
UNIVERSIDAD
DE GRANADA

TRABAJO FIN DE GRADO

INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

# Síntesis de voz a partir de bioseñales de habla usando técnicas de Machine Learning

**Autor**
Javier Lobato Martín

**Directores**
José Andrés González López (tutor 1)
José Luis Pérez Córdoba (tutor 2)

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Julio de 2023

# Síntesis de voz a partir de bioseñales de habla usando técnicas de Machine Learning

Javier Lobato Martín

**Palabras clave**: Síntesis de voz, Bioseñales, Unit Selection, PMA, WORLD,

## Resumen

Numerosas enfermedades o traumas conducen a la pérdida o la grave afectación del habla, como pueden ser el ictus cerebral, el Esclerosis Lateral Amiotrófica (ELA) o la laringectomía. Asimismo, muchas de estas enfermedades no tienen cura, por lo que la pérdida de la capacidad del habla es irreversible y, en ciertos casos, progresiva. Aunque existen en el mercado dispositivos que permiten a estas personas compensar los déficits de comunicación que padecen tras estas enfermedades, es común que estos dispositivos sean lentos y difíciles de usar, lo que afecta en gran medida la calidad de vida de estas personas.

Este proyecto aborda el problema de devolverle la comunicación oral a estas personas desde una perspectiva tecnológica. Así, en este trabajo fin de grado se desarrolla una serie de algoritmos para la síntesis de voz a partir de bioseñales de habla capturadas a partir del movimiento de los labios y lengua de participantes, haciendo uso de una técnica de captura conocida como *Articulografía por Imanes Permanentes* (PMA).

Para llevar a cabo el objetivo del trabajo se propone el uso de un algoritmo de síntesis de voz denominado Unit-Selection. Este algoritmo realiza la síntesis de voz usando la información de una base de datos en la que se tienen bioseñales PMA y voz (parámetros obtenidos de la misma en este caso), obtenidas de pacientes siguiendo un esquema de grabación simultánea. Los parámetros de la voz con los que se trabaja son Mel Frequency Cepstral Coefficients (MFCC)'s. Ambas señales son divididas en pequeñas secciones denominadas unidades. La predicción de la voz a partir de las bioseñales se lleva a cabo implementando distintas métricas de evaluación de distancia entre unidades que se incorporan y dan como resultado una secuencia de parámetros, a partir de la cual se sintetiza la voz. El proceso de síntesis a partir de los parámetros se lleva a cabo por medio del VoCoder WORLD.

Los resultados del trabajo muestran que es posible sintetizar voz inteligible a partir de las bioseñales PMA haciendo uso del algoritmo de Unit Selection. Los resultados obtenidos en términos de distorsión cepstral (MCD) se encuentran entre 10.8 dB y 12.4 dB para todos los datasets, mientras que para inteligibilidad (STOI) se encuentran entre 0.4 y 0.57. Escuchas subjetivas determinan que el algoritmo creado con este propósito tiene una inteligibilidad superior al método base implementado de regresión lineal.

El algoritmo diseñado es capaz de sintetizar voz inteligible para bioseñales PMA, tanto para dígitos como para oraciones completas, obteniendo mejoras de inteligibilidad con respecto a un método de predicción lineal.

# Speech Synthesis From Biosignals

Javier Lobato Martín

**Keywords**: Speech Synthesis, Biosignals, Unit Selection, PMA, WORLD

## Abstract

Numerous diseases or traumas have as a consequence the partial or total loss of speech, such as cerebral ictus, amyotrophic lateral sclerosis (ALS) or laryngectomy. Some of this diseases are incurable, making the loss of speech irreversible and, in some cases, progressive. There are commercial products available that allow affected people to compensate these communication deficits, although they are usually slow and difficult to use, which heavily affects the quality of life of these people.

This projects tackles the problem of restoring speech to these affected people through a technological approach. In this project, algorithms for speech synthesis will be developed to synthesise words and sentences using movement of the tongue and lips of participants, captured using a technique called Permanent Magnet Articulography (PMA).

To complete the objective of the project, a Unit-Selection speech synthesis algorithm is proposed. This algorithm synthesizes voice using information from a database that contains PMA biosignals and voice parameters. These voice parameters used are called Mel Frequency Cepstral Coefficients MFCC's. Both signals are divided into smaller pieces called units. The prediction of voice from these units is performed using different distance evaluation metrics between units. These metrics are combined and give as a result a sequence of voice parameters, which is then synthesized into voiced using a VoCoder (WORLD, in this case).

Project results show that it is possible to synthesize intelligible voice from PMA biosignals using the Unit-Selection approach created. The results obtained in terms of Mel Cepstral Distortion MCD lay between 10.8 db and 12.4 dB for all datasets. When it comes to intelligibility (measured with STOI), results lay between 0.4 and 0.57. Subjective hearings determine that the algorithm has a higher intelligibility than the Linear Regression base method.

The designed algorithm is capable of synthesizing intelligible voice using PMA biosignals for single digits and complete sentences, obtaining improvements compared with the base Linear Regression method.

Yo, **Javier Lobato Martín**, alumno de la titulación TITULACIÓN de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada**, con DNI 44739277E, 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: Javier Lobato Martín

Granada a 12 de julio de 2023 .

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

D. **José Luis Pérez Córdoba**, Profesor del Departamento Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada.

**Informan:**

Que el presente trabajo, titulado ***Síntesis de voz a partir de bioseñales de habla usando técnicas de Machine Learning***, ha sido realizado bajo su supervisión por **Javier Lobato Martín**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 12 de Julio de 2023 .

**Los directores:**

**José Andrés González López** **José Luiz Pérez Córdoba**

# Agradecimientos

A mis padres, que me han apoyado incondicionalmente en cada etapa de mi vida y sin los cuales no podría cumplir los objetivos que me propongo.

A mis amigos, que me acompañan en esta aventura y son una parte imprescindible de todo lo que hago.

A mi tutor José Andrés, por su infinita paciencia, profesionalidad y buen hacer.

# Siglas

**DCT** Discrete Cosine Transform. 33

**DFT** Discrete Fourier Transform. 32

**DNN** Deep Neural Network. 38, 39, 72

**DSS** Direct Speech Synthesis. 33–36, 41

**ECoG** Electrocorticografía. 27

**EEG** ElectroEncefaloGrafía. 26–29, 67

**ELA** Esclerosis Lateral Amiotrófica. 5, 17

**EMA** Electro Magnetic Articulography. 24, 25, 35, 39

**EMG** Electromiografía. 26

**EPG** Electro Palatografía. 24

**ERP** Event Related Potential. 27

**MCD** Mel Cepstral Distortion. 5, 7, 39, 57, 71

**MFCC** Mel Frequency Cepstral Coefficients. 5, 7, 32, 44

**MSE** Mean Square Error. 38

**OMS** Organización Mundial de la Salud. 17

**PMA** Permanent Magnet Articulography. 5, 7, 18, 25, 26, 35, 37, 41–43, 71

**SAAC** Sistemas Alternativos y Aumentativos de Comunicación. 17

**sEMG** Surface Electro-Myogram. 27, 35

**SSI** Silent Speech Interface. 18, 19, 21, 24, 27, 34

**STOI** Short Term Objective Intelligibility. 7, 40, 57

# Capítulo 1

# Introducción

El lenguaje es una capacidad inherente y esencial en los humanos, que nos permite una comunicación precisa y eficiente. Por medio de ella establecemos vínculos y relaciones con otras personas, podemos comunicar nuestros sentimientos, poner en común experiencias y transmitir conocimientos. Es uno de los fenómenos que nos ha permitido prosperar como especie.

Por desgracia, existen numerosas condiciones que pueden tener como consecuencia la pérdida de la voz. Estas afecciones pueden tener diversos orígenes, como pueden ser lesiones traumáticas, enfermedades neurodegenerativas como la Esclerosis Lateral Amiotrófica (ELA), el ictus cerebral o laringectomía.

Este problema no es trivial para la sociedad, numerosos estudios evidencian que afecta a un gran número de personas. Un estudio llevado a cabo por la agencia europea Eurostat [6] concluyó que un 0.4 % de la población europea tiene un impedimento en el habla. Un estudio llevado a cabo en 2011 por la Organización Mundial de la Salud (OMS) 2011 en 70 paises [16], concluyó que el 3.6 % de la población sufría una dificultad grave o extrema en la participación en la comunidad. Al ser el lenguaje de vital importancia, las consecuencias que esto tiene para las personas que lo sufren son significativas: La comunicación diaria se ve gravemente dificultada, así como la asistencia médica (debido a la inefectividad del intercambio de información). Esta imposibilidad puede provocar un sentimiento de aislamiento social, e incluso puede derivar, según [3], en depresión clínica. A nivel económico y de participación en el mercado laboral las consecuencias también son notables. Según [6], un 78 % de la población europea con una discapacidad severa se encuentran fuera del mercado laboral, en contraposición al 27 % para la población que no sufre esa condición.

Desafortunadamente, a día de hoy no existen soluciones para revertir las afecciones que causan la pérdida del lenguaje, de ahí la enorme importancia de la búsqueda de soluciones que puedan restaurar esta capacidad. Hay dispositivos conocidos como Sistemas Alternativos y Aumentativos de

Comunicación (SAAC) que pueden ayudar a reestablecer la capacidad de comunicación. Algunos ejemplos pueden ser desde la simple escritura hasta sistemas de conversión de texto a habla. Estos sistemas tienen sus limitaciones y no es posible su uso en todas las casuísticas.

## 1.1. Interfaces SSI

En los últimos años ha habido un creciente interés en las interfaces orales silenciosas Silent Speech Interface (SSI), que permiten la comunicación oral sin la necesidad de vocalización de palabras. Este tipo de interfaces permiten una comunicación oral por medio de la interpretación de bioseñales de distinta naturaleza [12]. Las bioseñales que se utilizan como fuente de información para la síntesis de voz provienen de distintos sectores del cuerpo humano que toman parte en el proceso de producción de voz. Ejemplos de obtención de bioseñales incluyen la lectura del movimiento de los labios labios, registro del movimiento del tracto vocal u obtención de la actividad neuronal del cerebro relacionada con el habla.

Las SSI's cuentan con numerosas aplicaciones potenciales, como pacientes que han sufrido una laringectomía o pacientes de avanzada edad en los que el habla requiere un esfuerzo significativo pero el movimiento asociado posibilita el uso de estas interfaces.

Existe una gran diversidad de aproximaciones que hacen uso de la información aportada por las bioseñales para sintetizar voz o texto, en función de la naturaleza de las bioseñales que se utilizan, la filosofía que se siga para la implementación del algoritmo, etc.

Las interfaces SSI tienen el potencial de producir voz con un resultado natural y una dinámica de uso intuitiva y sencilla, por lo que son un área de estudio en auge y que puede mejorar la calidad de vida de millones de personas.

## 1.2. Objetivos

Los objetivos de este trabajo se dividen en principales y secundarios. El objetivo principal del proyecto consiste en conseguir la **síntesis de voz inteligible a partir de bioseñales obtenidas con la técnica de *Articulografía por Imanes Permanentes* PMA**. La síntesis de voz se llevará a cabo mediante la implementación de un algoritmo de Unit Selection. Los detalles del algoritmo se detallarán en secciones posteriores. El objetivo principal incluye la síntesis de voz inteligible para una base de datos que contiene dígitos individuales.

Los objetivos secundarios incluyen:

- Síntesis de voz para una base de datos de oraciones completas fonéticamente balanceadas, de mayor complejidad.

- Implementación de un método base de síntesis de voz por Regresión Lineal para establecer comparaciones.

## 1.3. Estructura de la memoria

La memoria de este trabajo se estructurará en capítulos, siguiendo la siguiente distribución:

- **Estado del arte** 2: Descripción más en profundidad de las interfaces orales silenciosas SSI y revisión de los trabajos de la rama más destacados de los últimos años.
- **Método propuesto** 3: Especificación del caso de uso concreto para el trabajo, así como presentación del algoritmo diseñado para la síntesis de voz a partir de las bioseñales disponibles.
- **Resultados obtenidos** 4: Presentación de los resultados experimentales obtenidos para la evaluación del método propuesto para síntesis de voz.
- **Conclusiones y líneas futuras** 5: Conclusiones finales a las que se llega una vez se han analizado los resultados obtenidos. Presentación de las líneas de investigación y de mejoras futuras en base al trabajo realizado y los resultados.
- **Anexos**: Temporización del proyecto A.1 y presupuesto A.2 del mismo.

# Capítulo 2

# Estado del arte

En este capítulo, se realiza una presentación más exhaustiva de las interfaces orales silenciosas (SSIs), incluyendo los bloques que componen un sistema de estas características, así como una revisión de los trabajos más destacados en los últimos años. De esta manera, la sección 2.1 se encuentra dedicada a la especificación de un sistema de síntesis de voz por medio de bioseñales SSI. El resto de las secciones se dedican a explicar cada uno de los componentes de dicho sistema, incluyendo los procesos físicos involucrados con el habla, las técnicas de obtención de bioseñales, los parámetros que se extraen y la decodificación del habla a partir de bioseñales.

## 2.1. Esquema de un sistema SSI

Esta sección se va a centrar en la espeficicación genérica de los distintos componenentes que constituyen un sistema de síntesis de voz. La siguiente figura ilustra, por medio de un digrama de bloques, las principales etapas de procesamiento y bloques funcionales de un sistema de comunicación SSI:

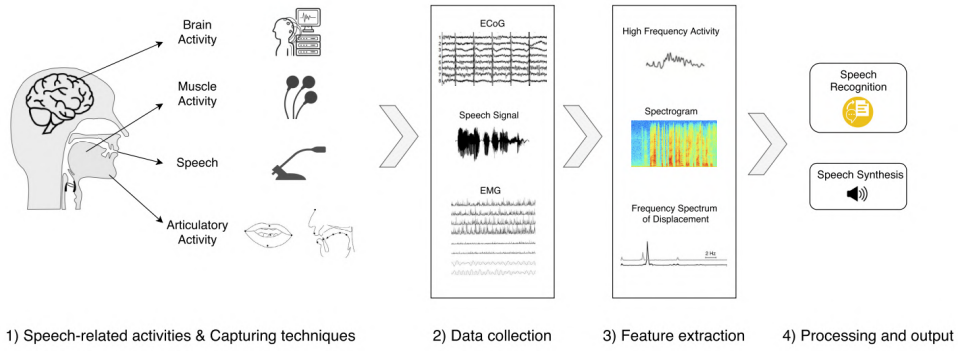


Figura 2.1: Sistema de comunicación basado en SSI. Fuente: [41]

Como se observa en la figura 2.1, el proceso de síntesis de voz sigue un flujo que se puede dividir en cuatro secciones diferenciadas:

1. El primer paso es encontrar qué actividad relacionada con el habla se va a registrar.
2. Se obtienen las bioseñales de interés haciendo uso de un método en concreto.
3. Por medio de procesamiento de señales se obtienen distintos atributos de las bioseñales
4. La voz se decodifica haciendo uso de los atributos obtenidos de las bioseñales.

Para estructurar esta nueva sección nos vamos a basar en el diagrama de bloques 2.1, ya que sigue la lógica organizativa de los sistemas en los que se centra este proyecto. Se hará una revisión de las técnicas y aspectos más relevantes por su uso o por su abundancia bibliográfica.

## 2.2. Procesos fisiológicos involucrados en el habla

En esta sección vamos a revisar los distintos procesos fisiológicos que juegan un papel en la producción del lenguaje. Son la fuente de las bioseñales que se utilizan en el proceso de síntesis de voz y comprenden el primer paso para abordar el problema.

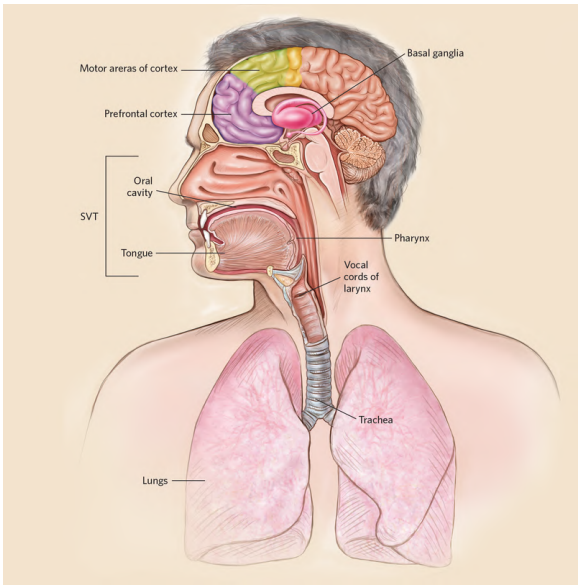


Figura 2.2: Fisionomía de los distintos componentes implicados en la producción del lenguaje. Fuente: *thescientist.com*

Para ilustrar estos procesos, la figura 2.2 indica las distintas partes del cuerpo humano que juegan un papel en el lenguaje. La génesis de la voz humana es la actividad cerebral, fruto de la cual aparece la actividad muscular que la articula.

### 2.2.1. Actividad Cerebral

Las señales que se producen en el cerebro son el origen de toda comunicación verbal. La enorme ventaja de utilizar este tipo de actividad para la síntesis de voz radica en que puede servir como solución para un amplio abanico de afecciones, desde diversos trastornos que no suponen la pérdida total de la voz (disastria, apraxia, laringectomía) hasta los casos más graves como la afasia (pérdida total de la capacidad del habla). Hay numerosos estudios que relacionan los procesos de producción del habla con una zona concreta del cerebro: el giro temporal superior [40]. Obteniendo señales provinientes de esta zona del cerebro se puede llegar a obtener el habla, aunque existen inconvenientes notables. Los procesos del cerebro que originan la voz son de una complejidad muy alta y a día de hoy no son comprendidos en su totalidad. Adicionalmente, estas señales se tienen que obtener con una resolución muy grande para obtener resultados aceptables, lo que comprende otro reto en sí mismo. Todos estos problemas dificultan notablemente la síntesis de voz a partir de señales cerebrales.

### 2.2.2. Actividad muscular

Para hablar, es necesario la existencia de movimientos en los músculos de la cara, así como en la laringe y la lengua. Es posible sintetizar la voz muestreando y procesando los fenómenos fruto de esta actividad muscular. En el proceso de producción de voz, una vez han tenido lugar los procesos cerebrales de conceptualización del mensaje y planificación de la actividad motora, dicha planificación se traduce a impulsos eléctricos que se transmiten por las neuronas motoras del sistema nervioso periférico, que inervan los músculos asociados son el proceso de producción de voz. Estos impulsos eléctricos coordinan la contracción y relajación de los músculos, que resultan en movimientos concretos de los mismos. La coordinación de estos movimientos con el flujo de aire por el tracto vocal origina la voz. De esta manera, existen dos fuentes de información diferenciadas de las que se pueden obtener las bioseñales: Muestreo de los impulsos eléctricos que originan los movimientos musculares o muestreo del propio movimiento muscular. El muestreo de la actividad muscular constituye un proceso de gran interés para la síntesis de voz ya que son una familia de señales mucho más acotadas que las cerebrales y cuya conversión a voz es factible y explotada desde hace años. Como desventaja, no todas las casuísticas de pacientes son compatibles con este método de obtención, ya que en muchos casos los articuladores del habla

no generan movimiento o señales eléctricas a partir de las cuales se pueda sintetizar voz.

## 2.3. Obtención de bioseñales

En el contexto de la síntesis de voz a partir de bioseñales, se definen las bioseñales como señales fisiológicas que tienen relación con diversos aspectos del proceso de generación de voz humana. Estas señales pueden ser o no de naturaleza eléctrica. La obtención de las bioseñales se lleva a cabo haciendo uso de sensores especializados para cada tipo de bioseñal. En este apartado se van a revisar las técnicas de obtención de bioseñales más relevantes para el uso en SSIs

### 2.3.1. Movimiento Articular

La producción del habla requiere del movimiento de los articuladores del habla: labios, lengua, paladar y laringe. Este tipo de bioseñales se obtienen colocando sensores magnéticos o visuales en distintas zonas del tracto vocal. Este tipo de métodos no están diseñados para capturar la actividad de la glotis (cuyo funcionamiento influye en el tono y la intesidad de la voz), por lo que su campo de aplicación suele ser el de personas con desórdenes en la voz como pacientes que han sufrido una laringectomía. Cuatro métodos destacan:

#### EPG

Electro Palatografía (EPG). Un array de electrodos se coloca en el paladar para registrar la secuencia de contactos de la lengua con el paladar. Se registra un patrón de contactos que aporta información sobre la pronunciación de fonemas. Su utilidad para nuestro caso es limitada, puesto que requiere movimiento de la lengua y principalmente se utiliza para investigación acerca de la fonética. [4]

#### Técnicas de Imagen

Se registran imágenes y vídeo de los articuladores vocales. Es una solución sencilla y práctica, haciendo uso de distintos sensores radar, ultrasonidos, ópticos, etc. Por lo general, necesita ser complementado con otro tipo de información para mostrar una imagen completa del proceso del habla. [13]

#### EMA

En la técnica de Electro Magnetic Articulography (EMA), sensores de campo magnético se colocan en los articuladores, a su vez conectados por

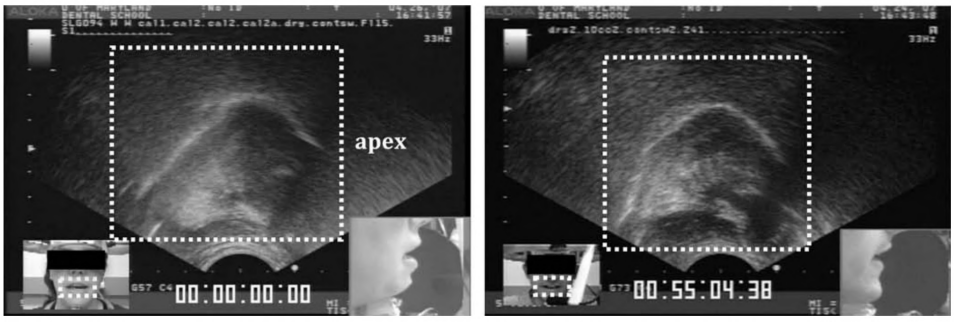


Figura 2.3: Imagen del tracto vocal y labios para dos pacientes. Fuente: [13]

cableado a monitores externos. Se disponen bobinas transmisoras cerca de la cabeza del paciente que generan un campo magnético alternante. El registro continuo de esta señal electromagnética permite rastrear la posición espacial de los receptores y obtener parámetros de la voz. Se obtiene una resolución temporal muy alta, pero no es posible registrar el movimiento de la glotis, dificultando el proceso. Adicionalmente, la necesidad de aparatos externos hace menos práctica esta solución. [19]

### PMA

Permanent Magnet Articulography (PMA) una solución similar a EMA, pero cuenta con ventajas claras. En esta técnica se disponen numerosos imanes de pequeño tamaño en zonas específicas de los articuladores vocales. La suma de los campos magnéticos generados por el movimiento de estos imanes es registrada por sensores magnéticos fuera de la boca. Esta suma registra la posición y temporización del movimiento de múltiples articuladores, como pueden ser labios, mandíbula o lengua. La principal ventaja radica en que no se necesitan conexiones cableadas a un monitor y que los sensores tienen una colocación sencilla. Esta técnica es más cómoda para el usuario y facilita la portabilidad de la misma. Por contra, la señal que se registra es una suma del movimiento del campo magnético generado por varios imanes, por lo que la relación de los datos con la posición exacta de los imanes es menos explícita que en EMA. [41]. En la siguiente imagen 2.4 se puede observar una configuración típica para la obtención de bioseñales por medio de PMA.

En la parte izquierda de la imagen2.4 (a) se puede comprobar un ejemplo de colocación de los distintos imanes. En las imagenes se observa que se encuentran dispuestos en labios y lengua. La parte derecha de la imagen2.4 (b) nos muestra un ejemplo de distribución de los sensores. El sistema completo de captura de señales es compacto y no requiere de conexiones cableadas con máquinas externas.

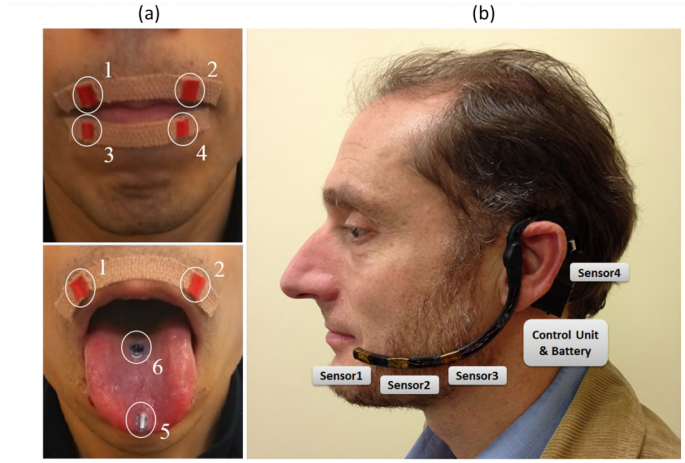


Figura 2.4: Disposición de sensores para PMA. Fuente: [31]

### 2.3.2. EMG

Electromiografía (EMG). Esta señal registra los potenciales eléctricos que se generan en los músculos faciales en la fase de contracción. Se puede obtener por métodos invasivos o no invasivos. Los más convenientes son los segundos, aunque tienen la inconveniencia de que a la complejidad de la señal (dependiente del sistema nervioso y de las propiedades anatómicas y fisiológicas del músculo) se le suma el ruido que supone el viaje de la señal eléctrica a través de la piel o la posible interferencia de otros músculos. Una ventaja fundamental de los sistemas que utilizan esta bioseñal radica en que la señal eléctrica aparece  60ms antes de la contracción muscular, este adelanto permite reducir notablemente la latencia y facilita la implementación de sistemas en tiempo real. Una de las desventajas de estos sistemas es que los resultados varían significativamente con cada sesión de entrenamiento, al variar la posición exacta de los sensores sesión a sesión. Un ejemplo de disposición de sensores se puede comprobar en la siguiente imagen

### 2.3.3. Bioseñales Cerebrales

Hay una variedad de sensores diseñados para registrar actividad cerebral. Para nuestro cometido, la principal diferenciación para métodos electrodinámicos se hace entre métodos invasivos y no invasivos. También existen métodos hemodinámicos, aunque están fuera del foco de este estudio.

#### EEG

ElectroEncefaloGrafía (EEG). Es uno de los métodos más utilizados para la obtención de la actividad cerebral, al ser una práctica no invasiva y cuyo uso tiene un largo recorrido. Su funcionamiento consiste en la colocación


Figura 2.5: Disposición de sensores para sEMG. Fuente: [23]

de electrodos en el cuero cabelludo para el registro de señales eléctricas. Con ello, lo que se obtiene es una señal con buena resolución temporal pero con una resolución espacial pobre, al obtener una versión muy suavizada del patrón de disparo de las neuronas de la zona, añadido a ello el patrón de filtro paso-baja que provoca la piel y el cráneo por los que la señal pasa. Es por ello que el uso principal de esta señal sea para obtener patrones amplios del disparo de las neuronas. Un ejemplo muy usado es el potencial P300. El P300 es un tipo concreto de bioseñal, obtenida por medio de métodos EEG y que se engloba dentro que lo que conocemos como Event Related Potential (ERP). Esto es una respuesta eléctrica cerebral que se obtiene como consecuencia de un estímulo, sea cognitivo, sensorial o motor. El P300 se obtiene por medio del conocido como 'paradigma de la rareza', en el que se mezclan estímulos (visuales en el caso que nos atañe) que se repiten continuamente con otros menos usuales, que deben ser señalados por el sujeto y que es lo que hace disparar el potencial P300. [42]. En la figura 2.6 se observa un esquema de colocación de sensores para obtención de bioseñales EEG.

### ECOG

Electrocorticografía (ECoG). Es un método invasivo, por medio del cual se coloca un array de electrodos directamente sobre el córtex cerebral. De esta manera se obtiene una muy buena resolución tanto temporal como espacial, además de una alta portabilidad, por lo que se anticipa como una solución plausible para la implantación de prótesis para restaurar el habla. Numerosos estudios han analizado la capacidad de crear modelos SSI con este método, obteniendo distintos niveles de éxito. [36] , [37]

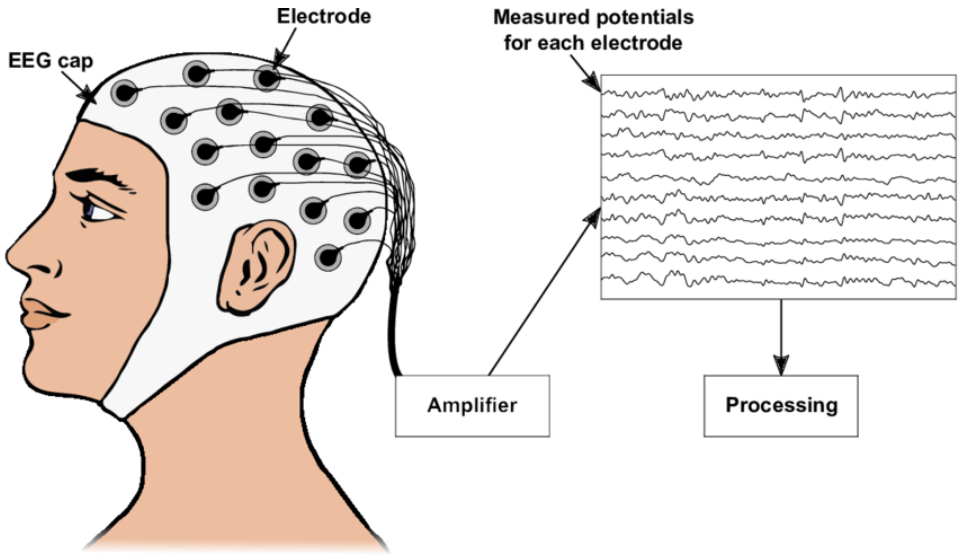


Figura 2.6: Disposición de sensores para obtención de bioseñales por medio de EEG. Fuente: [39]

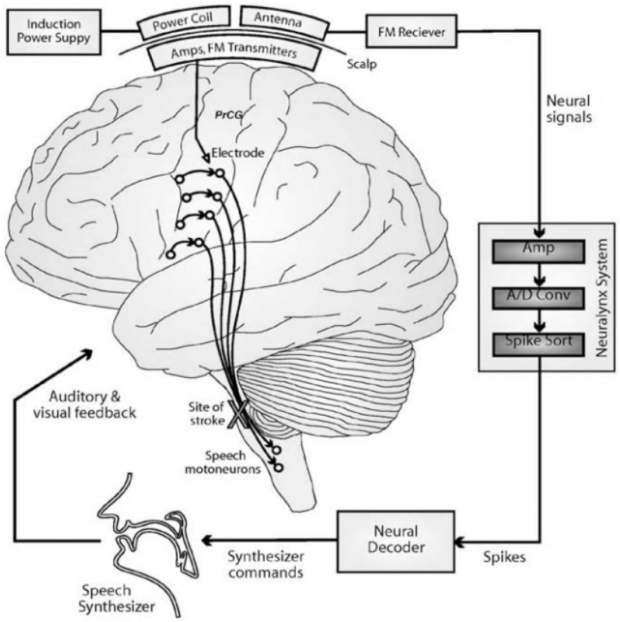


Figura 2.7: Esquema de funcionamiento para un sistema de síntesis de audio neuronal. Fuente: [11]

## 2.4. Extracción de parámetros

Este apartado se va a centrar en los parámetros que más se utilizan en la bibliografía, con especial interés en los que se van a usar en nuestro estudio.

### 2.4.1. Parámetros de EEG

#### Ondas Alpha

Las señales eléctricas que tienen lugar en el cerebro son oscilaciones que pueden adoptar distintas frecuencias. En el caso de las *Ondas Alpha*, estas se encuentran en el rango de frecuencia entre 8-12 Hz y tienen su origen en la combinación síncrona y coherente de la actividad eléctrica de las células del tálamo. Son una de las ondas cerebrales que se pueden detectar con técnicas de EEG. Se utilizan en síntesis de voz porque se ha hipotetizado que estas toman un papel en la comunicación. [17]

#### Ondas Beta

Similar a las Ondas Alpha, estas ondas se encuentran en el rango de frecuencia entre 12.5-30 Hz y se obtienen por métodos como EEG. Su uso en esta casuística se debe a que se asocian estas ondas con contracciones musculares isotónicas, así como con el proceso de pensamiento activo y concentración. [9] [8]

### 2.4.2. Parámetros de la voz

#### Espectrograma

Usado extensivamente para el análisis de señales, en especial para audio. Consiste en una representación visual de la variación del espectro de una señal con respecto al tiempo. Consta de tres dimensiones: tiempo, frecuencia y densidad espectral. Aplicado a la voz, permite identificar las palabras individuales que se pronuncian, así como los componentes espectrales de la voz y su evolución temporal.

El espectrograma es una representación interesante en tanto en cuanto nos puede permitir evaluar de una forma clara la similitud entre señales originales y sintetizadas, pudiendo comprobar la cercanía para las componentes espectrales de cada instante de tiempo con un vistazo.

#### Vocoder WORLD

Un VoCoder es un dispositivo software utilizado para sintetizar la voz humana. El término se origina de la combinación de las palabras *'voice'* y *'coder'*. Un VoCoder analiza las componentes espectrales de la voz y las aplica a una portadora. WORLD fue creado con el objetivo de disponer de

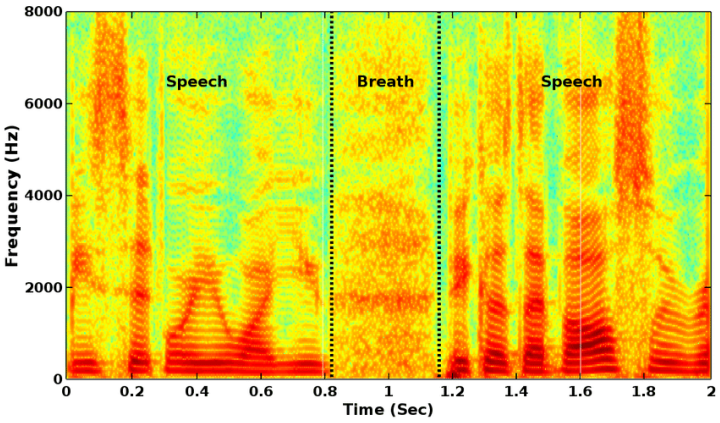


Figura 2.8: Espectrograma típico para el habla. Fuente: [30]

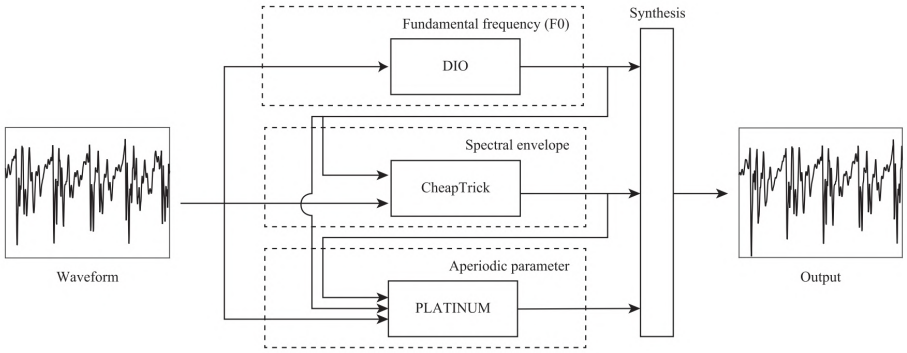


Figura 2.9: Esquema del Vocoder WORLD para análisis y síntesis. Fuente: [28]

un VoCoder que sintetizase voz de alta calidad y que pudiese ser utilizado en tiempo real. En [28], se afirma que en cuanto a velocidad de procesamiento, es 10 veces más veloz que los métodos convencionales para síntesis de voz. Contiene 3 algoritmos de análisis y un algoritmo de síntesis.

En la figura 2.9 se observa que los 3 algoritmos de análisis obtienen la frecuencia fundamental *F0*, la envolvente espectral y el parámetro de aperiodicidad. El algoritmo de síntesis utiliza estos 3 parámetros para obtener la voz sintetizada. Como se ha comentado en apartados anteriores, el objetivo de este trabajo será el de obtener la envolvente espectral de la señal (representada por medio de los coeficientes espectrales MFCC), por lo que el análisis del VoCoder se centrará únicamente en la obtención de esta envolvente, así como en la síntesis de voz.

La voz humana está compuesta por una superposición de ondas de frecuencia única. La menor de todas ellas es lo que se conoce como la frecuencia fundamental (F0) y nos permite caracterizar la voz.

El parámetro de aperiodicidad se utiliza para indicar la existencia de componentes no periódicos en la voz, de diversos orígenes y que contribuyen a la calidad del habla.

La envolvente espectral es un parámetro clave para hacer posible la síntesis de voz, los algoritmos por los que se suele obtener son *Cepstrum* [29] y *LPC* [2]. El problema principal se encuentra en que el resultado de estos algoritmos depende de la posición temporal, por lo que es clave eliminar en la medida de lo posible la componente de variación temporal sin perder calidad en la estimación. Para obtener la estimación de la envolvente espectral se hace uso de un algoritmo llamado CheapTrick [18] en varios pasos.

1. Se calcula el espectro de potencia de la forma de onda. Previamente se le aplica una ventana de Hanning.
2. La potencia de la forma de onda enventanada se estabiliza temporalmente por medio de una integral.
3. El espectro de potencia se suaviza haciendo uso de una ventana rectangular.
4. Se elimina la componente variante con el tiempo por medio de una operación de *liftering* (aplicación de una ventana en el dominio cepstral)

En WORLD se utiliza un algoritmo de síntesis que hace uso de los mínimos productos de convolución posibles para la obtención de la voz.

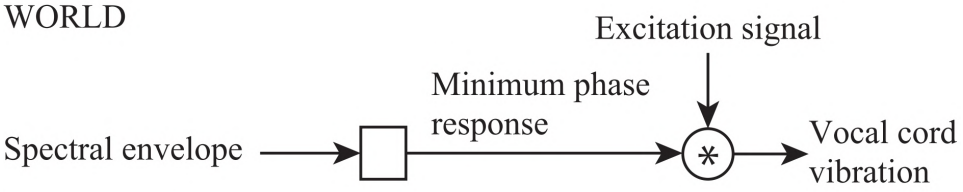


Figura 2.10: Esquema del Vocoder WORLD para síntesis de voz. Fuente: [28]

Lo que se obtiene al sintetizar voz es vibración de las cuerdas vocales en cada momento. Para ello, la vibración se calcula como la convolución de la respuesta de fase mínima de la envolvente espectral y una señal de excitación. El parámetro F0 se utiliza para determinar las posiciones temporales del inicio de cada vibración de las cuerdas vocales. El VoCoder WORLD cuenta con menos operaciones de convolución para la función de síntesis de voz por lo que el coste computacional de este proceso se reduce, permitiendo su aplicación en casos de uso de tiempo real.

### Cepstrum

Es el resutlado de calcular la transformada inversa de Fourier (IFT) al logaritmo del espectro de una señal. Se utiliza en el análisis de voz y aporta información acerca del ritmo de cambio en las franjas del espectro de una señal.

### Mel Frequency Cepstral Coefficients (MFCC)

Son coeficientes que se utilizan para caracterizar de forma compacta el habla y se basan en la percepción humana de la audición. En conjunto conforman un MFCC (Mel-frequency Cepstrum Coefficient), esto es, una representación del espectro de potencia de un sonido en un corto periodo de tiempo. Estos coeficientes representan propiedades de la función de transferencia del tracto vocal y son ampliamente utilizados en aplicaciones de reconocimiento de voz. En nuestro trabajo componen los 'targets' con los que vamos a trabajar. Es decir, en vez de trabajar con la señal de audio directamente, obtenemos los MFCC's de la voz por medio del vocoder WORLD. Para la síntesis de voz se utilizan estos parámetros como entrada a WORLD. Su cálculo se realiza de la siguiente manera:

- Se divide la señal en secciones cortas, típicamente de 20 a 40 ms, con cierto solapamiento para no perder la continuidad.

- A cada una de estas secciones se le practica la Discrete Fourier Transform (DFT).

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N} \quad 1 \leq k \leq K \tag{2.1}$$

  Donde $s_i(n)$ es la señal original divida en secciones, y $h(n)$ es la ventana que se debe aplicar a $s(n)$. Para este cálculo se suele utilizar la ventana de Hanning. $K$ indica la longitud de la DFT. A continuación, se calcula el valor absoluto al cuadrado de la DFT para obtener la potencia espectral.

$$P_i(k) = |\frac{1}{N}S_i(k)|^2 \tag{2.2}$$

- Se aplica al espectro un banco de filtros de la escala *Mel* haciendo uso de ventanas superpuestas triangulares o de Hanning. El interés de utilizar la escala de *Mel* es que es perceptual (es decir, está basada en la sensibilidad del oído humano a las distintas frecuencias). Esta escala es aproximadamente lineal hasta 500 Hz, a partir de los cuales

se establecen intervalos de frecuencia cada vez más amplios para incrementos iguales en la percepción humana del tono. Para la conversión se utiliza la siguente fórmula:

$$M(f) = 1127 \ln(1 + f/700) \tag{2.3}$$

Mientras que para el cambio inverso, se utiliza:

$$M^{-1}(m) = 700(e^{m/1127} - 1) \tag{2.4}$$

- Por último, se practica el logaritmo a las energías de cada frecuencia *Mel* y se realiza la Discrete Cosine Transform (DCT).

#### Transformada de Hilbert

Utilizado en el campo de procesamiento de señales además de en las matemáticas, la transformada de Hilbert es un operador lineal que toma una función y la transforma siguiendo la siguiente fórmula:

$$\hat{x} = x(t) \circledast \frac{1}{\pi t} \tag{2.5}$$

Donde $\circledast$ denota la operación de convolución. La transformada de Hilbert tiene una representación en frecuencia muy simple. Desplaza la fase de las componentes espectrales positivas $-90°$ y para las componentes espectrales negativas $+90°$, mientras que el espectro se mantiene inalterado en magnitud. Por medio de esta operación se puede obtener la envolvente compleja de una señal.

## 2.5. Decodificación del habla a partir de bioseñales

Tal y como se muestra en la figura 2.1, el objetivo final de un sistema SSI es decodificar el mensaje que el usuario quiere expresar a partir de las bioseñales de habla. Por lo general, no se trabaja con las bioseñales directamente, sino que se procesan para disponer de una representación más compacta y que tenga una correlación lo más alta posible con el proceso del habla (*features o características*), tal y como se ha visto en el apartado anterior. El cómputo de la voz a partir dependerá del tipo de bioseñal con el que se esté trabajando. Como se indica en 2.1, existen dos alternativas para decodificar el habla a partir de bioseñales: Conversión a texto (de bioseñales a texto) y síntesis directa de voz (de bioseñales a voz) [41]. En lo que respecta al foco de este estudio, se trabajará con la conversión directa a voz, conocida como Direct Speech Synthesis (DSS). Bajo esta premisa, lo que se busca es realizar realizar una transformación $\boldsymbol{f}: x \rightarrow y$, donde $\boldsymbol{x}$ representa el vector de *features* extraídos de las bioseñales e $\boldsymbol{y}$ representa vector de

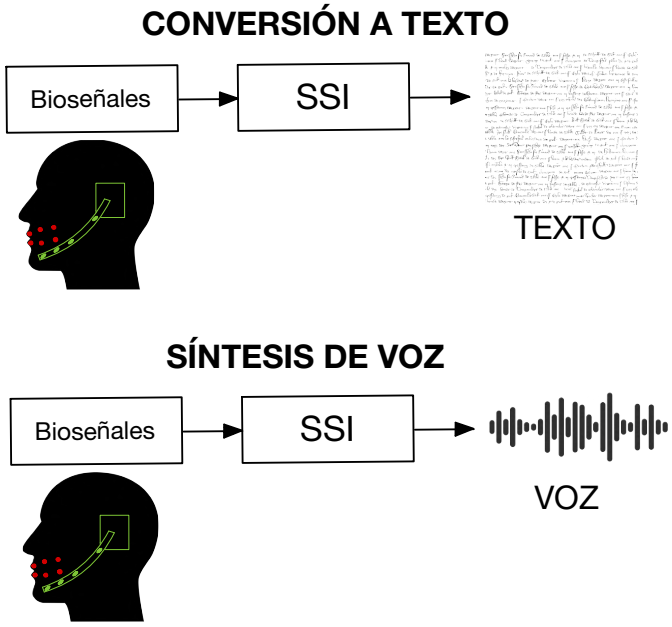


Figura 2.11: Aproximaciones para decodificación del habla a partir de bioseñales

*features* extraído de la señal acústico de voz. Para cada instante de tiempo, se computará lo siguiente:

$$y_t = f(x_t) + \epsilon_t \tag{2.6}$$

Los retos de DSS consisten en que esta función no es lineal y, adicionalmente, la transformación que se realiza no es unívoca, esto es, los mismos *features* acústicos pueden asociarse a múltiples *features* de la bioseñal. Por último, el proceso de registro de las bioseñales no cuenta con resolución infinita temporal o espacial, por lo que es inevitable que parte de la información se pierda.

En función de si el objetivo del SSI es la síntesis de voz o de texto, se puede hacer una distinción entre dos enfoques alternativos.

La figura 2.11 ilustra el concepto. Los siguientes apartados profundizan en cada aproximación.

### 2.5.1. Conversión a texto

El proceso de sintetizar texto a partir de bioseñales tiene la ventaja de que puede ser predicho de una manera más precisa, gracias a los modelos de lenguaje y de pronunciación. Sin embargo, no son capaces de reconocer palabras que no han sido reconocidas durante la fase de entrenamiento.

El proceso de grabado de bioseñales suficiente como para entrenar dichos modelos requiere una masa muestral de entrenamiento significativa. Todo el contexto paralingüístico del proceso del habla (entonación según el ánimo o la identidad del usuario) se pierden en la síntesis de texto. Estas desventajas no suelen ser determinantes, sin embargo, la mayor problemática de este tipo de sistemas es la desconexión entre el proceso de generación de bioseñales y el feedback sonoro del texto siendo enunciado. Esto supone que el sistema no es capaz de funcionar en tiempo real, lo que tiene consecuencias negativas para su uso por pacientes. Según [5], en la comunicación oral, un delay de propagación de 100 a 300 ms causa dubitación por parte del hablante; para retardos superiores a 300 ms los usuarios comienzan a evitar hablar para no interrumpir. En cuanto al feedback auditivo en el proceso de síntesis de texto, los efectos negativos se empiezan a dar a los 50 ms de retardo, considerando como un retardo aceptable hasta 100 ms. [1]. Estos retardos no están al alcance de la técnica de síntesis de texto.

### 2.5.2. Síntesis de voz

La síntesis directa de voz trabaja con las bioseñales para producir habla de forma directa. Esta técnica permite que la latencia pueda ser mucho más baja y que se puedan dar las condiciones de tiempo real. La síntesis directa se han implementado en la bibliografía consultada con sEMG, PMA y EMA. Se abre, además, la posibilidad de que este feedback que recibe el usuario con retardo mínimo abra la puerta a la asimilación de lo escuchado como si fuese su propia voz. Esto permite una mejor modulación por parte del usuario de los parámetros acústicos, además de mejorar la asimilación de este tipo de dispositivos [7].

Dentro de la técnica de DSS, se puede hacer una división en dos metodologías principales, en función de la aproximación que se tome [41]:

#### Conversión basada en modelos

El mapeo de los *features* sensoriales a los *features* acústicos se divide en dos etapas: Estimación de la forma del tracto vocal (modelo que lo represente represente) y síntesis de voz simulando el flujo de aire a través del tracto. [10]. Una desventaja clara de esta técnica de síntesis consiste en que el modelo creado debe tener una alta precisión para obtener buenos resultados. La creación de modelos precisos es compleja y tiene una demanda computacional muy alta. [25].

#### Conversión basada en datos

La metodología de síntesis de voz basada en datos es la más utilizada a día de hoy. En ella, el mapeo se modela como una función paramétrica del

tipo $\boldsymbol{f(x;\theta)}$, donde $\theta$ son los parámetros de la función. Una forma razonable de estimar la tranformación que aparece en 2.6 es haciendo uso de un conjunto de datos que estén etiquetados por pares $(x, y)$, es decir, se usa una aproximación estadística en la que los parámetros de $f(x)$ se estiman para minimizar una función de coste. Esta es la aproximación que más se utiliza a día de hoy en las técnicas de *Machine Learning* y es la adoptada en este trabajo por medio de Unit-Selection. El proceso cuenta con dos etapas principales:

1. **Etapa de entrenamiento:** En la etapa de entrenamiento los parámetros de la voz se estiman usando un set de datos con pares de vectores *source* y *target* $D = (x_1, y_1),....,(x_N, y_N)$ . Este set de pares se obtiene de la captura simultánea de voz y bioseñales en una fase en la que el paciente todavía conserva la voz intacta o no lo suficientemente dañada. En estos casos, las características de la voz se obtienen por medio de una parametrización acústica compacta, típicamente por medio de *MFCC's*.

2. **Etapa de síntesis:** Una vez de han estimado los parámetros $\theta$, se puede utilizar la función de mapeo para sintetizar la voz del paciente por medio de la predicción de las características acústicas contando únicamente con las bioseñales. La voz como tal se sintetiza a partir de estas características como *MFCC's* por medio de *VoCoders* como WORLD [28].

## 2.6. Técnicas de aprendizaje automático para la síntesis de voz a partir de bioseñales

A continación, se van a comentar las técnicas para síntesis de voz directa DSS más relevantes de la bibliografía considerada.

### 2.6.1. Métodos lineales

Como se ha comentado previamente, la condición de trabajo en tiempo real es clave a la hora de determinar la viabilidad de la implementación de un DSS para su uso en pacientes. Para simplificar el proceso, se puede establecer una relación lineal entre las bioseñales obtenidas de los sensores y la voz o los parámetros derivados de la voz como MFCC's. En los modelos de regresión lineal se busca la ecuación lineal que mejor describe la relación entre las señales de las que se dispone y los parámetros de la voz observados. El ajuste más usado para la regresión lineal es el de mínimos cuadrados. Para este ajuste, se obtiene la ecuación de la recta $y = mx + b$ para **N** Parejas de puntos $(x, y)$ tal que el error cuadrático entre la nube de puntos y la recta se minimice. En la ecuación, $x$ comprende los parámetros de la bioseñal e $y$ los

parámetros de la voz. Para obtener la expresión, se deriva y se iguala a cero la expresión del error cuadrático, cuya expresión se muestra a continuación.

$$SSE = \sum (y - \hat{y})^2 \tag{2.7}$$

SSE hace referencia a Sum of Squared Errors. En la siguiente fórmula se indica la ecuación que se obtiene tras la derivación para obtener la expresión de la recta por medio de mínimos cuadrados.

$$m = \frac{N\sum(xy) - \sum x \sum y}{N\sum(x^2) - (\sum x)^2} \tag{2.8}$$

$$b = \frac{\sum y - m\sum x}{N} \tag{2.9}$$

Para que la implementación de un método de regresión lineal funcione con un rendimiento aceptable es necesaria la asunción de que esta relación es de muy baja complejidad, ya que como se ha comentado previamente esta relación no suele ser lineal. La ventaja principal de este método radica en que la implementación se puede llevar a cabo en tiempo real e incluso en dispositivos móviles, debido. En [15] se utiliza un modelo de regresión lineal para sintetizar voz directamente desde la señal obtenida de un sensor PMA, sin disponer de la etapa intermedia de reconocimiento de características y adaptación de la señal sensorial. El modelo correspondiente tiene una complejidad computacional baja, pero los resultados son mejorables, con coeficientes de correlación que se encuentran entre [0.48 , 0.72] para el ajuste lineal. La viabilidad de esta implementación es limitada.

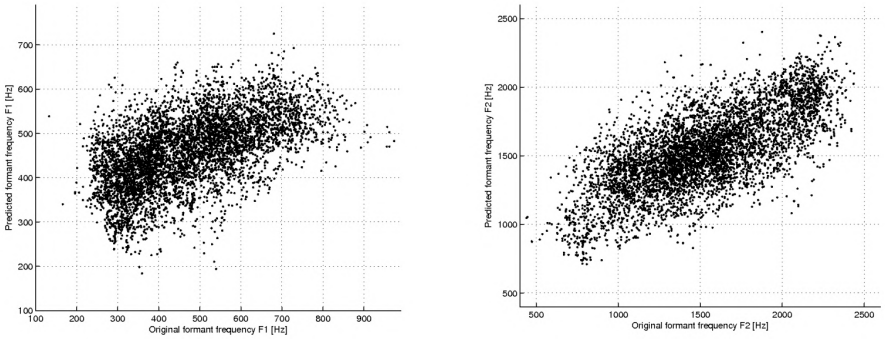


Figura 2.12: distribución de las predicciones para cada paciente en el estudio. Fuente: [15]

### 2.6.2. Métodos no Lineales

Como se ha comentado en apartados anteriores, la tarea de asociación de pares de parámetros de bioseñales y parámetros de la voz suele tener

una distribución no lineal. Es por ello que los métodos que parten de la asunción de que esta relación no es lineal tienen una importancia reseñable en este campo de estudio. Dentro de las técnicas no lineales para síntesis de voz, la que cuenta con un mayor nivel de popularidad en los últimos años es la de Deep Neural Network (DNN). Esta técnica se diferencia de las demás en que utiliza un mayor número de capas para la construcción de la red y que implementa la *propagación hacia atrás (backpropagation)*, por medio de la cual se implementa un método de realimentación en el que las salidas dan información a las capas anteriores en función de la corrección de la predicción. En una DNN se tienen numerosas capas, cada una formada por una serie de neuronas, cuyo esquema se muestra en la siguiente figura:

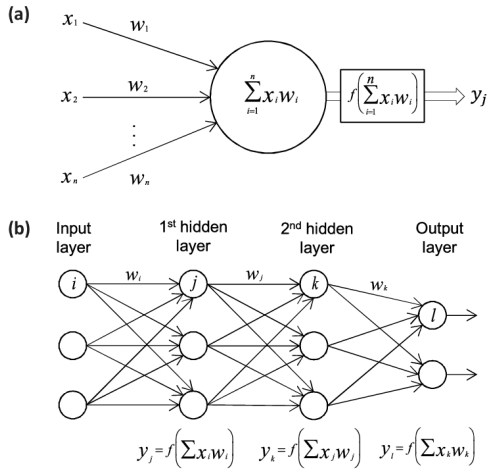


Figura 2.13: Esquema de funcionamiento de una Red Neuronal. Fuente: [35]

En la parte de superior (a) tenemos el esquema de una neurona. Cada neurona toma una serie de entradas, cada una con un peso correspondiente. A la suma de todas las entradas se le aplica un peso y todo pasa por una función de activación no lineal, que determina la velocidad de aprendizaje y precisión del modelo creado, a parte de normalizar la salida. La no linealidad de esta función es clave para modelar relaciones complejas y para el aprendizaje en sí. De esta manera, la salida de cada neurona quedaría:

$$y_j = f(\sum X_i W_i) \tag{2.10}$$

Donde $X_i$ indica cada una de las entradas de la neurona y $W_i$ el peso que se le aplica. En la fase de aprendizaje se determina un valor para cada peso $w$ y se realizan múltiples iteraciones en las que se comprueba para cada una si la salida es la deseada o no. Utilizando una métrica del error (como el Mean Square Error (MSE)) se realiza la propagación hacia atrás para que las neuronas cuenten con un feedback para ajustar los pesos $w$ en futuras

iteraciones. En la parte inferior (b) podemos comprobar el aspecto general que tiene una DNN.

El uso de DNN en síntesis de voz es una opción cada vez más popular debido a que se pueden modelar características multi-dimensionales de una manera eficiente y con unos resultados satisfactorios, por lo que las Redes Neuronales comprenden una herramienta potente en este campo de estudio. Como inconveniente principal, una DNN que incluya un número considerable de capas para potenciar los resultados puede tener asociado una complejidad computacional excesiva, por lo que el uso de DNN dependerá de la aplicación concreta.

En cuanto a su uso en estudios concretos, en [33] se consigue implementar un sistema de conversión a texto a partir de bioseñales EMA haciendo uso de redes neuronales. En [22] se hace uso de una DNN para síntesis directa de voz por medio de bioseñales EMA. La solución por medio de DNN obtiene una mejora significante de la velocidad de síntesis comparado frente aun método estándar de mapeado por distribución gaussiana.

## 2.7. Métricas

Como conclusión al capítulo, se van a exponer a continuación las distintas métricas objetivas que serán utilizadas para evaluar la calidad de inteligibilidad de la voz sintetizada por las técnicas de síntesis de voz directa a partir de bioseñales.

### 2.7.1. Distorsión cepstral en escala Mel

La Mel Cepstral Distortion (MCD) es una métrica que se utiliza para cuantificar la diferencia entre dos sets de MFCC's. Su campo de uso principal es el procesamiento y reconocimiento de voz. Para obtener el MCD se comparan los MFCC's del set de unidades original y el set de unidades sintetizadas. El resultado es una métrica de distancia (menos es mejor). Se puede obtener el MCD haciendo uso de la siguiente ecuación.

$$MCD = \frac{10}{ln(10)} * \sqrt{2 * \sum_{d=1}^{D_a} (mc_d^t - mc_d^s)^2}(dB) \tag{2.11}$$

Donde $mc_d^t$ comprende la unidad de MFCC's objetivo (la original) y $mc_d^s$ comprende la sintetizada. La sumatoria recorre la dimensionalidad completa de las unidades. El MCD se mide en escala decibelios (dB).

### 2.7.2. STOI

En la rama de procesamiento y síntesis de voz, nace la necesidad de medir la inteligibilidad de los audios de una forma robusta y con repetitibilidad.

Este proceso es inherentemente subjetivo ya que está sujeto a la valoración del oyente del audio. Este proceso es lento y costoso, ya que implica que un individuo evalúe el audio, de ahí la necesidad de crear métricas de evaluación automáticas. Las métricas como Short Term Objective Intelligibility (STOI) tratan de evaluar el grado de comprensión del habla en presencia de un ambiente ruidoso. Para ello, se toman las características acústicas y lingüísticas de la voz. El proceso de obtención del STOI se lleva a cabo para dos señales, original y sintetizada. Cuenta con los siguientes pasos (Fuente: [14]):

1. Se obtiene una representación Tiempo-Frecuencia de la señal segmentando sendas señales en frames solapados al 50 % y procesados con una ventana de Hanning. A cada frame se le añade un padding (se añaden ceros) y se le practica la transformada de Fourier.

2. Se lleva a a cabo un análisis para cada tercio de octava en cada frame. Una octava representa un duplicado o una división por dos de la frecuencia. En total se analizan 15 tercios de octavas.

3. La medida de inteligibilidad para un frame de Tiempo-Frecuencia depende de una región de frames consecutivos. Para dicha región de frames, se practica una normalización para que la energía de esa región coincida con la región correspondiente del audio original. A continuación se obtiene una métrica conocida como SDR (Signal to Distorsion Ratio), similar a la SNR.

4. La métrica de la inteligibilidad se obtiene finalmente como una estimación del coeficiente de correlación lineal entre los frames originales y los frames modificados anteriormente.

# Capítulo 3

# Método Propuesto

El objetivo de este capítulo va a ser la especificación del caso de uso concreto con el que se trabaja, así como la descripción en profundidad del algoritmo diseñado para la síntesis de voz a partir de las bioseñales de habla obtenidas. Una vez estudiados los distintos métodos para la obtención de bioseñales y las técnicas para cada situación, se concluyó que se trabajaría con una base de datos de bioseñales obtenidas por medio de la técnica PMA [26] [31]. Esta base de datos contiene registros simultáneos de bioseñales y voz, por lo que es apto para diseñar un modelo *basado en datos*. Partiendo de este set de datos, el objetivo es la implementación de un algoritmo de Unit-Selection (Selección de unidades), que relaciona unidades de bioseñales PMA con unidades de MFCC's de la voz. Se conforma, por tanto, un sistema de Síntesis Directa de Voz (DSS). Para la implementación en código del algoritmo se hará uso del lenguaje de programación *Python*. En la siguiente figura 3.1 se muestra un diagrama de bloques del método propuesto:

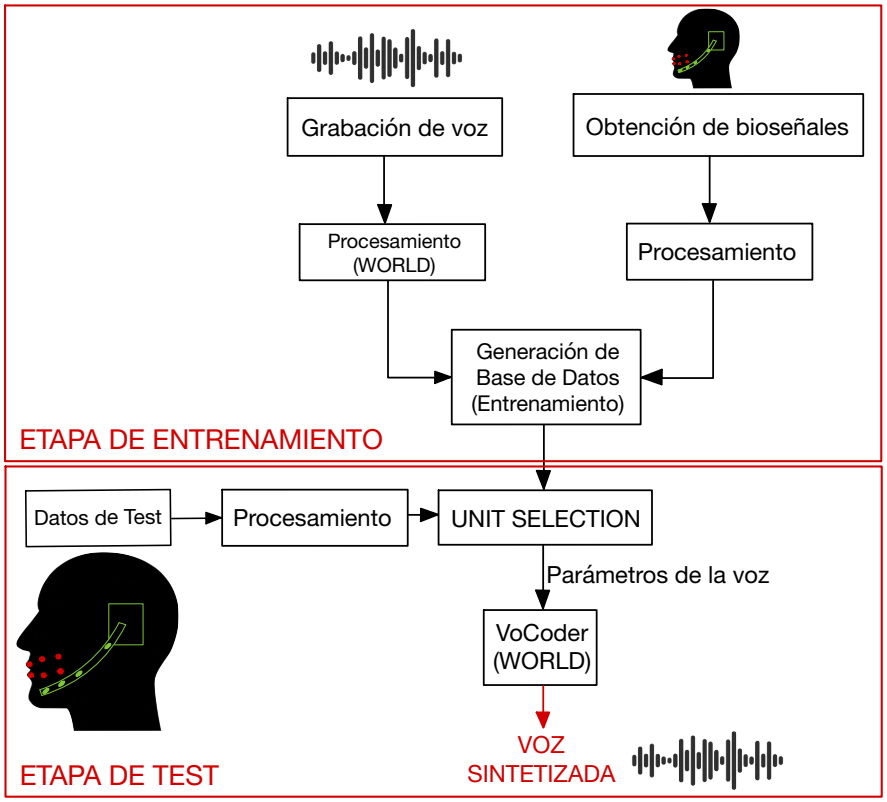


Figura 3.1: Esquema de la metodología aplicada.

El primer paso consiste en la obtención síncrona de la voz y las bioseñales PMA. Ambas se procesan y se modifican en base a lo establecido por el algoritmo para formar la base de datos. A continuación, el algoritmo de Unit Selection utiliza la información disponible en la base de datos para predecir los parámetros de la voz que van asociados a los *features* de la voz. Por último, el VoCoder WORLD toma estos parámetros de la voz para sintetizar voz natural. En las siguientes secciones se profundizará en los detalles de cada fase.

## 3.1. Obtención de datos

En este capítulo se detallará la naturaleza y la obtención de las bioseñales, así como el trabajo de preprocesado. Para este trabajo se contaba con datos PMA obtenidos de dos estudios pasados.

### 3.1.1. Adquisición y procesado de señales PMA

Para el primer estudio, la obtención de bioseñales de individuos sanos. El dispositivo diseñado para la obtención de los datos PMA 3.2 se muestra en la siguiente figura.

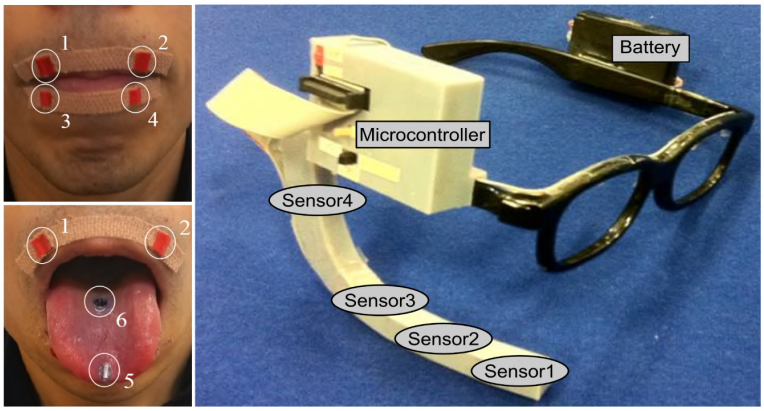


Figura 3.2: Descripción general de la técnica PMA para obtención de bioseñales articulares. Fuente: [31]

En la parte izquierda se puede observar la disposición de los sensores. Para este estudio se eligieron dos articuladores del habla: labios y lengua. Se disponen 6 imanes de Neodimio-Hierro-Boro (NdFeB) de 1 mm de diámetro en 6 posiciones distintas. Los imanes se fijan utilizando un adhesivo quirúrgico (en aplicaciones permanentes, este imán se implanta por medio de una operación quirúrgica).

Para la percepción del campo magnético de los imanes, se disponen 4 sensores en las posiciones indicadas. La colocación de los sensores se diseñó a medida de la anatomía de uno de los sujetos. Los sensores 1,2,3 están dispuestos para obtener el campo magnético proveniente del movimiento de los articuladores del habla, mientras que el cuarto sensor se utiliza como referencia del campo magnético de fondo para compensar el efecto del campo magnético de la tierra. Los datos PMA se muestrean a 100 Hz y se retransmiten por medio de un transmisor *Bluetooth* a un PC para su procesamiento. Adicionalmente, se lleva a cabo un filtrado paso-baja a 50 Hz para eliminar el posible ruido eléctrico. Todos los pacientes utilizaron el mismo dispositivo para la captura de las señales.

La sesión de grabado es síncrona, se grabó de forma simultánea el audio y las señales PMA de 9 canales, utilizando frecuencias de muestreo de 16 kHz y 100 Hz, respectivamente.

Para el segundo estudio, se utilizó un apartado de captura similar al de la figura 3.2, en el que se cuentan con 3 sensores y un cuarto sensor para la eliminación del campo magnético de fondo.

En cuanto al procesamiento de las grabaciones de audio, en este trabajo no se trabaja con las señales de voz sin tratar, si no que los vectores están compuestos de coeficientes espectrales (MFCC's) obtenidos de la voz, de dimensión 25. Para obtener los MFCC's de la voz, se utiliza el VoCoder WORLD, cuyo funcionamiento quedó especificado en 2.4.2.

Para procesar las señales de PMA, la única modificación que se practicó fue la eliminación del campo magnético terrestre de las señales PMA. Para ello, se le indicaba a cada paciente que realizase unos movimientos de cabeza determinados mientras mantenía los articuladores del habla (labios y lengua) estáticos. De esta manera se podía sustraer el efecto del campo magnético terrestre.

Se ha de tener en cuenta que lo que se obtiene de los sensores PMA no es una representación cartesiana de la posición de los articuladores, sino una suma de las señales obtenidas a raíz del movimiento de los mismos, a partir de la cual se obtendrán los coeficientes cepstrales MFCC's de la voz. Esta relación del movimiento recogido por los sensores y los coeficientes cepstrales de la voz sigue una lógica no-lineal. Adicionalmente, la técnica PMA implementada no permite obtener información acerca de parámetros adicionales como la frecuencia fundamental $F_0$ y la *aperiodicidad*, por lo que el audio sintetizado sería *unvoiced* (como un susurro). Por simplicidad, se ha optado por obtener estos parámetros de los audios originales grabados, dejando como trabajo futuro la obtención de estos parámetros por medio de otros métodos.

El procesado de las señales obtenidas es idéntico de un artículo a otro. La frecuencia de muestreo de las señales se mantiene inalterada a lo largo de los dos estudios.

## 3.2. Algoritmo de síntesis de voz a partir de bioseñales

Este capítulo se va a centrar en la especificación del algoritmo de Unit Selection creado, incluyendo la generación de la base de datos que se utiliza para la síntesis. El algoritmo parte de una división de las señales PMA y de voz en pequeñas secciones llamadas unidades. Estas unidades forman una base de datos en las que se tienen parejas de unidades PMA y de voz, existe una relación unívoca entre estas parejas de unidades. En la etapa de síntesis de voz, en la que se cuenta únicamente con las bioseñales PMA, se hace uso de la información de la base de datos para predecir la secuencia de unidades de parámetros de voz (camino de Viterbi) a partir de las de PMA, aplicando técnicas de evaluación de distancia

### 3.2.1. Etapa de Entrenamiento

Una vez se cuenta con las señales PMA procesadas, así como con los MFCC's de la voz, se procede a generar la base de datos que utilizará el método de Unit Selection para la síntesis de voz. El algoritmo de Unit Selection (especificado más adelante) hace uso de unidades (pequeñas porciones de señales PMA y coeficientes MFCC), por lo que la base debe contener unidades de MFCC's y de PMA. La creación de cada tipo de unidades se lleva a cabo por separado.

#### Generación de la base de datos

Para cada uno de los 2 sets que componen la base de datos, el trabajo llevado a cabo para su creación queda especificado a continuación.

- **MFCC's:** Por un lado, tras procesar las señales de audio por el VoCoder WORLD, se obtienen los MFCC's de la voz con un tamaño de trama de 5 o 10 ms (el parámetro se puede modificar al ejecutar el script de WORLD). Este parámetro deberá estar de acuerdo con la frecuencia de muestreo de las señales PMA, de forma que el tamaño de trama coincida con el periodo de muestreo de la señal PMA. Por ejemplo, una frecuencia de muestreo de 100 Hz corresponde con un tamaño de trama de 10 ms. La justificación a esto reside en que tiene que haber una correspondencia temporal entre los MFCC's y las señales PMA. Ambas han sido obtenidas de forma síncrona y es clave para el correcto funcionamiento del algoritmo que el sincronismo se mantenga. Las distintas unidades que componen la base de datos formada por los MFCC's se obtienen directamente de la salida del VoCoder, no necesitan procesamiento adicional.

- **PMA:** Para la creación de las unidades PMA que van a formar la base de datos, se creó una función que lleva a cabo la conversión a unidades de cada fichero individual. Esta función toma como parámetros de entrada un fichero que contenga la señal PMA y su frecuencia de muestreo asociada, así como el tamaño de la ventana y el solapamiento. Como indicado previamente, el solapamiento será forzosamente la inversa de la frecuencia de muestreo de la señal. Las unidades que se obtengan del fichero serán segmentos solapados del mismo, contando todos con el mismo tamaño. La siguiente figura ilustra el proceso de creación de la base de datos de unidades.

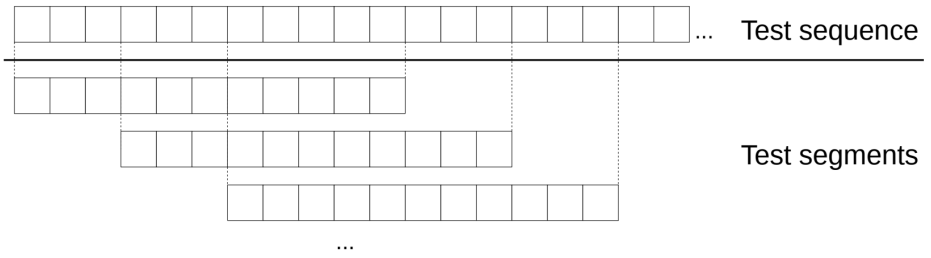


Figura 3.3: Obtención de las unidades para señales PMA. Fuente: [20]

La parte superior *Test Sequence* conforma el fichero PMA, mientras que en la parte inferior *Test segments* se tienen las unidades individuales. Se puede denotar el tamaño unitario de la ventana como $w_u$ y el solapamiento unitario como $s_u$. De esta manera, en la figura 3.3 contamos con $w_u = 11$ muestras y $s_u = 3$ muestras. El parámetro $s_u$ nos viene determinado por la señal obtenida, pero $w_u$ es un parámetro que podemos modificar y que afectará al resultado de la síntesis de voz. El efecto que tiene aumentar el tamaño de la ventana $w_u$ consiste en que para cada unidad se cuenta con un mayor contexto temporal (se tiene una mayor cantidad de información de las unidades vecinas) por lo que puede mejorar el resultado. La modificación de este parámetro será evaluada y puesta en contexto con los resultados obtenidos.

Una vez se han obtenido todas las unidades de un fichero, el proceso se repite para todos los ficheros que conforman la secuencia de entrenamiento y de test.

### Normalización de las unidades

Para poder establecer las métricas de forma correcta es conveniente normalizar todos los parámetros con los que se esté trabajando. La normalización de los parámetros es una práctica de suma importancia en este tipo de trabajos por dos razones.

- Diferencias de escala: Los distintos valores dentro de una misma unidad pueden tener escalas distintas y muy dispares. Esta disparidad puede desvirtuar el cálculo a la hora de obtener las distancias, por lo que practicando una normalización a la base de datos se evita esa problemática.
- Diferencias entre unidades: Los rangos de valores en los que trabaja cada unidad (de sensor y de MFCC) pueden moverse en zonas distintas. Debido a que en la evaluación de métricas se incorporan medidas

de sensor así como de MFCC, la normalización aun rango concreto evita que una distancia tenga una prevalencia inadecuada sobre otra.

Para este método, se ha optado por una normalización en el rango [0,1] haciendo uso de la librería de *sklearn MinMaxScaler*. La librería transforma las unidades por medio de un escaladado al rango especificado. Para llevarlo a cabo, se realiza un escalado de la base de datos de entrenamiento en base a la estadística de la misma. Para las unidades de test, se aplica el mismo escalado usando la estadística de la base de datos de train. Cada parámetro se escala de manera individual de manera que se encuentre en el rango del set de datos. Este tipo de transformaciones son habitualmente utilizadas como alternativa al escalado de media nula y varianza unitaria.

### 3.2.2. Síntesis de Voz

En esta sección se detallará el funcionamiento del algoritmo Unit Selection creado para la síntesis de voz. Esta URL[1] enlaza a un repositorio en el que se han subido los principales scripts de el algoritmo implementado.

El método de Unit Selection ha sido el estándar durante muchos años para la síntesis de voz por medio de bioseñales, debido a que es capaz de generar voz con una calidad aceptable con un método que no tiene una complejidad prohibitiva. El método original de Unit Selection parte de la premisa de que se pueden sintetizar pronunciaciones naturales seleccionando pequeñas unidades (secciones de palabras pronunciadas) obtenidas de una base de datos de voz. La siguiente figura ilustra el concepto.

[1] https://github.com/javilobato/speech-synthesis.git

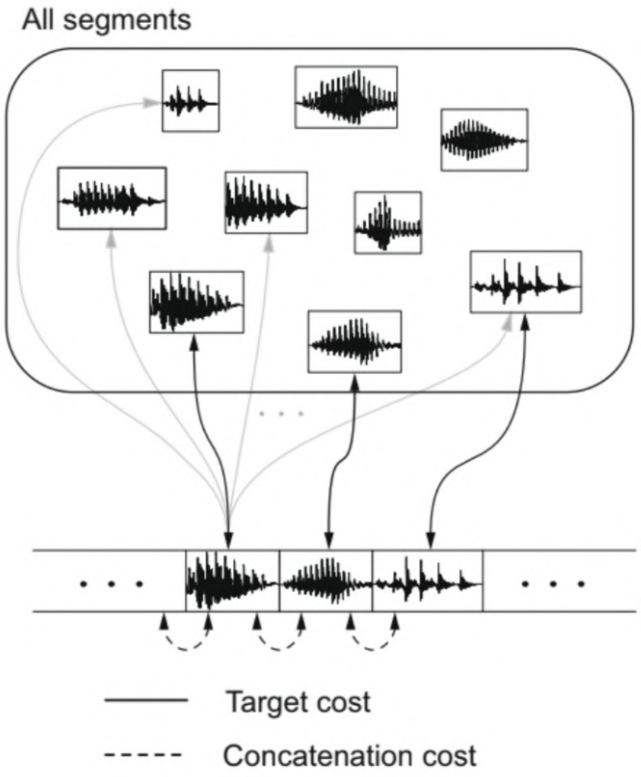


Figura 3.4: Esquema de funcionamiento para Unit Selection. Fuente: [21]

La métrica principal que se utiliza para la selección de la unidad óptima dentro de la base de datos es el *coste objetivo*, esta indica la similitud entre la unidad que se está evaluando y la unidad correspondiente de la base de datos. El resultado obtenido utilizando únicamente esta métrica no es suficiente para un buen resultado, por lo que se incorpora el coste de concatenación, que indica cómo de bien se combinan dos unidades seleccionadas [21]. Podemos definir el *coste objetivo* como:

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \tag{3.1}$$

Donde $j$ se itera sobre todos los componentes que componen la unidad en cuestión, $t$ simboliza *target* y $u$ simboliza una unidad específica de la base de datos. El parámetro $w$ implementa un peso que se utiliza para priorizar cualquiera de los dos costes. El coste de concatenación se define como:

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i)$$

(3.2)

Donde $k$ se itera sobre los componentes de la unidad en evaluación. Se implementa también un peso $w$ para establecer prioridades.

Ambos costes deben ser optimizados para obtener la secuencia de unidades $u_{1:n} = u_1, ..., u_n$ de la base de datos que optimice el coste total, tal que:

$$\hat{u}_{1:n} = argmin_{1:n} C(t_{1:n}, u_{1:n}) \tag{3.3}$$

Donde, $C(t_{1:n}, u_{1:n})$ denota la función de coste total. La solución de la ecuación superior 3.3 conforma el camino de Viterbi, en el sentido de que indica, para cada unidad, la secuencia más probable de unidades. En cuanto a la ecuación que determina el coste total:

$$C(t_{1:n}, u_{1:n}) = \sum_{i=1}^{n} C^{(t)}(t_i, u_i) + \sum_{i=2}^{n} C^{(c)}(u_{i-1}, u_i) \tag{3.4}$$

La tarea de elección de los pesos dependerá del caso de estudio concreto y no existe un criterio fijo para la determinación del mismo. También es determinante la longitud de la unidad. Cuanto más grande sea la unidad, más grande tendrá que ser la base de datos para cubrir el dominio necesario. Unidades más pequeñas pueden dar un mejor rendimiento al ofrecer más puntos potenciales de unión.

En referencia a el trabajo realizado y siguiendo el orden de los subapartados, una vez terminado el proceso de creación de la base de datos, se tiene un array de unidades que conforman la base de datos de entrenamiento y un array de unidades PMA. El objetivo es la obtención de un array de unidades de coeficientes MFCC a partir del array de unidades PMA haciendo uso del algoritmo de Unit Selection.

### Evaluación de la distancia (coste objetivo)

La evaluación de la distancia implementa la ecuación 3.1 . Esta métrica de la similitud de dos unidades se puede conseguir calculando la distancia euclídea de dos vectores.

$$C^t = \sum_{k=1}^{w_u} \sqrt{\sum_{d=1}^{D_s} (s^t_{test}(k,d) - s^t_{train}(k,d))^2}, \tag{3.5}$$

donde $k$ se itera a lo largo de todos los frames que componen una unidad y $d$ se itera para cada dimensión de la señal PMA. $D_s$ denota la dimensionalidad de la señal PMA (s indica source). El término *train* hace referencia al número de unidades de entranamiento en la base de datos.

El cálculo de la distancia objetivo definida en (3.5) en la fase de test puede ser un proceso costoso computacionalmente, ya que para cada unidad de test se debe evaluar la distancia con todas y cada una de las unidades que componen la base de datos. Es por ello que, para optimizar dicho cálculo, se optó por utilizar una estructura de datos en árbol para particionado de los datos de entrenamiento que permitiese realizar la evaluación de la distancia de la forma más eficiente posible. En concreto, la estructura de datos usada se conoce como *Ball Tree* y permite organizar vectores de datos en un espacio multidimensional, optimizando con ello la búsqueda del vecino más próximo [24]. Esta búsqueda es la que implementamos en nuestro trabajo al evaluar la distancia y obtener la menor posible. Como se ilustra en la figura 3.5, el algoritmo construye un árbol jerárquico que parte de una nube de puntos (que contiene los elementos de la base de datos). El árbol se construye creando *'esferas'*, primero se toma un punto cualquiera como centro de una esfera. Esta esfera tendrá un tamaño (*leaf size*). El tamaño de la esfera determina la velocidad de búsqueda y el tiempo que tarda en crearse la estructura, si bien no afecta a la calidad de los resultados obtenidos. A continuación, se dividen los puntos en dos grupos: Los del interior de la esfera y los del exterior. En el siguiente paso se vuelven a tomar dos puntos, uno dentro de la esfera y otro fuera, estos dos puntos comprenden los centros de otras dos esferas. Este proceso se itera hasta que se cumple una condición de finalización (que se tenga un tamaño de dao de esfera, por ejemplo). Para nuestro caso, esta estructura permite que se puedan evaluar de forma eficiente los $N$ vecinos más próximos de una unidad concreta. De esta manera, se modificarían los limites de la ecuación 3.4 , ya que no se tendría que evaluar la distancia de cada unidad con las M unidades de entrenamiento, sino sobre los N vecinos más próximos. Quedaría, por tanto:

$$C(t_{1:N}, u_{1:N}) = \sum_{i=1}^{N} C^{(t)}(t_i, u_i) + \sum_{i=2}^{N} C^{(c)}(u_{i-1}, u_i) \tag{3.6}$$

En nuestro caso, usamos la implementación que hace la librería de *Python scikit-learn* de la estructura *BallTree*.

### Coste de concatenación

Para obtener la métrica del coste de concatenación se hace uso de la distancia cepstral, ya que ahora se trabaja con la similaridad de dos unidades de MFCC's consecutivas. Esta métrica se obtiene como la distancia euclídea entre los MFCC's de la unidad óptima anterior y la unidad actual en evaluación.

$$C_c(t_{1:N}) = \sum_{k=1}^{w_{u(MFCC)}} \sqrt{\sum_{d=1}^{DA} (t_{train}^{t+1}(k,d) - t_{train}^{t}(k,d))^2} \tag{3.7}$$

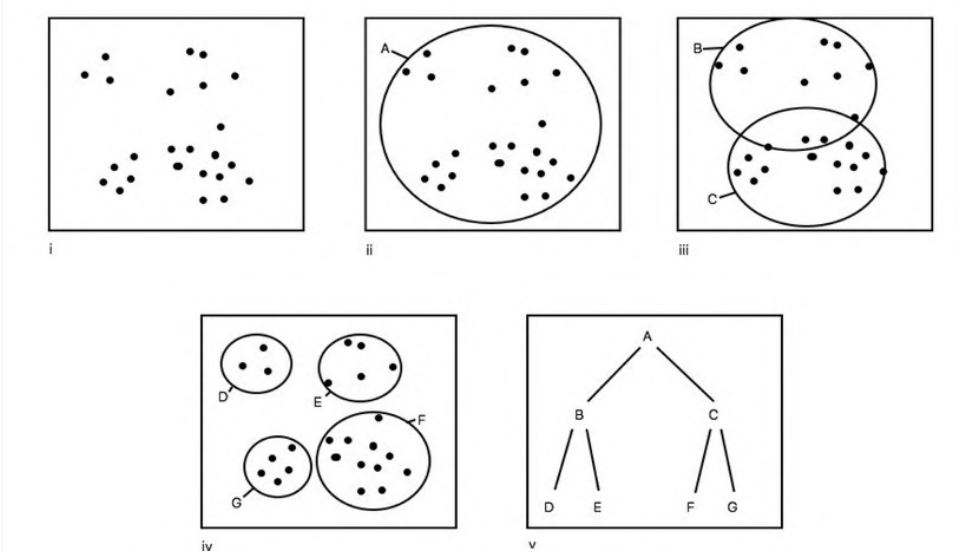


Figura 3.5: Construcción del árbol para la técnica *Ball Tree*. Fuente: [34]

Donde $k$ se itera a lo largo del tamaño (número de elementos) de la unidad de MFCC's y $d$ se itera a lo largo de la dimensionalidad de la unidad.

### Incorporación de métricas

Al evaluar el coste objetivo, obtenemos las $N$ unidades de la base de datos más similares a la unidad de test que se evalua. Esto se hace porque la existencia del coste de concatenación puede dar como mejor unidad candidata una que no sea la de menor coste objetivo. Por ejemplo, se puede dar el caso de una unidad que sea la más próxima a la de test siendo evaluada pero que la transición de una unidad MFCC a otra sea muy abrupta. De esta manera, se toman varios candidatos a ser la unidad más próxima y se evalúa el coste de concatenación para cada uno de los candidatos. El menor coste indicará la unidad que se debe elegir. Para la incorporación del peso $w$ en la métrica, se incorpora un factor multiplicativo al coste de concatenación que pueda hacer que este sea más o menos importante a la hora de determinar la unidad óptima. La consideración de varias unidades candidatas provoca que dispongamos de otro parámetro que puede afectar en los resultados obtenidos: el número de vecinos que se tienen en consideración. Este parámetro será modificado y tenido en cuenta para la obtención de resultados.

El proceso de la elección de las unidades óptimas se va conformando en una matriz de costes que se va rellenando con los distintos costes. Las unidades de MFCC's anteriores son tenidas en cuenta para el coste de concatenación. Por el funcionamiento del algoritmo, al recorrer todas las unidades PMA de test y obtener los costes correspondientes, se dispone de un array

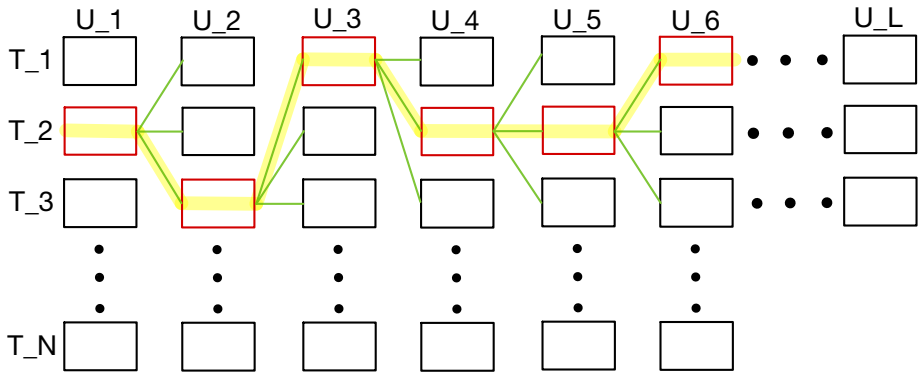


Figura 3.6: Ilustración del proceso de elección de la unidad óptima.

de unidades MFCC que minimiza la función de coste para cada una de las unidades. El camino de menor coste se va determinando conforme se rellena la matriz de costes. En la primera iteración, la unidad óptima es la unidad de menor coste objetivo, ya que no hay concatenación con una unidad anterior. Para la segunda unidad, se obtiene el coste objetivo para cada uno de los N vecinos y se calcula el coste de concatenación con la unidad óptima anterior. Esto da como resultado una unidad óptima (que minimiza ambos costes). Este proceso se repite de forma iterativa para ir construyendo el camino óptimo. Esta manera de obtener los MFCC's se conoce como *backtracking*. Cabe notar que el proceso de minimización de coste se puede llevar a cabo haciendo uso del algoritmo de Viterbi. Como se ha comentado en apartados anteriores, la salida final del algoritmo diseñado obtiene una secuencia de unidades que coincide con el camino de Viterbi, entendido como la secuencia de unidades de menor coste asociado.

La siguiente figura 3.6 ilustra el proceso de elección de la unidad óptima.

En la figura se puede observar una serie de unidades candidatas. Para cada unidad PMA se tienen $N$ unidades candidatas MFCC (dispuestas en vertical en la figura). Estas unidades se encuentran dispuestas en orden descendente en función del coste objetivo. Para la primera unidad de todas, la decisión de la unidad óptima se hace únicamente en función de este coste objetivo (aparece en rojo en la figura). Las líneas verdes que parten al resto de unidades indican que el coste de concatenación se calcula entre la unidad óptima anterior (en rojo) y las $N$ unidades candidatas. Evaluando la unidad de las $N$ candidatas que tiene el menor coste asociado (sumando el objetivo y el de concatenación) se determina la nueva unidad óptima. Este proceso se repite hasta que se han determinado todas las unidades completas para las $L$ unidades en las que se divide la señal. La secuencia de unidades óptimas (El camino de Viterbi) aparece resaltada en amarillo en la figura 3.6.

Una vez terminado el proceso, se dispone con una secuencia de unidades

de parámetros MFCC's, cuyo número coincide con el número de unidades PMA de test de las que se disponía. A partir de estos coeficientes MFCC se sintetiza la voz haciendo uso del VoCoder WORLD, especificado en la sección 2.4.2.

A modo de ejemplo, se muestra a continuación la representación temporal de un audio sintetizado por medio de este método, así como el espectrograma asociado al audio. La señal utilizada proviene del estudio correpondiente a una base de datos de dígitos.

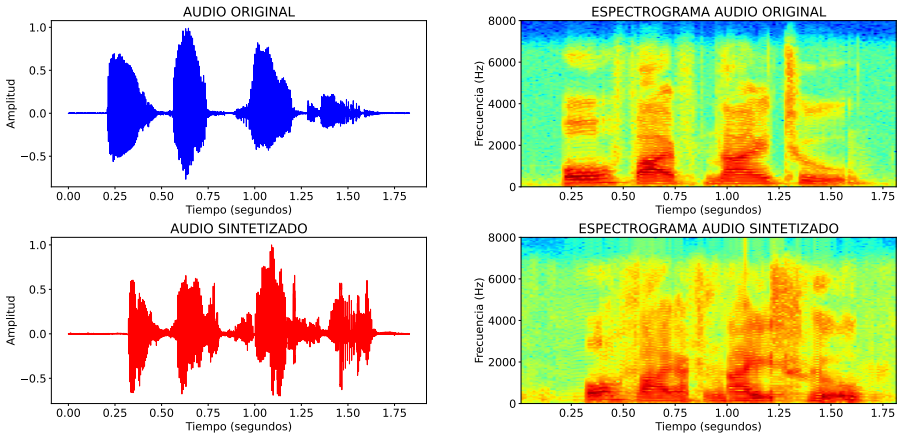


Figura 3.7: Ejemplo de audio sintetizado por medio del algoritmo de Unit Selection implementado. Representación temporal y espectrograma

El ejemplo de 3.7 corresponde a una serie de dígitos, en la que se pronuncia la secuencia de números en inglés *'zero five nine two'*.

## 3.3. Ventajas del algoritmo

El uso del algoritmo implementado cuenta con una serie de ventajas sobre el resto de implementaciones comentadas en el capítulo 2 que hacen que su uso sea más apropiado en determinadas casuísticas. Quedan enumeradas a continuación:

- Como se ha indicado previamente, la relación que existe entre los parámetros obtenidos de las bioseñales y los parámetros de la voz, suelen no seguir una distribución lineal. El funcionamiento del algoritmo de unit-selection no aplica un modelo lineal a los datos, por lo que los resultados obtenidos deben ser mejores que usando técnicas no lineales, como por ejemplo regresión lineal.
- El algoritmo implementado de Unit-Selection es un método no paramétrico: utiliza para la síntesis de voz una base de datos de co-

eficientes MFCC de personas reales hablando. Esto se encuentra en contraposición con los modelos paramétricos, que sintetizan la voz por medio de modelos matemáticos y estadísticos. El uso de modelos paramétricos requiere utilizar una cantidad significativa de datos de entrenamiento que en ocasiones debe provenir de variedad de hablantes. La ventaja de un método no paramétrico consiste en que la secuencia de entrenamiento puede ser corta, se puede conseguir síntesis de voz con una calidad aceptable haciendo uso de una base de datos de un tamaño relativamente contenido.

- Al poder contar con datasets pequeños para la implementación de la base de datos, el algoritmo que se implementa es ligero y tiene un coste computacional asociado bajo en comparación con otros métodos, por lo que esto facilita su implementación en tiempo real, característica clave en los sistemas de síntesis de voz.

- Al hacer uso de porciones reales de audio que se concatenan para obtener el audio sintetizado, la distorsión asociada a la voz es baja en comparación con el resto de métodos que la sintetizan a través de un modelo con parámetros.

# Capítulo 4

# Resultados obtenidos

Este capítulo está destinado a presentar los resultados experimentales obtenidos tras evaluar el método propuesto para síntesis de voz a partir de bioseñales propuesto. Así, se mostrarán los resultados obtenidos en cuanto a métricas objetivas de calidad e integilibilidad de la voz obtenidas por nuestro método, así como una comparación con el método base basado en regresión lineal.

## 4.1. Marco de evaluación experimental

Previo a la exposición de los resultados, se presenta el marco de evaluación experimental, en el que se especifica las bases de datos utilizada, procesamiento de las mismas, métricas utilizadas y métodos evaluados.

### 4.1.1. Datasets para síntesis de voz

En este trabajo se han utilizado señales provenientes de otros estudios [26] [31] en los que se registra voz y bioseñales PMA de individuos sanos. En cuanto a las palabras/oraciones que pronuncian los pacientes, es conveniente recurrir a bases de datos de dígitos u oraciones diseñada con anterioridad. Estas aportan la ventaja de que están formadas y reguladas, por lo que permiten disponer de ficheros adicionales como dígitos pronunciados o la frecuencia de aparición de cada uno. Es de interés que las bases de datos estén balanceadas fonéticamente, en el sentido de que todos los fonemas aparezcan en proporciones similares.

En referencia a la base de datos para los pacientes del primer estudio, se eligió la base de datos *TiDigits*, en la que cada grabación consiste en una secuencia de uno a siete dígitos pronunciados en inglés. El vocabulario total contiene 11 palabras: los dígitos de 'uno' a 'nueve', además de 'cero' y 'oh' (esta última, pronunciación alternativa para el cero). En total, hay 21 fonemas, 11 vocales y 10 consonantes.

Para cada paciente, se disponen de 308 oraciones grabadas. La información de este estudio comprende los dígitos pronunciados por dos locutores.

En cuanto a la base de datos usada en el segundo estudio, los pacientes pronunciaban oraciones provenientes de la base de datos del *Carnegie Mellon University (CMU)*. Esta base de datos, de nombre *Arctic* contiene un rango de oraciones fonéticamente ricas que permite evaluar la reconstrucción del habla en un rango fonético amplio. Se cuentan con 1132 oraciones seleccionadas de libros ingleses. Se disponen de 470 y 510 oraciones grabadas de dos sujetos sanos de la misma manera que para el primer estudio. Ambos eran hombres.

Para tener una referencia de los distintos registros bioseñales/voz con las que se trabaja, la siguiente tabla 4.1.1 refleja el origen y características de cada una de los datasets con las que se sintetiza voz.

| Base de Datos | Locutor | Nº de ficheros | Duración total (minutos) |
|---|---|---|---|
| TiDigit | LC | 308 | 8 |
| TiDigit | TP | 308 | 8 |
| Arctic | JG | 470 | 26 |
| Arctic | RM | 510 | 28 |

Cada uno de los datasets contiene un set de bioseñales PMA y audios de voz grabados de forma síncrona. En cada fichero del dataset se enuncian unos dígitos u oraciones concretas. Para la creación de la base de datos se toman el 90 % de los ficheros para entrenamiento, mientras que el 10 % restante se utilizan para sintetizar voz. El proceso se repite 10 veces, ya que el parámetro de la partición *K-Fold* es $K = 10$.

La presentación de los resultados se dividirá en varias secciones en las que se irá modificando cada uno de los parámetros del algoritmo creado que pueden afectar al resultado.

### 4.1.2. Procesamiento de los datos

Como se ha indicado en previos apartados, los datos contenidos de la base de datos comprenden bioseñales PMA y audios de voz. El procesamiento aplicado a cada uno de ellos:

- **Bioseñales PMA**: Como se ha especificado en 3.1.1, las bioseñales que se utilizan para crear las unidades del algoritmo cuentan con un procesado en el que se elimina el campo magnético terrestre haciendo uso de un sensor de referencia, eliminando así posibles influencias externas.

- **Audios de voz**: Como se indica en 3.2.1, los audios de voz son procesados por medio de VoCoder WORLD, especificado en 2.4.2, a partir

de los cuales se obtienen los coeficientes MFCC's que el algoritmo utilizará para la síntesis de voz. Estos coeficientes pasan a formar las unidades del algoritmo.

### 4.1.3. Evaluación y métricas

En el trabajo realizado se lleva a cabo la partición de la carpeta de datos de cada paciente siguiendo la filosofía de Validación cruzada (*K-Fold*). Esta técnica se utiliza para poder tener una estimación robusta del rendimiento del algoritmo creado al trabajar con datos desconocidos (datos de *test*). El funcionamiento es simple: La carpeta de cada paciente se particiona en $K$ subconjuntos iguales. El algoritmo se evalúa $K$ veces. De esta manera, la mayoría de los subconjuntos se utilizan para conformar la base de datos de entrenamiento, mientras que los restantes conforman la base de datos de test, por medio de la cual se evaluará el rendimiento del algoritmo. Al realizar $K$ iteraciones, el procedimiento habrá llegado a término y todos los componentes de la carpeta habrán formado parte de la base de datos de entrenamiento y habrán sido utilizados para test. Este hecho permite que los resultados que se obtengan de las métricas asociadas a los resultados del algoritmo sean mucho más robustas que si se hubiese realizado una partición única, ya que se tiene representación de la distribución completa de los datos.

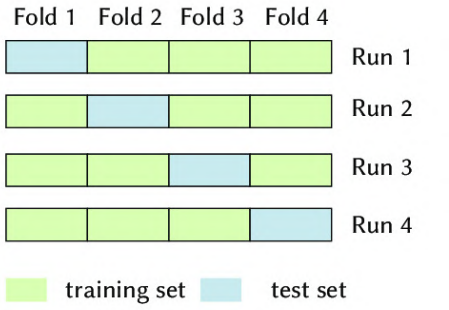


Figura 4.1: Ejemplo de funcionamiento de *K-fold* con $K = 4$. Fuente: [43]

Para el trabajo desarrollado, se toma $K = 10$, de manera que en cada iteración 9 *folds* van destinados a la creación de la base de datos y 1 *fold* se utiliza para test.

En cuanto a las métricas utilizadas para la evaluación del algoritmo, se utilizará la distorsión Mel-Cepstral MCD para evaluar la similitud de las unidades de MFCC's originales y sintetizadas, mientras que para la evaluación de la inteligibilidad se hará uso de la métrica STOI. Ambas han sido descritas en 2.7.

### 4.1.4. Métodos evaluados

Para la evaluación de los resultados, se obtendrán las métricas descritas para dos algoritmos distintos de síntesis de voz a partir de bioseñales: Unit Selection, descrito en profundidad en , así como Regresión lineal, descrito a continuación en .

## 4.2. Método Base: Regresión Lineal

Los resultados obtenidos por medio de Unit Selection se evaluarán por medio de métricas objetivas usadas ampliamente en aplicaciones de síntesis de voz. Para tener un mayor nivel de contexto, se ha implementado un método alternativo para poder comparar los resultados frente al método implementado. Se eligió el método de regresión lineal por la simpleza de su implementación y la compatibilidad con el análisis por unidades.

Para la creación del método base se tomó como referencia el código que lleva a cabo la síntesis por medio de Unit Selection, es decir, la regresión lineal sigue trabajando con unidades de entrenamiento y de test, con particionado K-fold. Para la implementación en Python se hizo uso de la librería *LinearRegression* de *sklearn*. La obtención de resultados por medio de regresión lineal se divide en dos secciones.

1. **Creación del modelo:** El objetivo es ajustar un modelo lineal con unos coeficientes determinados a los datos de entrenamiento, de forma que se minimice la suma cuadrática residual entre las unidades de la base de datos y las predichas por la aproximación lineal. Una vez se tiene la base de datos de entrenamiento, se crea el modelo de ajuste lineal. Las fórmulas que rigen la creación del modelo lineal se presentan en 2.7, 2.8, 2.9.

2. **Predicción de las unidades MFCC:** La biblioteca usada permite realizar predicciones en base al modelo creado, por lo que para obtener el resultado de las unidades de entrenamiento, simplemente se hace uso del método *predict* de la biblioteca, que devuelve una predicción las unidades de parámetros MFCC haciendo uso del modelo creado. De igual manera que con el algoritmo de Unit Selection, se hace uso del VoCoder WORLD para sintetizar audio a partir de los resultados obtenidos.

## 4.3. Resultados para Regresión Lineal y Unit-Selection

A continuación, los resultados de los algoritmos diseñados se reflejan para los distintos sets de datos con los que se trabaja en el estudio.

En referencia al método base de Regresión lineal, por la naturaleza del código implementado para síntesis de voz sólo cuenta con un único parámetro que se puede modificar: la longitud de la unidad. Por lo tanto, se mostrarán los resultados para todos los datasets sintetizados por medio de regresión lineal modificando la longitud de unidad. De igual manera que para los resultados anteriores, se mostrará la comparación de audios y espectrogramas.

Como se ha mencionado en el capítulo 3, para Unit Selection contamos con 3 parámetros que pueden afectar al rendimiento del sistema:

- Longitud de la Unidad
- Peso de Concatenación
- N°de Vecinos de *BallTree*

Para evaluar los resultados, se ha ejecutado el algoritmo diseñado para cada uno de los datasets y cada una de las configuraciones posibles. Se cuenta con 4 datasets y 3 parámetros por datasets. Para disponer de una información más completa, se incluyen comparaciones entre audios originales y sintetizados, haciendo uso de una representación temporal y de espectrograma. Previo al estudio de la modificación de los parámetros se comprobó el funcionamiento con todos los datasets. Por medio de una escucha subjetiva se determinó que el algoritmo de Unit-Selection era capaz de sintetizar voz inteligible para todos los datasets.

En cada una de las tablas de resultados, la configuración que de mejor resultado aparecerá resaltada en negrita. En el apartado de discusión 4.4 se estudiarán los resultados expuestos a continuación, para poder realizar un análisis en conjunto.

### 4.3.1. TiDigit - LC

A continuación se muestran los resultados para el método base de Regresión Lineal.

**Longitud de Unidad** = VARIABLE

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 10.950 | **10.909** | 10.937 | 11.034 |
| STOI | **0.529** | 0.517 | 0.514 | 0.502 |

El mejor rendimiento se da para una longitud de 0.04 para MCD y 0.02 para la el STOI.

La siguiente figura 4.2 ilustra una comparación entre una señal original y sintetizada por medio de Regresión Lineal. Se incluye una representación temporal y espectrograma.

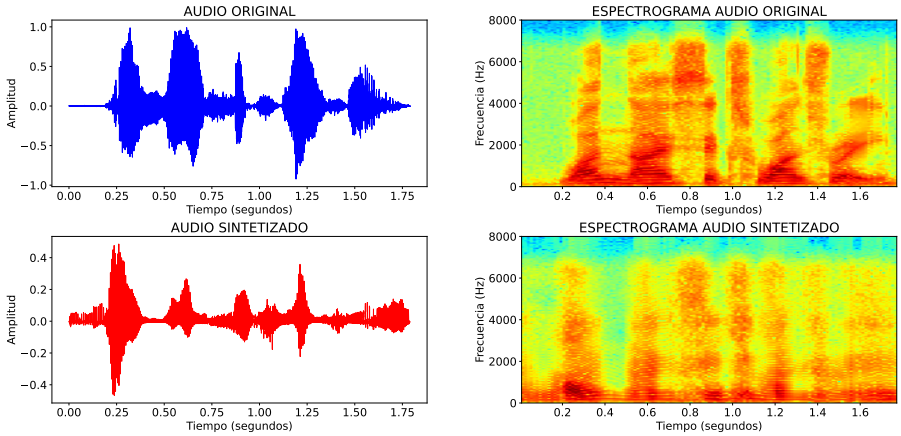


Figura 4.2: Comparación entre señal original y sintetizada para Regresión Lineal. TiDigit - LC

La figura 4.2 contiene señal original y sintetizada para la sucesión de dígitos ingleses 'One Nine Six One Three' (19613). En la figura se pueden comprobar las limitaciones del método de Regresión Lineal. Se puede comprobar en la representación temporal, así como en el espectrograma, que en la señal sintetizada hay ruido de baja frecuencia a lo largo de todo el audio. De igual manera, en el espectrograma se puede comprobar una clara falta de detalle en las zonas en las que se pronuncian los dígitos (aparecen difusas). Estos dos fenómenos se deben a las limitaciones del propio método de regresión lineal.

A continuación se muestran los resultados para el algoritmo de Unit Selection, en los que se incluye la modificación individual de cada parámetro.

Para longitud de unidad variable:

**Longitud de Unidad** = VARIABLE ; **Peso** = 0.1 ; **N°Vecinos** = 10

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 11.389 | 11.363 | **11.032** | 10.819 |
| STOI | 0.494 | 0.517 | **0.521** | 0.506 |

El mejor rendimiento se da para una longitud de unidad de 0.08 segundos.

Para peso de concatenación variable:

**Longitud de Unidad** = 0.08 ; **Peso** = VARIABLE ; **N°Vecinos** = 10

| Peso de concatenación | 0 | 0.01 | 0.1 | 0.5 | 1 | 2 |
|---|---|---|---|---|---|---|
| MCD | 11.195 | 11.111 | **10.819** | 10.912 | 10.938 | 10.945 |
| STOI | **0.572** | 0.566 | 0.521 | 0.504 | 0.500 | 0.491 |

Para MCD, el mejor resultado lo da el peso 0.1, mientras que la mejor inteligibilidad la da el peso nulo.

Para N°de vecinos variable:

**Longitud de Unidad** = 0.08 ; **Peso** = 0.1 ; **N°Vecinos**= VARIABLE

| N°de Vecinos | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| MCD | 11.195 | 10.845 | **10.819** | 11.134 |
| STOI | **0.572** | 0.549 | 0.521 | 0.471 |

Para la distorsión, el mejor número de vecinos es 10, mientras que para inteligibilidad es 1 único vecino. 1 vecino equivale a utilizar un peso 0 (como hemos obtenido anteriormente).

La siguiente figura 4.3 muestra sendas representaciones de un audio original y su homólogo sintetizado por Unit Selection.

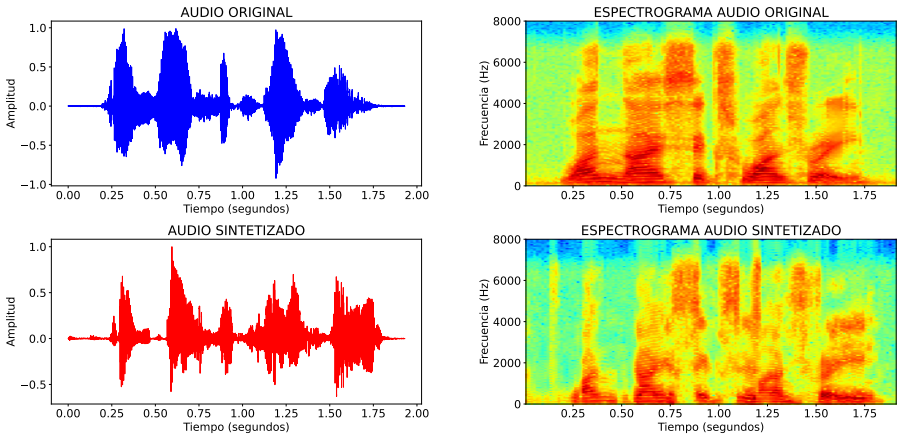


Figura 4.3: Comparación entre señal original y sintetizada para Unit Selection. TiDigit - LC

La figura 4.3 muestra la comparación de señales original y sintetizada para la secuencia de dígitos ingleses 'One Nine Six One Three' (19613) . La figura muestra las ventajas de Unit Selection Frente a Regresión Lineal: En el espectrograma se puede comprobar como los dígitos se encuentran diferenciados y no están difusos. El ruido de baja frecuencia que aparecía hasta en las zonas de silencio no tiene lugar usando el método de Unit Selection. Por lo general, la representación temporal y espectrograma nos muestra que el algoritmo funciona mejor que el método base.

### 4.3.2. TiDigit - TP

Los resultados para Regresión Lineal:

**Longitud de Unidad** = VARIABLE

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 10.968 | 10.816 | **10.712** | 10.756 |
| STOI | 0.562 | 0.559 | **0.566** | 0.554 |

El mejor resultado se da para una longitud de unidad de 0.08 segundos.

A continuación se muestran los resultados para el algoritmo de Unit-Selection.

Para longitud de unidad variable:

**Longitud de Unidad** = VARIABLE ; **Peso** = 0.1 ; **N°Vecinos**=10

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 11.389 | 11.363 | **11.032** | 10.819 |
| STOI | 0.490 | 0.505 | **0.517** | 0.520 |

De igual manera, la longitud 0.08 segundos es la que mejor resultado da.

Para peso de concatenación variable:

**Longitud de Unidad** = 0.08 ; **Peso** = VARIABLE ; **N°Vecinos** = 10

| Peso de concatenación | 0 | 0.01 | 0.1 | 0.5 | 1 | 2 |
|---|---|---|---|---|---|---|
| MCD | 11.546 | 11.467 | 11.318 | **11.309** | 11.338 | 11.359 |
| STOI | **0.566** | 0.542 | 0.521 | 0.504 | 0.500 | 0.492 |

La mejor distorsión la da el peso 0.5, mientras que la mayor inteligibilidad la da el peso 0.01.

Para N°de vecinos variable:

**Longitud de Unidad** = 0.08 ; **Peso** = 0.1 ; **N°Vecinos**= VARIABLE

| N°de Vecinos | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| MCD | 11.546 | **11.237** | 11.318 | 11.526 |
| STOI | **0.542** | 0.538 | 0.517 | 0.496 |

5 vecinos es el mejor valor en términos de distorsión y un único vecino para la mejor inteligibilidad.

### 4.3.3. Arctic - RM

Los resultados para Regresión Lineal:

**Longitud de Unidad** = VARIABLE

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 11.175 | 11.145 | **11.116** | 11.123 |
| STOI | 0.468 | 0.477 | 0.481 | **0.487** |

Para 0.08 segundos se da la mejor métrica de distorsión, mientras que la mejor inteligibilidad la da la longitud 0.16 segundos.

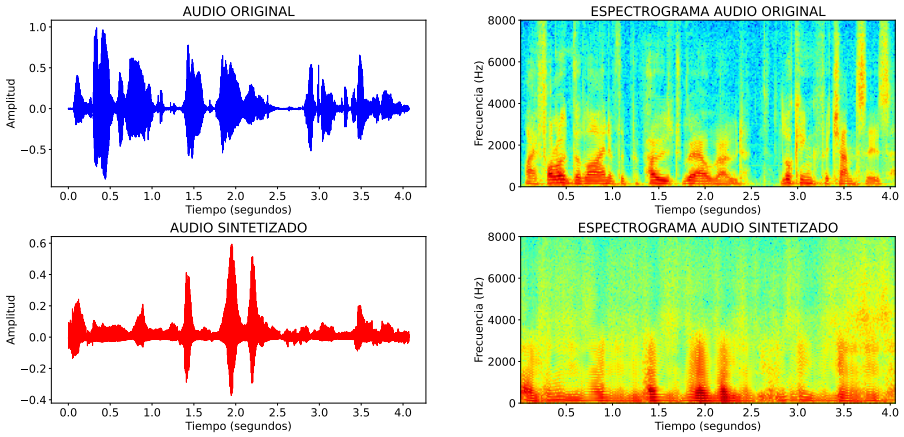


Figura 4.4: Comparación entre señal original y sintetizada para Regresión Lineal. Arctic - RM

En la figura 4.4 podemos comprobar las diferencias entre audio original y sintetizado para una oración completa: 'I followed the line of the proposed railroad, looking for chances'. Se observan los mismos problemas que para la síntesis de dígitos. Las zonas que corresponden a palabras en el espectrograma aparecen difusas, también se aprecia un ruido a lo largo de todo el audio. Los problemas de inteligibilidad de la Regresión Lineal se verán acentuados para una situación más compleja como la síntesis de voz para oraciones completas.

A continuación se muestran los resultados para Unit-Selection.

Para longitud de unidad variable:

**Longitud de Unidad** = VARIABLE ; **Peso** = 0.1 ; **N°Vecinos**=10

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 12.255 | 12.057 | **12.060** | 12.638 |
| STOI | 0.396 | 0.404 | **0.405** | 0.373 |

El mejor rendimiento total se da de nuevo para la longitud de unidad 0.08 segundos.

Para peso de concatenación variable:

**Longitud de Unidad** = 0.08 ; **Peso** = VARIABLE ; **N°Vecinos** = 10

| Peso de concatenación | 0 | 0.01 | 0.1 | 0.5 | 1 | 2 |
|---|---|---|---|---|---|---|
| MCD | 12.640 | 12.447 | **12.068** | 12.130 | 12.142 | 12.144 |
| STOI | 0.413 | **0.414** | 0.404 | 0.389 | 0.387 | 0.387 |

La mejor distorsión la da el peso 0.1, mientras que la mejor inteligibilidad la da el peso 0.01.

Para N°de vecinos variable:

**Longitud de Unidad** = 0.08 ; **Peso** = 0.1 ; **N°Vecinos**= VARIABLE

| N°de Vecinos | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| MCD | 12.640 | 12.127 | **12.068** | 12.193 |
| STOI | 0.413 | **0.414** | 0.405 | 0.384 |

10 vecinos da el mejor resultado en términos de distorsión, mientras que un 5 vecinos da el mejor rendimiento para inteligibilidad.

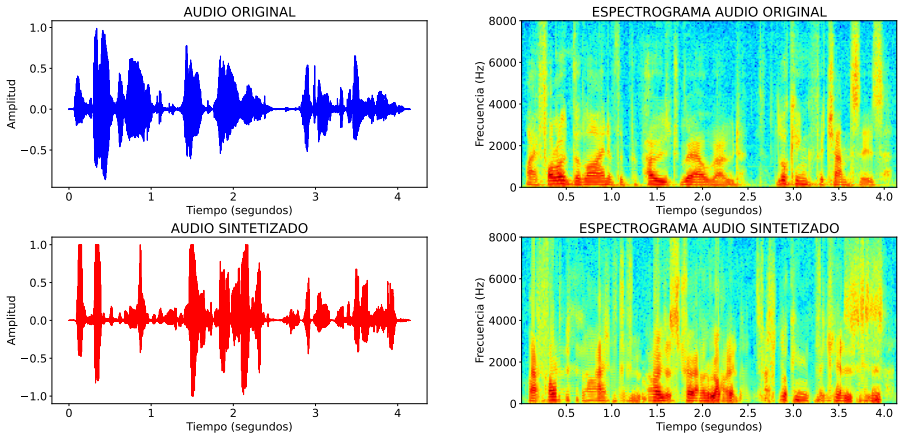


Figura 4.5: Comparación entre señal original y sintetizada para Unit Selection. Arctic - RM

En la figura 4.5 podemos comprobar el rendimiento del algoritmo de Unit Selection para la síntesis de voz de oraciones completas. Se observan las diferencias entre audio original y sintetizado para una oración completa: 'I followed the line of the proposed railroad, looking for chances'. En la representación temporal podemos comprobar que hay correspondencia entre la señal original y la sintetizada, así como en el espectrograma, donde comprobamos que los fonemas se encuentran mucho más definidos.

### 4.3.4. Arctic - JG

Los resultados para Regresión Lineal:

**Longitud de Unidad** = VARIABLE

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 11.216 | **11.188** | 11.142 | 11.143 |
| STOI | 0.501 | 0.505 | 0.506 | **0.511** |

0.04 segundos de longitud de unidad da la mejor distorsión, 0.16 da la mejor inteligibilidad.

A continuación se muestran los resultados para Unit-Selection.

Para longitud de unidad variable:

**Longitud de Unidad** = VARIABLE ; **Peso** = 0.1 ; **N°Vecinos**=10

| Longitud de Unidad (s) | 0.02 | 0.04 | 0.08 | 0.16 |
|---|---|---|---|---|
| MCD | 12.408 | 12.142 | **12.070** | 12.382 |
| STOI | 0.408 | 0.430 | **0.433** | 0.421 |

De nuevo, 0.08 segundos es la longitud de unidad óptima.

Para peso de concatenación variable:

**Longitud de Unidad** = 0.08 ; **Peso** = VARIABLE ; **N°Vecinos** = 10

| Peso de concatenación | 0 | 0.01 | 0.1 | 0.5 | 1 | 2 |
|---|---|---|---|---|---|---|
| MCD | 12.428 | 12.299 | **12.070** | 12.148 | 12.161 | 12.1745 |
| STOI | **0.455** | 0.453 | 0.433 | 0.420 | 0.420 | 0.418 |

El peso óptimo es 0.1 para distorsión y 0 para inteligibilidad.

Para N°de vecinos variable:

**Longitud de Unidad** = 0.08 ; **Peso** = 0.1 ; **N°Vecinos**= VARIABLE

| N°de Vecinos | 1 | 5 | 10 | 20 |
|---|---|---|---|---|
| MCD | 12.428 | **12.018** | 12.069 | 12.266 |
| STOI | **0.455** | 0.451 | 0.433 | 0.415 |

5 vecinos da la mejor distorsión y un único vecino da la mejor inteligibilidad.

## 4.4. Discusión

El objetivo de esta sección va a ser el de discutir los resultados obtenidos en cuanto a métricas y escuchas subjetivas.

### 4.4.1. Evaluación subjetiva

Con el fin de evaluar de forma auditiva la calidad de las señales de voz sintetizadas, se dispusieron escuchas informales de los distintos audios sintetizados. En lo que respecta a el objetivo principal del proyecto, este se ha cumplimentado de forma satisfactoria. La síntesis de voz por medio de señales PMA haciendo uso de la versión del algoritmo de Unit-Selection implementada produce voz inteligible y de una calidad aceptable. La evaluación determina que los audios sintetizados haciendo uso del algoritmo de Unit Selection tienen un rendimiento notablemente superior a los sintetizados por medio del método base de Regresión Lineal. La inteligibilidad del algoritmo implementado también es superior a la del método base, siendo bastante pobre para Regresión Lineal, en la que la identificación de los dígitos pronunciados es complicada, en ocasiones imposible. En este aspecto, la diferencia de los audios sintetizados por medio de Unit Selection y Regresión Lineal se hace evidente, siendo el primero mucho más efectivo a la hora de generar audios inteligibles.

Llevando a cabo una escucha subjetiva de los diferentes resultados para cada dataset, se llega a la conclusión de que la inteligibilidad que se consigue para la base de datos *TiDigit* es superior a la de *Arctic*. Esto es esperable, ya que la síntesis de oraciones completas es un proceso más complejo, que además requiere de bases de datos más grandes. Se debe tener en cuenta también la variabilidad fonética de las oraciones de la base de datos de *Arctic*, en la que el número de fonemas que se pronuncian es mucho mayor. Otro factor, identificado en otros trabajos como [32], es el de las limitaciones de la propia técnica de PMA. En [32] se concluye que existen fonemas que la técnica de PMA no es capaz de capturar de manera fiel, por lo que de facto se imposibilita su correcta síntesis sin contar con herramientas de captura adicionales o con información extra del contexto. El alcance de este trabajo no suficientemente específico como para analizar qué fonemas no son capturados de manera fidedigna. La inteligibilidad para las oraciones sintetizadas de la base de datos de *Arctic* varía en función de la oración que se esté pronunciando, mientras que para *TiDigit* la inteligibilidad era total prácticamente para todos los ficheros.

### 4.4.2. Evaluación con métricas objetivas

En lo que se refiere a síntesis de voz a partir de bioseñales, el algoritmo de Unit-Selection presenta una opción más potente que la regresión lineal, esto

se debe a que es un método no paramétrico que, al contrario que la regresión lineal, no distorsiona la voz. Esta es sintetizada con porciones inalteradas de la voz de la base de datos. Como se comentó en 3.3, la relación entre las bioseñales PMA y la voz suelen no seguir una lógica lineal, por lo que es una ventaja adicional de la técnica de Unit Selection sobre Regresión Lineal. Unit Selection también constituye un método flexible ya que puede ser utilizado con multitud de bioseñales, por ejemplo EEG [38]. En teoría, también sería posible con Regresión Lineal, pero la naturaleza no lineal es más acusada para una casuística compleja como la de síntesis EEG-voz.

Si bien las métricas no hacen visible una diferencia de rendimiento del algoritmo de Unit Selection frente a regresión lineal, las diferencias son apreciables al revisar las diferencias de las figuras. Por ejemplo, los resultados para Unit Selection 4.3 4.5 muestran una representación temporal limpia, las zonas de silencio y de habla están claramente diferenciadas y, por lo general, se asemejan claramente al audio original. La escucha subjetiva, como se ha comentado previamente, deja claro la inteligibilidad de los audios. Por su parte, las figuras obtenidas para regresión lineal 4.2 4.4 muestran una representación temporal mucho más 'sucia' en el sentido de que se aprecia audio en las zonas en las que en el audio original hay silencio. En el espectrograma se observa una línea de potencia en la totalidad de la duración de todos los audios en la zona de frecuencias más bajas, esto indica la presencia de ruido. Adicionalmente, las zonas en las que se pronuncian los fonemas y palabras aparecen mucho más difusas, lo que indica la baja calidad del resultado.

En lo que se refiere al rendimiento del algoritmo de Unit Selection conforme se modifican los parámetros expuestos, obtenemos lo siguiente:

- **Longitud de Unidad Variable:** Para los distintos datasets, la longitud de unidad que ha reportado mejores resultados ha sido 0.08 segundos. En principio, una longitud de unidad más grande debe reportar mejores resultados, ya que incluye mayor contexto temporal. En el estudio realizado se ha observado que esto es así hasta cierto punto, ya que las métricas empeoran al usar longitudes de unidad superiores a la óptima. Esto se debe, probablemente, a que una unidad de gran longitud obtiene demasiado contexto temporal (se puede estar tomando información del comienzo de otro sonido o de un silencio). Esto hace errónea la evaluación de la unidad con las de la base de datos y puede dar como óptima una unidad que termina dando peores resultados. Por último, cabe destacar que la variación en la longitud de la unidad tiene un efecto significativo en el tiempo de cómputo del algoritmo, sobre todo para la creación de la base de datos, aunque también para la síntesis. Esto se debe tener en cuenta para poder tener un equilibrio entre calidad y coste computacional.

- **Peso del coste de concatenación:** Para los pesos variables, en todos los casos estudiados se ha observado que el mejor rendimiento se

da para un peso de 0.1. Tras una depuración del código, comprobando qué valores adoptaban los costes objetivo y de concatenación, se comprobó que los costes de concatenación tomaban el valor más 'adecuado' cuando se les aplicaba un peso de 0.1. El objetivo del coste de concatenación es que modifique ligeramente el valor del coste objetivo, para incentivar aquellas unidades que tienen una transición más suave y desincentivar las que no. Para ello, lo ideal es que el coste de concatenación adopte valores que se encuentren en el mismo orden de magnitud que el coste objetivo, pero que sean ligeramente menores. Si estos valores son demasiado bajos en comparación con el coste objetivo, se pierde el efecto. Es por eso que se habla del valor más 'adecuado'. Los resultados obtenidos respaldan esto, al ser 0.1 el peso óptimo para todos los datasets.

- **N°de vecinos variable:** Para la modificación de este parámetro hay que tener en consideración el coste computacional de cada opción, ya que a menor n°de vecinos, más rápida será la síntesis. Para la modificación del número de vecinos, se obtuvo que el número óptimo se encontraba entre 5 y 10, dependiendo del dataset, para los mejores resultados de MCD. No obstante, según la métrica STOI, que hace una evaluación objetiva de la inteligibilidad de la voz sintetizada, el mejor resultado se da para un único vecino en todos los datasets (a excepción del dataset Arctic - RM). Esto puede ser indicativo a un funcionamiento errático del coste de concatenación en ocasiones, quizá por tener una influencia excesiva. La influencia del coste de concatenación varía en función del dataset concreto ya que las distancias, aún estando en rangos similares, varían. En este caso, la solución pasaría por tener un menor coste de concatenación, pero como se ha comentado, la variación debería de personalizarse a cada casuística específica, lo que sería prohibitivo en términos de tiempo invertido. Cabe esperar que para un número de vecinos altos las métricas se estabilice. Sin embargo, se observa cierta variación entre los números más altos (10 y 20 vecinos). Esto se puede deber a una influencia de el coste de concatenación (quizá al tener un valor algo superior a lo ideal) que haga que se favorezcan unidades cuyo coste objetivo no sea el más óptimo.

Por lo general, los valores obtenidos para MCD son altos en comparación con la literatura [26] [31], en los que este valor se encuentra entre 4.4 y 6 dB. Una posible causa de esta diferencia reside en que el VoCoder que se ha utilizado en la literatura mencionada es el STRAIGHT, mientras en este trabajo se ha usado el VoCoder WORLD. El funcionamiento de ambos es distinto y ha podido afectar al rendimiento del sistema de síntesis. Paralelamente, se ha observado que la modificación de los parámetros no reporta variaciones significativas en los resultados obtenidos por métricas objetivas

(tampoco para escuchas objetivas). Adicionalmente, la diferencias observadas en las métricas objetivas entre datasets, así como frente al método base son mínimas. Se concluye por tanto que las métricas objetivas utilizadas en este trabajo no son una fuente de información definitiva en lo que se refiere al rendimiento del algoritmo, ya que en las escuchas subjetivas (así como en imágenes y espectrogramas) las diferencias son visibles.

# Capítulo 5

# Conclusiones y líneas futuras

El foco de este trabajo ha sido la implementación de un algoritmo de síntesis de voz por medio de bioseñales. Para cumplimentar los objetivos, se implementó un algoritmo de Unit Selection, de manera que tomase la información de una base de datos con bioseñales PMA y registros de voz para dígitos y oraciones. A partir de esta base de datos se realizaba la síntesis de voz. Los resultados han demotrado que es posible sintetizar voz inteligible a partir de las bioseñales PMA haciendo uso del algoritmo de Unit Selection.

En cuanto a las métricas asociadas a los resultados obtenidos, en términos de distorsión cepstral (MCD) se encuentran entre 10.8 dB y 12.4 dB para todos los datasets, mientras que para inteligibilidad (STOI) se encuentran entre 0.4 y 0.57. Escuchas objetivas determinan que el algoritmo creado con este propósito tiene una inteligibilidad superior al método base implementado de regresión lineal.

Tras el trabajo realizado en este proyecto, la principal conclusión a la que se llega es que es posible sintetizar voz por medio de bioseñales PMA haciendo uso de la versión del algoritmo de Unit-Selection implementada en el proyecto. La información que aportan las señales PMA está lo suficientemente correlada con la voz como para obtener una voz inteligible y de calidad aceptable.

La razón por la que se ha elegido esta técnica frente a otras se debe principalmente a que es una técnica no paramétrica, por lo que permite llegar a resultados (síntesis de voz) aceptables con bases de datos de tamaños manejables, lo que hace posible su implementación y experimentación con un hardware común. Adicionalmente, la abundancia en la literatura de implementaciones de síntesis de voz que hacen uso de Unit-Selection fue un factor a tener en cuenta.

En este proyecto se han implementado dos métodos distintos para sintetizar voz a partir de bioseñales PMA, obteniendo con ambas el objetivo fijado. La implementación de un método base de funcionamiento simple ha

servido como base para la implementación y depuración del algoritmo principal, teniendo la oportunidad de comprobar de primera mano el rendimiento que se obtiene al sintetizar voz por medio de un algoritmo más simple en las mismas condiciones y para los mismos datasets.

Para el algoritmo principal de Unit-Selection se implementó una versión ligeramente distinta de lo que se puede encontrar en la literatura, debido principalmente a que la síntesis de voz no era totalmente directa (existía el paso intermedio de los MFCC's) y el cálculo del coste de concatenación se realizaba en base a estos parámetros MFCC's. Se añade por tanto este factor diferencial al trabajo realizado, en el que se ha podido comprobar la posibilidad de sintetizar voz por medio de esta variante del algoritmo.

En cuanto a los resultados obtenidos, se obtuvo voz sintética para todos los datasets de los que se disponía. Si bien las métricas utilizadas (MCD sobre todo) no justifican los resultados obtenidos, sobre todo en cuanto a la mejora con respecto al algoritmo base. Adicionalmente, la variación de los distintos parámetros no ha resultado en alteraciones significativas de los resultados obtenidos. Más allá de los resultados obtenidos en las métricas, las escuchas objetivas demuestran que el audio sintetizado es inteligible y que el algoritmo puede ser válido para ser usado en sistemas de síntesis de voz para bioseñales PMA.

En cuanto a las líneas futuras que se pueden tomar, cabe destacar principalmente que una vía de desarrollo para la investigación y la implementación de algoritmos podría ser la de métodos no lineales, especialmente DNN's. Los resultados observados en la amplia literatura al respecto indican que el desarrollo de este tipo de algoritmos puede ser prometedor en cuanto a las oportunidades que abre para síntesis de voz por su versatilidad y potencia. Todo esto estaría sujeto, por supuesto, a la disponibilidad de capacidad computacional suficiente y a bases de datos crecientemente grandes.

Adicionalmente, una línea de estudio de interés consistiría en la combinación de la aproximación de Unit Selection con una DNN. Este tipo de combinaciones híbridas tratan de suplir desventajas de Unit Selection (como transiciones abruptas en ciertas casuísticas) manteniendo los puntos fuertes de la misma. Estudios al respecto confirman que la síntesis de voz es posible haciendo uso de esta aproximación [27].

Otro campo de estudio de sumo interés sería el de síntesis de voz a partir de bioseñales cerebrales, sobre todo por medio de DNN's. La síntesis por medio de señales cerebrales implica que pueden ser utilizadas en pacientes de todo tipo de casuísticas, incluso en aquellos que no tengan ningún tipo de movilidad articular, por lo que podría ser utilizado en virtualmente todas las situaciones imaginables. Sin embargo, los retos aparejados a esta rama también son significativos, principalmente la calidad de las bioseñales en sí. Muchas veces la obtención de bioseñales de este tipo depende de criterios médicos (no técnicos) por lo que en numerosas situaciones se tiene una flexibilidad nula en lo que a obtención de la bioseñal se refiere. Adicionalmente,

el campo de estudio que relaciona las señales cerebrales con la producción de voz es una rama del conocimiento que todavía no tiene una respuesta sólida a cómo se pueden obtener bioseñales altamente correladas con la voz, existiendo una variabilidad muy alta inter-paciente.

# Apéndice A

# Temporización y Presupuesto

## A.1. Temporización

Para mostrar la temporización del proyecto se ha creado un diagrama de Gantt, reflejado en la siguiente figura.

| | MES | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Tarea** | **Septiembre** | **Octubre** | **Noviembre** | **Diciembre** | **Enero** | **Febrero** | **Marzo** | **Abril** | **Mayo** | **Junio** | **Julio** |
| Documentación | | | | | | | | | | | |
| Obtención de unidades del dataset | | | | | | | | | | | |
| Implementación de Unit-Selection | | | | | | | | | | | |
| Implementación de Regresión Lineal | | | | | | | | | | | |
| Modificaciones y mejoras a algoritmos | | | | | | | | | | | |
| Memoria | | | | | | | | | | | |

Figura A.1: Temporización del trabajo realizado

Los primeros meses de trabajo se dedicaron prácticamente en exclusiva a la documentación bibliográfica: estudio del fenómeno de la voz, fenómenos fisiológicos relacionados con la misma y obtención de bioseñales, así como las distintas aproximaciones que existían para abordar el problema de la síntesis de voz por medio de bioseñales. Una vez se había establecido el algoritmo que se quería implementar y qué tipo de bioseñales iban a servir el propósito, los siguientes meses se dedicaron a la implementación completa del mismo, así como a la comprobación e implementación de mejoras pertinentes. Por último, en los últimos meses de trabajo se obtuvieron las métricas y se redactó la memoria al completo.

## A.2. Presupuesto del proyecto

En este apéndice se va a indicar el presupuesto estimado para la realización del trabajo. Como se indica en A.1, en total, son 10 meses completos de trabajo, con la adición de una porción del último mes (julio). Se estima que el tiempo total invertido en la realización del proyecto asciende a 350 horas (Aproximadamente un 20 % del tiempo invertido en el trabajo). Un ingeniero trabajando a tiempo completo trabaja 160 horas mensuales, por lo que el número de meses de un ingeniero de telecomunicaciones que tomaría el proyecto:

$$N°meses = \frac{Horas\_trabajadas}{Horas\_mensuales\_ingeniero} = \frac{350h}{160h} = 2.18meses \quad \text{(A.1)}$$

Tomando como un sueldo común para un ingeniero de telecomunicaciones junior la cifra de 1500€, y teniendo en cuenta que éste habría sido el único coste monetario asociado (todo el software y material utilizado es gratuito), Obtenemos lo siguiente.

| Tipo de puesto | Meses | Sueldo/Mes (€) | Coste total (€) |
|---|---|---|---|
| Ingeniero de Telecomunicaciones Junior | 2.18 | 1500 | 3270 |

A parte del tiempo invertido en el proyecto, la realización del proyecto se ha llevado a cabo en un ordenador portátil con un coste aproximado de unos 1000€, así como en un ordenador de sobremesa con un coste aproximado de 800€. El software utilizado para la implementación del algoritmo es *Python*, sin coste asociado. Adicionalmente, se han consultado todos los artículos que aparecen reflejados en la bibliografía, así como las dos bases de datos de bioseñales/voz indicadas a lo largo de la memoria, sin coste asociado para la obtención de ninguno de los recursos.

# Bibliografía


[1] Aubrey J Yates. "Delayed auditory feedback." En: *Psychological bulletin* 60.3 (1963), pág. 213.

[2] Bishnu S Atal y Suzanne L Hanauer. "Speech analysis and synthesis by linear prediction of the speech wave". En: *The journal of the acoustical society of America* 50.2B (1971), págs. 637-655.

[3] Elaine Smith et al. "Effect of voice disorders on quality of life". En: *Journal of Medical Speech-Language Pathology* 4.4 (1996), págs. 223-244.

[4] William J Hardcastle y Fiona Gibbon. "Electropalatography and its clinical applications". En: *Instrumental clinical phonetics* (1997), págs. 149-193.

[5] Songun Na y Seungwha Yoo. "Allowable Propagation Delay for VoIP Calls of Acceptable Quality". En: ago. de 2002, págs. 47-56. ISBN: 978-3-540-43968-4. DOI: 10.1007/3-540-45639-2_6.

[6] Antti Karjalainen Didier Dupré. "Employment of disabled people in Europe in 2002". En: *Eurostat* 1.1 (2003), págs. 1-4.

[7] Mikhail Lebedev y Miguel Nicolelis. "Brain-Machine Interfaces: Past, Present and Future". En: *Trends Neurosci.* 29 (oct. de 2006), págs. 536-. DOI: 10.1016/j.tins.2006.07.004.

[8] Stuart N Baker. "Oscillatory interactions between sensorimotor cortex and the periphery". En: *Current opinion in neurobiology* 17.6 (2007), págs. 649-655.

[9] Jochen Baumeister et al. "Influence of phosphatidylserine on cognitive performance and cortical activity after induced stress". En: *Nutritional neuroscience* 11 (jun. de 2008), págs. 103-10. DOI: 10.1179/147683008X301478.

[10] Paul Taylor. "The text-to-speech problem". En: (2009), págs. 26-51. DOI: 10.1017/CBO9780511816338.005.

[11] Jonathan Brumberg et al. "Brain-Computer Interfaces for Speech Communication". En: *Speech communication* 52 (abr. de 2010), págs. 367-379. DOI: 10.1016/j.specom.2010.01.001.

[12] Bruce Denby et al. "Silent speech interfaces". En: *Speech Communication* 52.4 (2010), págs. 270-287.

[13] Thomas Hueber et al. "Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips". En: *Speech Communication* 52 (abr. de 2010), págs. 288-300. DOI: 10.1016/j.specom.2009.11.004.

[14] Cees Taal et al. "A short-time objective intelligibility measure for time-frequency weighted noisy speech". En: abr. de 2010, págs. 4214-4217. DOI: 10.1109/ICASSP.2010.5495701.

[15] Robin Hofe et al. "Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA." En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (ago. de 2011), págs. 3009-3012. DOI: 10.21437/Interspeech.2011-753.

[16] World Health Organization et al. "World report on disability." En: *World report on disability.* (2011).

[17] Laura Marzetti et al. "Magnetoencephalographic alpha band connectivity reveals differential Default Mode Network interactions during focused attention and open monitoring meditation". En: *Frontiers in Human Neuroscience* 8 (sep. de 2014). DOI: 10.3389/fnhum.2014.00832.

[18] Masanori Morise. "CheapTrick, a spectral envelope estimator for high-quality speech synthesis". En: *Speech Communication* 67 (ene. de 2014). DOI: 10.1016/j.specom.2014.09.003.

[19] Chris Neufeld y Pascal van Lieshout. "Tongue kinematics in palate relative coordinate spaces for electro-magnetic articulography". En: *The Journal of the Acoustical Society of America* 135.1 (2014), págs. 352-361.

[20] M. Zahner et al. "Conversion from facial myoelectric signals to speech: A unit selection approach". En: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (ene. de 2014), págs. 1184-1188.

[21] Heiga Zen. "Statistical Parametric Speech Synthesis". En: (2014). Tutorial.

[22] Sandesh Aryal y Ricardo Gutierrez-Osuna. "Data driven articulatory synthesis with deep neural networks". En: *Computer Speech Language* 36 (mar. de 2015). DOI: 10.1016/j.csl.2015.02.003.

[23] Lorenz Diener, Matthias Janke y Tanja Schultz. "Direct conversion from facial myoelectric signals to speech using Deep Neural Networks". En: (jul. de 2015), págs. 1-7. DOI: 10.1109/IJCNN.2015.7280404.

[24] Mohamad Dolatshah, Ali Hadian y Behrouz Minaei-Bidgoli. "Ball*-tree: Efficient spatial indexing for constrained nearest-neighbor search in metric spaces". En: *arXiv preprint arXiv:1511.00628* (2015).

[25] Marc Arnela et al. "Influence of lips on the production of vowels based on finite element simulations and experiments". En: *The Journal of the Acoustical Society of America* 139 (mayo de 2016), págs. 2852-2859. DOI: 10.1121/1.4950698.

[26] Jose Gonzalez Lopez et al. "A Silent Speech System based on Permanent Magnet Articulography and Direct Synthesis". En: *Computer Speech Language* 39 (mar. de 2016). DOI: 10.1016/j.csl.2016.02.002.

[27] Thomas Merritt et al. "Deep neural network-guided unit selection synthesis". En: (mar. de 2016), págs. 5145-5149. DOI: 10.1109/ICASSP.2016.7472658.

[28] Masanori MORISE, Fumiya YOKOMORI y Kenji Ozawa. "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications". En: *IEICE Transactions on Information and Systems* E99.D (jul. de 2016), págs. 1877-1884. DOI: 10.1587/transinf.2015EDP7457.

[29] R.B. Randall. "A History of Cepstrum Analysis and its Application to Mechanical Problems". En: *Mechanical Systems and Signal Processing* 97 (dic. de 2016). DOI: 10.1016/j.ymssp.2016.12.026.

[30] Sri Harsha Dumpala y K N R K Alluri. "An Algorithm for Detection of Breath Sounds in Spontaneous Speech with Application to Speaker Recognition". En: (ago. de 2017), págs. 98-108. DOI: 10.1007/978-3-319-66429-3_9.

[31] Jose Gonzalez Lopez et al. "Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning". En: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (dic. de 2017), págs. 2362-2374. DOI: 10.1109/TASLP.2017.2757263.

[32] Jose Gonzalez Lopez et al. "Evaluation of a Silent Speech Interface Based on Magnetic Sensing and Deep Learning for a Phonetically Rich Vocabulary". En: (ago. de 2017), págs. 3986-3990. DOI: 10.21437/Interspeech.2017-802.

[33] Myungjong Kim et al. "Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network". En: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25 (dic. de 2017), págs. 2323-2336. DOI: 10.1109/TASLP.2017.2758999.

[34] Conor Ransome y C Ransome. "The Fermi Paradox and Galactic Habitability (Masters thesis)". Tesis doct. Abr. de 2017.

[35] Sandra Vieira, Walter Pinaya y Andrea Mechelli. "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications". En: *Neuroscience Biobehavioral Reviews* 74 (ene. de 2017). DOI: 10.1016/j.neubiorev.2017.01.002.

[36] Miguel Angrick et al. "Speech Synthesis from ECoG using Densely Connected 3D Convolutional Neural Networks:" en: (nov. de 2018). DOI: 10.1101/478644.

[37] Gopala Anumanchipalli, Josh Chartier y Edward Chang. "Speech synthesis from neural decoding of spoken sentences". En: *Nature* 568 (abr. de 2019), págs. 493-498. DOI: 10.1038/s41586-019-1119-1.

[38] Christian Herff et al. "Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices". En: *Frontiers in Neuroscience* 13 (nov. de 2019), pág. 1267. DOI: 10.3389/fnins.2019.01267.

[39] Sebastian Nagel. "Towards a home-use BCI: fast asynchronous control and robust non-control state detection". En: (dic. de 2019). DOI: 10.15496/publikation-37739.

[40] Han Gyol Yi, Matthew K Leonard y Edward F Chang. "The encoding of speech sounds in the superior temporal gyrus". En: *Neuron* 102.6 (2019), págs. 1096-1110.

[41] Jose A Gonzalez-Lopez et al. "Silent speech interfaces for speech restoration: A review". En: *IEEE access* 8 (2020), págs. 177995-178021.

[42] Miguel Angrick et al. "Towards closed-loop speech synthesis from stereotactic eeg: a unit selection approach". En: (2022), págs. 1296-1300.

[43] Filip Boltuzic. "Computational methods for argumentation mining of claims in internet discussions". Tesis doct. Jun. de 2022.