EEG-a-voz: Síntesis de voz a partir de registros de la actividad cerebral

Autor:
Jesús del Castillo Cabrera

Supervisores:
José Andrés González López
Ana Belén Chica Martínez

Un trabajo de investigación presentado como
requisito parcial para optar al título de Magíster
en Ingeniería de Telecomunicación

Departamento de Teoría de la Señal, Telemática y Comunicaciones
E.T.S. de Ingenierías Informática y de Telecomunicación

Trabajo Fin de Máster

Universidad de Granada
Granada, España

—

Septiembre de 2024

Autor:
Jesús del Castillo Cabrera

Supervisores:
José Andrés González López
Ana Belén Chica Martínez

—

Universidad de Granada, Granada, España

EEG-a-voz: Síntesis de voz a partir de registros de la actividad cerebral

Palabras clave: Inteligencia artificial, aprendizaje profundo, electroencefalografía, vocoder.

Resumen

La capacidad de comunicación verbal es una de las habilidades más esenciales para el ser humano. Sin embargo, esta capacidad puede verse afectada por lesiones traumáticas o enfermedades neurodegenerativas como la esclerosis lateral amiotrófica (ELA), las lesiones cerebrales o de médula espinal y la parálisis cerebral, entre otras. De estas enfermedades quizás la más devastadora es la ELA. Esta enfermedad neurodegenerativa progresiva provoca una pérdida gradual de la función muscular, incluyendo la capacidad de hablar y comunicarse verbalmente. A medida que la enfermedad progresa, las personas afectadas pierden gradualmente la capacidad de comunicarse de forma verbal y deben recurrir a dispositivos que dependen de señales no verbales. En casos extremos, estas enfermedades pueden dejar al individuo en un estado conocido como síndrome de enclaustramiento, en el que las capacidades cognitivas se mantienen intactas, pero el paciente no puede moverse ni comunicarse verbalmente debido a la parálisis de casi todos los músculos que soportan los movimientos voluntarios.

Ante la falta de soluciones clínicas efectivas para mejorar la sintomatología asociada a estos trastornos, las interfaces cerebro-computador (ICC) se presentan como una alternativa prometedora para proporcionar un nuevo canal de comunicación y control no muscular a las personas paralizadas. En este contexto, el presente estudio propone un enfoque innovador basado en la codificación automática variacional (VAE, por sus siglas en inglés) para la síntesis de voz a partir de registros de actividad cerebral obtenidos mediante métodos de electroencefalografía (EEG) invasivos en pacientes epilépticos implantados con electrodos profundos durante tareas de producción del lenguaje. Este enfoque tiene como objetivo ofrecer una solución de comunicación más natural y efectiva para personas con discapacidades del habla.

EEG-to-voice: Speech synthesis from brain activity recordings

Keywords: Artificial intelligence, deep learning, electroencephalography, vocoder.

Abstract

Verbal communication ability is one of the most essential skills for humans. However, this ability can be impaired by traumatic injuries or neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS), brain or spinal cord injuries, and cerebral palsy, among others. Of these conditions, ALS is perhaps the most devastating. This progressive neurodegenerative disease causes gradual loss of muscle function, including the ability to speak and communicate verbally. As the disease progresses, affected individuals gradually lose the ability to communicate verbally and must rely on devices that depend on non-verbal signals. In extreme cases, these diseases can leave the individual in a state known as locked-in syndrome, where cognitive abilities remain intact, but the patient cannot move or communicate verbally due to the paralysis of nearly all the muscles that support voluntary movement.

In the absence of effective clinical solutions to repair these disorders, brain-computer interfaces (BCIs) offer a promising alternative to provide a new channel for non-muscular communication and control for paralyzed individuals. In this context, the present study proposes an innovative approach based on variational autoencoding (VAE) for speech synthesis from brain activity recordings obtained via electroencephalography (EEG) in epileptic patients implanted with deep electrodes during language production tasks. This approach aims to offer a more natural and effective communication solution for individuals with speech disabilities.

Yo, **Jesús del Castillo Cabrera**, alumno de la titulación Ingeniería de Tecnologías de Telecomunicación de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 77147313T, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Máster en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Jesús del Castillo Cabrera

Granada a 6 de Septiembre de 2024.

D. **José Andrés González López**, Profesor del área de Procesado de Señal del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada.

Dña. **Ana Belén Chica Martínez**, Catedrática del área de Neurociencia Cognitiva y del Comportamiento del Departamento de Psicología Experimental de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado ***EEG-a-voz: Síntesis de voz a partir de registros de la actividad cerebral***, ha sido realizado bajo su supervisión por **Jesús del Castillo Cabrera**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 6 de Septiembre de 2024.

Los directores:

José Andrés González López **Ana Belén Chica Martínez**

A mi familia, por su apoyo y cariño incondicional.

Gracias por estar siempre ahí.

Índice

CAPÍTULO

1 Introducción

La esclerosis lateral amiotrófica (ELA), las lesiones cerebrales o de médula espinal y la parálisis cerebral, entre otras enfermedades, pueden interrumpir los canales neuromusculares a través de los cuales el cerebro se comunica y controla su entorno. De estas enfermedades quizás la más devastadora es la ELA. Esta enfermedad neurodegenerativa progresiva provoca una pérdida gradual de la función muscular, incluyendo la capacidad de hablar y comunicarse verbalmente. Aunque la ELA es relativamente rara (en España se diagnostican unos 900 casos nuevos cada año (FUNDELA, 2014)), la carga económica y social es considerable. A medida que esta enfermedad progresa, los individuos pierden la capacidad para comunicarse verbalmente, requiriendo de sistemas de comunicación alternativos que dependen de señales no verbales (p. ej. el movimiento de los dedos o de los ojos). En el peor de los casos la enfermedad puede dejar al sujeto en un estado denominado como síndrome de enclaustramiento, caracterizado por una parálisis motora completa pero con un procesamiento cognitivo y emocional intacto. Las personas en este estado son totalmente conscientes pero no pueden moverse o comunicarse debido a la parálisis de casi todos los músculos voluntarios del cuerpo. En el mejor de los escenarios, la tecnología moderna de soporte vital permite que estas personas sobrevivan durante años. Sin embargo, las consecuencias personales, sociales y económicas de sus discapacidades suelen ser graves y prolongadas.

Ante la ausencia de métodos para reparar el daño causado por estas enfermedades, las interfaces cerebro-computador (ICC) han surgido como una alternativa prometedora para proporcionar un nuevo canal de comunicación y control no muscular a los individuos paralizados. Una ICC es un dispositivo neuroprotésico que utiliza señales neurofisiológicas capturadas del cerebro para controlar dispositivos externos. La idea principal es capturar las manifestaciones eléctricas, magnéticas o de otro tipo de la actividad cerebral de los deseos de comunicación del usuario y traducirlas en órdenes que son interpretadas y ejecutadas por una computadora u otro dispositivo. Hoy en día los sistemas ICC se consideran una herramienta con un enorme potencial para establecer alternativas de comunicación, restablecer funciones y ofrecer procesos de rehabilitación a pacientes con discapacidad neuromotora. Así, las ICC se han utilizado para controlar dispositivos que sustituyen el control mo-

tor (p. ej. mediante el uso de extremidades robóticas) o los órganos sensoriales[1]. Las ICC también se han empleado con éxito para restaurar la comunicación a personas que, como los pacientes de ELA, presenten serias dificultades para comunicarse oralmente. Por ejemplo, estos sistemas se han empleado para controlar un teclado virtual mediante un cursor movido por ondas cerebrales o monitorizando la actividad cerebral para detectar potenciales evocados que se generan cuando la persona mira la letra deseada o cuando la persona parpadea[1]. Estos enfoques, sin embargo, son lentos y poco intuitivos: en la literatura se reportan tasas de comunicación de hasta 12 palabras/minuto[2]. Asimismo, dominar el control de estos dispositivos es una tarea ardua que puede llevar varios meses de práctica.

Como enfoque más eficiente y natural, diversos estudios recientes han investigado la posibilidad de decodificar el habla directamente del electroencefalograma del sujeto. Para ello se emplean algoritmos de reconocimiento automático del habla (RAH) que permiten obtener una representación textual de la actividad cerebral obtenida durante el proceso de producción del habla. Así, por ejemplo, Lotte[3] demostró que características fonéticas tales como el lugar y modo de articulación y el modo de fonación pueden distinguirse a partir de señales cerebrales grabadas usando electrocorticografía (ECoG). Ésta es una técnica invasiva para registrar la actividad eléctrica de la corteza cerebral mediante la colocación de una manta de electrodos directamente sobre la superficie expuesta del cerebro. En otro estudio reciente, Herff y Schultz[4] demostraron que también es posible distinguir sílabas y palabras aisladas a partir de grabaciones intracraneales. El primer estudio que abordó la tarea de decodificar todos los fonemas (las unidades básicas del lenguaje hablado) de una lengua fue Mugler et al. [5], en el que obtuvieron tasas de reconocimiento de hasta un 36 % al clasificar los fonemas del inglés americano dentro de las producciones de palabras y hasta un 63 % para fonemas aislados. Más recientemente, en un estudio realizado por Herff y Schultz[6], se demostró por primera vez que es posible decodificar voz continua a partir de grabaciones del electroencefalograma (ECoG). En este estudio siete participantes fueron implantados con sensores intracraneales y, posteriormente, la voz y la actividad cerebral de los mismos fue registrada simultáneamente mientras estos leían varios textos en voz alta. Las señales cerebrales adquiridas se utilizaron entonces para entrenar algoritmos de RAH para cada sujeto. Estos algoritmos fueron capaces de descifrar correctamente hasta el 75 % de las palabras dichas por los participantes a partir de su electroencefalograma cuando el vocabulario constaba de 10 palabras posibles, mientras que si el vocabulario se ampliaba a 100 palabras la tasa de clasificación disminuyó hasta el 40 %.

Aunque los enfoques anteriores para restaurar la comunicación son muy atractivos, estos presentan también varias limitaciones importantes. En primer lugar, la creación de una neuroprótesis del habla requiere una comprensión sólida del proceso de producción del habla en el cerebro. La producción del habla es un proceso complejo que implica la coordinación de múltiples áreas cerebrales, incluyendo el cortex prefrontal, el cortex temporal, el cortex parietal y el cerebelo. Aunque se han propuesto varios modelos sobre el proceso de producción del habla, aún no se ha comprendido completamente el papel preciso de todas las áreas involucradas. En segundo lugar, los sistemas de RAH usados introducen retrasos significativos que hacen imposible descifrar el habla en tiempo real a partir de la actividad cerebral capturada. En tercer lugar, el uso de un algoritmo de RAH limita el idioma y el vocabulario con el que el usuario puede comunicarse a aquél para el que el sistema está configurado. Esto restringe la flexibilidad del sistema para hablar nuevas palabras

o en un idioma diferente. Por último, algunos aspectos paralingüísticos del habla (p. ej. el estado de ánimo de la persona), que son importantes para la comunicación oral, son difíciles de capturar y modelar usando software de RAH.

Para una completa restauración de la comunicación oral sería deseable, por un lado, que la ICC proporcionase feedback acústico del habla decodificada y, por otro, que el retardo de la señal acústica fuese mínimo. En concreto, estudios en feedback auditivo retardado indican que si la síntesis de voz se realiza en menos de 200 ms, es posible restaurar la retroalimentación auditiva normal sin causar disfluencias al sujeto [7]. Además, debido a la plasticidad neuronal, podría producirse que el cerebro del sujeto asimilara como propia la voz sintética generada por la ICC al recibir feedback auditivo instantáneo de la misma, tal y como han demostrado otros estudios similares para el caso del control de extremidades robóticas [8]. Esto permitiría al paciente producir una mejor voz con la práctica y uso de la ICC. Para llegar a esta situación ideal, no obstante, se hace necesario explorar un enfoque radicalmente distinto para generar voz a partir del electroencefalograma de la persona.

Aprovechando los últimos avances en aprendizaje profundo, estudios recientes han explorado el uso de modelos de regresión basados en redes de memoria a largo y corto plazo[9] y redes generativas adversariales[10] para generar habla sintética directamente desde señales cerebrales, obteniendo resultados prometedores. Sin embargo, estos resultados son todavía preliminares y requieren más investigación para ser considerados robustos y aplicables en situaciones prácticas. La mayoría de los estudios realizados hasta la fecha se han centrado en la calidad de los parámetros sintetizados, dejando aspectos importantes como la inteligibilidad, naturalidad y calidad general del habla generada en segundo plano. Para que estas tecnologías alcancen una verdadera utilidad y sean ampliamente adoptadas, es fundamental abordar estas limitaciones y llevar a cabo evaluaciones rigurosas en diversos escenarios y con diferentes grupos de usuarios.

En este contexto, el presente estudio propone un enfoque innovador basado en la codificación automática variacional para la síntesis de voz a partir de registros de actividad cerebral obtenidos mediante métodos de electroencefalografía (EEG) invasivos en pacientes epilépticos implantados con electrodos profundos durante tareas de producción del lenguaje. El objetivo principal de este trabajo es desarrollar una solución de comunicación más natural y efectiva para personas con discapacidades del habla, aprovechando los avances más recientes en aprendizaje profundo. La propuesta presentada busca superar las limitaciones de los enfoques anteriores, centrándose no solo en la calidad de los parámetros sintetizados, sino también en la inteligibilidad, la naturalidad y la calidad general del habla generada.

1.1 Objetivos

Este trabajo propone el desarrollo de un sistema de síntesis de voz basado en registros de actividad cerebral obtenidos mediante métodos de electroencefalografía (EEG) invasivos en pacientes epilépticos. El objetivo principal es ofrecer una solución de comunicación más natural y efectiva para personas con discapacidades del habla, como aquellas afectadas por enfermedades neurodegenerativas o el síndrome de enclaustramiento. Para lograrlo, se utilizan modelos de aprendizaje profundo

para mapear las señales cerebrales a características acústicas del habla, siguiendo un enfoque de síntesis paramétrica que busca mejorar la calidad, inteligibilidad y naturalidad de la voz generada en comparación con trabajos previos.

Se plantean diversas métricas para evaluar tanto los parámetros sintetizados como la calidad perceptual de la voz sintetizada. Aunque los resultados no alcanzaron la calidad esperada, este trabajo sienta las bases para futuras investigaciones en la síntesis de voz a partir de señales cerebrales, abriendo nuevas posibilidades en el campo de las interfaces cerebro-computadora y la comunicación asistida.

1.2 Organización de la memoria

Para el desarrollo de este trabajo, se ha llevado a cabo una organización de la memoria por capítulos, cada uno dedicado a un aspecto específico de la investigación:

- **Fundamentos**: Este capítulo establece el marco teórico de la investigación, introduciendo los conceptos clave sobre la adquisición de señales de actividad cerebral. Se explican los principios y técnicas empleados para capturar dichas señales, aportando el contexto necesario para el desarrollo de la investigación.
- **Aprendizaje Profundo**: En este capítulo se examinan los principios del aprendizaje profundo y su aplicación en la investigación. Se analizan los modelos neuronales utilizados para la síntesis de voz a partir de registros de la actividad cerebral, detallando los algoritmos empleados y su relevancia para la resolución del problema planteado en el estudio.
- **Metodología**: Este capítulo describe el enfoque metodológico seguido en la investigación, desde la recolección de datos hasta la síntesis y evaluación de la calidad de la voz generada. Se detallan los procedimientos experimentales, las técnicas de procesamiento de señales empleadas, los modelos de aprendizaje profundo utilizados, y los métodos de evaluación implementados para garantizar la validez y fiabilidad de los resultados.
- **Resultados Experimentales**: En este capítulo se presentan y analizan los resultados obtenidos durante la investigación. Se exponen los hallazgos relacionados con la síntesis de voz a partir de registros de la actividad cerebral, evaluando la calidad y precisión de la voz generada. Asimismo, se discuten las implicaciones de estos resultados en relación con los objetivos del estudio y se señalan las limitaciones encontradas en el proceso experimental.
- **Conclusiones**: Este capítulo final ofrece una síntesis comprensiva del trabajo realizado, destacando las principales contribuciones y limitaciones encontradas en el proceso experimental. Se evalúa el cumplimiento de los objetivos propuestos y se sugieren posibles líneas de investigación futura para mejorar la calidad y precisión de los modelos implementados.

CAPÍTULO

2 Fundamentos

A finales del siglo XIX, una serie de descubrimientos fundamentales establecieron las bases de la neurociencia moderna. En 1870, Gustav Fritsch y Eduard Hitzig demostraron la excitabilidad eléctrica de la corteza cerebral y localizaron áreas motoras específicas, mejorando así la comprensión del control cerebral sobre el cuerpo. Paralelamente, Santiago Ramón y Cajal Santiago Ramón y Cajal desarrolló la teoría neuronal, identificando a las neuronas como las unidades fundamentales del sistema nervioso. Los hallazgos de Cajal, que le valieron el Premio Nobel de Medicina en 1906, junto con los de otros pioneros, impulsaron el desarrollo de técnicas especializadas como la electrofisiología y las tecnologías de neuroimagen. Estos avances dieron lugar a disciplinas como la neurociencia cognitiva, la neuropsicología y la neurobiología, proporcionando una comprensión más profunda del funcionamiento cerebral y su relación con el comportamiento humano.

2.1 El cerebro

El cerebro es el núcleo central del Sistema Nervioso Central y su principal función es el procesamiento de la información que recibe a través de los sentidos para preparar y ejecutar respuestas adecuadas a los estímulos del entorno. Este órgano complejo está dividido en dos hemisferios: el derecho y el izquierdo, los cuales están separados por una fisura longitudinal y conectados a través del cuerpo calloso. Generalmente, el hemisferio derecho se encarga de recibir sensaciones y controlar movimientos del lado izquierdo del cuerpo, mientras que el hemisferio izquierdo realiza funciones similares pero con el lado derecho del cuerpo[11].

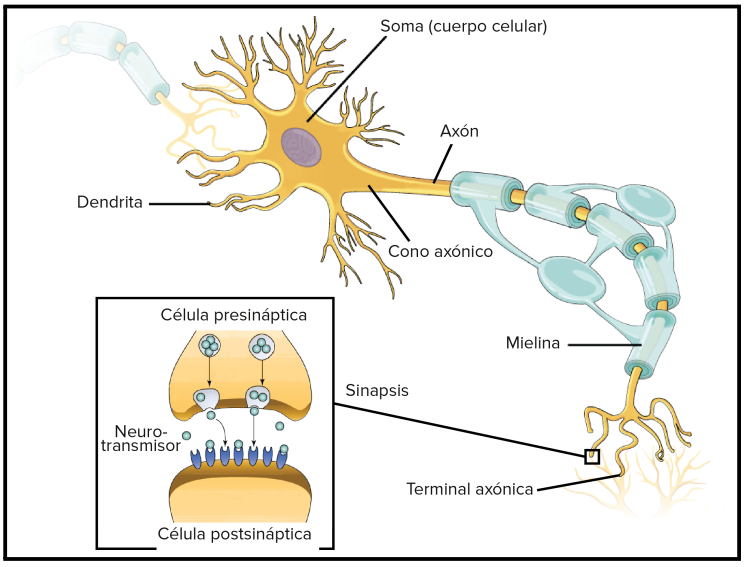
2.1.1 Las neuronas

Dentro del cerebro, las neuronas son las células especializadas responsables de la comunicación neuronal y del procesamiento de la información. Estas células transmiten señales mediante procesos electroquímicos, lo que les permite enviar infor-

mación a grandes distancias dentro del cuerpo de forma rápida y eficiente. Cada neurona posee ramificaciones conocidas como dendritas, que actúan como receptores de señales provenientes de otras neuronas, las cuales son luego transmitidas al soma o cuerpo celular. En el núcleo de la neurona, se procesa la información y, si la respuesta es apropiada, se genera un impulso eléctrico que viaja a lo largo del axón, una prolongación que permite la comunicación con otras neuronas.

La transmisión de la señal de una neurona a otra ocurre en las sinapsis, puntos de conexión entre el axón de una neurona y las dendritas de la siguiente. Estas sinapsis están formadas por un espacio líquido que contiene concentraciones específicas de iones, principalmente sodio y potasio, los cuales permiten la transmisión del impulso eléctrico o, en algunos casos, lo inhiben. De esta manera, las sinapsis funcionan como moduladores de la señal, regulando la intensidad de la transmisión y posibilitando una alta plasticidad en la respuesta del sistema nervioso.

Figura 2.1: Estructura de la neurona.



FUENTE: Khan Academy. (s.f.). Función y estructura de la neurona. Recuperado de es.khanacademy.org.

2.1.2 Divisiones funcionales del cerebro

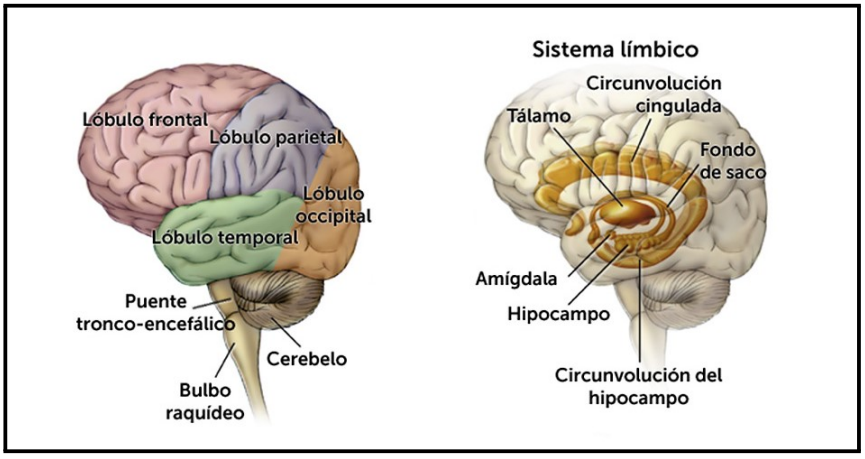
Desde una perspectiva global, las estructuras conformadas por neuronas en el cerebro pueden percibirse como tres grandes estructuras cerebrales: la corteza cerebral, el cerebelo y el sistema límbico (ver Figura 2.2):

- **Corteza Cerebral**: La corteza cerebral, también conocida como neocorteza, es la estructura más externa del cerebro y se encarga de funciones superiores y del control de los actos voluntarios. Esta región del cerebro se divide en cuatro lóbulos: frontal, parietal, temporal y occipital. Cada uno de estos lóbulos tiene funciones específicas. El lóbulo frontal está relacionado con funciones ejecutivas, como la planificación y el control de movimientos voluntarios. Los lóbulos parietal, temporal y occipital se encargan de procesar la información

sensorial. Por ejemplo, el lóbulo occipital es fundamental para la visión, mientras que el lóbulo temporal está implicado en el procesamiento auditivo.

- **Cerebelo**: El cerebelo es una estructura situada en la parte posterior del cerebro y es crucial para la coordinación y el control de los movimientos. Procesa los impulsos eléctricos provenientes de los sistemas esquelético y muscular, incluyendo los sentidos. Sus funciones incluyen la coordinación avanzada, la organización de la ejecución de movimientos, la optimización de la adquisición sensorial de información, y la recalibración de los sistemas motores.
- **Sistema Límbico**: El sistema límbico es una red de estructuras cerebrales que se encargan de procesar las emociones y la memoria a largo plazo. Este sistema incluye estructuras como el hipocampo, la amígdala y el hipotálamo. El hipocampo es esencial para la formación de nuevas memorias y su consolidación a largo plazo, mientras que la amígdala está asociada con las respuestas emocionales, especialmente el miedo y la agresión.

Figura 2.2: Anatomía del cerebro humano.



FUENTE: Recuperado de BrightFocus Foundation.

2.1.3 Ondas cerebrales

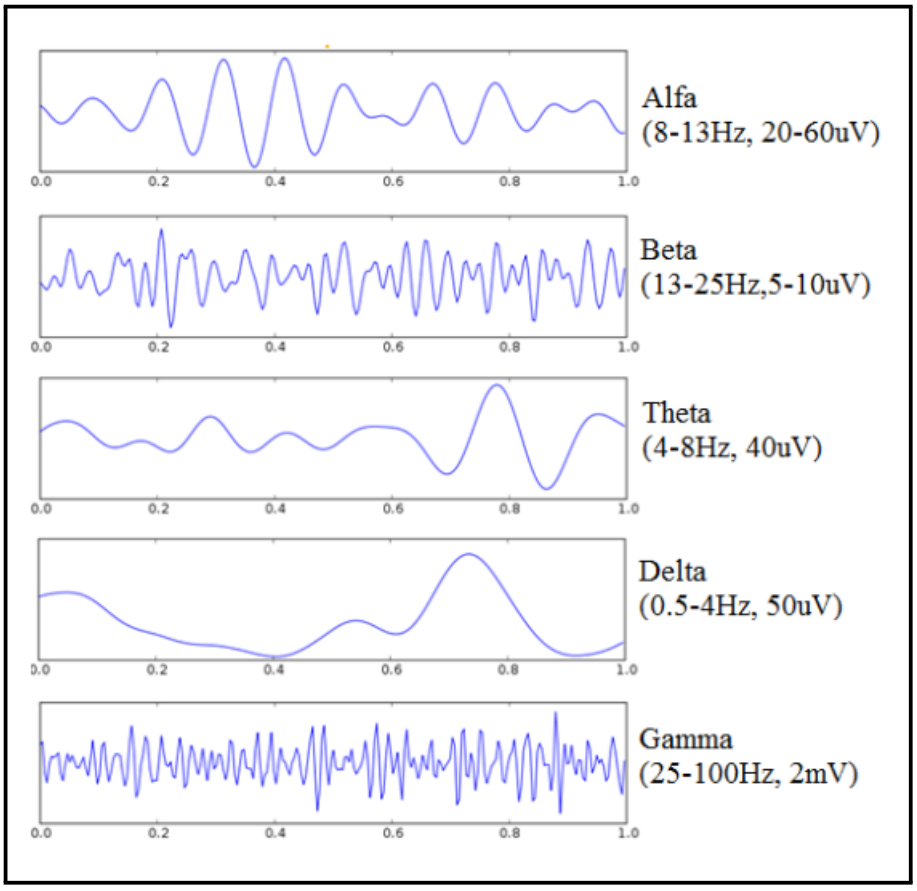
La actividad eléctrica del cerebro se organiza en distintas bandas de frecuencia, cada una de las cuales está relacionada con diferentes estados y funciones cognitivas. Estas bandas de frecuencia se miden en microvoltios y pueden graficarse mediante un logaritmo natural como una línea ascendente [12]. Bandas de frecuencia que actúan en redes específicas de neuronas han sido asociadas a diferentes estados de consciencia y procesos cognitivos [13] y dentro de su comportamiento se anota que varias de estas ondas pueden actuar juntas en las mismas redes y estructuras neuronales [12][13]. Se identifican cinco tipos principales de ondas cerebrales, cada una con funciones específicas que contribuyen a distintos procesos mentales y capacidades cognitivas del ser humano (ver Figura 2.3):

- **Ondas Delta**: Comprende ondas con frecuencias entre 0.5 y 4 Hz. Se observa principalmente durante el sueño y en estados de meditación profunda. Debi-

do a su baja frecuencia, es fácil confundir este tipo de señales con artefactos producidos por el movimiento de los músculos del cuello o la mandíbula.

- **Ondas Theta**: Comprende ondas con frecuencias entre 4 y 8 Hz. Tiende a aparecer de forma aislada con predominio hacia las zonas centrales y temporales. Se asocia con estados de somnolencia y sueño ligero. Es crucial para la consolidación de la memoria y el aprendizaje.
- **Ondas Alfa**: Presenta frecuencias entre 8 y 13 Hz. Se observa cuando la persona está despierta, en reposo y con los ojos cerrados. Se identifica mejor en la parte posterior de la cabeza, aunque en algunos casos se visualiza mejor en las regiones parietales o temporales posteriores.
- **Ondas Beta**: Comprende ritmos con una frecuencia mayor de 13 Hz. La forma más común presenta un máximo frontal con extensión a las regiones centrales. Tiende a disminuir durante la somnolencia, aunque puede aumentar de manera transitoria en niños y adultos.
- **Ondas gamma**: Incluye ondas con frecuencias superiores a 25 Hz. Se asocia con procesos cognitivos de alto nivel, como la atención focalizada y la integración de información sensorial. Es más prominente en las regiones frontales y parietales durante tareas mentales complejas.

Figura 2.3: Rangos de frecuencia y amplitud de las ondas cerebrales.



FUENTE: Urgilés Cárdenas DF, Vásquez Rodríguez GJ. (2017)[14].

2.2 Métodos de adquisición de señales

En la actualidad, existen diversas tecnologías que permiten observar cambios en la actividad cerebral, ya sean eléctricos, magnéticos u ópticos. Estas tecnologías se pueden clasificar en dos tipos principales: invasivas y no invasivas. Las tecnologías invasivas implican la introducción de dispositivos o materiales en el cuerpo mediante cirugía o procedimientos similares, proporcionando datos con alta resolución espacial y temporal debido a su proximidad a las estructuras cerebrales. Sin embargo, conllevan riesgos significativos, como infecciones, daño cerebral o complicaciones quirúrgicas. Por otro lado, las tecnologías no invasivas permiten observar la actividad cerebral sin necesidad de intervenciones físicas dentro del cuerpo, siendo más seguras y cómodas para el paciente, aunque generalmente ofrecen menor resolución espacial y temporal en comparación con las invasivas, evitando los riesgos asociados a procedimientos quirúrgicos o la exposición a sustancias radiactivas.

Dentro de las tecnologías de adquisición de la actividad cerebral no invasivas más destacadas se encuentran los electroencefalogramas (EEG, del inglés *Electroencephalogram*), la resonancia magnética funcional (fMRI, del inglés *Functional Magnetic Resonance Imaging*), la espectroscopia de infrarrojo cercano funcional (fNIRS, del inglés *Functional Near-Infrared Spectroscopy*), la magnetoencefalografía (MEG, del inglés *Magnetoencephalography*), y la tomografía por emisión de positrones (PET, del inglés *Positron Emission Tomography*):

- **Electroencefalografía** (EEG): Registra la actividad eléctrica cerebral mediante electrodos colocados en el cuero cabelludo, ofreciendo una visión continua y detallada de la actividad neuronal con alta resolución temporal.
- **Imagen por Resonancia Magnética Funcional** (fMRI): Mide los cambios en el flujo sanguíneo cerebral relacionados con la actividad neuronal a través de técnicas de imagen magnética, ofreciendo alta resolución espacial para visualizar áreas activas del cerebro.
- **Espectroscopía de Infrarrojo Cercano Funcional** (fNIRS): Utiliza luz infrarroja para medir los cambios en la concentración de oxihemoglobina y desoxihemoglobina en el cerebro, proporcionando información sobre la actividad hemodinámica cerebral con alta portabilidad y menor complejidad operativa.
- **Magnetoencefalografía** (MEG): Captura los campos magnéticos generados por la actividad neuronal en el cerebro con sensores externos, proporcionando información precisa sobre la localización y el *timing* de la actividad cerebral.
- **Tomografía por Emisión de Positrones** (PET): Emplea trazadores radiactivos para medir la actividad cerebral detectando radiación emitida, permitiendo observar procesos metabólicos y químicos en el cerebro, pero implica la exposición a radiación y la necesidad de inyecciones.

En cuanto a las tecnologías invasivas, que requieren procedimientos quirúrgicos, se incluyen la estereoelectroencefalografía (SEEG, del inglés *Stereoelectroencephalography*), la electrocorticografía (ECoG, del inglés *Electrocorticography*) y los potenciales de campo local (LFP, del inglés *Local Field Potentials*):

- **Estereoelectroencefalografía** (SEEG): Utiliza electrodos profundos implantados quirúrgicamente en áreas específicas del cerebro para registrar la actividad eléctrica con alta precisión espacial y temporal. Se utiliza principalmente en la evaluación prequirúrgica de pacientes con epilepsia farmacorresistente, permitiendo una localización precisa del foco epileptógeno.
- **Electrocorticografía** (ECoG): Mide la actividad eléctrica cerebral mediante electrodos colocados directamente sobre la superficie cortical del cerebro, ofreciendo alta resolución espacial y temporal, pero conlleva riesgos quirúrgicos.
- **Potenciales de Campo Local** (LFP): Utiliza electrodos implantados en el cerebro para medir la actividad eléctrica local en áreas específicas, proporcionando datos detallados sobre la actividad neuronal en esas regiones.

La Tabla 2.1 muestra una comparación entre las diferentes tecnologías de adquisición de la actividad cerebral anteriormente mencionadas:

Tabla 2.1: Tecnologías de adquisición de la actividad cerebral.

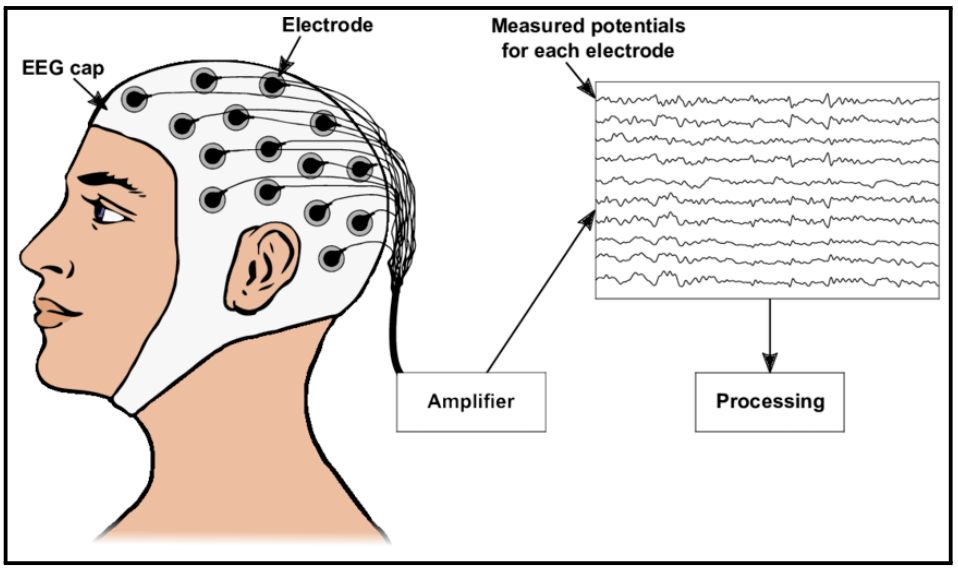
Técnica	Invasivo	Origen	Medición
EEG	No	Neuronal	Eléctrica
fMRI	No	Hemodinámica	Magnética
fNIRS	No	Hemodinámica	Óptica
MEG	No	Neuronal	Magnética
PET	No	Farmacocinético	Química
SEEG	Sí	Neuronal	Eléctrica
ECoG	Sí	Neuronal	Eléctrica
LPF	Sí	Neuronal	Eléctrica

2.3 Electroencefalografía

La electroencefalografía es una técnica de exploración funcional del sistema nervioso central mediante la cual se obtiene el registro de la actividad eléctrica cerebral en tiempo real. En 1929 Hans Berger acuñó el término *electroencefalograma*, en abreviatura EEG, para describir el registro de las fluctuaciones eléctricas en el cerebro captadas por unos electrodos fijados al cuero cabelludo (ver Figura 2.4, 2.5).

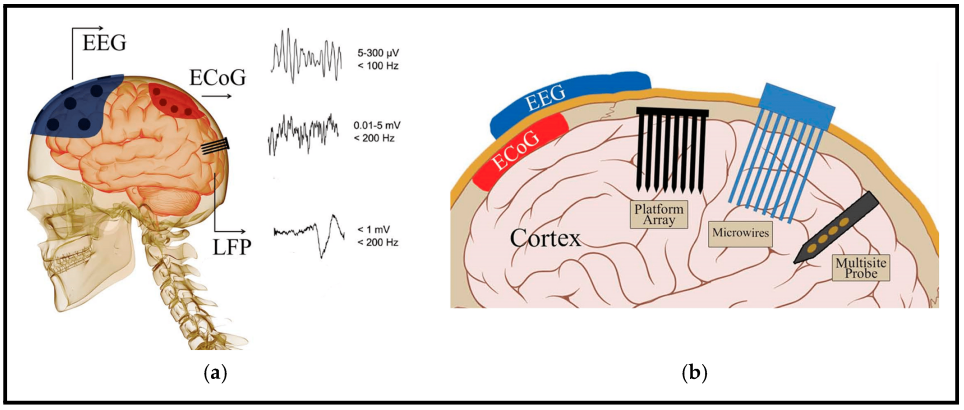
Cada electrodo registra la actividad de un conjunto de neuronas, siendo esta actividad el resultado de la existencia de dipolos eléctricos generados por la suma de potenciales post-sinápticos excitatorios (PPSE) o potenciales post-sinápticos inhibitorios (PPSI) que se generan en el soma y las dendritas de las neuronas piramidales. El EEG amplifica la diferencia de potencial entre los electrodos para evaluar la topografía, polaridad y variación espacial y temporal de estos campos eléctricos cerebrales. Dependiendo del tipo de EEG, los electrodos pueden colocarse en el cuero cabelludo (EEG estándar), en la superficie cortical (ECoG), o dentro del cerebro mediante electrodos intracerebrales (EEG de profundidad).

Figura 2.4: Diagrama esquemático del proceso de registro de señales de EEG. Los electrodos colocados en el cuero cabelludo miden los potenciales eléctricos generados por la actividad neuronal, los cuales son amplificados y procesados digitalmente para su posterior análisis.



FUENTE: Nagel, S. (2019)[15].

Figura 2.5: (a) Representación esquemática de las señales neuronales electroencefalográficas (EEG), electrocorticográficas (ECoG) y de potencial de campo local (LFP), (b) Ubicación de los electrodos: EEG sobre el cuero cabelludo, ECoG sobre la superficie cerebral y LFP mediante electrodos intracraneales. Se incluyen ejemplos de matrices de electrodos, microhilos y sondas multisitio utilizados para capturar la actividad neuronal LFP en diversas profundidades del córtex cerebral.



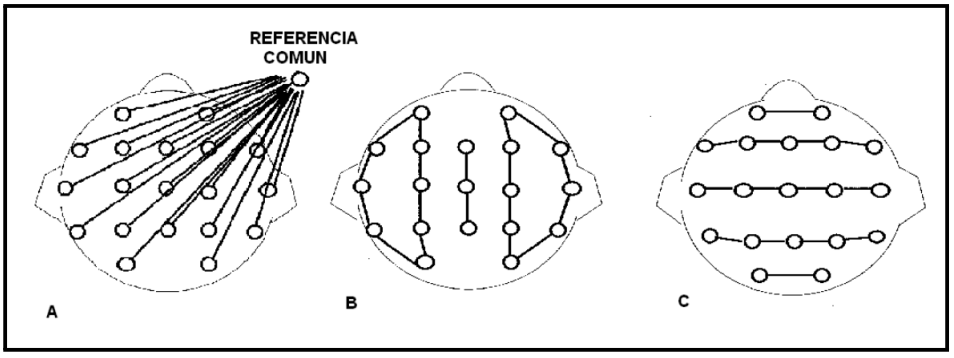
FUENTE: Lago N, Cester A. (2017)[16].

2.3.1 Configuración de los electrodos

La actividad eléctrica cerebral puede ser registrada sobre el cuero cabelludo por medio de electrodos superficiales. Estos electrodos realizan la transducción de corrientes iónicas, generadas y propagadas por células nerviosas, a corrientes electrónicas, permitiendo obtener señales factibles de ser procesadas electrónicamente. Por lo general, estos electrodos son pequeños discos de plata clorurada de aproximadamente 5mm de diámetro, los cuales se adhieren con una pasta conductora. Aplicados correctamente, dan resistencias de contacto muy bajas ($< 4k\Omega$).

Asociadas a los modos de configuración de los electrodos, existen dos tipos principales de configuraciones de registro: bipolar (transversal y longitudinal) y monopolar (o referencial). La configuración bipolar implica el registro de la diferencia de voltaje entre dos electrodos colocados en áreas de actividad cerebral, lo que permite detectar cambios locales en la actividad eléctrica. Esta configuración puede ser transversal o longitudinal, dependiendo de la orientación de los electrodos. Por otro lado, la configuración monopolar, también conocida como referencial, mide la diferencia de potencial entre un electrodo en una zona activa y otro en una zona neutra o sin actividad, como el lóbulo de la oreja. Alternativamente, puede registrar la diferencia de voltaje entre un electrodo activo y el promedio de todos o algunos de los electrodos activos. Esta configuración es útil para obtener una referencia estable y evaluar la actividad general en una región específica del cerebro.

Figura 2.6: Diagrama de configuración de 21 electrodos EEG en los arreglos a) monopolar, b) bipolar transversal y c) bipolar longitudinal.



FUENTE: Recuperado de www.cti.hc.edu.uy.

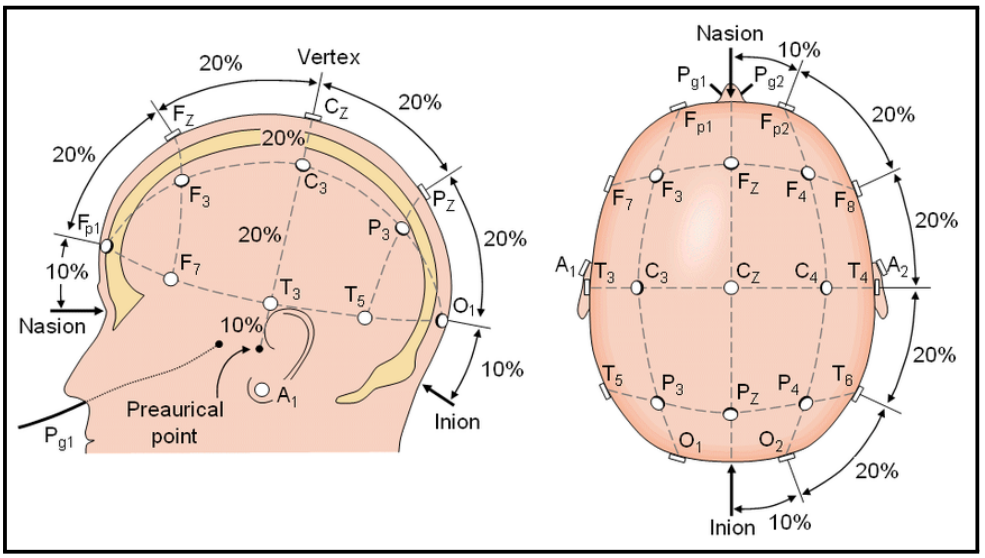
2.3.2 Disposición de los electrodos

Desde 1958, la Federación Internacional de Sociedades de Electroencefalografía y Neurofisiología Clínica estableció la norma estándar de colocación de electrodos. La norma 10-20 divide las distancias del cráneo en incrementos que corresponden al 10 % de la longitud de arco entre el inion y el nasion, con una trayectoria a través del vertex; el 20 % se refiere a la distancia entre dos electrodos, como una separación equidistante. En esta convención, cada punto posee una letra y un subíndice; las letras hacen referencia al lóbulo cerebral al que corresponden y los subíndices

'z' indican la línea media del cráneo, mientras que los subíndices numéricos pares indican el hemisferio derecho y los impares el hemisferio izquierdo. De esta forma, se puede decir que: F corresponde al lóbulo frontal, T al temporal, O al occipital y P al parietal; C corresponde al vertex y A, a la zona del pabellón auditivo [17]. Este sistema utiliza 21 electrodos para cubrir las áreas clave del cuero cabelludo.

Aunque el sistema 10-20 es ampliamente aceptado y utilizado, tiene algunas limitaciones. Por ejemplo, la resolución espacial en áreas específicas del cerebro puede no ser suficiente, y el sistema puede necesitar ajustes en pacientes con anatomías craneales atípicas. En estos casos, se pueden utilizar sistemas de mayor densidad como el 10-10 y el 10-5. Estos enfoques subdividen aún más el patrón 10-20, añadiendo más electrodos para mejorar la resolución espacial. El sistema 10-10 utiliza aproximadamente 64 electrodos y permite una cobertura más detallada del cerebro, mientras que el sistema 10-5 emplea alrededor de 128 electrodos, proporcionando una captura aún más precisa de la actividad cerebral. Estos sistemas avanzados permiten una mejor localización de áreas específicas y una mayor resolución en el análisis de la actividad cerebral, abordando algunas de las limitaciones del sistema 10-20.

Figura 2.7: Sistema de colocación de electrodos 10-20 para EEG estándar.



FUENTE: Novo-Olivas C, Gutiérrez L, Bribiesca J. (2010)[18].

CAPÍTULO

3 Aprendizaje Profundo

El siglo XXI ha sido testigo de una revolución en la Inteligencia Artificial (IA), impulsada en gran medida por los avances en el aprendizaje profundo. Desde su surgimiento, el aprendizaje profundo ha transformado radicalmente la capacidad de las máquinas para comprender, procesar y generar información de manera autónoma.

En el corazón del aprendizaje profundo se encuentran las redes neuronales, inspiradas en la estructura y el funcionamiento del cerebro humano. Estas redes son una de las técnicas más utilizadas en el aprendizaje automático (en inglés, *machine learning*). Este campo de la inteligencia artificial incluye algoritmos basados en modelos matemáticos que se ajustan a los datos disponibles para resolver problemas específicos. Desde el reconocimiento de patrones complejos hasta la generación de contenido creativo, las redes neuronales han demostrado una capacidad innata para abordar tareas que anteriormente se consideraban exclusivas del intelecto humano.

En los inicios de la inteligencia artificial, los esfuerzos se centraban en desarrollar algoritmos que permitieran a las computadoras realizar tareas humanas, como razonar, hablar y reconocer imágenes. Para muchas de estas tareas, se encontraron algoritmos efectivos. Un ejemplo claro es el algoritmo A* para la búsqueda del camino más corto. Sin embargo, para la mayoría de los problemas que los humanos resuelven, no existe un algoritmo evidente que permita a las computadoras hacerlo. Por ejemplo, determinar si un objeto específico está presente en una imagen ha demostrado ser extremadamente difícil para una máquina.

Debido a esta limitación, surgió el interés en técnicas que permitieran obtener estos algoritmos sin detallar los pasos a seguir, similar a cómo los humanos aprenden ciertas tareas: mediante un proceso de aprendizaje basado en datos de ejemplo. Aquí es donde entra en juego el aprendizaje automático, que, a diferencia de las técnicas tradicionales de inteligencia artificial, ajusta modelos predefinidos en base a los datos disponibles del problema para resolverlo.

Cuando se trabaja con técnicas de aprendizaje automático, se utiliza un conjunto de datos para ajustar un modelo predefinido que resuelve un problema específico.

Los datos con los que trabaja el modelo no suelen ser los datos originales en bruto del problema. En su lugar, se definen una serie de características, conocidas como "*features*" en inglés, que se obtienen a partir de los datos originales y son las que utiliza el modelo para resolver el problema.

Este enfoque requiere un estudio previo del problema por parte de expertos que ayuden a definir las mejores características para el modelo, lo que supone una gran cantidad de tiempo y dinero. Esta ha sido la forma tradicional de trabajar con los modelos de aprendizaje automático. Una vez que se dispone de las características más relevantes para solucionar el problema, se elige un modelo de aprendizaje automático (como una red neuronal) y este se encarga de aprender la transformación que hay que aplicar a las características elegidas para obtener la salida esperada.

En contraste con el enfoque tradicional del aprendizaje automático, las técnicas de aprendizaje profundo se encargan de extraer, por sí solas, las características más relevantes de los datos originales para resolver el problema, además de aprender la transformación que hay que aplicar a estas características para dar la salida esperada. Sin embargo, ante problemas más sencillos y con un menor número de datos disponibles, las técnicas de aprendizaje profundo pueden tener dificultades para entrenarse correctamente. Tienden a sobreajustarse al conjunto de datos, en comparación con modelos más sencillos, y su entrenamiento suele ser mucho más costoso. Esto explica en parte por qué, hasta prácticamente el año 2012, no se había utilizado de forma más generalizada este tipo de técnicas para resolver problemas. El abaratamiento del almacenamiento, el aumento de la capacidad de cómputo y el uso de las GPUs para el entrenamiento de los modelos de aprendizaje profundo, junto con los resultados que estos consiguen, han propiciado el auge de este campo.

3.1 Redes neuronales

Existen numerosas formas de definir lo que son las redes neuronales; desde las definiciones cortas y genéricas hasta las que intentan aclarar qué es una «red neuronal» o la «computación neuronal»:

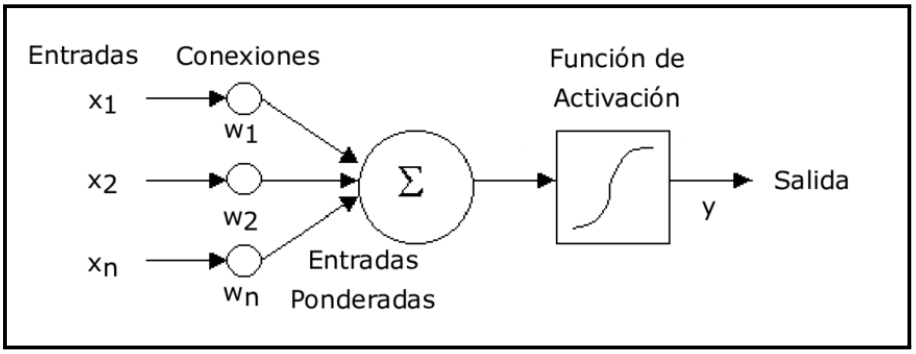
> *"... un sistema de computación hecho por un gran número de elementos simples, elementos de proceso altamente interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas.*
>
> — Robert Hecht-Nielsen, *Neurocomputing: Picking the Human Brain*

> *"... son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico."*
>
> — Teuvo Kalevi Kohonen, *An Introduction to Neural Computing*

Así pues, una Red de Neuronas Artificial (RNA) puede ser concebida como un grafo computacional que emula la estructura y funcionamiento del cerebro humano, compuesto por unidades de proceso idénticas, denominadas neuronas artificiales o nodos. Estas unidades, inspiradas en el sistema nervioso biológico, procesan y transmiten información a través de los arcos en respuesta a estímulos externos.

La Figura 3.1 muestra un ejemplo de una unidad típica de proceso o neurona de una Red Neuronal Artificial. En este ejemplo, la neurona recibe múltiples entradas $(x_1, \ldots, x_n)$, ya sea de otras neuronas o de fuentes externas, cada una vinculada a un peso específico $(w_1, \ldots, w_n)$, que representa la intensidad de la conexión entre la neurona emisora y la receptora. Estas entradas ponderadas se suman y el resultado pasa por una función de activación, que determina la salida de la neurona.

Figura 3.1: Neurona artificial.



FUENTE: Ardila J, Suárez Mantilla L. (2019)[19].

3.2 Aprendizaje e Inferencia

Cuando se trabaja con redes neuronales, o cualquier modelo de *machine learning*, suele hacerse una distinción entre la obtención del modelo y su uso. Estas dos etapas se conocen como etapa de entrenamiento y etapa de inferencia:

- **Etapa de entrenamiento**: En esta etapa, la red aún no tiene los pesos definidos. Se utilizan los datos disponibles para que la red ajuste lo mejor posible la función que se desea aproximar. El ajuste de los pesos puede ser manual, aleatorio o automático mediante un algoritmo de entrenamiento. En el caso de las redes neuronales, suele denominarse a esta etapa como etapa de aprendizaje. El resultado de esta etapa es una red con unos pesos definidos.
- **Etapa de inferencia**: Cuando no se actualizan los pesos de la red, se dice que la red se encuentra en la etapa de inferencia. Esta etapa es útil para validar el entrenamiento de la red y decidir si seguir o no con el entrenamiento en base a los resultados que obtiene. Una de las ventajas de las redes neuronales es que permiten obtener la salida para cualquier valor de entrada, no se limitan a los datos que han utilizado durante la etapa de entrenamiento. Por lo tanto, es posible darle a la red datos que nunca ha visto para hacer predicciones sobre ellos. El resultado de esta etapa es únicamente la salida de la red para los datos de entrada que se le suministren.

La primera etapa en el proceso de implementación de un modelo de machine learning es el entrenamiento, fundamental para definir los pesos de la red neuronal. Una vez seleccionado el modelo adecuado para abordar el problema en cuestión, se procede a entrenarlo con los datos disponibles. Al completar esta fase, se inicia la etapa de inferencia, en la que el modelo no sufre modificaciones y se utiliza para generar salidas basadas en los datos, ya sea del conjunto de datos original o de nuevos datos.

3.3 Paradigmas de aprendizaje

Durante la etapa de entrenamiento se pueden ajustar los pesos de la red de muchas formas diferentes. En el aprendizaje automático, se suele utilizar algún algoritmo de aprendizaje que se encarga de ajustar los pesos del modelo utilizando un conjunto de datos. En base al tipo de dato que se utilice para ajustar el modelo se distinguen tres paradigmas de aprendizaje: el aprendizaje supervisado, no supervisado y por refuerzo.

- **Aprendizaje supervisado**: En este tipo de aprendizaje, el conjunto de datos utilizado para entrenar el modelo está etiquetado, es decir, cada ejemplo de entrenamiento tiene asociada una salida deseada. El objetivo del modelo es aprender la relación entre las entradas y las salidas de los ejemplos de entrenamiento, de manera que pueda generalizar y predecir correctamente las salidas de nuevos ejemplos que no ha visto antes.
- **Aprendizaje no supervisado**: En este caso, el conjunto de datos de entrenamiento no está etiquetado, y el objetivo del modelo es encontrar patrones, estructuras o relaciones ocultas en los datos. El modelo trata de agrupar los ejemplos de entrenamiento en función de sus similitudes o diferencias, sin tener en cuenta una salida deseada. Este tipo de aprendizaje es útil para tareas como la reducción de dimensionalidad, la detección de anomalías o el descubrimiento de conocimiento.
- **Aprendizaje por refuerzo**: En este paradigma, el modelo aprende a través de la interacción con un entorno dinámico y incierto. En lugar de disponer de un conjunto de datos como en los casos anteriores, el modelo recibe una señal de recompensa o castigo en función de las acciones que realiza en el entorno, y su objetivo es aprender una política de acción que maximice la recompensa acumulada a largo plazo. Este tipo de aprendizaje es adecuado para tareas como el control de robots, la optimización de procesos industriales o los juegos.

3.4 Preparación de los datos

Como se ha visto en secciones anteriores, cuando se utiliza una red neuronal, siempre se manejan unos datos de entrada, incluso se consideran una capa de la red. Ya sea para verificar que el modelo funciona correctamente o para ajustar sus pesos, se va a necesitar un conjunto de datos para trabajar con las redes.

Normalmente, el conjunto de datos original o *data set* suele ser dividido en tres conjuntos diferentes: entrenamiento (*train* en inglés), validación (*validation* en inglés) y prueba (*test* en inglés):

- **Entrenamiento** (*train*): Este conjunto de datos se utiliza para ajustar los parámetros de la red neuronal durante el proceso de entrenamiento. El objetivo es que la red aprenda a partir de estos datos, reconociendo patrones y ajustando sus pesos para minimizar el error en las predicciones.
- **Validación** (*validation*): Este conjunto de datos se utiliza durante el entrenamiento para evaluar el desempeño de la red neuronal y ajustar hiperparámetros, como la tasa de aprendizaje o el número de capas ocultas. La validación ayuda a prevenir el sobreajuste, es decir, evitar que la red se vuelva demasiado específica con los datos de entrenamiento y tenga un rendimiento pobre con datos nuevos.
- **Prueba** (*test*): Este conjunto de datos se reserva para evaluar el desempeño final de la red neuronal una vez que el entrenamiento y la validación han sido completados. Los datos de prueba permiten obtener una medida imparcial de la capacidad de generalización de la red neuronal, demostrando cómo se comporta con datos que no ha visto durante el entrenamiento.

Debido a que los conjuntos de datos para el entrenamiento de redes neuronales deben ser especialmente grandes, se suele dejar en torno al 5 % de los datos para validar y otro 5 % para pruebas, dejando el resto de los datos para entrenar la red.

3.5 Aprendizaje basado en el gradiente

Durante el entrenamiento de una red neuronal, la red utiliza el conjunto de datos de entrenamiento para ajustar sus parámetros de acuerdo con un algoritmo de optimización. El objetivo es que la red aprenda a partir de estos datos, identificando patrones y ajustando sus pesos para minimizar el error en las predicciones.

Uno de los métodos más conocidos y utilizados para el aprendizaje en redes neuronales es el algoritmo de retropropagación de errores, o "backpropagation". Este algoritmo se basa en el cálculo del gradiente de la función de error, que evalúa la diferencia entre las predicciones de la red y los valores esperados, para determinar la contribución de cada nodo al error total. En cada iteración, el algoritmo retropropaga el gradiente hacia atrás a través de cada capa de la red, utilizando la regla de la cadena para calcular la contribución relativa de cada nodo al error total. Estas contribuciones se utilizan luego para ajustar los pesos de las conexiones según un algoritmo de optimización, como el descenso de gradiente, con el objetivo de minimizar el error de la red.

En el caso de descenso en gradiente, este proceso de minimización se realiza iterativamente, ajustando los pesos del modelo en la dirección opuesta al gradiente de la función de costo, hasta alcanzar un mínimo local o satisfacer un criterio de convergencia. Matemáticamente, este proceso se puede expresar como:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_n) \tag{3.1}$$

donde η es la tasa de aprendizaje, que controla el tamaño del paso de ajuste, y $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_n)$ es el gradiente de la función de error en la iteración n. A medida que el algoritmo recorre el conjunto de entrenamiento, realiza la actualización de los parámetros en la dirección opuesta al gradiente para cada muestra de entrenamiento, pudiendo realizar múltiples pasadas, o épocas, a través del conjunto de entrenamiento, hasta que el algoritmo converja.

3.6 Capacidad de generalización

A la hora de evaluar el comportamiento de una red neuronal, no sólo es importante saber si la red ha aprendido con éxito los patrones utilizados durante el aprendizaje, sino que es imprescindible, también, conocer el comportamiento de la red ante patrones que no se han utilizado durante el entrenamiento. Es decir, de nada sirve disponer de una red que haya aprendido correctamente los patrones de entrenamiento y que no responda adecuadamente ante patrones nuevos. Es necesario que durante el proceso de aprendizaje la red extraiga las características de las muestras, para poder así responder correctamente a patrones diferentes.

En general, a medida que aumentamos la cantidad de parámetros ajustables en un modelo, este se vuelve más flexible y puede adaptarse mejor a un conjunto de datos de entrenamiento. Se dice que tiene menor error o sesgo. Sin embargo, con modelos más flexibles que poseen un mayor número de parámetros, se observa una mayor variabilidad en el ajuste del modelo entre conjuntos de datos de entrenamiento. Idealmente, se busca un equilibrio entre sesgo y varianza. Se desea un modelo con bajo sesgo, para que las predicciones sean precisas en general, y baja varianza, para que el modelo generalice bien a nuevos datos. Sin embargo, en la práctica, suele existir un compromiso entre ambos, y reducir uno puede aumentar el otro.

En este contexto, es fundamental implementar estrategias que ayuden a gestionar y mitigar el sobreajuste (*overfitting*) y asegurar que el modelo no solo se ajuste bien a los datos de entrenamiento, sino que también mantenga una buena capacidad de generalización. Diversas técnicas y métodos se utilizan para lograr este equilibrio, cada una abordando diferentes aspectos del proceso de entrenamiento y evaluación del modelo. Entre estas técnicas se incluyen:

- **Validación cruzada** (*Cross-validation*)[20]: Esta técnica divide el conjunto de datos en varios subconjuntos o *folds*. El modelo se entrena con una combinación de estos subconjuntos y se valida con el restante. Este procedimiento se repite múltiples veces, alternando el subconjunto de validación en cada iteración. Finalmente, se promedia el rendimiento obtenido en todas las iteraciones, lo que ofrece una estimación más precisa del desempeño del modelo en datos nuevos y ayuda a identificar posibles problemas de sobreajuste.
- **Dilución** (*Dropout*)[21]: Esta técnica implica la desactivación aleatoria de ciertos nodos de la red durante el entrenamiento para mejorar la capacidad de generalización y evitar una dependencia excesiva de los datos de entrenamiento. *Dropout* utiliza una distribución de Bernoulli con probabilidad p_l, donde l es el índice de la capa, para desactivar aleatoriamente neuronas en cada capa, generando una combinación distinta de neuronas activas, o subredes, en cada paso del entrenamiento. Una vez finalizado el entrenamiento, los pesos del

modelo se ajustan por su correspondiente valor esperado $(1 - p_l)$ para reflejar la proporción de neuronas activas durante el entrenamiento. Este ajuste evita que algunas neuronas se vuelvan demasiado dominantes y promueve una distribución más equitativa de la carga entre todas las neuronas de la red.

- **Parada temprana** (*Early stopping*)[22]: Esta técnica previene el sobreajuste al interrumpir el entrenamiento del modelo antes de que haya terminado, basándose en el rendimiento del conjunto de validación. Durante el entrenamiento, se monitorea el desempeño del modelo en el conjunto de validación. Si el rendimiento no mejora después de un número definido de iteraciones, conocido como el "parámetro de parada", el entrenamiento se detiene. Esto ayuda a evitar que el modelo se ajuste demasiado a los datos de entrenamiento, manteniendo así su capacidad de generalización y mejorando su rendimiento en datos no vistos previamente.

3.7 Modelos neuronales

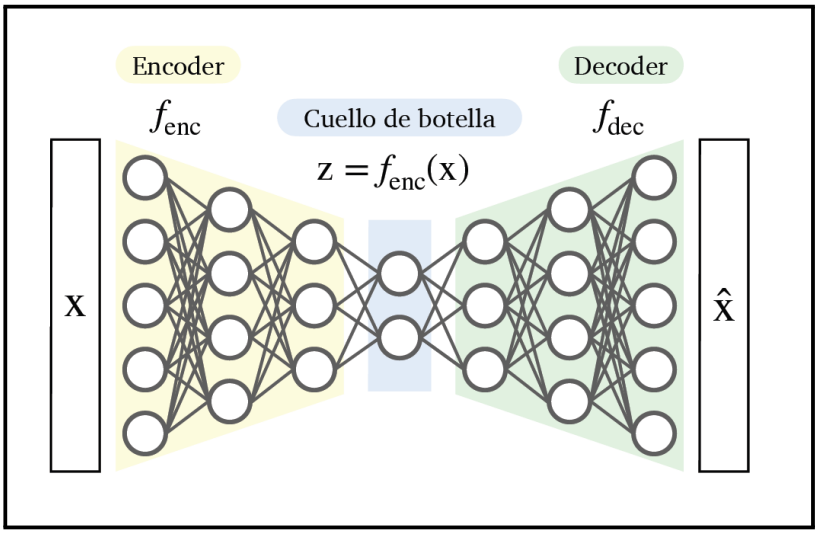
Antes de presentar los VAEs, consideramos apropiado brindar una breve introducción a la estructura general de los autocodificadores. Esta arquitectura precede en la historia a los variacionales y comparte muchos aspectos en común, aunque sus aplicaciones no se enfoquen en la generación de datos como sucede con los VAEs.

3.7.1 Autocodificadores

Los autocodificadores son un tipo de red neuronal, especializada en la representación de un espacio dimensional de origen en otro más pequeño. El objetivo de esta transformación es representar la información observada en un vector de menor dimensión que capture la estructura subyacente de los datos, preservando la información relevante y descartando la redundante o superflua para su reconstrucción. Se pretende que dicha compresión sirva para codificar en la red una representación significativa de los verdaderos factores explicativos de las señales observadas. A la representación vectorial reducida se le denomina espacio latente.

Los autocodificadores (ver Figura 3.2) son una propuesta de arquitectura para llevar a cabo este proceso. Este tipo de redes neuronales se compone de dos partes principales: un codificador f_{enc}, que proporciona una representación latente $\mathbf{z}$ de los datos de entrada $\mathbf{x}$; y un decodificador f_{dec}, que reconstruye los datos de entrada a partir de su representación latente $\mathbf{z}$, devolviendo una estimación de la entrada $\hat{\mathbf{x}}$. Durante el entrenamiento, la red se optimiza para minimizar la diferencia entre la entrada original $\mathbf{x}$ y la salida reconstruida $\hat{\mathbf{x}}$. Este proceso permite que el codificador aprenda a extraer y comprimir la información más relevante de los datos de entrada, y que el decodificador aprenda a reconstruir la entrada original a partir de esta representación comprimida. El cuello de botella en el espacio latente obliga a la arquitectura a comprimir en representaciones más sencillas la información relevante, que será utilizada para la reconstrucción del original.

Figura 3.2: Ejemplo de arquitectura de un autocodificador.



FUENTE: González Muñiz A. (2022)[23].

3.7.2 Autocodificadores Variacionales (VAEs)

Los autocodificadores variacionales (VAEs, por sus siglas en inglés, *Variational Autoencoders*) heredan la estructura de los autocodificadores tradicionales, añadiendo restricciones adicionales en el cuello de botella que transforman la arquitectura determinista de los autocodificadores en un modelo probabilístico. A diferencia de los autocodificadores convencionales, que generan representaciones latentes de manera determinista, los VAEs adoptan un enfoque probabilístico para capturar la distribución subyacente de los datos. Este enfoque no solo permite generar nuevas muestras similares a los datos de entrenamiento, sino también inferir representaciones latentes para datos no vistos.

Una red codificadora, también conocida como red de reconocimiento o *encoder*, transforma la distribución original de los datos de entrada a un espacio latente estocástico, con probabilidad $q_\phi(\mathbf{z}|\mathbf{x})$, representado por una distribución más sencilla (por ejemplo, una distribución normal). A partir de esta distribución, una red decodificadora (o red de generación) genera los datos de reconstrucción, utilizando la información latente para producir los datos de salida.

Durante el entrenamiento, la red se optimiza para maximizar la evidencia marginal de los datos de entrada bajo el modelo, con el objetivo de aproximar la verdadera distribución a posteriori de los datos. Este proceso se centra en maximizar la probabilidad marginal de los datos $p(\mathbf{x})$, que se obtiene integrando sobre todas las posibles variables latentes. Sin embargo, calcular directamente esta integral suele ser computacionalmente intratable debido a la alta dimensionalidad del espacio latente. En su lugar, se emplea una aproximación basada en la descomposición de la evidencia. Se introduce una distribución aproximada, $q_\phi(\mathbf{z}|\mathbf{x})$, para estimar la verdadera distribución a posteriori de los datos de entrada, $p(\mathbf{z}|\mathbf{x})$, y se calcula la divergencia de Kullback-Leibler[1] ($\mathcal{KL}$) entre ellas.

$$\underbrace{\mathcal{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x}))}_{\mathcal{KL}\geq 0} = \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} \tag{3.2}$$

$$= \int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})}\, d\mathbf{z} \tag{3.3}$$

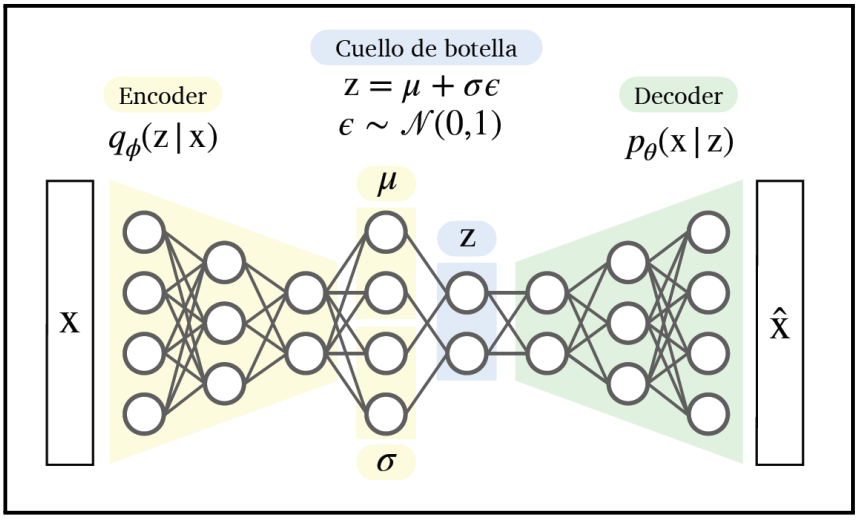
$$= \log p(\mathbf{x}) + \underbrace{\int q_\phi(\mathbf{z}|\mathbf{x}) \log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})} d\mathbf{z}}_{\text{ELBO}} \tag{3.4}$$

El objetivo del entrenamiento se convierte entonces en maximizar un límite inferior de la evidencia marginal, conocido como ELBO (*Evidence Lower BOund*), que implica minimizar la divergencia $\mathcal{KL}$ entre la distribución aproximada y la verdadera posterior, mientras se maximiza la probabilidad de reconstrucción de los datos.

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \mathcal{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \tag{3.5}$$

En esta ecuación (3.5), el primer término representa la pérdida de reconstrucción, que mide la capacidad del decodificador para generar muestras similares a las entradas originales, mientras que el segundo término representa la pérdida de similitud, que cuantifica la divergencia entre la distribución a posteriori del codificador y la distribución a priori.

Figura 3.3: Diagrama representativo de la arquitectura VAE.



FUENTE: González Muñiz, A. (2022)[23].

[1]En teoría de la probabilidad y teoría de la información, la divergencia de Kullback-Leibler ($\mathcal{KL}$) es una medida no simétrica de la similitud o diferencia entre dos funciones de distribución de probabilidad P y Q. Representa la cantidad de información perdida cuando Q se utiliza para aproximar P, y siempre es no negativa, siendo cero solo si P y Q son idénticas.

CAPÍTULO

4 Metodología

En este capítulo se describe la metodología empleada para abordar la síntesis de voz a partir de registros de actividad cerebral, organizada en tres etapas principales: recolección de datos, síntesis de voz y evaluación de calidad. La primera etapa abarca la captura sincronizada de señales de voz y actividad cerebral, garantizando la precisión y representatividad de los datos. La segunda etapa se enfoca en la generación de voz a partir de los registros cerebrales, empleando técnicas avanzadas de aprendizaje profundo y síntesis paramétrica para reconstruir el habla de manera efectiva. Finalmente, en la tercera etapa, se evalúa la calidad de la voz sintetizada mediante diversas métricas, tanto objetivas como subjetivas, con el fin de analizar los resultados obtenidos y validar la eficacia del sistema desarrollado.

4.1 Grabación de datos

En esta primera etapa, se llevó a cabo la captura sincronizada de señales de voz y registros de actividad cerebral de pacientes con electrodos profundos implantados en el Hospital Universitario Virgen de las Nieves (HUVN) de Granada. El propósito de esta fase fue identificar los patrones cerebrales relacionados con la producción del habla articulada, buscando una comprensión más profunda de los procesos cognitivos implicados en la generación y comprensión del lenguaje en condiciones de voz articulada. Para ello, se diseñó un experimento utilizando la herramienta Psychopy[24], con el fin de recolectar datos simultáneamente de la voz y la actividad cerebral de los pacientes. Durante el experimento, se registró el habla de los participantes y su actividad cerebral mediante electrocorticografía intracraneal (iEEG).

4.1.1 Participantes

Se proporcionó información detallada sobre la investigación a los pacientes ingresados en la Unidad de Epilepsia Refractaria del Hospital de Neurotraumatología y Rehabilitación de Granada, invitándolos a participar en el estudio que se llevó a

cabo en la unidad de Video-EEG del Hospital Universitario Virgen de las Nieves de Granada. Antes de proceder con la recolección de datos, cada participante recibió una Hoja Informativa y un Consentimiento Informado, los cuales debían ser firmados para formalizar su participación en el proyecto. Dichos documentos fueron previamente aprobados por el Comité Ético de Investigación Humana de la Universidad de Granada (UGR). Se informó a los participantes que podían retirarse del estudio en cualquier momento, sin necesidad de ofrecer explicaciones y sin que esto afectara de ninguna manera su atención médica. Además, los participantes fueron compensados económicamente por su tiempo y colaboración.

La Tabla 4.1 presenta una descripción detallada de los participantes, incluyendo su identificación, sexo, edad, el tipo de electrodo implantado (ECoG o SEEG) y la distribución de los electrodos. En esta tabla, el subíndice asociado a cada electrodo indica el número de contactos o canales presentes en el mismo.

Tabla 4.1: Información de los pacientes y distribución de electrodos[2].

ID	Sexo	Edad	Electr.	Distrib. Electr.
F01	F	27	ECoG	$PF1_{18}$, $PF2_{22}$, $PF3_{12}$, $PF4_{12}$
F05	F	48	SEEG	R'_{18}, TP'_{8}, NA_{12}, HC'_{12}
F07	F	36	SEEG	FP_{10}, HC_{10}, HT_{10}, Pi_{15}, GC_{10}, FCA_{12}, OT_{8}, GL_{8}, CU_{8}
F08	F	46	SEEG	NA_{12}, NA'_{12}, HC_{12}, HC'_{12}, HT_{12}, HT'_{10}
F10	F	49	SEEG	HT_{10}, PI_{8}, IA_{10}, HC_{10}, NA_{12}, IP_{10}, NA'_{10}, HC'_{12}
M06	M	54	SEEG	CR_{15}, CC_{15}, NA_{12}, HC_{10}, H_{8}, HT_{12}, OT_{10}

4.1.2 Diseño del experimento

Durante el experimento, se solicitó a los participantes que leyeran en voz alta una serie de pseudopalabras con la estructura vocal-consonante-vocal (VCV). Para asegurar una adecuada variedad en los pares mínimos VCV, las consonantes se seleccionaron siguiendo el Alfabeto Fonético Internacional (IPA)[25], abarcando distintos puntos y modos de articulación. Las pseudopalabras se organizaron en seis bloques de 50 elementos cada uno, compuestos por dos vocales y una consonante (ver Listado 4.1). El objetivo del experimento era evaluar la capacidad de los participantes para decodificar y pronunciar correctamente estas pseudopalabras[3], mientras se registraban las señales de audio correspondientes (ver Figura 4.1).

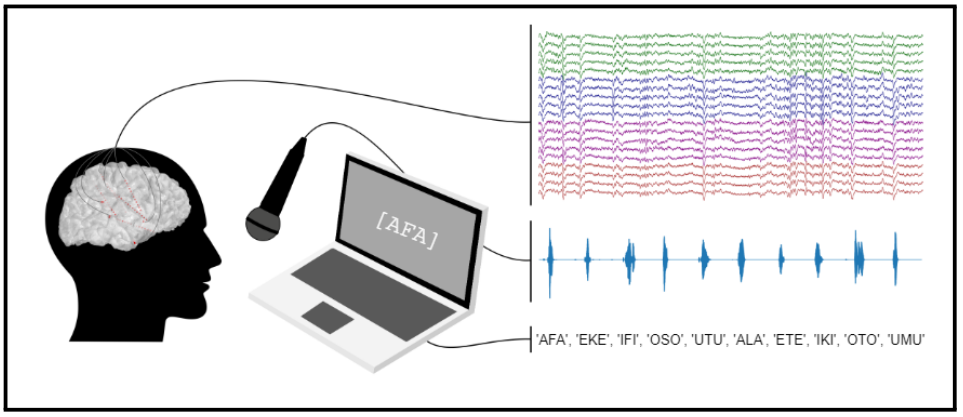
Listado 4.1: Corpus de pseudopalabras.

```
AFA AJA AKA ALA AMA ANA APA ARA ASA ATA
EFE EJE EKE ELE EME ENE EPE ERE ESE ETE
IFI IJI IKI ILI IMI INI IPI IRI ISI ITI
OFO OJO OKO OLO OMO ONO OPO ORO OSO OTO
UFU UJU UKU ULU UMU UNU UPU URU USU UTU
```

El proceso experimental comenzó con una pantalla de inicio que indicaba el comienzo del experimento, seguido de un ciclo iterativo para cada pseudopalabra,

compuesto por cuatro fases. Primero, se mostraba un cuadro vacío ([ ]) en la pantalla durante un intervalo de 1.5 a 2.0 segundos, con el fin de evitar que los participantes anticiparan la duración del bloque y mantener su atención. A continuación, se presentaba una pseudopalabra en el centro de la pantalla durante 1.5 segundos (por ejemplo, [APA]), tiempo durante el cual el participante debía decodificarla mentalmente sin verbalizarla. Después, se volvía a mostrar un cuadro vacío por un período aleatorio de 1.5 a 2.0 segundos, seguido de un cuadro con tres asteriscos ([***]) durante 1.5 segundos, momento en el que el participante debía pronunciar en voz alta la pseudopalabra (ver Figura 4.2). Este ciclo se repitió para cada una de las 50 pseudopalabras del corpus, con seis repeticiones por pseudopalabra presentadas en orden aleatorio para evitar sesgos. La actividad cerebral se registró utilizando un amplificador Natus Quantum de 256 canales a una frecuencia de muestreo de 512 Hz. Simultáneamente, se grabó el habla de los participantes con un micrófono Blue Yeti a 44.1 kHz, utilizando la herramienta Lab Streaming Layer (LSL)[26] para sincronizar los datos de audio e iEEG. Para reducir la fatiga de los participantes, los datos se recopilaron en sesiones de 20 a 30 minutos, acumulando un total de 300 registros de pseudopalabras por participante, almacenados en un único archivo XDF (*Extensible Data Format*), con los datos de iEEG y audio sincronizados.

Figura 4.1: Resumen del experimento: Registro simultáneo de la actividad eléctrica intracraneal (iEEG) y los datos de voz en formato WAV. En la figura se muestra la actividad correspondiente a un registro de 20 segundos de iEEG y audio. Los colores presentes en los trazos de iEEG representan cada uno de los electrodos, y los canales del mismo color, cada uno de los contactos que componen el electrodo.

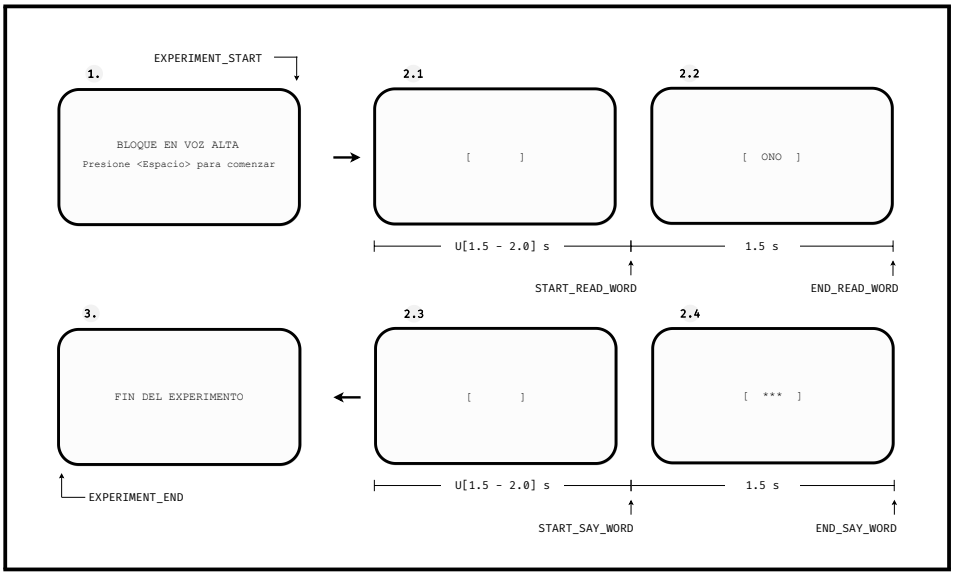


FUENTE: Adaptado de Herff C et al. (2022)[27].

[2]Glosario de acrónimos: NA (Amígdala), HC (Cabeza del Hipocampo), OT (Occipital Temporal), PI (Subparietal), FCA (Fisura Calcarina Anterior), HT (Cola del Hipocampo), R (Sous Rostral), TP (Polo Temporal), CR (Cajal-Retzius), CC (Corpus Callosum), H (Hipotálamo), PF (Cortex Prefrontal), FP (Frontal Precentral), GC (Giro Cingulado), GL (Giro Lingual), CU (Corteza Uncinada), IA (Área Infralímbica), IP (Área Parietal Inferior).

[3]El uso de pseudopalabras resulta esencial en este contexto, ya que permite a los investigadores controlar las variables léxicas y eliminar los efectos de la familiaridad y el conocimiento previo sobre palabras reales. Esto facilita una evaluación más precisa de las estructuras fonológicas y morfológicas, libres de significados preexistentes, y promueve el estudio del procesamiento subléxico, ayudando a comprender cómo las personas analizan y producen sílabas y sonidos en el lenguaje.

Figura 4.2: Diseño del experimento. Para asegurar una sincronización precisa entre la voz y los datos de EEG, se implementó un sistema de marcadores de hardware a través del puerto serial del equipo de registro. Los triggers utilizados son: `EXPERIMENT_START` (inicio del experimento), `START_READ_WORD` (inicio de la lectura), `END_READ_WORD` (fin de la lectura), `START_SAY_WORD` (inicio de la lectura en voz alta), `END_SAY_WORD` (fin de la lectura en voz alta) y `EXPERIMENT_END` (fin del experimento).



FUENTE: Elaboración propia.

4.2 Síntesis de voz

Una vez obtenidos los registros de audio y de electrocorticografía intracraneal (iEEG) del experimento, se procedió a la síntesis de voz. En primer lugar, se segmentaron los datos de iEEG y de audio para aislar las secciones correspondientes a la voz. Posteriormente, se extrajeron y preprocesaron las características necesarias para entrenar y validar los modelos de síntesis de voz desarrollados.

4.2.1 Segmentación de los datos

Previo a la segmentación de los datos, se llevó a cabo un proceso preliminar de eliminación de artefactos con el fin de de excluir aquellos segmentos del experimento en los que los registros de voz o iEEG pudieran estar comprometidos por interferencias, como artefactos de iEEG, ruido externo o problemas de pronunciación, tales como palabras incompletas, pronunciaciones ininteligibles o errores en la selección de palabras. Posteriormente, se utilizaron los marcadores registrados durante el experimento para segmentar bloques de lectura en voz alta de 1.5 segundos, asegurando que cada segmento de voz estuviera correctamente alineado con su correspondiente registro de actividad cerebral. A pesar de las diferencias en las frecuencias de muestreo, tanto los registros de audio como los de iEEG se truncaron para coincidir en duración, ajustándose al segmento más corto. Como resultado, se generaron dos tipos de registros por paciente: uno con los registros de

voz en formato WAV, nombrados como `XXX_YYY.wav`, donde `XXX` representa el índice de aparición de la pseudopalabra en el experimento y `YYY` la pseudopalabra; y otro correspondiente a los registros individuales de iEEG en formato de array de NumPy, siguiendo el mismo esquema de nomenclatura (por ejemplo, `180_APA.wav` para el archivo de audio y `180_APA.npy` para el registro de iEEG).

4.2.2 Extracción de características

Tras la segmentación de los datos, se procedió a la extracción de características de los registros de audio y de electrocorticografía intracraneal (iEEG). Para los registros de audio, se realizó un proceso de decimación, reduciendo la frecuencia de muestreo de 44,1 kHz a 16 kHz, con el fin de facilitar el procesamiento. En el caso de los registros de iEEG, se aplicó un procedimiento de detrending para eliminar desplazamientos o tendencias en las señales de cada canal. Posteriormente, se utilizaron dos filtros IIR de cuarto orden con rechazo de banda y anchos de banda de 4 Hz, centrados en 100 Hz y 150 Hz, con el objetivo de suprimir el ruido de línea del amplificador y su primer armónico, también aplicados por canal.

Para las señales de iEEG, se extrajeron características gamma[28], asociadas con oscilaciones de alta frecuencia vinculadas a la actividad cerebral. Estas características se obtuvieron para cada canal y para cada pseudopalabra registrada, abarcando la banda de frecuencia de 70 a 170 Hz. Para este proceso, se utilizaron ventanas de 25 milisegundos de duración, con un desplazamiento equivalente.

En el caso de los registros de audio, se extrajeron los siguientes parámetros utilizando el *vocoder* WORLD[29]:

- **Coeficientes Cepstrales de Frecuencia Mel** (CCFM): Describen la envolvente espectral de una señal de audio, capturando las características más relevantes del espectro de frecuencias para la percepción humana.
- **Aperiodicidad de Banda** (AB): Evalúa la relación de potencia promedio entre la potencia total y la potencia en cada índice de frecuencia por tramo de voz, para proporcionar información sobre la estructura espectral de la señal[30].
- **Logaritmo de la Frecuencia Fundamental** (Log F0): Representa el logaritmo natural de la frecuencia fundamental de la señal de voz por tramo de voz.
- **Sonoro/No Sonoro** (SNS): Clasifica los segmentos de voz como sonoros, producidos con vibración de las cuerdas vocales, o no sonoros, generados sin vibración de las cuerdas vocales, con una estructura más aleatoria o ruidosa.

Este proceso resultó en matrices de características para cada pseudopalabra. Para el iEEG, la matriz resultante tuvo un tamaño de $N \times P$, donde N corresponde al número de tramas (38 en este caso, considerando la duración de los registros, la frecuencia de muestreo, y el tamaño de la ventana y el desplazamiento utilizados) y P al número de canales iEEG registrados (ver Tabla 4.1). De manera similar, para el audio se generó una matriz de características de tamaño $N \times 28$, con 38 tramas y 28 características por trama, incluyendo 25 coeficientes CCFM y un coeficiente para AB, Log F0 y SNS.

4.2.3 Preprocesamiento

Previo al entrenamiento de los modelos, se implementó un esquema de validación cruzada con 5 pliegues para cada paciente, dividiendo los datos en conjuntos de entrenamiento (70 %) y de prueba (30 %), por separado para las características de iEEG y audio. Estos conjuntos se concatenaron horizontalmente, dando lugar a un total de 4 matrices: $\mathbf{X}_t \in \mathbb{R}^{M\times 28}$, $\mathbf{X}_v \in \mathbb{R}^{D\times 28}$, $\mathbf{Y}_t \in \mathbb{R}^{M\times P}$ y $\mathbf{Y}_v \in \mathbb{R}^{D\times P}$, correspondientes a las características de iEEG $(\mathbf{X}_t, \mathbf{X}_v)$ y audio $(\mathbf{Y}_t, \mathbf{Y}_v)$ para los conjuntos de entrenamiento y validación, respectivamente.

Para las características de iEEG, el preprocesamiento incluyó la creación de una ventana de contexto de 11 tramas. A cada trama se le concatenó verticalmente la información de contexto de las 5 tramas inmediatamente anteriores y las 5 tramas inmediatamente posteriores. Luego, se aplicó una normalización[4] centrada en la media, un análisis de componentes principales (ACP) que retuvo el 99 % de la varianza (ver Apéndice A), y, posteriormente, una normalización estándar en media y varianza.

En el caso de las características de audio, se adoptó un enfoque distinto. Se extrajeron las velocidades[31] asociadas a cada trama, capturando la dinámica temporal de la señal, seguido de una normalización[4] estándar en media y varianza. Esto resultó en matrices de dimensiones $\hat{\mathbf{Y}}_t \in \mathbb{R}^{M\times 55}$, $\hat{\mathbf{Y}}_v \in \mathbb{R}^{D\times 55}$ para los conjuntos de entrenamiento y validación de los datos de audio, integrando los CCFM, AB, log F0 y sus respectivas velocidades, seguido del parámetro SNS; y $\hat{\mathbf{X}}_t \in \mathbb{R}^{M\times P}$, $\hat{\mathbf{X}}_v \in \mathbb{R}^{D\times P}$ para los conjuntos de entrenamiento y validación de los datos de iEEG, con P el número de componentes principales extraídas tras aplicar ACP.

4.2.4 Descripción de los modelos

Se entrenaron autocodificadores variacionales (VAE) para cada parámetro de audio, excepto el SNS, utilizando una arquitectura de red de siete capas ocultas. El codificador consistía en capas con 512, 256 y 128 neuronas, seguido de un cuello de botella con 8 neuronas. El decodificador, por su parte, incluía capas con 128, 256 y 512 neuronas, generando una salida que representaba el doble de la dimensión de la entrada, correspondiente a la media y la log-varianza predichas. La función de pérdida combinaba la divergencia de Kullback-Leibler con la log-verosimilitud negativa. Posteriormente, se entrenó un segundo VAE para predecir los parámetros de audio a partir de características iEEG. Este segundo VAE utilizó el decodificador del primer VAE para estimar la media y la log-varianza, siguiendo una estructura similar de capas ocultas al codificador del primero. El objetivo de este segundo VAE era mejorar la predicción de los parámetros de audio a partir de la representación latente aprendida en el primer VAE.

Para el parámetro SNS, se empleó un modelo de red neuronal de tipo perceptrón multicapa con 3 capas de 256 neuronas cada una. Las capas ocultas utilizaban funciones de activación ReLU, mientras que la capa de salida empleaba una función de activación sigmoide. La función de pérdida aplicada fue la *binary cross entropy*, y se

[4]En el caso de los conjuntos de validación, las normalizaciones se realizaron utilizando la media y desviación estándar (si procede) obtenidas del conjunto de entrenamiento para cada tipo de dato.

incorporaron capas de *dropout* con una probabilidad de 0.1 para mejorar la generalización y evitar el sobreajuste.

Todos los modelos se entrenaron durante 250 épocas con una tasa de aprendizaje de 10^{-4} y el optimizador *Stochastic Gradient Descent*, con un parámetro de momento de 0.9. Una vez completado el entrenamiento, se validaron los resultados utilizando muestras individuales del subconjunto de prueba. Para cada parámetro, se calcularon los vectores de medias y logaritmos de varianzas asociados a cada entrada para el segundo VAE. Estos vectores se denormalizaron y se introdujeron en el algoritmo MLPG (*Maximum Likelihood Parameter Generation*)[31] para suavizar las trayectorias de los parámetros a partir de sus velocidades. En el caso del parámetro SNS, el vector de valores se redondeó al entero más próximo, resultando en un vector binario que representaba los tramos con y sin voz. Finalmente, los parámetros predichos se concatenaron en una matriz de dimensiones $N \times 28$, y estas características sintetizadas se compararon con las originales para evaluar el rendimiento de los modelos.

4.3 Métricas de evaluación

Para evaluar la calidad de la voz sintetizada, se emplearon dos tipos de métricas que abordan distintos aspectos de la síntesis de voz. La primera clase, denominada **Métricas de Calidad de Parámetros Sintetizados**, se enfoca en medir la precisión y fidelidad de los parámetros de la voz en comparación con los originales. Dentro de esta categoría se incluyen:

- **Distorsión Mel-Cepstral (DMC)**[32]: Mide la diferencia entre los coeficientes mel-cepstrales de la señal original y su versión sintetizada en decibelios. Un valor más bajo de DMC indica una mayor fidelidad en la síntesis, reflejando una menor discrepancia entre ambas señales. En la fórmula (4.1), $\mathbf{x}$ y $\hat{\mathbf{x}}$ representan los coeficientes mel-cepstrales de la señal original y la sintetizada, respectivamente; N es el número total de tramas de la señal sintetizada, y d es el número de coeficientes mel-cepstrales calculados por trama.

$$\text{DMC}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{10}{\ln(10)} \frac{1}{N} \sum_{i=1}^{N} \sqrt{2 \sum_{j=1}^{d} (\mathbf{x}_{ij} - \hat{\mathbf{x}}_{ij})^2} \tag{4.1}$$

- **Distorsión de Aperiodicidades de Banda** (DAB)[32]: Evalúa la distorsión en las aperiodicidades de la señal de voz en diferentes bandas de frecuencia, expresada en dB, entre las aperiodicidades de banda de la señal original y y la sintetizada $\hat{y}$. Un valor bajo de DAB indica una mayor similitud en las características de aperiodicidad entre la señal sintetizada y la original.

$$\text{DAB}(y, \hat{y}) = \frac{10}{\ln(10)} \frac{1}{N} \sum_{i=1}^{N} \sqrt{2(y_i - \hat{y}_i)^2} \tag{4.2}$$

- **Raíz del Error Cuadrático Medio en Log F0** ($\text{RECM}_{\ln F_0}$)[32]: Evalúa la raíz del error cuadrático medio del logaritmo natural de la frecuencia fundamental (F_0) entre la señal original f y la señal sintetizada $\hat{f}$. Un valor bajo de

$\text{RECM}_{\ln F0}$ indica una mayor precisión en la reproducción de la frecuencia fundamental, lo que se traduce en una mejor calidad y fidelidad de la síntesis en términos de la variación tonal y prosódica de la señal original.

$$\text{RECM}_{\ln F_0}(f, \hat{f}) = \sqrt{\frac{\sum_{i=1}^{N}(\ln(f_i) - \ln(\hat{f}_i))^2}{N}} \tag{4.3}$$

- **Tasa de Error Sonoro/No Sonoro** (TESNS)[32]: Mide la precisión en la clasificación de segmentos de voz y no voz en la señal sintetizada, como el porcentaje de segmentos mal clasificados respecto al total. Una tasa baja indica mayor exactitud en la síntesis de zonas vocales y no vocales. En la fórmula (4.4), v_i y $\hat{v}_i$ representan los segmentos de la señal original y sintetizada, respectivamente, y N es el número total de segmentos analizados.

$$\text{TESNS}(\,\%) = \frac{1}{N}\sum_{i=1}^{N} S_i \times 100, \qquad S_i = \begin{cases} 0, & \text{si } v_i = \hat{v}_i \\ 1, & \text{si } v_i \neq \hat{v}_i \end{cases} \tag{4.4}$$

La segunda categoría de métricas, las **Métricas de Calidad Perceptual**, se centra en evaluar la inteligibilidad y la calidad percibida de la señal de voz sintetizada desde el punto de vista del oyente. Estas métricas proporcionan una medida de cómo se percibe la voz sintetizada en comparación con la señal original, abordando aspectos de la percepción humana que las métricas objetivas no siempre capturan completamente. Esto incluye la naturalidad, claridad y fluidez del habla, ofreciendo así una evaluación más completa y centrada en la experiencia auditiva del usuario.

- **STOI** (*Short-Time Objective Intelligibility*) [33]: Esta métrica evalúa la inteligibilidad del habla procesada, proporcionando un valor entre 0 (completamente ininteligible) y 1 (completamente inteligible). STOI compara las envolventes temporales de corta duración entre la señal de habla limpia y la degradada, utilizando segmentos de aproximadamente 386 ms. Este método emplea un banco de filtros y opera con una frecuencia de muestreo de 10 kHz, estimando la correlación promedio entre las envolventes de ambas señales.

- **PESQ** (*Perceptual Evaluation of Speech Quality*) [34]: Desarrollado por la Unión Internacional de Telecomunicaciones (ITU) bajo la recomendación P.862, PESQ es una herramienta objetiva que evalúa la calidad de la voz en sistemas de telecomunicaciones. Utiliza un enfoque psicoacústico para analizar la calidad del habla transmitida, considerando aspectos como la naturalidad, claridad y la inteligibilidad del discurso. Esta métrica compara la señal de voz original con la señal de voz transmitida, proporcionando una puntuación en una escala que va de -0.5 a 4.5, donde 4.5 representa la máxima calidad perceptual.

CAPÍTULO

Resultados experimentales 5

En este capítulo se presentan los resultados del análisis de la calidad de la voz sintetizada a partir de registros cerebrales para los seis pacientes estudiados (ver Tabla 4.1). Se emplearon diversas métricas objetivas y perceptuales para evaluar la precisión y naturalidad de la voz generada, incluyendo la distorsión cepstral Mel, la distorsión de aperiodicidades de banda, el error cuadrático medio de log f0, la tasa de error sonoro/no sonoro, y las métricas perceptuales STOI y PESQ. Para asegurar la validez y robustez de los resultados, se implementó un esquema de validación cruzada que mantuvo constantes los conjuntos de prueba para cada paciente a lo largo de los diferentes pliegues del análisis.

5.1 Evaluación de la calidad de la voz sintetizada

La evaluación de la calidad de la voz sintetizada mostró resultados consistentes entre los seis pacientes, sugiriendo que la selección de conjuntos de entrenamiento y validación no afectó significativamente los resultados obtenidos. Sin embargo, las métricas evaluadas indicaron que la calidad de la voz sintetizada no fue satisfactoria en ninguno de los casos analizados. La distorsión mel-cepstral varió entre 8 y 15 dB, con valores particularmente altos para el paciente F10, cercanos a 17 dB. De manera similar, la distorsión de aperiodicidades de banda fluctuó entre 7 y 12 dB, con resultados menos favorables también para el paciente F10 (ver Figura 5.5).

Las métricas adicionales también reflejaron un desempeño insatisfactorio. El error cuadrático medio de log f0 se mantuvo alrededor de 2.9 log Hz, indicando una predicción moderada de la frecuencia fundamental. La tasa de error SNS mostró una variabilidad considerable, con promedios entre 25 % y 35 %, evidenciando dificultades en la clasificación de segmentos sonorizados y no sonorizados. Además, las métricas perceptuales STOI y PESQ arrojaron valores consistentemente bajos, con STOI entre 0.3 y 0.4, y PESQ por debajo de 1.5, confirmando la insuficiencia en la calidad de la voz sintetizada para todos los pacientes evaluados.

Figura 5.1: Resultados de calidad de voz sintetizada para el paciente F01.

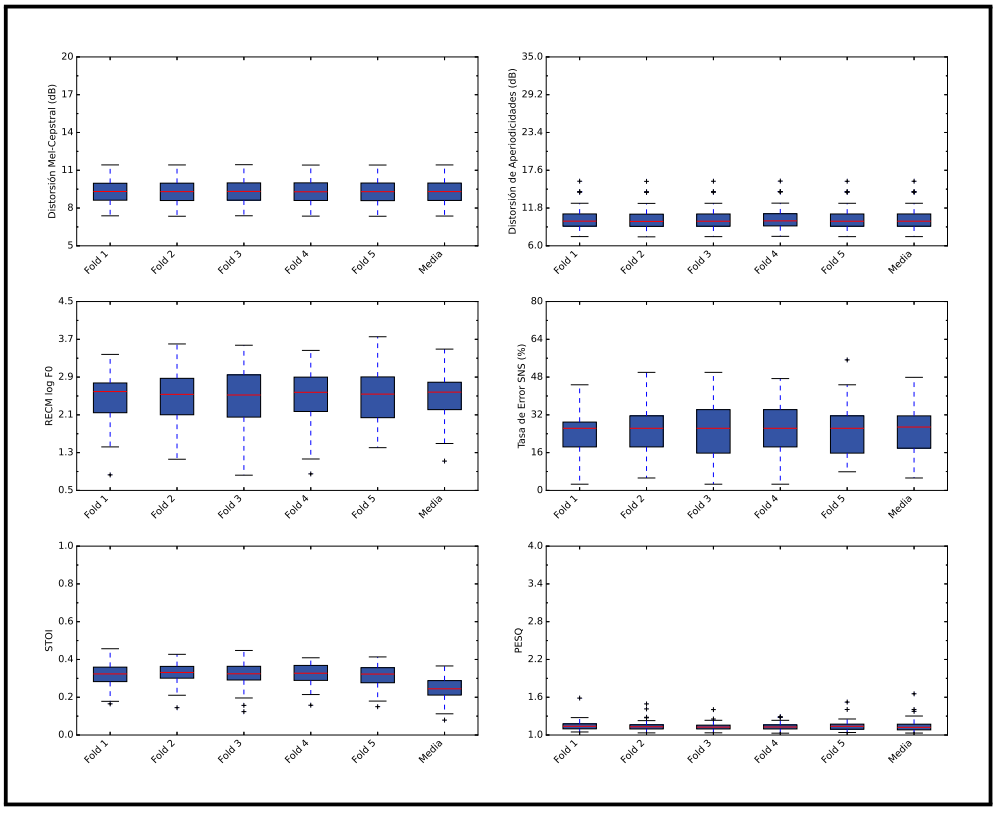


Figura 5.2: Resultados de calidad de voz sintetizada para el paciente F05.

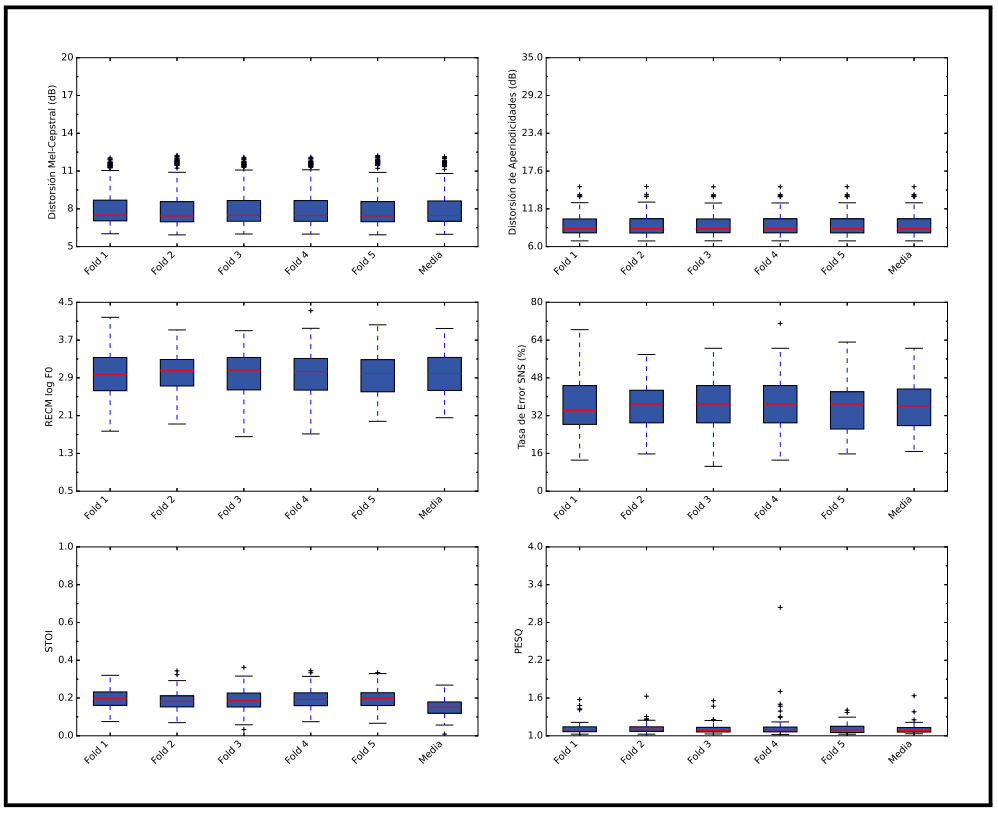


Figura 5.3: Resultados de calidad de voz sintetizada para el paciente F07.

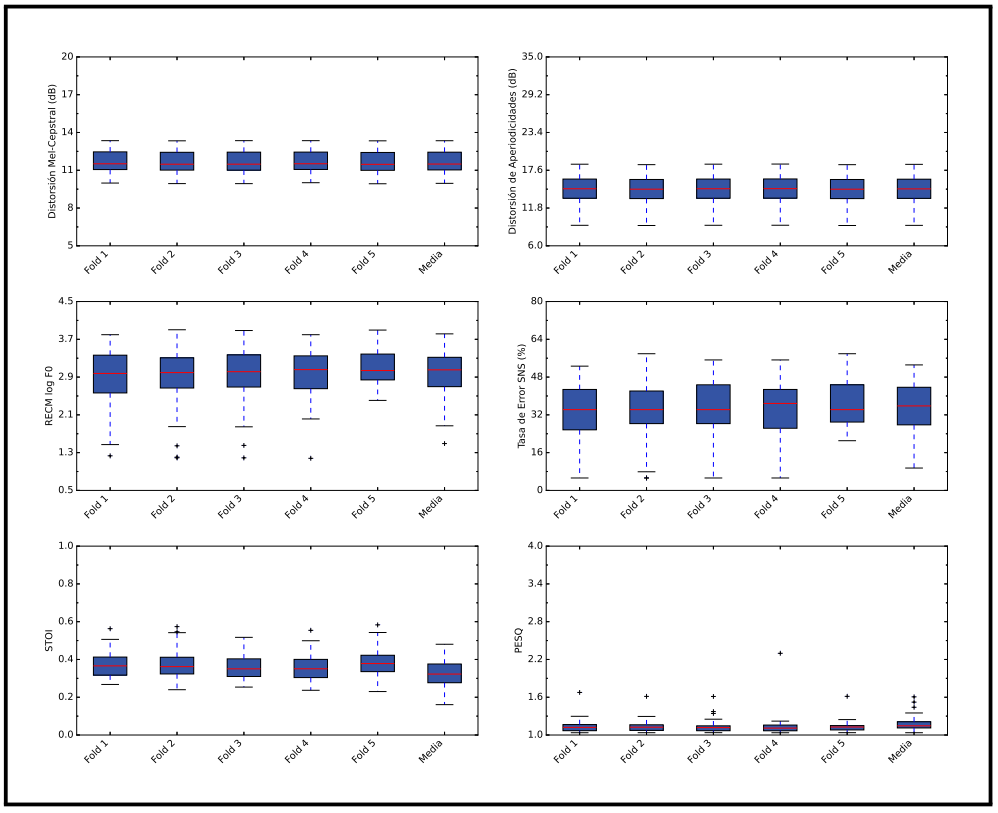


Figura 5.4: Resultados de calidad de voz sintetizada para el paciente F08.

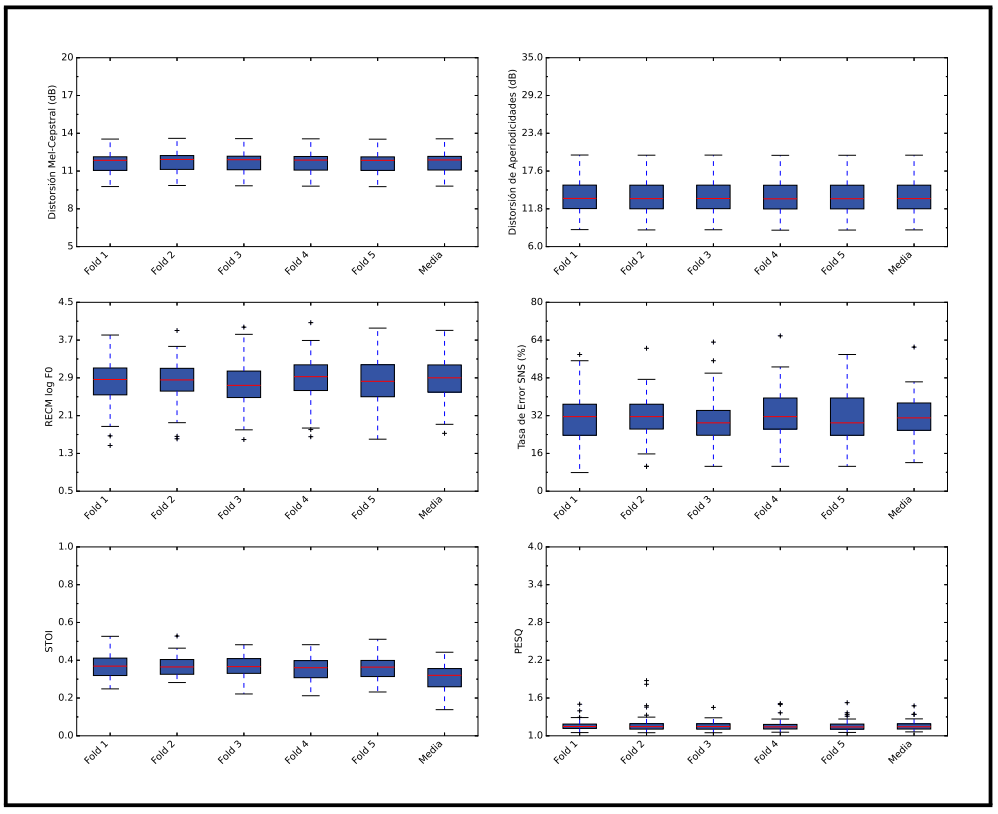


Figura 5.5: Resultados de calidad de voz sintetizada para el paciente F10.

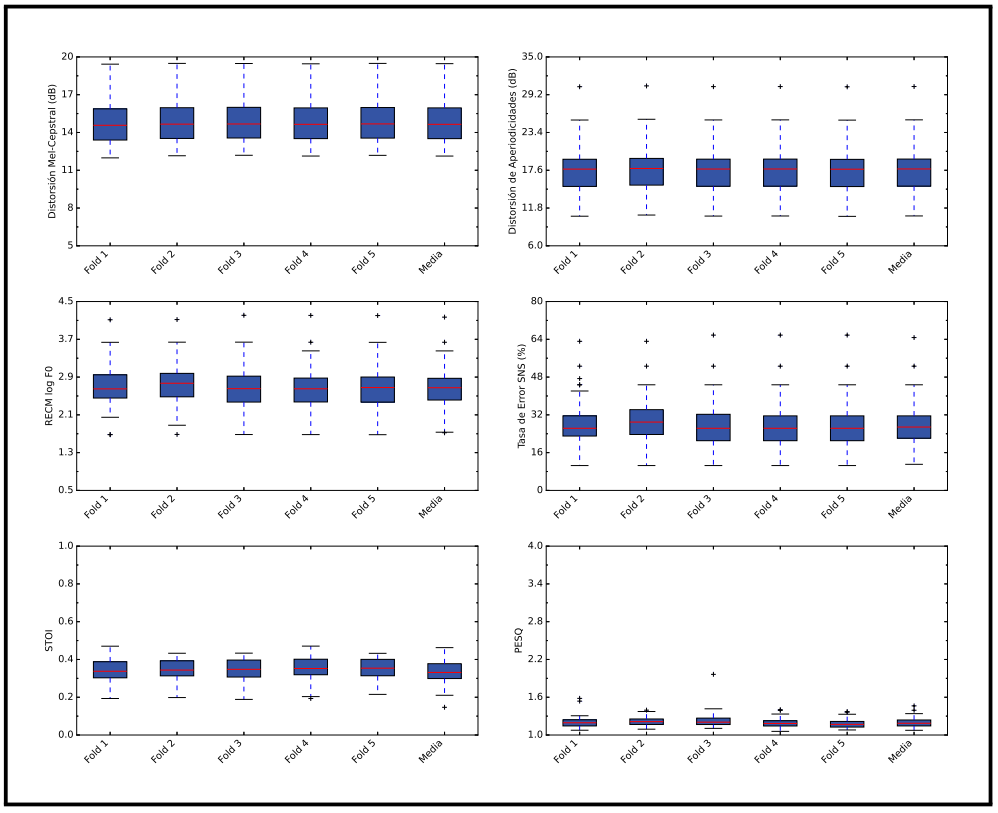
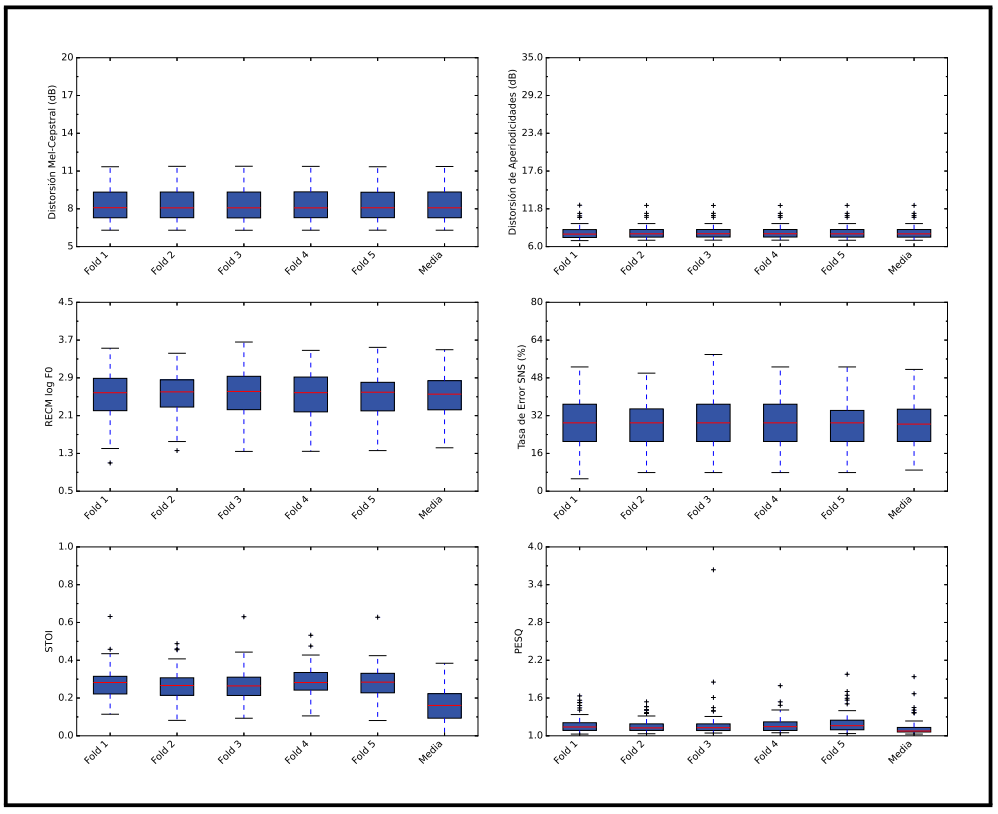


Figura 5.6: Resultados de calidad de voz sintetizada para el paciente M06.



5.2 Análisis de Correlación Canónica (ACC)

Para abordar el desempeño insuficiente en la síntesis de voz, se realizó un análisis de correlación canónica (ACC)[35] con el fin de investigar la relación entre las características del audio original y las características del iEEG registrado. Este análisis, descrito en el Apéndice B, se centró en la asociación entre los coeficientes cepstrales de frecuencia Mel (CCFM) del audio y las características gamma del iEEG, con el objetivo de evaluar la linealidad entre ambos conjuntos de datos para cada paciente individual a partir del primer coeficiente de correlación canónica.

Los resultados, presentados en las Figuras 5.7 a 5.12, mostraron una correlación moderada general (rango de 0.21 a 0.56), la cual se incrementó significativamente (superior a 0.9 en algunos casos) al analizar datos más específicos, como aquellos correspondientes a una misma vocal o pseudopalabra específica. Esta variabilidad en los coeficientes sugiere que ciertos patrones de iEEG se correlacionan mejor con las características del audio que otros, posiblemente debido a la complejidad fonética de las pseudopalabras o a la capacidad del paciente para reproducirlas de manera consistente. No obstante, el bajo rendimiento en la síntesis de voz sugiere una ausencia de una relación clara, ya sea lineal o no lineal, entre los datos, lo que podría explicarse por la colocación de los electrodos en áreas cerebrales no directamente relacionadas con la producción del habla y sus procesos asociados.

Figura 5.7: Análisis de correlación canónica para el paciente F01.

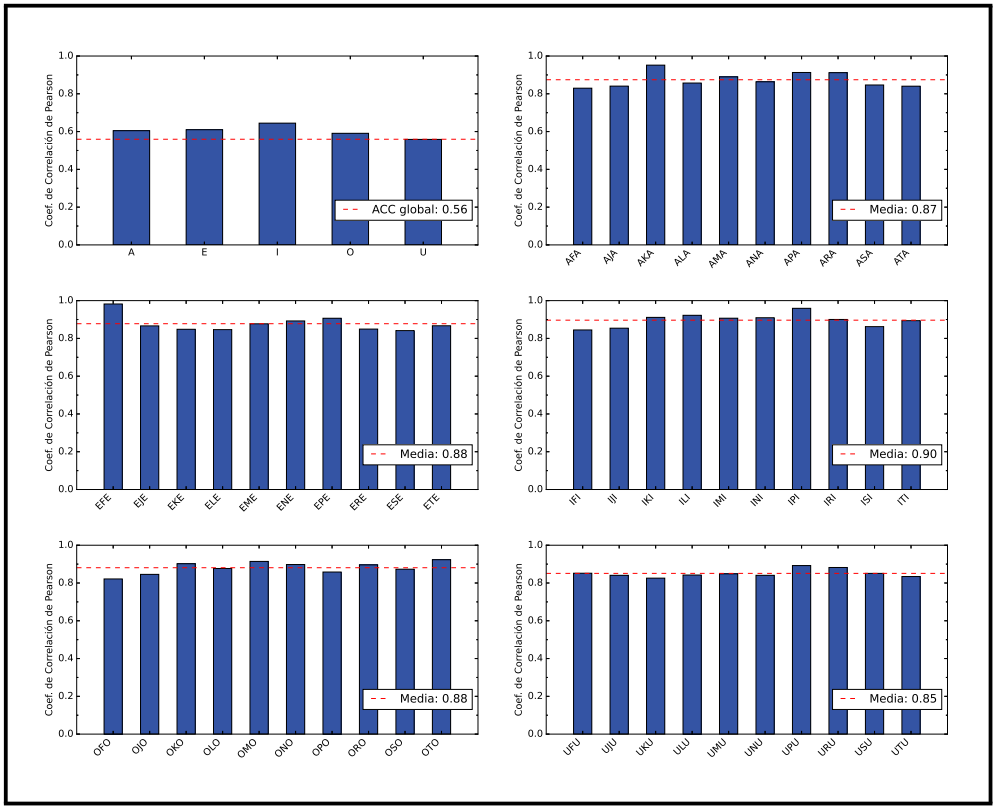


Figura 5.8: Análisis de correlación canónica para el paciente F05.

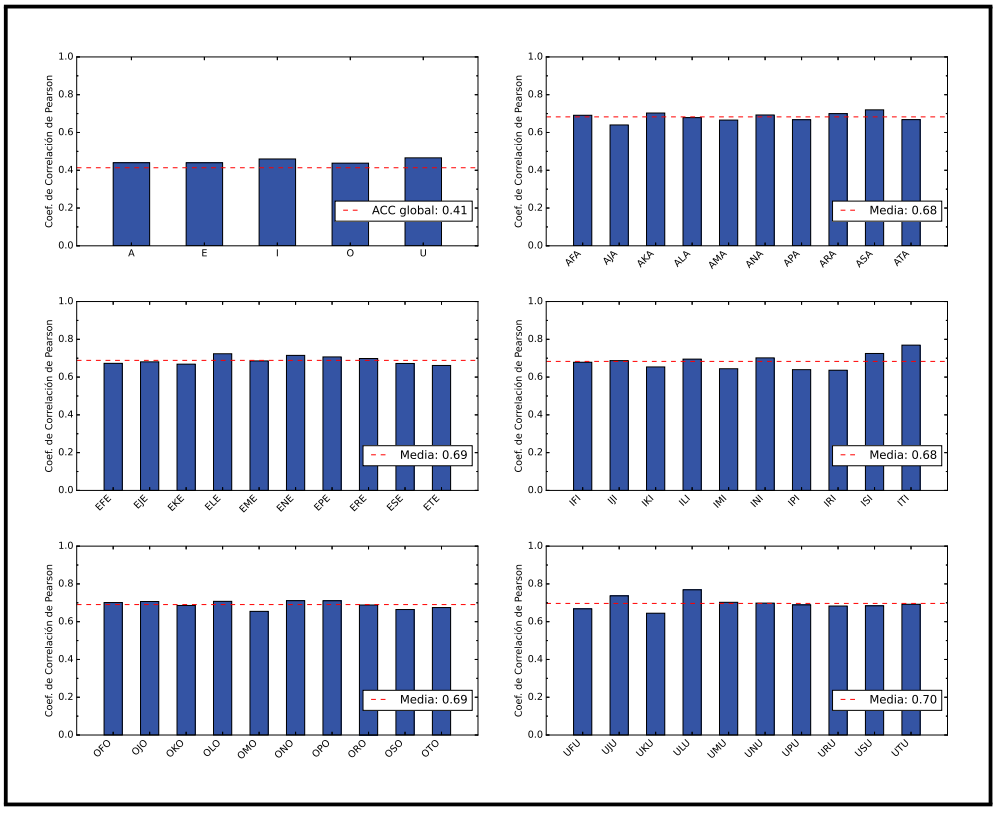


Figura 5.9: Análisis de correlación canónica para el paciente F07.

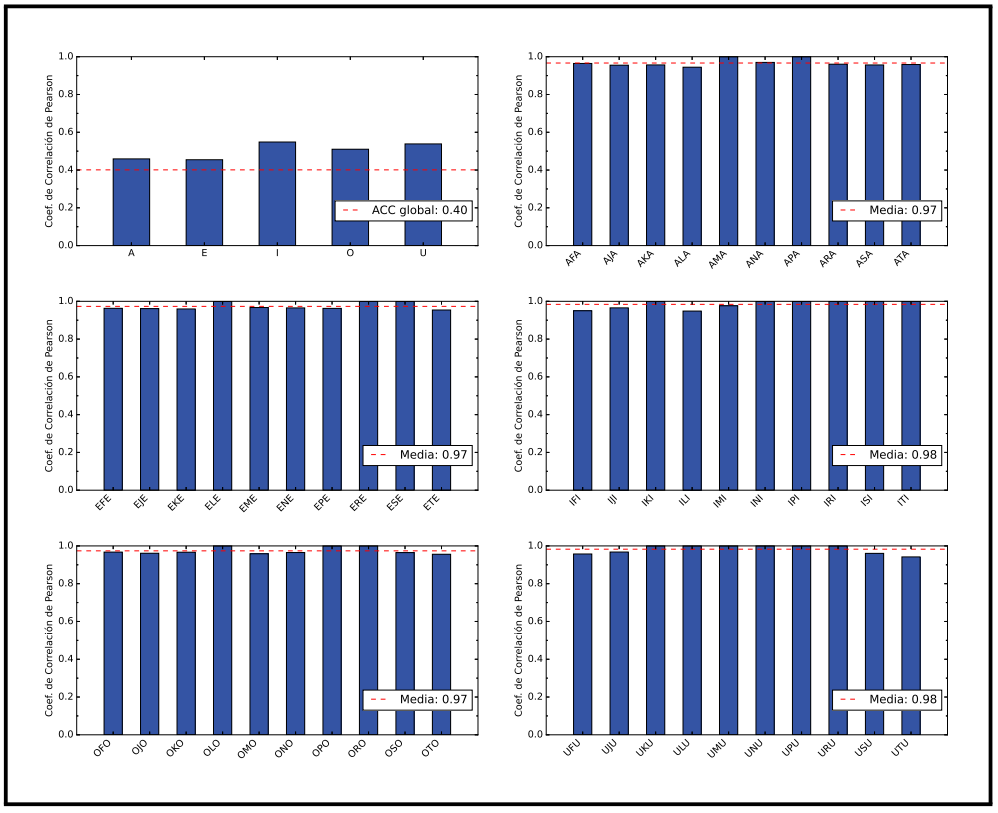


Figura 5.10: Análisis de correlación canónica para el paciente F08.

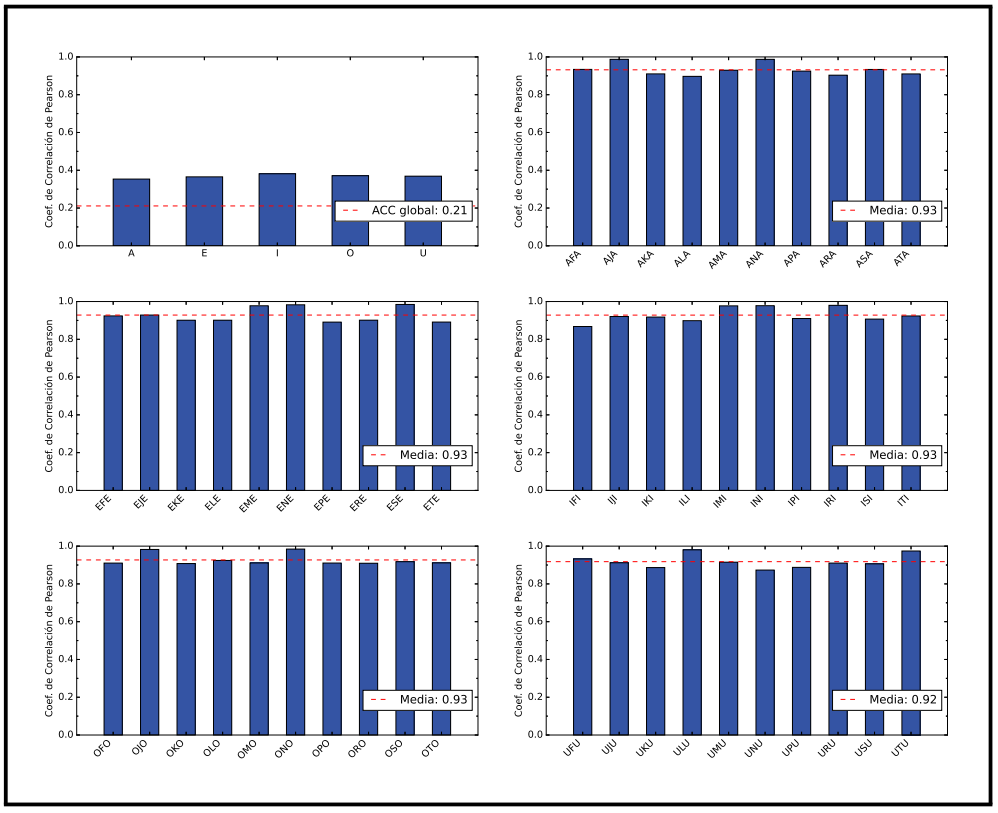


Figura 5.11: Análisis de correlación canónica para el paciente F10.

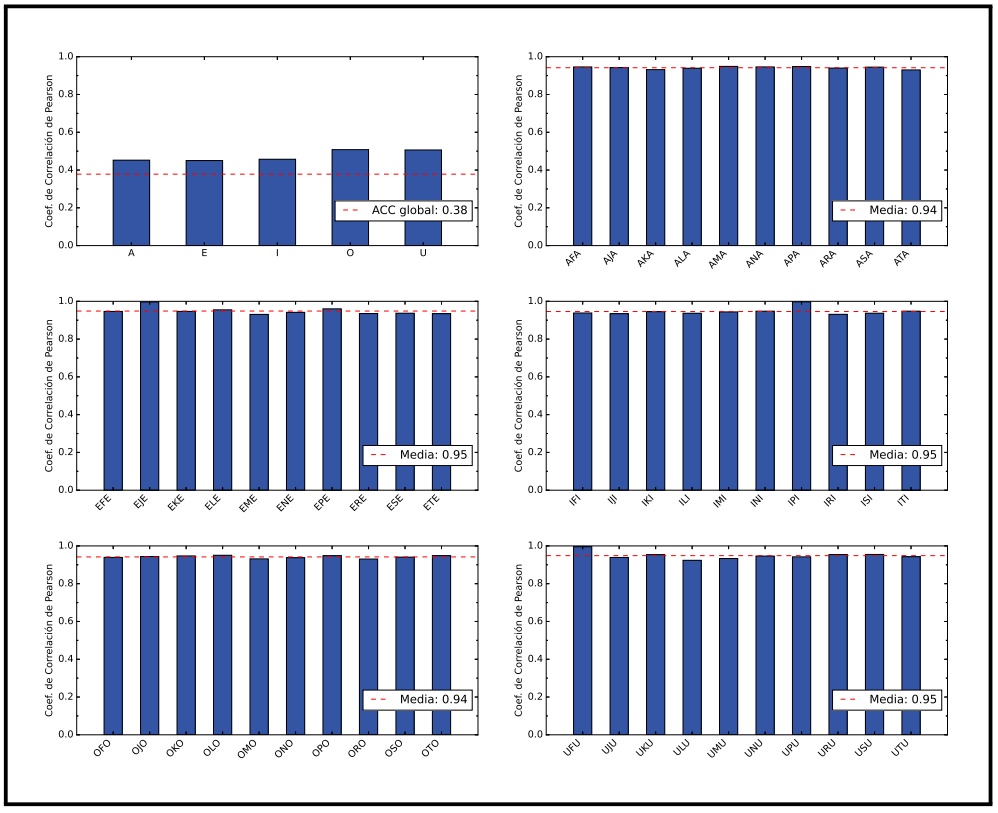
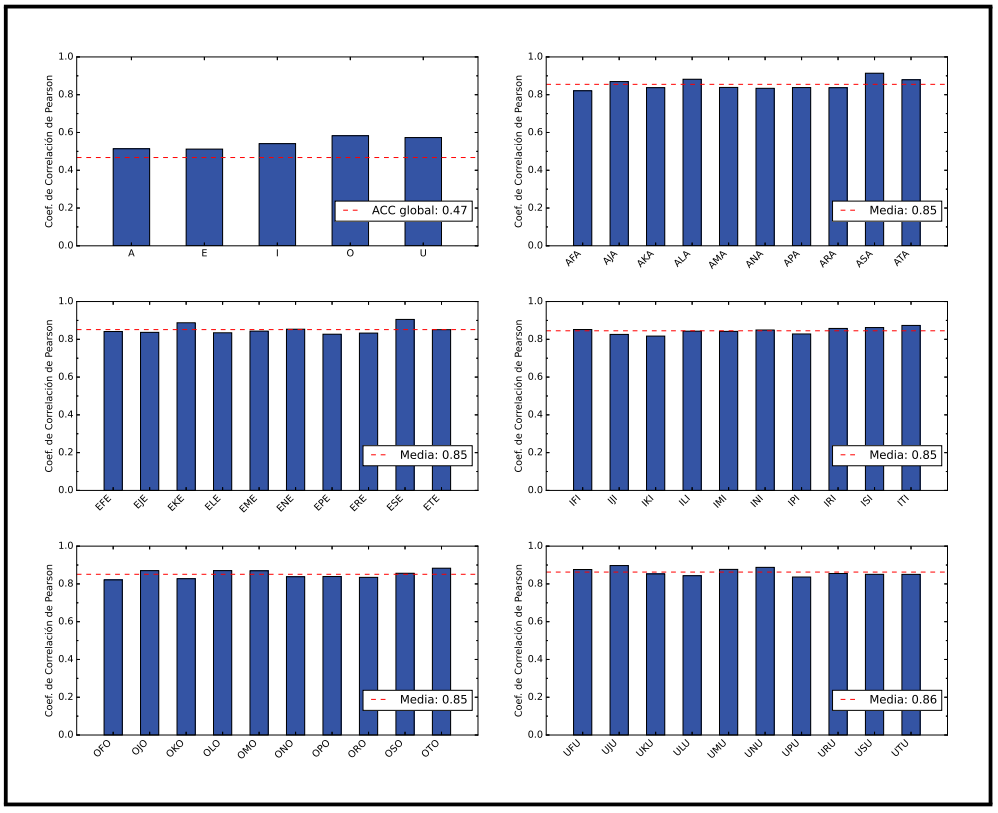


Figura 5.12: Análisis de correlación canónica para el paciente M06.



CAPÍTULO

6 Conclusiones

Los resultados obtenidos en esta tesis proporcionan una visión detallada sobre la viabilidad de la síntesis de voz a partir de métodos de electroencefalografía (EEG) invasivos, evidenciando los desafíos significativos que enfrenta este campo. A pesar de los esfuerzos realizados para mapear las características cerebrales registradas mediante iEEG a las características acústicas del habla, la calidad de la voz sintetizada no alcanzó niveles satisfactorios en ninguno de los pacientes analizados. Las mediciones, como la distorsión cepstral Mel y la distorsión de aperiodicidades vocales, revelaron discrepancias notables entre la voz generada y la referencia, con valores de distorsión cepstral Mel que oscilaron entre 8 y 15 dB, llegando incluso a alcanzar 17 dB en casos extremos. Las métricas perceptuales, como el índice de inteligibilidad del habla (STOI) y la evaluación de la calidad del habla (PESQ), reflejaron una inteligibilidad limitada y una baja calidad percibida en la voz sintetizada. Además, el análisis de correlación canónica (ACC) reveló una relación débil entre las características iEEG y las características de los coeficientes cepstrales de frecuencia Mel (CCFM) del audio, lo que sugiere que la predictibilidad de las características de audio a partir de iEEG podría no ser suficientemente robusta en casos, especialmente en análisis globales.

Estos hallazgos evidencian la complejidad y no linealidad inherentes a la relación entre las señales cerebrales y las características acústicas del habla, justificando el uso de modelos basados en redes neuronales para abordar este desafío. Sin embargo, la baja calidad de la voz sintetizada indica que el enfoque actual no es lo suficientemente robusto, debido a limitaciones en la captura de información relevante con los electrodos iEEG. Aunque los resultados no alcanzaron el nivel de calidad deseado, este estudio establece una base sólida para futuras investigaciones. Estas podrían explorar nuevas configuraciones de electrodos, técnicas avanzadas de captura de iEEG, o la incorporación de información adicional, como datos articulatorios o contextuales, para mejorar la calidad de la síntesis de voz. Tales enfoques podrían conducir a avances significativos en la síntesis de voz basada en la actividad cerebral, abriendo nuevas posibilidades en el campo de la comunicación asistida y las interfaces cerebro-computadora (ICCs).

Apéndice A

Análisis de Componentes Principales

En estadística, el Análisis de Componentes Principales (ACP) es una técnica utilizada para describir un conjunto de datos en términos de nuevas variables no correlacionadas. El objetivo principal del PCA es transformar un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas componentes principales. Estos componentes principales se ordenan de manera que el primer componente tiene la mayor varianza posible, y cada componente subsiguiente tiene la mayor varianza posible bajo la restricción de ser ortogonal a los componentes anteriores.

En su forma mas básica, tal y como describe Hotelling (1931) en su trabajo *Analysis of a Complex of Statistical Variables into Principal Components*, ACP busca encontrar la proyección $\mathbf{u}$ de la matriz de datos $\mathbf{X}$ que maximiza la varianza de sus puntos, sujeta a la restricción de que la norma al cuadrado de $\mathbf{u}$ sea igual a 1:

$$\arg\max_{\mathbf{u}} \text{Var}(\mathbf{u}^T\mathbf{X}) \quad s.t \quad \mathbf{u}^T\mathbf{u} = 1 \tag{F.1}$$

o, de manera equivalente:

$$\arg\max_{\mathbf{u}} \ \mathbf{u}^T\mathbf{S}\mathbf{u} \quad s.t \quad \mathbf{u}^T\mathbf{u} = 1 \tag{F.2}$$

donde $\mathbf{S}$ es la matriz de covarianzas de los datos de entrada $\mathbf{X}$.

Optimizando (F.2) por Lagrange:

$$\mathcal{L}(\mathbf{u}, \lambda) = \mathbf{u}^T\mathbf{S}\mathbf{u} - \lambda(\mathbf{u}^T\mathbf{u} - 1) \tag{F.3}$$

El vector de derivadas parciales de la función de Lagrange (F.3) es:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = 2\mathbf{S}\mathbf{u} - 2\lambda\mathbf{u} \tag{F.4}$$

que, igualando a cero, da lugar a:

$$\mathbf{S}\mathbf{u} = \lambda\mathbf{u} \tag{F.5}$$

En esta ecuación F.5, los vectores propios $\mathbf{u}$ de la matriz de covarianzas $\mathbf{S}$ establecen los ejes de proyección óptimos, o componentes principales, que maximizan la

varianza de los datos proyectados. Los autovalores λ asociados a estos autovectores representan la cantidad de varianza explicada por cada componente principal.

$$\begin{aligned} \arg\max_{\mathbf{u}} \ \mathbf{u}^T \mathbf{S}\mathbf{u} &= \arg\max_{\mathbf{u}} \ \mathbf{u}^T \lambda \mathbf{u} \\ &= \arg\max_{\mathbf{u}} \ \lambda \mathbf{u}^T \mathbf{u} \\ &= \arg\max_{\mathbf{u}} \ \lambda \end{aligned} \tag{F.6}$$

ACP utiliza esta información para seleccionar un subconjunto de componentes principales que conserven la mayor parte de la variabilidad de los datos originales. La cantidad de componentes seleccionados, k, depende del umbral de varianza explicada acumulado (η) que se desee conservar:

$$\eta \geq \frac{\sum_{i=0}^{k} \lambda_i}{\sum_{j=0}^{M} \lambda_j} \quad s.t \quad 0 \leq \eta \leq 1 \tag{F.7}$$

donde M indica el número total de componentes principales, equivalente al número de eigenvalores de la matriz de covarianzas $\mathbf{S}$, y k representa la cantidad de componentes principales elegidas para representar los datos.

Apéndice B

Análisis de Correlación Canónica

En estadística, el análisis de correlación canónica (ACC), también conocido como análisis de variables canónicas, es un método de análisis multivariante desarrollado por Harold Hotelling en 1936. A diferencia de otros enfoques estadísticos como la regresión lineal, que se enfocan principalmente en la predicción de una variable dependiente a partir de una o varias variables independientes, el ACC busca identificar y cuantificar la relación entre dos conjuntos de variables.

El objetivo es encontrar los vectores de pesos $\mathbf{w}_a$ y $\mathbf{w}_b$ que maximicen la correlación entre las combinaciones lineales $\mathbf{z}_a = \mathbf{X}_a\mathbf{w}_a$ y $\mathbf{z}_b = \mathbf{X}_b\mathbf{w}_b$ de las variables originales $\mathbf{X}_a$, $\mathbf{X}_b$. Aquí, $\mathbf{X}_a$ y $\mathbf{X}_b$ son matrices de datos con dimensiones $N \times P$ y $N \times M$, respectivamente. Estas combinaciones lineales, llamadas variables canónicas, representan las direcciones de mayor varianza conjunta entre ambos conjuntos.

Matemáticamente, esto se puede expresar como:

$$\arg\max_{\mathbf{w}_a, \mathbf{w}_b} \operatorname{corr}(\mathbf{z}_a, \mathbf{z}_b) \tag{F.8}$$

Utilizando el coeficiente de correlación de Pearson, esto es equivalente a:

$$\arg\max_{\mathbf{w}_a, \mathbf{w}_b} \frac{\mathbf{w}_a^T \boldsymbol{\Sigma}_{ab} \mathbf{w}_b}{\sqrt{\mathbf{w}_a^T \boldsymbol{\Sigma}_{aa} \mathbf{w}_a} \sqrt{\mathbf{w}_b^T \boldsymbol{\Sigma}_{bb} \mathbf{w}_b}} \tag{F.9}$$

Aquí, $\boldsymbol{\Sigma}_{aa} \in \mathbb{R}^{P \times P}$, $\boldsymbol{\Sigma}_{bb} \in \mathbb{R}^{M \times M}$ representan las matrices de covarianza de las variables originales $\mathbf{X}_a$ y $\mathbf{X}_b$, y $\boldsymbol{\Sigma}_{ab}$ la matriz de covarianza cruzada entre $\mathbf{X}_a$ y $\mathbf{X}_b$.

Sujeto a la condición de invarianza de escala $|\mathbf{z}_a|_2 = |\mathbf{z}_b|_2 = 1$ y asumiendo que las variables se encuentran normalizadas en media, se cumple $\mathbf{w}_i^T \boldsymbol{\Sigma}_{ij} \mathbf{w}_j = \mathbf{z}_i^T \mathbf{z}_j$.

Bajo esta condición, la función objetivo (F.9) se reduce a:

$$\arg\max_{\mathbf{w}_a, \mathbf{w}_b} \{\mathbf{z}_a^T \mathbf{z}_b\} \;\; s.t \;\; \|\mathbf{z}_a\|_2 = \|\mathbf{z}_b\|_2 = 1 \tag{F.10}$$

Optimizando (F.10) por Lagrange:

$$\mathcal{L}(\mathbf{w}_a, \mathbf{w}_b) = \mathbf{w}_a^T \boldsymbol{\Sigma}_{ab} \mathbf{w}_b - \lambda_1(\mathbf{w}_a^T \boldsymbol{\Sigma}_{aa} \mathbf{w}_a - 1) - \lambda_2(\mathbf{w}_b^T \boldsymbol{\Sigma}_{bb} \mathbf{w}_b - 1) \tag{F.11}$$

El vector de derivadas parciales de la función de Lagrange (F.11) es:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_a} = \boldsymbol{\Sigma}_{ab}\mathbf{w}_b - \lambda_1 \boldsymbol{\Sigma}_{aa}\mathbf{w}_a \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}_b} = \boldsymbol{\Sigma}_{ba}\mathbf{w}_a - \lambda_2 \boldsymbol{\Sigma}_{bb}\mathbf{w}_b \end{cases} \tag{F.12}$$

que, igualando a cero, da lugar a:

$$\begin{cases} \boldsymbol{\Sigma}_{ab}\mathbf{w}_b = \lambda_1 \boldsymbol{\Sigma}_{aa}\mathbf{w}_a \\ \boldsymbol{\Sigma}_{ba}\mathbf{w}_a - \lambda_2 \boldsymbol{\Sigma}_{bb}\mathbf{w}_b \end{cases} \tag{F.13}$$

lo que conduce a:

$$\underbrace{\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}}_{\mathbf{A}\in\mathbb{R}^{P\times P}}\mathbf{w}_a = \frac{1}{\lambda_1\lambda_2} I\mathbf{w}_a \tag{F.14}$$

$$\underbrace{\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}}_{\mathbf{B}\in\mathbb{R}^{Q\times Q}}\mathbf{w}_b = \frac{1}{\lambda_1\lambda_2} I\mathbf{w}_b \tag{F.15}$$

En esta ecuación, los eigenvectores $\mathbf{w}_a$ y $\mathbf{w}_b$ de las matrices $\mathbf{A}$ y $\mathbf{B}$, respectivamente, representan los ejes de proyección óptimos que maximizan la correlación entre las variables canónicas $\mathbf{z}_a$ y $\mathbf{z}_b$. Los valores propios asociados $\lambda_1^{-1}\lambda_2^{-1}$ indican la fuerza de las relaciones canónicas entre los conjuntos de variables $\mathbf{X}_a$ y $\mathbf{X}_b$.

Bibliografía

[1] Brumberg JS, Nieto-Castanon A, Kennedy PR, Guenther FH. Brain-Computer Interfaces for Speech Communication. Speech Commun. 2010 Apr 1;52(4):367-379. doi: 10.1016/j.specom.2010.01.001. PMID: 20204164; PMCID: PMC2829990.

[2] Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM. Brain-computer interfaces for communication and control. Clin Neurophysiol. 2002 Jun;113(6):767-91. doi: 10.1016/s1388-2457(02)00057-3. PMID: 12048038.

[3] Lotte F, Congedo M, Lécuyer A, Lamarche F, Arnaldi B. A review of classification algorithms for EEG-based brain-computer interfaces. J Neural Eng. 2007 Jun;4(2):R1-R13. doi: 10.1088/1741-2560/4/2/R01. Epub 2007 Jan 31. PMID: 17409472.

[4] Herff C, Heger D, de Pesters A, Telaar D, Brunner P, Schalk G, Schultz T. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front Neurosci. 2015 Jun 12;9:217. doi: 10.3389/fnins.2015.00217. PMID: 26124702; PMCID: PMC4464168.

[5] Mugler EM, Goldrick M, Slutzky MW. Cortical encoding of phonemic context during word production. Annu Int Conf IEEE Eng Med Biol Soc. 2014;2014:6790-3. doi: 10.1109/EMBC.2014.6945187. PMID: 25571555.

[6] Herff C, Schultz T. Automatic Speech Recognition from Neural Signals: A Focused Review. Front Neurosci. 2016 Sep 27;10:429. doi: 10.3389/fnins.2016.00429. PMID: 27729844; PMCID: PMC5037201.

[7] YATES AJ. Delayed auditory feedback. Psychol Bull. 1963 May;60:213-32. doi: 10.1037/h0044155. PMID: 14002534.

[8] Lebedev MA, Nicolelis MA. Brain-machine interfaces: past, present and future. Trends Neurosci. 2006 Sep;29(9):536-46. doi: 10.1016/j.tins.2006.07.004. Epub 2006 Jul 21. PMID: 16859758.

[9] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997; 9 (8): 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735

[10] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative Adversarial Nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, editors. Advances in Neural Information Processing Systems. Vol 27. Curran Associates, Inc.; 2014.

[11] Milner D. Cognitive neuroscience: the biology of the mind and findings and current opinion in cognitive neuroscience. Trends Cogn Sci. 1998 Nov 1;2(11):463. doi: 10.1016/s1364-6613(98)01226-1. PMID: 21227278.

[12] Buzsáki G, Draguhn A. Neuronal oscillations in cortical networks. Science. 2004 Jun 25;304(5679):1926-9. doi: 10.1126/science.1099745. PMID: 15218136.

[13] Engel AK, Fries P, Singer W. Dynamic predictions: oscillations and synchrony in top-down processing. Nat Rev Neurosci. 2001 Oct;2(10):704-16. doi: 10.1038/35094565. PMID: 11584308.

[14] Alvarado Cando OS, Vásquez Rodríguez GJ, Urgilés Cárdenas DF. Implementación de un sistema BCI para el análisis del comportamiento de bioseñales neurológicas [Bachelor's thesis]. Cuenca, Ecuador: Universidad del Azuay; 2017. Available from: `http://dspace.uazuay.edu.ec/handle/datos/7306`.

[15] Nagel S. Towards a home-use BCI: fast asynchronous control and robust non-control state detection. PhD thesis. 2019 Dec. doi: 10.15496/publikation-37739.

[16] Lago N, Cester A. Flexible and organic neural interfaces: a review. Appl Sci. 2017;7(12):1292. doi: 10.3390/app7121292.

[17] Farfán F. Control Cerebral de Interfases: Análisis Exploratorio de Técnicas Paramétricas Digitales para la Detección y Cuantificación de Estados Mentales. PhD thesis. 2005. doi: 10.13140/RG.2.1.1649.1123.

[18] Novo-Olivas C, Guitiérrez L, Bribiesca J. Mapeo Electroencefalográfico y Neurofeedback. In: [Book Title]. [City]: [Publisher]; 2010 Feb. p. 371-412. ISBN: 978-970-764-911-8.

[19] Ardila J, Suárez Mantilla L. Clasificación de registros fonocardiográficos usando descomposición empírica en modos y redes de gran memoria de corto plazo [dissertation]. 2019 Jun. doi: 10.13140/RG.2.2.24258.61125.

[20] Mosteller F, Tukey J. Data Analysis, including Statistics. In: Lindzey G, Aronson E, editors. Revised Handbook of Social Psychology. Addison Wesley; 1968. p. 80–203.

[21] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15(56):1929-1958.

[22] Prechelt L. Early Stopping — But When? In: Proceedings of the 25th International Conference on Machine Learning. Berlin, Heidelberg: Springer; 2012. p.53-67. doi: 10.1007/978-3-642-35289-8_5.

[23] González-Muñiz A. Aplicación de técnicas de aprendizaje profundo (deep learning) al análisis y mejora de la eficiencia en sistemas de ingeniería [dissertation]. 2023 Feb. doi: 10.13140/RG.2.2.29722.82888.

[24] Peirce JW. PsychoPy–Psychophysics software in Python. J Neurosci Methods. 2007 May 15;162(1-2):8-13. doi: 10.1016/j.jneumeth.2006.11.017. Epub 2007 Jan 23. PMID: 17254636; PMCID: PMC2018741.

[25] International Phonetic Association. The International Phonetic Alphabet (Revised to 2020). 2020. Available from: `https://www.internationalphoneticassociation.org/IPAcharts/IPA_chart_orig/pdfs/IPA_Kiel_2020_full.pdf`.

[26] Kothe C, Shirazi SY, Stenner T, Medine D, Boulay C, Crivich M, Mullen T, Delorme A, Makeig S. The Lab Streaming Layer for Synchronized Multimodal Recording. bioRxiv [preprint]. 2024. doi:10.1101/2024.02.13.580071.

[27] Verwoert M, Ottenhoff MC, Goulis S, et al. Dataset of Speech Production in intracranial.Electroencephalography. Sci Data. 2022;9(1):434. Published 2022 Jul 22. doi:10.1038/s41597-022-01542-9.

[28] Kaplanoglu E. Decoding Brain's Electrical Activity: Leveraging Hilbert Transforming Techniques for EEG Analysis. COJ Electronics & Communications. 2024 Apr;3:10.31031/COJEC.2024.04.000552.

[29] Morise M, Yokomori F, Ozawa K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Trans Info Syst. 2016;E99.D(7):1877-1884. doi:10.1587/transinf.2015EDP7457.

[30] Morise M. D4C, a band-aperiodicity estimator for high-quality speech synthesis. Speech Communication. 2016;84:57-65. doi: 10.1016/j.specom.2016.09.001.

[31] Toda T, Black AW, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans Audio Speech Lang Process. 2007;15(8):2222-2235. doi:10.1109/TASL.2007.907344.

[32] Moon S, Kim S, Choi YH. MIST-Tacotron: end-to-end emotional speech synthesis using mel-spectrogram image style transfer. IEEE Access. 2022;10:25455–63. doi: 10.1109/ACCESS.2022.3156093.

[33] Yu D, Kolbæk M, Tan ZH, Jensen J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2017 Mar 5-9; New Orleans, LA, USA. New York, NY: IEEE Press; 2017. p. 241–5. doi: 10.1109/ICASSP.2017.7952154.

[34] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221). Salt Lake City, UT, USA; 2001. p. 749-752 vol.2. doi: 10.1109/ICASSP.2001.941023.

[35] Gundersen G. Canonical Correlation Analysis in Detail. Available from: `https://gregorygundersen.com/blog/2018/07/17/cca/`. Published July 17, 2018.