Universidad del País Vasco Euskal Herriko Unibertsitatea
BILBOKO INGENIARITZA ESKOLA
ESCUELA DE INGENIERÍA DE BILBAO

MÁSTER UNIVERSITARIO EN INGENIERÍA DE TELECOMUNICACIÓN

TRABAJO FIN DE MÁSTER

CONVERSIÓN DE PARÁMETROS ELECTROMIOGRÁFICOS EN VOZ

Estudiante *Román, Agra, Xabier*
Directora *Navas, Cordón, Eva*
Departamento **Ingeniería de Comunicaciones**
Curso académico *2020-2021*

Bilbao, 22 de septiembre de 2021

Agradecimientos

Transmitir mi más sincero agradecimiento a todos aquellos que me han ayudado a lo largo de esta etapa y han colaborado en esta investigación.

En primer lugar, a mi tutora, la Doctora Eva, por su ayuda y dedicación en este Trabajo de Fin de Máster.

En segundo lugar, a mi familia y amigos que han estado apoyándome en todo momento.

También, expresar mi más sentido agradecimiento al grupo de investigación Aholab por acogerme tanto en el TFG como en el TFM.

Por último, agradecer al profesorado completo del Máster en Ingeniería de Telecomunicación, especialmente a los de este curso por su dedicación, compromiso y trabajo.

Resumen trilingüe

En este trabajo nos centraremos en la generación de un sistema de referencia de generación de voz a partir de parámetros articulatorios. Para ello, se explican los diferentes métodos para la captación de señales y sistemas que permiten la comunicación de voz sin señal acústica audible. El objetivo principal del trabajo es obtener el mejor sistema de referencia para predicción de diferentes parámetros de la voz para el proyecto ReSSint, para lo que se realizarán diferentes experimentos. En ellos, se han probado diferentes redes neuronales profundas para generar los valores de los parámetros acústicos a partir de parámetros captados por sensores electromiográficos. Finalmente, se ha realizado un análisis para comprobar la precisión de las predicciones.

Lan honetan ahotsa sortzeko erreferentzia sistema sortzeko artikulazio parametroetatik abiatuko gara. Horretarako, seinale akustiko entzungarririk gabeko ahots bidezko komunikazioa ahalbidetzen duten sistemak eta seinaleak harrapatzeko metodo desberdinak azalduko dira. Lanaren helburu nagusia ReSSint proiektuarentzako ahots parametro desberdinak iragartzeko erreferentziako sistema onena lortzea da eta horretarako esperimentu desberdinak egingo dira. Horietan, sare neuronal sakon desberdinak probatu dira sentsore elektromiografikoek harrapatutako parametroetatik parametro akustikoen balioak sortzeko. Azkenik, iragarpenen zehaztasuna egiaztatzeko analisia egin da.

In this work we will focus on the generation of a voice generation reference system from articulatory parameters. The different methods for capturing signals and systems that allow voice communication without audible acoustic signal will be explained. The main objective of the work is to obtain the best reference system for prediction of different voice parameters for the ReSSint project, for which different experiments will be carried out. In them, different deep neural networks have been tested to generate the values of the acoustic parameters from parameters captured by electromyographic sensors. Finally, an analysis has been carried out to check the precision of the predictions.

Palabras clave

Interfaces de habla silenciosa, Redes Neuronales, Voz sintética, Coeficientes Cepstrales, Espectrograma, Frecuencia Fundamental

Índice

Índice de figuras

Índice de tablas

1. Introducción

El habla es la forma más importante de comunicación entre personas, de manera eficiente y natural. Por lo que la discapacidad para producir mensajes hablados produce un aislamiento social de las personas que la sufren, ya que su comunicación con el entorno se ve seriamente afectada, empeorando las relaciones personales y generando problemas de integración en el entorno escolar y social.

Los avances tecnológicos han permitido a estas personas interactuar con habla de manera compleja, por ejemplo, utilizando las aplicaciones de comunicación aumentativa y alternativa (CAA). Estas aplicaciones integran conversores de texto a voz (TTS, Text To Speech) para producir mensajes orales.

Otra de las alternativas, son las interfaces de habla silenciosa (SSI, Silent Speech Interfaces), se tratan de interfaces de voz que son capaces de generar voz incluso cuando no hay una señal acústica audible.

Entre los tipos de SSI destaca la conversión de señales electromiográficas (EMG) a voz. Se trata de la conversión directa de la actividad muscular electrofisiológica facial, medida mediante electromiografía de superficie, en habla audible. Aparte de los usos clínicos, otras aplicaciones potenciales de esta tecnología incluyen proporcionar privacidad en las conversaciones y mejorar la comunicación hablada normal en entornos ruidosos a personas sin capacidad para producir mensajes hablados.

El habla audible se genera directamente desde la señal biológica, mediante el uso de electrodos no invasivos para la adquisición de señales EMG. Normalmente la generación de la señal hablada a partir de los datos adquiridos por los sensores EMG, se realiza mediante redes neuronales profundas (DNN, Deep Neural Networks), entrenando con las grabaciones de las señales biológicas y de la voz alineadas en el tiempo.

En este trabajo se pretende desarrollar un sistema de referencia de generación de parámetros acústicos a partir de señales electromiográficas, usando una base de datos ya existente que permita la generación de una voz sintética a partir de la predicción de Coeficientes Cepstrales y la Frecuencia Fundamental o el Espectrograma, mediante redes neuronales.

2. Contexto

Este trabajo, se enmarca dentro del proyecto ReSSint (Aholab, 2020)[1], cuyo objetivo es desarrollar SSIs para restaurar la comunicación en personas que se han visto privadas de la capacidad de hablar en castellano. Es de decir, que el principal objetivo de ReSSint es desarrollar interfaces de voz silenciosas, que son dispositivos que capturan señales biológicas no acústicas generadas durante el proceso de producción de voz y las utilizan para predecir el mensaje deseado.

En concreto este trabajo, consiste en el primer paso del desarrollo del proyecto ReSSint, con el principal objetivo de obtener un sistema de referencia de predicción obteniendo unos resultados similares a los obtenidos por la Universidad de Bremen (Diener, 2021) que, aunque obtienen buenas medidas objetivas no fueron capaces de reproducir los mensajes orales inteligibles. Por ello, para el proyecto ReSSint es necesario disponer de un sistema de referencia para obtener mensajes audibles utilizando SSIs, el cual se mejorará en el futuro a lo largo del desarrollo del proyecto.

Se ha realizado un análisis previo del impacto que puede tener este proyecto a nivel europeo y nacional. En este análisis, se destaca una encuesta realizada en 2011 por Eurostat, que concluyó que el 0.5 % de los europeos presentaban dificultades para la comunicación (Eurostat, 2021). Centrándonos en España y en la discapacidad para producir mensajes hablados, según la última Encuesta de Discapacidad, Autonomía Personal y Situaciones de Dependencia realizada en 2008 por el INE, en España hay 410.600 personas afectadas por una discapacidad que les impide generar mensajes orales (Instituto Nacional de Estadística, 2008). En la Figura 1, se muestra la distribución por edades de las personas con discapacidad para producir mensajes orales en España.

[1] http://ressint.eus/

Figura 1. Población española con discapacidad para producir mensajes hablados por rango de edad.

La comunicación vocal humana es un proceso extremadamente complejo, donde se involucran múltiples órganos, entre los que se encuentran la lengua, los labios, la mandíbula, las cuerdas vocales y los pulmones. Además, requiere una coordinación precisa entre estos órganos para producir sonidos específicos.

En la Figura 2, se muestra las estructuras involucradas en la producción de la voz además de una breve descripción el movimiento de las cuerdas vocales que juegan un papel fundamental en la generación del habla.

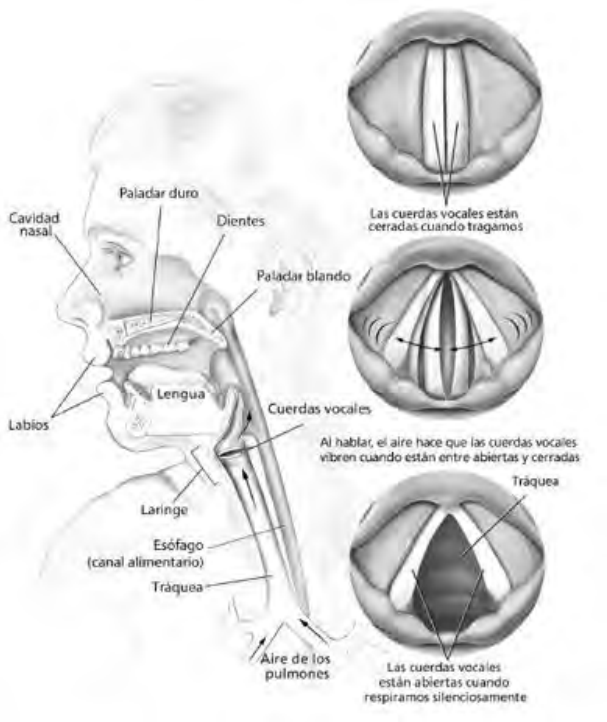


Figura 2. Estructuras involucradas en la generación de voz. Imagen extraída del sitio web de National Institute of Deafness an Other Communications Disorders del Departamento de Salud de EEUU.

Debido a esta complejidad, la comunicación vocal, puede volverse difícil o incluso imposible cuando estos órganos, las áreas anatómicas del cerebro involucradas en la producción del habla o las vías neurales por las que el cerebro controla los músculos están dañados o alterados.

Por ello, estas personas, principalmente, utilizan aplicaciones CAA. Estas aplicaciones facilitan la elaboración rápida de mensajes utilizando símbolos y esquemas, mensajes que son después convertidos en voz por un TTS que suele estar integrado con la aplicación. Además, permiten el control de la ejecución de otros programas mediante interfaces adaptadas a las limitaciones del usuario, siendo la última tecnología el control mediante el movimiento de los ojos, como el comunicador que se muestra en la Figura 3. El cual es muy útil para aquellas personas con movilidad reducida que a su vez no pueden comunicarse oralmente.

Figura 3. Comunicador VOX 12 EYE PRO, Imagen extraída del sitio web de BJ Adaptaciones.

Estos sistemas presentan algunos inconvenientes entre los que se encuentran la dificultad de comunicación en ambientes hostiles o con alto nivel de ruido, además de la falta de privacidad en las conversaciones y la limitación de vocabulario de estos comunicadores. Por ello, una alternativa interesante para mejorar la comunicación oral de estas personas son las SSIs.

Las SSIs, son sistemas que permiten que la comunicación de voz tenga lugar cuando una señal acústica audible no está disponible, adquiriendo los datos de sensores que captan señales de diversos elementos relacionados con el proceso de producción del habla humana, como pueden ser los articuladores, sus vías neurales o el cerebro. Estos sensores producen una representación digital del habla que se puede sintetizar directamente, interpretándola como datos.

Las SSI aún se encuentran en la etapa experimental, pero parece evidente una serie de posibles implementaciones para mejorar la calidad de vida de algunas personas, entre las que se encuentran personas que se han sometido a una laringectomía o ciudadanos mayores para quienes hablar requiere un esfuerzo, y quienes, de esta manera, serían capaces de producir mensajes hablados mediante la gesticulación vocal.

Las señales biológicas utilizadas en estos sistemas, se dan por diferentes procesos eléctricos, físicos y biológicos que tienen lugar durante la producción del habla. Estos procesos incluyen actividad neuronal en las regiones anatómicas del cerebro involucradas en la planificación del habla y la actividad en el sistema nervioso periférico que proporciona control motor sobre los músculos articuladores, gestos articulatorios como la apertura de la boca o los movimientos

de la lengua, la vibración de las cuerdas vocales y la actividad pulmonar de los pulmones durante la respiración.

En función de los tipos de trastornos del habla, son aplicables diferentes tecnologías de sensores para adquirir los datos de entrada de las SSI. En la Tabla 1, se muestran diferentes desórdenes que impiden la capacidad para producir mensajes hablados, entre los que se encuentran las afasia, la apraxia, la disartria y la laringectomía.

Desorden	Descripción	Sensores
Afasia	Trastorno causado por lesiones en las partes del cerebro que controlan el lenguaje	Sensor de actividad cerebral
Apraxia	Trastorno del cerebro y del sistema nervioso en el cual una persona es incapaz de llevar a cabo tareas o movimientos cuando se le solicita	Sensor de actividad cerebral
Disartria	Trastorno nervioso, cerebral o muscular dificulta el uso o control de los músculos de la boca, la lengua, la laringe o las cuerdas vocales.	Sensor de actividad cerebral Sensor de actividad muscular
Laringectomía	Operación quirúrgica en la que se extirpa total o parcialmente la laringe.	Sensor de actividad cerebral Sensor de actividad muscular Captura del movimiento del articulador

Tabla 1. Resumen de los desórdenes que afectan a la capacidad de producir mensajes hablados.

En este proyecto, se ha seleccionado los sensores de adquisición de señales bioeléctricas, que permiten registrar la actividad eléctrica generada por diferentes partes del cuerpo humano, como pueden ser el corazón, el cerebro o los músculos. Concretamente en este caso se ha seleccionado sensores electromiográficos, que adquieren la señal generada por los músculos esqueléticos cuando realizan algún movimiento. Aunque también se podrían utilizar otro tipo de sensores como los PMA, que utilizan magnetómetros para capturar cambios en el campo magnético generado por el movimiento de los imanes (Beiming, Nordine, Ted, Omer T, & Jun, 2019).

Durante la producción del habla los músculos en la cara y la laringe son responsables de los movimientos. Como se mencionó anteriormente, el cerebro controla la activación de estos músculos por medio de señales eléctricas transmitidas a través de las neuronas motoras del sistema nervioso periférico. Estas señales eléctricas hacen que los músculos se contraigan y relajen, produciendo así los movimientos y gestos articulatorios requeridos.

Se pueden utilizar dos tipos de electrodos para la adquisición de señales EMG, invasivos y no invasivos. Los métodos no invasivos emplean electrodos superficiales directamente adheridos a la piel. Las señales del EMG adquiridas se tratan del sumatorio del potencial de acción de los músculos localizados debajo del área cubierta por el sensor. Los métodos invasivos miden principalmente potenciales de acción localizados y requieren realizar implantación o incluso inserción de los sensores en la cavidad bucal, por lo que en este caso es más interesante la utilización de métodos no invasivos.

En este proyecto, se trabaja con sensores similares al que se muestra en la Figura 4, que consiste en un array de electrodos (Diener, Janke, & Schultz, 2015) que ofrece el sumatorio de todos los potenciales de acción de los músculos localizados debajo el área cubierta por el sensor.

Figura 4. Array de electrodos. Imagen extraída del artículo (Gonzalez, Gomez, Martín, Pérez, & Gomez, 2020)

La base de datos utilizada en este proyecto pertenece a Cognitive Systems Lab de la Universidad de Bremen y fue adquirida con un sistema de adquisición compuesto por cuarenta sensores, colocados como se indica en la Figura 5. Consiste en una base de datos grabada en inglés por ocho locutores diferentes no profesionales.

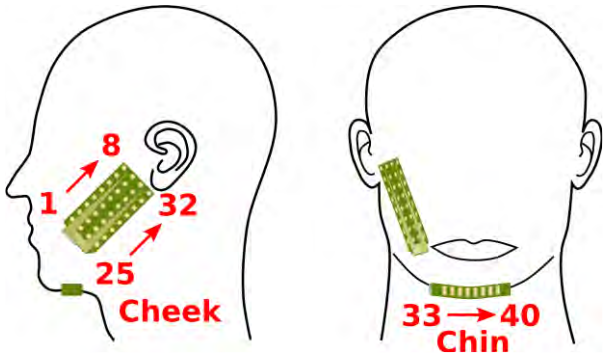


Figura 5. Colocación de los sensores. Imagen extraída del artículo (Diener, Vishkasougheh, & Schultz, 2020)

Los primeros cuarenta sensores corresponden al EMG correspondiente a su sensor físico. El último sensor, corresponde a la señal de sincronización que se muestra en la Figura 6, como se puede observar la señal es un pulso rectangular que muestra un 0 en modo inactivo y un 16'62 en modo activo. Además, el tramo activo corresponde con la duración del audio permitiendo así eliminar la zona del audio sin señal deseada, ya sea de silencio o de ruido.

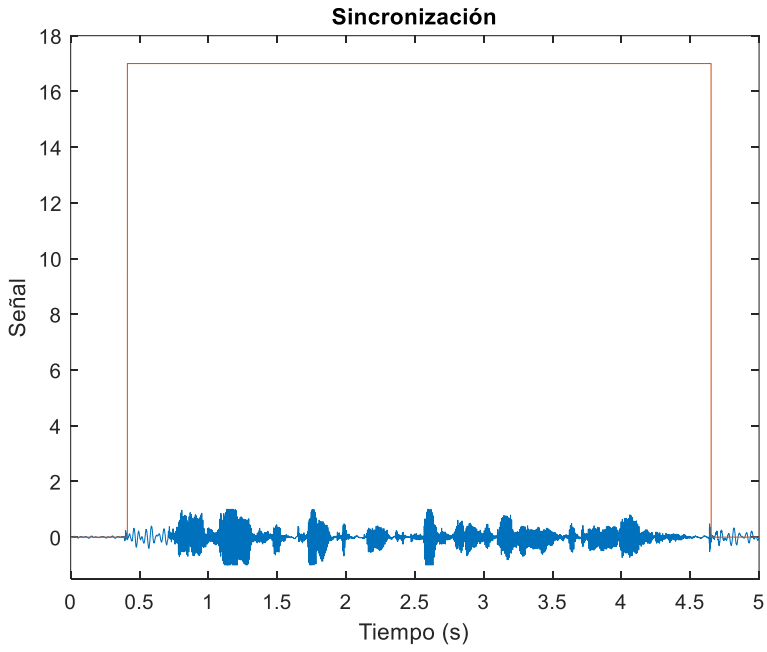


Figura 6. Señal de sincronización (rojo) junto con la señal de audio grabada (azul).

Una vez seleccionados los sensores y capturadas las señales, se pueden utilizar dos principales enfoques SSI para decodificar los parámetros de los sensores (Meltzner, y otros, 2017).

El primer enfoque, es el que utiliza reconocimiento de voz, mediante algoritmos entrenados en datos de voz silenciosa para decodificar el habla a partir de los vectores de características extraídos de la señal biológica. Si se desea escuchar el mensaje obtenido, se puede utilizar un programa de conversión de texto en habla para sintetizar voz a partir del texto decodificado si es necesario.

En el segundo enfoque, que es el que se utilizará en este proyecto, el habla audible se genera directamente desde la señal biológica. Mediante la utilización de redes neuronales profundas (DNN) (LeCun, Bengio, & Hinton, 2015) (Schmidhuber, 2015), entrenando la red con las grabaciones de las señales biológicas y de la voz alineadas en el tiempo.

3. Objetivos y alcance

El objetivo principal de este proyecto es generar un sistema de referencia para la predicción de parámetros de la voz para el proyecto ReSSint, partiendo del experimento realizado por Cognitive Systems Lab de la Universidad de Bremen y adaptando los resultados obtenidos para obtener el mejor sistema de referencia posible realizando diferentes experimentos.
Desde un punto de vista técnico, los objetivos principales de este proyecto son:

- Procesado de datos del EMG para convertirlos en datos de entrada en una red neuronal de entrenamiento.
- Procesado de señales de audio, para convertirlos en datos de salida en la red neuronal de entrenamiento.
- Diseño de redes neuronales capaces de predecir Coeficientes Cepstrales.
- Diseño de redes neuronales capaces de predecir valores de la Frecuencia Fundamental.
- Diseño de redes neuronales capaces de predecir el Espectrograma de la señal.
- Análisis de los resultados obtenidos mediante las redes neuronales, mediante diferentes métricas.
- Procesado de datos para generación de voz sintética.
- Evaluación comparativa con los resultados de Cognitive Systems Lab.

4. Beneficios

El principal beneficio que aporta este trabajo se trata de la posibilidad de generar una voz sintética a partir de la señal capturada por sensores electromiográficos, que adquieren la señal generada por los músculos esqueléticos.

Entre los beneficios técnicos se encuentran el desarrollo, generación y entrenamiento de redes neuronales capaces de predecir los Coeficientes Cepstrales o el Espectrograma a partir de la señal adquirida de los sensores o capaces de predecir la frecuencia fundamental a partir de los Coeficientes Cepstrales con unos resultados subjetivamente aceptables como para considerarlo el sistema de referencia del proyecto. Además de la posibilidad de usar SSIs gracias a la predicción de ambos valores y la decodificación de los mismos.

Entre los beneficios sociales, se puede destacar la capacidad para producir mensajes orales para aquellas personas que han perdido o no tienen la capacidad para ello mediante un sistema más complejo. Aparte del beneficio principal se encuentran otros beneficios:

- Aumento de la autonomía de estas personas.
- Posibilitan la socialización de la persona y mejoran la interacción comunicativa.
- Evita el aislamiento.
- Facilitan la comprensión.

5. Análisis del estado del arte

5.1. Introducción

La comunicación verbal mediante dispositivos asume un papel fundamental en la comunicación y actualmente se encuentra en una amplia gama de escenarios, desde asistentes personales telefónicos hasta dispositivos de entretenimiento.

La comunicación humana, no se trata únicamente de la onda de sonido audible resultante. Estudios previos, en particular el que descubrió el efecto McGurk (McGurk & MacDonald, 1976), mostró que los humanos, son capaces de emplear los sentidos auditivo y visual en la percepción del habla. Concretamente, los humanos son de hecho capaces de interpretar información contextual relacionada, como el movimiento de los labios, la cabeza y el cuerpo. Por ejemplo, el gesto de la mano o expresiones faciales. En la actualidad, se ha visto la importancia de esta característica no audible y su supresión debido al uso generalizado de las mascarillas por la Covid-19. Por último, también se tienen en cuenta las características específicas de la señal del habla como la prosodia, que se convierte en una parte integral del proceso de comunicación verbal.

Algunas de estas señales de información pueden proporcionar cierta redundancia o ser menos importantes para la comunicación, sin embargo, otras pueden ser esenciales para comprender completamente el mensaje.

En muchas de las situaciones el habla audible es suficiente, pero hay una multitud de escenarios para los que es inadecuada debido al ruido ambiental, la necesidad de privacidad o el resultado de las deficiencias del habla existentes.

5.2. Investigación SSI

Los investigadores están comenzando a explorar cómo obtener la información de las diferentes fases del proceso de producción del habla, entre las que se encuentran los articuladores, sus vías neurales o el cerebro, debido a que esta información puede ser adquirida y utilizada en el desarrollo de modalidades mejoradas de entrada de voz para interacción de humanos con máquinas.

Esta área de investigación, comúnmente designada como las SSI, ofrece una solución potencial y alternativa para una interacción humano-computadora (HCI, Human Computer Interaction) en ausencia de habla audible. Esta alternativa, teóricamente es capaz de abordar varios problemas inherentes al reconocimiento automático de voz (ASR, Automatic Speech Recognition), como podría ser el ruido ambiente.

Las SSI, se puede decir, que se trata del proceso de producción del habla humana mediante la exploración de señales biométricas diferentes a la voz, generadas en el proceso de producción de la misma, medida por dispositivos de detección como ultrasonidos, electromiografía, visión u otros tipos de sensores.

Actualmente, existe un trabajo notable de literatura sobre SSI, pero el estudio del mismo, aún carece de una visión integrada de aspectos clave, tecnologías y hallazgos que permiten un enfoque sistemático y aplicado.

En la década de 1990, con el uso masivo de teléfonos móviles, las SSI comenzaron a aparecer como una posible solución a problemas como la privacidad en las comunicaciones personales, y para los usuarios que habían perdido la capacidad de producir un habla sonora. También, con la evolución de las cámaras de video, comenzaron a aparecer más estudios sobre lectura de labios. Un ejemplo relevante es el trabajo de Hasegawa y Ohtani que lograron un reconocimiento del 91%, basado en información de los labios y la lengua adquirido mediante video (Hasegawa & Ohtani, 1992).

A principios de la década de 2000, DARPA (Agencia de Proyectos de Investigación Avanzada de Defensa) se centró en recuperar las señales de excitación glotal del habla sonora en entornos ruidosos, con el Programa de codificación avanzada del habla (Ng, Burnett, Holzrichter, & Gable, 2000).

En 2002, en Japón, NTT DoCoMo, anunció un prototipo de teléfono celular silencioso mediante la utilización de electromiografía y captura óptica del movimiento de los labios (Fitzpatrick, 2002), especialmente dirigido a escenarios que abarcan el ruido ambiental y usuarios con problemas de habla.

En 2012, la Universidad de Bremen, realizó su primera publicación, donde utiliza un enfoque de conversión de voz basado en el modelo de mezcla de gaussianas (GMM, Gaussian Mixture Model). Los resultados de la evaluación experimental indicaron que el número óptimo de

mezclas gaussianas depende de la cantidad de datos de entrenamiento. Por último, realizaron una evaluación independiente de la sesión, que utilizó sesiones de grabación, entre las cuales se retiraron y volvieron a colocar los electrodos EMG. Los resultados de estos datos mostraron un rendimiento bastante razonable en comparación con conversiones realizadas sobre una misma sesión, es decir, sin retirar los electrodos (Janke, Wand, Nakamura, & Schultz, 2012).

En 2020, la Universidad de California, demostró una alta inteligibilidad de las frases de salida con datos de un vocabulario cerrado, obteniendo una tasa de error de palabra de transcripción del 3,6% y una reducción del error relativo del 95% desde su sistema de base (Gaddy & Klein, 2020).

En 2021, la Universidad de Bremen, ha publicado una tesis doctoral (Diener, 2021), donde presentan diferentes enfoques para hacer frente a algunos problemas que plantea la conversión de parámetros EMG en voz. Han desarrollado un sistema de baja latencia útil para utilizar en tiempo real, también han analizado diferentes Redes Neuronales y Corpus para seleccionar la configuración más adecuada para la conversión de parámetros EMG en voz. En esta conversión de parámetros han obteniendo finalmente una distorsión de los coeficientes cepstrales en escala Mel (Mel Cepstral Distortion, MCD) de 8'5 dB aproximadamente en el mejor de los casos.

En los últimos años, el concepto SSI se hizo más prominente en la investigación del habla, de diferentes modalidades, es decir, diferentes formas en las que los usuarios pueden interactuar con el sistema. Como consecuencia, supuso un crecimiento en la investigación de SSI.

5.3. Sensores y modalidades

Debido las diversas etapas de la producción del habla y sus diferentes resultados medibles y extraíbles, existe una amplia variedad de modalidades de transmisión de información entre el usuario y una máquina, con el objetivo principal de obtener un discurso silencioso. En la Tabla 2, se muestra un pequeño análisis de las diferentes etapas que forman parte del proceso de producción del habla relacionadas con un ejemplo tecnológico de adquisición de datos.

Etapa de producción del habla	Ejemplo
Cerebro y Sistema Nervioso	Interpretación de señales de implantes en la corteza motora del habla
Control articulatorio (Músculos)	Electromiografía de superficie (SEMG, Surface Electromiography) del articulador muscular
Articulación (Movimiento)	Caracterización en tiempo real del tracto vocal mediante ultrasonido (US, Ultrasound)
Efecto articulatorio (Trato Vocal)	Transformación digital de señales de un micrófono no audible de murmullo (NAM, Non Audible Microphone)

Tabla 2. Análisis de las etapas de producción del habla.

Las modalidades actualmente en uso, cubren todas las etapas de la producción del habla humana. Partiendo de la interpretación de señales de implantes en la corteza cerebral hasta la medición de efectos visibles en el rostro.

Todas las modalidades de SSI, tienen como objetivo común capturar información sobre la actividad del habla, concretamente, actividad que no es visible ni medida por los sentidos humanos.

Actualmente, existen diferentes tecnologías de adquisición de información. Pero cada una de ellas es más apropiada en función del caso para el que se desee obtener dicha información, como puede ser un caso extremo de parálisis, donde se debe captar la información de la producción del habla en el cerebro o en el sistema nervioso.

5.4. Bases de Datos

Uno de los factores que está frenando el desarrollo de la tecnología SSI es la falta de grandes conjuntos de datos, que son necesarios para desarrollar herramientas de voz y su recopilación requiere mucho tiempo y esfuerzo. En consecuencia, la mayoría de los estudios realizados en este campo han utilizado pequeños conjuntos de datos registrados por diferentes grupos de investigación, utilizando dispositivos internos de registro de señales biológicas. Esta diversidad de enfoques ha llevado a la fragmentación de la investigación, lo que dificulta la comparación de los avances técnicos y algorítmicos de diferentes tecnologías.

Actualmente, la única base de datos grande disponible y con características adecuadas, es TORGO (Rudzicz, Namasivayam, & Wolff, 2012), que contiene alrededor de 23 horas de información acústica alineada en el tiempo. y señales articulatorias EMA obtenidas de ocho hablantes disártricos y siete hablantes sin patologías para utilizar como control.

También existen otros conjuntos de datos de menor tamaño, como:

- El corpus EMG-UKA para el procesado de EMG en habla (Wand, Janke, & Schultz, 2014), está formado por datos en inglés adquiridos con varios modos de habla, audible, susurrada y articulada en silencio. Además de los datos EMG, que son capturados mediante 6 sensores individuales, también cuenta con los datos acústicos síncronos. El corpus consta de 63 sesiones grabadas por 8 locutores, con un total de 7 horas y 32 minutos.

- La base de datos articulatoria MOCHA-TIMIT (Wrench, 2020), se trata de una base de datos en inglés, que está formada por grabaciones de 2 locutores con 460 frases cada uno. Para las grabaciones, se utilizó una frecuencia de muestreo de micrófono de 16 kHz y para la captura de las señales biológicas, se utilizaron 10 sensores de 2 mm.

- CSL-EMG Array de la Universidad de Bremen (Diener, Vishkasougheh, & Schultz, 2020), se trata de una base de datos, que está formada por grabaciones de 8 locutores no profesionales y en inglés. Cuenta con diferentes bloques de entrenamiento y de evaluación y forman un total de 10 horas de grabaciones aproximadamente. Para la adquisición de las señales biológicas, se utilizaron 40 sensores, distribuidos en 2 arrays independientes, uno de 32 sensores y otros de 8. Esta es la base de datos que se va a utilizar en este trabajo.

5.5. Evaluación

Los análisis realizados en la gran mayoría de los desarrollos de las SSI se han validado mediante datos pregrabados. En estos análisis, se utiliza un corpus de datos pregrabados tanto para el entrenamiento del sistema como para la evaluación. Los resultados obtenidos en este tipo de evaluación, son válidos para optimizar algunos parámetros del sistema, como la latencia del sistema, la calidad de salida y la solidez del sistema. Para analizar el rendimiento y desempeño de estos sistemas es necesario realizar análisis en línea (online). Los análisis en línea, evalúan la eficacia del SSI mientras está en uso, posiblemente mientras el usuario recibe retroalimentación de audio en tiempo real.

Idealmente, el sistema debe probarse en escenarios de la vida real, durante un período prolongado y con un número adecuado de usuarios que presentan una diversidad de problemas del habla en diferentes etapas de evolución para medir la eficacia real de los sistemas.

El parámetro de medida más utilizado, es la distorsión Mel-Cepstral o MCD, que mide cuánto de diferentes son dos secuencias de coeficientes Mel-Cepstrum. De esta manera es fácil conocer cuál es la distorsión entre la señal original y la señal predicha. Las últimas publicaciones realizadas sobre SSIs (Diener, Felsch, Angrick, & Schultz, 2018) (Diener, 2021), obtuvieron unos valores entre 8 dB y 9 dB.

6. Descripción de la solución propuesta

Para conseguir el objetivo principal que se plantea en este proyecto, se han seguido una sucesión de pasos como se muestran en la Figura 7.

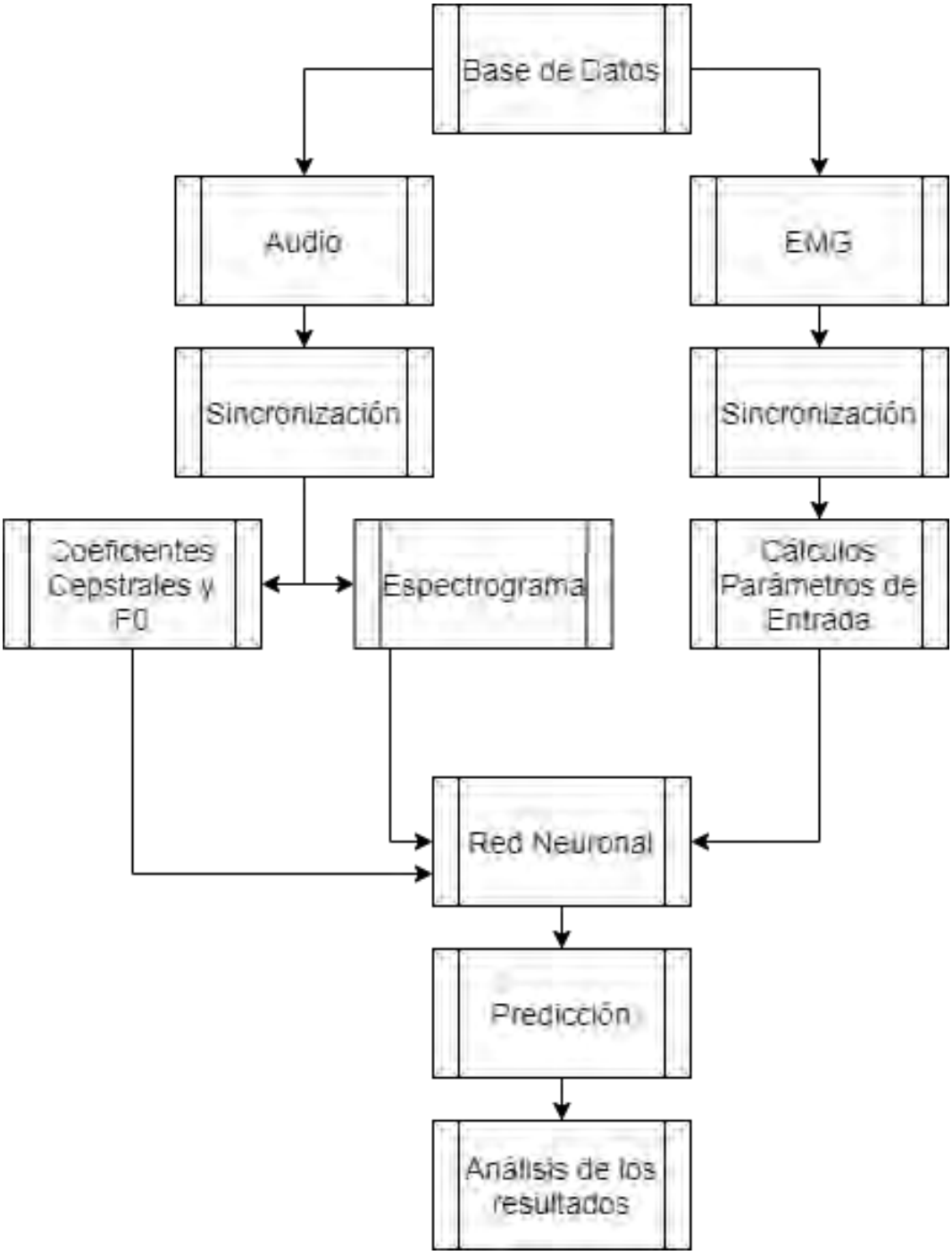


Figura 7. Arquitectura del Sistema

En este proyecto se parte desde de la Base de Datos donde se encuentran las señales de audio y su correspondiente señal electromiográfica obtenidas en el mismo instante de tiempo. Es necesaria la sincronización entre ambas señales para trabajar con la parte de la señal que realmente es necesaria, evitando de esta manera los tramos sin señal audible.

Cuando ambas señales están sincronizadas, se calculan los parámetros necesarios para cada una de ellas. Por un lado, de los sensores se obtendrán diferentes parámetros calculados de cada ventana estimada. Por otro lado, de las señales de audio grabadas se obtendrán los Coeficientes Cepstrales y los valores de la Frecuencia Fundamental o el Espectrograma que serán la salida de las Redes Neuronales.

Para procesar los datos se utilizarán scripts programados mediante el lenguaje de programación Python y Matlab, con el objetivo de obtener datos adecuados para introducir en las redes neuronales, entre los que se encuentran el filtrado, media, ZCR o la agrupación de los propios datos.

Se diseñarán, crearán y entrenarán diferentes modelos de Redes Neuronales, con el objetivo de predecir diferentes parámetros, que son, los Coeficientes Cepstrales, la Frecuencia Fundamental y el Espectrograma.

Una vez obtenidas las Redes Neuronales tanto de los Coeficientes Cepstrales como de la Frecuencia Fundamental, se realizará la evaluación de diferentes bloques para comprobar la precisión de las predicciones y se compararán con los resultados obtenidos y publicados por Cognitive Systems Lab mediante el MCD.

Sin embargo, en el caso alternativo de la Red Neuronal de predicción de los Espectrogramas, se realizará una evaluación mediante la distancia euclídea.

7. Metodología

7.1. Descripción de la Base de Datos

En este trabajo se ha utilizado la Base de Datos proporcionada por Cognitive Systems Lab de la Universidad de Bremen. Está formada por grabaciones de voz y sensores EMG de 8 locutores no profesionales en inglés no nativo. En Tabla 3, se muestra la estructura de las grabaciones de cada locutor, que se encuentran distribuidas en un bloque inicial formado por 340 frases con la señal audible (Frecuencia de muestreo 16000 Hz) y su correspondiente señal EMG (Frecuencia de muestreo 2048 Hz) que está formada por 40 sensores enumerados del 0 al 39. Además, cuenta con otros 3 bloques de evaluación formados por 50 frases con la señal audible y su correspondiente señal EMG. Finalmente, ambos bloques tanto el inicial como el de evaluación cuentan con la señal no audible, que consiste en una señal de audio en silencio y su correspondiente señal EMG.

En este proyecto, se ha utilizado únicamente el primer locutor, debido a que el objetivo es conseguir un sistema base para poder recrearlo en castellano con unas grabaciones propias.

Bloque	Número de Bloques	Número de Frases
Inicial	1	340
Evaluación	3	50

Tabla 3. Estructura de la Base de Datos.

7.2. Procesado de datos

7.2.1. Procesado de datos de entrada

En este apartado, se detallan las especificaciones seguidas para generar los parámetros de entrada a las Redes Neuronales que parten de las señales EMG.

Para cada sensor se ha realizado un procesado previo para obtener los datos de entrada a la red neuronal deseados, siguiendo el trabajo realizado en la Universidad de Bremen (Diener, 2021). En la Figura 8, se muestra de manera esquemática los pasos seguidos para obtener las cinco métricas deseadas en cada ventana temporal.

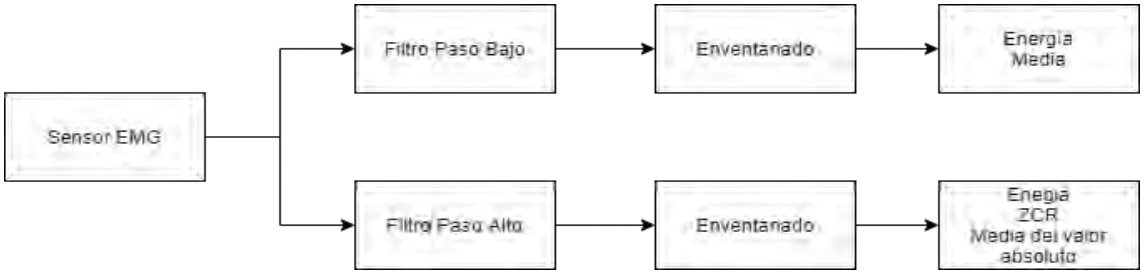


Figura 8. Esquema del procesado de datos.

El filtrado se ha realizado mediante dos métodos diferentes con las mismas características, uno en Python y otro en Matlab.

Los filtros diseñados son, un filtro paso alto y otro paso bajo con las siguientes características.

- Filtro Paso Bajo, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.
- Filtro Paso Alto, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.

Filtro diseñado mediante Matlab

- Filtro Paso Bajo, como se muestra en la Figura 9, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.

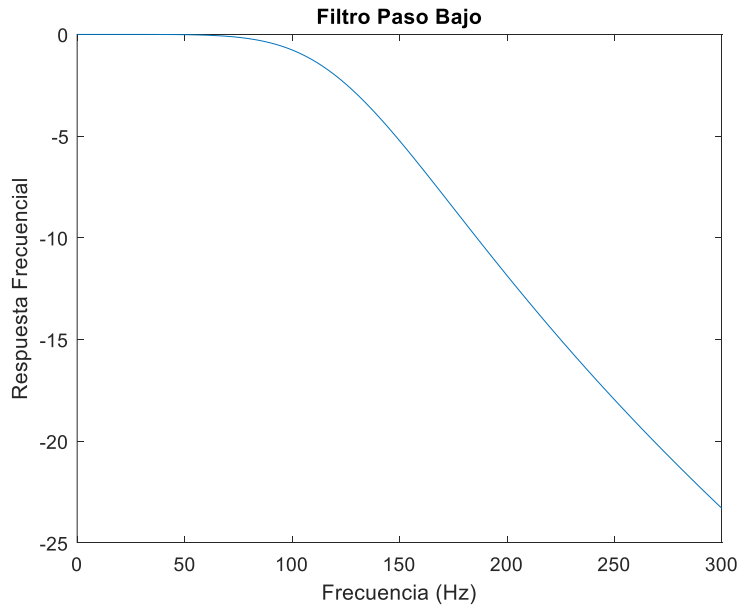


Figura 9. Filtro paso bajo Matlab.

- Filtro Paso Alto, como se muestra en la Figura 10, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.

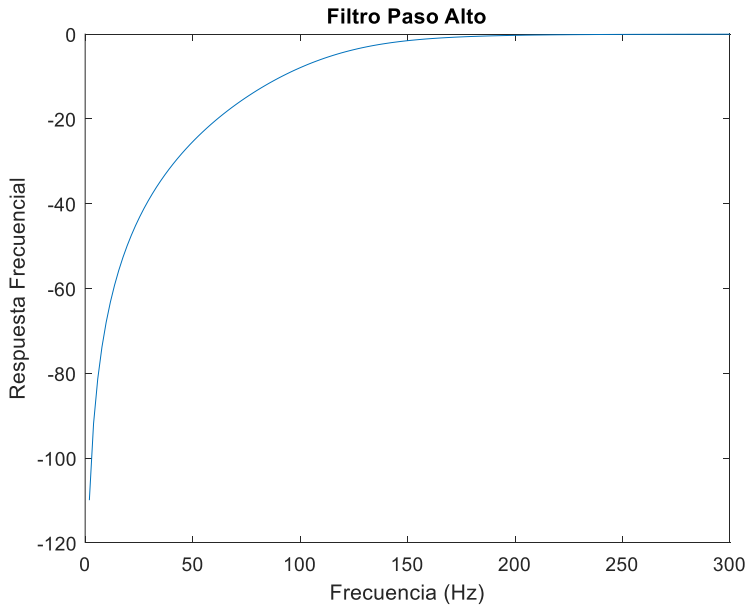


Figura 10. Filtro paso alto Matlab.

Filtro diseñado mediante Python

- Filtro Paso Bajo, como se muestra en la Figura 11, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.

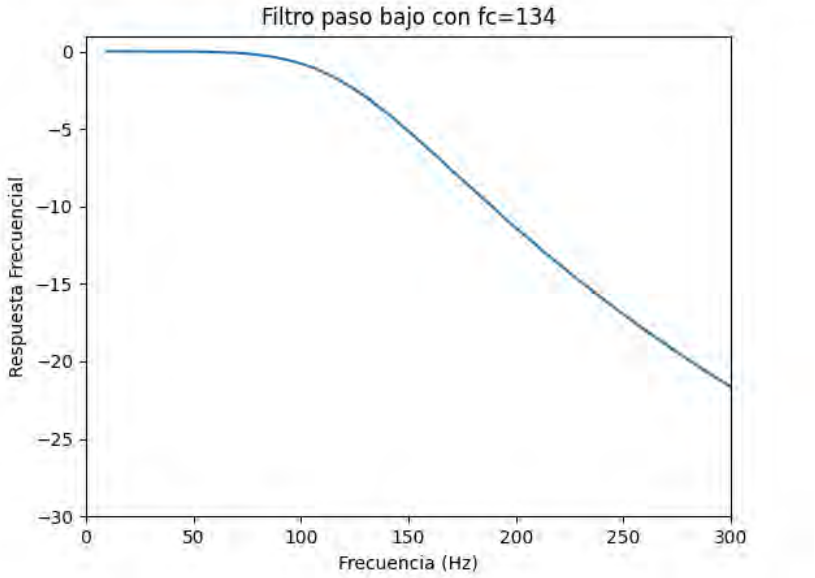


Figura 11. Filtro paso bajo Python.

- Filtro Paso Alto, como se muestra en la Figura 12, se trata de un filtro Butterworth de orden 3 y frecuencia de corte de 134 Hz.

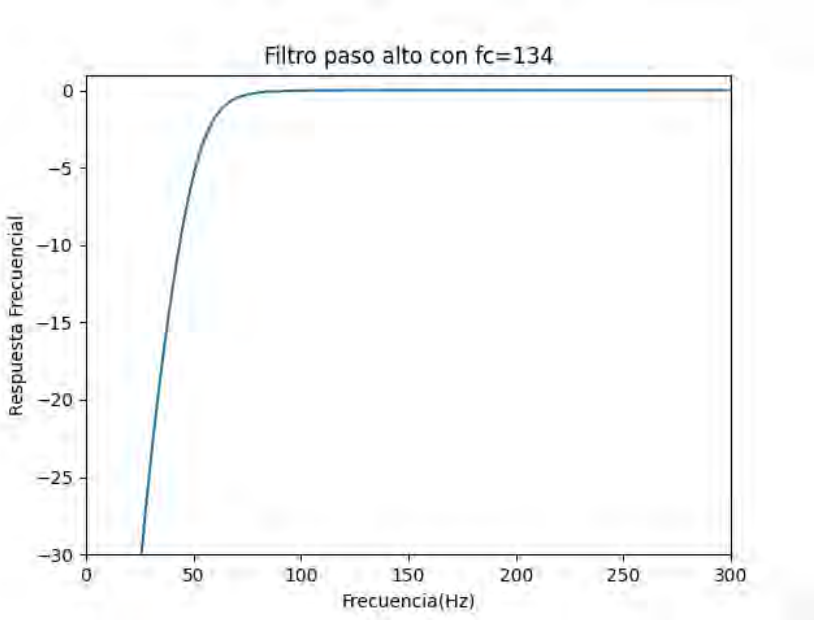


Figura 12. Filtro paso alto Python.

Para obtener más información, se ha realizado el espectrograma del mismo sensor de la Base de Datos. Como se puede comprobar en la Figura 13, en la señal del sensor EMG original mediante ambos filtros, la potencia de la señal se encuentra distribuida aunque predomina la potencia en las bajas frecuencias. Cuando esta señal es filtrada ocurre un cambio de la distribución de la potencia.

La potencia de la señal del sensor que ha sido filtrada mediante el filtro paso bajo queda distribuida en las bajas frecuencias, sin embargo, la potencia de señal del sensor que ha sido filtrada mediante el filtro paso alto, está almacenada en las altas.

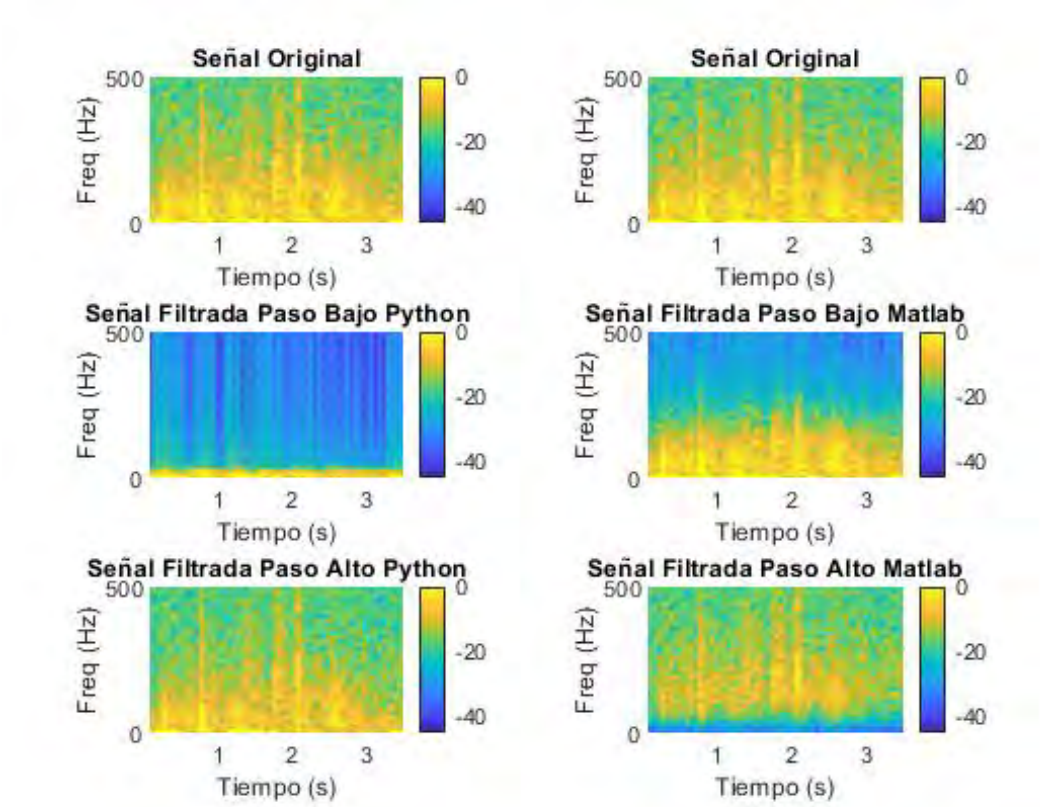


Figura 13. Espectrograma de la señal del Sensor.

En la Figura 14, se muestra la señal original en tiempo de un sensor elegido aleatoriamente filtrada tanto por el Filtro Paso Bajo como por el Filtro Paso Alto mediante el método de Python. El resultado obtenido en Matlab es similar, pero ofrece menor resolución. Como se puede comprobar, tras pasar por el filtro paso bajo disminuye la velocidad de la señal, tal y como era de esperar.

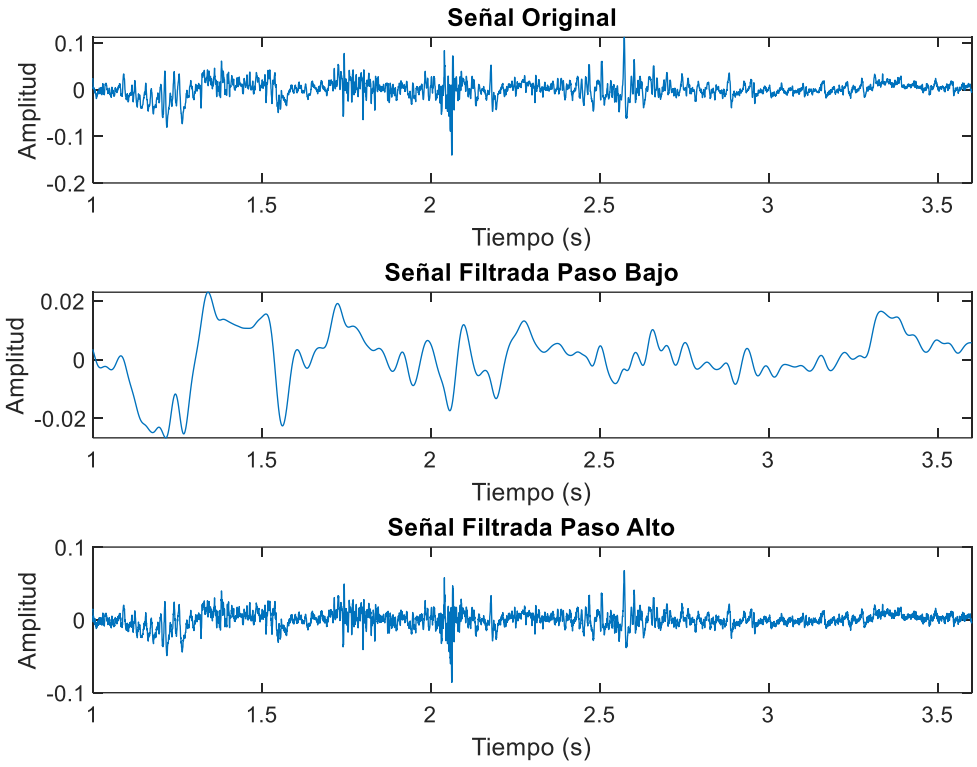


Figura 14. Señal del Sensor filtrado.

Siguiendo con el trabajo de la universidad de Bremen, se ha seleccionado una ventana Blackman de 32 ms y un desplazamiento entre ventanas de 10 ms. Con la peculiaridad de que la frecuencia de muestreo de la señal EMG (2048 Hz) y la señal de audio (16 KHz) no son múltiplo una de la otra. Por lo tanto, para solucionar este problema, se ha realizado la conversión temporal a muestra, redondeando a la muestra más cercana desde la cual seleccionar 40 muestras correspondientes a los 32 ms de la ventana temporal y realizando un desplazamiento temporal de 10 ms. De esta manera, se aprovecha la ventaja de que la duración temporal si es la misma.

Antes de realizar el cálculo correspondiente, se multiplica la trama de señal filtrada por la ventana Blackman. De esta manera, se calcula en su caso el valor de la energía, la media y Tasa de Cruce por Cero (ZCR, Zero Crossing Rate) para la señal filtrada y enventanada.

Cálculo: para cada trama de la señal filtrada se han calculado los siguientes parámetros:

- En la señal filtrada con el Paso Bajo se han calculado la energía y la media de la señal en cada ventana.
- En la señal filtrada con el Paso Alto se han calculado la energía, la Tasa de Cruce por Cero (ZCR) y la media del valor absoluto de la señal en cada ventana.

A continuación, se detalla cómo se han calculado cada uno de estos parámetros.

- **Energía media a corto plazo**

La energía asociada con el habla varía con el tiempo y se calcula la energía asociada con la región del habla a corto plazo.

Siendo x la trama de la señal, n la muestra correspondiente y N la longitud de la trama o número de observaciones, la energía media se calcula como:

$$\boldsymbol{E = \frac{\sum |x[n]|^2}{N}}$$

- **Media**

Se trata del valor medio que tiene la señal en una trama concreta.

Se calcula conforme a la siguiente expresión, en la que x es la trama de la señal, n la muestra correspondiente y N la longitud de la trama o número de observaciones:

$$\boldsymbol{Media = \frac{\sum_1^N x[n]}{N}}$$

- **Tasa de cruce por cero (ZCR)**

La tasa de cruce por cero proporciona información sobre el número de cruces por cero presentes en una señal determinada, en este caso la señal de voz. Así mismo, el ZCR proporciona información indirecta sobre el contenido de frecuencia de la señal.

Siendo x la trama de la señal, n la muestra correspondiente y N la longitud de la trama o número de observaciones, el ZCR se calcula según la siguiente expresión:

$$\boldsymbol{ZCR = \frac{1}{2N} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|}$$

En la Figura 15, se muestra la energía y la media de la señal EMG filtrada paso bajo de un único sensor. Como se ha comentado, esta señal filtrada se divide en ventanas para calcular la media y la energía correspondiente para la ventana.

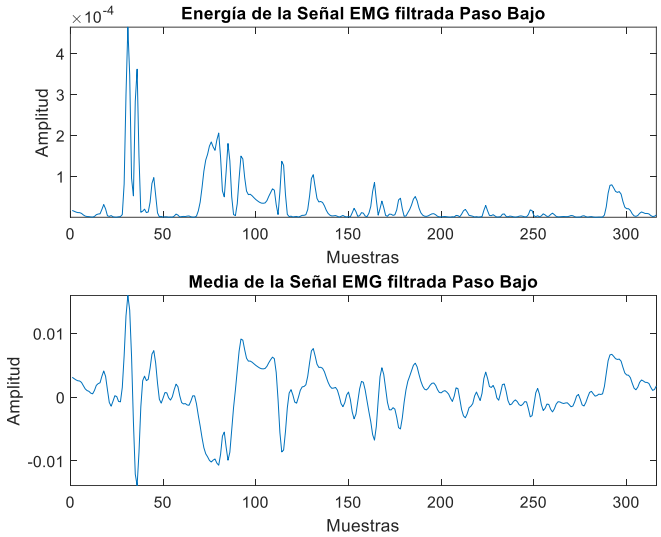


Figura 15. Parámetros obtenidos a partir de la banda baja de la señal.

En la Figura 16, se muestra la media, la energía y el ZCR de la señal EMG filtrada paso alto de un único sensor. Estas señales se dividen en ventanas para calcular la media, la energía y el ZCR correspondiente para la ventana.

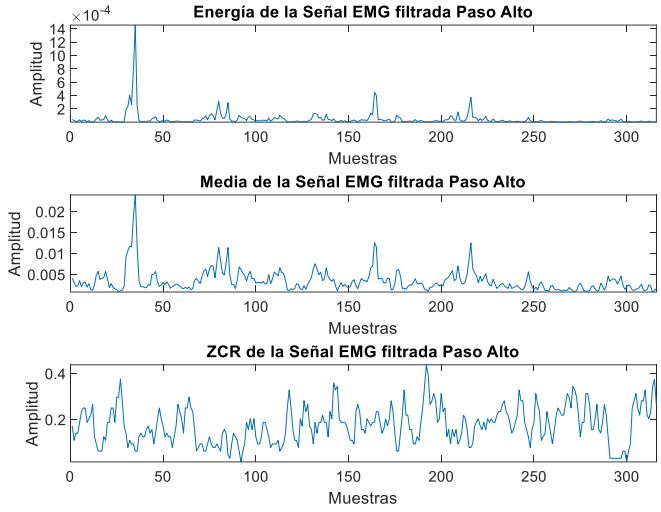


Figura 16. Parámetros obtenidos a partir de la banda alta de la señal.

Además, es necesario mencionar que la variabilidad entre sensores contiguos es muy pequeña como se muestra en la Figura 17, cuyo coeficiente de correlación es 0'64. Para comprobar la variabilidad entre los sensores consecutivos, se ha calculado el coeficiente de correlación medio entre las señales de los sensores para el bloque de entrenamiento, obteniendo un valor de 0'2. El coeficiente de correlación más alto que se ha obtenido ha sido 1, obtenido entre sensores contiguos, es decir, que en este caso se ha captado la misma señal. Por otro lado, el coeficiente de correlación más bajo que se ha obtenido ha sido 1'11e-05, obtenido entre sensores no consecutivos.

La correlación entre sensores que complica el aprendizaje de las Redes Neuronales, ya que los valores de los parámetros calculados serán muy similares entre sensores. Además de la posibilidad de poder trabajar con menos sensores.

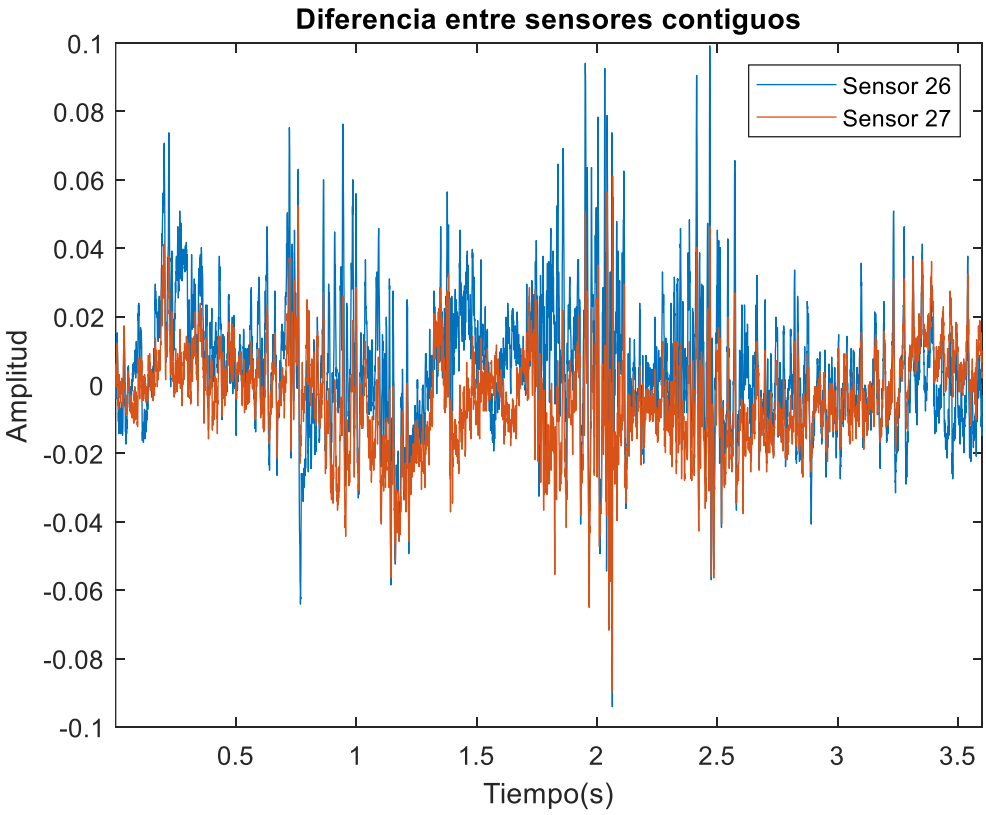


Figura 17. Diferencia entre sensores contiguos.

7.2.2. Procesado de datos de salida

En este apartado, se detalla la obtención de los diferentes parámetros de salida para el aprendizaje de las Redes Neuronales, como son los Coeficientes Cepstrales y el Espectrograma correspondientes a las señales de entrada procesadas del EMG y la Frecuencia Fundamental correspondientes a los Coeficientes Cepstrales.

Coeficientes Cepstrales y Frecuencia Fundamental

La red neuronal también necesita datos de salida para el entrenamiento. En este caso para extraer parámetros acústicos a partir de la señal de voz se ha seleccionado el vocoder Ahocoder (Erro, Sainz, Navas, & Hernaez, 2014), desarrollado por el grupo de investigación Aholab para obtener los Coeficientes Cepstrales y la Frecuencia Fundamental de cada señal de audio. En la Figura 18, se muestran los Coeficientes Cepstrales y la Frecuencia Fundamental de una señal de audio seleccionada aleatoriamente de la Base de Datos.

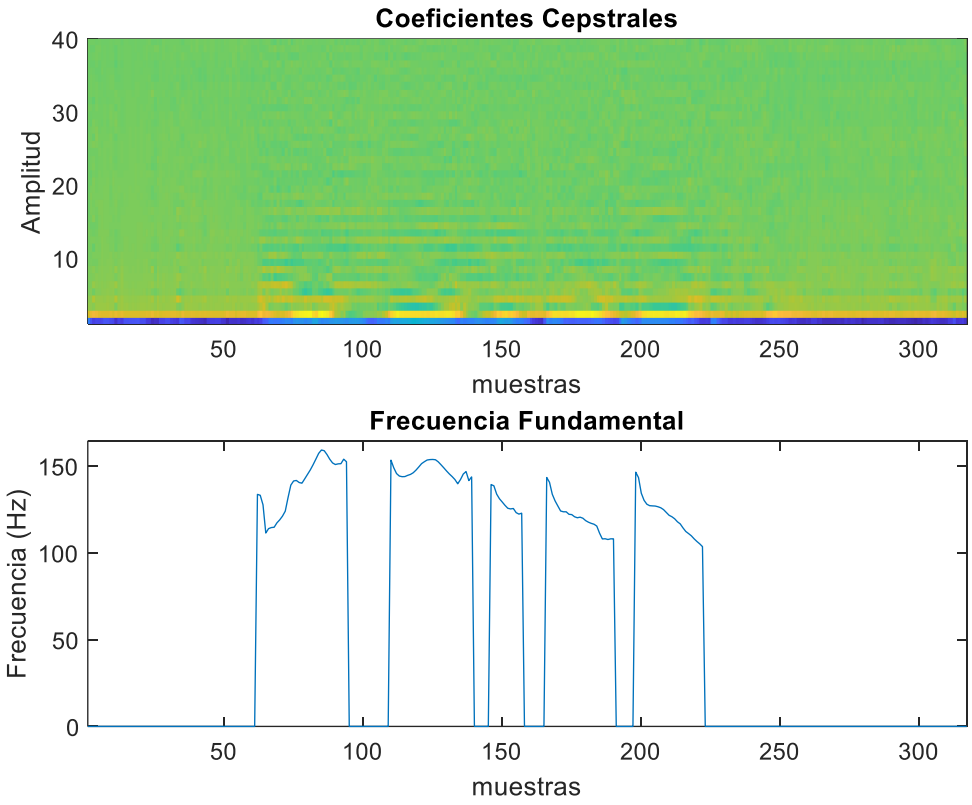


Figura 18. Parámetros de salida de Coeficientes Cepstrales y F0.

Espectrograma

Para la segunda alternativa que se ha planteado, se ha utilizado el espectrograma de las señales de audio para realizar la predicción del mismo. Para ello, se han utilizada las propias herramientas de Matlab. En la Figura 19, se muestra la señal de audio con su correspondiente Espectrograma.

Para calcular el espectrograma, primero se ha diezmado la señal de audio a 8000 Hz, con el objetivo de centrar el análisis en los primeros 4000 Hz que contienen la mayor parte de información. Para ajustar el espectrograma a la señal del EMG, se ha utilizado la siguiente configuración:

- Ventana hamming de 32ms
- Desplazamiento de 10ms
- 256 puntos de la FFT

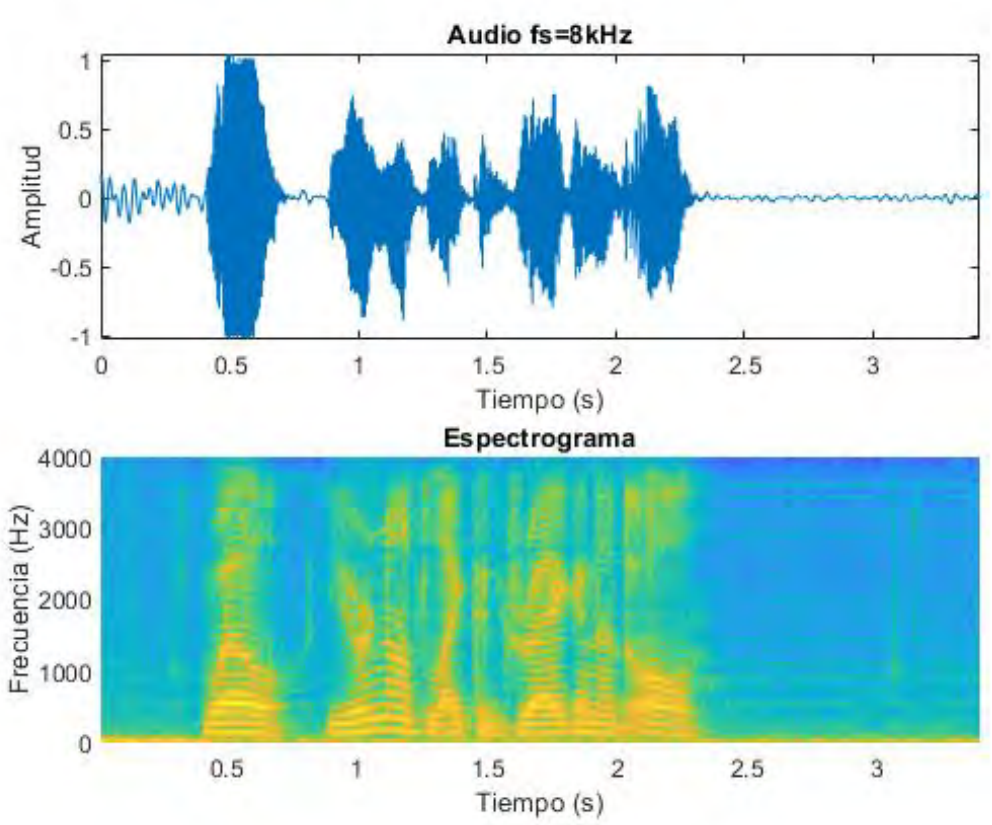


Figura 19. Espectrograma de salida.

7.3. Experimentos

En este bloque se presentan los diferentes experimentos realizados con las Redes Neuronales, con el objetivo de analizar mediante métricas cuál ofrece una mayor precisión y decidir cuál será el sistema de referencia seleccionado y obtener unas conclusiones.

7.3.1. Redes Neuronales de Coeficientes Cepstrales

Se han realizado tres Redes Neuronales para predecir los coeficientes Cesptrales con ligeras diferencias.

- Red Neuronal Básica
- Red Neuronal con Regularización
- Red Neuronal con BatchNormalization

Todas las Redes Neuronales utilizadas para predecir Coeficientes Cepstrales, se tratan de un modelo secuencial, con una Estructura de Cuellos de Botella (Bottleneck Network Structure). Como se muestra en la Tabla 4, están formadas por tres capas de 2048, 512 y 1024 valores de entrada a cada capa y una última capa de salida de 40 valores.
Las Redes Neuronales han sido entrenadas 500 épocas, usando descenso de gradiente estocástico con una tasa de aprendizaje de 0'01 y la función de pérdidas utilizada es la del Error Cuadrático Medio.

Modelo Secuencial		
Capa	**Forma de Salida**	**Parámetros**
Dense	(None, 2048)	155648
Dropout	(None, 2048)	0
Dense	(None, 512)	1049088
Dropout	(None, 512)	0
Dense	(None, 1024)	525312
Dropout	(None, 1024)	0
Dense	(None, 40)	41000
Parámetros totales: 1771048 Parámetros entrenados: 1771048 Parámetros no entrenados: 0		

Tabla 4. Resumen del modelo de la Red Neuronal de Coeficientes Cepstrales

Los datos de entrada, denominados los ficheros CTD-15, requieren una estructura compleja, que viene explicada a continuación. Cada bloque de datos de entrada a la red neuronal está formado por 3000 valores. Estos valores corresponden a la energía de la señal filtrada paso bajo, la media de la señal filtrada paso bajo, la energía de la señal filtrada paso alto, el ZCR de la señal filtrada paso alto y la media del valor absoluto de la señal filtrada paso alto de la ventana actual, además de los valores de las 14 ventanas anteriores de cada uno de los sensores. En el caso de que la trama actual no cuente con alguna de las 14 ventanas será sustituido por ceros. Para mejorar los resultados, estos valores serán normalizados antes de entrar a la Red Neuronal.

Como se muestra en la Figura 20, la estructura de datos está compuesta por 75 valores de cada sensor (5 valores corresponden a la ventana actual y 70 a las 14 ventanas anteriores). De esta manera, para cada ventana temporal se almacenan estos 75 valores de cada sensor, formando un array de 75x40 =3000 valores.

	Ventana 20	Ventana 21	Ventana 22	...	Ventana n
Sensor 0	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	...	Energía Media Energía ZCR Media ABS
Sensor 1	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	...	Energía Media Energía ZCR Media ABS
...					
Sensor 39	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	Energía Media Energía ZCR Media ABS	...	Energía Media Energía ZCR Media ABS

Ventana 20 (3000)	Ventana 21 (3000)	Ventana 22 (3000)	...	Ventana n
Ventana 20 (5) Ventana 6:19 (70)	Ventana 21 (5) Ventana 7:20 (70)	Ventana 22 (5) Ventanas 8:21 (70)	...	Ventana n Ventanas n-14:n-1
Ventana 20 (5) Ventana 6:19 (70)	Ventana 21 (5) Ventana 7:20 (70)	Ventana 22 (5) Ventanas 8:21 (70)	...	Ventana n Ventanas n-14:n-1
...	...	...	...	...
Ventana 20 (5) Ventana 6:19 (70)	Ventana 21(5) Ventana 7:20 (70)	Ventana 22 (5) Ventanas 8:21 (70)	...	Ventana n Ventanas n-14:n-1

Figura 20. Estructura de datos de datos de entrada CTD-15.

Por último, en la Figura 21, se muestra una representación no fiel de la arquitectura de la Red Neuronal de los Coeficientes Cepstrales. Como se ha mencionado anteriormente, los datos de entrada están formados por 3000 valores, la Red Neuronal está formada por tres capas ocultas de 2048, 512 y 1024 valores de entrada respectivamente y finalmente la salida formado por 40 valores que corresponden a los Coeficientes Cepstrales.

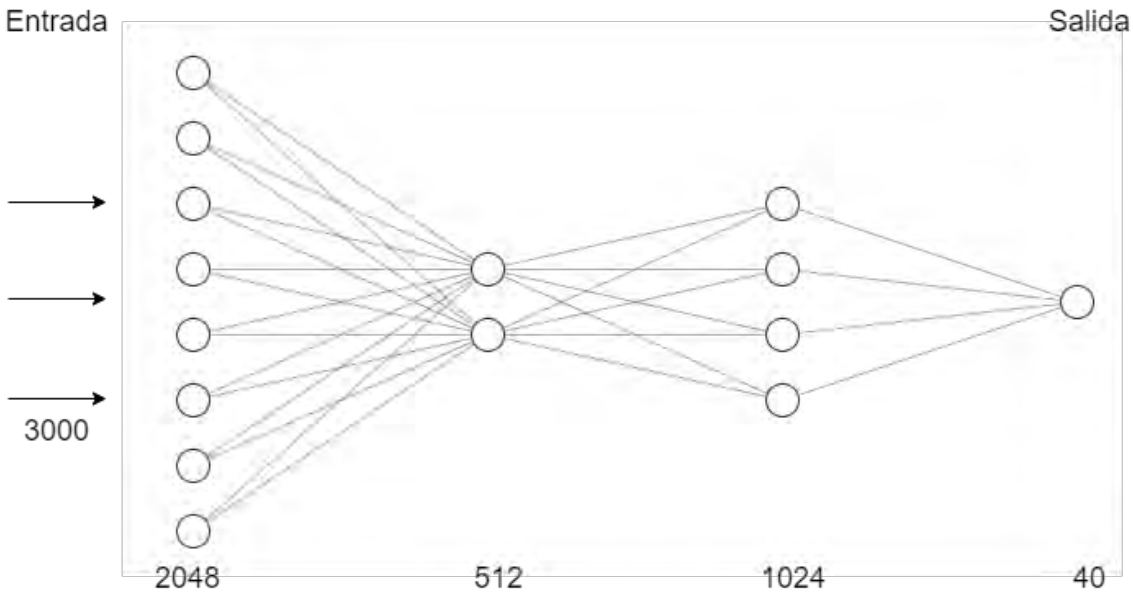


Figura 21. Representación gráfica de la Red Neuronal de CC.

Red Neuronal Básica

En esta alternativa, se plantea la Red Neuronal como se ha explicado previamente.

Red Neuronal con Regularización

En esta alternativa, se ha añadido regularización L2 de 0'001 en la salida de cada capa, con el objetivo de obtener un resultado más adaptable en la predicción.

Red Neuronal con BatchNormalization

En esta alternativa, se ha añadido la normalización por lotes, que es un método que se utiliza para hacer que las redes neuronales sean más rápidas y estables mediante la normalización de las entradas de cada capa al volver a centrar y escalar la salida de cada capa antes de usarlas como entrada a la capa siguiente.

7.3.2. Red Neuronal de Frecuencia Fundamental Completa

Para predecir la Frecuencia Fundamental se han diseñado dos redes neuronales diferentes, una de ella es la encargada de predecir el valor de la Frecuencia Fundamental para cada bloque de 40 Coeficientes Cepstrales. La otra red neuronal es la encargada de predecir si la señal es sorda (0) o sonora (1) para cada bloque de 40 Coeficientes Cepstrales. Una vez obtenidos ambos valores se multiplican para obtener el valor final de la Frecuencia Fundamental.

Red Neuronal de Frecuencia Fundamental

La red neuronal utilizada para predecir Coeficientes Cepstrales, se trata de un modelo secuencial, con una Estructura de Cuellos de Botella (Bottleneck Network Structure). Como se muestra en la Tabla 5, está formada por tres capas de 2048, 512 y 1024 valores de entrada a cada capa y con regularización de abandonos después de cada capa y una última capa de salida de 1 valor. La red neuronal ha sido entrenada 500 épocas, usando descenso de gradiente estocástico con una tasa de aprendizaje de 0'01 y la función de pérdidas utilizada es la del Error Cuadrático Medio.

Modelo Secuencial		
Capa	**Forma de Salida**	**Parámetros**
Dense	(None, 2048)	83968
Dropout	(None, 2048)	0
Dense	(None, 512)	1049088
Dropout	(None, 512)	0
Dense	(None, 1024)	525312
Dropout	(None, 1024)	0
Dense	(None, 1)	1025
Parámetros totales: 1659393 Parámetros entrenados: 1659393 Parámetros no entrenados: 0		

Tabla 5. Resumen del modelo de la Red Neuronal de Frecuencia Fundamental.

En la Figura 22, se muestra una representación no fiel de la arquitectura de la Red Neuronal de Frecuencia Fundamental. Se puede observar la entrada formada por 40 valores y la salida de 1 valores correspondiente al valor de la frecuencia fundamental.

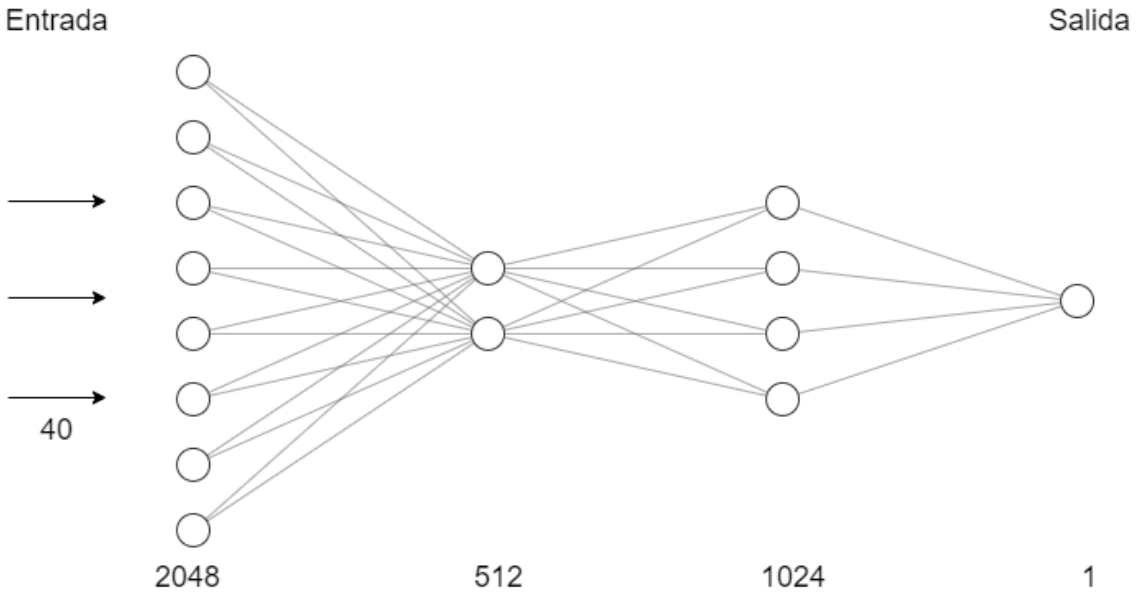


Figura 22. Representación gráfica de la Red Neuronal de Frecuencia Fundamental.

Red Neuronal de Sonidos Sordos y Sonoros

La red neuronal utilizada para predecir los sonidos sordos y sonoros, se trata de un modelo secuencial. Como se muestra en Tabla 6, está formada por la capa de entrada, una capa intermedia y la capa de salida y con regularización de abandonos después de cada capa intermedia. La red neuronal ha sido entrenada 20 épocas, usando como optimizador el modelo Adam. Para la función de pérdidas se ha realizado un análisis entre la función del Error Cuadrático Medio y la función de la Entropía Binaria Cruzada. Finalmente, tras realizar la evaluación se ha decidido que la función con la que se alcanza una mayor precisión es la función de pérdidas de la Entropía Binaria Cruzada, con un 0'1% más de acierto.

Cuando se obtiene la predicción los valores superiores a 0'5 se toman como sonoros (1) y cuando es menor que 0'5 se toma como sordos (0). Luego, el valor predicho (0 o 1) se multiplica por la predicción del valor de F0.

Modelo Secuencial		
Capa	**Forma de Salida**	**Parámetros**
Dense	(None, 64)	2624
Dropout	(None,64)	0
Dense	(None, 1)	65
Parámetros totales: 2689 Parámetros entrenados: 2689 Parámetros no entrenados: 0		

Tabla 6. Resumen del modelo de la Red Neuronal de Sonidos Sordos y Sonoros.

En la Figura 23, se muestra una representación de la arquitectura de la Red Neuronal de Sonidos Sordos y Sonoros. En este caso la entrada también está formada por 40 Coeficientes Cepstrales, pero la Red Neuronal únicamente cuenta con una capa y finalmente convergerá en el único valor que predecirá si la trama corresponde a un sonido sordo o sonoro.

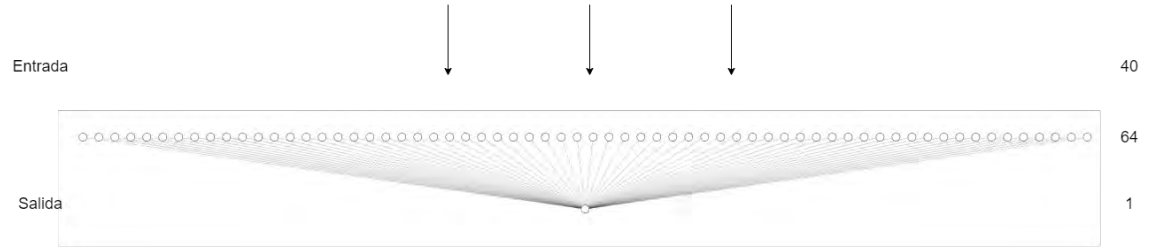


Figura 23. Representación gráfica de la Red Neuronal de Sonidos Sordos y Sonoros.

7.3.3. Red Neuronal de Espectrograma

La red neuronal utilizada para predecir el espectrograma, se trata de un modelo secuencial, con una Estructura de Cuellos de Botella (Bottleneck Network Structure). Como se muestra en la Tabla 7, está formada por tres capas de 2048, 512 y 1024 valores de entrada a cada capa y con regularización de abandonos después de cada capa y una última capa de salida de 257 valores. Además, se ha añadido regularización L2 con valor 0'001 para obtener un resultado más adaptable en la predicción.

La red neuronal ha sido entrenada 500 épocas, usando descenso de gradiente estocástico con una tasa de aprendizaje de 0'01 y la función de pérdidas utilizada es la del Error Cuadrático Medio.

Modelo Secuencial		
Capa	**Forma de Salida**	**Parámetros**
Dense	(None, 2048)	6146048
Dropout	(None, 2048)	0
Dense	(None, 512)	1049088
Dropout	(None, 512)	0
Dense	(None, 1024)	525312
Dropout	(None, 1024)	0
Dense	(None, 257)	263425
Parámetros totales: 7983873 Parámetros entrenados: 7983873 Parámetros no entrenados: 0		

Tabla 7. Resumen del modelo de la Red Neuronal de Espectrograma.

Por último, en la Figura 24, se muestra una representación no fiel de la arquitectura de la Red Neuronal del Espectrograma. La Red Neuronal está formada por tres capas ocultas de 2048, 512 y 1024 valores de entrada respectivamente y finalmente la salida formado por 257 valores que corresponden al espectrograma.

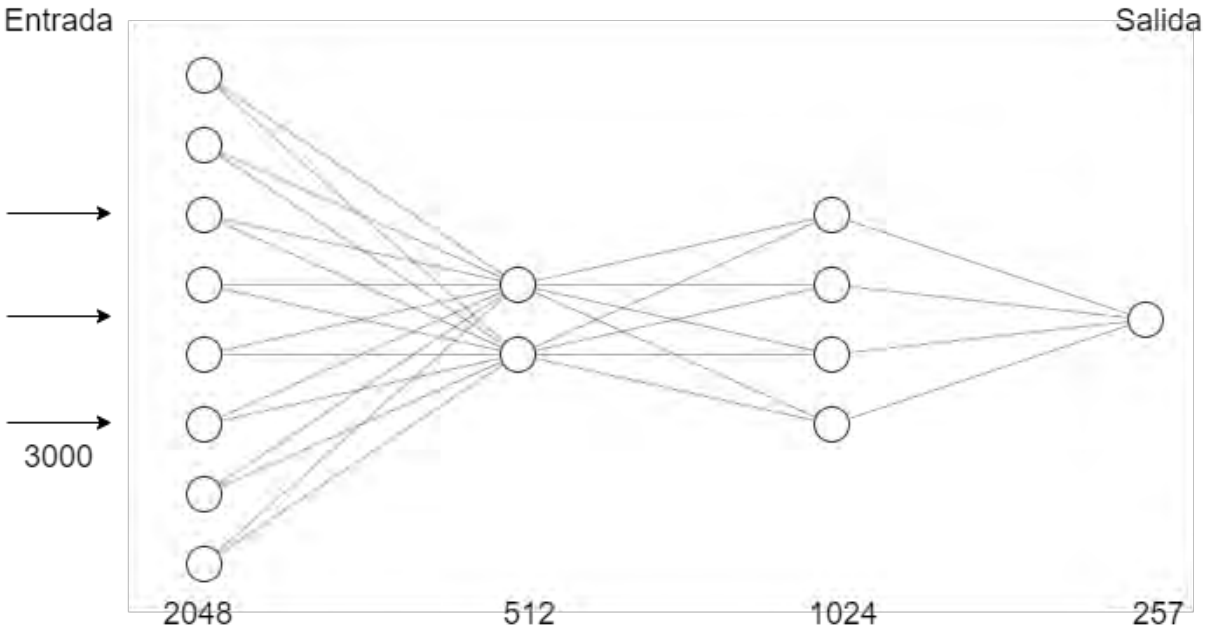


Figura 24 .Representación gráfica de la Red Neuronal del Espectrograma.

8. Cálculos y algoritmos

En esta sección, se recogen los parámetros de medida que se van a utilizar en la evaluación de los resultados obtenidos en las predicciones.

- **Algoritmo para el cálculo de Mel-Cepstral Distortion**

La distorsión Mel-Cepstral, mide cuánto de diferentes son dos secuencias de coeficientes Mel-Cepstrum. Se utiliza para evaluar la calidad de los sistemas de síntesis de voz paramétrica, incluidos los sistemas de síntesis de voz paramétrica estadística, con la idea de que cuanto más pequeño es el MCD entre las secuencias Mel-Cesptrales sintetizadas y naturales, más cerca está la voz sintética de reproducir la voz natural.

Siendo mgc_orig los coeficientes cepstrales de la trama de la señal original, mgc_predichos los coeficientes cepstrales de la trama de la señal predicha, n la trama correspondiente e i la posición dentro de la trama o número índice del coeficiente, la MCD se calcula de acuerdo a la siguiente expresión:

$$MCD = \frac{10}{ln\ 10} \cdot \sqrt{2 \cdot \sum_{i}^{40} |mgc\_orig(n,i) - mgc\_predichos(n,i)|^2}$$

- **Error Cuadrático Medio**

Mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. Siendo i la muestra a analizar y M el número total de muestras el error cuadrático medio se calcula de acuerdo a la siguiente expresión.

$$ECM = \frac{1}{M}\sum_{i=1}^{M}(real[i] - predicción[i])^2$$

- **Matriz de Confusión**

La matriz de Confusión está formada, por cuatro métricas que son TP, FP, FN y TN, que hacen referencia a las diferentes posibilidades que puede ofrecer un sistema de predicción binario.

TP (Verdadero Positivo), se trata de un 1 que el sistema ha calificado como 1.

FP (Falso Positivo), se trata de un 0 que el sistema ha calificado como 1.

TP (Falso Negativo), se trata de un 1 que el sistema ha calificado como 0.

TP (Verdadero Negativo), se trata de un 0 que el sistema ha calificado como 0.

Matriz de Confusión		
Original / Predicción	**1**	**0**
1	TP (Verdadero Positivo)	FP (Falso Positivo)
0	FN (Falso Negativo)	TN (Verdadero Negativo)

- **Precisión**

Mediante la métrica de precisión podemos medir la calidad del modelo de clasificación. Indica la cantidad de aciertos entre los ejemplos clasificados como pertenecientes a la clase de interés.

Siendo TP los verdaderos positivos y FP los falsos positivos la precisión se calcula como:

$$Precisión = \frac{TP}{TP + FP}$$

- **Recall o Exhaustividad**

Mediante la métrica de exhaustividad, se obtiene información sobre la fracción de instancias relevantes que el modelo es capaz de identificar.

Siendo TP los verdaderos positivos y FN los falsos negativos, la recall se calcula de acuerdo a la siguiente ecuación:

$$Recall = \frac{TP}{TP + FN}$$

- **Fscore**

El valor Fscore se utiliza para combinar las medidas de precisión y recall en un solo valor. Es útil para comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.

$$Fscore = 2 \cdot \frac{Precisión \cdot Recall}{Precisión + Recall}$$

- **Distancia Euclídea**

La distancia euclídea se trata de la distancia más corta entre dos puntos en cualquier dimensión.

Siendo x e y las muestras correspondientes a un punto y n el número total de puntos.

$$distancia[x, y] = \sqrt{\sum_{i=1}^{n} (y[i] - x[i])^2}$$

9. Análisis de los resultados

9.1. Red Neuronal de Coeficientes Cepstrales Filtrado Matlab

En el siguiente apartado, se presentan los resultados de entrenamiento y resultados obtenidos para los datos procesados mediante el filtro diseñado en Matlab.

Una vez diseñadas las redes neuronales es necesario comprobar el valor de las pérdidas. En la Figura 25, Figura 26 y Figura 27, se muestran las pérdidas en el conjunto de datos de entrenamiento y validación para la red básica, la red con regularización y la red con BatchNormalization respectivamente. Todas ellas tienen un valor de pérdidas finales muy similares.

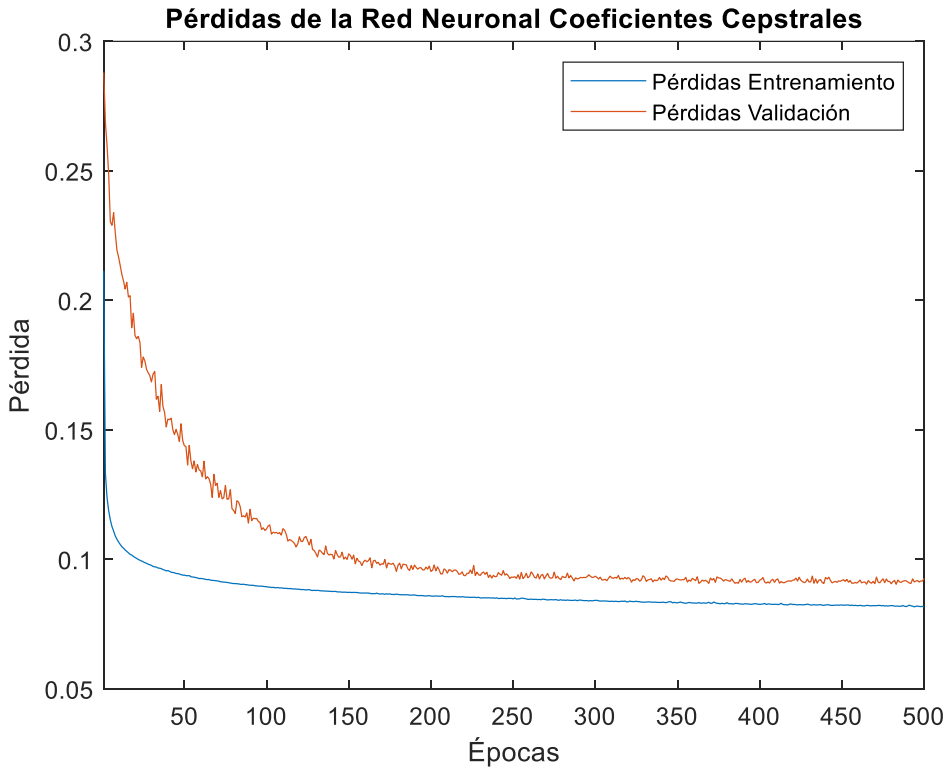


Figura 25. Pérdidas Red Neuronal Inicial.

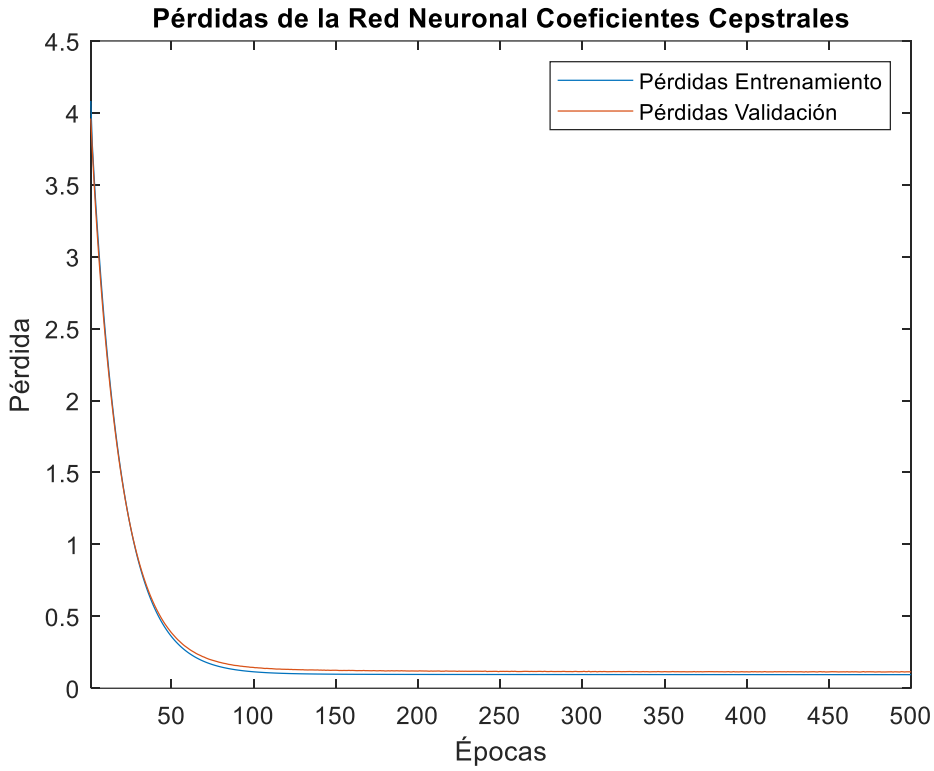


Figura 26. Pérdidas Red Neuronal con Regularización.

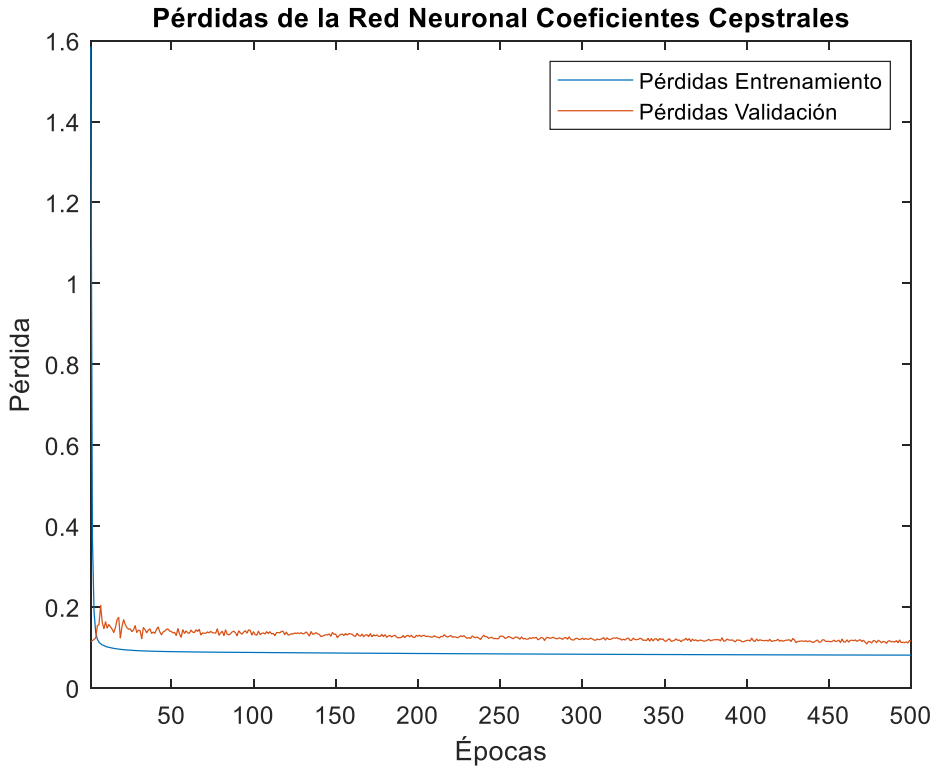


Figura 27. Pérdidas Red Neuronal con BatchNormalization.

Para medir con más exactitud la precisión de la predicción de los Coeficientes Cepstrales, se ha calculado el valor de MCD, para tres bloques de evaluación (formados por cincuenta frases) del Locutor 1. En la Figura 28, se muestra el resultado del análisis mencionado anteriormente sobre el Locutor 1 para los tres modelos de redes, la red básica (azul), la red con regularización (rojo) y la red con BatchNormalization (verde). El mejor resultado se da en el segundo bloque de evaluación con un valor de distorsión de 9'24 dB. En todos los casos la red básica obtiene resultados ligeramente mejores que las otras dos opciones. Por esta razón y además de que el promedio de los tres bloques es el más bajo para esta red, siendo 10'86 dB, se ha decidido que el Modelo de Red Neuronal más adecuado en el caso de aplicar el filtrado con MATLAB es el Modelo Básico.

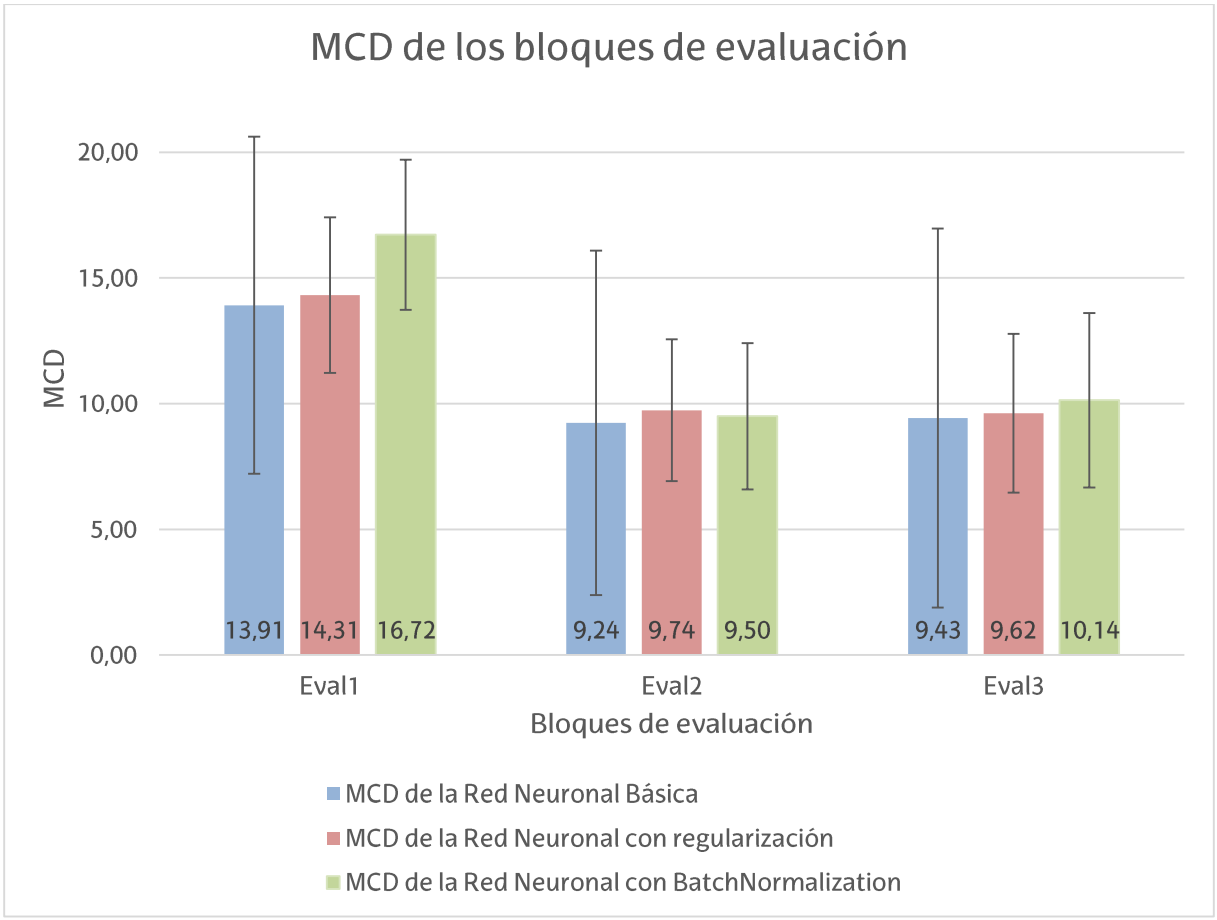


Figura 28. Mel-Cepstral Distortion promedio de los bloques de evaluación. Las barras de error muestran la desviación estándar.

Mediante el Modelo de Red Neuronal Básico, se ha realizado una predicción de los Coeficientes Cepstrales para una señal elegida aleatoriamente de un bloque de evaluación. En la Figura 29, se muestra la comparación gráfica entre los Coeficientes Cepstrales originales y predichos. Como se puede observar en la banda baja existen diferencias de potencia, sobre todo en las primeras tramas donde la potencia debería ser menor que la predicha.

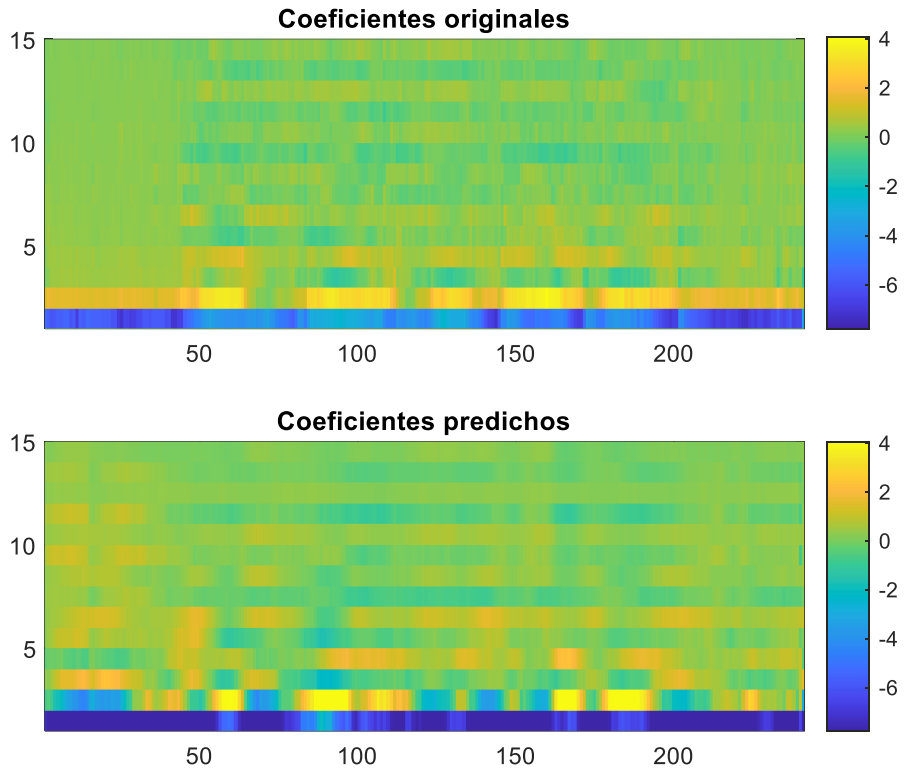


Figura 29. Comparación CC originales y predichos.

Por último, en la Figura 30, se han comparado las pérdidas de las tres variantes analizadas en el proyecto. Tanto la Red Neuronal inicial como la de BatchNormalization tiene un aprendizaje más rápido, sin embargo, la Red Neuronal con regularización aprende más despacio para adaptarse más a las muestras de validación, pero finalmente obtiene un resultado similar.

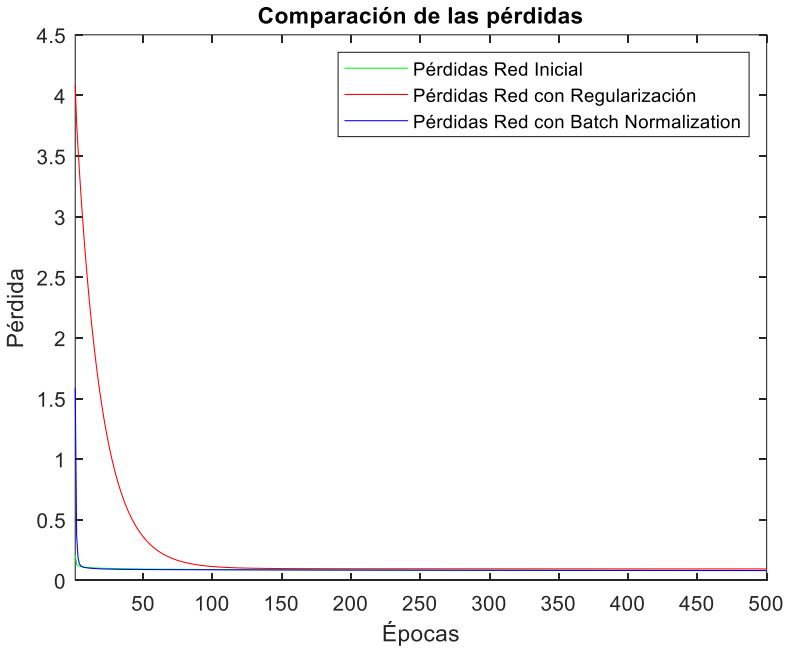


Figura 30. Comparativa de las pérdidas de los tres modelos.

9.2. Red Neuronal de Coeficientes Cepstrales Filtrado Python

En el siguiente apartado, se presentan los resultados de entrenamiento y resultados obtenidos para los datos procesados mediante el filtro diseñado en Python.

Una vez diseñadas las redes neuronales es necesario comprobar el valor de las pérdidas. En la Figura 31, Figura 32 y Figura 33, se muestran las pérdidas en el conjunto de datos de entrenamiento y validación para la red básica, la red con regularización y la red neuronal con BatchNormalization respectivamente. En este caso la red neuronal básica no se adapta bien al conjunto de datos de validación debido a que se sobre ajusta al conjunto de datos de entrenamiento. Sin embargo, tanto la red con regularización como la red neuronal con BatchNormalization, presentan unas perdidas similares.

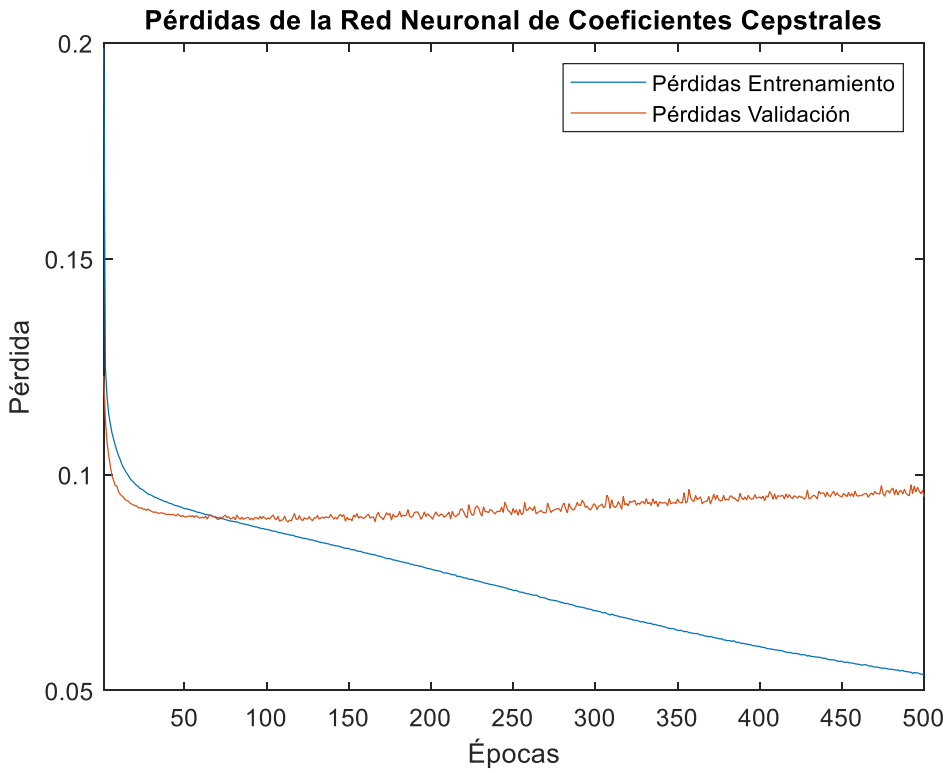


Figura 31. Pérdidas Red Neuronal Inicial.

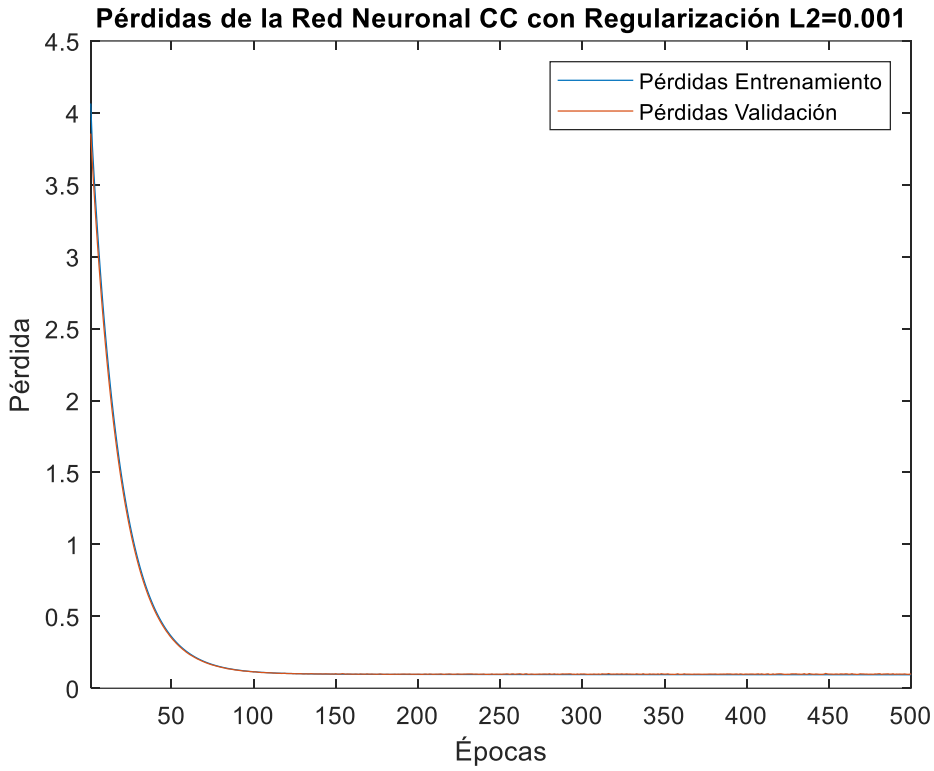


Figura 32. Pérdidas Red Neuronal con Regularización.

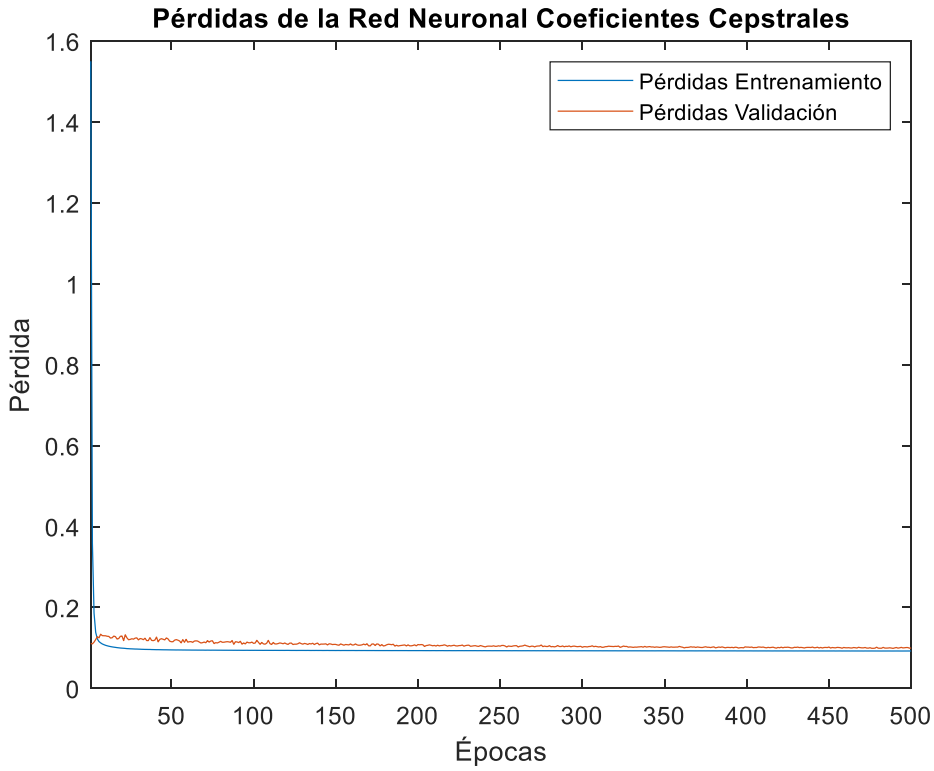


Figura 33. Pérdidas Red Neuronal con BatchNormalization.

Para evaluar con más precisión la predicción de los Coeficientes Cepstrales, se ha calculado el valor de MCD, para tres bloques de evaluación (formados por cincuenta frases) del Locutor 1. En la Figura 34, se muestra el resultado de este análisis. Como ninguna de las redes se obtiene el menor MCD en todos los bloques de evaluación, se ha calculado el promedio de las tres, siendo 9'5 dB, 9'37 dB y 9'44 dB respectivamente.

Por lo tanto, se ha decidido que el Modelo de Red Neuronal más adecuado en el caso de aplicar el filtro implementado en Python es el Modelo con regularización, debido a que en promedio es el que menor MCD obtiene.

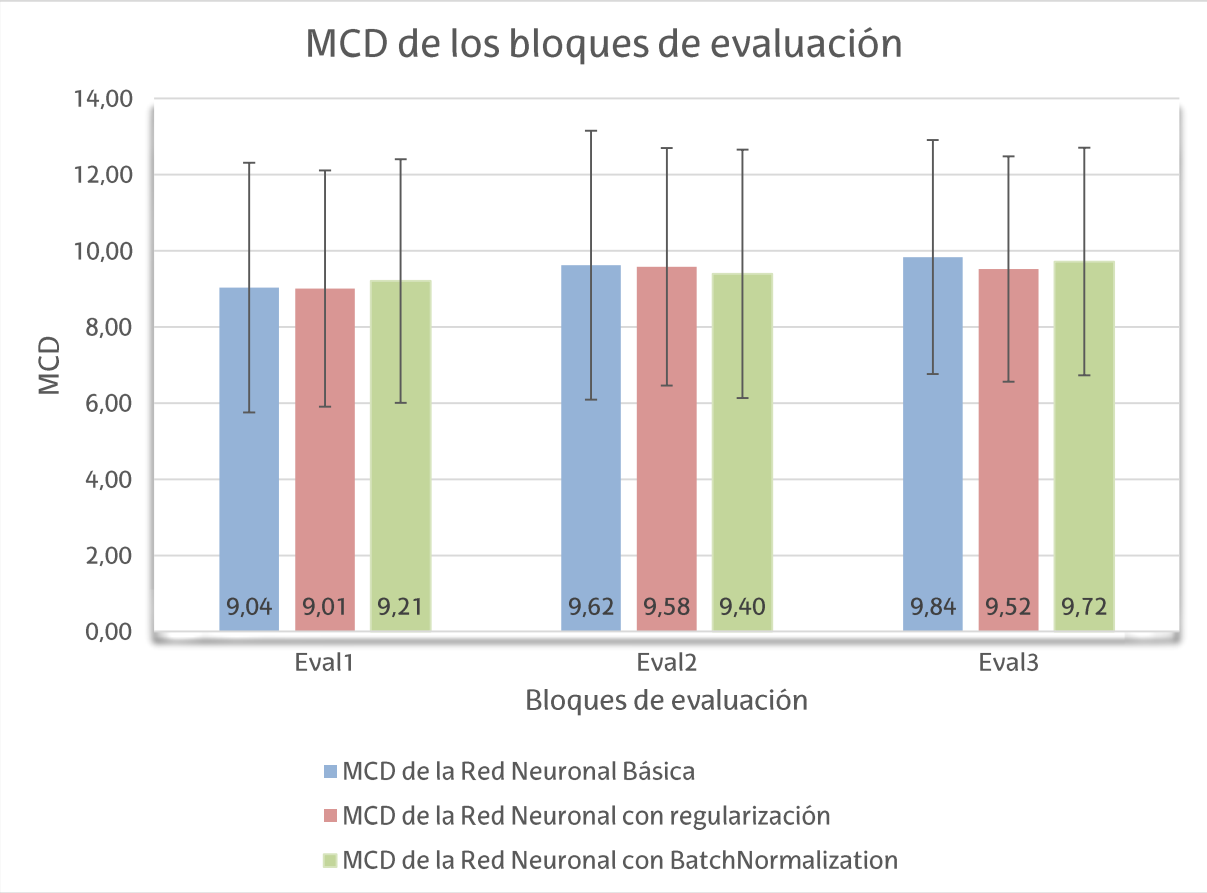


Figura 34. Mel-Cepstral Distortion promedio de los bloques de evaluación. Las barras de error muestran la desviación estándar.

Mediante el Modelo de Red Neuronal con regularización, se ha realizado una predicción de los Coeficientes Cepstrales para la misma señal elegida que en la Figura 29. En este caso, se ha obtenido un valor del MCD de 8'59 dB, que se trata de un resultado inferior al promedio de todos los bloques de evaluación.

En la Figura 35, se muestra la comparación gráfica entre los Coeficientes Cepstrales originales y predichos. En este caso, en la banda baja la predicción es bastante similar. Principalmente las diferencias se encuentran en la banda intermedia donde no se ajusta lo suficiente al detalle de los Coeficientes Cepstrales originales.

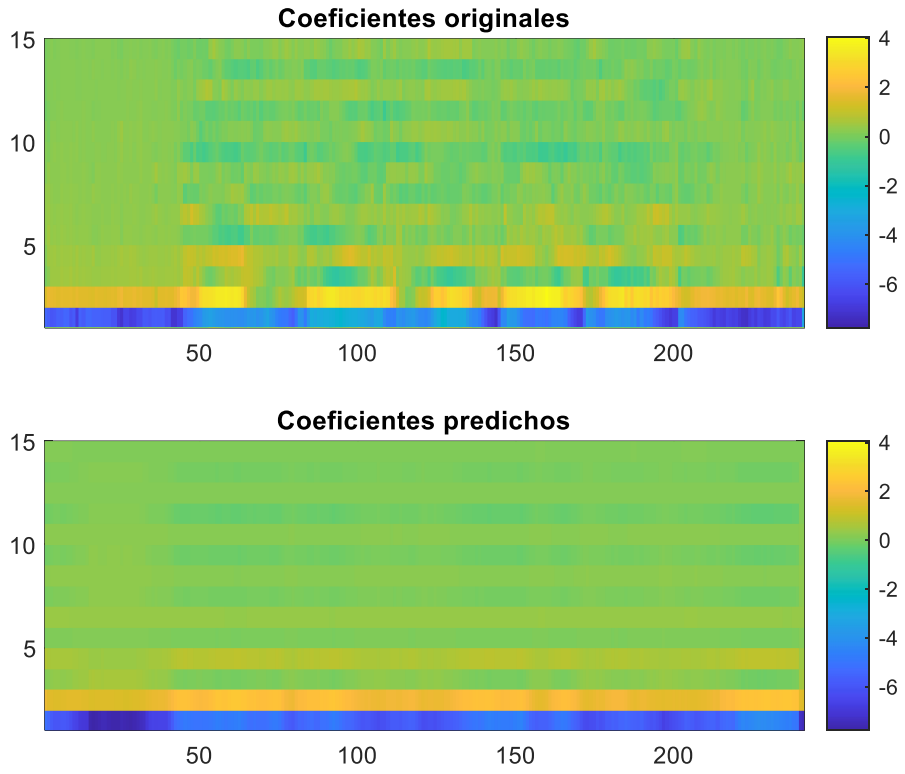


Figura 35. Comparación CC originales y predichos.

9.3. Red Neuronal de Coeficientes Cepstrales Comparativa

En este apartado se muestran los mejores resultados obtenidos para cada modo de filtrado con los publicados por la Universidad de Bremen.

Comparando los resultados promedios obtenidos en este trabajo y que se han presentado en la Figura 28 y la Figura 34. con los publicados por la Universidad de Bremen que se muestran en la Figura 36. Se puede concluir finalmente que la mejor combinación se trata del filtrado mediante Python con el modelo de red neuronal con regularización.

El filtrado de Matlab queda descartado debido a que al no ser tan preciso las señales filtradas mediante el filtro paso bajo no ofrecen tanta precisión como la obtenida con Python que se trata de un filtro con una bajada más abrupta.

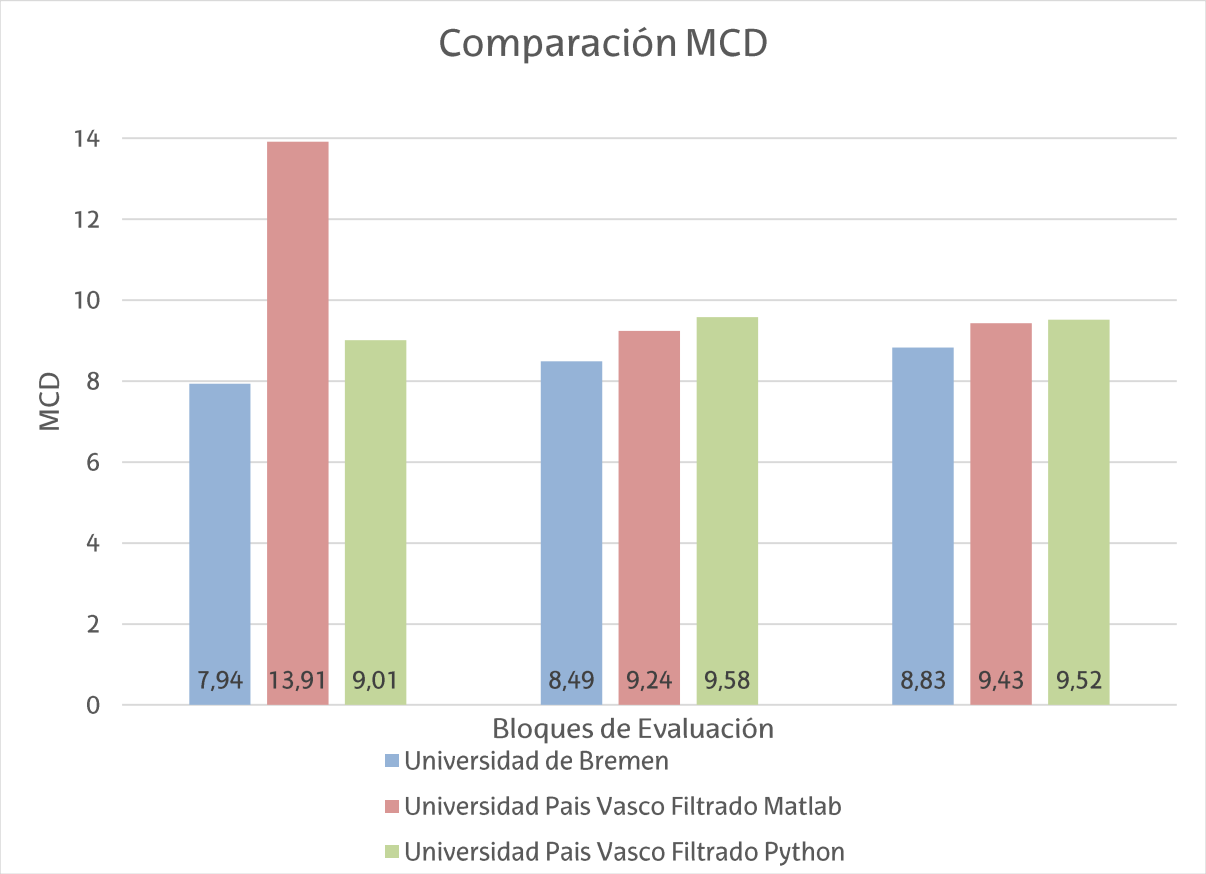


Figura 36. Comparación de resultados MCD promedio.

9.4. Red Neuronal de Frecuencia Fundamental

Como se puede comprobar en la Figura 37, las perdidas en el conjunto de datos de entrenamiento disminuyen hasta obtener un valor de pérdidas de 0'1877 tras realizar las 500 épocas de entrenamiento. Además, se puede observar que el entrenamiento da un poco de underfitting entre las pérdidas de entrenamiento y las de validación, en este caso al ser prácticamente igual no presenta problemas significativos en la predicción.

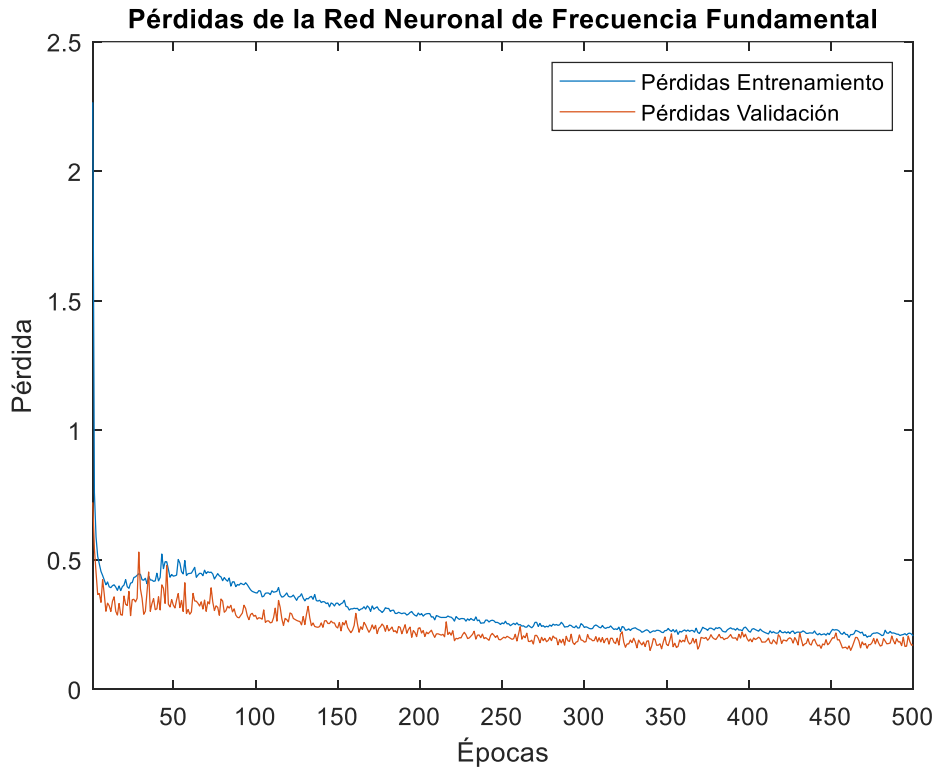


Figura 37. Resultado de las pérdidas de la Red Neuronal de Frecuencia Fundamental.

Para comprobar la precisión del valor, se ha calculado el Error Cuadrático Medio para cada uno de los bloques de evaluación en la parte sonora de la señal. En la Figura 38, se muestran los valores del Error Cuadrático Medio para cada uno de los bloques. Entre los tres bloques se obtiene una media de 0'2856 de error.

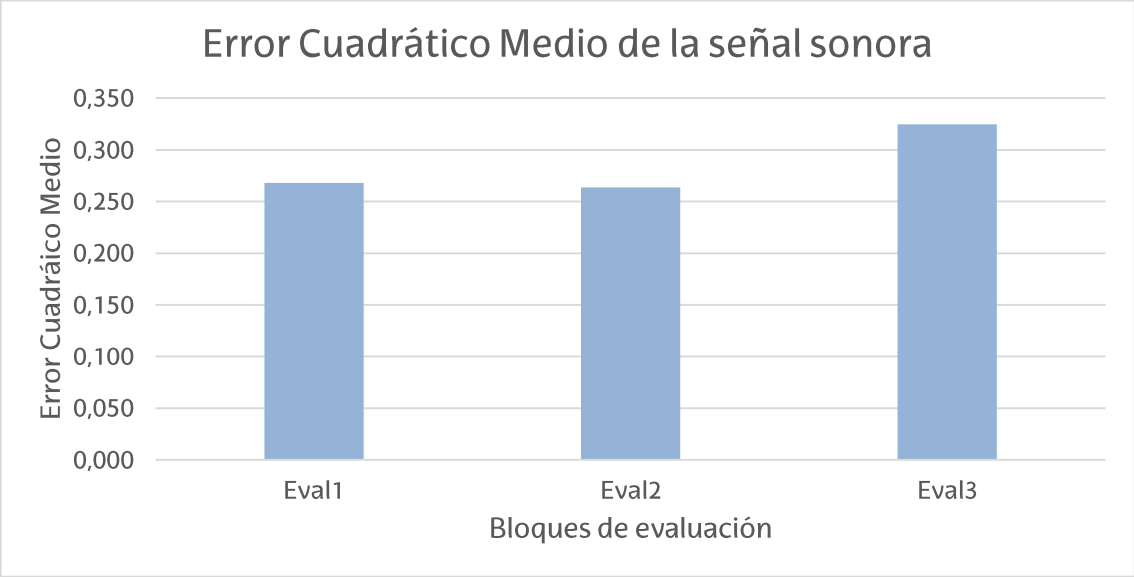


Figura 38. Error Cuadrático Medio de la Frecuencia Fundamental.

9.5. Red Neuronal de Sonidos Sordos y Sonoros

Una vez diseñada ambas Redes Neuronales, se ha calculado la matriz de confusión para comprobar cuál de las dos funciones de pérdidas ofrece una mayor precisión de acierto. En la Tabla 8 y la Tabla 9, se muestran las matrices de confusión mencionadas anteriormente. Con los datos de las matrices de confusión se han calculado la precisión, el Recall y el Fscore, que se muestran en la Tabla 10. Tras realizar el análisis, se ha concluido que es más adecuado utilizar la función de pérdidas de la Entropía Binaria Cruzada, con un Fscore de 0'990.

Matriz de Confusión Error Cuadrático Medio		
Original / **Predicción**	**Sonoro**	**Sordo**
Sonoro	TP = 10764	FP = 173
Sordo	FN = 60	TN = 9688

Tabla 8. Matriz de Confusión Error Cuadrático Medio.

Matriz de Confusión Entropía Binaria Cruzada		
Original / **Predicción**	**Sonoro**	**Sordo**
Sonoro	TP = 10739	FP = 128
Sordo	FN = 85	TN = 9733

Tabla 9. Matriz de Confusión Entropía Binaria Cruzada.

Función de pérdida	**Precisión**	**Recall**	**Fscore**
Error Cuadrático Medio	0'984	0'994	0'989
Entropía Binaria Cruzada	0'988	0'992	0'990

Tabla 10. Comparativa de calidad de las Redes Neuronales.

Tras seleccionar la función de pérdidas adecuada, es necesario comprobar el valor de las pérdidas del entrenamiento y la validación. Como se puede comprobar en la Figura 39, las pérdidas en el conjunto de datos de entrenamiento disminuyen hasta obtener un valor de pérdidas de 0'0369 tras realizar las 20 épocas de entrenamiento.

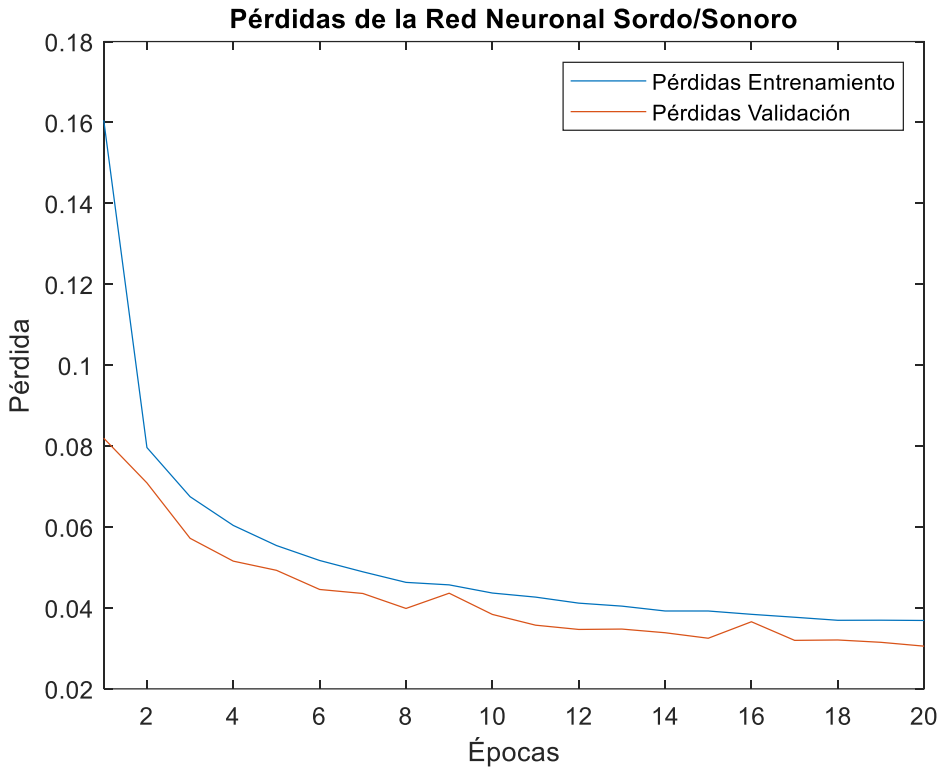


Figura 39. Resultado de las pérdidas de la Red Neuronal de los sonidos sordo o sonoros.

9.6. Red Neuronal de Espectrograma

En este apartado, se analiza la Red Neuronal diseñada para predecir el espectrograma de una señal. En la Figura 40, se muestra el espectrograma de la señal original y el espectrograma predicho de esa misma señal con un distancia euclídea entre ellos de 2.060'85. Como se puede comprobar la predicción tiene muy poca precisión frecuencial, aunque temporalmente si se obtiene una mayor precisión.

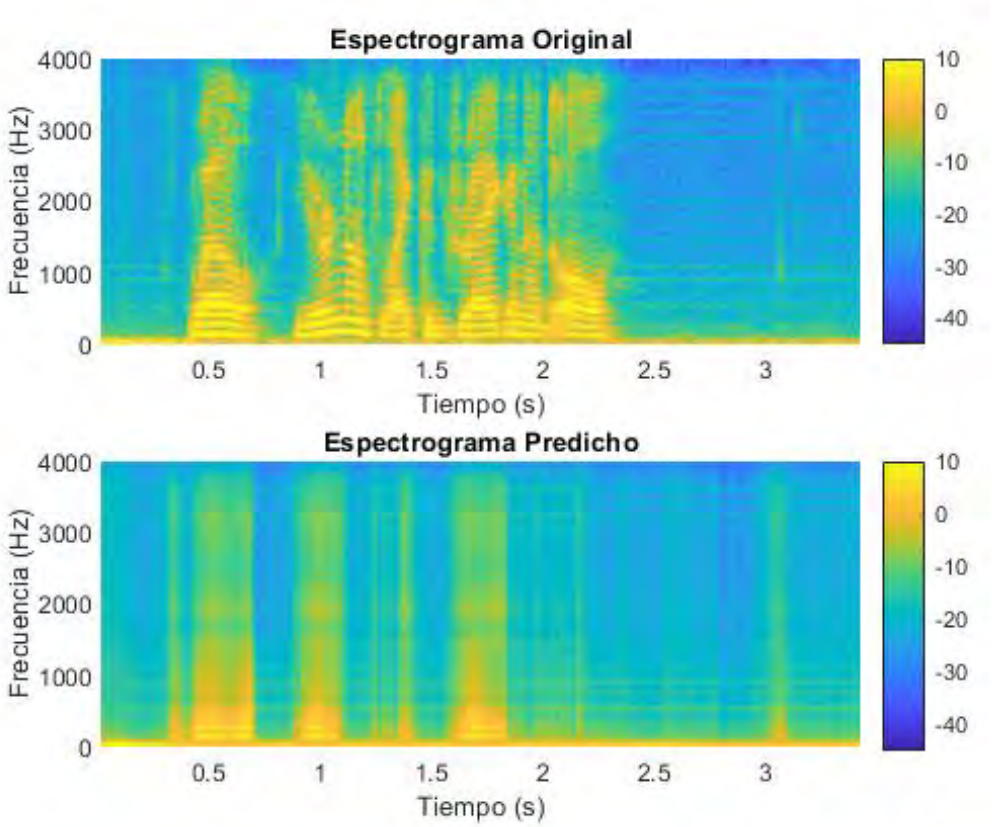


Figura 40. Original y predicción del espectrograma.

Por último, se ha calculado la distancia euclídea promedio para cada bloque de evaluación como se muestra en la Tabla 11.

Bloque de Evaluación	Distancia Euclídea Promedio	Desviación estándar
Eval 1	2.550'56	±956'66
Eval 2	2.250'30	±446'83
Eval 3	2.441'43	±382'08

Tabla 11. Cálculo de las distancias euclídeas.

10. Plan de trabajo

Fase 1 - Contextualización del proyecto: Fase inicial donde se contextualiza el proyecto. El objetivo principal de esta fase consiste en adquirir conocimientos necesarios para el desarrollo del proyecto como son el lenguaje de programación Python y la biblioteca Keras utilizada para la creación y entrenamiento de las redes neuronales. También conocimientos generales sobre Machine Learning y Deep Learning.

- Curso Python: esta tarea consiste en la introducción al lenguaje de programación Python, donde se conocen las principales librerías y funciones que ofrece este lenguaje.
- Curso Machine Learning: esta tarea consiste en la introducción a Machine Learning, donde se estudia cómo procesar datos y obtener información sobre ellos con el objetivo de realizar predicciones.
- Curso TensorFlow: esta tarea consiste en la introducción a Deep Learning, donde se estudian técnicas y algoritmos avanzados para trabajar con grandes conjuntos de datos.

Fase 2 – Estudio de la Base de Datos: Esta fase ha sido dedicada al estudio y conocimiento de la Base de Datos utilizada en el proyecto. Ha sido necesario el conocimiento de la estructura de la misma y los puntos de sincronización entre las señales de audio y las de los sensores EMG.

- Base de Datos: esta tarea consiste en el análisis de la Base de Datos, estructurada en diferentes bloques, por una parte, los bloques de entrenamiento y por otro, los bloques de evaluación de los diferentes locutores.
- Sincronización: esta tarea consiste en la obtención del script de sincronización entre las señales electromiográficas y las señales de audio.

Fase 3 – Procesado de datos: Esta fase ha sido dedicada al procesado de los datos y los cálculos necesarios para introducirlos en la red neuronal. También se han estructurado los datos sincronizados de los diferentes sensores. El principal objetivo de esta fase ha sido obtener los ficheros de entrada para entrenar las redes neuronales.

- Algoritmos de los cálculos: esta tarea consiste en la obtención de los algoritmos necesarios para obtener los valores de entrada de las Redes Neuronales.
- Ficheros CTD-15: esta tarea consiste en la obtención de los ficheros CTD-15, que están

formados por los valores de los algoritmos calculados en la tarea anterior, con el objetivo de estructurar los ficheros de manera adecuada.

- Ficheros CC y F0: esta tarea consiste en la obtención de los ficheros CC y F0 correspondientes a los valores de salida de las Redes Neuronales. Para ello se emplea el vocoder ahocoder desarrollado por el grupo de investigación Aholab.
- Ficheros Espectrograma: esta tarea consiste en la obtención de los ficheros correspondientes a los espectrogramas que se utilizarán para la salida de las Redes Neuronales.

Fase 4 – Desarrollo de las Redes Neuronales: Esta fase ha sido dedicada a la generación, entrenamiento y predicción de las diferentes Redes Neuronales, probando diferentes estructuras que se ajusten más al modelo para obtener unos resultados más precisos.

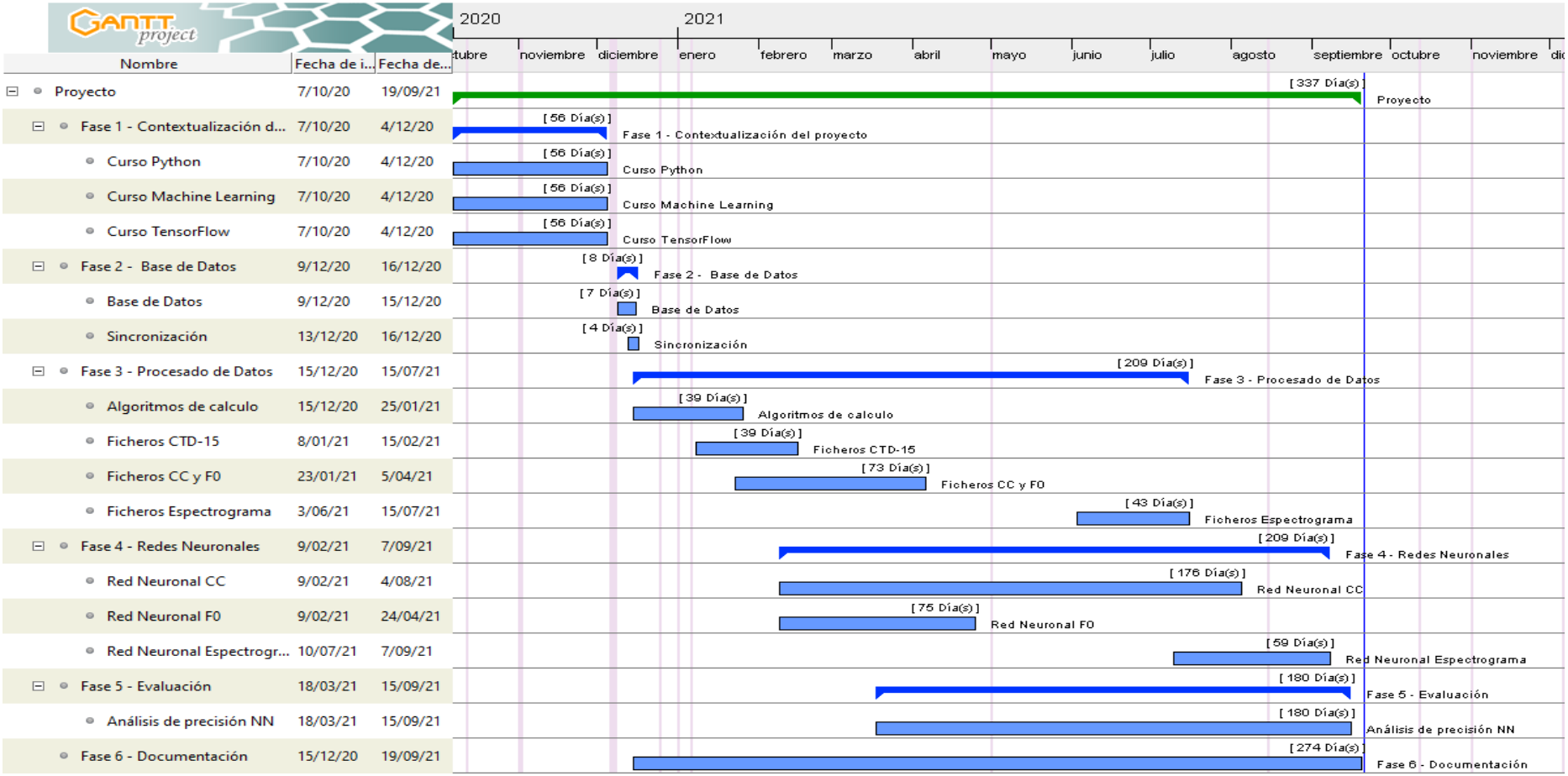
- Red Neuronal CC y espectrograma: esta tarea consiste en la obtención de una Red Neuronal capaz de predecir los Coeficientes Cepstrales o los valores del espectrograma a partir de los parámetros EMG.
- Red Neuronal F0: esta tarea consiste en la obtención de una Red Neuronal capaz de predecir la Frecuencia Fundamental y otra para la predicción de la sonoridad de cada trama.
- Red Neuronal Espectrograma: esta tarea consiste en la obtención de una Red Neuronal capaz de predecir el Espectrograma a partir de los datos de los sensores.

Fase 5 - Evaluación: Esta fase ha sido dedicada a la evaluación de los resultados obtenidos mediante las Redes Neuronales.

- Análisis de precisión: esta tarea consiste en analizar los resultados obtenidos por ambas Redes Neuronales, con el objetivo de evaluar la precisión y el error en la predicción.

Fase 6 - Documentación: Esta fase ha sido dedicada a la elaboración de la documentación relacionada con el proyecto.

11. Diagrama Gantt



12. Presupuesto

En este apartado se detallan los aspectos económicos desglosados en diferentes partidas, tratándose de horas internas, amortizaciones y costes complementarios.

HORAS INTERNAS			
Concepto	**Horas**	**Precio**	**Total**
Desarrollar del Proyecto	600	30 €/h	18.000 €
Directora del Proyecto	80	60 €/h	4.800 €
		SUBTOTAL	22.800 €

AMORTIZACIONES				
Concepto	**Compra**	**Cantidad**	**Vida Útil**	**Total**
Ordenador	1.200 €	1	6 años	200 €
Licencia Microsoft Word	135 €	1	5 años	27 €
Licencia Microsoft Excel	135 €	1	5 años	27 €
Licencia WinSCP	0 €	1		0 €
Licencia PuTTY	0 €	1		0 €
Licencia Notepad++	0 €	1		0 €
NVIDIA TITAN RTX GPU	2.700 €	1	5 años	540 €
			SUBTOTAL	794 €

COSTES COMPLEMENTARIOS			
Concepto	**Horas**	**Precio (€/kWh)**	**Total**
Coste eléctrico	680	0'15 €	102 €
		SUBTOTAL	102 €

PRESUPUESTO	
Horas internas	22.800 €
Amortizaciones	794 €
Costes Complementarios	102 €
TOTAL	23.696 €

13. Conclusiones

Tras realizar este trabajo, se ha obtenido un sistema de referencia con margen de mejora a lo largo del tiempo. Al trabajar con diferentes tipos de datos y de Redes Neuronales, se ha podido realizar un pequeño estudio a futuro para el proyecto ReSSint, para el que se han concluido importantes aspectos clave para el desarrollo del mismo. Entre los aspectos destacables se encuentran los siguientes:

- La frecuencia de muestreo de las señales del EMG y la señal de audio es conveniente que sean múltiplos para facilitar la sincronización de las ventanas y no perder muestras.

- Uno de los aspectos más destacables es la utilización de sensores independientes, debido a que ofrecen señales más variantes y puede ser muy útil para el aprendizaje de las Redes Neuronales. Además, aparece de la posibilidad de seleccionar en que zona colocar el sensor para obtener señales más relevantes y variantes entre los diferentes movimientos vocales.

- La Red Neuronal base utilizada se ajusta mejor a predecir los Coeficientes Cepstrales que el Espectrograma, debido a que debe predecir más valores y se trata de una predicción similar al de una imagen

- El filtrado de la señal tiene un papel fundamental en el resultado final, ya que se obtienen unos mejores resultados con el filtro más preciso como es de esperar. Además, determina la selección del modelo de la Red Neuronal.

- Otro de los aspectos fundamentales en un buen aprendizaje de la Red Neuronal se trata de la cantidad de muestras de entrada debido a la variabilidad de la voz, el movimiento de la cara y por lo tanto la señal captada por los sensores.

14. Referencias

Aholab. (2020). *ReSSint*. Obtenido de ReSSint: https://aholab.ehu.eus/ressint/

Beiming, C., Nordine, S., Ted, M., Omer T, I., & Jun, W. (2019). Permanent Magnetic Articulograph (PMA) vs Electromagnetic Articulograph (EMA) in Articulation-to-Speech Synthesis for Silent Speech Interface. *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 17-23.

Diener, L. (2021). *The Impact of Audible Feedback on EMG-to-Speech Conversion.*

Diener, L., Janke, M., & Schultz, T. (2015). Direct Conversion from Facial Myoelectric Signals. *IEEE*.

Diener, L., Vishkasougheh, M. R., & Schultz, T. (2020). CSL-EMG Array: An Open Access Corpus for EMG-to-Speech Conversion. *Interspeech*.

Erro, D., Sainz, I., Navas, E., & Hernaez, I. (2014). Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 184-194.

Eurostat. (2 de 8 de 2021). *Population by type of basic activity difficulty, sex and age*. Obtenido de https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=hlth_dp040&lan%g=en

Fitzpatrick, M. (2002). Lip-reading cellphone silences loudmouths. *New Scientist*.

Gaddy, D., & Klein, D. (2020). Digital Voicing of Silent Speech. *Association for Computational Linguistics*, 5521-5530.

Gonzalez, J. A., Gomez, A., Martín, J. M., Pérez, J. L., & Gomez, A. M. (2020). Silent Speech Interfaces for Speech. *IEEE*, 177995-178021.

Hasegawa, T., & Ohtani, K. (1992). Oral image to voice converter-image input microphone. *IEEE*, 617–620.

Instituto Nacional de Estadística. (2008). *Encuesta de Discapacidad, Autonomía Personal y Situaciones de Dependencia 2008*. Obtenido de

http://www.ine.es/jaxi/Datos.htm?path=/t15/p418/a2008/hogares/p01/modulo1/l0/&file=01002.px

Janke, M., Wand, M., Nakamura, K., & Schultz, T. (2012). Futher investigations on EMG-To-Speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 365-368.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 436.

M, W., M, J., & T, S. (2014). The EMG-UKA corpus forelectromyographic speech processing. *Interspeech.*

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 746-748.

Meltzner, G., Heaton, J., Deng, Y., De Luca, G., Roy, S., & Kline, J. (2017). Silent speech recognition as an alternative communication device for persons with laryngectomy. *IEEE/ACMTrans. Audio, Speech, Language Process.*, 2386-2398.

Ng, L., Burnett, G., Holzrichter, J., & Gable, T. (2000). IEEE International Conference On Acoustics, Speech, And Signal Processing., (págs. 229-232). Estambul.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatoryspeech from speakers with dysarthria. *Lang Resources & Evaluation*, 523–541.

Schmidhuber, J. (2015). Deep learning in neural networks. *Neural Networks*, 85-117.

VentureBeat. (17 de 05 de 2017). Obtenido de https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/

Westbury, J. R., Turner, G., & Dembowski, J. (1994). *X-ray microbeamspeech production database user's handbook*. Wisconsin.

Wrench, A. (03 de 07 de 2020). *The MOCHA-TIMIT articulatory database*. Obtenido de https://www.cstr.ed.ac.uk/research/projects/artic/mocha.html