UNIVERSIDAD DE GRANADA

TRABAJO FIN DE GRADO

INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

Interfaces cerebro-ordenador basadas en EEG para la síntesis de voz

Autor
Miguel Díaz Martín

Director
José A. González López

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Septiembre de 2023

Interfaces cerebro-ordenador basadas en EEG para la síntesis de voz

Miguel Díaz Martín

Palabras clave: ELA, actividad cerebral, EEG, red neuronal, SincNet, MFCC, síntesis, vocoder, señal de voz predicha, PMA

Resumen

Las lesiones que afectan al cerebro, o daños producidos por enfermedades como la esclerosis lateral amiotrófica (ELA) pueden producir deterioros en el cerebro irreversibles que pueden provocar la pérdida total de la capacidad de hablar.

En este trabajo se ha desarrollado un algoritmo capaz de generar voz a partir de registros de la actividad cerebral de dos pacientes implantados con electrodos profundos invasivos (es decir electroencefalografía intracraneal). En concreto se ha adaptado y entrenado un modelo de red neuronal del estado del arte denominado SincNet que, al contrario que otros tipos de redes neuronales artificiales, permite trabajar directamente con las señales de EEG en bruto sin la necesidad de pasar por una etapa intermedia de extracción de parámetros.

Para realizar la síntesis de voz, la red neuronal diseñada obtiene los coeficientes MFCC predichos a partir de los datos de EEG en crudo, es decir sin realizar un procesamiento previo, y haciendo uso del vocoder WORLD, se realiza la síntesis de la señal de voz predicha.

Para la evaluación del algoritmo de síntesis de voz propuesto, se ha evaluado la calidad e inteligibilidad de la voz obtenida usando tanto métricas objetivas y subjetivas. Concretamente se han comparado de forma objetiva los coeficientes MFCC originales y predichos mediante la métrica MCD y las señales de voz originales y predichas mediante la métrica STOI. Además se han realizado escuchas subjetivas para cotejar los resultados de las métricas objetivas. Con el fin de contrastar los resultados obtenidos se ha comparado la red neuronal diseñada con una red DNN simple y se han utilizado datos procedentes de Articulografía por Imanes Permanentes o PMA para comprobar el funcionamiento y la calidad de síntesis de la red SincNet propuesta respecto a una red DNN simple.

Para el caso de procesado de las señales de EEG en crudo no se obtiene voz inteligible debido a la escasez de los datos disponibles y a la limitación de extracción de los mismos, la cual depende de criterios médicos. Para este caso se han obtenido valores de MCD entre 21,55 dB y 24,65 dB y valores de STOI de 0,001 para el caso de procesado mediante la red SincNet propuesta. Por otro lado la red neuronal diseñada presenta resultados positivos

realizando la decodificación del habla de manera más precisa a partir de las señales de PMA en comparación a una red DNN estándar. Se han obtenido unos resultados medios de MCD de 10,61 dB frente a 11,26 dB, además de resultados favorables mediante escuchas subjetivas de los resultados a pesar de presentar unos peores resultados en la métrica STOI teniendo una puntuación de 0,54 frente a 0,57 obtenido con la red neuronal simple. El modelo ha demostrado ser funcional con un tipo de bioseñales, aunque ha demostrado la necesidad de realizar el entrenamiento de la red neuronal con un mayor volumen de datos para realizar un ajuste más preciso en el caso de EEG, o utilizar unas técnicas de extracción de señales más precisas.

EEG-based brain-computer interfaces for speech synthesis

Miguel Díaz Martín

Keywords: ALS, brain activity, EEG, neural network, SincNet, MFCC, synthesis, vocoder, predicted speech signal, PMA

Abstract

Injuries that affect the brain, or damage caused by diseases such as amyotrophic lateral sclerosis (ALS), can cause irreversible damage to the brain that can cause the total loss of the ability to speak.

In this work, an algorithm has been developed capable of generating speech from recordings of the brain activity of two patients implanted with invasive deep electrodes (that is, intracranial electroencephalography). Specifically, a state-of-the-art neural network model called SincNet has been designed and trained which, unlike types of artificial neural networks, allows working directly with the raw EEG signals without the need to go through an intermediate stage of parameter extraction.

To perform voice synthesis, the designed neural network obtains the predicted MFCC coefficients from the raw EEG data, that is, without performing prior processing, and using the WORLD vocoder, the synthesis of the predicted voice signal is performed.

For the evaluation of the proposed speech synthesis algorithm, the quality and intelligibility of the obtained voice has been evaluated using both objective and subjective metrics. Specifically, the original and predicted MFCC coefficients have been objectively compared using the MCD metric and the original and predicted voice signals using the STOI metric. In addition, subjective listening has been carried out to compare the results of the objective metrics. In order to contrast the results obtained, the designed neural network has been compared with a simple DNN network and data from Permanent Magnet Articulography or PMA has been used to verify the operation and synthesis quality of the proposed SincNet network with respect to a simple DNN network.

In the case of processing raw EEG signals, intelligible voice is not obtained due to the scarcity of available data and the limitation of data extraction, which is limited by medical criteria. For this case, MCD values between 21.55 dB and 24.65 dB and STOI values of 0.001 have been obtained for the case of processing through the proposed SincNet network. On the other hand, the designed neural network presents positive results by decoding speech more accurately from PMA signals compared to a standard DNN network. Average MCD results of 10.61 dB compared to 11.26 dB have been obtained, in addition to favorable results through subjective listening

of the results despite presenting worse results in the STOI metric, having a score of 0.54 compared to 0.57 obtained with the simple neural network. The model has proven to be functional with a type of biosignals, although it has demonstrated the need to train the neural network with a larger volume of data to perform a more precise adjustment, or use more precise signal extraction techniques.

Yo, **Miguel Díaz Martín**, alumno de la titulación Grado en Ingeniería de Tecnologías de Telecomunicación de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 77945195N, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Miguel Díaz Martín

Granada a 03 de Septiembre de 2023.

D. **José A. González López**, Profesor del Área de Teoría de la Señal y Comunicaciones del Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada.

Informa:

Que el presente trabajo, titulado ***Interfaces cerebro-ordenador basadas en EEG para la síntesis de voz***, ha sido realizado bajo su supervisión por **Miguel Díaz Martín**, y autorizo la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expide y firma el presente informe en Granada a 03 de Septiembre de 2023.

El director:

José A. González López

Agradecimientos

En primer lugar me gustaría agradecer a todos mis seres queridos el apoyo que me han proporcionado, no solo durante el desarrollo de este trabajo sino durante toda mi etapa universitaria.

A todos mis amigos, por todos los momentos compartidos conmigo. Agradeceros que, a pesar de haber sido un arduo camino, lo habéis hecho más ameno. Gracias tanto por los buenos ratos que hemos pasado juntos como por acompañarme en aquellos más complicados.

Gracias a José Andrés por darme la oportunidad de llevar a cabo este proyecto, por toda su dedicación y atención puesta en mí.

En especial quiero agradecer a mis padres toda la confianza depositada en mí, todo el esfuerzo que han realizado y el apoyo incondicional que me han proporcionado desde el primer momento. Gracias por hacerme crecer como persona y por todo el cariño que me habéis dado siempre. Gracias también a mi hermano, Diego, por alegrarme todos los días de mi vida. Sabes que llegarás tan lejos como te propongas.

Gracias a todos vosotros me he dado cuenta de que, realmente, en la vida todo esfuerzo y sacrificio tiene su recompensa.

Índice general

Siglas

ADN Ácido desoxirribonucleico. 33

ANN Artificial Neural Network. 19, 44, 45, 47

BCE Binary Cross Entropy. 50, 51

CAA Comunicación Aumentativa y Alternativa. 29

CNN Convolutional Neural Network. 15, 20, 30, 47, 48, 61–63

DCT Discrete Cosine Transform. 38, 39

DDM Drift-Diffusion Model. 65

DNN Deep Neural Network. 5, 6, 20, 21, 47, 59, 60, 72–86

ECoG Electrocorticografía. 15, 19, 36, 37, 40, 41

EEG Electroencefalografía. 5, 6, 15, 16, 19–21, 23, 29, 30, 34–37, 40, 57–61, 63–66, 69–71, 73–77, 79–86, 89, 90

ELA Esclerosis Lateral Amiotrófica. 5, 19, 25–29, 87

FC Fully Connected. 64–66

ICC Interfaz Cerebro-Computadora. 19, 28, 29, 41

LFP Local Field Potential. 35, 36

LPC Linear Predictive Coding. 38

MAE Mean Absolute Error. 50

MCD Mel Cepstral Distorsion. 5, 6, 15, 23, 53, 76, 77, 79, 80, 83–86

MEG Magnetoencefalografía. 19, 35–37

MFCC Mel Frequency Cepstral Coefficients. 5, 15, 20, 38, 39, 53, 55, 58–60, 64–68, 72–74, 76, 77, 79

MSE Mean Squared Error. 43, 46, 50, 76

PCA Principal Component Analysis. 16, 59, 60, 73, 76

PESQ Perceptual Evaluation of Speech Quality. 15, 19, 54, 76

PMA Permanent Magnet Articulography. 5, 6, 16, 20, 23, 59, 67, 69–78, 81, 83–86, 89, 90

ReLU Rectified Lineal Unit. 48, 50

SNC Sistema Nervioso Central. 32

SNP Sistema Nervioso Periférico. 32, 34

STOI Short-Time Objective Intelligibility. 5, 6, 15, 19, 23, 53, 54, 76–80, 83–85

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto previo

Las interacciones sociales son la base de nuestro crecimiento como personas y como especie. Una comunicación en el ámbito de la interacción social entre humanos está formada por un emisor o persona que quiere comunicar, un receptor dispuesto a escuchar y un mensaje que contiene la información que se quiere transmitir. El problema aparece cuando la persona que quiere realizar una interacción con su entorno no puede expresar su mensaje.

Este es el caso de todas las personas que sufren afecciones cerebrales que impiden desarrollar las acciones motoras relacionadas con el proceso de producción del habla. Personas con parálisis cerebral, lesiones cerebrales derivadas de un accidente u otras causas, lesiones medulares o personas con enfermedades cerebrales entran dentro del colectivo en cuestión. Todas estas afecciones tienen la capacidad de impedir las comunicaciones de los canales neuromusculares. Se debe hacer especial mención a la ELA, la cual posee una enorme capacidad neurodegenerativa. Esta enfermedad provoca la muerte paulatina de las células de Betz en la corteza motora primaria y de las neuronas motoras ubicadas en la médula espinal y el tronco encefálico [1].

Consecuentemente esto provoca que las personas que la padecen en primer lugar sientan espasmos musculares espontáneos y rigidez muscular. Posteriormente se produce un atrofiamiento gradual de los músculos (inicialmente los pertenecientes a las extremidades), provocando un incremento de la debilidad muscular y perdiendo poco a poco movilidad y capacidad del habla entre otras cosas. En los peores casos, (aproximadamente 1/3 de las personas que sufren la ELA) es capaz de provocar un estado de parálisis completa del paciente denominado síndrome de enclaustramiento [2]. En este estado el paciente conserva plenamente sus capacidades cognitivas, además de controlar los esfínteres, el movimiento vertical de los ojos o el parpadeo (en ocasiones se pierden estas capacidades). A pesar de ello el paciente experimenta una

parálisis motora total [3].

Esto provoca una carga tanto económica como social para todos los familiares y la sociedad, puesto que las personas que padecen síndrome de enclaustramiento derivado de la ELA se convierten en personas totalmente dependientes, necesitando terapias y cuidados permanentes hasta el fallecimiento del paciente puesto que es una enfermedad crónica sin cura actualmente. Teniendo en cuenta que la vida media de una persona que padece ELA es de entre 17,7 meses a 91,0 meses a partir del diagnóstico según predictores analizados [4], y que en España el coste medio del mantenimiento clínico de un paciente es de aproximadamente 50000 euros anuales [5] además del coste derivado de otro tipo de cuidados, se puede afirmar que supone un reto para las familias, (y para la sociedad) el mantenimiento de un ser querido que padece dicha enfermedad.

Estudios afirman que, a pesar de depender de la geografía, la incidencia global de pacientes con ELA ha subido desde los últimos años, teniendo una cifra de incidencia global aproximada de 2,39 casos por cada 100.000 habitantes para estudios prospectivos y de 1,52 casos por cada 100.000 habitantes para casos retrospectivos [6]. En la figura 1.2 se puede observar una tendencia alcista en los casos de ELA (aunque los datos varían según el país).

En España existen muy pocos estudios epidemiológicos, por lo que no se pueden arrojar con certeza datos actuales, aunque se estima una incidencia de 1,4 y una prevalencia de 5,4 por cada 100.000 habitantes [5]. En la figura 1.1 se puede observar un mapa donde se muestran los casos de ELA en España por comunidades autónomas. El hecho de que el factor genético tenga cierta importancia a la hora de desarrollar la ELA puede explicar la variación de pacientes entre comunidades.

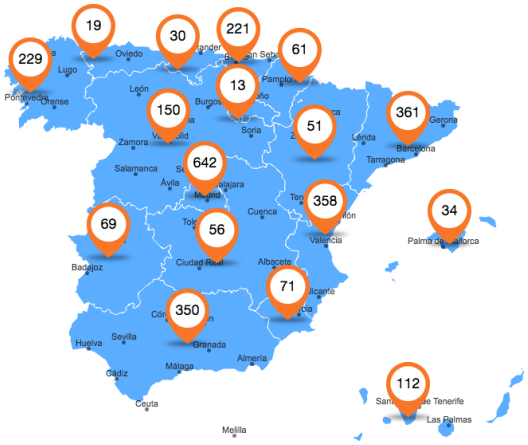

Figura 1.1: Pacientes con ELA en España (por comunidades) en el año 2016.

Existen otro tipo de enfermedades que pueden afectar al habla de la persona. Una de las más comunes es el ictus, el cual se produce debido al estrechamiento u obstrucción de los vasos sanguíneos encargados de regar el cerebro y proporcionar oxígeno al mismo. Existen dos tipos de ictus: el producido debido a la obstrucción total de dichos vasos sanguíneos (denominado ictus isquémico) y el producido debido a la ruptura de los vasos sanguíneos en cuestión. Estos pueden conllevar diferentes complicaciones médicas en el individuo que lo sufre, pero una de las posibles consecuencias es la falta de riego sanguíneo a partes concretas del cerebro relacionadas con el proceso de producción del habla, produciendo la muerte de las células cerebrales de dicha área en concreto y provocando una incapacidad de hablar a las personas que lo sufren.

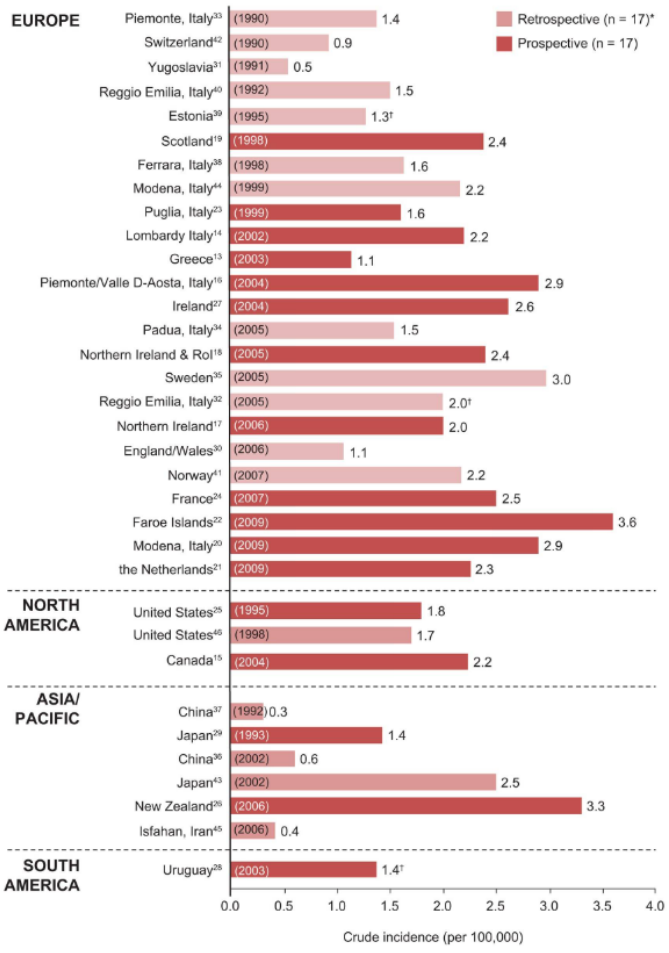

Figura 1.2: Incidencia de ELA según el país.

1.2. Avances tecnológicos y mejora de la calidad de vida

Actualmente existen numerosas terapias y tecnologías que permiten mejorar la calidad de vida de las personas que sufren problemas relacionados con la producción del habla. Debido a que la mayoría de ellas quedan totalmente aisladas socialmente deben de buscar alternativas para comunicarse con el resto de personas puesto que la capacidad del habla se pierde con el tiempo (en el caso de la ELA), o permanentemente (en el caso de enfermedades cerebrovasculares como por ejemplo un ictus). Con ello únicamente pueden expresar decisiones binarias, por ejemplo mediante el parpadeo o el movimiento de los ojos. Es por ello que muchos de estos pacientes recurren a técnicas basadas en interfaces cerebro-computadora o ICC para poder comunicarse con las personas. En esencia una ICC es un dispositivo capaz de captar señales cerebrales para controlar dispositivos externos y obtener una respuesta deseada, en este caso, poder producir voz a partir de dichas señales cerebrales [7].

Algunas de los sistemas basados en ICC para restablecer la comunicación de estas personas se comentan en [8]; en la figura 1.3 se puede observar un sistema ICC basado en los potenciales corticales lentos. El sistema es capaz de permitir la selección de letras para formar palabras a partir de la lecturas de la actividad cortical, de forma que a través de una pantalla se puede visualizar las palabras formadas.

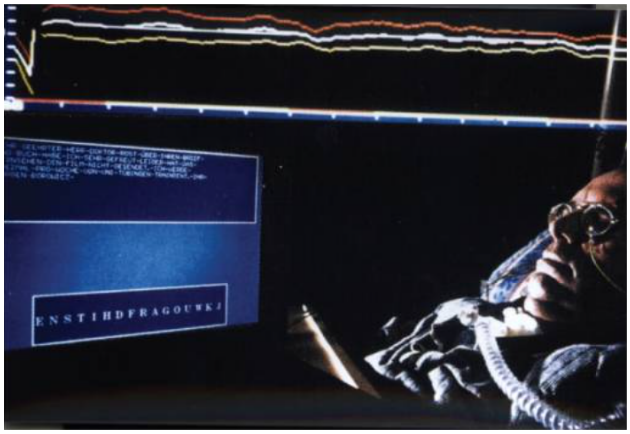

Figura 1.3: ICC basada en potenciales corticales lentos [8].

Otro ejemplo de ICC comentado en [8] es la ICC P300. Ésta se basa en la lectura de los potenciales P300 cerebrales para la selección de letras contenidas en una matriz visible a través de una pantalla. Esto potenciales cerebrales P300 se generan ante la aparición de estímulos de naturaleza tanto visual como acústica, de forma que en caso de tener un estímulo posi-

tivo, el potencial P300 manifiesta un pico de amplitud de voltaje de latencia corta (entorno a los 300 ms), de modo que permite una obtención de palabras rápida (en el caso de los potenciales P300 auditivos se tienen peores eficacias).

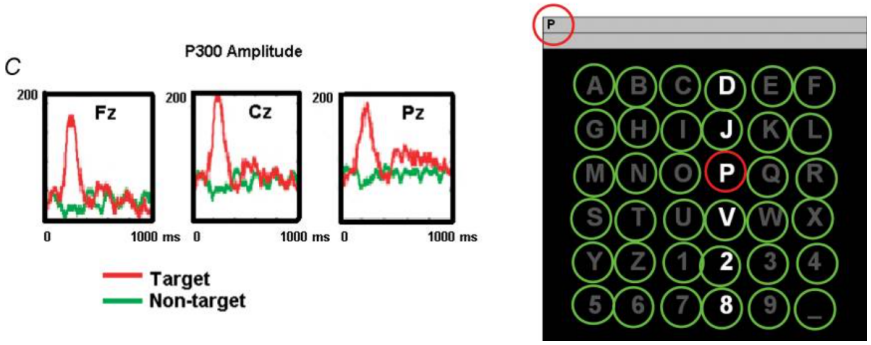

Figura 1.4: ICC basada en potenciales P300 [8].

Por último cabe comentar el beneficio que introduce el uso de CAA en la población que padece de problemas para producir el habla. Estos dispositivos electrónicos establecen un canal de comunicación alternativo dependiendo de las necesidades de cada persona. Estos sistemas generalmente son capaces de permitir al usuario seleccionar iconos para que el dispositivo diga las palabras en voz alta, permiten escribir texto para que el dispositivo lo diga, o permite realizar dibujos para mostrar a otras personas, entre otros muchos usos.

Estos son solo algunas de las facilidades con las que actualmente cuentan estas personas y que permiten facilitar en cierto grado la comunicación con el resto de la sociedad, rompiendo la barrera de la comunicación impuesta por las diferentes enfermedades comentadas.

1.3. Objetivos

El **objetivo** de este proyecto es el **diseño de un dispositivo de comunicación alternativa que permita decodificar el habla de una persona a partir de registros de su actividad cerebral** (capturada usando electroencefalografía, EEG) realizados mientras realiza tareas de producción de lenguaje.

Por último cabe comentar las posibles contribuciones que se esperan realizar a la sociedad con la realización de este proyecto. Este proyecto supone un avance en la investigación en el marco de la calidad de vida de los pacientes con ELA y síndrome de enclaustramiento, posibilitando una mejora sustancial en la forma que tienen de comunicarse y brindando la oportunidad de evitar ese aislamiento social al que quedan sometidos. Existen numerosos

campos de investigación abiertos en cuanto a la interpretación y decodificación de las señales cerebrales relacionadas con el habla, por lo que este proyecto pretende formar parte del conjunto de estudios que pretenden dar una mejor vida a un colectivo de personas que han perdido la principal forma de comunicación con otros seres humanos y con su entorno. Actividades cotidianas que implican interactuar con la sociedad pueden volver a estar disponibles para estas personas.

Para ello se tienen dos objetivos específicos:

- Diseño de un modelo de conversión de datos de EEG a voz usando los últimos avances en redes neuronales profundas. Para ello se diseñará y entrenará una red neuronal que permita diseñar un modelo de conversión de extremo a extremo, trabajando con la señal de EEG en crudo.
- Diseño de un sistema de evaluación mediante diferentes métricas tanto objetivas como subjetivas para la valoración de la calidad de las predicciones obtenidas por el modelo diseñado.
- Adaptación de una arquitectura de red neuronal CNN del estado del arte, concretamente la adaptación de la red SincNet para la tarea de decodificación del habla a partir de señales EEG

1.4. Estructura de la memoria

Este informe queda estructurado de la siguiente manera:

- El capítulo **2** se expone el estado del arte, es decir, la base teórica del proyecto. Se exponen los estudios relacionados con este proyecto y los métodos utilizados por los mismos.
- Posteriormente en el capítulo **3** se describe el método propuesto en este proyecto para realizar la síntesis de voz a partir de las muestras de EEG y de audio (procedimiento seguido para el preprocesado de las señales o entrenamiento de la red neuronal utilizada entre otros).
- En el capítulo **4** se exponen los resultados obtenidos a partir de las métricas utilizadas y se analiza la calidad de las señales de audio predichas a partir de señales de EEG.
- Finalmente en el capítulo **5** se concluye con una reflexión acerca de los resultados obtenidos y posibles trabajos futuros derivados de este proyecto.

Capítulo 2

Estado del arte

En este capítulo se expone la teoría que ha servido previamente como base del proyecto, realizando una presentación de toda la bibliografía relacionada con este proyecto.

En primer lugar en la sección 2.1 se comenta el proceso de producción del habla en general. Se aborda el proceso desde el punto de vista neurológico (cómo el cerebro genera determinadas ondas cerebrales durante el proceso del habla y como pueden ser obtenidas mediante diferentes técnicas).

Seguidamente en la sección 2.2 se presentan diferentes estudios relacionados con este proyecto y que han servido como base del proyecto. Se indican tanto los métodos utilizados como los resultados obtenido por cada uno de ellos.

En la sección 2.2.3 se indican y exponen los parámetros acústicos utilizados para la parametrización de las señales de voz y que son utilizados en el método propuesto en este proyecto.

A continuación en la sección 2.3 se comentan las diferentes técnicas de machine learning utilizadas en la bibliografía y que son de interés para la aplicación a este proyecto pues las usa como base para el método propuesto.

En las secciones 2.4 y 2.5 se comentan las diferentes métricas de evaluación. Se introducen las y exponen tanto las métricas objetivas como las métricas subjetivas de interés para la evaluación de las señales de audio predichas.

Finalmente en la sección 2.6 se indica el funcionamiento del vocoder utilizado para la parametrización de las señales de voz y su posible reconstrucción a partir de los parámetros extraídos.

2.1. Proceso de producción del habla

Durante el proceso de producción del habla, el cerebro emite ciertas ondas cerebrales características, por lo que para poder hablar de los estudios

realizados es necesario primeramente conocer la fisiología del cerebro y entender cómo éste funciona a nivel interno.

2.1.1. Anatomía y actividad cerebral

El cerebro es uno de los órgano más importantes del cuerpo humano y probablemente el más complejo de todos ellos. A pesar de que su funcionamiento no es trivial y que contiene inmensas cantidades de detalles, su funcionamiento básico puede describirse sin mucha complejidad. El sistema nervioso central o SNC está formado por el encéfalo (formado a su vez por el cerebro, el cerebelo y el tronco encefálico) y la médula espinal, mientras que el sistema nervioso periférico o SNP, de forma que ambas partes se comunican mutuamente de forma constante intercambiando información sensorial y/o motora entre otras.

Respecto al cerebro, éste se encuentra protegido por el cráneo, e internamente por el líquido cefalorraquídeo, encargado de amortiguarlo en caso de sufrir una lesión. Su capa más externa se denomina corteza cerebral y está formada por materia gris (la cual contiene la mayor parte de los cuerpos de las neuronas), se encuentra dividida (aunque no anatómicamente, si se encuentra dividida en cuanto a funciones que desempeñan cada una de las partes) en diferentes lóbulos. Cada uno de ellos se encarga de ciertas funciones concretas como por ejemplo las funciones motoras o cognitivas, entre otras muchas [9]. A continuación se indican los diferentes lóbulos en los que se encuentran dividida la corteza y las funciones que desempeñan cada uno de ellos:

- Lóbulo frontal: encargado de la memoria, las funciones motoras y asociado a tareas relacionadas con el lenguaje y el razonamiento.
- Lóbulo parietal: administra las funciones sensoriales de la persona ya que contiene la corteza somatosensorial, encargada de proporcionar información al individuo acerca de los estímulos sensoriales como la temperatura o la propiocepción.
- Lóbulo occipital: se encarga de las tareas realizadas con la visión ya que contiene la corteza visual.
- Lóbulo temporal: contiene la corteza auditiva de forma que es responsable de procesar todas las señales auditivas.
- Otros autores [10] consideran también el lóbulo de la ínsula, asociado a tareas emocionales y de atención y el lóbulo límbico asociado al aprendizaje y a la consciencia afectiva del individuo, y por tanto, al manejo de sentimientos.

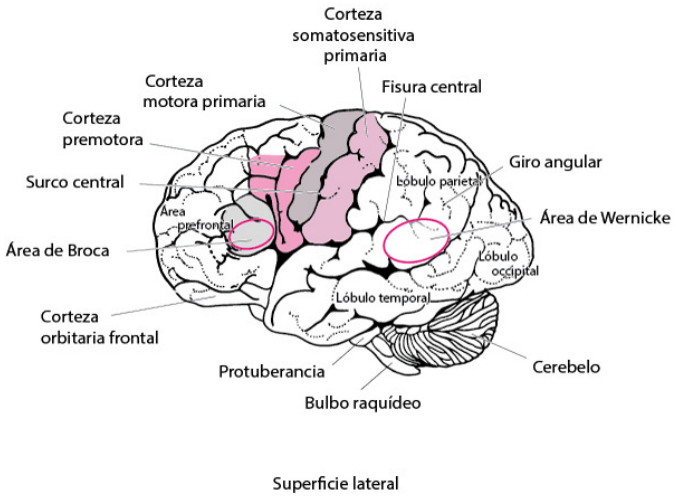

Figura 2.1: Anatomía del encéfalo [11].

Dentro del cerebro, la unidad mínima es la neurona, cuyo cometido es procesar y transmitir la información funcional del sistema nervioso en forma de impulsos electroquímicos. Existen multitud de tipos de neuronas dependiendo del tipo de información que conducen [12] (sensitiva o motora por ejemplo), aunque la mayoría de ellas comparte una fisiología común (con excepciones dependiendo del tipo de neurona en cuestión).

La mayoría de las neuronas están formadas por los siguientes elementos:

- Cuerpo neuronal o soma: es el cuerpo de la neurona, el cual contiene todos los orgánulos que permiten su correcto funcionamiento, además de albergar el núcleo y el ADN de la misma. En el soma es donde se generan los neurotransmisores utilizados por la neurona para la comunicación con otras neuronas. Los impulsos eléctricos recorren la neurona desde las dendritas hasta el axón atravesando el soma.

- Dendritas: constan de prolongaciones utilizadas para la conexión con otras neuronas y por tanto para el intercambio de información a través de la sinapsis. Sirven como entrada de información procedente de otras neuronas.

- Axón: mientras que las dendritas disminuyen su diámetro a medida que aumenta su longitud, el axón es una prolongación que tiene un diámetro fijo, de forma que permite la comunicación con otras neuronas, liberando los neurotransmisores necesarios.

Una vez expuesta la anatomía del cerebro se puede describir el proceso de producción del habla. El primer paso para la producción del habla tiene lugar en el cerebro, lugar donde se generan las señales que posteriormente

se propagarán a zonas concretas del cuerpo involucradas en el proceso del habla. Estas señales son transportadas y procesadas por las neuronas motoras del SNP, las cuales generan los impulsos eléctricos que posteriormente se transformarán en contracciones y relajaciones de los músculos involucrados en la tarea del habla.

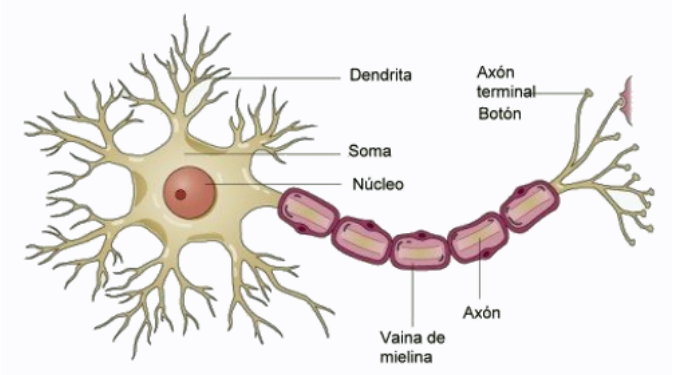

Figura 2.2: Anatomía de la neurona [13].

2.2. Decodificación del habla a partir de señales cerebrales

A continuación se presenta un breve resumen de estudios previos llevados a cabo por otros investigadores con el mismo objetivo con el que surgió este trabajo: proporcionar una vía de comunicación a personas que han perdido la capacidad del habla aunque conservan plenamente la cognición y emociones. Estos estudios han servido como base de este proyecto y exponen diferentes formas de establecer una forma alternativa de comunicación a partir de las lecturas de las señales obtenidas mediante diferentes técnicas, cada cual con unos resultados diferentes.

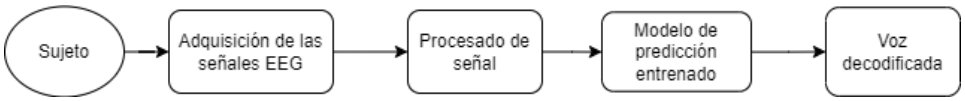

Figura 2.3: Estructura general de un sistema de decodificación del habla a partir de señales de EEG.

2.2.1. Extracción de señales cerebrales

A grandes rasgos, estas técnicas se pueden clasificar como invasivas o no invasivas dependiendo de si requieren intervención quirúrgica para su

implementación o no. A continuación se presentan las diferentes técnicas utilizadas.

Análisis mediante EEG

Se trata de una técnica no invasiva ampliamente utilizada para medir la actividad cerebral, concretamente los potenciales de campo local (LFP en inglés), los cuales se pueden definir como la suma de la influencia de campo todas las neuronas individuales. Estos potenciales tienen un gran ancho de banda (desde corriente continua hasta varios cientos de hercios). Esta técnica utiliza un conjunto de electrodos para medir los LFP mediante su adhesión al cuero cabelludo.

Un problema que tiene esta técnica es que el cráneo y las sucesivas capas que envuelven el cerebro (como el cráneo o la duramadre entre otras) actúan como un filtro paso alto [17], de forma que las bajas frecuencias se atenúan, por tanto al tratarse de una técnica no invasiva las señales medidas mediante esta técnica se encuentran muy atenuadas. Además cada uno de los electrodos utilizados para medir los LFP capta información procedente de poblaciones numerosas de neuronas, de forma que no se puede captar información de forma selectiva procedente de neuronas individuales o grupos de neuronas reducidos.

En función del tipo de electrodos utilizados es necesario o no utilizar un gel conductor (existen de diferentes tipos) para facilitar la transmisión de las ondas cerebrales a los electrodos. Los electrodos no se colocan de forma arbitraria, sino que al proporcionar información procedente de áreas relativamente extensas de neuronas deben colocarse un número determinado de ellos en posiciones específicas para mejorar la resolución espacial siguiendo el sistema 10-20 (aunque existen otros esquemas de montaje).

En definitiva se tiene que al tratarse de una técnica relativamente sencilla, no invasiva y que proporciona una buena resolución temporal, por lo que la obtención de ondas cerebrales mediante EEG es una técnica muy usada y útil.

Análisis mediante MEG

La magnetoencefalografía o MEG se trata de un método no invasivo similar a EEG pero con la diferencia de que este método realiza una medición del campo magnético de grupos reducidos de neuronas producidos durante la sinapsis neuronal, por lo que combina la buena resolución temporal que se obtiene mediante EEG con una buena resolución espacial, lo cual presenta ventajas respecto al uso de EEG. A pesar de ello se trata de un método que requiere de mayor inversión económica debido a la maquinaria necesaria para medir dichos campos magnéticos cerebrales ya que MEG no utiliza

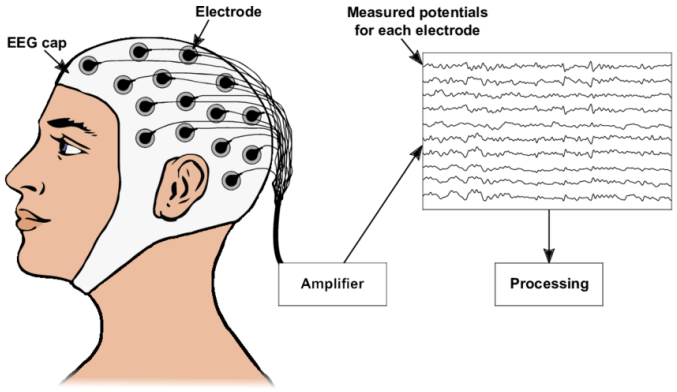

Figura 2.4: Obtención de ondas cerebrales a partir de EEG [14].

electrodos, sino que utiliza un equipo especial que consta de numerosos sensores.

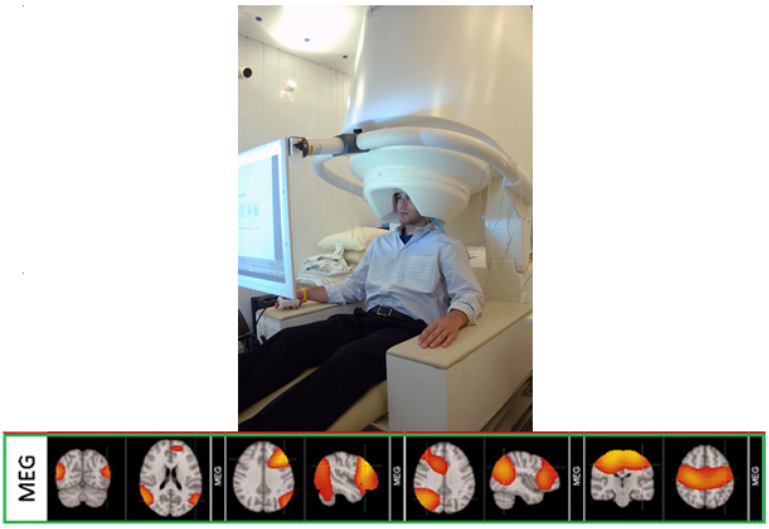

Figura 2.5: Obtención de ondas cerebrales a partir de MEG [15].

Análisis mediante ECoG

Este método se considera invasivo puesto que requiere intervención quirúrgica para poder realizar el análisis de las ondas cerebrales.

El análisis mediante ECoG consiste en la implantación de un array de electrodos sobre el cerebro para realizar mediciones del campo eléctrico y de los potenciales LFP de igual forma que se hace mediante EEG con la excepción de que al realizar las mediciones directamente sobre la corteza cerebral, los potenciales LFP no se atenúan y por tanto se tienen señales de

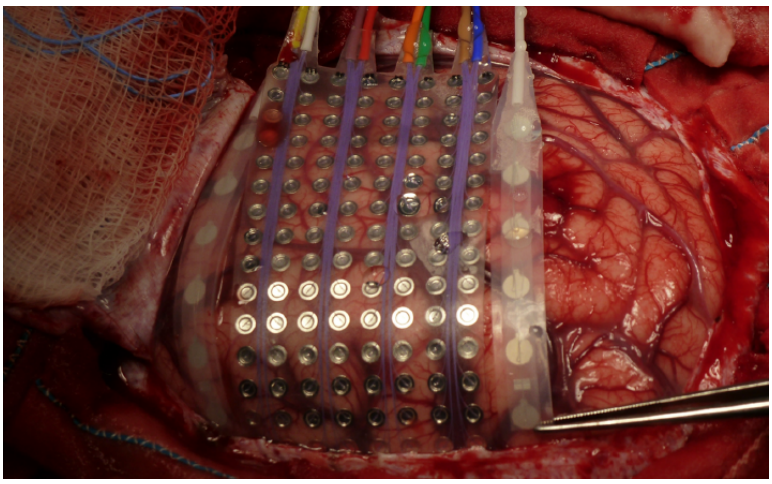

Figura 2.6: Obtención de ondas cerebrales a partir de ECoG [16].

mayor amplitud. El análisis mediante ECoG presenta buenas resoluciones temporales y espaciales. El hecho de tener que realizar una craneotomía para la implementación de la rejilla hace que puedan tenerse complicaciones médicas, por lo que esta técnica suele aplicarse con frecuencia a personas que sufren epilepsia resistente a medicamentos.

2.2.2. Análisis de las ondas cerebrales

Durante la producción del habla se activan ciertas áreas del cerebro relacionadas con el lenguaje como el área de Wernicke o el área de Broca [22], de forma que las neuronas ubicadas en dicha área se activan, produciendo un campo eléctrico en la corteza cerebral medible mediante diferentes técnicas como por ejemplo mediante EEG y MEG o ECoG.

Estos campos eléctricos generados tienen comportamientos periódicos, es decir, las diferentes ondas cerebrales tienen diferentes frecuencias características, cuya amplitud se encuentra en el orden de los microvoltios (μV) y es proporcional al número de neuronas activadas, de modo que a mayor número de neuronas activadas que generen campo a la misma frecuencia, mayor potencia de campo eléctrico se tendrá a esa cierta frecuencia.

En la tabla 2.1 se indica una clasificación de las distintas ondas cerebrales medibles mediante EEG [17].

Tipo de onda cerebral	Ondas delta (δ)	Ondas theta (θ)	Ondas alfa (α)	Ondas beta (β)	Ondas gamma (γ)
Frecuencia (Hz)	$\leq$1-4	4-8	8-13	13-30	30-200

Tabla 2.1: Tipos de ondas cerebrales en función de su frecuencia.

Para medir estos campos como se ha comentado se pueden utilizar diferentes técnicas, donde cada una de ellas ofrece particularidades específicas para cada uso concreto. Estos métodos, y por tanto la posibilidad de detectar ondas cerebrales, aprender sus características y en general acerca del funcionamiento del cerebro nos brinda la posibilidad de poder caracterizar el comportamiento del mismo y así poder detectar actividades anómalas que puedan ser vinculadas a diferentes enfermedades cerebrales.

2.2.3. Coeficientes cepstrales en escala Mel (MFCC)

Las características acústicas de una señal de voz permiten realizar una parametrización de la misma. De esta forma una señal de audio se puede expresar en función de sus coeficientes MFCC o LPC, entre otros.

Los MFCC son coeficientes muy utilizados para la parametrización de señales de audio ya que permiten seleccionar únicamente las componentes más relevantes de la señal de audio y descartar la información menos relevante. Estos coeficientes son ampliamente usados en el campo del reconocimiento automático de locutor, entre otras muchas aplicaciones, siendo los MFCC base de nuevos estudios que buscan nuevas características acústicas [18] [19]. Los coeficientes MFCC contienen información del espectro de la señal de audio de la cual se han obtenido, pero se basan en una escala de frecuencias (denominada escala de frecuencias Mel) que tiene en cuenta la percepción auditiva humana. A continuación se detalla el proceso llevado a cabo para su obtención:

1. Entramado: en primer lugar se debe de segmentar la señal de voz en tramas de determinada duración. Para escoger la duración de las tramas es necesario tener en cuenta que durante pequeños intervalos de tiempo suponemos que la señal de voz es cuasi-estacionaria (la forma del tracto vocal permanece invariante). Típicamente el valor utilizado es de entre 20 y 40 milisegundos. Además en el proceso de entramado se deberá de tener en cuenta que debe existir un solapamiento entre tramas, típicamente de 10 ms.

2. Enventanado: una vez realizado el entramado se debe de aplicar una función de ventana a cada trama. Suele aplicarse una ventana de Hamming, aunque existen muchos tipos de funciones de ventana como Hanning o Blackman.

3. Obtención del espectro de potencia por tramas: para ello se aplica la DCT a cada trama de acuerdo a la ecuación 2.1.

$$X(k) = \sum_{n=1}^{N} x(n) e^{-2\pi jkn/N} \qquad 1 \leq k \leq K \tag{2.1}$$

Cabe comentar que $x(n)$ en la ecuación 2.1 se corresponde a la señal enventanada.

4. Puesto que para su obtención se realiza mediante un análisis de la señal basado en bancos de filtros, la forma de los filtros, el número de los mismos y su ancho de banda influirán en el análisis. Diseño del banco de filtros en escala Mel y aplicación: a continuación puesto que se pretende realizar un análisis basado en un banco de filtros, la forma de los filtros, el número de los mismos y su ancho de banda influirán en dicho análisis. Los MFCC se obtienen haciendo uso de la escala Mel, la cual establece una relación logarítmica entre frecuencia lineal y frecuencia en escala mel, teniendo en cuenta el comportamiento natural del oído humano (se debe tener en cuenta que el oído humano percibe mejor las bajas frecuencias que las altas frecuencias). En las ecuaciones 2.2 y 2.3 se pueden observar las equivalencias entre frecuencia lineal y frecuencia en escala mel:

$$f_{mel} = 1127ln(1 + f/700) \tag{2.2}$$

$$f = 700(e^{f_{mel}/1127} - 1) \tag{2.3}$$

De esta forma se tiene que el banco de filtros aplicado tiene la forma que se puede apreciar en la figura 2.7. Típicamente se suelen utilizar 26 filtros triangulares. Una vez definido el filtro se aplica al espectro obtenido anteriormente para obtener el espectro a la salida de cada filtro.

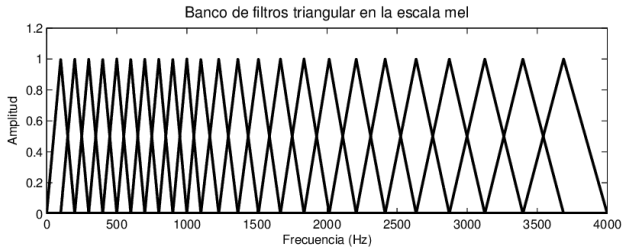

Figura 2.7: Ejemplo de bancos de filtros en escala Mel.

5. Logaritmo de la energía a la salida de cada filtro y DCT: Se calcula el logaritmo del la energía a la salida de cada filtro y posteriormente se calcula la transformada discreta del coseno (DCT) para obtener finalmente los coeficientes MFCC. Tras esto normalmente se suelen escoger los primeros N coeficientes MFCC puestos que son los que contienen la mayor parte de la información de la señal.

2.2.4. Estudios de conversión EEG/ECoG a texto

- En el estudio realizado por Christian Herff [20] se propone el uso de ECoG para realizar registros de la actividad cerebral y por tanto los potenciales eléctricos procedentes de 7 sujetos diferentes mientras se realizaban registros de las ondas acústicas correspondientes de los sujetos (cada sujeto leyó unas frases en voz alta un número determinado de veces) de forma sincronizada. A continuación se extrajo la actividad correspondiente a la banda gamma alta y se segmentaron las señales de voz en tramas de 50 ms con desplazamientos de ventana de 25 ms, para posteriormente realizar un etiquetado de los datos. Además se realizaron filtrados de las señales cerebrales en el intervalo de frecuencias de 58 a 62 Hz para eliminar el ruido de la línea eléctrica. Finalmente la decodificación de las palabras se obtuvo encontrando la secuencia de palabras que tiene la mayor probabilidad de generar la secuencia de datos neuronales de entrada a partir del modelo creado en el estudio. Como resultado se tuvo que para un menor número de palabras en el diccionario, se tienen predicciones más precisas. La mejor de las predicciones tuvo una precisión promedio superior al 50 % en las frases predichas. Para un tamaño del diccionario de 10 palabras, se tiene que el 75 % de las palabras han conseguido ser decodificadas de forma correcta. En general las predicciones promedio obtenidas con este modelo son mucho mayores que los modelos aleatorios lo cual abre el camino a posibles implementaciones en sistemas de escritorio y además abre también paso al diseño de nuevos sistemas que permitan que las personas se puedan comunicar únicamente a partir de las señales cerebrales

2.2.5. Estudios de conversión EEG/ECoG a voz

- En el estudio realizado por Chakrabartiy otros autores [22] se indica que al hablar existen evidencias de variaciones de potencia en la banda gamma alta correspondiente a la banda de frecuencias desde 70 Hz hasta 170 Hz aproximadamente, en las partes superior y media del lóbulo temporal, el área de Wernicke, el área de Broca y la corteza motora primaria, hecho importante para posibilitar una decodificación neuronal lineal y predecir espectrogramas con los que sintetizar la voz.
- Un estudio interesante es el realizado por Christian Herff y otros autores [23] donde se plantea la posibilidad de realizar una síntesis de voz a partir de registros neuronales en partes del cerebro involucradas en los procesos del habla haciendo uso de ECoG.

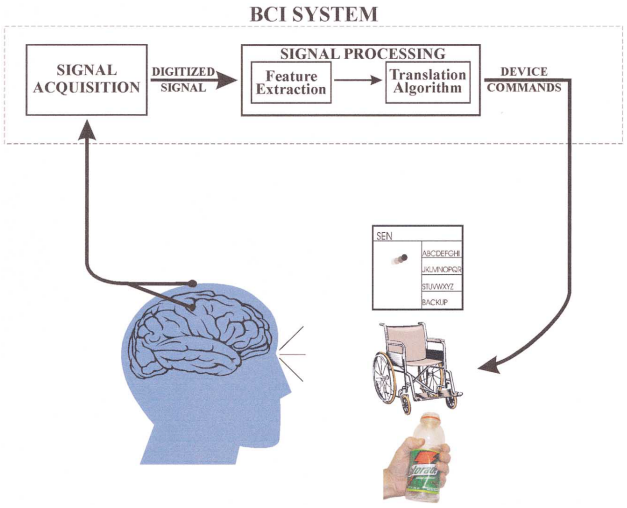

Figura 2.8: Esquema de una ICC propuesto en [21].

Para el estudio se realizaron grabaciones tanto de audio como de ECoG de entre 8,3 a 11,7 minutos de palabras monosilábicas a 6 pacientes a los que se les implantaron unas rejillas de electrodos en el hemisferio izquierdo. Centrándose en la potencia gamma alta, consiguieron obtener reconstrucciones de señales de audio. Con el fin de comparar las reconstrucciones con los audios originales realizaron en cálculo de la correlación cruzada entre ambas señales de audio, obteniendo correlaciones por encima del 95 %. Además realizaron pruebas de escucha en las que sometieron unos 55 voluntarios (40 hombres y 15 mujeres) a evaluar las señales de audio reconstruidas, obteniendo una precisión de aproximadamente el 66,1 %.

- Por último en el estudio realizado por Gopala K. Anumanchipalli [25] se propone el diseño de una herramienta de síntesis de voz a partir de señales cerebrales obtenidas con ECoG. Para ello se contó con 5 pacientes que fueron sometidos a registros de señales ECoG mientras además obtenían señales de audio de los mismos pacientes. En este estudio para la decodificación del habla se realizan estimaciones con una red neuronal de las características articulatorias de la actividad neuronal. A partir de ellas se decodifican las características acústicas de la señal de audio con las características articulatorias, para posteriormente realizar la síntesis. Para evaluar los resultados se diseñaron pruebas de audición a voluntarios que fueron sometidos a evaluar la inteligibilidad tanto de palabras simples como de oraciones. Se obtuvo como resultados que el 43 % de las palabras sueltas se consiguie-

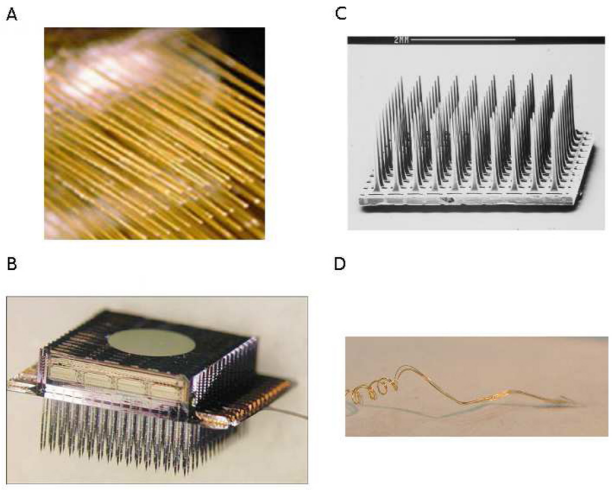

Figura 2.9: Ejemplos de implante de microelectrodos intracorticales [24].

ron transcribir perfectamente, junto con el 21 % de las oraciones.

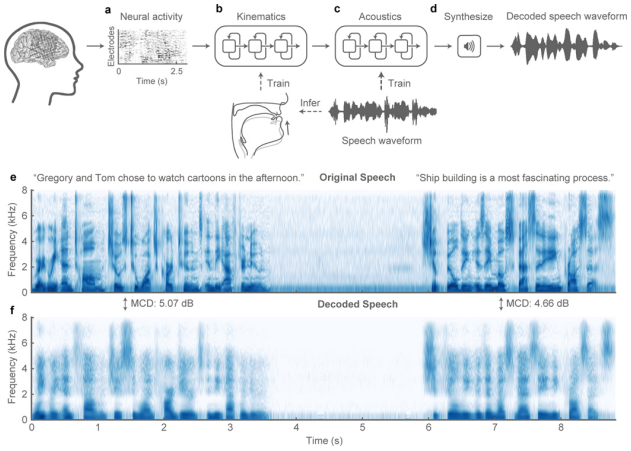

Figura 2.10: Esquema de síntesis de voz a partir de actividad neuronal en [25].

2.3. Fundamento de técnicas de Machine Learning

Las técnicas de Machine Learning tienen como objetivo la creación y diseño de modelos que permitan realizar estimaciones a partir de unos datos de entrada, de forma que estos modelos sean capaces de encontrar patrones que puedan aplicarse a nuevos datos de entrada y así poder obtener

predicciones fiables y precisas.

2.3.1. Método de regresión lineal y logística

El método de regresión lineal simple tiene como objetivo realizar la estimación de los valores de una variable dependiente en función de los valores de una variable independiente. Para ello este método se encarga de realizar el cálculo de los parámetros de una función lineal que permita realizar una correspondencia entre unos valores de entrada de la variable independiente y unos valores de salida de la variable dependiente conocidos previamente.

Dicha función lineal se obtiene ajustando una función lineal a los datos iniciales mediante una función de pérdidas que normalmente se trata del MSE. Una vez obtenida dicha función que establece la relación entre las variables dependiente e independiente, este método permite realizar predicciones en función de los datos de entrada.

Por ello este método funciona mejor cuanto mayor es el conjunto de datos iniciales con el que se obtiene la función que establece la relación entre ambas variables. En la figura 2.11 se puede observar una gráfica con una nube de puntos iniciales y la función obtenida mediante regresión lineal.

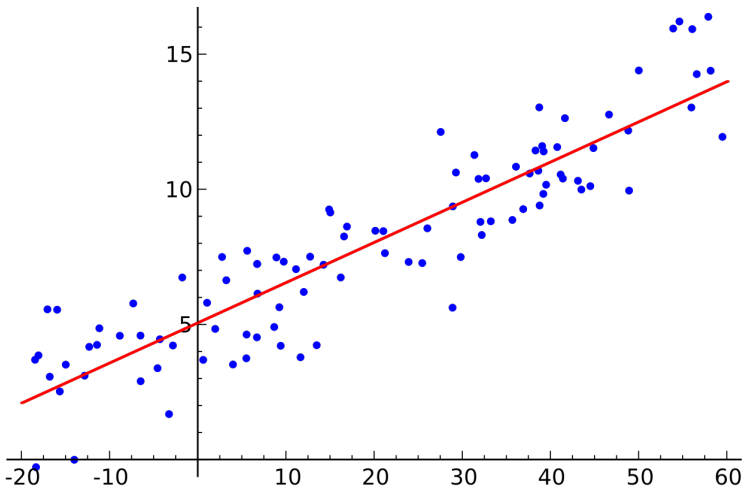

Figura 2.11: Ejemplo de función lineal (en rojo) obtenida mediante regresión lineal a partir de los datos (en azul) [26].

Por otro lado también es interesante el método de regresión logística. El objetivo es similar al método de regresión lineal, con la salvedad de que en este caso la función a ajustar es aquella que intente realizar predicciones finitas y no infinitas como en el caso de la regresión lineal. Un ejemplo de su uso es por ejemplo el ajuste de una función que proporcione la probabilidad de pertenencia de una entrada a un cierto grupo o conjunto. Es por ello que

en tareas de clasificación es bastante usada. En la expresión 2.4 se tiene la forma de la función que la regresión ajusta a los datos iniciales para ajustar el modelo.

$$P\left(\frac{Y=1}{X}\right) = \frac{1}{1+e^{-z}} \tag{2.4}$$

Donde z es la combinación lineal de las variables independientes que se tengan (para casos más simples solo se tienen una variable independiente). Esta función devuelve probabilidades por lo que devuelve valores entre 0 y 1.

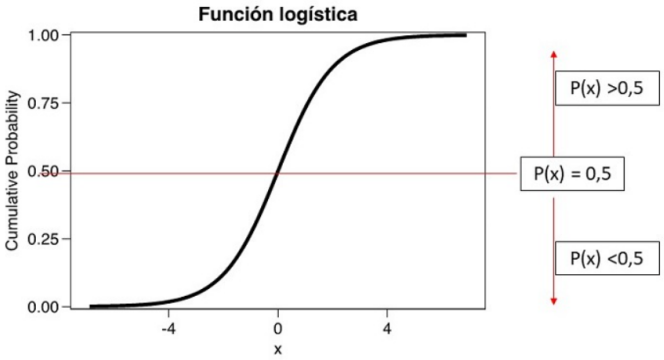

Figura 2.12: Ejemplo de función obtenida mediante regresión logística [27].

En definitiva los métodos de regresión lineal y logístico son unos métodos básicos y sencillos de implementar para realizar predicciones paramétricas en función de unos datos de partida.

2.3.2. Redes neuronales

Las ANN son estructuras que simulan el funcionamiento de las neuronas biológicas del cerebro humano, siendo capaces de procesar información de múltiple naturaleza y de mejorar de forma iterativa su propio funcionamiento y respuesta.

Las redes neuronales están organizadas por capas, donde cada una de las capas está formada por nodos que se conectan a otros nodos de otras capas, de forma que las neuronas o nodos realizan múltiples interconexiones (simulando las estructuras cerebrales).

Habitualmente las redes neuronales se componen de las siguientes capas:

- Capa de entrada: compuesta por nodos que reciben los datos a procesar del exterior, encargados de realizar los procedimientos iniciales y mandarlos a capas intermedias.

- Capas intermedias u ocultas: encargadas de realizar procesamientos intermedios necesarios antes de proporcionar una salida adecuada a la aplicación con la que ha sido creada.

- Capa de salida: última capa de la red neuronal, la cual se encarga de proporcionar información de salida.

De esta forma en la figura 2.13 se puede observar la estructura de una red neuronal.

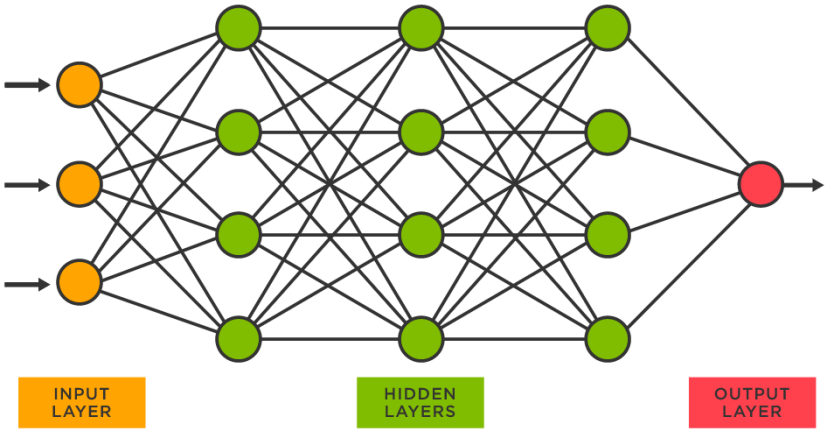

Figura 2.13: Estructura de una red neuronal artificial (ANN).

La unidad más simple de una ANN es la neurona artificial, la cual se puede modelar de forma matemática como un perceptrón. El perceptrón tiene ciertas simetrías a una neurona biológica: tiene unas dendritas encargadas de transmitir información de entrada a la neurona y un axón para la salida de la información procesada hacia otras neuronas. Estas entradas tienen unos pesos asociados de forma que unas entradas tienen una mayor relevancia que otras en la ponderación y por tanto unas influyen más que otras en la obtención de la respuesta a la salida del perceptrón. A continuación se obtiene una suma ponderada de las entradas en función de dichos pesos, y una función de activación encargada de procesar la entrada para proporcionar una salida determinada.

Esta función de activación actúa como limitadora de la neurona, de forma que establece un umbral para decidir si la salida de dicha neurona debe influir o no en el resultado, proporcionando o no una salida. En la figura 2.14 puede observarse un esquema de funcionamiento de un perceptrón, el cual modela una neurona artificial:

La salida toma la forma dada por la expresión 2.5 en caso de superar cierto umbral establecido. En caso contrario la función de activación no proporcionaría una salida.

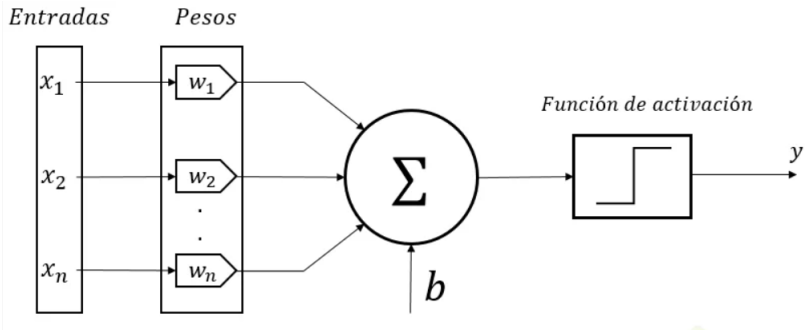

Figura 2.14: Ejemplo de perceptrón con n entradas.

$$y = f\left(b + \sum_{i=1}^{n} x_i w_i\right) \tag{2.5}$$

Teniendo conocimiento de la salida esperada para cada conjunto de entradas, se puede realizar un ajuste los pesos de las entradas de manera iterativa para obtener, en cada una de las iteraciones, unas mejores predicciones de las salidas correspondientes respecto de las salidas correctas. De esta forma la neurona es capaz de ajustar sus pesos para proporcionar salidas más precisas.

Para el ajuste de los pesos, las salidas obtenidas se comparan con las salidas correctas, intentando minimizar una función de error o pérdida establecida, por ejemplo el error cuadrático medio o MSE (que se obtiene a partir de la expresión 2.6, donde $\hat{y}_i$ es la salida predicha y y_i la salida correcta).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{2.6}$$

Las redes neuronales están formadas por numerosas de estas neuronas artificiales, y para la actualización de los pesos de las neuronas se aplica el algoritmo *backpropagation* o retropropagación, el cual proporciona un aprendizaje supervisado a la red. Este algoritmo realiza el calculo de la derivada parcial de la salida de la función de coste (o costo) con respecto a cada peso de correspondiente a cada entrada para obtener la contribución de cada peso en el cálculo del error, y de esta forma tener una idea de los pesos que contribuyen a aumentar el error, de forma que estas derivadas se retropropagan entre capas para actualizar cada uno de los pesos de forma que tomen valores que contribuyan a disminuir la salida de la función de coste establecida, técnica denominada descenso en gradiente (aunque pueden aplicarse otros métodos de optimización). En la figura 2.15 se puede observar la corrección de los pesos aplicada al perceptrón por el algoritmo de *backpropagation*, haciendo uso de una técnica de optimización.

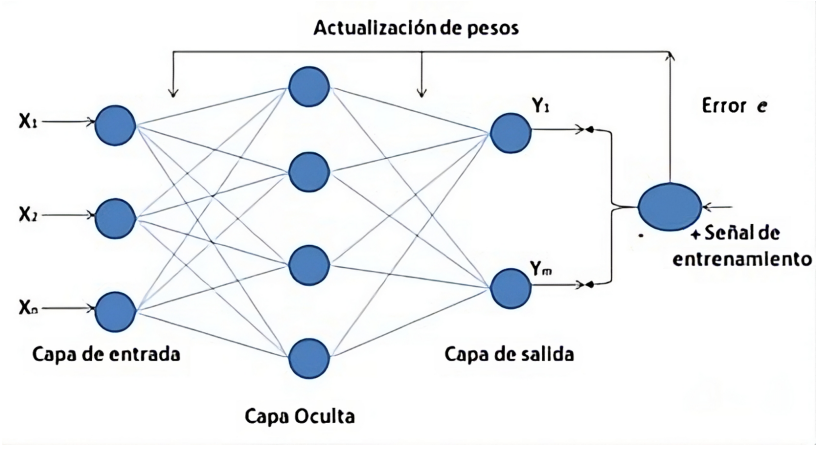

Figura 2.15: Ejemplo de aplicación del algoritmo *backpropagation* mostrado en [28].

Para redes neuronales completas el algoritmo se aplica de forma análoga, aplicando el descenso en gradiente o cualquier otra técnica de optimización capa a capa, retropropagando a capas anteriores los resultados obtenidos y actualizando cada peso de acuerdo con la técnica elegida. Se puede consultar información adicional sobre las ANN en la revisión realizada por Nikolaus Kriegeskorte y otros autores [29].

Estas redes tienen capacidades de aprendizaje interesantes para la aplicación a este proyecto, puesto que una vez diseñados y entrenados los modelos de redes escogidos, tienen la capacidad de realizar predicciones a partir de los datos de entrada de forma que actúen como una 'caja negra' de acuerdo al entrenamiento realizado, sin tener que preocuparnos por lo que pasa internamente en ella. La finalidad es realizar un entrenamiento de una red neuronal con datos iniciales preprocesados de EEG para realizar predicciones de los parámetros acústicos de las señales (concretamente de los MFCC de cada trama de voz), para posteriormente realizar una evaluación de los resultados obtenidos.

2.3.3. Redes neuronales convolucionales (CNN)

Las redes neuronales convolucionales (CNN por sus siglas en inglés), son un tipo de redes que constan con capas de filtrado convolucional, por lo que suelen usarse en tareas que requieran extracción de características o reconocimiento de patrones, lo que las hace muy útiles en tareas de reconocimiento de voz o procesamiento de imágenes [30]. La estructura de las CNN suele presentar el siguiente tipo de capas: capas convolucionales, capas de agrupación y capas totalmente conectadas o DNN:

- Capas convolucionales: son capas que calculan una convolución de los

datos de entrada a dichas capas. Para ello emplean filtros o kernels (dependiendo del propósito y el tamaño de los datos de entrada, el tamaño de estos kernels toma diferentes tamaños) para obtener un mapa de convolución o mapa de características, resultado de la aplicación de la convolución entre los datos de entrada y los kernels. Para la obtención del mapa de convolución los filtros o kernels se van desplazando un número determinado de muestras (denominado paso o *stride*) y se calculan sucesivamente las convoluciones de los datos de entrada con los filtros. Estas capas pueden aplicar una función de activación (usualmente una función de activación ReLU). Por ello estas capas son las encargadas de realizar la extracción de las características de los datos a partir de las convoluciones. En la expresión 2.7 se puede observar la forma de la convolución de una señal y un filtro unidimensionales, donde $x(n)$ se refiere a la señal de entrada, $y(n)$ se refiere a la señal de salida y $h(n)$ es el filtro de convolución de tamaño L.

$$y(n) = x(n) * h(n) = \sum_{l=0}^{L-1} x(l) \cdot h(n-l) \tag{2.7}$$

- Capas de reducción de la dimensión o 'subsampling': el resultado de las convoluciones puede dar lugar a una cantidad muy elevada de datos dependiendo del número de filtros aplicados, por lo que el objetivo de esta capa es realizar un submuestreo de los resultados obtenidos en las capas convolucionales de forma que las características más importantes obtenidas por ellas se conserven. Para ello existen diferentes métodos:

  - Max-pooling: Usualmente es el método elegido para realizar la reducción de convolución. El método consiste en establecer un tamaño de submuestreo NxN, de forma que esta matriz recorra cada mapa de características de cada filtro aplicado y de esta forma de cada bloque de NxN muestras, únicamente preserva el valor de la muestra del mapa que mayor valor tiene. De esta forma se realiza una reducción del tamaño del mapa de características y demás se preserva la información más relevante.
  - Average-pooling: El funcionamiento es análogo a la técnica de Max-pooling, pero realizando la media del bloque de NxN muestras escogido, de forma que en lugar de guardar el valor de la muestra de mayor valor dentro de cada cuadro de NxN muestras, se guarda la media de los valores dentro del cuadro de muestras de tamaño NxN del mapa de características.

- Capa de neuronas totalmente conectadas: Por último las CNN suelen contener una capa de neuronas totalmente conectadas para trabajar

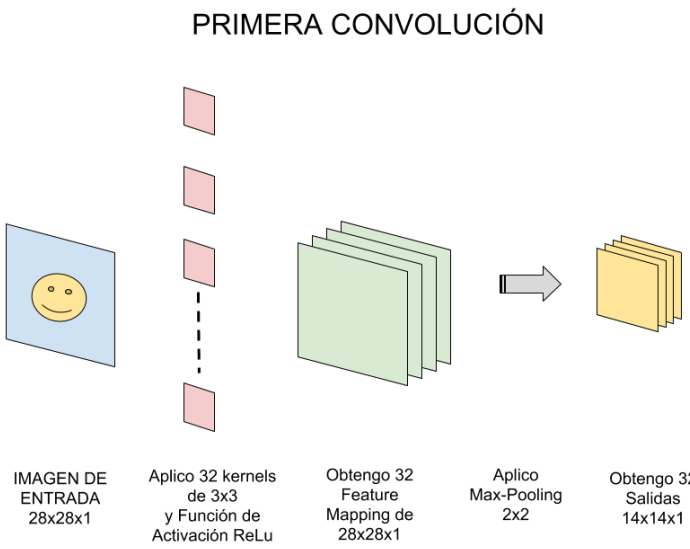

Figura 2.16: Ejemplo de capa convolucional y 'subsampling' para procesamiento de imágenes [31].

con las características extraídas de las capas de convolución y submuestreo y realizar tareas de regresión o clasificación.

2.3.4. Funciones de activación

Como se comenta en la sección 2.3.2, las funciones de activación se encargan de activar o no la neurona de forma que se transmita o no una salida por parte de la neurona en cuestión, y en caso de transmitir una salida, pueden actuar de filtros, haciendo que la suma ponderada de las entradas sea modelada por dicha función de activación y se proporcione una cierta respuesta, dependiendo de la función de activación escogida. Por ello las funciones de activación determinan en cierto modo el funcionamiento de la red neuronal, optimizándola para tareas concretas dependiendo de la función escogida. A continuación se comentan algunas de estas funciones de activación:

- Función sigmoidal: Esta función proporciona valores de salida entre 0 y 1, transformando los valores de entrada de la función de activación de acuerdo a la expresión 2.8. Esta función presenta asíntotas horizontales en *y=0* y en *y=1* por lo que valores de entrada muy bajos tienden a 0 y valores altos tienden a 1.

$$f(x) = \frac{1}{1 - e^{-x}} \tag{2.8}$$

- Función rectificadora (ReLU): proporciona valores de salida iguales que los de entrada en caso de ser positivos, y proporciona una salida nula en caso de tener valores a la entrada. En la ecuación 2.9 se puede observar la salida de la función ReLU.

$$f(x) = max(0, x) = \begin{cases} 0 \; si \; x < 0 \\ x \; si \; x \geq 0 \end{cases} \tag{2.9}$$

- Función Softmax: Se utiliza en las capas de salida de las redes neuronales y proporciona salidas que se corresponden con probabilidades de pertenencias a clases distintas, por lo que es muy usada en clasificadores.

- Función Softplus: Útil para la aplicación de la técnica de *backpropagation* debido a la facilidad de obtener su derivada (coincidente con la propia función sigmoidal). En la expresión 2.10 se puede observar la forma de la misma.

$$f(x) = ln\,(1 + e^x) \tag{2.10}$$

2.3.5. Funciones de coste

Existen diferentes funciones de coste para determinar distancias de error entre las salidas correctas y las salidas predichas por las redes neuronales. A continuación se enumeran algunas de las más utilizadas:

- MSE (Mean Square Error): Como se puede observar en la expresión 2.6, esta función calcula el sumatorio del cuadrado de las diferencias entre la salida esperada y la salida predicha.

- MAE (Mean Absolute Error): Esta función calcula el valor absoluto de la diferencia entre la salida esperada y la predicha por el modelo (véase la expresión 2.11).

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{2.11}$$

  El término $\hat{y}_i$ se refiere a la salida predicha, y_i a la salida correcta N al número de datos de correspondencia para los que se está calculando el error.

- BCE (Binary Cross Entropy): suponiendo que la red neuronal proporciona valores de probabilidad de pertenencia a una clase u otra, es decir de clasificación (suponiendo únicamente dos posibles clases), la

función de entropía cruzada binaria se encarga de proporcionar distancias entre funciones de probabilidad estimada y correcta, produciendo valores de salida en el rango [0,1]. En la expresión 2.12 se puede observar la función BCE.

$$BCE = -\left(y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i)\right) \tag{2.12}$$

De nuevo el término $\hat{y}_i$ se refiere a la salida predicha, y_i a la salida correcta.

2.3.6. Funciones de optimización

Como se ha comentado anteriormente, las funciones de optimización ajustan los pesos de las redes neuronales para optimizar la función de coste seleccionada. A continuación se indican algunas de las más utilizadas:

- Optimizador por descenso en gradiente: Este optimizador toma la función de coste y trata de realizar un cálculo del paso que hay que dar para minimizar dicha función. Para ello calcula el gradiente de la función error escogida respecto de las entradas y se da un paso en el sentido negativo del mismo para llegar al mínimo de la función de coste. La tasa de aprendizaje o velocidad con la que varían los pesos para que converja el algoritmo es fija para este método.

- Optimizador Adam: es un algoritmo relativamente novedoso que calcula de forma adaptativa las tasas de aprendizaje de cada parámetro de la red de forma que permite un aprendizaje y convergencia más rápidos, utilizando un menor coste computacional para el entrenamiento de los modelos de redes neuronales.

$$\hat{m_t} = \frac{m_t}{1 - \beta_1^t} \qquad \hat{v_t} = \frac{v_t}{1 - \beta_2^t} \tag{2.13}$$

$$w_{t+1} = w_t - \mu \frac{\hat{m_t}}{\sqrt{\hat{v_t}} + \epsilon} \tag{2.14}$$

En las expresiones 2.13 y 2.14 se tiene que m_t, v_t se corresponden con la suma acumulada de gradientes y con la suma de los cuadrados de los gradientes anteriores respectivamente. Por otro lado ϵ es una constante de valor muy reducido para evitar divisiones entre 0, μ se corresponde con la tasa de aprendizaje, β_1^t y β_2^t se corresponden con las tasas de influencia de la media móvil de los gradientes y con la tasa de influencia de la media móvil de los cuadrados de los gradientes, respectivamente.

2.3.7. Método de validación cruzada (cross-validation)

La validación cruzada es un algoritmo utilizado para comprobar el rendimiento de un método aplicado sobre un número determinado de datos u observaciones. Esta técnica es adecuada cuando las observaciones o datos que se poseen son reducidas.

Este algoritmo divide el conjunto de datos es N subconjuntos, y de forma iterativa utiliza N-1 subconjuntos como datos de entrenamiento y el subconjunto restante como datos de test. En cada una de las iteraciones (en total se realizan N iteraciones) se escoge uno de los subconjuntos como datos de test, diferente en cada iteración y dejando el resto de los N-1 subconjuntos como datos de entrenamiento. En cada una de las iteraciones se evalúa el modelo desarrollado con la clasificación de datos de entrenamiento-test realizada, y una vez realizadas las evaluaciones en cada iteración, la precisión total de las evaluaciones se obtiene como el promedio de las precisiones obtenidas en cada iteración. Para el caso de utilizar un solo dato de test y el resto de datos de entrenamiento, se tiene que existen tantas iteraciones como número de datos haya. Este caso particular se denomina *leave-one-out cross-validation*.

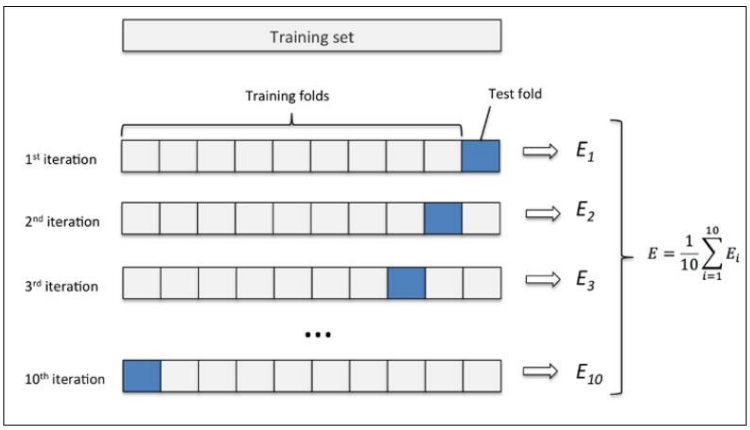

Figura 2.17: Esquema de operación del algoritmo de validación cruzada.

Este método se aplica a numerosos estudios relacionados con el actual [32] [33], demostrando su funcionamiento y utilidad. En el caso de este proyecto, se realizará la evaluación del modelo propuesto utilizando la base de datos indicada en la sección 4.1.1, en la que se tienen datos de 10 pacientes, con 100 muestras de voz cada uno, es decir, 100 palabras grabadas (excepto el paciente 10, del que se tienen 94 muestras de voz).

2.4. Métricas objetivas de evaluación

En esta sección se describen las métricas objetivas utilizadas para la evaluación de las predicciones obtenidas haciendo uso del método propuesto. Se presenta el funcionamiento de cada una de ellas, distinguiendo entre las que se aplican directamente a las secuencias de coeficientes MFCC y aquellas que se aplican sobre las señales de audio.

2.4.1. Mel Cepstral Distorsion (MCD)

Esta métrica objetiva permite obtener una medida de similitud entre dos secuencias de coeficientes MFCC proporcionados, y por tanto evaluar el grado de similitud entre dos señales de audio. Esta métrica calcula un coeficiente a partir de ambas secuencias a partir de la expresión 2.15:

$$MCD = \frac{10}{ln(10)} \sqrt{2 \sum_{i=1}^{K} (mc_i^{(t)} - mc_i^{(e)})^2} \tag{2.15}$$

donde $mc_i^{(t)}$ se corresponde con los coeficientes MFCC de la señal original y $mc_i^{(e)}$ se corresponde con los coeficientes MFCC predichos, y K se define como el número de coeficientes MFCC a comparar. Esta operación nos devuelve un escalar, que cuanto menor sea, menor será la distancia entre ambas secuencias de MFCC y por tanto mejor será la predicción realizada. Por contrario un coeficiente MCD demasiado alto indica una mayor distancia entre ambas secuencias, y por tanto una peor semejanza entre ambas secuencias de MFCC y por tanto entre ambas secuencias de audio.

2.4.2. Short-Time Objective Intelligibility (STOI)

Este algoritmo establece un grado de similitud entre la señal de voz original (o limpia) y la señal de voz predicha (o contaminada). De esta forma el algoritmo devuelve un escalar que indica el grado de inteligibilidad de la señal predicha respecto de la original. Para ello realiza el proceso que se puede observar en la figura 2.18: Se obtienen las envolventes de las señales para ciertas bandas de frecuencias y a continuación se segmentan en tramas de 386 ms de duración para, aplicando una normalización y un clipping (eliminado de muestras por debajo de un cierto umbral) calcular el coeficiente de correlación de Pearson entre las tramas por un lado y entre las bandas de frecuencias de las envolventes por otro. Finalmente se toma la media de los coeficientes de correlación, obteniendo un valor 0 para señales totalmente diferentes y 1 para señales de entrada exactamente iguales.

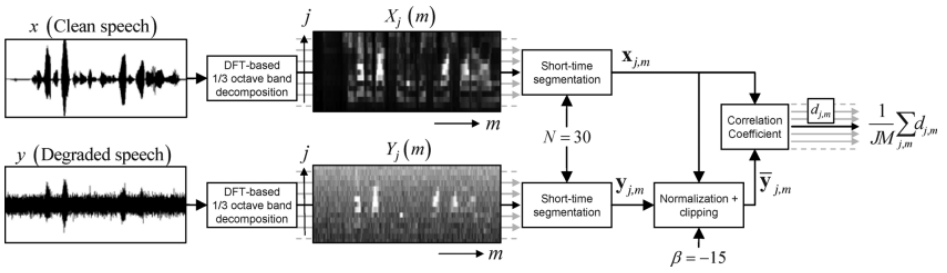

Figura 2.18: Esquema de funcionamiento de la métrica STOI [34].

2.5. Métricas subjetivas de evaluación de la predicción

En esta sección se comenta la métrica subjetiva utilizada para llevar a cabo las comparaciones entre las señales de audio originales y las señales de audio predichas.

2.5.1. Perceptual Evaluation of Speech Quality (PESQ)

Esta métrica propuesta por Antony W. Rix y otro autores [35] realiza una evaluación subjetiva de la señal de audio predicha, de forma que si esta métrica devuelve una alta puntuación (la máxima puntuación es de 4,5), la calidad de la señal predicha será mejor que si la métrica devuelve una puntuación baja (la mínima puntuación es de -0,5). En la figura 2.19 se puede observar el funcionamiento de la métrica PESQ.

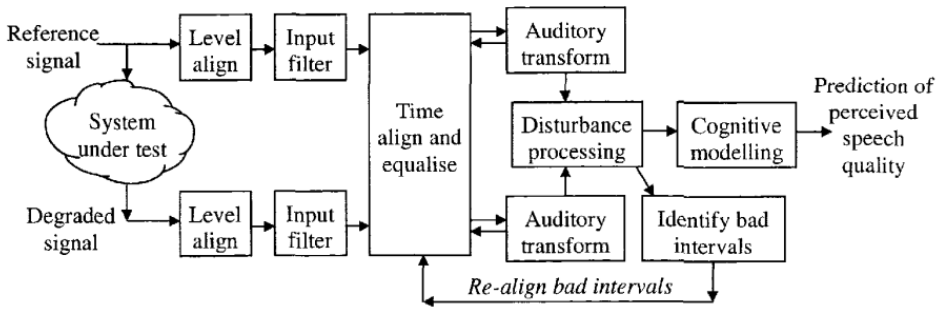

Figura 2.19: Esquema de funcionamiento de la métrica PESQ [35].

Durante el proceso este algoritmo lleva a cabo un filtrado de la señal de entrada tal y como haría un sistema telefónico, de forma que esta métrica proporciona medidas muy similares a las obtenidas mediante opiniones subjetivas.

2.6. Vocoder WORLD

El vocoder WORLD es un programa utilizado para la parametrización y el análisis de las señales de voz. Este software es capaz de extraer parámetros de la señal de audio de entrada, concretamente la frecuencia fundamental, relacionada con el timbre de la persona en cuestión y definida como la frecuencia de resonancia menor de las cuerdas vocales; la matriz de aperiodicidades la cual contiene información acerca de la existencia de tramos de voz sonoros o sordos y que contribuye de forma directa en la calidad de la voz predicha; y por último el espectrograma suavizado de la señal de audio analizada, el cual contiene información de la contribución de los armónicos y resto de frecuencias en la señal de voz.

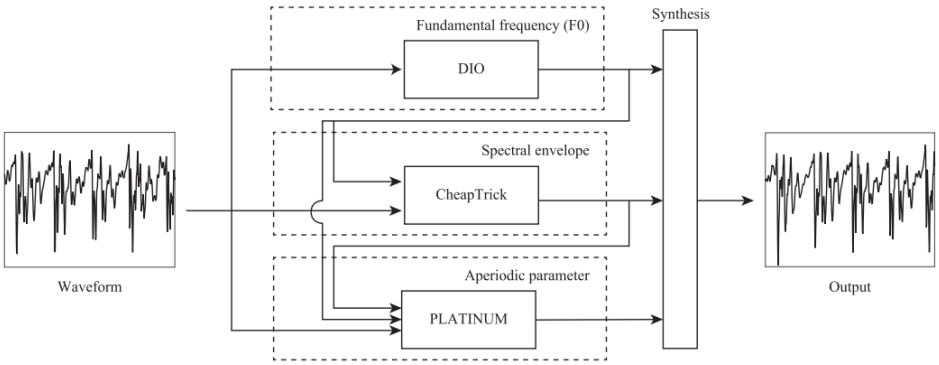

Figura 2.20: Funcionamiento del vocoder WORLD propuesto en [40].

El vocoder como se puede observar en la figura 2.20 es capaz de realizar la operación inversa, es decir la síntesis de la señal de audio a partir de los parámetros comentados. Adicionalmente se puede sintetizar la voz en formato whispered o susurrada en caso de disponer únicamente de los coeficientes MFCC. El uso del vocoder WORLD frente a otras opciones radica en la posibilidad de poder ser usado en tiempo real debido a su velocidad de ejecución.

Capítulo 3

Metodología propuesta

En este capítulo se describe el método propuesto para permitir decodificar el habla de una persona a partir de registros de su actividad cerebral captadas con EEG. El método se centra en el enfoque de síntesis directa de voz (conversión de EEG a voz) debido a las ventajas que presenta frente a la conversión de EEG a texto. La conversión de EEG a texto necesita un mayor número de datos de entrenamiento que el método de conversión de EEG a voz para la obtención de buenos resultados. Además la conversión de EEG a texto no permite el reconocimiento de palabras que no fueron consideradas durante el entrenamiento. Se puede observar la implementación del sistema diseñado en [37].

En la sección 3.1 se presenta el preprocesado inicial que es necesario aplicar a los datos de partida para realizar posteriormente la síntesis. En la figura 3.1 se puede observar el procesado de los datos necesario para la síntesis de los mismos. En esta sección también se comentan las diferencias que toma el esquema de la figura 3.1 dependiendo de la red neuronal utilizada para el procesado de los datos puesto que la red neuronal propuesta está diseñada para procesar los datos de EEG en crudo.

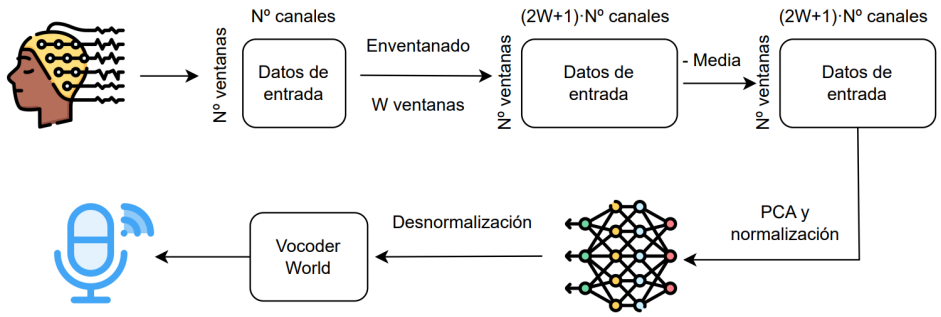

Figura 3.1: Procedimiento para la síntesis de voz.

En la sección 3.2 se comenta el funcionamiento de la red neuronal SincNet, utilizada para la inclusión de una capa convolucional en el modelo propuesto basada en dicha red. A continuación se detalla la estructura y funcionamiento de la red neuronal propuesta para el procesamiento de las bioseñales, basada en la red neuronal propuesta en [38].

Finalmente en la sección 3.3 se detallan la forma en la que se ha realizado el procesado de los datos por parte de la red neuronal propuesta, así como los parámetros establecidos para el entrenamiento de la misma. También se comenta la forma en la que se ha utilizado el vocoder WORLD para realizar la síntesis de las señales de voz a partir de los coeficientes MFCC predichos por la red.

3.1. Preprocesado de los datos

En primer lugar se debe de realizar un preprocesado de los datos iniciales. Partiendo de una base de datos en la que se tienen señales de audio y de EEG sincronizadas, se deben adaptar los datos al sistema de predicción diseñado. A cada tipo de dato se le debe de aplicar un preprocesado diferente debido a la diferente naturaleza de los mismos.

3.1.1. Procesado inicial de los datos de audio

Los datos de audio deben de tener una frecuencia de muestreo de 16 kHz para poder tratarlos de manera cómoda. Para ello en caso de contar con señales de audio con una frecuencia de muestreo superior, se debe realizar un submuestreo o diezmado de las señales de audio de acuerdo a la expresión 3.1 (en caso de hacer un diezmado de factor M). También puede resultar interesante incluir previamente un filtro anti-aliasing.

$$y(m) = x(Mm) = \sum_{k=-\infty}^{\infty} x(Mk)\delta(m-k) \tag{3.1}$$

Una vez realizado el diezmado de las señales de audio es interesante realizar una parametrización de la voz y obtener los coeficientes MFCC de las señales de voz segmentadas en tramas. Para ello se ha usado el vocoder WORLD para parametrizar la voz en ventanas temporales con desplazamiento de 10 milisegundos (en la sección 2.6 se detalla el funcionamiento del mismo). Para la utilización de este vocoder se debe especificar tanto la frecuencia de muestreo de las señales de audio que van a parametrizarse (en este caso las señales decimadas tienen una frecuencia de muestreo de 16 kHz). También debe indicarse el desplazamiento de las ventanas temporales en las que se segmentan las señales de audio (el solapamiento entre muestras permite mantener el contexto temporal ya que de esta forma una determinada ventana de muestras contiene información de la ventana o ventanas

anteriores); en este caso se ha indicado un paso entre ventanas de 10 milisegundos. Finalmente otro de los parámetros que deben modificarse para adaptar el vocoder a nuestros datos es el número de coeficientes MFCC que se obtienen de cada ventana de cada señal de audio procesada y en nuestro caso se han seleccionado 25 coeficientes por trama.

3.1.2. Procesado inicial de los datos de EEG

Para el caso de síntesis utilizando una red DNN simple utilizando los datos de EEG se debe hacer un preprocesado para extraer las características high-gamma (no para las señales PMA utilizadas para comparar resultados). Para ello se realizan dos filtrados rechaza-banda para atenuar las bandas de frecuencias correspondientes a los 2 primeros armónicos de la línea eléctrica y un filtrado adicional para la extracción de las características high-gamma (en el caso de procesado mediante la red propuesta se usan las señales en crudo como se comenta más adelante, por lo que se omite este último filtrado). Las frecuencias críticas de este filtrado dependen de los datos analizados. Se pueden ver los detalles del filtrado en la tabla 3.1. De esta forma se puede comparar si es mejor utilizar las características high-gamma o realizar procesados en crudo con la red SincNet propuesta.

Propósito	**Tipo de filtro**	**Frecuencias críticas (Hz)**	**Orden del filtro**
Atenuación del primer armónico de la línea eléctrica	Rechaza-banda	98-102	4
Atenuación del segundo armónico de la línea eléctrica	Rechaza-banda	148-152	4
Extracción de características high-gamma	Paso-banda	70-170 ó 70-127	4

Tabla 3.1: Filtros utilizados para el preprocesado de las señales de EEG.

3.1.3. Enventanado, aplicación de PCA y normalización de los datos

Puesto que se pretende realizar una predicción de los coeficientes MFCC a partir de las señales EEG/PMA, resulta útil realizar un enventanado de la señal. Esto se debe a que si se asignan varias ventanas de los datos (tanto anteriores a la ventana central como posteriores) a una ventana de coeficientes MFCC se logra conservar el contexto temporal de las ventanas adyacentes, lo cual aporta ventajas a la hora de realizar las predicciones.

El problema es que esto genera un aumento de la dimensionalidad de los datos. Una solución a este problema es la aplicación de PCA a los datos enventanados para así almacenar únicamente aquellas componentes que representan la mayor parte de la varianza de los datos, y con ello realizar una reducción de la dimensionalidad de los datos, obteniendo un menor coste computacional y complejidad en el entrenamiento de la red neuronal. Cabe mencionar que primeramente antes de aplicar PCA a los datos de entrada se les debe restar la media (de cada canal).

Por otro lado los datos de entrada a la red neuronal deben de ser normalizados (tras aplicar PCA) antes de ser introducidos en la misma para evitar los efectos de escala en el entrenamiento de la red. El hecho de realizar el entrenamiento con los datos sin normalizar puede llevar a un aprendizaje no válido de la red, provocando la generación de predicciones incorrectas. Por tanto para evitar problemas con las escalas de los datos tanto de entrada como de salida de la red, se debe realizar una normalización. Puesto que se tienen distintos canales, la normalización se realiza canal por canal de forma independiente puesto que podrían tener escalas diferentes (diferentes medias o desviaciones en cada canal). La normalización de los datos de entrada y salida de la red (es decir de los datos de EEG y los MFCC) deben de realizarse por separado de acuerdo a la expresión 3.2

$$X_{normalizado} = \frac{X - \overline{X}}{\sigma} \tag{3.2}$$

Donde $\overline{X}$ se refiere a la media de las muestras pertenecientes a cada canal y σ se refiere a la desviación típica de las muestras.

Obtenidas las salidas de la red neuronal y puesto que se ha realizado una normalización previa, es necesario desnormalizar los datos de salida, es decir los MFCC predichos, por lo que cada uno de los MFCC se desnormaliza de acuerdo a la normalización aplicada previamente. En la figura 3.1 se puede observar el procesado de los datos aplicado tras el preprocesado inicial de la sección 3.1.

Cabe comentar que el método propuesto para el análisis de las señales EEG mediante una red neuronal basada en SincNet realiza un procesado de las señales en crudo, por lo que el proceso de enventanado se mantiene, aunque no se aplica ni PCA ni normalización a los datos de entrada a la red (señales EEG). En cambio si se usa una red DNN sí que se aplican las técnicas expuestas.

3.2. Arquitectura de la red neuronal SincNet

La arquitectura SincNet es un tipo de red neuronal convolucional en la que la primera capa de convolución o extracción de características está diseñada específicamente para aprender un banco de M filtros parametrizados

por funciones tipo *sinc* (véase expresión 3.3). SincNet incluye una primera capa de convolución que filtra las señales de entrada con un banco de filtros tipo *sinc* parametrizados, de forma que evita aplicar filtros más genéricos y por tanto evitar resultados incongruentes (especialmente cuando se tienen pocas muestras con las que se pueda entrenar la red), además de evitar tener que procesar y aprender todos los elementos de todos los filtros. De esta forma la convolución que propone la arquitectura SincNet se puede observar en la expresión 3.4, en la que se observa que ϕ se refiere a los parámetros de los kernels predefinidos que se aprenden.

$$sinc(x) = \frac{\sin x}{x} \tag{3.3}$$

El hecho de incluir capas con filtros *sinc* parametrizados presenta ventajas a la hora de realizar tareas de extracción de características de las señales de entrada frente a CNN convencionales, obteniendo mejores resultados en las tareas de identificación de locutor. Otra ventaja es que permite operar con la señal en crudo de forma que no necesita un preprocesado previo. Estas ventajas hacen que sea interesante su aplicación al caso de decodificación del habla a partir de señales EEG ya que permite adaptar el banco de filtros de tipo sinc en función de las frecuencias que puedan aportar una mayor información.

$$y(n) = x(n) * h(n, \phi) \tag{3.4}$$

La forma que tienen estos filtros tipo *sinc* tanto en el dominio de la frecuencia como en el dominio del tiempo se pueden contemplar en la expresiones 3.5 y 3.6. Se tratan de filtros paso-banda rectangulares en frecuencia (funciones *sinc* en el dominio del tiempo), de forma que los filtros optimizan las frecuencias críticas, es decir, las frecuencias de corte inferior y superior, quedando totalmente caracterizados a partir de las frecuencias en cuestión, y estableciendo un contexto previo en la definición de las formas de los kernels o filtros. En la figura 3.2 se pueden ver una comparativa de filtros prendidos por una CNN estándar y por la red SincNet, tanto en el dominio del tiempo como en el dominio de la frecuencia.

$$H(f, f_1, f_2) = rect\left(\frac{f}{2f_2}\right) - rect\left(\frac{f}{2f_1}\right) \tag{3.5}$$

$$h(n, f_1, f_2) = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{3.6}$$

También es interesante contemplar una comparativa entre la señal filtrada mediante una CNN estándar y la red SincNet tras cierto tiempo de entrenamiento 3.3, se puede observar que la red SincNet aprende mucho más rápido a evitar el ruido blanco introducido artificialmente en la señal de voz.

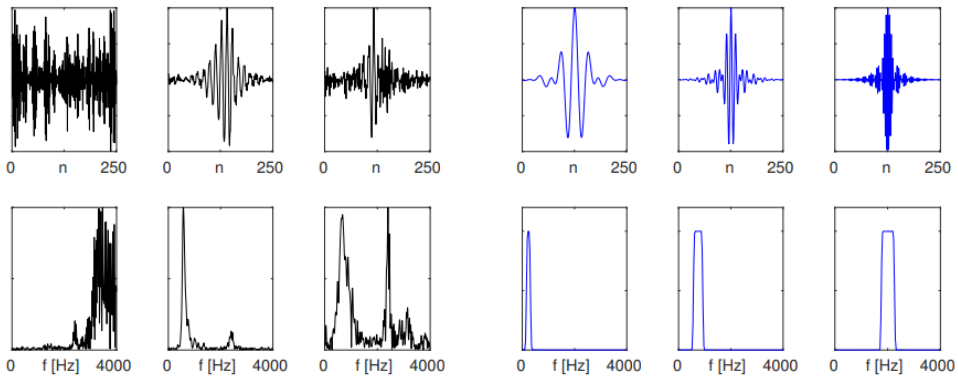

Figura 3.2: Filtros aprendidos por una CNN estándar (izquierda) y filtros aprendidos por SincNet en el dominio del tiempo y la frecuencia [36].

El hecho de que los filtros únicamente cuenten con dos parámetros seleccionables hace que presente una enorme ventaja frente a las CNN convencionales. En el caso de una CNN convencional con F número de filtros de longitud L cada uno de ellos, el número de parámetros con los que cuenta es de $F \cdot L$, mientras que SincNet únicamente cuenta con $2F$ parámetros. Esto conlleva una reducción de parámetros en uso muy ventajosa puesto que supone el manejo de un número considerablemente menor de parámetros, mientras que se logra una gran selectividad en frecuencia de modo que su uso es adecuado para aplicaciones en las que se dispongan de pocos datos de partida.

Es debido a la primera capa que implementa estos filtros el hecho de que permite una convergencia de la red mucho más rápida frente a otras opciones debido a la adaptabilidad que presenta la red a los datos disponibles ya que estos filtros tienen en cuenta la naturaleza de las señales de entrada. Además, la forma simétrica de los filtros permiten un ahorro del 50 % de costes computacionales en el cálculo de las convoluciones [36].

En su concepción inicial propuesta en [36], la arquitectura SincNet sólo permitía procesar señales de audio monocanal. Esto se debe a que la red neuronal SincNet se creó con el fin de tratar audio sin procesar para tareas de reconocimiento de locutor y de reconocimiento de voz. El resultado del estudio fue la creación de SincNet, una red capaz de realizar una extracción de las características de las señales de audio provenientes de un locutor gracias a la inclusión de filtros significativos de y eficientes que tienen en cuenta las frecuencias de la voz humana. En [36] se comentan las tasas de error de clasificación obtenidas para ambas tareas. Para ello entrenaron tanto una CNN estándar como la SincNet con 12-15 segundos de datos de cada locutor, teniendo que para la tarea de reconocimiento de locutor la SincNet era capaz de obtener una menor tasa de error de clasificación tanto

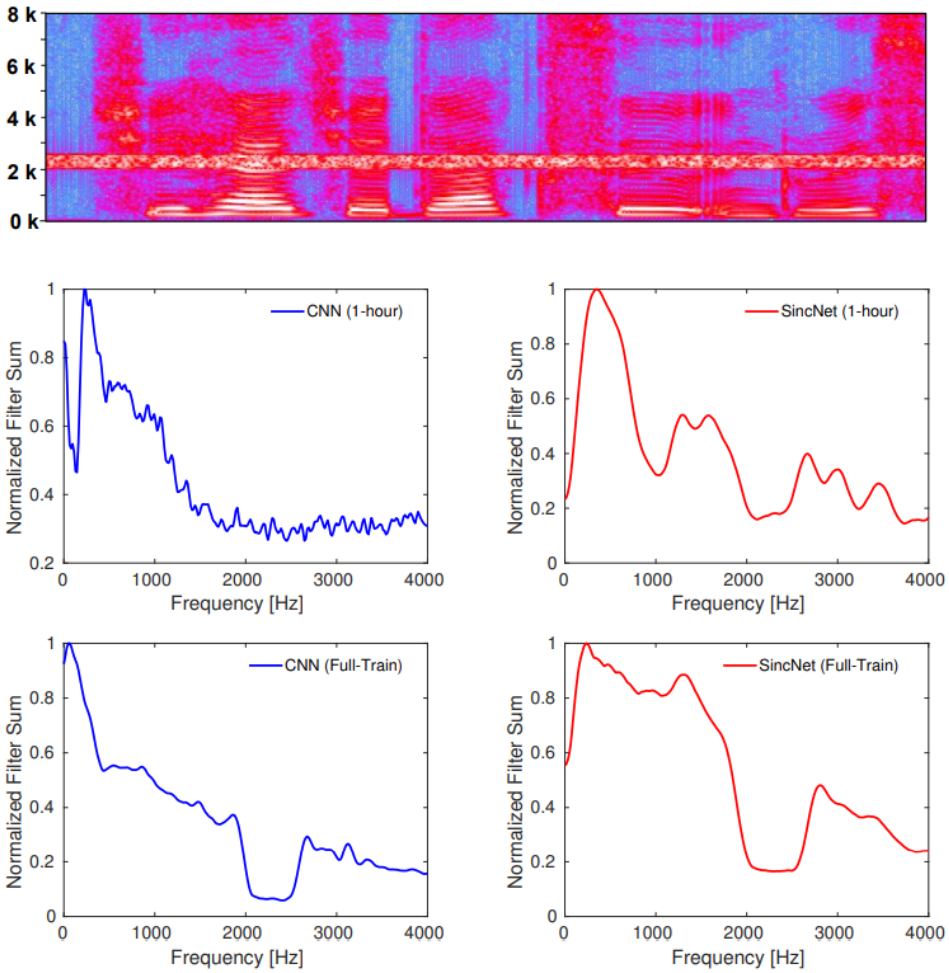

Figura 3.3: Filtrado realizado por una red CNN (izquierda) y por la red SincNet (derecha) tras diferentes tiempos de entrenamiento sobre una señal de voz con ruido blanco introducido artificialmente [36].

para las bases de datos TIMIT y LibriSpeech frente al uso de otras redes convencionales, por lo que tareas donde se disponen de pocos datos SincNet presenta una ventaja de implementación. Para tareas de reconocimiento de voz SincNet también ha arrojado mejores resultados que redes CNN estándar en ambientes ruidosos.

En este proyecto se pretende utilizar la capa de convolución SincNet para trabajar con señales multicanal, en este caso con señales EEG compuestas por numerosos canales, por lo que será necesario realizar una adaptación para permitir el procesamiento de señales formadas por mas de un canal.

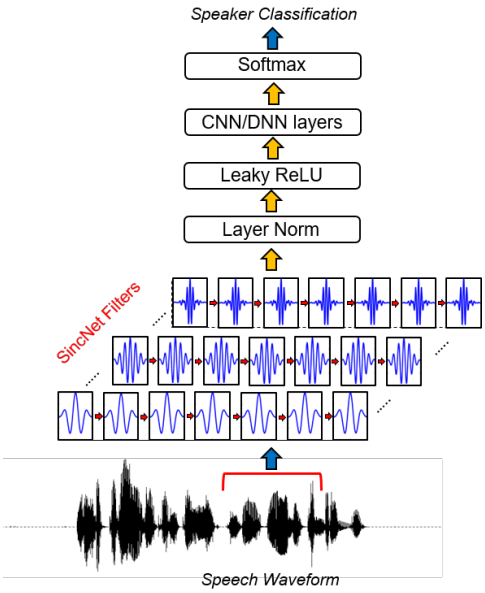

Figura 3.4: Estructura interna de la red neuronal SincNet [36].

3.2.1. Estructura de la red neuronal propuesta

En esta sección se detalla la arquitectura de red neuronal propuesta para realizar la síntesis de los MFCC. Puesto que SincNet incluye una capa convolucional para la extracción de características como se ha comentado anteriormente, es interesante combinarla con diferentes capas que ayuden a realizar la síntesis propuesta. Por ello se ha decidir combinar la primera capa convolucional de la red SincNet con capas convolucionales y una capa totalmente conectada (FC por sus siglas en inglés).

Se pretenden analizar ventanas temporales de N muestras con K canales de EEG, por lo que se debe de adaptar la capa convolucional SincNet para realizar convoluciones bidimensionales en lugar de unidimensionales (puesto que originalmente esta capa estaba diseñada para procesar señales de audio).

Para crear la red neuronal se ha partido del estudio realizado por Srinivasan y otros autores en [38] donde proponen un tipo de estructura de red neuronal en la que incluyen una capa de convolución SincNet adaptada y orientada al uso sobre señales de EEG con varios canales. Los modelos de red neuronal utilizados en este estudio están disponibles en [39]. En la figura 3.5 se puede observar la estructura de la red neuronal propuesta en [38] para el análisis de las señales de EEG.

Esta red neuronal tiene como objetivo la decodificación de series temporales de datos de EEG para tareas de clasificación. Concretamente en dicho estudio se somete a los pacientes a una tarea de decisión binaria frente a un estímulo, en particular discriminar si un parche de Gabor tenía una frecuen-

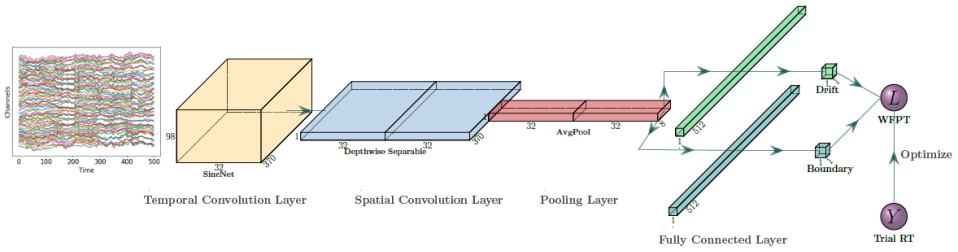

Figura 3.5: Estructura de la red red neuronal tipo SincNET usada en nuestro trabajo para la síntesis de voz a partir de señales de EEG [38].

cia espacial más alta o más baja [38]. Para ello en el estudio se utiliza el modelo de deriva-difusión o DDM, el cual se basa en inicializar una variable de decisión, la cual se actualiza en función de las evidencias sensoriales a favor de una de las dos posibles decisiones. Dicho estudio pretende vincular los procesos de elección con unos parámetros, concretamente con la tasa de deriva δ (la cual expresa el aumento promedio del cambio al acumular las evidencias de decisión) y los parámetros de límite α (los cuales expresan distancias entre ambas opciones en el proceso de elección o decisión).

Lo interesante de esta red es que contiene capas de convolución basadas en SincNet, utilizadas para que a partir de los datos de EEG sin procesar se puedan obtener aquellos filtros más significativos. Una primera capa de convolución SincNet se aplica sobre muestras en el mismo instante temporal de diferentes canales puesto que el kernel de los filtros de convolución son unidimensionales. A continuación se aplica una nueva capa de convolución espacial, la cual se encarga de realizar la convolución entre muestras de diferentes instantes de tiempo. Seguidamente se aplica una tercera capa, en este caso se trata de una capa de pooling, donde se puede variar la longitud de la ventana de pooling como el tamaño de su paso. Finalmente y a continuación de la capa de pooling se aplican dos capas FC separadas que tienen como entrada las salidas de la capa de pooling y ofrecen como salida los parámetros de deriva δ y límite α.

Puesto que se pretende adaptar la red neuronal para el caso en cuestión (síntesis de coeficientes MFCC), se deben de realizar cambios en función del tipo de datos que se deben analizar. A continuación se indican los cambios generales realizados sobre la estructura de la red neuronal:

- Puesto que únicamente se pretende realizar la predicción de un conjunto de parámetros (25 MFCCs), se elimina el procesado de la señal realizado para la obtención de los parámetros δ y α y se sustituye por el procesado para un único parámetro; los coeficientes MFCC.

- Se eliminan las capas de clasificación, sustituyéndolas por una capa

FC tras la capa de pooling con un número de salidas igual al número de MFCC correspondientes a una trama.

- Debido a la distinta naturaleza de los datos utilizados, se deben de ajustar los parámetros propios de la red neuronal como el tamaño de la ventana temporal analizada o las frecuencias de muestreo por ejemplo. Estas particularidades se detallan más adelante en el capítulo 4.

3.3. Síntesis de la voz empleando la red neuronal propuesta y el vocoder WORLD

Diseñada la red neuronal, debe entrenarse desde el principio haciendo uso de los datos disponibles para adaptarla a los mismos y por tanto capacitarla para realizar la síntesis de los coeficientes MFCC a partir de los datos de entrada. Teniendo en cuenta que los coeficientes MFCC de los audios originales se obtienen cada 10 milisegundos (obtenidos mediante el uso del vocoder WORLD explicado en la sección 2.6), y teniendo en cuenta las frecuencias de muestreo de los datos (en el caso de las señales EEG la frecuencia de muestreo es de 512 Hz para un paciente y 256 para el otro que forma la base de datos), para el procesado de de las señales cerebrales deben cogerse ventanas de muestras de un cierto tamaño, dependiendo de cada frecuencia de muestreo. Estas ventanas son bidimensionales puesto que se tienen un número determinado de canales. Estas ventanas se procesan cada 10 milisegundos, por lo que se deben apilar para formar posteriormente el dataset completo. También deben de apilarse los datos de salida de la red, es decir, los coeficientes MFCC.

Una vez creado el dataset, y teniendo en cuenta que se utiliza validación cruzada por el número reducido de datos disponibles se divide el dataset en conjuntos de entrenamiento, validación y test. Teniendo en cuenta que se han establecido 10 iteraciones para la validación cruzada, en cada una de las iteraciones se establecen un 10 % de los datos disponibles para test mientras que el 90 % restantes quedan para el entrenamiento. De ese 90 %, para validación se establece el 10 %, por lo que finalmente en cada iteración se establece el 10 % para test, el 9 % para validación y el 81 % para entrenamiento de la red. De esta forma se tiene que en cada iteración se realizará la predicción de un 10 % de los datos disponibles.

Se ha optado por una función de coste basada en el error cuadrático medio y como método de optimización se ha escogido el optimizador Adam para la actualización de los parámetros de la red neuronal. Adicionalmente se ha incluido el uso de la técnica de early stopping para evitar el overfitting o sobreajuste de los pesos de la red neuronal. Se han establecido 100 épocas de procesamiento de los datos de entrenamiento y evaluación antes de realizar la predicción de los datos de test. La paciencia se ha establecido en 8 épocas;

esto quiere decir que en el caso de que en 8 épocas sucesivas las pérdidas del conjunto de validación no disminuyan, se cargan los pesos de la red de la época en la que se haya conseguido una menor pérdida.

Obtenidos todos los coeficientes MFCC de todos los audios correspondientes, se hace uso del vocoder WORLD para realizar la síntesis de los audios predichos.

El uso del vocoder WORLD posibilita el entrenamiento de una red neuronal, ya que es usado para obtener los coeficientes MFCC de las señales de voz originales en pasos de 10 milisegundos, y de estar forma poder entrenar la red neuronal. Una vez entrenada y obtenidos los coeficientes MFCC correspondientes a cada señal de audio, usando el vocoder se puede realizar la síntesis de las señales de audio correspondientes a las predicciones realizadas. La ventaja del vocoder es que permite almacenar los parámetros adicionales de las señales de voz, por lo que usando únicamente los coeficientes predichos se puede sintetizar la voz en modo whispered, aunque si se almacenan los parámetros adicionales (frecuencias fundamentales, espectrogramas suavizados y matrices de aperiodicidades), es posible mejorar la calidad del audio sintetizado.

En cuando a su implementación en Python, el vocoder posee 4 funciones principales encargadas de realizar la parametrización de las señales de voz de entrada (2 de ellas se encargan de obtener los MFCC, y resto de parámetros comentados a partir de las señales de audio) y de realizar la operación inversa (es decir, de obtener las señales de voz correspondientes a los coeficientes MFCC de entrada con la posibilidad de proporcionar el resto de parámetros para evitar la síntesis de voz whispered), de forma que una vez obtenidas las predicciones con el modelo de red propuesto, se volverá a hacer uso del vocoder para obtener las señales de audio predichas. En la figura 3.6 se puede observar un ejemplo de señales sintetizadas con el vocoder WORLD a partir de los coeficientes MFCC predichos por el método propuesto a partir de señales de PMA.

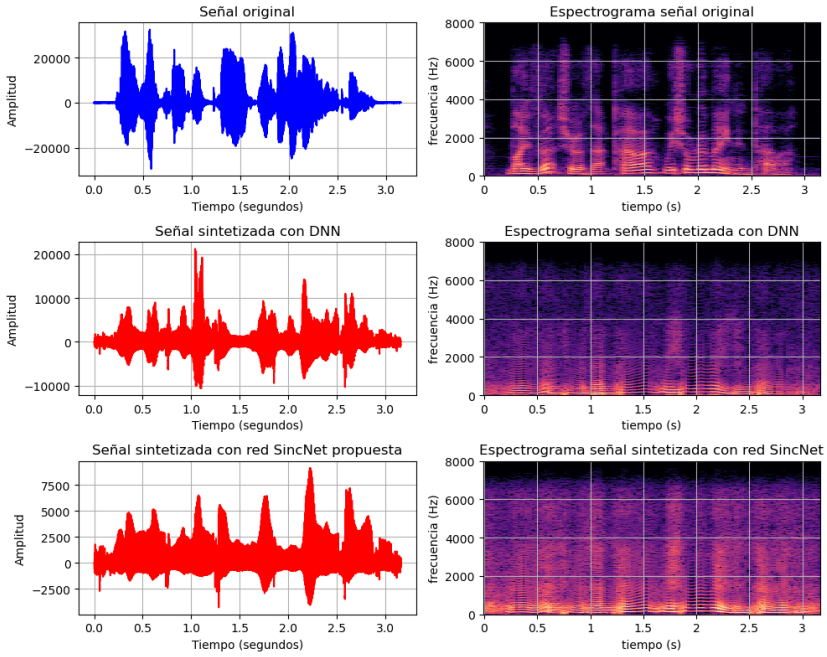

Figura 3.6: Ejemplo de señal original y señales sintetizadas a partir de los coeficientes MFCC haciendo uso del vocoder WORLD.

Capítulo 4

Resultados obtenidos

El objetivo de este capítulo es evaluar de forma experimental el método de síntesis de voz a partir de señales de EEG propuesta en el capítulo 3. Para ello se parte de una base de datos con grabaciones de voz y EEG realizadas a pacientes epilépticos implantados con electrodos profundos mientras leen en voz alta una serie de palabras. Además de esta base de datos, también se evalúa nuestro método en una segunda base de datos con grabaciones de sujetos ingleses realizadas con una técnica conocida como Permanent Magnet Articulography (PMA), que captura el movimiento de los órganos del habla mediante imanes permanentes (ver detalles en sección 4.1.1). Para evaluar nuestro método, se han usado tanto medidas objetivos de la calidad e inteligibilidad de la voz resultante, como medidas subjetivas obtenidas mediante test de escucha por parte de oyentes humanos.

4.1. Marco experimental

En esta sección se indican los detalles de la implementación realizada del método propuesto en el capítulo 3, concretamente las bases de datos utilizadas para la evaluación del método (estructura y características de las mismas) o la configuración utilizada para la red neuronal propuesta.

4.1.1. Bases de datos utilizadas

A continuación se comentan las bases de datos utilizadas para la evaluación del método propuesto, indicando la estructura y características de las mismas.

Base de datos de señales de EEG

Esta base de datos utilizada contiene señales de audio y de EEG sincronizadas de 2 pacientes diferentes. Las señales de audio están tomadas a una frecuencia de muestreo de 44,1 kHz, mientras que las señales de EEG están

tomadas a 512 Hz (paciente M11) y 256 Hz (paciente F09). Las señales EEG han sido tomadas con un numero de canales determinado (cuyo número depende de cada uno de los pacientes) ubicados en diferentes partes del cuero cabelludo. Cada una de las señales de audio tiene una duración de aproximadamente 2 segundos. Las señales de EEG son tomadas de forma síncrona con las señales de voz, y se registran los datos tomados por cada uno de los canales.

Para este estudio se reclutarán pacientes con epilepsia fármaco-resistente implantados con electrodos intracraneales invasivos ingresados en el Hospital Universitario Virgen de las Nieves (HUVN) de Granada, como parte del proyecto estatal 'Voice restoration with brain-computer interfaces' [41]. Estos pacientes presentan crisis epilépticas muy frecuentes que les limitan su vida diaria. Para identificar y delimitar el foco epileptógeno se usan distintas pruebas, entre ellas la hospitalización en una unidad de Video-EEG en la que se registra el vídeo del paciente junto con su electroencefalograma durante 5 días aproximadamente. Debido a la mayor precisión temporal y espacial que aportan, se suelen usar técnicas invasivas como los electrodos profundos para registrar el electroencefalograma del paciente.

Se ha registrado una base de datos que contiene EEG intracraneal y audio para 2 pacientes epilépticos con electrodos de EEG intracraneal implantados mientras realizaban una tarea de producción de pseudopalabras de lenguaje. Las pseudopalabras se forman combinando las 5 vocales en español con las consonantes p, m, f, t, n, rr, s, l, k y j. Se escogieron esas 10 consonantes para tener una cobertura de los distintos lugares de articulación en español, pero sin saturar al paciente con muchas palabras. El corpus consta de 50 pseudopalabras siguiendo el patrón vocal-consonante-vocal como se muestra en la tabla 4.1

Combinación de letras utilizada									
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

Tabla 4.1: Pseudopalabras utilizadas en la producción de lenguaje para la formación de la base de datos de EEG.

A continuación se muestran las características de los pacientes partícipes en la elaboración de la base de datos:

Base de datos de PMA

Aunque el objetivo de este proyecto es la síntesis de voz a partir de señales de EEG, para depurar el sistema de síntesis de voz se usan también

Sujeto	**Sexo**	**Nº de archivos**	**Duración total estimada (minutos)**
F09	Femenino	263	7
M11	Masculino	336	9

Tabla 4.2: Características de la base de datos de EEG.

datos de señales de PMA obtenidos de estudios previos [42], [43]. Dichos datos contienen información del campo magnético generado por unos imanes implantados en varias zonas de la boca. Los datos de los sensores están muestreados a una frecuencia de 100 Hz, por lo que se tiene una medida del campo magnético generado por los mismos cada 10 milisegundos. Además se dispone de las señales de audio generadas por los participantes, muestreadas a 16 kHz. Respecto a los datos de los sensores, se tienen 9 canales diferentes.

Para esta base de datos se disponen de datos de 4 participantes, 3 hombres y 1 mujer, dos de ellos mientras leían en voz alta secuencias de dígitos en inglés extraídos del corpus TiDigit [42] y otros dos mientras leían en voz alta frases fonéticamente balanceadas extraídas del corpus en inglés Arctic [43].

	Sujeto	**Nº archivos**	**Duración total estimada (minutos)**
TiDigit	LC	308	8
	TP	308	8
Arctic	JG	470	19
	RM	509	21

Tabla 4.3: Características de la base de datos de PMA.

Las señales de PMA contienen información acerca de la suma de los campos magnéticos generados por unos imanes implantados en distintas partes de la boca, de forma que la persona que los posee genera un cierto campo magnético al realiza los movimientos articulatorios de la boca y lengua. De esta forma se tiene que los movimientos articulatorios quedan registrados por medio de las sumas de los campos magnéticos generados por los imanes.

La ventaja de dicha técnica radica en la facilidad en la implantación del sistema de registro en las personas, además de su portabilidad, por otro lado el uso de esta técnica no proporciona una alta correlación de las señales obtenidas con la posición y el desplazamiento de los imanes debido a que lo que se registra como se ha comentado se trata de la suma de los campos magnéticos.

En este caso las señales PMA se han obtenido de 9 canales diferentes, con una frecuencia de muestreo de 100 Hz para las señales PMA y de 16 kHz para las señales de audio grabadas de forma sincronizada.

El único preprocesado realizado a los datos de PMA es el eliminado de la

Figura 4.1: Ejemplo de obtención de señales PMA [43].

contribución terrestre a las medidas de las sumas de los campos magnéticos. Para ello uno de los sensores durante la tarea de adquisición de las señales se dedicó a realizar una medida de la contribución de dicho campo magnético terrestre. Con ello se pudo paliar dicha contribución.

4.1.2. Preprocesado de los datos

Para cada base de datos se tienen unos detalles de implementación específicos debido a la propia naturaleza de las señales analizadas. El hecho de tener diferentes frecuencias de muestreo limitan en gran medida el número de ventanas de muestras que se pueden analizar de forma conjunta, lo cual hace que algunos parámetros de la red neuronal deban cambiar para así posibilitar el entrenamiento de las mismas. Por ello a continuación se van a indicar las particularidades de cada base de datos:

Procesado de las señales de PMA

Se tiene que las señales de PMA tienen 9 canales. También se tiene que la frecuencia de muestreo es de 100 Hz por lo que se tiene una muestra cada 10 milisegundos y que los coeficientes MFCC se calculan cada 10 milisegundos, por lo que se concluye que a una ventana de tamaño 1x9, es decir a una muestra de cada canal le corresponde una ventana de 1x25 coeficientes MFCC. Esto es importante de cara a configurar los datasets de las redes neuronales. Con el fin de comparar resultados, esta base de datos se procesa tanto con una red neuronal DNN sencilla como con la red neuronal propuesta en el capítulo 3.

Por ello para el caso de análisis de las bases de datos TiDigit y Arctic

mediante una red neuronal DNN, se realiza un enventanado de las señales, se aplica PCA y se realiza una normalización de los datos de entrada (señales PMA) y de los datos de salida correspondientes (coeficientes MFCC de las señales de audio originales) tal y como se comenta en la sección 3.1.3. Tras aplicar PCA, se ha comprobado que se tienen mejores resultados para un total de 25 componentes, por lo que finalmente se tiene que una ventana de tamaño 1x25 corresponde a una fila de 1x25 coeficientes MFCC.

Por otro lado en caso de analizar las bases de datos TiDigit y Arctic mediante la red SincNet, no se aplica PCA ni se realiza normalización de los datos de entrada de la red (señales PMA), pero sí de los coeficientes MFCC de las señales de audio originales correspondientes (tras la síntesis de los MFCC predichos se realiza la desnormalización de los mismos). Previamente a la normalización se realiza un enventanado de las señales PMA de entrada para que en el proceso de entrenamiento de la red se mantenga el contexto temporal de la señales. Se ha establecido un número total de 10 ventanas de muestras para el contexto temporal, 5 ventanas anteriores y 5 posteriores, de forma que se mantenga el contexto temporal correspondiente a 100 milisegundos de señal. De esta forma a cada ventana de 10x9 muestras le corresponde una ventana de 1x25 coeficientes MFCC.

Procesado de las señales de EEG

En este caso se tiene que las señales de la base de datos de EEG tienen 77 canales en el caso del sujeto M11 y 90 en el caso del sujeto F11. La frecuencia de muestreo de las mismas es de 512 Hz y 256 Hz, respectivamente. Dado que se desean procesar tramas temporales de 50 milisegundos, a cada ventana de 1x25 coeficientes MFCC de la señal de voz original le corresponden aproximadamente 26 muestras de EEG de cada uno de los canales en el caso del paciente M11 y 13 muestras para el caso del paciente F09. Para el caso de utilizar la red neuronal propuesta, no se realiza normalización de los datos de entrada (señales EEG) pero si se realiza normalización de los datos de salida (coeficientes MFCC de la señal de voz original asociados), para posteriormente desnormalizar los coeficientes MFCC predichos.

Puesto que las señales de audio están muestreadas a 44,1 kHz, se deben de diezmar como se comenta en la sección 3.1.1 para establecer una frecuencia de muestreo de 16 kHz.

4.1.3. Implementación de modelos

Implementación de la red neuronal propuesta basada en SincNet

Para entrenar y usar la red neuronal propuesta se deben de establecer unos valores a diversos parámetros los cuales pueden permitir unas mejores o peores predicciones. En primer lugar se define el número de filtros que se van a aplicar para analizar las señales. Un mayor número de filtros implica

una mayor complejidad de la capa que implementa las convoluciones basadas en SincNet y por tanto de la red neuronal. Por otro lado la longitud de los filtros está relacionada con la mínima frecuencia central posible para los filtros, y se debe escoger teniendo en cuenta el número de canales a analizar y las muestras por canal en cada ventana de análisis. Los tamaños de las ventanas de pooling y el paso de las mismas es importante, puesto que la ventana de pooling permite conservar la característica más relevante de una ventana determinada de N milisegundos con pasos de K milisegundos (depende del caso). Por otro lado el número de canales es diferente para cada base de datos y es un parámetro que debe ajustarse, así como la frecuencia de muestreo.

Para la base de datos de EEG el número de filtros se ha establecido en 40, mientras que la longitud de los mismos se ha establecido en 19 muestras para el sujeto M11 y en 11 para el sujeto F09 debido a que se tienen 77 canales para el sujeto M11 y 90 para el sujeto F09 (recordar que la longitud de los filtros debe ser menor que el número de canales y de muestras por canal analizadas). En cada ventana se analizan fragmentos de 50 ms de señal, correspondientes a 26 muestras para el paciente M11 y 13 muestras para el paciente F09 debido a que las señales de EEG tienen frecuencias de muestreo de 512 Hz y 256 Hz respectivamente. Se ha establecido una longitud para la ventana de pooling de 4 milisegundos con un paso de 2 milisegundos. Las frecuencias analizadas por los filtros SincNet van desde los 0 a los 200 hercios para ambas bases de datos.

Por su parte para la base de datos de PMA el número de filtros se ha mantenido en 40 pero su longitud se ha reducido a 7 puesto que se tiene un menor número de canales y de muestras por canal. Las ventanas escogidas para el análisis tienen una longitud de 100 ms, lo que equivale a 10 muestras ya que la frecuencia de muestreo es de 100 Hz. Las ventanas de pooling y su paso tienen valores de 40 y 10 milisegundos, respectivamente.

Además respecto al script diseñado para el entrenamiento de la red neuronal, se debe establecer un tamaño de batch para, una vez construido el dataset, procesar los datos en pequeños lotes. Se ha escogido un tamaño de bath de 64, es decir que para el caso de las señales PMA procesadas con la DNN, se introducen conjuntos de 64 muestras por 25 componentes a la entrada de la red (a la que le corresponden 64 ventanas de 25 MFCC cada una de ellas). En el caso de procesar dichos datos con la red SincNet propuesta de nuevo el tamaño de batch es de 64, aunque en este caso se procesan 64 ventanas de 10x9 muestras antes de realizar un ajuste de los pesos de la red.

Mientras tanto puesto que el tamaño de los mini-lotes o batches es de 64, en el caso de las señales de EEG se procesan 64 ventanas de 23x77 muestras para el paciente M11 (o 13x90 para el paciente F09) antes de actualizar los pesos de la red.

A continuación en la tabla 4.4 se puede observar el número de parámetros utilizados para el caso de las señales de EEG, concretamente para el

sujeto M11 y un esquema de la red neuronal propuesta para el procesado de las mismas en la figura 4.2. Puesto que los parámetros cambian para el procesado de las señales de PMA (y para diferentes sujetos de la base de datos de EEG), varían los tamaños de las entradas y salidas, aunque la estructura de la red es análoga. Para el caso del paciente M11, la red neuronal configurada cuenta con un total de 9347 parámetros entrenables.

Bloque	**Capa**	**Filtros**	**Tamaño**	**Número de parámetros**	**Salida**	**Activación**
1	Input	-	-	-	(77,26)	
	BatchNorm	-	-	2	(77,26)	
	SincConv2D	40	(1,11)	80	(40, 77, 8)	
2	BatchNorm	-	-	80	(40, 77, 8)	
	SeparableConv2D	40	(77,1)	3080	(40,1,8)	ReLU
3	BatchNorm	-	-	80	(40,1,8)	
	AvgPool2D	-	(1,2)	-	(40,1,6)	
	Dropout	-	-	-	(40,1,6)	
4	FC	-	-	6025	25	

Tabla 4.4: Arquitectura de la red SincNet para el paciente M11 de la base de datos de EEG.

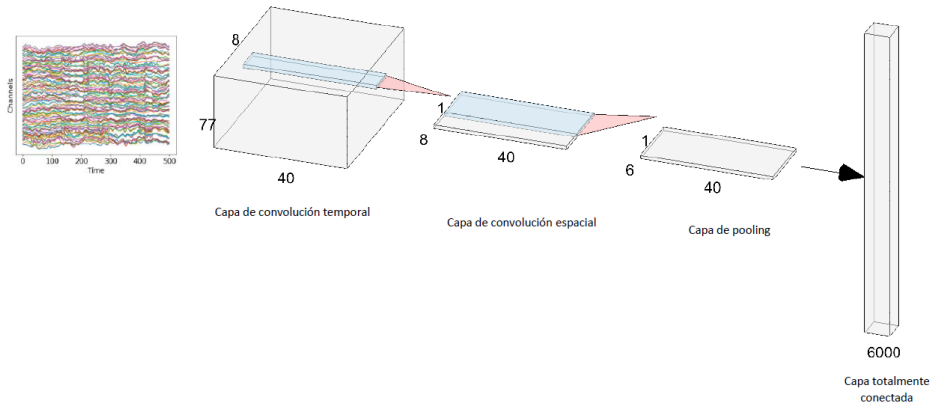

Figura 4.2: Red neuronal diseñada para el sujeto M11 de la base de datos de EEG.

Implementación de la red DNN

Para el caso de síntesis haciendo uso de la red DNN, se han extraído las características high-gamma de las señales de EEG de ventanas temporales de 50 milisegundos con desplazamiento de 10 milisegundos. La red DNN toma diferentes formas dependiendo del tipo de dato analizado, de forma que su estructura es diferente para el caso de análisis de señales EEG y PMA.

Para el caso de análisis de señales de EEG, la red cuenta con 150 entradas para el caso de paciente M11 (cuyas señales de EEG cuentan con 77 canales) y para el caso del paciente F09 la red cuenta con 170 entradas (ya que tiene sus señales cuentan con 90 canales). Se han determinado ese número de entradas ya que al aplicar PCA, ese número de canales presenta mejores resultados al mantener la mayor parte de la información (varianza) de las señales. Se ha establecido como función de coste la función MSE y se ha utilizado el optimizador Adam para el ajuste de los pesos de la red. A continuación en la tabla 4.5 se puede observar el número de neuronas de cada capa, además del número de parámetros con el que cuenta cada una de ellas para el caso de la red DNN que procesa los datos de EEG del paciente F09, teniendo un total de 14649 parámetros entrenables:

	Numero de neuronas	**Numero de parámetros entrenables**
Capa de entrada	170	-
Capa oculta 1	32	5472
Capa oculta 2	128	4224
Capa oculta 3	32	4128
Capa de salida	25	825

Tabla 4.5: Estructura de la red DNN utilizada.

Por otro lado para el procesado de los datos de PMA, la función de pérdidas y el optimizador y los número de neuronas en cada capa de la red DNN son las mismas, aunque cambian el número de entradas de la red, en este caso se tienen 25 entradas. El tamaño del batch es de 64 (el mismo que para el caso de la red neuronal propuesta). La tasa de aprendizaje está establecida en 0,0005 para todas las bases de datos tanto para la red DNN como para la red neuronal propuesta.

4.1.4. Evaluación y métricas

Para la evaluación de la calidad de las señales de audio predichas y sus parámetros (coeficientes MFCC) se han usado tanto técnicas objetivas como subjetivas.

En el caso del análisis de resultados mediante métricas objetivas, para la evaluación mediante el uso de la métrica MCD se han realizado comparaciones directas entre los coeficientes MFCC originales y sintetizados, de forma que se ha obtenido el valor de la métrica para cada par de coeficientes originales-sintetizados y se ha obtenido la media para cada sujeto. Lo mismo se ha realizado para las comparaciones entre señales de audio originales-sintetizadas en el caso de las métricas STOI y PESQ. Los valores de las métricas obtenidos son en cierto modo orientativos puesto que objetivamente una síntesis puede presentar mejores resultados de MCD que otra,

aunque podría ser que subjetivamente se escuche peor o menos nítido, por lo que los resultados objetivos son orientativos y pueden validarse con los resultados subjetivos.

Para evaluar los resultados de forma subjetiva se ha realizado una escucha para cada caso concreto de síntesis. Además se ha elaborado un test subjetivo con el fin de probar la calidad de los audios sintetizados haciendo escuchas subjetivas y valorando aquellos resultados que proporcionan síntesis más nítidas y parecidas a las originales.

4.2. Resultados obtenidos

En esta sección se muestran los resultados obtenidos de las métricas utilizadas para la evaluación de la síntesis realizada en cada caso. Se presentan los resultados obtenidos tanto para la base de datos de EEG como de la base de datos de PMA tras su procesado con una DNN simple como con la red SincNet propuesta.

4.2.1. Resultados obtenidos con los datos de PMA

En esta sección se exponen los resultados obtenidos al evaluar las métricas MCD y STOI sobre los coeficientes MFCC predichos y los audios sintetizados, respectivamente. Los resultados se muestran tras procesar las bases de datos descritas en el apartado 4.1.1 tanto con una red DNN simple como con la red neuronal propuesta en la sección 3.2.1. Además se muestran las formas de onda en el dominio temporal de las señales de voz original y sintetizada correspondientes, así como sus espectrogramas con el fin de poder realizar comparaciones entre ellas.

En la tabla 4.6 se muestran los resultados de la métrica MCD obtenidos sobre los coeficientes MFCC originales y sintetizados correspondientes para las bases de datos TiDigit y Arctic (la media de las puntuaciones obtenidas para cada base de datos):

	TiDigit			**Arctic**			
	LC	TP	Media	JG	RM	Media	**Media global**
DNN	10,98	10,65	10,82	12,10	11,28	11,69	11,26
SincNet	10,10	10,25	**10,18**	10,91	11,14	**11,03**	**10,61**

Tabla 4.6: Resultados medios de la métrica MCD obtenidos para los datos de PMA para cada red neuronal utilizada.

Se puede observar que el mejor resultado medio de MCD (mínima distorsión) se obtiene al procesar la base de datos TiDigit con la red neuronal

propuesta, mientras que el peor resultado medio (máxima distorsión) se obtiene al procesar la base de datos Arctic con la red DNN simple. En general se tienen peores resultados para la base de datos Arctic debido a que, además de tener un vocabulario mucho más amplio y ser una tarea más compleja, la técnica PMA tienes algunas limitaciones a la hora de capturar la información articulativa de ciertos fonemas [44], lo cual empeora los resultados obtenidos. A pesar de ello el uso de la red SincNet propuesta arroja mejores resultados tanto para la base de datos Arctic como para la base de datos TiDigit. También se observan diferencias entre los resultados obtenidos para diferentes locutores, teniendo que dependiendo de la red utilizada para síntesis se tienen mejores valores para unos pacientes o para otros.

A continuación en la tabla 4.7 se muestran los resultados de la métrica STOI obtenidos sobre las señales originales y sintetizadas correspondientes para las bases de datos TiDigit y Arctic (la media de las puntuaciones obtenidas para cada base de datos):

	TiDigit			**Arctic**			
	LC	TP	Media	JG	RM	Media	**Media global**
DNN	0,60	0,62	**0,61**	0,54	0,53	**0,53**	**0,57**
SincNet	0,56	0,60	0,58	0,53	0,47	0,50	0,54

Tabla 4.7: Resultados medios de la métrica STOI obtenidos para los datos de PMA para cada red neuronal utilizada.

Se puede observar que el mejor resultado de STOI se obtiene al procesar la base de datos TiDigit con la red DNN simple, mientras que el peor resultado se obtiene al procesar la base de datos Arctic con la red propuesta. De nuevo se tienen unos peores resultados para el caso de la base de datos Arctic. En este caso la métrica arroja unos mejores resultados al procesar ambas bases de datos con la red DNN.

A continuación en la figura 4.3 se pueden observar las diferencias entre un ejemplo de señal original y señales sintetizadas pertenecientes a la base de datos TiDigit al realizar un procesado de los datos mediante una red DNN simple y la red SincNet propuesta.

Se puede observar que la forma de onda obtenida con la red neuronal propuesta es más parecida a la señal original. Además se puede visualizar un espectrograma menos ruidoso y más nítido con respecto a la síntesis mediante una DNN simple.

Por último en la figura 4.4 se realiza la misma comparativa para los audios obtenidos tras el procesado de la base de datos Arctic.

La señal contiene una oración hablada formada por varias palabras en idioma inglés. Se puede observar que la forma de onda reconstruida con ambas redes tiene una peor calidad que las predicciones obtenidas a partir

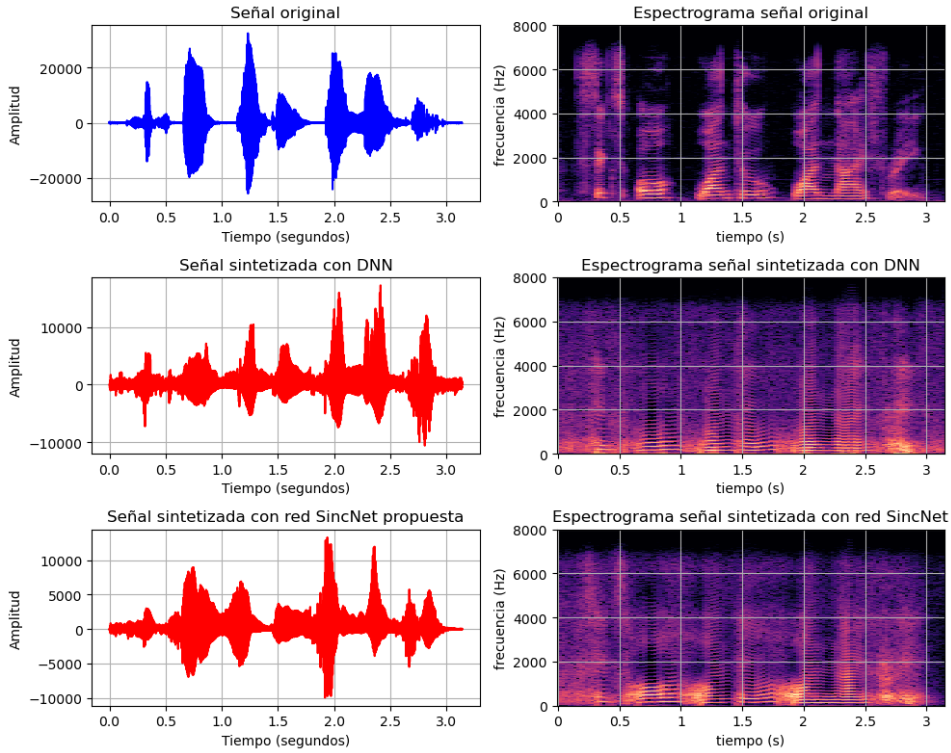

Figura 4.3: Señal original de la base de datos TiDigit junto a su espectrograma (fila superior), señal sintetizada obtenida tras el procesado con una red DNN simple junto a su espectrograma (segunda fila) y señal sintetizada obtenida tras el procesado con la red SincNet propuesta junto a su espectrograma (fila inferior). La señal contiene la sucesión de dígitos 'Six O One Two One Nine Three' (6012193) en idioma inglés.

de la base de datos TiDigit. Por otro lado se puede visualizar de nuevo un espectrograma menos ruidoso y más nítido con respecto a la síntesis mediante una DNN simple.

4.2.2. Resultados obtenidos con los datos de EEG

En esta sección se exponen los resultados obtenidos al evaluar las métricas MCD y STOI sobre los coeficientes MFCC predichos y los audios sintetizados tras procesar la base de datos de EEG descrita en la sección 4.1.1 con la red neuronal propuesta.

En las tablas 4.8 y 4.9 se muestran los resultados de las métricas MCD y STOI obtenidos respectivamente sobre los coeficientes MFCC originales y sintetizados correspondientes y sobre las señales de voz originales y sintetizadas para ambos sujetos:

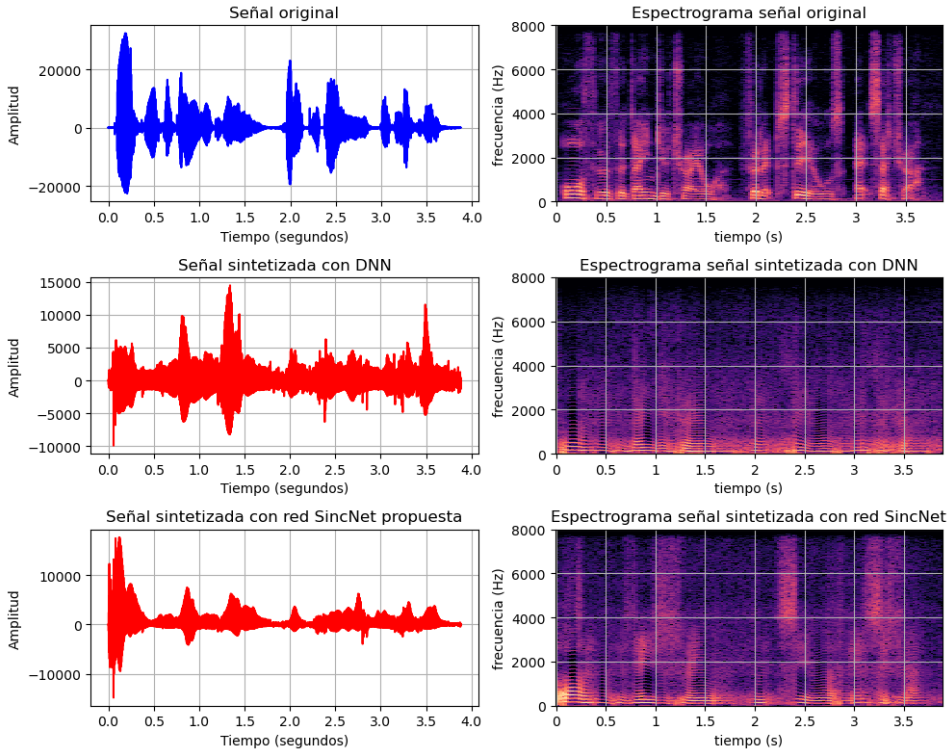

Figura 4.4: Señal original de la base de datos Arctic junto a su espectrograma (fila superior). señal sintetizada obtenida tras el procesado con una red DNN simple junto a su espectrograma (segunda fila) y señal sintetizada obtenida tras el procesado con la red SincNet propuesta junto con su espectrograma (fila inferior). La señal contiene la oración 'Author of The Danger Trail, Philip Steels, etc' en idioma inglés.

	M11	F09	**Media Global**
DNN	7,82	6,45	**7,14**
SincNet	21,61	24,64	23,13

Tabla 4.8: Resultados de la métrica MCD obtenidos para los datos de EEG.

	M11	F09	**Media Global**
DNN	0,007	0,005	**0,006**
SincNet	0,001	0.003	0,002

Tabla 4.9: Resultados de la métrica STOI obtenidos para los datos de EEG.

Se puede observar que los resultados evaluados con ambas métricas son mucho peores en el caso del procesado de la base de datos de EEG que en el

caso del procesado de la base de datos de PMA. Se tiene en este caso un mejor resultados de las métricas para el caso del procesado de los datos con la red DNN, aunque como posteriormente se comenta, a pesar de obtener mejores resultados en las métricas objetivas, la evaluación subjetiva mediante una escucha auditiva de las señales sintetizadas indican que el resultado no es acertado de igual forma que al realizar el procesado con la red diseñada. En las figuras 4.5 y 4.6 se pueden observar las diferencias entre un ejemplo de señal original y las señales sintetizadas mediante cada red neuronal para ambos sujetos de la base de datos:

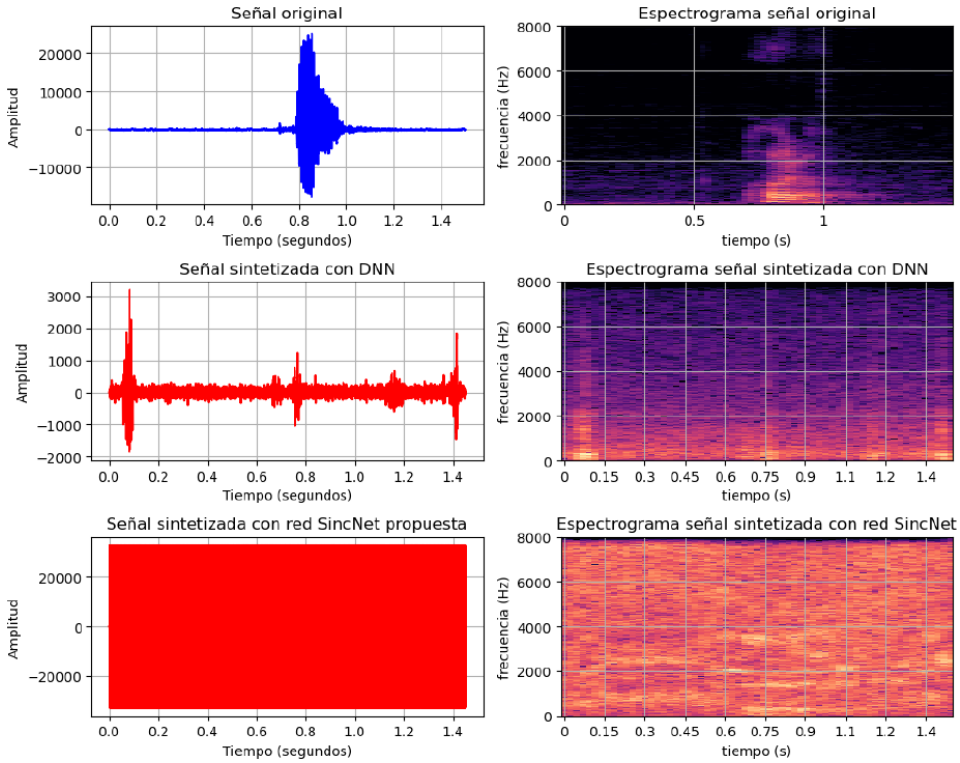

Figura 4.5: Señal original del paciente M11 de la base de datos de EEG junto a su espectrograma (fila superior), señal sintetizada obtenida tras el procesado con una red DNN simple junto a su espectrograma (segunda fila) y señal sintetizada obtenida tras el procesado con la red SincNet propuesta junto con su espectrograma (fila inferior). La señal contiene la pseudopalabra 'IFI' en idioma castellano.

Se puede observar que la forma de onda reconstruida para ambos pacientes es absolutamente ruidosa para el caso de procesado con la red SincNet propuesta, presentando un espectrograma totalmente aleatorio. Esto podría deberse a una falta de muestras en la base de datos utilizada, o a la falta de correlación entre las señales de EEG obtenidas y la información lingüística.

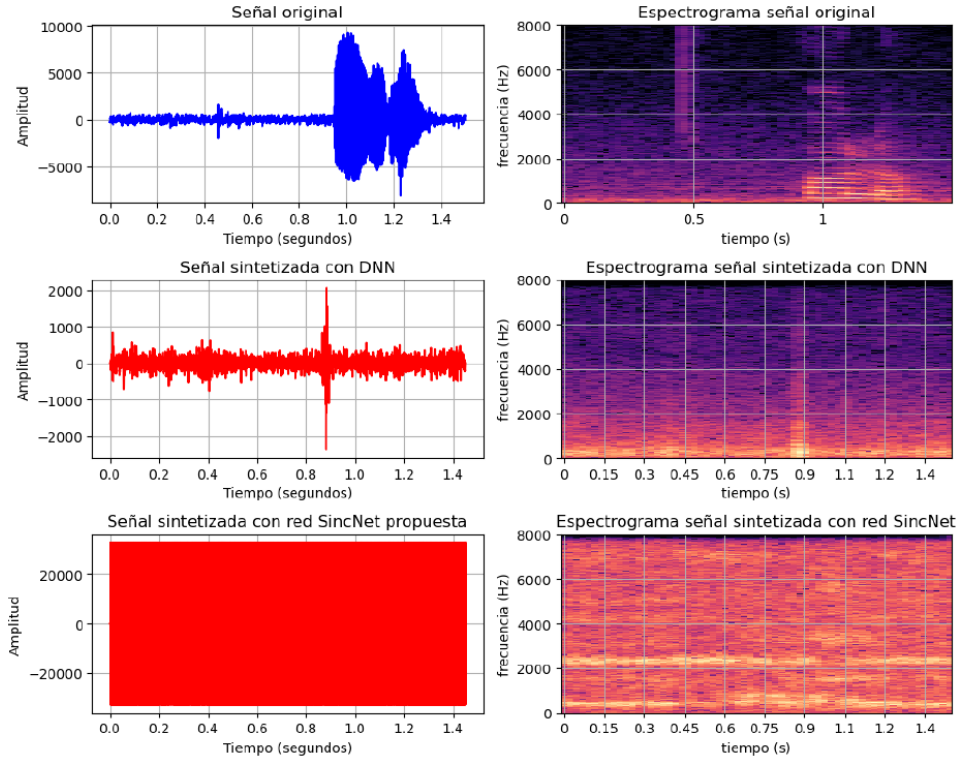

Figura 4.6: Señal original del paciente F09 de la base de datos de EEG junto a su espectrograma (fila superior), señal sintetizada obtenida tras el procesado con una red DNN simple junto a su espectrograma (segunda fila) y señal sintetizada obtenida tras el procesado con la red SincNet propuesta junto con su espectrograma (fila inferior). La señal contiene la pseudopalabra 'OLO' en idioma castellano.

En el caso de las señales sintetizadas con la red DNN no se tiene un espectrograma aleatorio, aunque como posteriormente se comenta en la evaluación subjetiva, las señales sintetizadas con ambas redes tienen un mal parecido acústico a la señal original.

4.3. Discusión

En esta sección se va a realizar una reflexión de los resultados obtenidos y expuestos en las secciones 4.2.1 y 4.2.2.

4.3.1. Evaluación subjetiva

debido a restricciones temporales, se ha realizado una evaluación informal subjetiva de los audios obtenidos, a falta de poder realizar una evalua-

ción subjetiva formal en el futuro. La idea de este apartado es presentar los resultados obtenidos a partir de la escucha por parte de terceras personas de las predicciones, de forma que se pueda arrojar un dato de inteligibilidad humana por parte de personas ajenas al proyecto y así poder valorar la calidad de la síntesis realizada y por tanto del funcionamiento del método propuesto.

A partir de la evaluación informal subjetiva se concluye que los audios obtenidos tras el procesamiento de la base de datos de PMA con la red neuronal propuesta es más nítido con respecto a las señales acústicas generadas tras el procesamiento de los datos con una red DNN simple. Para el caso de la base de datos de EEG, el audio generado no aporta ningún tipo de información audible.

El resultado acústico es el esperado respecto a la señal de voz generada para el caso de las señales PMA. La base de datos TiDigit contiene un número reducido de fonemas con respecto a la base de datos Arctic la cual contiene numerosos fonemas adicionales, lo cual dificulta el aprendizaje al aumentar el diccionario fonético. Por esta razón, los audios sintetizados para la base de datos TiDigit tiene una calidad mucho mayor que los audios sintetizados para la base de datos Arctic en general. A pesar de ello se tiene que comparando el procesado realizado de ambas bases de datos con la red DNN sencilla y la red propuesta, en general se tiene una mayor calidad de síntesis haciendo uso de la red neuronal propuesta. Esta diferencia es más notable en el caso de la base de datos Arctic, donde los audios sintetizados con la red neuronal propuesta resultan mucho más nítidos que los sintetizados con la red DNN.

Respecto al audio de la base de datos de EEG, el hecho de sintetizar el audio en modo whispered (para los datos de PMA se sintetizó teniendo en cuenta parámetros adicionales como matrices de aperiodicidades y espectrogramas) también contribuye a la aparición de ruido constante en el audio, enturbiándolo.

4.3.2. Evaluación objetiva

Objetivamente hablando la red neuronal propuesta es capaz de generar mejores resultados que la red DNN para realizar síntesis de voz a partir de bioseñales, hecho que queda demostrado al observar que tanto la métrica MCD como la evaluación acústica presenta mejores resultados al analizar las señales de PMA (a pesar de que la métrica STOI arroja resultados ligeramente peores). Para el caso de la síntesis de voz a partir de señales EEG haciendo uso de la base de datos indicada en la sección 4.1.1 ha demostrado no ser eficaz, lo cual queda plasmado al analizar las señales de voz y los espectrogramas correspondientes.

Sin embargo el uso de la red neuronal propuesta si que arroja unos buenos resultados en el análisis de las señales de PMA. Si se observan los

espectrogramas, la red neuronal propuesta (junto con el vocoder WORLD) es capaz de generar una señales de voz cuyos espectrogramas son mucho más nítidos para las bases de datos TiDigit y Arctic. Concretamente en bases de datos en las que se tiene una gran diversidad fonética como el caso de la base de datos Arctic, la red neuronal propuesta es capaz de generar audios con una calidad acústica notablemente mejor que para el caso de una red DNN simple. En los resultados se puede observar que para las señales de PMA, la métrica MCD presenta unos mejores resultados al usar la red propuesta. A pesar de que la métrica STOI presente peores resultados, la evaluación subjetiva corrobora el hecho de que se obtienen mejores resultados con la red propuesta.

Respecto a estudios del estado del arte como [42] o [43] se tiene que la métrica MCD obtenida es peor (en estos estudios se tiene un valor de entre 5 y 7 dB), lo cual se puede deber a la diferencia entre métodos utilizados o a la diferencia de vocoders utilizados. Si bien el número de filtros y tamaño de los mismos puede influir a la hora de obtener mejores o peores resultados en las métricas utilizadas, tras numerosos ensayos se han escogido aquellos parámetros que proporcionan un mejor valor de las métricas.

Las métricas arrojan muy malos resultados para los análisis de las señales de EEG. Una de las principales causas puede ser que los electrodos en la configuración de EEG usada pueden no estar localizados en las mejores áreas para capturar información sobre el lenguaje debido a que la localización de estos electrodos se determinó atendiendo a razones puramente clínicas). Este hecho ligado a la falta de datos para procesamiento y la fuerte no-linealidad que presentan las señales de EEG con el proceso de producción de voz pueden explicar los resultados obtenidos al realizar la síntesis de la voz a partir de las señales de EEG.

Capítulo 5

Conclusiones y vías futuras

El objetivo principal del proyecto era el demostrar la viabilidad del diseño de un sistema de conversión de señales de EEG obtenidas por medios invasivos (electrodos profundos) a voz. Para ello se exploró el uso de una arquitectura de red neuronal del estado del arte, denominada SincNet, para implementar la conversión extremo a extremo para la conversión de señales EEG sin procesar en señales de voz. Para ello se hizo una adaptación de la red SincNet por las ventajas que conllevaba su uso, concretamente la extracción de características más significativas, el uso de una cantidad de parámetros reducida en comparación con otras redes neuronales lo cual permite una mayor rapidez de convergencia de la red y la posibilidad de poder procesar ventanas temporales de muestras en crudo, es decir sin procesar.

Para poder realizar una evaluación del método propuesto en el capítulo 3, se decidió realizar una comparativa de resultados de las métricas seleccionadas con diferentes bases de datos haciendo uso tanto de la red propuesta como de una red DNN simple de forma que se tuviera un método base que sirviera de referencia para poder comparar resultados. Concretamente se ha utilizado una base de datos de señales de PMA para comprobar el correcto funcionamiento de la red diseñada, la cual está compuesta por señales provenientes de 4 pacientes diferentes, y por otro lado se ha utilizado una base de datos de EEG formada por señales procedentes de 2 pacientes del Hospital Universitario Virgen de las Nieves de Granada para evaluar el propósito de la red neuronal propuesta.

Concretamente las evaluaciones mediante la métrica objetiva MCD y la evaluación subjetiva nos aportaron información acerca del funcionamiento del modelo propuesto en el procesado de señales de PMA; resultados de MCD igual a 10,18 dB y 11,03 dB para las bases de datos TiDigit y Arctic respectivamente, aproximadamente 0,7 dB menos en ambos casos que mediante el uso de una red DNN sencilla. A pesar de que la métrica STOI aplicada sobre los audios originales y sintetizados tuvieron un resultado peor para la red propuesta, como se ha comentado, la evaluación acústica permi-

te corroborar el hecho de que la red propuesta funciona mejor que una red DNN simple con los datos que se tienen. A pesar de ello, el resultado no es satisfactorio para la tarea de la decodificación de las señales de EEG, donde la métrica MCD toma un valor de 21,55 dB y 24,65 dB para dos pacientes diferentes y donde los audios predichos son ininteligible.

Se concluye que no se tienen buenos resultados en la tarea de síntesis de voz a partir de señales de EEG para la base de datos en particular que se ha utilizado, aunque resulta interesante probar la red neuronal diseñada en otra base de datos con más grabaciones, más electrodos y/o con diferentes localizaciones. A pesar de ello, la red si que funciona para señales de distinta naturaleza como el caso de las señales PMA (y con mejores resultados comparado con una red DNN simple). Estos resultados en parte pueden deberse a las limitaciones con respecto a la base de datos de EEG; la fuerte no-linealidad de las señales EEG y la producción del habla podría conllevar la necesidad de bases de datos sustancialmente mayores para la obtención de resultados de síntesis aceptables. Además de ello, el tiempo de ejecución de esta red es muy elevado, por lo que el incremento de muestras de la base de datos utilizada para el entrenamiento de la misma conlleva la necesidad de una cantidad de recursos hardware muy superiores. Ello conllevaría también la necesidad de incrementar el número de parámetros de la red y por tanto la complejidad de la misma.

Por otro lado el método precisa de una desventaja, precisamente en la propia elaboración de la base de datos. Normalmente la adquisición de las señales EEG está condicionada respecto al ámbito médico, lo cual hace que las zonas de donde se extraen las señales EEG pueden no ser las mismas o el número de canales para la extracción de las señales puede ser diferente. Este hecho limita en gran parte la posibilidad de crear un conjunto de datos relativamente grande.

En cuanto a las líneas futuras de investigación es interesante la implementación de un método de extracción de señales EEG que facilite el entrenamiento de redes neuronales con datos provenientes de sujetos diferentes de forma que las bases de datos para el entrenamiento puedan ser sustancialmente mayores para asegurar un número suficiente de muestras. Además es interesante la búsqueda de nuevos algoritmos que permitan modelar la correlación existente entre las señales EEG y la producción de voz de forma que se pueda trabajar sobre ello para diseñar nuevos algoritmos basados en redes neuronales adaptados a señales EEG. La notable no-linealidad entre las señales de EEG en crudo y el proceso de producción del habla implica la necesidad de búsqueda de nuevos métodos para la extracción de características de las señales EEG. La misma tarea (síntesis de voz a partir de señales EEG) puede implementarse haciendo uso de nuevas estructuras de redes neuronales o haciendo uso de capas que lleven a cabo una extracción de características con mejores resultados. Es interesante el uso de redes neuronales para esta tarea puesto que una vez entrenada, pueden usarse en

dispositivos de bajo coste, lo cual podría permitir a personas con ELA y otras afecciones cerebrales mejorar de forma drástica su calidad de vida. También es interesante incluir en futuros trabajos una evaluación subjetiva exhaustiva (la cual se ha omitido por falta de tiempo) para poder corroborar la escucha informal realizada y por tanto someter a diversas personas a un test de inteligibilidad de la síntesis de voz obtenida en cada caso.

Apéndice A

Planificación y presupuesto

A.1. Planificación temporal del proyecto

A continuación se presenta en la tabla A.1 la temporización del proyecto, el tiempo dedicado a cada una de las tareas llevadas a cabo para el desarrollo y finalización del proyecto y las fechas en las que se ha llevado a cabo cada una de las tareas indicadas.

	Mes							
Tarea	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre
Elección del TFG								
Recopilación y lectura de la bibliografía								
Aprendizaje del lenguaje Python								
Preprocesado de los datos								
Diseño y ajuste de la red neuronal								
Evaluación de los resultados obtenidos								
Redacción de la memoria								
Corrección de la memoria								

Tabla A.1: Diagrama de Gantt.

A.2. Presupuesto económico del proyecto

El proyecto se ha desarrollado durante 8 meses de acuerdo a la planificación del proyecto mostrada en el apéndice A.1. De todo ese tiempo he empleado aproximadamente el 30 % del tiempo indicado, de forma que esto equivale a haber trabajado 432 horas de forma aproximada, o 2,4 meses a tiempo completo. Según [45] el salario medio de un recién graduado en Ingeniería de Telecomunicaciones es de 1716€ al mes, por lo que se puede obtener el coste total aproximado asociado al trabajo desarrollado.

Además del coste humano, se debe de tener el cuenta el coste asociado a los materiales para las tareas de adquisición de las señales EEG y PMA. Para ello se debe añadir el coste del dispositivo de adquisición de señales EEG, además del dispositivo utilizado para las señales PMA, incluyendo sus gastos

derivados (gorros de EEG), jeringas, gel conductor e imanes de neodimio para el caso de las señales PMA (además de las intervenciones quirúrgicas llevadas a cabo para la implantación de los sensores). Además debe tenerse en cuenta el coste del ordenador personal utilizado para el desarrollo del trabajo. En la tabla A.2 se puede observar el coste total asociado al trabajo.

Puesto	Meses de trabajo equivalentes	Salario mensual	Coste ordenador	Coste obtención de las señales EEG y PMA	Coste total
Ingeniero de telecomunicaciones júnior	2,4	1716€	1100€	25000€	30218€

Tabla A.2: Presupuesto del proyecto.

Bibliografía

[1] Grad, L. I., Rouleau, G. A., Ravits, J., & Cashman, N. R. (2017). Clinical Spectrum of Amyotrophic Lateral Sclerosis (ALS). Cold Spring Harbor perspectives in medicine, 7(8), a024117.

[2] Hulisz D. (2018). Amyotrophic lateral sclerosis: disease state overview. The American journal of managed care, 24(15 Suppl), S320–S326.

[3] M Das, J., Anosike, K., & Asuncion, R. M. D. (2022). Locked-in Syndrome. In StatPearls. StatPearls Publishing.

[4] Goutman, S. A., Hardiman, O., Al-Chalabi, A., Chió, A., Savelieff, M. G., Kiernan, M. C., & Feldman, E. L. (2022). Recent advances in the diagnosis and prognosis of amyotrophic lateral sclerosis. The Lancet. Neurology, 21(5), 480–493.

[5] Camacho, A., Esteban, J., & Paradas, C. (2018). Report by the Spanish Foundation for the Brain on the social impact of amyotrophic lateral sclerosis and other neuromuscular disorders. Informe de la Fundación Del Cerebro sobre el impacto social de la esclerosis lateral amiotrófica y las enfermedades neuromusculares. Neurologia, 33(1), 35–46.

[6] Chiò, A., Logroscino, G., Traynor, B. J., Collins, J., Simeone, J. C., Goldstein, L. A., & White, L. A. (2013). Global epidemiology of amyotrophic lateral sclerosis: a systematic review of the published literature. Neuroepidemiology, 41(2), 118–130.

[7] Mridha, M. F., Das, S. C., Kabir, M. M., Lima, A. A., Islam, M. R., & Watanobe, Y. (2021). Brain-Computer Interface: Advancement and Challenges. Sensors (Basel, Switzerland), 21(17), 5746.

[8] Birbaumer, N., & Cohen, L. G. (2007). Brain-computer interfaces: communication and restoration of movement in paralysis. The Journal of physiology, 579(Pt 3), 621–636.

[9] Thau, L., Reddy, V., & Singh, P. (2022). Anatomy, Central Nervous System. In StatPearls. StatPearls Publishing.

[10] Torrico, T. J., & Abdijadid, S. (2022). Neuroanatomy, Limbic System. In StatPearls. StatPearls Publishing.

[11] Huang, J. (2023, 20 mayo). Generalidades sobre la función cerebral. Manual MSD versión para profesionales. https://www.msdmanuals.com/es-es/professional/trastornos-neurol%C3%B3gicos/funci%C3%B3n-y-disfunci%C3%B3n-de-los-l%C3%B3bulos-cerebrales/generalidades-sobre-la-funci%C3%B3n-cerebral

[12] Serrano, C., & Dds, A. T. (2023, junio 6). Histología de las neuronas. https://www.kenhub.com/es/library/anatomia-es/neurona

[13] neuronas. (s.f.). bebrain. https://bebrainid.wixsite.com/bebrain/neuronas

[14] Nagel, S. (2019). Towards a home-use BCI: fast asynchronous control and robust non-control state detection.

[15] Wicho. (s.f.). Magnetoencefalografía, una ventana al interior del cerebro mediante sus campos magnéticos. Microsiervos. https://www.microsiervos.com/archivo/ciencia/magnetoencefalografia-ventana-interior-cerebro-mediante-campos-magneticos.html

[16] Rabbani, Q., Milsap, G., & Crone, N. E. (2019). The Potential for a Speech Brain-Computer Interface Using Chronic Electrocorticography. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics, 16(1), 144–165.

[17] Müller-Putz G. R. (2020). Electroencephalography. Handbook of clinical neurology, 168, 249–262.

[18] Arias-Londoño, J. D., Godino-Llorente, J. I., Markaki, M., & Stylianou, Y. (2011). On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. Logopedics, phoniatrics, vocology, 36(2), 60–69.

[19] Vilda, P.G., Fernández-Baíllo, R., Biarge, M.V., Lluis, V.N., Marquina, A.Á., Mazaira-Fernández, L.M., Martínez, R., & Godino-Llorente, J.I. (2009). Glottal Source biometrical signature for voice pathology detection. Speech Commun., 51, 759-781.

[20] Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. Frontiers in neuroscience, 9, 217.

[21] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology, 113(6), 767–791.

[22] Chakrabarti, S., Sandberg, H. M., Brumberg, J. S., & Krusienski, D. J. (2015). Progress in speech decoding from the electrocorticogram. Biomedical Engineering Letters, 5(1), 10–21.

[23] Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M. C., Goldrick, M. A., Krusienski, D. J., Slutzky, M. W., & Schultz, T. (2019). Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. Frontiers in neuroscience, 13, 1267.

[24] Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., & Guenther, F. H. (2010). Brain-Computer Interfaces for Speech Communication. Speech communication, 52(4), 367–379.

[25] Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568(7753), 493–498.

[26] Colaboradores de Wikipedia. (2023c). Regresión lineal. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal.

[27] Ortega Páez, E., Ochoa Sangrador, C., & Molina Arias, M. (2022). Regresión logística binaria simple. Evid Pediatr, 18, 11.

[28] O'Meara, N.M., Salort, E.V., & Lario, F.C. (2009). Metodologías de Inteligencia Artificial para la Toma de Decisiones en la Red/Cadena de Suministro en el Contexto de Incertidumbre*.

[29] Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. Current biology : CB, 29(7), R231–R236.

[30] LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[31] Na, & Na. (2020). Convolutional Neural Networks: La Teoría explicada en Español — Aprende Machine Learning. Aprende Machine Learning. https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/

[32] Ogino, M., Kanoga, S., Muto, M., & Mitsukura, Y. (2019). Analysis of Prefrontal Single-Channel EEG Data for Portable Auditory ERP-Based Brain-Computer Interfaces. Frontiers in human neuroscience, 13, 250.

[33] Nurse, E. S., Karoly, P. J., Grayden, D. B., & Freestone, D. R. (2015). A Generalizable Brain-Computer Interface (BCI) Using Machine Learning for Feature Discovery. PloS one, 10(6), e0131328.

[34] Taal, C.H., Hendriks, R.C., Heusdens, R., & Jensen, J.R. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. IEEE Transactions on Audio, Speech, and Language Processing, 19, 2125-2136.

[35] Rix, A.W., Beerends, J.G., Hollier, M., & Hekstra, A.P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 2, 749-752 vol.2.

[36] Ravanelli, M., & Bengio, Y. (2018). Interpretable Convolutional Filters with SincNet. ArXiv, abs/1811.09725.

[37] MDiazMartin. (s.f.). GitHub - MDiazMartin/EEG_SincNet: Interfaces cerebro-ordenador basadas en EEG para la síntesis de voz. GitHub. https://github.com/MDiazMartin/EEG_SincNet

[38] Sun, Q. J., Vo, K., Lui, K. K., Nunez, P. L., Vandekerckhove, J., & Srinivasan, R. (2022). Decision SINCNET: Neurocognitive models of decision making that predict cognitive processes from neural signals. 2022 International Joint Conference on Neural Networks (IJCNN).

[39] Jennyqsun. (s.f.). GitHub - jennyqsun/EEG-Decision-SincNet. GitHub. https://github.com/jennyqsun/EEG-Decision-SincNet

[40] Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A VoCODer-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions on Information and Systems, E99.D(7), 1877-1884.

[41] ReSSint. (2023, 19 enero). RESSINt. voice restoration with silent speech interfaces. https://aholab.ehu.eus/ressint/

[42] González, J. A., Cheah, L. A., Gilbert, J. M., Bai, J., Ell, S. R., Green, P., & Moore, R. K. (2016). A silent speech system based on permanent magnet articulography and direct synthesis. Computer Speech & Language, 39, 67-87.

[43] González, J. A., Cheah, L. A., Gomez, A. M., Green, P., Gilbert, J. M., Ell, S. R., Moore, R. K., & Holdsworth, E. (2017). Direct speech reconstruction from articulatory sensor data by machine learning. IEEE/ACM transactions on audio, speech, and language processing, 25(12), 2362-2374.

[44] González, J. A., Cheah, L. A., Green, P. D., Gilbert, J. M., Ell, S. R., Moore, R. K., & Holdsworth, E. (2017). Evaluation of a silent speech

interface based on magnetic sensing and deep learning for a phonetically rich vocabulary. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 3986-3990

[45] ¿Cuánto cobra un ingeniero de telecomunicaciones? (Sueldo 2023) — Jobted.es. (s.f.). https://www.jobted.es/salario/ingeniero-telecomunicaciones Para el apendice del salario es lo ultimo