a la visualización. Esto significa que los datos fisiológicos y del habla no estaban alineados en el tiempo en WEMAC. Sin embargo, ambos tipos de datos debían fusionarse en Bindi 2.0b para cada reacción emocional

por usuaria o experimento, a diferencia de Bindi 1.0 y Bindi 2.0a, donde la fusión estaba condicionada a la pre-alarma fisiológica (véase la sección 5.4 para la descripción de las estrategias de fusión de Bindi). Por lo tanto, obtuvimos una única $p^m{}_n$ por experimento y modalidad, según la ecuación 5.1; nótese que L es la longitud de los estímulos audiovisuales para la modalidad fisiológica y la longitud total de la grabación de audio para la modalidad del habla. Durante el etiquetado de voz, se pidió a las voluntarias que revivieran las emociones sentidas durante la elicitación del estímulo, por lo que se supuso que la correspondencia era suficientemente sólida entre ambos instantes temporales. Sin embargo, esta suposición necesitará una mayor validación cuando se disponga del resto de subconjuntos de WEMAC.

En segundo lugar, para la división entrenamiento-test, se aplicó una estrategia LASO. Se trataba de un procedimiento *speaker-adapted* y *speaker-semi-independent* (en última instancia, *speaker-dependent*) para entrenar los 42 modelos necesarios, es decir, uno por usuaria. Se eligió este enfoque debido a que la personalización de la persona que proporciona LASO es crucial para un modelo de detección de emociones como el nuestro [285]. Así, cada modelo se entrenó con todos los datos disponibles del resto de las usuarias y se afinó (con una estrategia de *fine-tuning*) con la mitad de las instancias de la persona que se iba a probar, en concreto, los datos adquiridos de los siete primeros estímulos audiovisuales (de un total de 14). El resto de los datos de los siete últimos vídeos de la sesión debían utilizarse como muestras de test. De este modo, los datos de prueba no se veían durante la fase de entrenamiento, pero el modelo obtenía cierta información sobre el sujeto, tal y como estaba previsto.

En tercer lugar, en cuanto a las particularidades específicas del entrenamiento, para el subsistema fisiológico monomodal, se consideró el mismo coste de clasificación errónea de 1,6 a la clase positiva para tratar el desequilibrio de clases comentado para todos los modelos fisiológicos generados. Este coste se fijó mediante un barrido experimental de parámetros. Además, el entrenamiento se validó mediante una estrategia de validación cruzada *k-fold* estratificada *(stratified k-fold cross-validation*), con $k = 5$. Por último, la normalización aplicada al conjunto de datos se basó en la técnica de z-score aplicada a las características extraídas de todos las voluntarias.

Para el subsistema monomodal del habla, el clasificador consistió en una red neuronal ligera poco profunda con capas de entrada, una capa oculta totalmente conectada (*fully-connected*) y de salida también totalmente conectada. La red tenía 38 unidades en su capa de entrada, es decir, una por cacterística. El número de unidades ocultas en la capa densa se fijó en 250 para evitar aumentar en gran medida el coste computacional pero lograr tasas de predicción bastante buenas. La capa de salida proporcionó una etiqueta predicha como salida. Todas las muestras, excepto las de la usuaria de interés, se utilizaron para entrenar el modelo durante 300 épocas (*epochs*), con una parada temprana tras una paciencia de 30 épocas en la pérdida de validación del modelo, una

función de pérdida de entropía cruzada binaria (*binary cross-entropy*), utilizando el optimizador Adam, y una tasa de aprendizaje de 0,001. A continuación, se utilizaron muestras de la usuaria de interés (la mitad de las disponibles según la estrategia LASO) para afinar el modelo durante un máximo de 100 épocas, con un enfoque de parada temprana (es decir, parada tras una meseta de 10 épocas en la pérdida del modelo). En cuanto a la normalización de la *z-score* utilizada, las características extraídas de las grabaciones del habla del sexto estímulo audiovisual se utilizaron como *baseline*, ya que se esperaba que este vídeo provocara una emoción de calma y se suponía que evocaba un estado neutro en la usuaria, por lo que utilizamos la normalización del estado basal mencionada anteriormente.

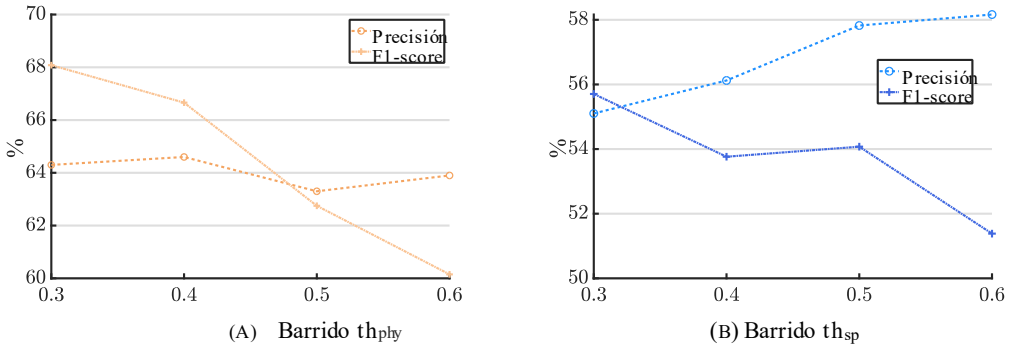
Por último, en cuanto al procedimiento de test, como se explica en la sección 5.4, las salidas del subsistema monomodal eran matrices de etiquetas binarias. Concretamente para WEMAC, la longitud de las matrices era igual a la división de la duración de cada estímulo por los periodos de muestreo monomodal, es decir, 10 y 1 s para los subsistemas fisiológico y del habla, respectivamente. Después, esas matrices recopiladas se procesaron calculando las probabilidades y sus correspondientes etiquetas binarias aplicando los umbrales fisiológicos (th_{phy}) y del habla (th_{sp}). Las estrategias de fusión de datos propuestas también generaron sus correspondientes etiquetas binarias, como se describe en la sección 5.4. Las métricas de evaluación seleccionadas, es decir, la precisión y la F1-score, utilizaron para las etiquetas binarias obtenidas. La precisión podía representar bastante bien los índices de predicción, ya que el desequilibrio de clases era bajo, de todos modos se consideró la F1-score en primer lugar para hacer frente al ligero desequilibrio observado y en segundo lugar debido a la mayor importancia de la clase positiva en nuestro caso de uso, ya que la F1-score es una buena métrica para un problema de detección en el que el número de positivos es menor en comparación con los negativos y sin embargo la detección de la clase positiva es crucial.

Resultados de reconocimiento del *miedo*

Esta sección presenta los resultados experimentales relativos a la predicción del *miedo* utilizando WEMAC para las diferentes configuraciones del sistema discutidas en la Sec. 5.5. Nótese que es la primera vez que se utiliza esta base de datos; por lo tanto, estos resultados representan el primer paso hacia la detección real (no actuada) de la emoción del *miedo* a partir de variables fisiológicas y auditivas para el problema de la violencia de género y pretenden ser un *baseline* para futuros desarrollos.

El primer análisis se refiere al rendimiento de los subsistemas fisiológico y del habla trabajando de forma independiente en un entorno continuo, es decir, teniendo en cuenta todas las muestras. Este experimento era esencial para determinar los umbrales, th_{phy} y th_{sp}, que convierten el conjunto de etiquetas binarias previstas durante una visualización de vídeo, en una única etiqueta

binaria para dicho periodo (véase la ecuación 5.2). Este paso era relevante para determinar si la arquitectura era más o menos propensa a las falsas alarmas, independientemente de la versión de Bindi que se considerara. Así, cada parámetro se barrió en el intervalo [0.3-0.6] con pasos de 0.1 mientras se generaban los 42 subsistemas monomodales correspondientes siguiendo el enfoque LASO. A este respecto, las Figs. 5.8a y 5.8b muestran los valores th_{phy} y th_{sp} frente a las métricas de precisión y F1-score media para los 42 grupos de prueba en los subsistemas fisiológico y del habla, respectivamente.



(A) Barrido th_{phy} (B) Barrido th_{sp}

Figura 5. 8: Barrido de parámetros para los subsistemas monomodales: thphy en el subsistema fisiológico y thsp en el subsistema del habla [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Analizando la Fig. 5.8a, observamos cómo la F1-score disminuye a medida que crece th_{phy}, mientras que la precisión permanece bastante estable. Obsérvese que la F1-score depende en gran medida del número de verdaderos positivos (TP) predichos, pero ignora sobre todo los verdaderos negativos (FN). Así, si los TP aumentan y la suma de los índices de falsos positivos (FP) y falsos negativos (FN) disminuye, la F1-score aumenta. Esta compensación ha provocado el comportamiento observado, en el que cuanto menor *es* th_{phy} , mayor es la F1-score. Según este análisis, th_{phy} se fijó en 0,40, obteniéndose un 66,66% y un 64,60% para la F1-score y la precisión, respectivamente. La razón de elegir este valor fue el buen compromiso observado entre ambas métricas y el hecho de que omitir un TP podría ser dramático para la GBV. El sistema multimodal combinado también debe abstenerse de disparar falsas alarmas para evitar abrumar a las instituciones encargadas de proteger a las usuarias, y por eso se eligió el subsistema del habla para ser más conservador en este sentido. La Fig. 5.8b muestra cómo la F1-score y la precisión empezaron a divergir a partir de 0.50 para el subsistema de voz. Por lo tanto, se fijó th_{sp} en este valor, obteniéndose un 54,07% y un 57,82% para la F1-score y la precisión, respectivamente. Tenga en cuenta que la precisión podría aumentarse eligiendo un th_{sp} más alto.

Una vez fijados$_{phy}$ y$_{sp}$, estudiamos la predicción del rendimiento medio en los 42 grupos de prueba para las distintas configuraciones de arquitectura, como se muestra en la Fig. 5.9. De izquierda a derecha, las configuraciones son: subsistema monomodal fisiológico, subsistema monomodal del habla, Bindi 1.0, Bindi 2.0a con fusión de datos de entropía mínima *(lowest entropy)*, Bindi 2.0a con fusión de datos de ponderación de entropía inversa (*inverse entropy weighting*), Bindi 2.0b con fusión de datos de entropía mínima *(lowest entropy)*, Bindi 2.0b con

fusión de datos de ponderación de entropía inversa (*inverse entropy weighting*), y Bindi 2.0b con fusión de datos OR lógica (*logical OR*). Nótese que el Bindi 2.0a no se combinó con fusión de datos OR lógica porque sería al equivalente al Bindi 1.0.

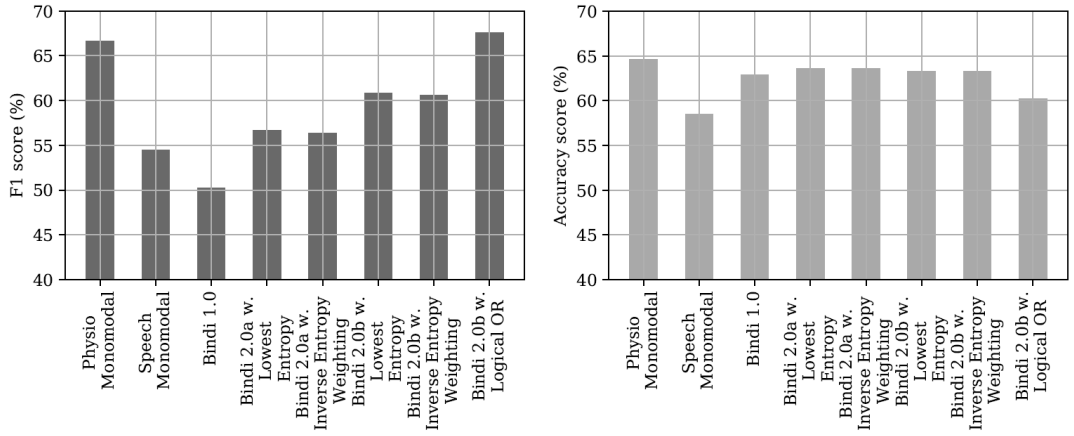


Figura 5. 9: Rendimiento medio utilizando la estrategia LASO para las distintas configuraciones de arquitectura: a) F1-score, b) Puntuación de precisión, [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

El subsistema monomodal fisiológico logró la mayor precisión, un 64,63%, superando incluso a los esquemas de fusión. Para la métrica de F1-score, este subsistema también proporcionó la segunda tasa más alta, un 66,67%. Este comportamiento podría estar relacionado en primer lugar con el sesgo introducido hacia la detección de la clase positiva con el coste de clasificación errónea del clasificador, y en segundo lugar con el barrido de parámetros de th_{phy} . El subsistema monomodal del habla proporcionó métricas significativamente más bajas que el subsistema fisiológico. Este hecho podría estar relacionado con el número limitado de muestras disponibles para entrenar la red neuronal y, posiblemente, con cierto desvanecimiento de la emoción elicitada en el momento en el que se grabaron als señales de habla. Esta situación hizo que Bindi 1.0 proporcionara las métricas más bajas, ya que la respuesta final del sistema depende del subsistema del habla. Tanto Bindi 2.0a como Bindi 2.0b proporcionaron precisiones similares cercanas a las del subsistema fisiológico en la mayoría de los casos. Sin embargo, Bindi 2.0b logró la F1-score más alta en todos los casos, especialmente con la fusión lógica de datos OR. Esta última estrategia proporcionó la F1-score más alta, del 67,59%, aunque la precisión fue limitada. Este rendimiento de la F1-score podría estar relacionado con el sesgo positivo aportado por los datos fisiológicos, ya que en este subsistema con el menor th_{phy} elegido, que introdujo un sesgo conservador hacia la no omisión de los TP a costa de aumentar los FP.

		Fisiológico Monomodal	Habla 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 Lógica OR
F1	media	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.	media	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

TABLA 5.2: Análisis del rendimiento medio para la predicción del reconocimiento binario del miedo en los 42 grupos de prueba independientes de la persona y adaptados al hablante [1].

Sin embargo, en cuanto a las otras arquitecturas con estrategias de fusión, el subsistema del habla puede haber deteriorado ligeramente el rendimiento del sistema en términos de F1-score y precisión, pero impidiendo que Bindi 2.0a y Bindi 2.0b produjeran demasiados FP. Además, se esperaba que la información auditiva desempeñara un papel importante en la detección de silencios, que podrían significar que la usuaria se encuentra en un estado de shock provocado por una situación de violencia de género, y proporcionara información acústica sobre el entorno. El significado y las consecuencias de estos indicadores sobre el rendimiento del sistema en la vida real deben analizarse a fondo a la luz de métricas más robustas, como en [286]. En la sección 5.6.2 damos un breve avance de este análisis y una discusión de las matrices de confusión obtenidas para cada configuración.

Para profundizar en los resultados mostrados en la Fig. 5.9, la tabla 5.2 presenta los resultados detallados de las distintas configuraciones, incluida la desviación típica media por voluntaria probada. Los índices bajos de desviación típica son buenos indicadores de una mejor capacidad de generalización siempre que los resultados sean comparables. Nótese, por ejemplo, que, aunque Bindi 1.0 presentó la desviación típica más baja (lo que podría considerarse una buena generalización) sus puntuaciones fueron superadas por la mayoría de las configuraciones, como ya se ha indicado. Además, puede observarse que los valores de desviación típica obtenidos son relativamente altos, especialmente para la F1-score. La causa se muestra en la Fig. 5.10, donde se proporcionan la F1-score y la precisión para cada una de las 42 pruebas y subsistemas monomodales. Puede observarse que algunos voluntarios tuvieron una F1-score de cero para el subsistema del habla. Esta situación se produce porque la F1-score depende de los TP detectados y no hubo predicciones positivas para algunas usuarias.

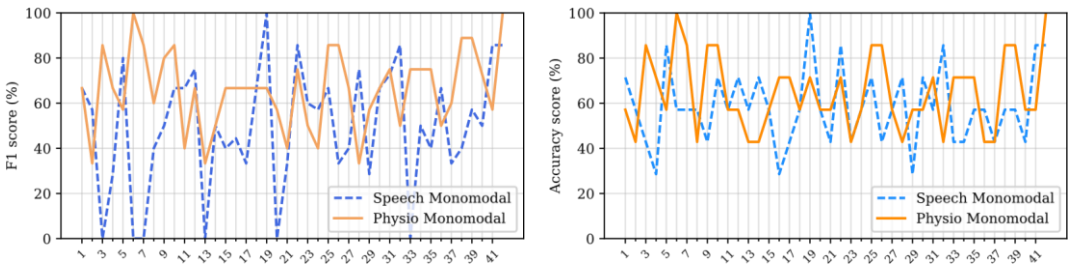


Figura 5. 10: Análisis del rendimiento individual para el reconocimiento binario del miedo para los dos subsistemas monomodales [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Matrices de confusión para los sistemas Monomodales y de Fusión

Las figuras 5.11, 5.12 y 5.13 muestran las matrices de confusión de las configuraciones evaluadas. En estas figuras, las filas corresponden a la clase predicha y las columnas a la clase verdadera o *ground-truth*. De izquierda a derecha y de arriba abajo, cada matriz de confusión muestra las tasas de TN, FP y *false omission* en la primera fila. La siguiente fila muestra las tasas de FN, TP y precisión. La última fila muestra la tasa FN, la especificidad y la precisión global.

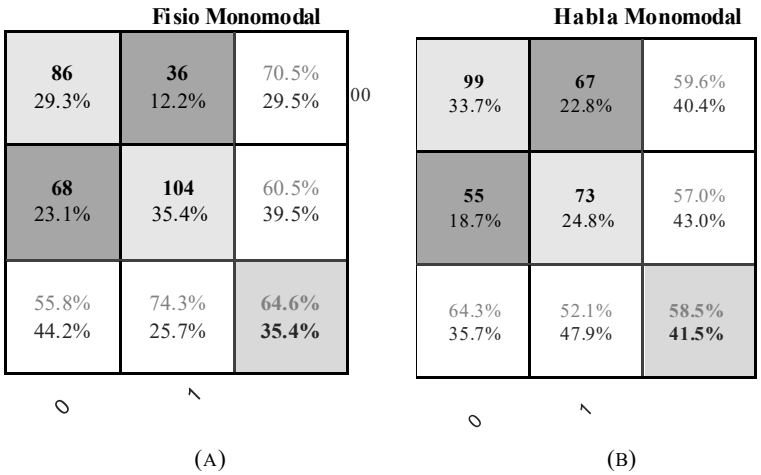


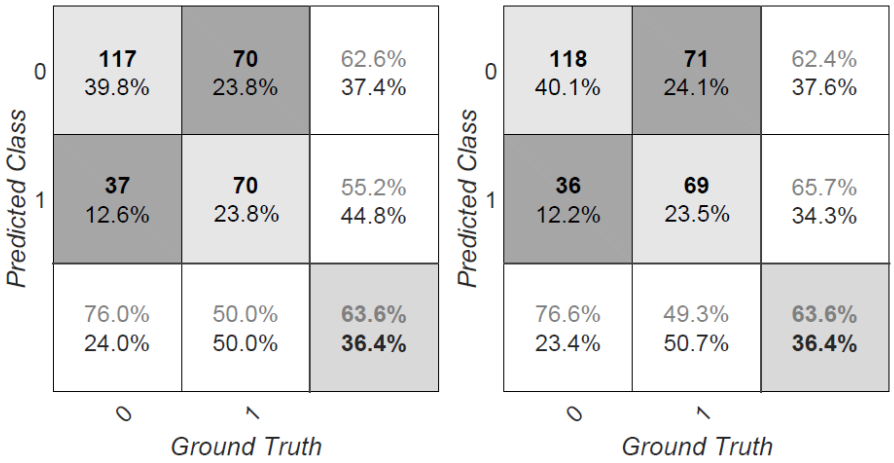
Figura 5. 11: Matrices de confusión monomodales para la detección binaria del miedo [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

La matriz de confusión del subsistema fisiológico refleja su tendencia a predecir la clase positiva a costa de perderse TN. Por su parte, el subsistema monomodal del habla obtiene tasas globales más bajas que los demás, pero logra una tasa de TN más alta. Encontrar un equilibrio entre estos dos comportamientos es muy importante en nuestra aplicación, donde las alertas perdidas pueden ser dramáticas para las usuarias, pero activar demasiadas alertas falsas podría abrumar a las instituciones encargadas de la protección. Por ello, el hecho de que el subsistema del habla pueda contener las FP activadas por el sistema monomodal fisiológico parece muy prometedor. En esta línea de trabajo, las estrategias de fusión cuyas matrices de confusión se muestran en las Figs. 5.12a, 5.12b, 5.13a y 5.13b sólo difieren en un par de casos, pero están más equilibradas entre TN y TP. Sin embargo, la estrategia mostrada en la Fig. 5.13c refleja tasas de FP y TP mucho más altas que las demás pero falla más TN que ninguna otra, y la Fig. 5.12c muestra cómo la toma de decisiones jerárquica de Bindi 1.0 tiene un rendimiento pobre, lo que demuestra que la fusión es realmente esencial.

Discusión

En cuanto a la arquitectura habitual de capas IoT (*edge*, *fog* y *cloud*) considerada en Bindi, una cuestión relevante del diseño del sistema se refiere a qué parte del sistema debe implementarse en cada una de las capas.

En primer lugar, la capa de computación *cloud* está pensada para recopilar y procesar grandes cantidades de datos sin limitaciones en cuanto a recursos informáticos, demanda de energía o tiempos de respuesta [287]. Esta definición se ajusta a las necesidades de los servicios informáticos centralizados de Bindi, que se sitúan por tanto en la capa de computación en nube para gestionar posibles pruebas delictivas e información histórica para el seguimiento a largo plazo de las víctimas.



(A, Izquierda) Bindi 2.0a. Fusión de entropía mínima (B, Derecha) Bindi 2.0a. Matriz de confusión de entropía inversa. matriz de confusión de fusión ponderada.

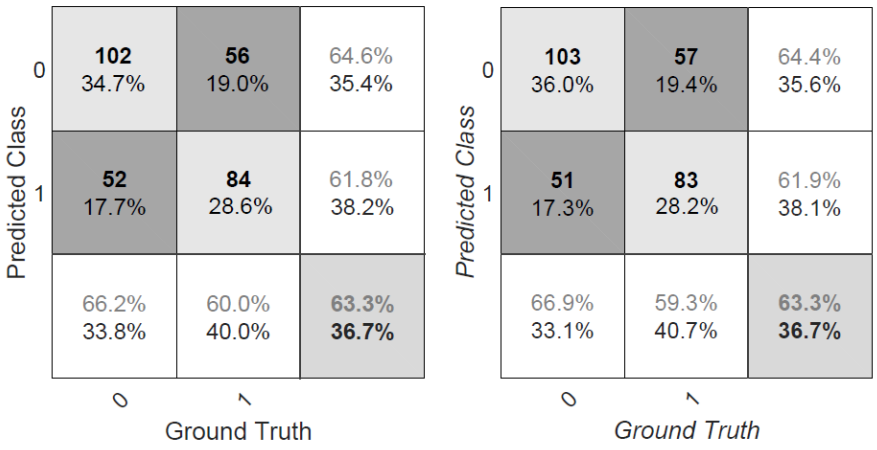
(C) Matriz de confusión Bindi 1.0.

Figura 5. 12: Matrices de confusión para estrategias de fusión de datos para Bindi 2.0a y Bindi 1.0 [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

En segundo lugar, la computación *edge* tiene lugar en los nodos IoT que capturan datos en el borde de la red. Estos dispositivos están limitados por sus capacidades informáticas y energéticas porque, en la mayoría de los casos, funcionan con baterías [288]. Esta definición encaja con los dispositivos mediante los que se capturan datos fisiológicos y auditivos a lo largo del tiempo en Bindi, es decir, una pulsera y un colgante.

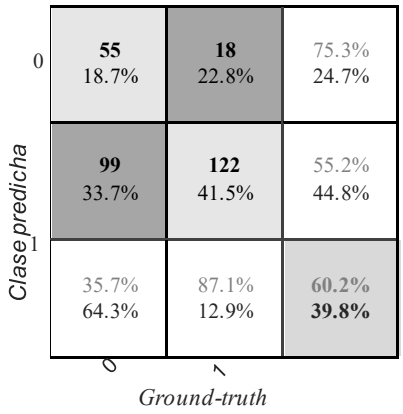
Por último, la capa de computación *fog* sigue un concepto similar al de la capa de computación *edge*. Sin embargo, los dispositivos *fog* están menos limitados en cuanto a capacidades informáticas y energéticas, al tiempo que permanecen cerca del origen de los datos [289]. Según

esta descripción, el *smartphone* de Bindi puede considerarse un dispositivo de *fog* porque no captura datos pero está cerca del origen de los datos, y tanto las capacidades de computación como las energéticas están menos limitadas que las de los dispositivos de *edge* (la pulsera y el colgante). Algunos autores afirman que la capa *fog* no existe, y entonces implementan las funcionalidades de la capa de *fog* descritas antes, dentro de la capa de *edge* [290]. Bajo este enfoque, sigue siendo posible estructurar los dispositivos en diferentes capas dentro del *edge*. Desde este punto de vista, el *smartphone* estaría en una capa superior dentro del *edge*, mientras que la pulsera y el colgante constituirían la capa inferior. Para profundizar y revisar las capas del *edge*, la *fog* y la *cloud*, se remite a los lectores a [291] y [292].



(A, Izquierda) Matriz de fusión de menor entropía de confusión

(B, Derecha) Matriz de confusión de fusión de ponderación de entropía inversa.



(C) Matriz de confusión de fusión de la función OR lógica.

Figura 5. 13: Matrices de confusión de las estrategias de fusión de datos para Bindi 2.0b [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Las técnicas de fusión de datos propuestas en este trabajo alcanzaron un máximo de hasta el 63,61% de precisión media para un caso de uso de reconocimiento del *miedo* independiente de la persona y adaptado al hablante. Este resultado se obtuvo utilizando señales multimodales del habla y fisiológicas y el enfoque de la estrategia de fusión de entropía más baja. La precisión

media obtenida se situó dentro del rango de los índices de precisión alcanzados por trabajos similares presentados en la sección 5.2.4 y superó al sistema propuesto en [280], que consideraba las mismas fuentes multimodales de información. Cabe destacar que, como característica diferenciadora de nuestro sistema, hacemos uso de la monitorización no invasiva de señales, en lugar de cascos de EEG o sensores de detección facial [276, 257]. Además, el número de usuarias consideradas (es decir, 42), proporciona una mayor variabilidad en los datos y, por tanto, produce un modelo más robusto.

Cabe destacar que las configuraciones aquí descritas para la detección del *miedo* a través de datos fisiológicos y del habla son sólo posibles formas de caracterizar las situaciones y contextos en los que podrían verse envueltos las usuarias de Bindi. Pretenden ser unos estudios *baseline* iniciales para futuros desarrollos y han permitido identificar importantes retos. Para empezar, encontrar un equilibrio adecuado entre TP y TN y PF y FN es crucial, ya que el coste de pasar por alto una verdadera necesidad de ayuda no es una posibilidad, pero también debemos evitar interferir en la vida cotidiana de las víctimas de violencia de género y saturar los servicios de protección con falsas alarmas.

Así pues, en este trabajo, intentamos reducir los FN en la medida de lo posible, mientras que los FP se mantenían en una tasa adecuada. Para ello, consideramos estrategias basadas en los costes de clasificación errónea y en la fijación de parámetros umbral. En concreto, fijamos *el* th_{phy} en el subsistema fisiológico para obtener un mayor resultado de predicciones positivas con este sistema para que, en una fase posterior, el habla (en Bindi 1.0) y las estrategias de fusión de datos (en Bindi 2.0a y Bindi 2.0b) ayudaran a corregir el sesgo mientras se intentaba mantener la predicción TP. Durante esta experimentación, el sistema monomodal de habla actual proporcionó tasas de rendimiento inferiores a las esperadas. Una posible explicación de este comportamiento podría ser la desalineación temporal de los datos fisiológicos y del habla en WEMAC. El desvanecimiento de la emoción elicitada en el momento en que se recoge la muestra de voz podría estar detrás de esta disminución del rendimiento. Además, sólo se han utilizado las técnicas clásicas de procesamiento y clasificación como línea de base para futuras exploraciones con este novedoso conjunto de datos. Una situación similar se aplica a las estrategias de fusión, concebidas para comprobar la fiabilidad de las pre-alarmas activadas por el modelo fisiológico y que actúan como moduladores para reducir la tasa de predicción de la clase FP.

En cuanto a los trabajos futuros, este estudio abre la puerta a nuevas investigaciones en muchas direcciones. Por ejemplo, el uso de redes neuronales recurrentes para explotar el contexto temporal de las señales, el análisis de otras alternativas de fusión o la evaluación de métricas de puntuación alternativas, como la información mutua o el área bajo la curva, podrían utilizarse para seguir encontrando un equilibrio adecuado entre las falsas alarmas y la probabilidad de fallo. Además, añadir datos adquiridos de más voluntarias en condiciones de laboratorio añadiría

solidez a los modelos. Del mismo modo, incluir datos de GBVV ayudaría a comprender mejor los mecanismos de activación de la GBVV en situaciones relacionadas con el *miedo*. Por último, cabe señalar que el desarrollo de técnicas de adaptación a cada usuaria es fundamental para nuestro caso de uso de la violencia de género.

5.7. Conclusiones

En este capítulo evaluamos la detección y clasificación de emociones relacionadas con el *miedo* desde una perspectiva multimodal para el desarrollo del sistema Bindi. Cabe destacar el alto grado de multidisciplinariedad del presente trabajo ya que las aportaciones se realizaron conjuntamente con otros miembros del equipo UC3M4Safety.

En la Sec. 5.3 describimos en profundidad los componentes y el funcionamiento del sistema -pulsera, colgante, app y servidor-. Acontinuación, en la Sec. 5.4, describimos la evolución de la arquitectura de prueba de concepto de Bindi 1.0 a Bindi 2.0. Analizamos los *pipeline* de datos monomodales disponibles y propusimos una arquitectura híbrida de fusión de datos combinando los enfoques a nivel de decisión (*late*) y a nivel de características (*early*) basados en la combinación de señales fisiológicas y de audio para detectar situaciones de violencia de género. Esta novedosa arquitectura incluye una tercera capa (es decir, fusión temprana) al sistema implementado de antemano en Bindi, aún por validar. Es necesario abordar técnicas de fusión alternativas, adaptadas específicamente a este problema y capaces de tener en cuenta sus limitaciones inherentes, como la necesaria optimización del ancho de banda, incluida la compresión de datos, las limitaciones de hardware y computacionales, y las compensaciones por el consumo de batería. Más adelante, en la sección 5.4, damos más forma a las arquitecturas de fusión para Bindi 2.0 y describimos la teoría en la que se basa cada arquitectura de fusión ideada. También detallamos cada uno de los conductos de procesamiento de datos -fisiológicos y del habla- en la Sec. 5.5.

En cuanto a la sección 5.6.1, los eventos acústicos estresantes con una correlación no determinista con el habla estresada demostraron ser beneficiosos hasta cierto punto para las clasificaciones del habla emocional. Este estudio deja muchas preguntas abiertas y líneas de trabajo futuras. La librería de programación utilizada para crear las mezclas sintéticas permite definir distribuciones de probabilidad para la aparición y duración de los eventos sonoros -como el procedimiento descrito en la Sec. 3.2.3-. Yestá preparada para realizar la adición de eventos de fondo cuando la etiqueta binaria es *Q2*. Y así los datos también podrían ampliarse procediendo de forma similar con eventos no estresantes siempre que la etiqueta binaria no sea *Q2*, haciendo que la mezcla resultante suene más realista. También los sonidos de fondo en el proceso de mezcla pueden adaptarse a cualquier tipo de problema, dando lugar a nuevas combinaciones de la BioS-DB y

otros conjuntos de datos. Dado que el objetivo principal de Bindi es detectar y prevenir la violencia de género, estos eventos de fondo podrían corresponder a clips de audio de escenas de películas que representen un escenario de violencia de género, seleccionados con el conocimiento y la orientación de expertos.

Por último, en relación con la Sec. 5.6.2, presentamos Bindi 2.0, un sistema multimodal autónomo de extremo a extremo que aprovecha la computación afectiva a través de sensores inteligentes comerciales listos para usar auditivos y fisiológicos, la fusión jerárquica de señales multisensoriales y una arquitectura de servidor segura, con el objetivo final de proporcionar seguridad y garantizar el bienestar de las víctimas de violencia de género. En concreto, en la sección 5.4.3 se propusieron tres arquitecturas de sistema para Bindi, consistentes en disposiciones específicas de los subsistemas de procesamiento de datos desarrollados, es decir, los subsistemas fisiológicos, de voz y de fusión de datos en el futuro próximo de Bindi. Estas arquitecturas se validaron y evaluaron utilizando el conjunto de datos WEMAC perteneciente a la base de datos UC3M4Safety. Nótese que el conjunto de datos se construyó específicamente para detectar el *miedo* en las mujeres en un entorno de laboratorio.

Los resultados experimentales muestran una precisión media de la tasa de reconocimiento del *miedo* de hasta el 63,61% con el método Leave-hAlf-Subject-Out (LASO). Las métricas obtenidas están en consonancia con sistemas multimodales similares del estado del arte, como los revisados en la sección 5.2.4. Además, nuestro sistema supera al único sistema de la literatura que trata la misma combinación bimodal que en este trabajo [280]. Que sepamos, es la primera vez que se presenta un modelo LASO que tiene en cuenta el reconocimiento del *miedo*, la fusión de señales multisensoriales y los estímulos de realidad virtual. Nótese que la importancia de los resultados está limitada por el número de participantes en el momento de la publicación [1], es decir, 47 mujeres.

Esta experimentación sirve como enfoque multimodal inicial para trabajar con el *miedo* real elicitado en las mujeres y su procesamiento adecuado. Bindi es un sistema muy complejo que requiere un minucioso equilibrio de muchos aspectos, como el consumo de batería, la potencia de cálculo, el uso de recursos y el rendimiento del algoritmo. Pretendemos señalar que el objetivo último de este trabajo es despertar el interés de la comunidad por desarrollar soluciones al problema tan desafiante de la violencia de género.

Todo este trabajo de reconocimiento de las emociones del *miedo* y las conclusiones recogidas pretenden allanar el camino y dar forma a la próxima versión de Bindi: Bindi 3.0.

Capítulo 6: Otras líneas de Investigación sobre el Audio y la Violencia de Género

Al mismo tiempo que realizábamos nuestra investigación en esta tesis, se abrieron líneas de investigación paralelas pero complementarias que podrían ayudar en la prevención de la violencia de género utilizando la modalidad auditiva. Al principio de este capítulo, hablamos de la caracterización afectiva del contexto acústico, en el contexto de la detección de situaciones de riesgo de violencia de género, incluyendo primero el análisis de eventos acústicos y después el análisis holístico de escenas o escenarios acústicos. Después, exploramos superficialmente el análisis de la fatiga en el habla, viendo que puede estar relacionado con el estrés en el habla. A continuación, realizamos un estudio preliminar de ablación sobre la detección de la condición de víctima de violencia de género sólo mediante el uso de datos del habla. Y por último, se comenta generalmente la relación entre el cambio climático y la violencia de género.

Estas líneas de investigación no forman parte del grueso de la tesis, pero nos pareció que eran campos importantes para investigar y que podían aportar información y contribuir a la prevención de la violencia de género a través de la tecnología de audio.

6.1. Caracterización Afectiva del Contexto Acústico

Dentro de la señal de audio que capturamos con Bindi tenemos varias fuentes de información, entre ellas: el habla de la usuaria, los silencios, el ruido ambiental, los eventos acústicos, los sonidos auxiliares, etc. Uniendo todas las fuentes podemos hacernos una idea del contexto en el que se encuentra la usuaria. En esta tesis hemos trabajado especialmente en la identificación del hablante y la detección de emociones en la voz principalmente, pero parece razonable pensar que los eventos acústicos y el ruido, al formar el contexto acústico, podrían darnos más información sobre la situación en la que se encuentra la usuaria. También nos interesa investigar la relación entre las escenas acústicas y las emociones que pueden suscitar.

Tenemos en cuenta esta modalidad acústica, porque todas las modalidades por sí solas son demasiado frágiles para dar un resultado fiable de la predicción de una situación de riesgo. Así que no podemos fijarnos sólo en una, sino que todas ellas contribuyen a una predicción más sólida y fiable. Por eso también es difícil aislar la detección de emociones o la identificación del hablante del análisis del resto del audio, y entre sí. Siempre están entrelazados en este reto de la violencia de género.

El estudio de la caracterización de los eventos y escenas acústicas para la detección de situaciones de riesgo de violencia de género es un campo muy desafiante y complejo, que requeriría una tesis

doctoral aparte en sí misma, pero queríamos hacer un trabajo preliminar inicial que pudiera iluminar el camino a seguir.

6.1.1. Caracterización de los Eventos Acústicos Afectivos

El campo de la detección de eventos acústicos (AED) es un campo de investigación de la AI en el que se han desarrollado y utilizado diferentes enfoques para la detección de eventos acústicos, a menudo imitando el sistema auditivo humano, e incluyendo diferentes conjuntos de características y algoritmos de detección. La detección de sonidos puede ayudarnos a caracterizar emocionalmente una señal de audio, relacionando los eventos acústicos que aparecen con la emoción que los audios tienden a elicitar.

Además, la comunidad DCASE[33] viene publicando desde 2013 varios conjuntos de datos para la "detección y clasificación de escenas y eventos acústicos". Esto ha fomentado una gran cantidad de contribuciones de investigación en este campo. Además, un conjunto de datos a gran escala de etiquetas manuales de eventos acústicos, AudioSet [170], desencadenó la investigación en modelos de aprendizaje profundo, varios abiertos a la comunidad investigadora como YAMNet [242]. Esto ofrece una alternativa robusta para la representación del entorno acústico que puede transferirse a otros dominios y tareas.

En esta sección presentamos el sistema de detección de eventos acústicos propuesto como prueba de concepto para Bindi 2.0, que se incluirá en Bindi 3.0.

Subsistema de información acústica en Bindi 2.0

Esta sección describe la *pipeline* preliminar de procesamiento de audio para la detección acústica de amenazas en la escena desarrollada por otros miembros del equipo UC3M4Safety. Lo utilizamos para proporcionar una caracterización afectiva de WEMAC en la sección 6.1.2. Este componente aún no se ha incluido en los dispositivos estudiados en [1] y se transmite aquí como prueba de concepto para futuras versiones de Bindi. Este subsistema se basa en la arquitectura presentada en [8]. Su tarea principal es detectar si los eventos sonoros registrados desde el micrófono representan una amenaza para la seguridad de la usuaria según nuestro caso de uso.El sistema de detección de eventos acústicos comienza procesando la señal de audio. En primer lugar, la señal de audio se normaliza, igual que para la canalización del habla (véase la sección 5.5). En segundo lugar, se computa un espectrograma log-Mel para obtener una representación tiempo-frecuencia de la señal en forma de imagen para alimentar posteriormente la red de detección de eventos. Así, se computa un espectrograma inicial mediante una transformada *short-time* de Fourier (STFT) con los siguientes parámetros: un tamaño de ventana de *25* ms, salto de

[33] https://dcase.community/

ventana de *10* ms y ventana *Hanning*. La dimensión de frecuencia del espectrograma se asigna a 64 bins Mel para cubrir las frecuencias que van de 125 a 7500Hz y la amplitud se transforma en una escala logarítmica con un *offset* de 0,001.Los espectrogramas tomados como características se encuadran en *frames* de 0,96 segundos con un solapamiento del 50%. Cada ejemplo abarca 96 fotogramas de 10 ms cada uno y 64 bandas de frecuencia Mel. Por lo tanto, las dimensiones de estas características son 96x64. Las características resultantes se introducen en una red neuronal convolucional (CNN) preentrenada para detectar los eventos de audio en una escena.

El modelo seleccionado para esta tarea es YAMNet. En concreto, se considera la arquitectura de convolución separable en profundidad MobileNet_v1 [293]. Este modelo ha sido preentrenado en 521 clases del corpus AudioSet YouTube [170], una base de datos de clasificación de eventos sonoros multietiqueta de uso general, y está preparado para realizar inferencia de detección de eventos acústicos. El rendimiento de este tipo de redes se ha estudiado ampliamente en el campo de la detección de eventos sonoros [294].

El procedimiento para alimentar la red es el siguiente: En primer lugar, los parches de 96 × 64 de la etapa de extracción de características se transforman en una matriz de 3 × 2 para los 1024 núcleos de la capa convolucional superior. Tras ser procesados a través de las capas de extracción de características, estos ejemplos se promedian para obtener un *embedding* de 1024 dimensiones. A continuación, una capa logística realiza la clasificación en 521 clases objetivo.

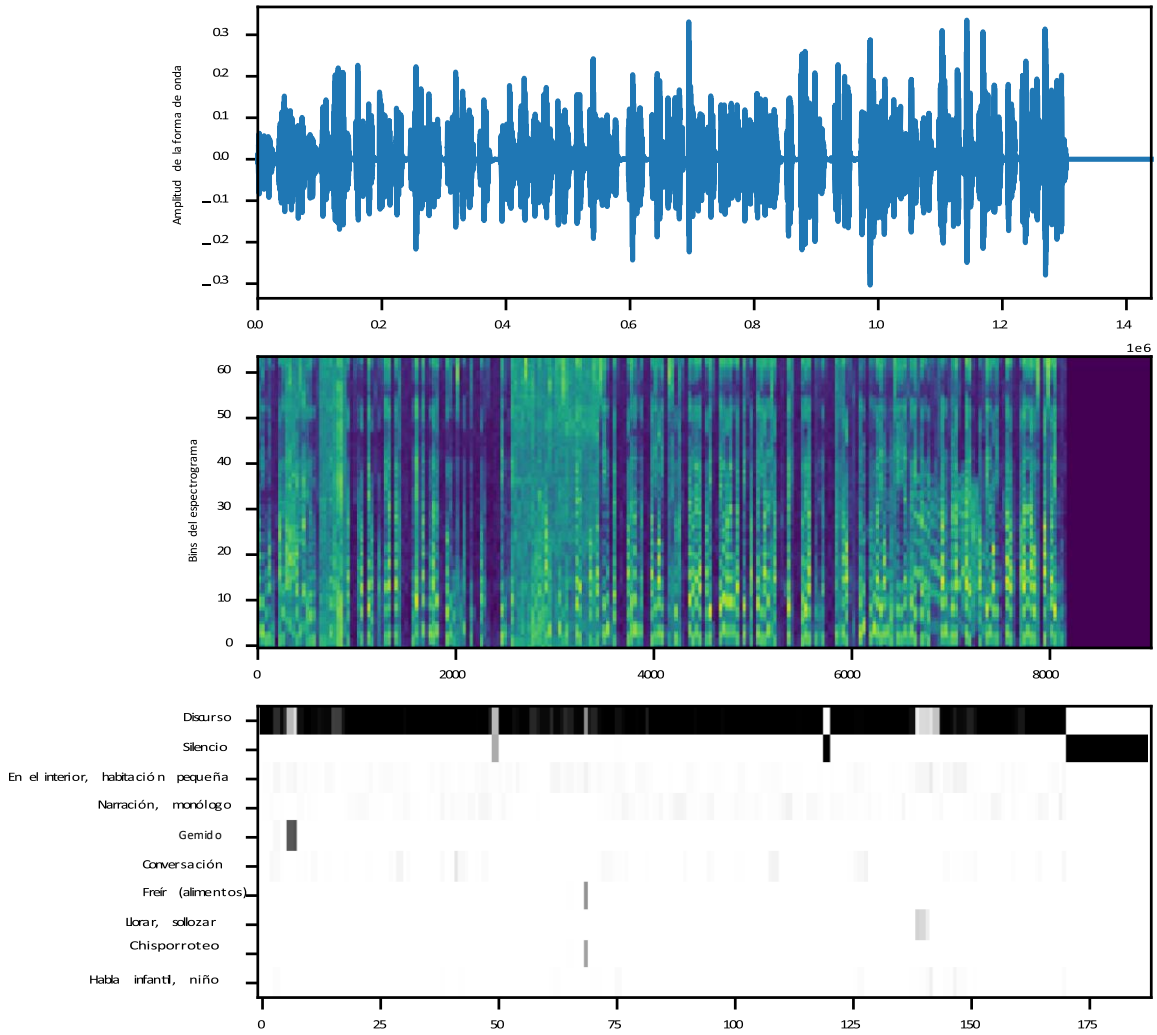


Figura 6. 1: YAMNet procesando una muestra de BioSpeech+. Representación temporal (arriba), espectrograma con bandas que van de 125 a 7500Hz (centro), y principales eventos encontrados (abajo) [8]. Reproducido con permiso del propietario del copyright, ISCA.

En cuanto a la tarea de detección y clasificación de eventos acústicos (AED/C), nos interesa analizar el conjunto de estímulos utilizados en WEMAC. En la Fig. 6.1 observamos el rendimiento de YAMNet clasificando un audio mixto de *90s* como parte de un pequeño análisis informal para la identificación de eventos acústicos presentes en una señal de audio de la base de datos Biospeech+ generada (véase la Sec. 3.2.3) [8]. El conjunto de datos Biospeech+ se preprocesa para que cumpla los requisitos de YAMNet (f_s = *16KHz*, mono, amplitud normalizada a [-1, 1]) y luego se introduce en el modelo. El único parámetro libre es patch_hop, que se fijó en 0.48s.

Este análisis pretende caracterizar el problema de la detección de la GBV desde una perspectiva acústica, ya que el desarrollo de una descripción empírica del problema es importante para su detección automática. Así, se aplicó el subsistema de información acústica a la señal de audio de los estímulos audiovisuales en WEMAC para analizar los eventos acústicos que conforman cada escena acústica en el contexto de la GBV.

Los resultados obtenidos aparecen en la Fig. 6.11, donde se representan todas las apariciones de las etiquetas de eventos acústicos de YAMNet en los estímulos audiovisuales de WEMAC. Curiosamente, algunas etiquetas se encontraron exclusivamente en los estímulos audiovisuales de *miedo*, como *los latidos del corazón*, las *explosiones* y la *respiración*, mientras que otras etiquetas nunca aparecieron para el *miedo*, como la *música tierna*, *las nanas* y el sonido de *multitudes*. También hubo casos intermedios en los que aparecen etiquetas para ambos tipos de estímulos, como etiquetas de contextualización espacial (relacionadas con interiores o exteriores), *animales*, *silencio* y *risa*. Por tanto, la clasificación automática de eventos acústicos parece prometedora, ya que pueden deducirse ciertos patrones a partir de casos extremos en los que aparecen exclusivamente etiquetas para uno de los dos tipos de estímulos audiovisuales. Hay que tener en cuenta que las etiquetas YAMNet son muy generales en sí mismas, es decir, pueden aparecer relacionadas con muchas circunstancias y escenas. Por lo tanto, deben analizarse como un conjunto, lo que es una forma factible de inferir algunas cualidades del contexto de una escena concreta, por ejemplo, la violencia.

A partir de este análisis exploratorio, podemos concluir que la información extraída de los eventos acústicos puede ser muy beneficiosa para desambiguar las posibles situaciones de violencia de género detectadas automáticamente en Bindi con el resto de sensores. Los eventos sonoros circundantes de una escena pueden ayudar a inferir su contexto, que es fundamental para determinar si la escena es o no violenta. Así pues, esperamos que el subsistema de información acústica desempeñe un papel clave en la evaluación de WE-LIVE, en la que las voluntarias realizan actividades cotidianas, fuera del entorno del laboratorio.

6.1.2. Caracterización Afectiva de la Escena Acústica

Tras el análisis de los eventos acústicos, en esta sección presentamos el trabajo preliminar realizado junto con otros miembros del equipo UC3M4Safety en el estudio de las escenas acústicas y los paisajes sonoros emocionales (*emotional soundscapes*). En la sección anterior describimos la detección de eventos acústicos sin relacionarlos entre sí, pero en esta sección queremos analizarlos conjuntamente para poder caracterizar una escena acústica holística de forma afectiva.El análisis y la interpretación de escenas acústicas es un campo de investigación que pretende explicar la información acústica del entorno captada a menudo por un sistema de adquisición multimicrófono [295]. Aunque existen en la literatura algunos trabajos sobre la relación entre las escenas acústicas y las emociones, no se han identificado colectivamente ni definido de forma específica. No existe un único título o acrónimo, como ocurre por ejemplo con el campo ampliamente conocido del reconocimiento de las emociones del habla (SER), en el que se está desarrollando un sólido corpus de trabajo. Así, encontramos trabajos relacionados con las

escenas acústicas y las emociones bajo diferentes nombres: "evaluación de entornos acústicos por las emociones", "emociones en paisajes sonoros" [296], "comprensión de escenas (acústicas) emocionales", "emociones inducidas en la sonificación" [297], "reconocimiento de emociones mediante eventos sonoros generales", "teoría del diseño sonoro" [298] o "diseño acústico de entornos virtuales" [299], entre otros. Sin embargo, a pesar del escaso número de trabajos, sigue habiendo investigaciones prometedoras en este campo. La motivación de estos trabajos en la literatura es dotar a las máquinas de la capacidad de comprender lo que experimenta una persona desde su marco de referencia acústico. Esto incluye su información contextual acústica, es decir, la situación y el entorno auditivo de la persona. Y nuestro propósito con este trabajo [4] es proporcionar una visión global de este subcampo, reuniéndolo bajo el nombre de "Análisis Afectivo de Escenas Acústicas" (AASA).

Este innovador trabajo [300] pretendía desarrollar modelos informáticos exhaustivos del afecto en el sonido. En él, un alto grado de coherencia entre dominios indicaba que la codificación de las dos dimensiones principales de la emoción (*arousal* y valencia) resultaba de la evolución de la voz y la música juntas de forma multimodal, incluyendo la combinación de sonidos de la naturaleza para conseguir efectos expresivos. Sin embargo, estos hallazgos se establecieron sobre la base del habla emocional actuada y espontánea, la música y los eventos sonoros generales [301] de forma aislada. Con el objetivo de crear un modelo holístico capaz de explicar el afecto que elicitan los sonidos, pretendemos caracterizar las emociones elicitadas al estar una persona inmersa una escena acústica específica, teniendo en cuenta la información acústica del entorno en su conjunto.

La representación subyacente común de los desencadenantes de emociones a partir de sonidos, música y habla se discute en [300], pero a pesar de la abundante literatura, que apunta hacia la relevancia del entorno acústico y las emociones humanas en las ciencias cognitivas (por ejemplo,[302]), hay muy pocos estudios que investiguen la relación entre los eventos acústicos y la elicitación de emociones [301] y apenas ninguno investiga la relación entre *el miedo* y los sonidos [303].

Como hemos mencionado anteriormente, nos interesa especialmente analizar cómo los entornos acústicos del mundo real pueden afectar e influir en las emociones y, por tanto, analizarlos y caracterizarlos. Estas tareas podrían englobarse en un subcampo de la computación afectiva que denominamos "Análisis Afectivo de Escenas Acústicas".

Que los autores sepan, ningún trabajo anterior se centra en la información emocional intrínseca de un paisaje sonoro y propone un método para encontrar relaciones directas y no supervisadas entre los eventos de audio de una escena acústica y su emoción elicitada. Por ello, en la siguiente subsección, presentamos una metodología para el *análisis de la escena acústica afectiva* y, a continuación, adoptamos una configuración basada en los métodos clásicos de recuperación de

información para producir una representación de la *escena acústica afectiva* basada en el conocido algoritmo TF-IDF (término-frecuencia-frecuencia inversa del documento) [304], [305], en el que construimos el espacio vectorial de los eventos acústicos que se producen en una escena equilibrando la *frecuencia del evento acústico* y la *frecuencia inversa de la escena*.

Metodología para un enfoque basado en la recuperación de información

En esta sección detallamos paso a paso nuestra metodología propuesta para el *Análisis Afectivo de Escenas Acústicas (*AASA) [4]. Proponemos que se trata de una alternativa más completa a la configuración clásica de aprendizaje automático que extrae características de las señales de audio y luego las introduce directamente en un modelo de aprendizaje automático para su inferencia, lo que también facilita la interpretabilidad. La Fig. 6.2 ilustra esta metodología en un diagrama de bloques.

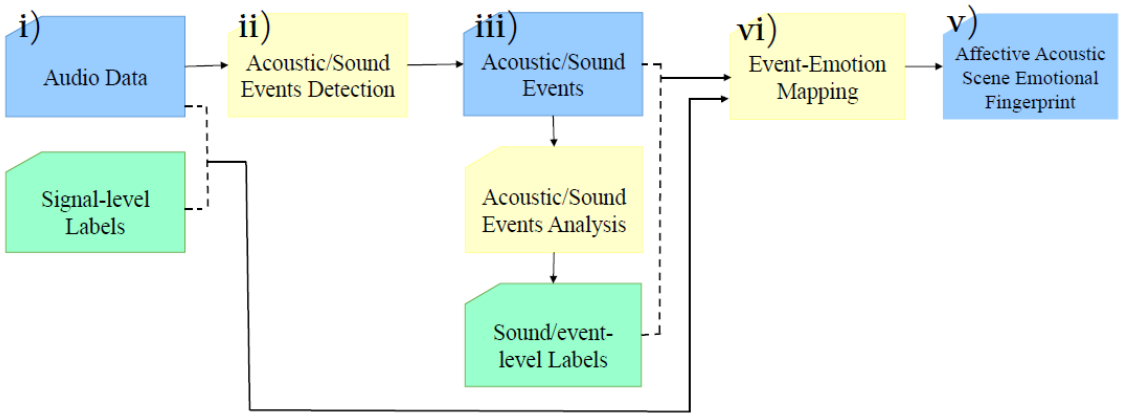


Figura 6. 2: Diagrama de bloques de la metodología de análisis de escenas acústicas afectivas [4].

Empezando por los bloques de la izquierda, i) el primer paso consiste en utilizar datos de audio, especialmente útiles si se graban en condiciones realistas o en una mezcla sintética que imite dichas condiciones (por ejemplo, un entorno de realidad virtual, un clip de película, un videojuego realista, etc.). Idealmente, estos datos se etiquetarían según los estados afectivos o las emociones percibidas por las personas que los escuchan (o estan inmersas en esas señales de audio) activamente. Podrían utilizarse etiquetas afectivas como *arousal*, valencia o dominancia, placer, emociones categóricas e interés. El objetivo de estas etiquetas es reflejar la emoción o el estado afectivo percibido por una persona que está inmersa en ese entorno acústico.

Como paso siguiente, que podría ser opcional, ii) puede aplicarse un módulo de detección o clasificación de eventos acústicos, que identifique los eventos acústicos o sonidos a partir de la señal de audio. Dicho módulo podría ser un modelo de aprendizaje automático preentrenado con bases de datos que incluyan etiquetas emocionales de los sonidos, iii) de modo que podría encontrarse una relación o alineación entre los sonidos detectados y su componente emocional etiquetado.

Una vez que los eventos o sonidos acústicos tienen una etiqueta emocional correspondiente, iv) es necesario analizar el mapeo entre ambos, de forma supervisada o no supervisada, con un algoritmo que pueda evaluar la relación entre los eventos o sonidos acústicos y las etiquetas emocionales. Este paso puede realizarse con cualquier par o etiqueta de datos, por ejemplo, los eventos acústicos separados junto con toda la etiqueta emocional original a nivel de señal. Por último, v) se extrae del análisis una huella emocional acústica (*emotional acoustic fingerprints)* o *embedding*, que condensa la información emocional del audio analizado.

Un reto importante que se plantea está relacionado con la intensidad del acontecimiento emocional, es decir, su saliencia emocional. Esta es una señal biológica adaptativa que influye en cómo se recuerdan los acontecimientos y cómo se incorporan a la memoria.

Además, diferentes sonidos o entornos acústicos pueden provocar diferentes emociones en los oyentes en función de su experiencia previa y de las asociaciones de memoria que los sonidos evocan. Así, puede haber una reacción emocional mayoritaria, pero no debemos olvidar las diferencias individuales de cada persona, concretamente en el caso de las mujeres que han sufrido o sufren violencia de género.

Como ya hemos mencionado, el componente de memoria asociativa y relacional de un acontecimiento acústico también puede desempeñar un papel en la reacción emocional de una persona. El sonido de unas llaves abriendo una puerta puede ser un sonido de alegría que signifique dar la bienvenida a una persona querida, pero para una víctima de violencia de género puede significar que ha llegado su maltratador. El efecto emocional puede ser completamente diferente aunque el acontecimiento acústico sea el mismo. Por ello, la necesidad de un método que pueda adaptarse y personalizarse es de suma importancia en este campo con un nivel tan alto de subjetividad.

Como parte del diagrama de bloques representado en la Fig. 6.2, opcionalmente podemos aspirar a clasificar los eventos acústicos que se producen en los datos de audio disponibles. Para dicha clasificación, podríamos emplear modelos de clasificación de eventos acústicos preentrenados, capaces de detectar eventos o sonidos acústicos. Nos referimos a este paso como opcional porque también podría realizarse un análisis directo de la señal acústica completa y su etiqueta emocional, pero creemos que este paso es clave para identificar los eventos acústicos que componen una señal de audio, de modo que nuestra interpretación posterior sea más transparente y directa, más explicable.

La parte central de la metodología es un algoritmo que analiza la relación entre los eventos acústicos o sonidos y las emociones elicitadas. De alguna manera, tenemos que extraer de las señales de audio los momentos más destacados o relevantes, que nos permitan condensar la información, dentro de la señal de audio, que puede desencadenar una emoción en un oyente. Una forma sería extraer características de audio de la señal sonora, como si se tratara de una tarea de

reconocimiento de emociones del habla, para su posterior clasificación emocional mediante modelos de predicción ML.

Otro tipo de proceso que se puede utilizar y que es el que empleamos en nuestro caso de uso, es el algoritmo TF-IDF, como explicaremos en la siguiente subsección.

Una vez extraídas los embeddings o huellas emocionales acústicas (*emotional acoustic fingerprints*), podrían utilizarse como entrada para modelos de aprendizaje automático. Estos pueden ser supervisados -reutilizando las etiquetas emocionales como etiquetas de *ground-truth*, como para los modelos de regresión o clasificación ML- o no supervisados, utilizando algún tipo de agrupación (*clustering*) o métrica de similitud que se aplique. Consideramos también clave que los resultados puedan visualizarse, con la ayuda de modelos de explicabilidad (XAI), para verificar e interpretar la responsabilidad de los resultados recogidos.

Configuración experimental del conjunto de datos de estímulos audiovisuales UC3M4Safety

Para inferir el espacio de *embeddings* de las emociones, utilizamos el conjunto de datos de estímulos audiovisuales UC3M4Safety, parte del conjunto de datos multimodales WEMAC publicado recientemente [11] y diseñado específicamente para representar la emoción del *miedo*. Utilizando la función de similitud del coseno, comprobamos que los *embeddings* de la representación TF-IDF muestran la similitud acústica a las emociones que pretenden elicitar, tal y como se expresan en el conjunto de datos. Nótese que esta categorización emocional es diferente (y podría ser complementaria) de la clásica clasificación y detección acústica de escenas, en la que las escenas suelen estar relacionadas con los lugares físicos que hay que caracterizar, por ejemplo, un aeropuerto, una estación de metro o un parque urbano.

En esta investigación se utilizan 42 de un total de 79 vídeos de la colección UC3M4Safety Audiovisual Stimuli Dataset [11.1] - véase la sección 3.3.1 - para crear una representación estándar de la información acústica y los eventos sonoros que inducen determinadas emociones. Cada estímulo dura entre 30 y 120 segundos, y la colección se compone de fragmentos de películas, escenarios ambientales y recopilaciones de vídeos. En este subconjunto de la primera versión, a cada vídeo se le asigna una etiqueta de emoción mediante *crowdsourcing*, correspondiente a la emoción que elicita en los espectadores. De esos vídeos, 19 están categorizados como *miedo* y los 24 restantes están etiquetados con categorías de otras 9 emociones discretas.

Los datos que utilizamos para el trabajo de esta sección son únicamente el componente de audio, de la colección de estímulos audiovisuales. Contienen diferentes tipos de sonidos -habla, música, eventos sonoros- que, junto con la información visual, inducen en los espectadores las emociones etiquetadas. Para identificar los eventos acústicos que se producen en los datos de audio empleamos un modelo de clasificación de eventos sonoros preentrenado: YAMNet [242].

Tomamos las etiquetas de eventos acústicos predichas por YAMNet como "palabras", y los clips de audio que suscitan las emociones como "documentos", donde nuestro conjunto de estímulos de audio equivale a la "colección de documentos". Obtenemos un vector de puntuaciones TF-IDF por clip -con un valor por etiqueta de evento acústico- que representa la *huella acústica afectiva* de los posibles desencadenantes emocionales de cada vídeo.

TF-IDF (*term frequency – inverse document frequency*) [304] es un método estadístico ampliamente aplicado en la recuperación de información que evalúa la importancia de una "palabra" en un "documento" de una "colección de documentos". Esta importancia viene dada por una puntuación, que resulta de "multiplicar dos métricas: el número de veces que dicha palabra aparece en un documento (TF), y la frecuencia documental inversa de la palabra en un conjunto de documentos (IDF)". La puntuación aumenta proporcionalmente al número de veces que una palabra aparece en un documento, pero disminuye cuando hay un elevado número de documentos que contienen dicha palabra. Cuando la puntuación TF-IDF de una palabra es alta, más relevante es la palabra en ese conjunto concreto de documentos.

Estas puntuaciones TF-IDF podrían introducirse en algoritmos de aprendizaje automático como vectores de palabras, ya que son una forma de representar los datos.

Con el fin de calcular lo similares que son cada par de vectores TF-IDF de cada vídeo de la colección de conjuntos de datos UC3M4Safety, utilizamos una métrica de similitud basada en la distancia coseno (detallada en la Ec. 6.1). La similitud coseno se utiliza ampliamente en la recuperación de información como una forma sencilla y eficaz de proporcionar una medida útil de lo similares que pueden ser dos documentos, independientemente de la longitud de los mismos. Así, como nuestros vídeos tienen longitudes diferentes, nos basamos en esta distancia para medir la similitud entre *los embeddings acústicas afectivos* representados por los vectores TF-IDF.

Resultados sobre la clasificación de escenas acústicas afectivas

Para realizar el análisis de la escena *acústica afectiva*, primero extraemos eventos acústicos de las señales audio que nos permiten caracterizar la escena acústica con YAMNet. Nuestro objetivo aquí es obtener un corpus de etiquetas que representen los eventos sonoros que se producen por ventana de tiempo. Así, posteriormente podremos establecer una métrica que mida lo próximas que están estas representaciones dentro de los estímulos de vídeo del conjunto de datos audiovisuales UC3M4Safety. En esta sección nos ocupamos de la construcción y evaluación del espacio vectorial de eventos acústicos y de los vectores que representan las distintas emociones. Para ello, se ha aplicado el siguiente *pipeline* de procesado, disponible públicamente en GitHub[34]

.

[3436] https://github.com/erituert/acoustic_information_retrievalLa emoción elegida fue la que la mayoría de los anotadores eligieron

Cada uno de los 42 vídeos de la colección elicita una emoción, validada por más de 50 usuarios cada uno[36]. Tanto la modalidad acústica como la visual son las que inducen estas emociones, por lo que primero extraemos el audio únicamente con la herramienta de línea de comandos *ffmpeg*. Además del habla, estos audios también contienen información sobre la escena acústica que induce dichas emociones. Es la escena acústica y el contexto lo que nos gustaría seguir analizando. En la fase de preprocesamiento, hemos utilizado el subsistema de información de audio de Bindi 2.0, ya descrito en la sección 6.1.1. A continuación, utilizamos YAMNet para detectar y clasificar los eventos acústicos presentes en las señales de audio de todos los estímulos de vídeo.

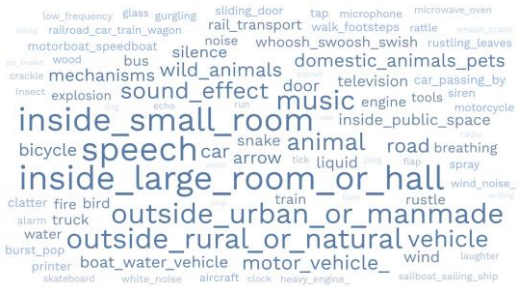


Figura 6. 3: Nube de palabras de las etiquetas acústicas emitidas por YAMNet para los estímulos audiovisuales anotados como "Miedo" [5]. Reproducido con permiso del propietario del copyright, ISCA.

Como YAMNet es un clasificador general de eventos acústicos, puede producir etiquetas de clase muy detalladas que no aporten información útil a nuestra caracterización acústica dados los estímulos audiovisuales utilizados, sino que sólo haga más complejas la tarea y las descripciones. Así pues, teniendo en cuenta la ontología Audioset, se filtran las etiquetas de las clases infantiles *Música* y *Animal*es, excepto las clases de *Música de estado de ánimo* y *Animales Salvajes*, donde se guardan todas las subclases.

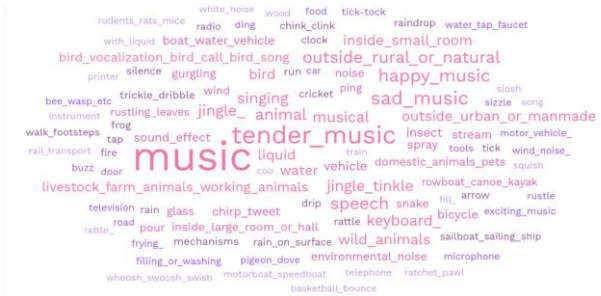


Figura 6. 4: Nube de palabras de las etiquetas acústicas emitidas por YAMNet para los estímulos audiovisuales anotados como "Ternura" [5]. Reproducido con permiso del propietario del copyright, ISCA.

Del total de 521 clases que clasifica YAMNet, las filtradas dan como resultado 351. Las figuras 6.3 y 6.4 representan la nube de palabras de las etiquetas acústicas emitidas por YAMNet para los estímulos audiovisuales anotados como '*miedo*' y *'ternura'*, respectivamente. La Fig. 6.5 representa como ejemplo un vídeo elegido para provocar *miedo* analizado mediante YAMNet. En la señal de audio, al principio se oye hablar a una mujer, después, en el segundo 18, aparece un

fuerte ruido similar a un chirrido seguido del sonido de un motor. Los latidos del corazón están presentes en la última parte del audio.

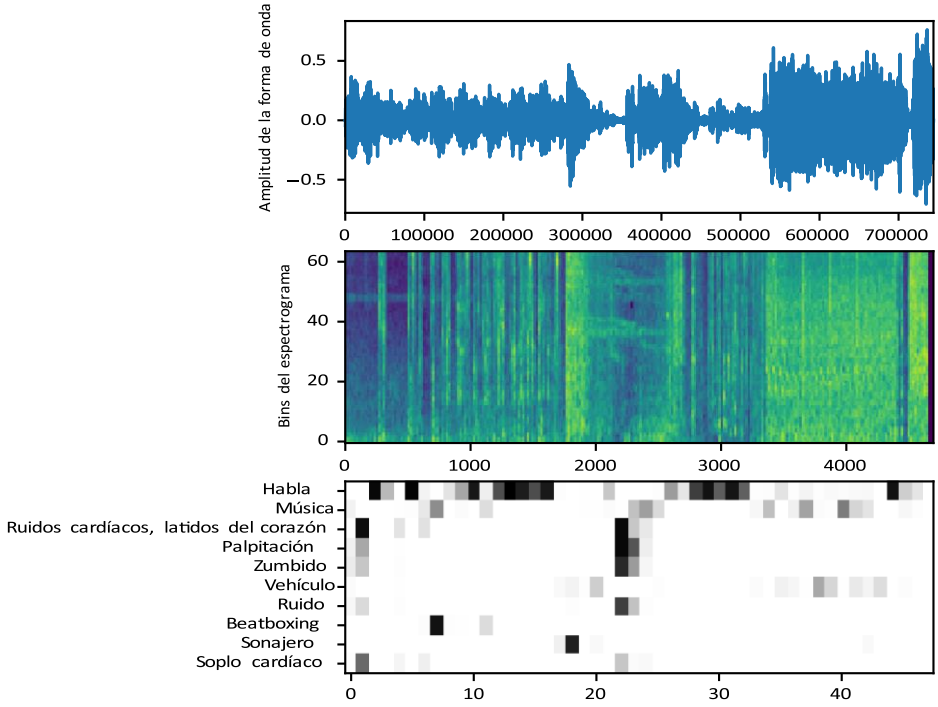


Figura 6. 5: YAMNet procesando una muestra del conjunto de datos de estímulos audiovisuales UC3M4Safety. Representación temporal (arriba), espectrograma con bandas que abarcan de 125 a 7500 Hz (centro) y principales eventos acústicos encontrados (abajo) [8]. Reproducido con permiso del propietario de copyrighy, ISCA.

Para que la comparación sea correcta, dado que las probabilidades de salida de YAMNet de detección de eventos acústicos pueden ser extremadamente bajas, todas las probabilidades se escalan logarítmicamente y luego se binarizan. El objetivo de la binarización es mantener sólo los eventos con una puntuación de salida lo suficientemente alta como para considerar que se han producido y no son una interpretación errónea de la red.

Por lo tanto, el umbral se fija en el valor medio global entre todas las probabilidades de salida de todos los audios, y sólo se tienen en cuenta los eventos sonoros cuya puntuación sea superior al umbral. De esta manera obtenemos un vector de identificadores de eventos acústicos por cada archivo de audio.

El siguiente paso en el proceso consiste en obtener un corpus en forma de texto de los eventos que se producen en el conjunto de datos. Así, cada archivo de audio se trata como un documento de texto compuesto por *mids*, que son las palabras de cada evento sonoro referenciadas a través del código de identificación interno proporcionado en la base de datos Audioset (*mid*).

Por último, utilizamos el algoritmo TF-IDF[35] de la biblioteca *sklearn* Python para obtener la matriz TF-IDF para cada uno de los 42 estímulos audiovisuales del conjunto de datos, lo que da como resultado una matriz de probabilidades de dimensiones (42, 351).

Con el objetivo de analizar la distancia entre los vectores TF-IDF o *embeddings acústicos afectivos* que representan cada una de las instancias del conjunto de datos para comprender los patrones subyacentes que relacionan las emociones, utilizamos una métrica de similitud basada en la distancia coseno:

$$similarity(\vec{x}, \vec{y}) = 1 - cos(\theta) = 1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \tag{6.1}$$

donde θ es el ángulo entre los dos vectores.

En la Fig. 6.7 representamos con un mapa de calor los resultados de calcular la Ec. 6.1 para cada estímulo audiovisual con su emoción etiquetada con respecto al resto de estímulos audiovisuales, con un total de 37, tras eliminar los valores atípicos (que seguían presentes en la Fig. 6.6). Los colores más claros en el mapa de calor representan una mayor similitud, y los colores más oscuros muestran una menor similitud, entre los *embeddings acústicos afectivos*.

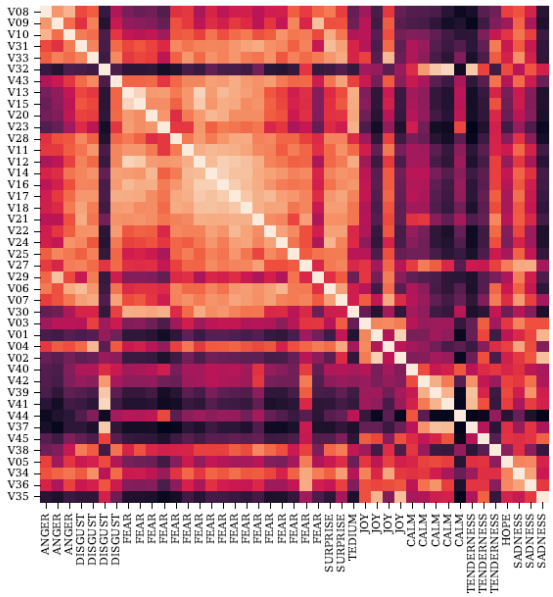


Figura 6. 6: Mapa de calor original de la similitud de las distancias de coseno entre los embeddings acústicos afectivos, ordenados por emociones [5]. Reproducido con permiso del propietario del copyright, ISCA.

[35] https://scikit-learn.org/stable/modules/generated/sklearn.feature_\extraction.text.TfidfVectorizer.html

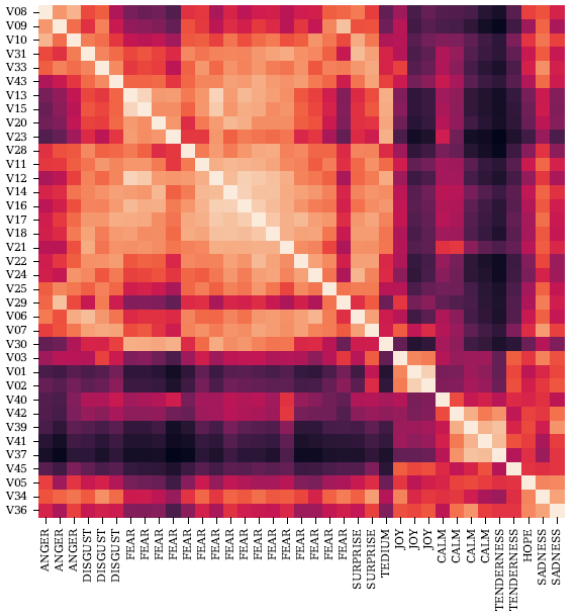


Figura 6. 7: Mapa de calor de los embeddings acústicos afectivos ordenadas por emociones tras eliminar los valores atípicos [5]. Reproducido con permiso del propietario del copyright, ISCA.

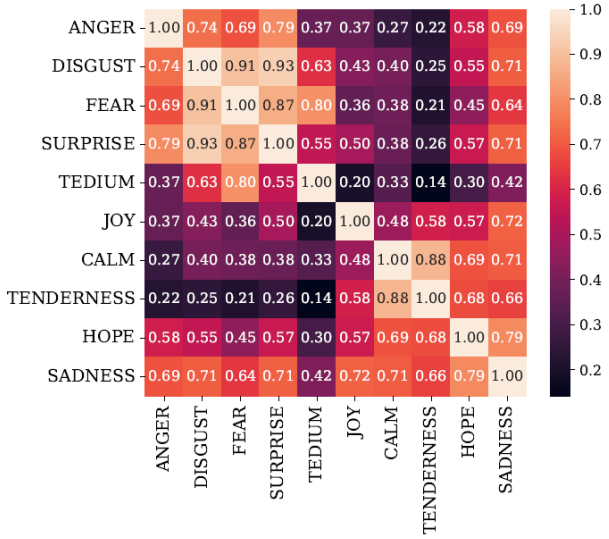


Figura 6. 8: Mapa de calor de las similitudes de las distancias coseno entre los embeddings de emoción [5]. Reproducido con permiso del propietario de los derechos de autor, ISCA.

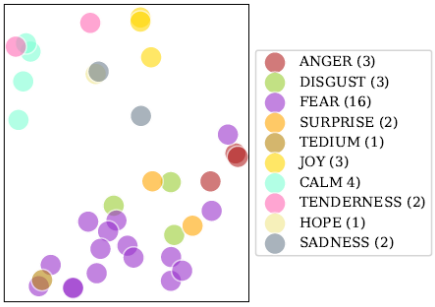


Figura 6. 9: Representación t-sne tf-idf de los embeddings de estímulos audiovisuales [5]. Reproducido con permiso de la página de copyright propietario, ISCA.

Como cada vídeo pretende desencadenar una única emoción, de esta forma podemos entender cómo se relaciona cada vídeo con la forma en que el resto de ellos representan cada una de sus emociones.

La detección y eliminación de valores atípicos se realizó tras comparar cada *embedding* acústico afectivo con el resto de la misma categoría de emoción. Por ejemplo, *V32* se identificó como un valor atípico teniendo en cuenta que su *embedding* presentaba una gran disimilitud con respecto al resto de *embedding*s etiquetadas con la emoción *asco*. Un análisis más detallado revela que el contexto acústico no coincide con la información visual, ya que V32 -que es una recopilación de vídeos- contiene sobre todo música clásica, similar a los vídeos etiquetados en la categoría de *calma*, y por tanto su *embedding* es similar a estos *embeddings* de emociones de *calma*.

Podemos observar que emociones similares presentan clusters de colores parecidos en la Fig. 6.7, lo que significa que los vídeos etiquetados con la misma emoción tienen una caracterización acústica similar. En la Fig. 6.7, se pueden observar a grandes rasgos cuatro grupos: una gran agrupación que incluye *la ira*, el *asco*, el *miedo*, *el tedio* y la *sorpresa*, otra agrupación para la *alegría*, y otra agrupación para la *calma* y la *ternura*, y la última que incluye la *esperanza* y la *tristeza*. Estas cuatro agrupaciones son hasta cierto punto coherentes con la similitud en el espacio PAD en los ejes valence y *arousal* [107] de estas emociones.

Después, realizamos la media de la matriz TF-IDF para cada estímulo audiovisual etiquetado con la misma categoría de emoción. De ese modo podemos comprender cómo influye cada etiqueta acústica en la clasificación de cada emoción. En la Fig. 6.8 presentamos el mapa de calor resultante, desde el punto de vista acústico, de los embeddings de emociones. Podemos observar cómo los resultados son prometedores, ya que las emociones similares presentan una mayor similitud entre sí (por ejemplo, la *calma* y *la ternura*), que las emociones que los humanos categorizan como más diferentes (por ejemplo, *el tedio* y *la alegría*). En particular, la categoría de *miedo* se sitúa cerca de las etiquetas de *asco* y *sorpresa*, lo que dificulta la discriminación entre ellas si sólo tenemos en cuenta el contexto acústico.

Discusión

En la Fig. 6.9 se dibujan los *embeddings* acústicos afectivos utilizando el algoritmo t-sne. Podemos observar que las distancias y la agrupación entre ellas son en cierto modo similares a la agrupación que se produce en la Fig. 6.7.

La relación entre *el miedo* y *la ira* es peculiar, ya que al contrario de lo que cabría esperar presentan una gran similitud. Esto podría explicarse teniendo en cuenta el sesgo de género [126],

que afirma que, en determinadas situaciones, las personas pueden sentir emociones diferentes ante los mismos estímulos en función de su sexo. Esto merece una investigación más profunda.

Dos factores pueden estar influyendo en la solidez de este análisis, en primer lugar el acuerdo entre los anotadores que etiquetaron cada vídeo y su género, y en segundo lugar, la cantidad de vídeos por cada categoría de emoción. Así pues, como trabajo futuro, se podría realizar un análisis más profundo con un estudio más profundo utilizando el conjunto original de vídeos -hasta 79- u otras bases de datos de escenas acústicas con anotaciones emocionales. También puede tenerse en cuenta el acuerdo de los anotadores por género como variable para estudiar su relevancia. Además, utilizando los vectores TF-IDF como características, se podrían alimentar modelos de aprendizaje automático con esos datos y predecir etiquetas emocionales en el aprendizaje supervisado.

Como nota final, trabajamos para intentar responder a la pregunta de si es posible caracterizar una escena acústica o un paisaje sonoro con respecto a las emociones que elicita. Partimos de la premisa de que caracterizar la escena acústica afectiva implica tener en cuenta el contexto acústico. Y respecto a los resultados presentados, conseguimos una caracterización emocional favorable de la escena acústica en material audiovisual, siendo un primer comienzo para el análisis de la escena *acústica afectiva* en entornos del mundo real.

Llegamos a la conclusión de que el uso de la metodología de análisis de escenas acústicas afectivas es un método prometedor, cuyos resultados pueden ser muy interpretables, para caracterizar una escena acústica con respecto a la información emocional. Los *embeddings* robustos que caracterizan acústicamente las emociones pueden utilizarse para medir la carga emocional de -o la emoción que debe elicitar- la información acústica en otras bases de datos.

Otros indicadores además del contexto acústico -como la información procedente de otras modalidades (por ejemplo, las bioseñales de la persona)- son cruciales para caracterizar con precisión una situación y detectar si la vida de la usuaria corre peligro.

6.2. Análisis de Equidad Interseccional en la Clasificación de la Fatiga

En línea con la detección del estrés en la voz, también publicamos un estudio sobre la detección de la fatiga a través de la voz y la respiración en las señales del habla [7], en una colaboración conjunta con la Cátedra de Inteligencia Integrada para la Atención Sanitaria y el Bienestar (Chair for Embedded Intelligence for Healthcare and Wellbeing, EIHW) de la Universidad de Augsburgo.

Este estudio tenía dos objetivos: en primer lugar comprender la fatiga o el estrés que puede provocar el ejercicio de correr para estudiarlo y caracterizarlo, desarrollado por miembros de

EIHW; y en segundo lugar, realizar un análisis de género sobre cómo se observa esta fatiga en cada sexo, la parte que realizamos.

Modelamos la escala de Percepción del Esfuerzo Recibido (RPE) de Borg [306], "una medida subjetiva de la fatiga bien validada", mediante señales de audio que se captaron en entornos reales al aire libre colocando un teléfono inteligente sujeto en los brazos de los corredores y utilizando modelos de aprendizaje automático. Mediante el ajuste fino (preentrenamiento) de una red neuronal convolucional (CNN14 [307]) en espectrogramas log-Mel, los investigadores de EIHW realizaron experimentos en función de la persona y obtuvieron un error medio absoluto (MAE) de 2,35, lo que demuestra que el audio puede adquirirse de forma más sencilla y no invasiva que las señales de otros sensores, a la vez que puede utilizarse eficazmente para modelar la fatiga.

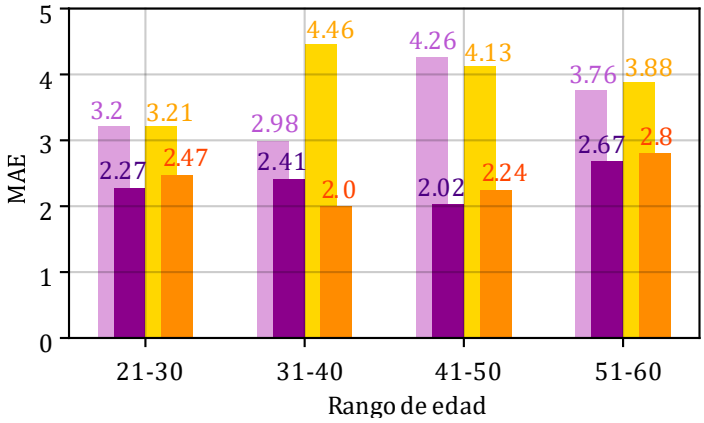


Figura 6. 10: Resultados en términos de MAE divididos por edad y sexo. El color naranja se refiere a las mujeres y el lila a los hombres. Los colores oscuros se refieren a CNN14-preentrenada y los claros a CNN14-random [7]. Reproducido con permiso del propietario del copyright, 2022 IEEE.

Entre los resultados derivados de este estudio, realizamos un análisis de género, en el que el modelo rinde casi por igual para corredores masculinos y femeninos. Y ello a pesar de que los datos utilizados (base de datos KIRun[36]) están sesgados hacia las mujeres (27 corredoras frente a 21 corredores). Esto indica que la cantidad relativa de datos puede no ser el único factor causante de desequilibrios en el rendimiento. Por último, observamos que, si bien el rendimiento preentrenado con CNN14 fue en general mejor para las mujeres en la mayoría de los grupos de edad, lo contrario ocurre en el grupo de edad (21 - 30), en el que las mujeres muestran un MAE de 2,37 frente al 1,86 de los hombres.

Otro patrón interesante es la diferencia de comportamiento entre *CNN14-random* y *CNN14-preentrenada*. Para algunos grupos de edad concretos, los dos modelos muestran un comportamiento muy diferente. Por ejemplo, para el grupo de edad (31 - 40) *CNN14-random* muestra un MAE mucho mayor para las mujeres, pero el rendimiento de *CNN14-preentrenada* es casi el mismo para ambos grupos de sexo. Esto demuestra que los modelos preentrenados no sólo

[36] https://www.uni-augsburg.de/de/fakultaet/fai/informatik/prof/eihw/forschung/proj ekte/vergangene-projekte/

mejoran el rendimiento absoluto, sino que también pueden cambiar el comportamiento del modelo en las distintas subdivisiones del conjunto de datos, lo cual es un efecto secundario no deseado del fenómeno de subespecificación observado en las arquitecturas ML [308].

Para concluir, el análisis de equidad interseccional realizado revela que el rendimiento difiere entre los grupos de edad y las combinaciones de sexos, y que los rendimientos a nivel individual son importantes. Esta conclusión debe reconocerse también para los sistemas de reconocimiento del *miedo*, y sería ideal realizar más trabajos que incluyan datos de estos estados afectivos diferentes pero también similares -fatiga, estrés y *miedo*- para separarlos y analizar sus diferencias, de modo que se pueda entrenar a Bindi para que detecte los estados adecuados y no los clasifique erróneamente.

6.3. Detección Automática de la Condición de Violencia de Género en el Habla

En el trabajo preliminar basado en [309] y publicado en [6] junto con otras componentes del equipo UC3M4Safety, exploramos si la condición de violencia de género podía detectarse a partir del audio sólo mediante un pequeño conjunto de características paralingüísticas del habla en la base de datos WEMAC [11]. El trabajo en [309] aborda el uso de técnicas de selección de rasgos para rasgos extraídos de claves paralingüísticas del habla, y a partir de esa base se realizó la presente clasificación.

Los datos utilizados comprenden 26 víctimas no GBVV y 26 GBVV de los mismos rangos de edad. El proceso de extracción de características está programado en Python[37] e incluye las características presentadas en la sección 3.3.2. Se aplicaron pruebas estadísticas como métodos de selección de rasgos entre ambos grupos (GBVV y no GBVV) para comprobar si había algún rasgo del habla que presentara diferencias significativas entre los grupos y permitiera así distinguirlos (más detalles en [309]). Los análisis estadísticos realizados condujeron al uso de diferentes conjuntos de características para su posterior clasificación (véase [6]).

A continuación, se creó una red neuronal poco profunda, un perceptrón multicapa (MLP) -programado en Python con la biblioteca *sci-kit learn*- para validar los resultados estadísticos. Se aplicaron dos enfoques: una estrategia dependiente de la persona y otra independiente. Los resultados muestran que el modelo MLP es capaz de distinguir entre GBVV y no GBVV con un enfoque dependiente de la persona mejor que con el enfoque independiente de la persona. Al eliminar la dependencia, las puntuaciones disminuyen significativamente, lo que creemos que podría explicarse por la existencia de valores atípicos y la pequeña cantidad de datos [6]. No obstante, debemos tener en cuenta que se trata de un trabajo preliminar que debe continuar y que tiene una limitación principal, que es el hecho de que la muestra contiene 52 usuarias en total.

[37] Disponible en: https://github.com/BINDI-UC3M/wemac_dataset_signal_processing/tree/master/speech_processing

Parte del trabajo futuro previsto consiste en volver a realizar este análisis con una muestra mayor una vez que se disponga de ella.

6.4. Cambio Climático y Violencia de Género

En línea con los diecisiete Objetivos de Desarrollo Sostenible (ODS) previstos y adoptados por todos los Estados miembros de las Naciones Unidas en la Agenda 2030, el ODS 13 es "un llamamiento a la acción para combatir el cambio climático por un mundo mejor", y aquí hemos explorado brevemente la literatura que vincula la violencia de género y el cambio climático.

El estudio más exhaustivo realizado sobre el tema hasta la fecha lo llevó a cabo la Unión Internacional para la Conservación de la Naturaleza (UICN), con la participación de más de 1.000 fuentes de investigación a lo largo de dos años en cambio climático y violencia de género [310]. El estudio sugiere que la violencia de género está aumentando debido al cambio climático, porque el incremento de la degradación medioambiental y la presión sobre los ecosistemas crea escasez de recursos, lo que a su vez genera estrés para las personas. Por tanto, cuando aumentan las presiones medioambientales, aumenta la violencia de género.

Este estudio [311] comparte un amplio plan del Índice de Violencia de Género (GBVI) para identificar la gravedad de los malos tratos en relación con la contaminación atmosférica y la cobertura vegetal, intentando encontrar un vínculo entre la contaminación atmosférica y las copas de los árboles con los niveles de agresión. Parece existir una correlación entre el cambio climático y la violencia de género, y puede que ayudar en uno reduzca los casos del otro.

En el marco de esta tesis, se han utilizado tecnologías de audio para la detección de situaciones de riesgo para las mujeres en el contexto de la violencia de género, y es esta misma tecnología -la audición por ordenador- la que también puede utilizarse para abordar el problema del cambio climático. En este trabajo [12], que es una colaboración conjunta con la Cátedra de EIHW de la Universidad de Augsburgo, ofrecemos una visión general de las áreas en las que la inteligencia auditiva -una tecnología potente pero hasta ahora poco considerada en este contexto- puede contribuir a superar los retos relacionados con el clima. Categorizamos las posibles aplicaciones de la audición por ordenador según los cinco elementos: *agua*, *aire*, *fuego*, *tierra* y *éter*, propuestos por los antiguos griegos en su teoría de los cinco elementos. Esta categorización sirve de marco para describir la audición por ordenador en relación con diferentes aspectos ecológicos. La *tierra* y el *agua se ocupan de* la detección precoz de los cambios medioambientales y, por tanto, de la protección de los seres humanos y los animales, así como de la vigilancia de los organismos terrestres y acuáticos. El *aire* se refiere al audio aéreo, que puede utilizarse para vigilar y obtener información sobre las poblaciones de aves e insectos. Además, las medidas acústicas pueden proporcionar información relevante para la vigilancia y previsión de fenómenos meteorológicos extremos. Por último, el elemento *fuego se ocupa* de la detección y clasificación

automáticas, basadas en el audio, de los incendios forestales, así como de la evaluación de los daños estructurales causados por el fuego. Este trabajo posiciona la audición por ordenador en relación con enfoques alternativos, discutiendo los puntos fuertes y las limitaciones metodológicas, así como los aspectos éticos. Concluimos con una llamada urgente a la acción a la comunidad en general para luchar colectivamente contra el cambio climático.

6.5. Conclusiones

En este capítulo hemos indicado algunas líneas de investigación que surgieron durante la realización de esta tesis. Se trata de trabajos preliminares de gran interés en los que hemos colaborado junto con otros miembros del equipo UC3M4Safety y la Cátedra EIHW. De hecho, requieren un trabajo más profundo en el futuro, ya que sus resultados son alentadores.

Detallamos los trabajos realizados en el campo del análisis de escenas y eventos acústicos y la importancia del análisis de eventos sonoros para la detección de situaciones de riesgo. De ahí surge el término de *Análisis Afectivo de Escenas Acústicas* y con él hacemos un llamamiento a futuras investigaciones bajo esta perspectiva y denominación. En nuestro estudio [5] presentamos resultados favorables, con *embeddings* acústicos robustos e interpretables que caracterizan las emociones en nuestro conjunto de datos de estímulos audiovisuales UC3M4Safety.

Además, realizamos una breve contribución al estudio de la fatiga, proporcionando un análisis de género de la expresión de la fatiga, y en futuras investigaciones sería interesante caracterizarla para ver las diferencias entre el estrés, el *miedo* y la fatiga en las variables fisiológicas y sus efectos en la voz. El trabajo realizado sobre la detección de la condición de GBVV a partir del habla indica de forma prometedora que es posible distinguir entre GBVV y no GBVV utilizando sus características paralinguisticas, abriendo una nueva linea de investigacion en computación afectiva y para la detección de víctimas de violencia de género con el uso de WEMAC. Por último, y siguiendo otro objetivo alineado con el *bien social*, relacionamos el cambio climático y la violencia de género.

Para terminar, como ya se ha mencionado, estas líneas de investigación requieren más atención y trabajos futuros para una comprensión más holística en el contexto de la detección y prevención de la violencia de género mediante la tecnología de audio.

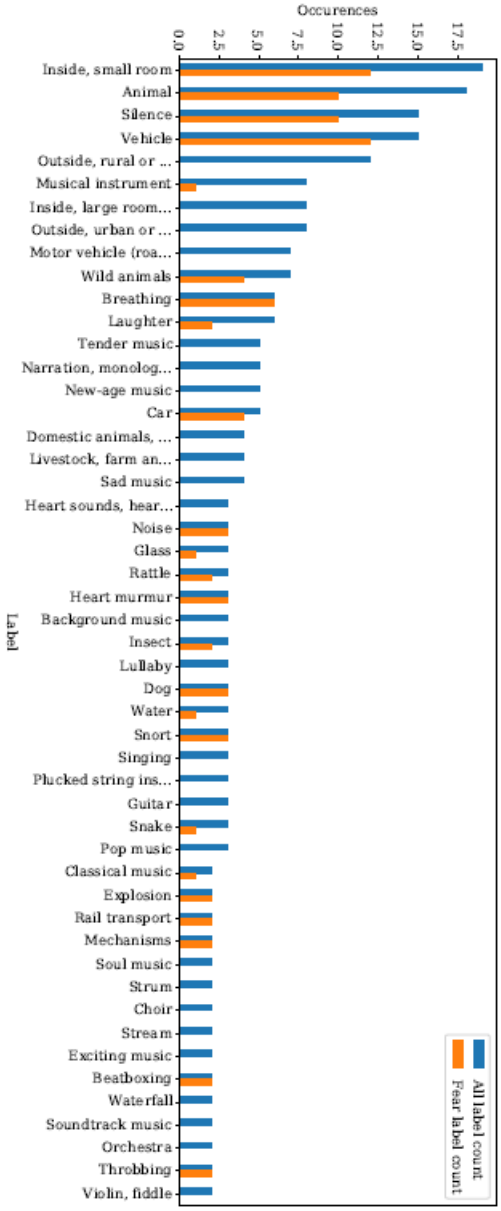


Figura 6. 11: Número absoluto de ocurrencias en las etiquetas acústicas de YAMNet en el miedo frente a todos los estímulos audiovisuales en WEMAC [1]. Reproducido con permiso del propietario del copyright, © 2022 IEEE.

Capítulo 7: Conclusiones y Trabajo Futuro

Este capítulo recoge las conclusiones extraídas de los trabajos de investigación realizados en esta tesis. Nuestro principal objetivo en esta tesis era comprender las reacciones de las mujeres ante situaciones de riesgo para poder detectarlas mediante mecanismos de detección automática utilizando la modalidad auditiva y algoritmos de aprendizaje automático. En este capitulo damos una visión general de las conclusiones obtenidas del trabajo realizado sobre el análisis de la violencia de género en el ámbito de la computación afectiva, el razonamiento de las bases de datos utilizadas y el trabajo en la tarea de reconocimiento de hablantes y emociones junto con la evolución de los sistemas de inferencia en Bindi. Esta tesis es multimodal y multidisciplinar porque ha habido un gran grado de colaboración y las contribuciones están profundamente entrelazadas con las de otros miembros del equipo de investigación. Detallamos cada una de las contribuciones en la parte introductoria de cada capítulo y sección. Una sólida perspectiva de género que guíe la tecnología es un trabajo pionero muy necesario, por lo que podemos considerar que esta investigación, junto con la de los demás miembros del equipo UC3M4Safety, sienta las bases de esta nueva perspectiva en la que pretendemos seguir trabajando en el futuro.

7.1. Conclusiones

La violencia de género la sufren 1 de cada 3 mujeres en el mundo en algún momento de su vida, ya sea física o sexualmente, según la Organización Mundial de la Salud (OMS) [21]. En concreto en España, más de 1.100 mujeres han sido asesinadas desde 2003 hasta 2022, víctimas de la violencia de género [22]. Además, la violencia de género se normaliza y reproduce debido a las desigualdades estructurales y sociales, por lo que es necesario abordar urgentemente este problema socialmente invisible para proteger a las mujeres, que constituyen más del 50% de la población mundial.

Desde la creación del equipo UC3M4Safety, hemos estado trabajando para desarrollar una solución tecnológica innovadora que pudiera ir de la mano de la inteligencia artificial para frenar la violencia contra las mujeres. Y así es como nació Bindi. Bindi consiste en un sistema formado por un *smartwatch* – pulsera inteligente - un colgante, una aplicación para *smartphone* y un servidor en la nube. Constituye un sistema de dispositivos cuyo objetivo es detectar automáticamente cuándo una situación puede poner en peligro la vida de una mujer, y ofrecerle apoyo y ayuda sobre la marcha. Bindi utiliza como principales fuentes de información señales fisiológicas y auditivas, además de otras variables situacionales como la hora del día y la ubicación GPS, para realizar dicho análisis.

La deteccion y clasificacion de emociones, dentro del campo de la computación afectiva, es una gran fuente de informacion ya que podria informar sobre el estado afectivo de una persona. Que una persona esté estresada, nerviosa, asustada, tenga *miedo* o genere la reacción de lucha-huida-congelación (*fight-flight-freeze*), podría ser un indicador de que se encuentra en una situación peligrosa. Por ello, el uso de la computación afectiva y la deteccion de emociones, podrian ser claves en la detección de la violencia de género. Las emociones pueden rastrearse a través de variables fisiológicas (véase [53]) y del habla, y en esta tesis nos centramos en estas últimas.

No sólo las emociones, sino también el contexto y la situación en la que se encuentra la persona pueden darnos información especialmente relevante para la confirmación de que una situación es de riesgo, como las ya mencionadas coordenadas GPS, la hora del día y el contexto acústico (coches que pasan, silencio o una pelea). Estos otros aspectos están siendo estudiados actualmente por otros miembros del equipo UC3M4Safety con los que colaboramos.

Pero la detección de las emociones y, en concreto del *miedo,* es muy difícil. Ya hemos hablado de las dificultades de las etiquetas emocionales debido a su naturaleza subjetiva y a la diferencia de percepción de los anotadores que las etiquetan en la sección 2.5. En cuanto a la falta de datos reales, hemos aplicado técnicas de aumento de datos generando habla sintética estresada -que consideramos una emoción real similar al *miedo*- además de contaminar aditivamente las señales de audio con ruido ambiental real presente en entornos reales, intentando imitar las condiciones en las que Bindi trabajaría.

Sin embargo, surgen varios problemas cuando el objetivo de un sistema es trabajar con datos reales, como se espera que haga Bindi. En primer lugar, la dificultad de encontrar datos de emociones reales (no actuadas) y, en segundo lugar, la escasa confianza en las arquitecturas desarrolladas si los datos utilizados para entrenarlas son sintéticos. Esta situación lleva a la necesidad de generar bases de datos con emociones elicitadas reales, lo que supone un gran reto y requiere mucho tiempo. Así es como surgieron los conjuntos de datos de estímulos audiovisuales UC3M [11.1] [11.2], WEMAC [11] y WE-LIVE, que se detallan por completo en el capítulo 3.

En particular, trabajar con la elicitación de emociones negativas fuertes, como las elicitsdas en WEMAC para la detección del *miedo* en las mujeres en un entorno de laboratorio, puede dar lugar a problemas éticos. Por ello, hay que dedicar muchos recursos a salvaguardar el bienestar de las voluntarias que participan en la recopilación de las bases de datos. Este problema concreto se agrava cuando el grupo objetivo de voluntarias está formado por mujeres que han sufrido violencia de género. Esto se debe a que los fallos del sistema tienen consecuencias críticas para ellas. Aunque la inversión de recursos para proporcionar seguridad y comodidad durante la grabación de nuestras bases de datos es considerable, nos comprometimos totalmente con el bienestar de las voluntarias, proporcionándoles asistencia psicológica constante, ya que la probabilidad de desencadenar un brote de su trastorno de estrés postraumático es alta. Cabe

señalar que el desarrollo de técnicas de adaptación al hablante es fundamental para nuestro caso de uso de la violencia de género. No obstante, consideramos una gran contribución la generación de estas bases de datos, en las que manejamos con sumo cuidado las cuestiones éticas y las limitaciones, con el propósito de servir para allanar el camino de la investigación sobre tecnología para combatir la violencia de género.

Pero antes de detectar el estado afectivo en el que se encuentra una persona, es necesario confirmar que el habla dentro de la señal de audio grabada por Bindi pertenece a esa usuaria, y esa es la razón de nuestro trabajo en el campo del reconocimiento o identificación del hablante en el capítulo 4, detectar la voz de la usuaria y por tanto su identidad de entre toda la información acústica presente en la señal de audio. En este sentido, en nuestra investigación tuvimos especial cuidado con las limitaciones de hardware que tenían nuestros dispositivos Bindi (véase la sección 1.2.2). Abordamos la tarea de SR teniendo en cuenta dos fuentes de variabilidad que podría incluir un audio grabado en un entorno real: las condiciones de estrés y el ruido ambiental. Se demostró que el estrés afectaba negativamente al rendimiento de la SR cuando sólo se utilizaba en la fase de test, por lo que aumentamos la base de datos con habla sintéticamente estresada para el entrenamiento de los modelos ML, lo que demostró mejorar el rendimiento. Por lo tanto, en ausencia de habla estresada emocional real podemos aumentar los datos para conseguir datos que se parezcan al estrés real y puedan ayudar a mantener una tasa de reconocimiento aceptable en los sistemas SR. Asimismo, la contaminación de las señales de audio con ruido ambiental empeora las tasas de SR, incluso en sistemas con limitaciones computacionales como Bindi. Así que nos esforzamos por desarrollar modelos robustos a ese ruido que sí se ajustaran a nuestras condiciones.

En el capítulo 5 detallamos el desarrollo del sistema Bindi. Evaluamos la detección y clasificación de las emociones *relacionadas con el miedo* desde una perspectiva multimodal y multidisciplinar, desde Bindi 1.0 hasta Bindi 2.0. Validamos el uso de los *pipeline* de datos monomodales y las arquitecturas de fusión de datos que combinan señales fisiológicas y de audio para detectar *el miedo* a partir de enunciados del habla mediante WEMAC y obtuvimos un resultado prometedor de una precisión global de clasificación del *miedo del* 63,61% para un enfoque dependiente de la persona adaptado al hablante. También describimos en profundidad los componentes y el funcionamiento del sistema -pulsera, colgante, aplicación y servidor-, y describimos el trabajo realizado en la tarea de detección de estrés a partir del habla. La experimentación llevada a cabo en el capítulo 5 sirve como aproximación multimodal inicial para trabajar con el *miedo* elicitado real en las mujeres y su procesamiento adecuado. Bindi es un sistema muy complejo que requiere un minucioso equilibrio de muchos aspectos, como el consumo de batería, la potencia de cálculo, el uso de recursos y el rendimiento del algoritmo. Todo este trabajo en el reconocimiento de

emociones de *miedo* pretende allanar el camino y dar forma a la próxima versión de Bindi: Bindi 3.0.

En el capítulo 6 damos voz a aquellas líneas de investigación que han surgido a lo largo del camino al tiempo que investigamos adicionalmente la detección del *miedo* a través del habla con una perspectiva de violencia de género. Detallamos los trabajos realizados en el campo de los eventos acústicos afectivos y el análisis de escenas y su importancia para la detección de situaciones de riesgo a través del análisis del contexto acústico. Además, el breve trabajo realizado en el estudio de la fatiga sería interesante utilizarlo para analizar las diferencias entre el estrés, el *miedo* y la fatiga y sus efectos en el habla. El trabajo preliminar realizado sobre la detección de la condición de violencia de género a partir del habla allana de forma prometedora el camino para futuros trabajos con aplicaciones en terapia psicológica. Por último, y siguiendo otro objetivo alineado con el *bien social*, relacionamos el cambio climático y la violencia de género.

En general, esta tesis explora el uso de la tecnología y la inteligencia artificial para prevenir y combatir la violencia de género. Esperamos haber allanado el camino para ello en la modalidad del habla y que nuestra experimentación, hallazgos y conclusiones puedan ayudar en futuras investigaciones. El objetivo último de este trabajo es despertar el interés de la comunidad por desarrollar soluciones al problema tan difícil de la violencia de género.

7.2. Trabajo Futuro

Específicamente en la línea del trabajo realizado en el campo de la identificación del hablante y las emociones utilizando la modalidad del habla, la mayor cantidad de datos realistas disponibles gracias a las bases de datos que registramos nos permite utilizar arquitecturas de aprendizaje profundo más elaboradas, por ejemplo, para utilizarlas en el desembiguamiento de la identidad del hablante y la información emocional en el futuro. Podríamos utilizar un modelo *adversarial* capaz de desenredar estas dos ramas de datos en diferentes vectores de baja dimensión (*embeddings*) para detectar de forma sincrona el hablante y la emoción juntos. De este modo, en cada muestra de habla podríamos inferir la identidad del hablante así como su emoción al mismo tiempo. Esto es en lo que pretendemos trabajar después de esta tesis, utilizando las bases de datos recopiladas a lo largo de la misma: WEMAC y WE-LIVE.

En términos generales, para el desarrollo del Bindi también debemos tener en cuenta que muchas mujeres permanecen en estado de *shock* cuando son agredidas o son víctimas de una agresión, en lugar de producir un habla. Debemos tener en cuenta este hecho para futuros desarrollos del sistema Bindi analizando la aparición de silencios en el audio, junto con las demás variables que ya hemos explorado.

En cuanto al análisis de los eventos acústicos y del contexto acústico dentro de Bindi, sería crucial incluir un módulo para el estudio de la información acústica, con su propio procesamiento y canalización de datos, y su rama de fusión junto con las modalidades fisiológica y del habla para disponer de un detector de situaciones de riesgo de violencia de género más completo y holístico. La detección de gruñidos, respiración agitada o chillidos, también es de especial interés para nuestra aplicación, así como la detección de eventos acústicos como golpes, choques o impactos, que probablemente denotarían que se está produciendo una situación de peligro.

Además, estadísticamente hablando, es más probable que sean los hombres, y no las mujeres, quienes cometan cualquier forma de violencia social [312], por tanto, distinguir voces masculinas bajo emociones dominantes como la ira con Bindi podría denotar una situación de riesgo, ya que la mayoría de las agresiones a mujeres son perpetradas por hombres.

En otro orden de cosas, los sistemas basados en AI están ganando popularidad en la atención sanitaria, pero se ven limitados por "los elevados requisitos de precisión, solidez y explicabilidad" [74]. La AI en la investigación sanitaria, un subcampo de la salud digital, explora muchos enfoques centrados en el ser humano. Hay muchos avances recientes en el ámbito del audio, hasta ahora poco estudiado pero también muy prometedor, con especial atención a los datos del habla presentes en las tecnologías más avanzadas. Este estudio [74] presenta "las últimas investigaciones sobre la detección automática de enfermedades a través de señales de audio" en un estilo de revisión, "desde enfermedades respiratorias agudas y crónicas, pasando por trastornos psiquiátricos, hasta trastornos del desarrollo y neurodegenerativos". El análisis de las emociones, en particular *el miedo*, y la condición de la violencia de género que se tratan en esta tesis, podrían ayudar a la investigación de la AI de audio orientada a la salud, en particular con aplicaciones en la atención a la salud mental y la psicoterapia.

Bibliografía

Las referencias [1-20] se encuentran desde la página *vii* hasta la *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-fre quently-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/i ntroduction.

[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-viole nce-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-con tinua/.

[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: htt ps://www.ohchr.org/en/instruments-mechanisms/instruments/declaration-elimination-violenc e-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.intersectional justice.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-re gulations.

[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.or g/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/10.101 6/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: 10.1145/3025453.302 5532.

[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/visu al-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/S0735-1 097(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.2 019.258.

[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: htt p://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-b ehaviour. Accessed: 2022-06-28.

[84] *Fight Or Flight Response. Psychologytools*. Accessed: 23-10-2022. URL: https://www.psychologytools.com/resource/fight-or-fligh-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://oratorv oice.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.healt hline.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.harva rd.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/understandingthe-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/S0 167876001001623.

[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-fli ght#1.

[96] *Anxiety and its affects on the auditory and vocal apparatus*. https://australianvoiceassociation.com.au/2 017/12/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/scien ce/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.or g/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.122 92.

[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: 10.145 69/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.1467-9280.2007.0 2024.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/B978012801 8514000094.

[116] Wikipedia. *Machine Learning*. Accessed: 23-10-2022. URL: https://en.wikipedia.org/wiki/Machine_lea rning.

[117] *What is Deep Learning? Techtarget.* Accessed: 23-10-2022. URL: https://www.techtarget.com/searchenter priseai/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/S15326985EP3 704_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/1660-4 601/17/22/8534.

[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://ww w.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.2 020.3021015.

[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/B978012801 8514000094.

[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.2019.0 6.001.

[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/10.1007/978-3-03042504-3_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/T AFFC.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/222 7-7390/9/8/807.

[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/art icle/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: 10.10 07/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/a rticle/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: 10.10 16/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 Rajcˇáni. "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 Rajcˇáni, 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/f comp.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: 10.3 3682/006b-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.p one.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/DCIS5133 0.2020.9268662.

[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,

[184] 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/z enodo.6097378. URL: https://doi.org/10.5281/zenodo.6097378.

[185] 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.

[186] 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.

[187] 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.

[188] 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”. In: *Interspeech 2017*. ISCA, Aug. 2017, pp. 3512–3516.

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: htt ps://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.1999.758 342.

[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: 10.1 109/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/S18770 5091731342X.

[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, Zbyneˇk 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/a rticle/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 Og˘uz. "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 Ba˘lan, 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/1424-8220/2 1/5/1587.

[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/a rticle/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.2018.8681 496.

[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.2020.297 3425.

[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/journa l.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: 10.10 73/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.

Colección "Premio de la Delegación del Gobierno contra la Violencia de Género a Tesis Doctorales sobre Violencia contra la Mujer"

TÍTULOS PUBLICADOS

1. Destrucción y reconstrucción de la identidad de mujeres maltratadas: análisis de discursos autobiográficos y de publicidad Institucional.

2. Autonomía personal y afrontamiento en mujeres en situación de maltrato.

3. Factores predictores del impacto psicopatológico en víctimas de agresión sexual.

4. Sexismo, amor romántico y violencia de género en la adolescencia.

5. Menores testigos de violencia entre sus progenitores: repercusiones a nivel psicoemocional.

6. Victimización en la Trata sexual: imaginarios e invisibilización.

7. La práctica judicial en los delitos de malos tratos. Patria potestad, guarda y custodia y régimen de visitas.

8. Menores y violencia de género: nuevos paradigmas.

9. El delito de stalking: análisis jurídico y fenomenológico.

10. La reproducción de la violencia sexual en las sociedades formalmente igualitarias: Un análisis filosófico de la cultura de la violación actual a través de los discursos y el imaginario de la pornografía.

11. Mujeres que se recuperan de la violencia de género en la pareja: Análisis de la relación entre el proceso de liberación psicosocial de las víctimas y su participación en el procedimiento judicial contra su agresor.

12. Arteterapia como vía de abordaje del trauma y la violencia hacia las mujeres: Diseño, aplicación y análisis de metodologías y registros de intervención.

13. Comunicación y violencia contra las mujeres. Análisis de la deontología periodística española (1999-2018) y latinoamericana (2004-2017) específica en violencia contra las mujeres.

14. Clasificación del miedo usando computación afectiva, información fisiológica, y dispositivos portables e inteligentes para ayudar a combatir la violencia de género.

15. La prostitución china en la Comunidad de Madrid. Un análisis desde la perspectiva de género.

16. La prostitución china en la Comunidad de Madrid. Un análisis desde la perspectiva de género.

17. Traducción de sentencias judiciales y perspectiva de género: estudio jurídico-traductológico de la traducción de sentencias del Tribunal Europeo de Derechos Humanos.