Análisis de la relación entre el grado de apertura velofaríngea y propiedades espectrales de la voz mediante el uso de vídeos de resonancia magnética

Moisés Alfonso Guerrero Jiménez, William Gilberto Almeida Mantilla

Trabajo de grado para optar al título de Ingeniero Electrónico

Director

Franklin Alexander Sepúlveda Sepúlveda

Phd Ingeniería Automática

Universidad Industrial De Santander

Facultad De Ingenierías Físico-Mecánicas

Escuela De Ingenierías Eléctrica, Electrónica y De Telecomunicaciones

Bucaramanga

2019

Dedicatoria

A mis padres Norbertina Jiménez y Moisés Guerrero, quienes han estado siempre apoyándome, ayudándome a ser la persona que soy gracias a su esfuerzo y compañía.

A mi hermana Nataly Guerrero, que me ha acompañado en muchos momentos, aconsejándome en muchas situaciones, siendo además de mi hermana, mi mejor amiga, y mi hermano Jesús David Guerrero, que me ha apoyado y enseñado, a darme cuenta de todo el potencial que tengo y que puedo dar.

A mis abuelos Marco Jiménez y Grace Martínez, que me acompañan desde el cielo y que sé que siempre me están acompañando en tantas situaciones.

A mis abuelos Mamá Ana Rosa y Papá Carlos, que quiero mucho y son muy importantes para mí.

Para todos mis primos y tíos que siempre los llevo en el corazón.

Moisés Alfonso Guerrero Jiménez

Dedicatoria

A Dios y a mi familia, por todo.

William Gilberto Almeida Mantilla

Agradecimientos

Doy gracias a Dios por todas las lecciones aprendidas, los conocimientos adquiridos y las personas conocidas durante este tiempo, que han jugado un papel fundamental en mi formación a nivel profesional.

A cada uno de los docentes y tutores que aportaron a mi formación y desarrollo como profesional y como ser humano.

A William Almeida mi compañero y amigo, con el que tuve la oportunidad de trabajar y del cual pude aprender una gran cantidad de cosas.

Al profesor Franklin Sepúlveda que fue mi orientador y del cual me llevo una gran cantidad de conocimiento.

A todos y cada uno de los compañeros y amigos que conocí a lo largo de mi carrera, que han sido fuente de inspiración y aprendizaje.

Moisés Alfonso Guerrero Jiménez

Agradecimientos

A Dios por guiarme en este camino.

A mi maestra, mi madre Socorro Mantilla por su templanza, valor y carácter y quien ha sido la más grande inspiración para incursionar en el mundo académico, a mi padre Gilberto Almeida por enseñarme el valor del trabajo, a mis hermanos Edwar y Wendy Almeida por convertirse en cómplices de este sueño, mi amada familia ha sido el núcleo de este camino con su infinito amor, comprensión y apoyo.

A mi compañero de trabajo, mi gran amigo Moisés Guerrero por su motivación, disciplina, entrega, compromiso, perseverancia y buen sentido del humor nuestro director Franklin Sepúlveda, quien nos inspiró para trabajar en un proyecto que captó nuestro interés desde el primer momento, y fue guía para fortalecer nuestras habilidades en el campo de la investigación con su constante supervisión y su incansable entrega.

A Maria Álvarez por ser fuente de comprensión, risas y alegría, por brindarme su constante apoyo y por encontrar siempre en ella los mejores consejos fruto de sus incontables virtudes.

A mis amigos de Casablanca con quienes viví inolvidables experiencias y he forjado lazos de amistad irrompibles, Carlos Lozano, Diego Corzo, Eduardo Muñoz, Valeria Romero, Maria Paula Guerrero, Yuri Gómez, Henry Velandia y todos los que formaron parte de esa hermosa familia.

A mis amigos y compañeros de carrera, con quienes compartí el sueño de ser ingeniero. A mis profesores, cuya labor de instruir, enseñar, guiar y corregir corresponden a una de las vocaciones más sublimes.

Finalmente, a todos aquellos que me enseñaron a conocerme, aquellos que me ayudaron a crecer personal, profesional y espiritualmente y descubrieron junto a mi lo maravilloso que puede ser un camino cuando vas bien acompañado, para todas esas personas mi gratitud esta con ustedes.

William Gilberto Almeida Mantilla

Contenido

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Lista de Anexos

Resumen

Título: Análisis de la relación entre el grado de apertura velofaríngea y propiedades espectrales de la voz mediante el uso de vídeos de resonancia magnética.*

Autores: Moisés Alfonso Guerrero Jiménez
William Gilberto Almeida Mantilla**

Palabras clave: Velo del paladar, Coeficientes cepstrales en las frecuencias de Mel, Redes neuronales artificiales, Imágenes de resonancia magnética.

DESCRIPCIÓN:

El velo del paladar o paladar blando es un órgano muy importante en la producción del habla; sin embargo, a diferencia de órganos como los labios, la posición de velo dificulta el análisis de su dinámica y su relación con la acústica de la voz. El poder inferir la dinámica del velo del paladar resulta importante en tareas relacionadas con terapias del habla y de aprendizaje de idiomas, entre otras. Aunque existen herramientas tales como el articulógrafo electromagnético y las mismas imágenes de resonancia magnética, estas son muy costosas, y además presentan inconvenientes particulares de tipo práctico a la hora de utilizarlas. Por tanto, se plantea el uso de la misma voz como medio para indagar acerca de la función velofaríngea mediante el uso de métodos de mapeo acústico-articulatorio [1].

La dinámica del velo se obtiene a partir de videos de resonancia magnética en tiempo real (rt–MRI) del tracto vocal, por medio de la base de datos USC-TIMIT [2] [3]. El procedimiento consiste, en primer lugar, en una extracción de la posición del velo del paladar por medio de un método semiautomático [4] que identifica los límites del tejido en imágenes de resonancia magnética, que luego son utilizados en la obtención del grado de apertura del velo. Para representar la acústica de la voz se utilizan los ampliamente conocidos parámetros MFCC (Mel Frequency Cepstral Coefficients). Finalmente, se utilizan redes neuronales artificiales para determinar un mapeo no lineal, el cual busca medir la capacidad de estimar el grado de apertura del velo del paladar a partir de la información acústica codificada en los MFCCs.

* Trabajo de grado

** Facultad de ingenierías Físico-Mecánicas. Escuela de ingenierías eléctrica, electrónica y de telecomunicaciones. Director: Franklin Alexander Sepúlveda Sepúlveda, PhD.

Abstract

Title: Analysis of the relation between velopharyngeal opening and voice spectral properties through the use of magnetic resonance videos.*

Authors: Moisés Alfonso Guerrero Jiménez

William Gilberto Almeida Mantilla**

Key Words: Soft palate, Mel Frequency Cepstral Coefficients, Artificial neural networks, real time Magnetic Resonance Imaging.

DESCRIPTION

The velum or soft palate is an important organ in the speech production; however, unlike organs like lips, the position of the velum makes difficult its dynamic analysis and its relationship with the voice acoustics. The option to infer the velum dynamics results so important in tasks related with speech therapies and language learning, among others. Although exist tools like Electromagnetic Articulography and Magnetic Resonance Images, these are too expensive, and they have some practical disadvantages at the moment to be used. Therefore, it is posed the use of the voice like a mean to inquire about the velopharyngeal function by using acoustic-articulatory mapping methods [1].

The velar dynamics are obtained from real time Magnetic Resonance Imaging (rt-MRI) of the vocal tract, through the USC-TIMIT database [2] [3]. The procedure consists, first of all, in an extraction of the velar position by a semi-automatic method [4] that identifies the tissue boundaries in magnetic resonance images, which later are used to obtain the velopharyngeal opening. To represent the voice acoustics, we used the well-known MFCC parameters (Mel Frequency Cepstral Coefficients). Finally, we used artificial neural networks to determine a non-linear mapping, looking for a measure of the capacity to estimate the velopharyngeal opening from the acoustic information encoded in the MFCCs.

* Bachelor Thesis

** Facultad de ingenierías Físico-Mecánicas. Escuela de ingenierías eléctrica, electrónica y de telecomunicaciones. Director: Franklin Alexander Sepúlveda Sepúlveda, PhD.

Introducción

El estudio de la señal acústica de la voz es un tema de gran interés para la comunidad científica [5]. Aunque existen diversas investigaciones enfocadas a la explicación del modelo de la producción del habla, incluyendo aquellas basadas en modelos teóricos de ecuaciones diferenciales [6], aún se tienen falencias a la hora de explicar la relación existente entre los órganos articulatorios (labios, lengua, velo del paladar) y la señal acústica emitida mediante datos empíricos.

El objeto de estudio del presente trabajo corresponde a uno de los órganos responsables de la producción del habla: el velo del paladar. En particular, el velo del paladar es un tejido blando que se encuentra en la terminación del paladar y es de notable importancia durante la producción de varios tipos de sonidos, entre ellos los sonidos del tipo nasal [7]. El análisis del velo del paladar merece una importancia especial por los siguientes motivos: 1) su análisis permite la compresión y tratamiento de patologías relacionadas con el habla tales como hipernasalidad e hiponasalidad [8]; 2) es de utilidad en aplicaciones de inversión articulatoria [1]; y, 3) permite la clasificación y segmentación de sonidos del tipo nasal.

Aun contando con dispositivos de reciente desarrollo para el análisis del habla tales como el Articulógrafo Electro Magnético y de Ultrasonido, realizar el análisis dinámico del velo del paladar no es una tarea sencilla, principalmente debido a su posición. Aunque sí se cuenta en el estado del arte con dispositivos de imágenes de resonancia magnética que permiten visualizar el movimiento del velo del paladar, estos dispositivos son de un coste elevado y no representan una solución práctica para la obtención de la dinámica de este órgano. En tal sentido, resulta de gran utilidad contar con métodos de bajo costo que permitan inferir el grado de apertura velofaríngea a partir de la información extraída de la señal acústica de la voz mediante métodos de mapeo acústico-articulatorio.

Para la aplicación de un método de mapeo acústico articulatorio, se pueden implementar redes neuronales artificiales que permitan inferir el grado de apertura del velo del paladar a partir de las propiedades de la señal de voz. Para ello, es necesario contar con la representación de la dinámica del velo del paladar de manera concurrente con la señal acústica de voz. Estudiar la relación entre la morfología de los órganos articulatorios del tracto vocal (particularmente el velo del paladar), con la señal acústica de la voz, se convierte en una herramienta esencial para la comprensión del modelo de la generación del habla. Para ello es importante contar con datos reales para la realización del análisis, como los obtenidos mediante la construcción de la base de datos USC-TIMIT [2], la cual consiste en una serie de videos de resonancia magnética del tracto vocal, en paralelo con una señal acústica.

En el presente trabajo se analiza la relación estadística entre el grado de apertura del velo del paladar y la señal acústica de la voz. Para ello se utiliza un conjunto de registros de videos de MRI y de audio grabados de manera concurrente para una serie de hablantes. A partir de los videos MRI se estima el grado de apertura del velo del paladar y, el espectro de la señal acústica es representada mediante los comúnmente utilizados parámetros MFCC. A modo de resultados se presenta el error cuadrático medio resultante de utilizar procesos de regresión no lineal mediante re- des neuronales, además del valor de correlación estadística entre las variables estimadas y medidas.

Morfología, generación de la voz y tracto vocal

La voz es una señal acústica producida de forma voluntaria, donde se necesitan tres sistemas de órganos para originarla: respiratorio, fonatorio y articulario [9]. El modelo acústico de la generación de la voz humana está formado por las cuerdas vocales, un par de membranas que oscilan

modulando el flujo de aire que proviene de los pulmones y constituye la generación del sonido madre; y el tracto vocal, que filtra y amplifica la señal determinando su contenido fonético [10].

El tracto vocal es un conducto que va desde la glotis hasta la boca y cuya sección transversal varía con la posición a lo largo de su eje [11]. Está conformado por la cavidad oral, nasal, la faringe y la laringe. En estas cavidades se encuentran los órganos de la articulación divididos en pasivos y activos. Siendo los primeros los dientes, paladar duro y maxilar superior y los segundos la lengua, mandíbula, velo del paladar y labios [12].

La articulación es el proceso por el que las partes del aparato fonatorio interponen obstáculos para la circulación del flujo del aire. Dependiendo de la posición que adoptan los órganos articulatorios, se varía la forma del tracto vocal, donde cada configuración corresponde a un filtro acústico para el sonido que se produce en la laringe. Por esta razón el sonido vocal escuchado será distinto [10]. Cada letra tiene una forma distinta relacionada en el tracto vocal, esto hace posible diferenciarlas a partir de la configuración del tracto [12]. En el caso de las vocales se trata de sonidos emitidos solamente por la vibración de las cuerdas vocales, sin obstáculos o constricciones en entre la laringe y las aberturas oral y nasal. Por otra parte, las consonantes se emiten interponiendo un obstáculo mediante los órganos articulatorios [13].

Los elementos en los que se basa la forma del tracto vocal son su largo y los distintos diámetros transversales a la largo de su longitud, dependiendo de estos parámetros el tracto actuará como un determinado filtro acústico [12]. Este filtro depende de los formantes, la concentración de energía que se encuentran en una frecuencia o banda de frecuencias específicas para cada tipo de sonido [9]. Los formantes se identifican por el centro de frecuencia, el ancho de banda y la energía. Al variar la forma del tracto vocal estos tres parámetros también lo hacen, por esta razón la función del filtro y el sonido cambian [12]. Ya que el largo del tracto vocal está sujeto al género y a la edad de la

persona, los valores de los formantes para un fonema particular varían según sus características [10].

Los fonemas son la mínima unidad lingüística que puede producir cambios de significado. Estos se pueden dividir en dos grupos: vocálicos y consonánticos, que a su vez se clasifican según la vibración de las cuerdas vocales en fonados (sonoros) o áfonos (sordos). Los fonemas vocálicos al contrario de los consonánticos siempre son sonoros, pero la principal diferencia entre ambos es el tamaño de combinaciones referentes a las articulaciones en el tracto vocal. Las vocales se caracterizan por el grado de apertura del tracto y por el grado de elevación de la lengua. Las consonantes se clasifican por el punto de articulación, modo de articulación, función de las cuerdas vocales y posición del velo del paladar [9]. Las consonantes son una clase compleja de sonido, razón por la cual no hay una medida para distinguir las sordas de las sonoras, basándose en la constricción del tracto vocal. En las consonantes sordas la cavidad oral está completa o estrechamente cerrada, el flujo de aire es detenido o se produce ruido fricativo, en este tipo de consonantes se encuentran las fricativas, africadas y oclusivas. Por su parte, en las consonantes sonoras el tracto vocal se encuentra relativamente abierto lo que produce resonancia en la cavidad, encontrando en este grupo las nasales, liquidas y semivocales. Las consonantes sonoras se pueden describir por patrones de formantes y antiformantes en estado estable y segmentos de transición [14].

Si un fonema se produce por aire que pasa por la cavidad nasal se denomina fonema nasal. Por otro lado, si el aire sale por la boca se trata de un fonema oral. La diferencia radica en el tipo de resonador principal por encima de la laringe [15]. Si al momento de producirse una vocal de deja pasar aire por la cavidad nasal se convierte en una vocal nasalizada. Por ejemplo, en las vocales orales el velo del paladar sube, impidiendo que el aire fluya por la cavidad nasal, en las vocales nasales el velo del paladar baja liberando el paso de aire a la nasofaringe e incorporando de ese modo la resonancia nasal [13]. Las consonantes que corresponde a fonemas nasales, en el idioma

inglés, son /m/, /n/ y /ŋ/, que pueden encontrarse al final de una sílaba (fonemas trabantes) o sucedidos de una vocal [16].

Las características acústicas con las que se puede identificar la nasalización son la presencia de formantes nasales, los cambios de frecuencia de los formantes, la reducción de su amplitud, el incremento de su ancho de banda y la presencia de antiformantes. Por tanto, extraer la representación paramétrica de señales acústicas es fundamental para obtener un mejor rendimiento en un sistema de estudio de la voz. Existen varios métodos enfocados a este propósito, haciendo necesario considerar cuál es el más adecuado con respecto a las necesidades del proyecto, además de brindar información relevante de la señal de la voz. Un método de extracción estudiado fueron los ampliamente conocidos MFCC (Mel Frequency Cepstral Coefficients) [17], que logran proporcionar información detallada de las propiedades de la voz para cada tipo de hablante, característica por la que son ampliamente conocidos en el campo del reconocimiento de voz.

La naturaleza del tracto vocal constituye una base fundamental para el estudio de la voz, la clasificación de sonidos y caracterización del hablante [18], propiedades que reflejan la relevancia que tiene el análisis de la voz para este trabajo.

1. Objetivos.

1.1 Objetivo general

Analizar la relación existente entre el grado de apertura velofaríngea y las propiedades acústicas de la voz en fonemas nasalizados mediante el uso de vídeos de resonancia magnética del tracto vocal.

1.2 Objetivos específicos

- Segmentar el velo del paladar en videos de resonancia magnética y estimar el grado de apertura velofaríngea mediante el uso de técnicas de segmentación de imágenes MRI reportadas en el estado del arte.

- Estimar parámetros acústicos relacionados con el velo del paladar sobre señales de voz pertenecientes a pronunciaciones que involucren ciclos de apertura y cierre del velo del paladar.

- Estimar una función que permita mapear desde el dominio de los parámetros acústicos del habla hacia el grado de apertura velofaríngea mediante el uso de técnicas de entrenamiento de regresores.

2. Datos

Los datos utilizados para el desarrollo del presente proyecto corresponden a una serie de vídeos de resonancia magnética en tiempo real (rt-MRI) del tracto vocal para cada hablante, los cuales están

sincronizados con las respectivas señales de audio. Estos datos están consignados en la base de datos USC-TIMIT [2].

La información contenida en la base de datos USC TIMIT permite realizar estudios sobre la anatomía y dinámica articulatoria del tracto vocal en los diferentes procesos asociados con la generación del habla. De esta base de datos se obtienen secuencias de imágenes de la sección midsagital de la cabeza y cuello del hablante donde se puede observar el movimiento de los órganos involucrados en la producción del habla: cuerdas vocales, labios, lengua, faringe, laringe, mandíbula, paladar blando. Además, cuenta con la información acústica producida por el hablante grabada de manera concurrente con los videos rt-MRI.

Los datos de resonancia magnética en tiempo real se adquirieron de diez hablantes nativos de inglés americano, donde, ninguno de ellos reportó padecer patologías asociadas a la audición y/o habla. Los hablantes corresponden a 5 hombres y 5 mujeres cuyas edades oscilan en un rango entre 20 a 46 años. Cada hablante articuló un conjunto de 460 oraciones (las mismas de la famosa base de datos TIMIT [19]), las cuales están diseñadas para obtener todos los fonemas del inglés estadounidense en una amplia gama de contextos prosódico y fonológico, con los procesos de habla conectados característicos del inglés hablado. Las características principales de los hablantes que conforman la base de datos USC-TIMIT se muestran en la Tabla 1.

Tabla 1.

Características de los hablantes

Hablante	Genero	Edad	Lugar de nacimiento
F1	Femenino	23	Commack, NY
F2	Femenino	32	Westfield, IN
F3	Femenino	20	Palos verdes, CA

F4	Femenino	46	Pittsburgh, PA
F5	Femenino	25	Brawley, CA
M1	Masculino	29	Buffalo, NY
M2	Masculino	33	Ann Arbor, MI
M3	Masculino	26	Madison, WI
M4	Masculino	26	St. Louis, MO
M5	Masculino	27	Mammoth, CA

El proceso de adquisición de datos fue realizado en Los Ángeles County Hospital en un escáner Signa Excite HD 1.5T (GE Healthcare, Waukesha, WI) sobre el que los participantes permanecieron en posición horizontal, al tiempo que las frases de estímulo fueron proyectadas de manera individual en una pantalla con la que los hablantes mantenían contacto visual.

La reconstrucción de la imagen MRI se realizó a través del software MATLAB (Mathworks, South Natick, MA). Los marcos de imagen se formaron utilizando la reconstrucción de cuadrícula a partir de datos muestreados a lo largo de trayectorias en espiral. Finalmente se obtuvieron videos con una velocidad de 23,18 fotogramas por segundo a una resolución de imagen de 68x68 píxeles (2,9 x 2,9 mm por pixel).

De manera simultánea se grabó el audio usando un micrófono de fibra óptica (Optoacoustics Ltd., Moshav Mazor, Israel) a una frecuencia de muestreo de 20kHz dentro del escáner MRI [2], los autores desarrollaron adicionalmente un algoritmo de procesamiento de señal del tipo adaptativo que considero las características periódicas del ruido generado por el equipo MRI de manera que este se pudiese identificar y suprimir.

En total se obtuvo una base de datos con 100 videos para cada hablante que contienen el total de 460 frases, cada video está conformado por cinco frases separadas entre sí por pausas [2].

2.1 Análisis preliminar de la base de datos y observación de los hablantes

En la secuencia de videos de la base de datos se aprecia el tracto vocal junto con los órganos involucrados en el proceso de articulación del habla, entre ellos el velo del paladar. En las imágenes se pueden observar también características físicas de los hablantes como la nariz, el mentón y la papada como se ilustra en la Figura 1.

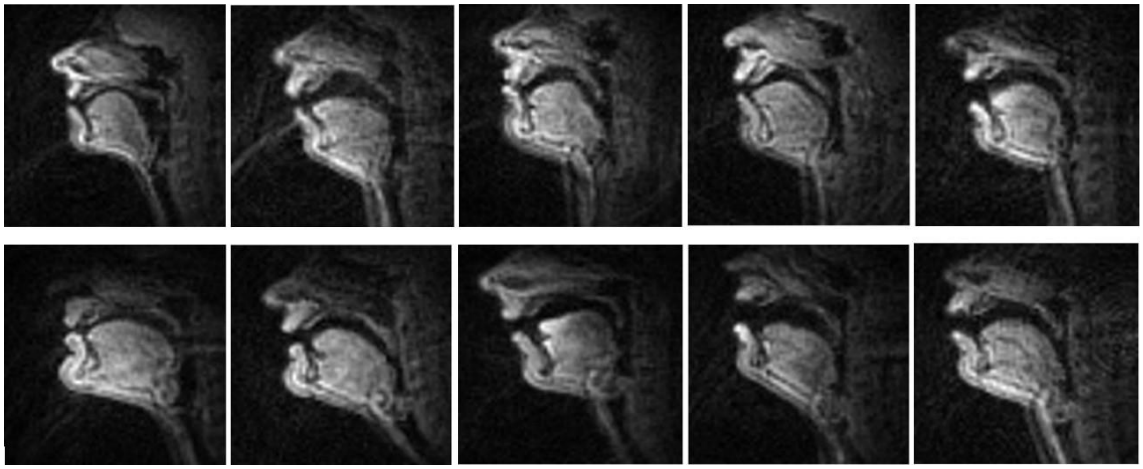


Figura 1. *Fotogramas de cada uno de los hablantes de la base de datos*

Durante el proceso de reconocimiento de la base de datos se examinó la dinámica del tracto vocal; el conjunto de movimientos representa un complejo sistema de articulación de fonemas para el habla inglesa en el que participan todos los órganos del aparato fonador.

Se prestó especial atención al comportamiento del velo del paladar. Tal como se expone en [20] el velo es un tejido delgado, flexible, unido al paladar duro que desvía y regula el flujo de aire hacia la cavidad nasal, principio físico para la generación de fonemas nasales. Si bien la función de apertura y cierre del velo es para todos los hablantes similar, se observó que las características físicas y la dinámica del movimiento de este órgano varían levemente en función de cada persona. Los

órganos que interactúan con el velo son la parte posterior de la faringe y la lengua; con ellos entra en contacto cuando se cierra o se abre respectivamente.

Durante el análisis preliminar de la base de datos se identificaron características de importante consideración para el desarrollo del proyecto, dos de las más relevantes fueron la resolución de la imagen y la presencia de ruido. Los fotogramas de cada video cuentan con una resolución de 68x68 pixeles (2.9x2.9 milímetros por pixel), característica que puede afectar directamente la percepción de las imágenes, en especial de un órgano delgado y pequeño como lo es el velo, generando un grado de incertidumbre sobre su posición. Por su parte la presencia de ruido ocasiona variación aleatoria de intensidad en los pixeles distorsionando el movimiento y la identificación de contornos. La secuencia de videos del hablante M5 presentó perturbaciones en la imagen de resonancia pro- ducto de factores externos, ocasionando una serie de ondulaciones que impidieron observar con claridad la dinámica del tracto vocal, este fenómeno solo fue observado en el hablante M5, mientras que en los demás hablantes se identificó solamente ruido de características aleatorias.

La velocidad de articulación fue otro aspecto importante observado en el estudio preliminar, se advirtió que esta variaba entre cada hablante y frase, mientras algunos hablantes articulaban una frase en 3 segundos, a otros les tomaba menos de 2 segundos, por lo tanto, la dinámica de órganos como el velo está sujeta también a la variación de la velocidad de articulación del hablante, tal cómo se presenta en [21].

En vista de que la información contenida en la base de datos representa la herramienta fuente para el análisis de la dinámica velar, las consideraciones resaltadas a priori generaron un importante campo de discusión durante el desarrollo del proyecto, ya que establecieron las condiciones de partida de la investigación.

3. Método

El proyecto partió de la selección del algoritmo de segmentación, pasando por la definición de parámetros acústicos de la señal de habla, hasta el estudio de las redes neuronales artificiales para la realización del mapeo no lineal. La Figura 2 resume el proceso de extracción y adecuación realizado.

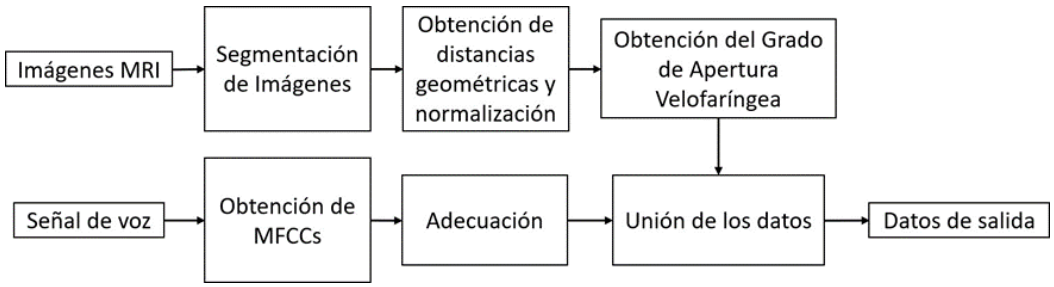


Figura 2. *Diagrama de bloques de la adecuación de datos para entrenamiento.*

3.1 Estimación del grado de apertura velofaríngea a partir de los videos de resonancia magnética

3.1.1 Segmentación de imágenes de resonancia magnética en tiempo real. La información contenida en las secuencias de imágenes MRI de la base de datos muestra la dinámica del velo del paladar; donde, su correcta extracción de interés notable para este proyecto, es un gran reto debido a características propias de este órgano y a la resolución de las imágenes. Al revisar el estado del arte se encontró que existen pocos métodos disponibles que permitan realizar esta tarea; aun así, se realizó un proceso de evaluación de algunos de estos métodos para determinar aquel más conveniente para la tarea en cuestión.

En [22] y [23] se presentan dos versiones de un software de segmentación sobre secuencias de imágenes de MRI del tracto vocal desarrolladas por el mismo laboratorio en el que se obtuvo la base de datos. Ambas versiones involucran el mismo paradigma basado en la construcción de una malla que cubre la totalidad del tracto vocal [4], tal como se ilustra en la Figura 3. Sobre cada uno de estos segmentos de línea que conforman la malla, el programa extrae los niveles de intensidad presentes en los píxeles cercanos, identificando los bordes del tejido, por medio de un umbral de intensidad (ver Figura 3) que permite delimitar las regiones de interés [4].

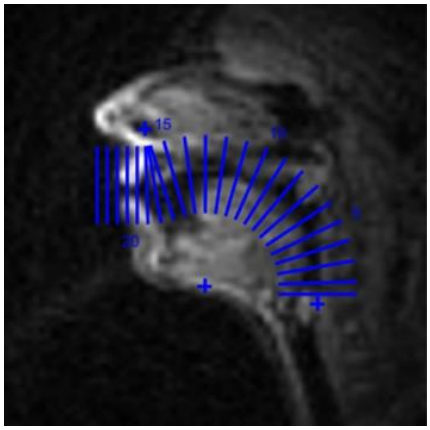


Figura 2. *Malla de referencia para la segmentación.*

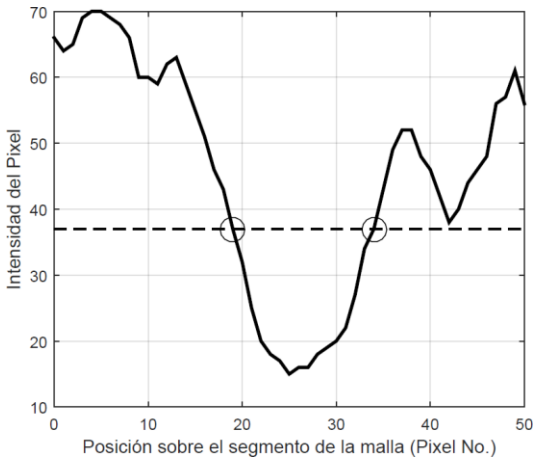


Figura 3. *Niveles de intensidad de los píxeles del tracto vocal sobre un segmento de referencia [4].*

Cada versión del software fue revisada, probada y estudiada advirtiendo las ventajas que cada herramienta proporcionó y la eficiencia respecto a los intereses del proyecto. A continuación, se exponen las principales funciones y características de cada una y los lineamientos de selección considerados para la elección del software de segmentación a usar en el proyecto.

Primera versión del software La primera versión del software [22] realiza una corrección de imagen y supresión de ruido granular, aumentando el contraste entre tejidos y las vías respiratorias. La malla se construye de forma semiautomática, teniendo como insumo de entrada 4 puntos de referencia correspondientes a la glotis, el punto más alto del paladar duro, la cresta alveolar y el punto medio entre los labios del hablante. Estos 4 puntos son seleccionados de forma manual y únicamente para la primera imagen del video. Una vez hecho esto, el software realiza la segmentación automática del tracto vocal [4]. Un ejemplo de resultado de esta segmentación se muestra en la Figura 4.

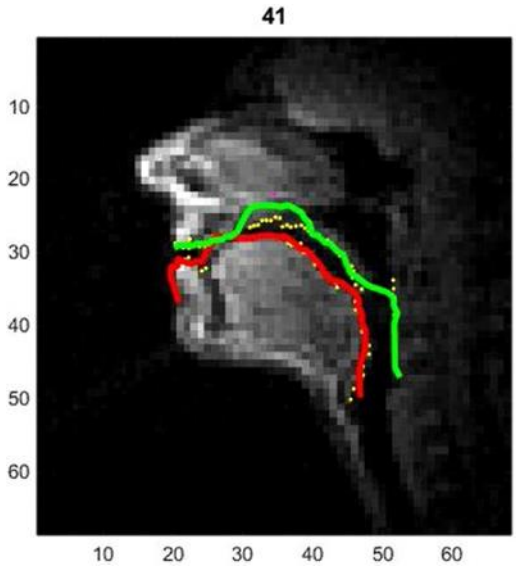


Figura 4. *Resultado de la segmentación dado por la serie de funciones rt-MRI.*

Segunda versión del software El segundo software evaluado consiste en una versión actualizada de [22], donde el proceso de segmentación es similar. Esta versión incluye una interfaz gráfica de usuario (GUI) que se ilustra en la Figura 5. La herramienta permite realizar análisis y observación *frame* por *frame* de la secuencia de videos MRI junto con la señal de audio y su respectivo espectrograma [23]. Adicionalmente, proporciona una herramienta con la que es posible corregir manualmente la segmentación de cada *frame*, en aquellos casos en los que resulte una segmentación incorrecta del tracto vocal. El software presenta ventajas tales como:

- Rápida segmentación semiautomática.

- Proporciona información de la voz a través de la interfaz, permitiendo observar la señal de audio junto con su espectrograma.

- Incluye herramientas que permiten modificar las propiedades de la malla como la ubicación o la cantidad de segmentos de referencia, para adaptarla a las características del tracto vocal de cada hablante.

- Corrección de intensidad que mejora las características de cada *frame* en términos de brillo y contraste, usando el método reportado en [24], además de un filtrado paso-bajo en 2 dimensiones.

- Función de corrección que permite modificar, manualmente, la posición de las líneas de contorno del tejido. Es útil en caso de encontrar errores en *frames* particulares, ver Figura 6.

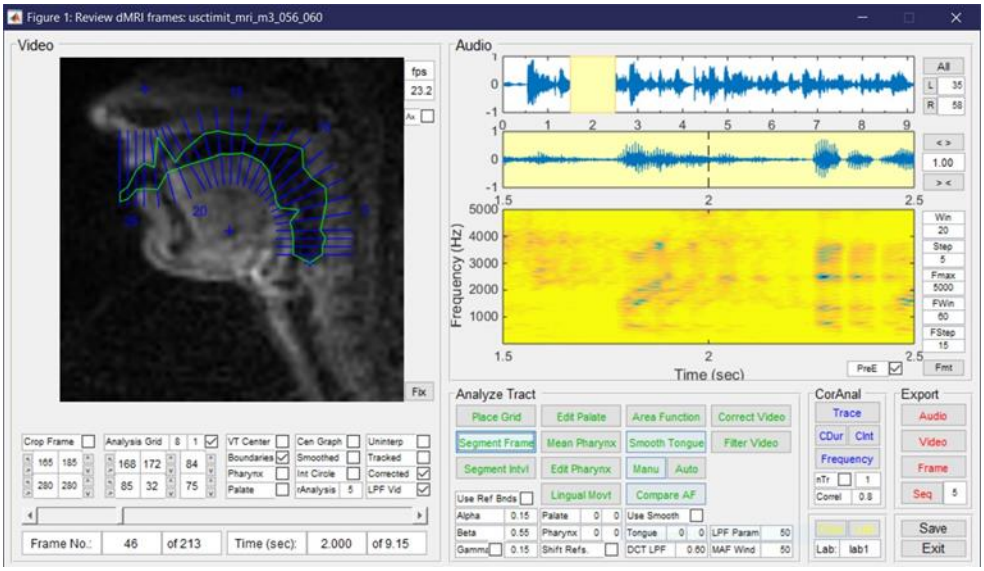


Figura 5. *Interfaz gráfica de usuario del segundo software.*

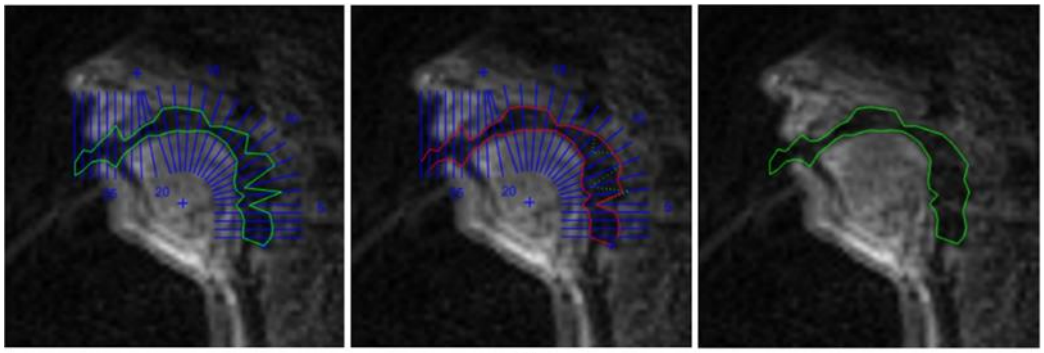


Figura 6. *(a) Frame con errores de segmentación. (b) Corrección del frame. (c) Frame corregido.*

Las ventajas de segmentación y las herramientas con las que cuenta el segundo software son características de gran valor para los objetivos del proyecto, por lo tanto se decidió que era la mejor versión para el proceso de segmentación y por consiguiente se trabajó con este. Como se mencionó anteriormente, cuenta una completa interfaz gráfica que reconoce y sincroniza los elementos que conforman la base de datos (imagen y audio). Estos elementos son cargados a la herramienta, donde se puede observar la articulación a la vez que se muestran las señales de audio y espectrograma en

función del tiempo y sobre las que se proporciona una completa serie de opciones y técnicas que facilitan el análisis del comportamiento del tracto vocal de los hablantes. La interfaz está diseñada con el propósito de analizar videos de resonancia magnética *frame* por *frame*, además de ver de manera sincronizada las características de la señal de audio [23].

El vídeo y el audio son cargados a través de la invocación del comando inspect_dmri () desde la ventana de comandos de MATLAB, cuyo argumento es el nombre de los archivos entre comillas simples. Adicionalmente, se debe asignar un nombre a la variable correspondiente a la estructura de datos en la cual se almacena la información. Un ejemplo de comando es:

usctimit_mri_m3_056_060 = inspect_dmri('usctimit_mri_m3_056_060')

Una vez validada la información de audio y video el software carga la interfaz con los datos de cada registro, un ejemplo del resultado de este proceso se ilustra en la Figura 7.

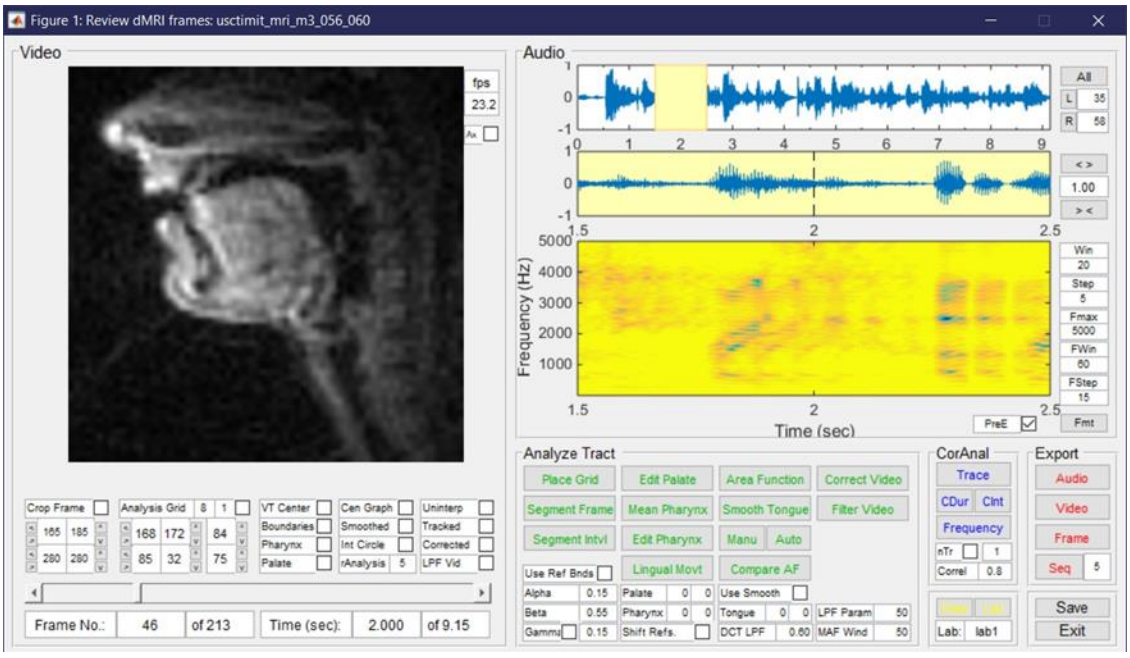


Figura 7. *Información cargada en el Software*

Tal como se ha mencionado, este software cuenta con una herramienta de corrección que consiste en un preprocesamiento de imagen para mejorar las características propias de los videos en términos de brillo y contraste de tal manera que mejoren los valores de intensidad de todos los *frames* y así se homogenice la secuencia que conforma el video, con el fin de facilitar el método de procesamiento y mejorar los resultados de segmentación, haciendo uso del método reportado en [24]. El preprocesamiento consiste en la identificación de las zonas más brillantes de la imagen MRI del tracto vocal, con el objetivo de realizar una corrección en estos pixeles y obtener una imagen con una función de intensidad suavizada. En la mayoría de los casos los pixeles más brillantes se encuentran en la zona de la nariz, debido a la cantidad de masa con respecto a la cabeza y el cuello [23].

Una vez realizado el proceso de corrección de imagen, se obtiene un grupo de *frames* más homogéneos y ecualizados, donde se reduce el riesgo de error por incorrecta detección, debido a cambios abruptos en los valores de intensidad de los *frames*.

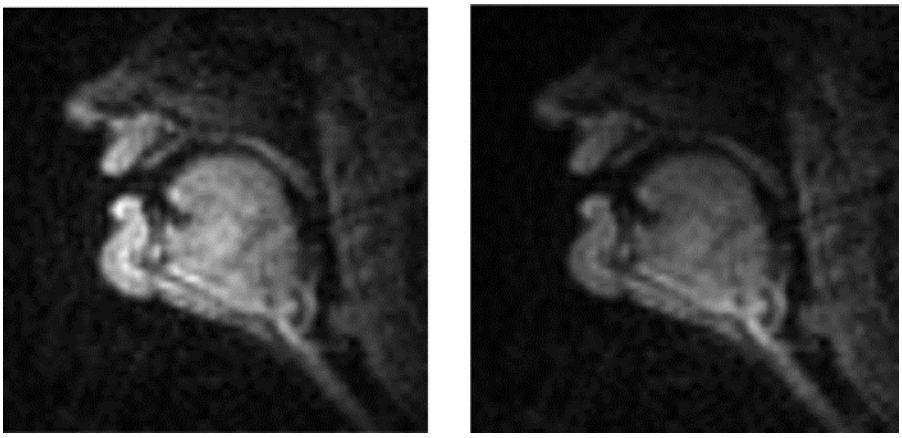


Figura 8. *Ejemplo de corrección de imagen en el frame 82 del primer vídeo, asociado al segundo hablante masculino. Izquierda: Frame sin corregir. Derecha: Frame corregido.*

En la Figura 8 se ilustra la transformación del frame número 82, del registro 1, para el segundo

hablante masculino, a través de la herramienta de corrección de video que presenta el software. Como se puede observar, se logra un proceso de ecualización seccionado a través de los criterios de comparación y dispersión por pixeles.

El proceso de generación de la malla se llevó a cabo a través de la ubicación de 4 puntos de orientación que corresponden a la pared de la glotis, el punto más alto del paladar duro, la cresta alveolar y la sección midlabial del hablante. Estos puntos fueron localizados sobre un frame ilustrativo, en el orden antes mencionado, haciendo clic izquierdo con el ratón en el punto específico.

Una vez situados de forma manual los puntos de referencia, se crea una malla que cubre toda la longitud del tracto vocal del hablante, desde la pared de la glotis en la parte inferior hasta la sección de los labios en la parte superior. Sin embargo, la correcta identificación y ubicación de los puntos de referencia no es suficiente para crear una malla acorde a la morfología del hablante, puesto que el movimiento de las estructuras que conforman el tracto vocal, en algunas ocasiones, queda por fuera de la malla creada, haciendo necesario realizar una revisión que garantice una cobertura total del movimiento articulatorio. Es importante resaltar que la morfología de cada persona es diferente, motivo por el que se hizo necesario crear una malla para cada hablante.

La interfaz cuenta con una sección que permite modificar en tiempo real las características de la malla, permitiendo aumentar su resolución, modificar su tamaño manteniendo la estructura inicial o cambiar la inclinación a partir de los puntos de referencia, todo esto con miras a obtener una mejor segmentación del tracto vocal.

La sección del velo del paladar es un área particularmente complicada de identificar y segmentar debido a su naturaleza blanda, delgada y elástica, por lo que se presentaron situaciones de considerable frecuencia, donde no se identificó de manera acertada la posición de velo tal como se ilustra en la Figura 9. Luego de realizar un proceso de inspección visual y en tiempo real, se

logró corregir manualmente, los frames donde se presentaron errores.

Figura 9. *Ejemplo de incorrecta segmentación del velo*

Se observó que las mallas con una pequeña cantidad de líneas (entre 10 y 18) reducían en gran medida de la resolución de la segmentación y proporcionaban datos de mapeo con cambios abruptos, caso que resultaba contraproducente para el velo del paladar debido a las pequeñas proporciones que tiene este órgano. Se determinó también que usar una gran cantidad de líneas para formar la malla de mapeo (superior a 40) aumentaba el error asociado a la identificación del velo, su movimiento y su grado de apertura. Esto es debido al aumento de la resolución, el hecho de que se implementaran más líneas implicaba un aumento en la complejidad para detectar con exactitud la línea donde se producía un cambio abrupto de la posición del velo. También se pudo notar que el tiempo de corrección de un proceso de segmentación automático evaluado como incorrecto era superior y aumentaba de manera directamente proporcional a la cantidad de líneas por malla, debido a la mayor cantidad de segmentos por corregirse. Esta situación se vió reflejada en un mayor gasto de tiempo, con base a este análisis se decidió que el rango óptimo de líneas dentro de la malla de mapeo oscilara entre 22 y 37, de forma tal que quedaran 4 líneas de medición situadas sobre la región velar, desde la punta del velo hasta el

punto más cercano al eje de movimiento del mismo.

3.1.2 Obtención del Grado de Apertura Velofaríngea. El software mencionado en el apartado anterior no entrega el valor del grado de apertura velofaríngea, razón por la cual, se propuso realizar una estimación de la apertura del velo del paladar por medio de la diferencia de cada una de las distancias geométricas medidas sobre el velo y las distancias geométricas definidas sobre un segmento de referencia localizado en la zona de la faríngе, de manera similar a lo observado en la Figura 10.

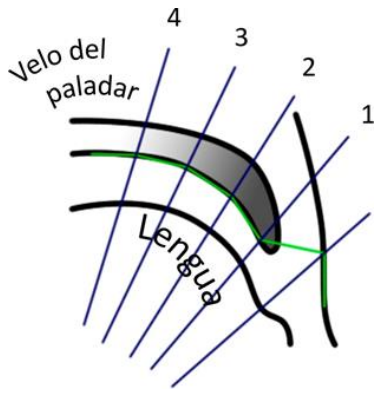


Figura 10. *Ilustración de los 4 segmentos sobre el velo además de 1 de las dos referencias medidas sobre la faríngе.*

El proceso de obtención de los diferenciales fue realizado con base a lo mostrado en la figura Fi gura 11, donde se puede observar la forma en que los datos sobre los puntos de medición del velo abierto presentan distancias geométricas menores (Figura izquierda) y aumentan su amplitud cuando el velo se encuentra cerrado (Figura derecha). Entre tanto, las distancias sobre la faringe presentan cierta tendencia a mantenerse en un valor fijo para cada frame del vídeo. Esta extracción de diferenciales permitió obtener 4 valores del grado de apertura respecto a la referencia en la faríngе.

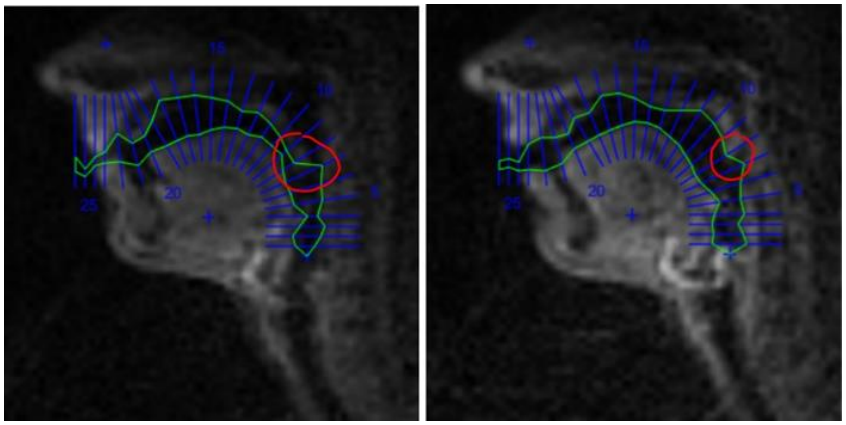


Figura 11. *Figura Izquierda: Distancias con menor amplitud para una apertura del velo. Figura Derecha: Distancias con mayor amplitud para un cierre del velo.*

Se partió del desarrollo un algoritmo en Matlab que, tal como se muestra en la Figura 12, toma la información de cada uno de los archivos generados por el software de segmentación y mide las distancias geométricas desde la parte inferior de cada uno de los segmentos de referencia y la parte superior asociada al velo.

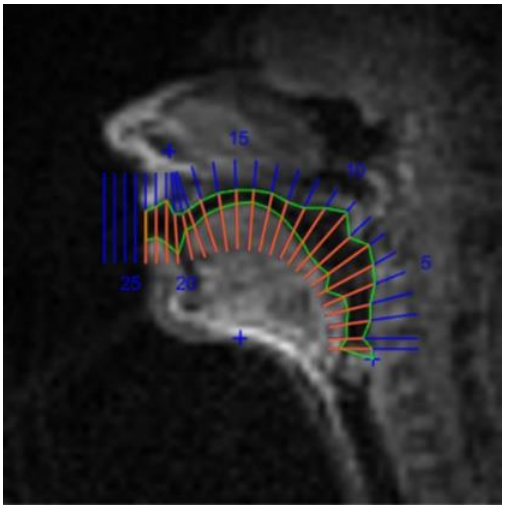


Figura 12. *Distancias geométricas obtenidas, por medio de la malla de referencia.*

Cada una de estas distancias, que se pueden visualizar en color anaranjado, fueron extraídas a partir del mismo archivo dado por el Software de segmentación, como se puede observar en la Figura 13, donde la parte superior del tracto vocal (Segmentos amarillos) y la parte inferior de la malla (Segmentos rojos), se observan en un plano con posiciones dadas por los píxeles de cada frame de un vídeo particular.

De acuerdo a lo mencionado anteriormente, se realizó la selección de los segmentos de interés que hacen parte de la región velofaríngea de forma que pudieran ser utilizados para extraer los diferenciales anteriormente referidos.

Cabe mencionar que los valores mayores a las medidas de 68x 68 del vídeo, se presentan debido a una interpolación de imagen realizada por la misma herramienta de segmentación, redimensionando cada frame hasta 340 x 340. La distancia fue medida entre cada uno de los puntos marcados en azul, con cada uno de los puntos marcados en rojo, que se encuentran ubicados sobre el mismo segmento de la malla, obtenidas por medio de la Ecuación de Distancia Geométrica (Ecuación 2.1):

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (2.1)$$

Donde d es la distancia entre dos puntos (x_1, y_1) y (x_2, y_2) cualesquiera.

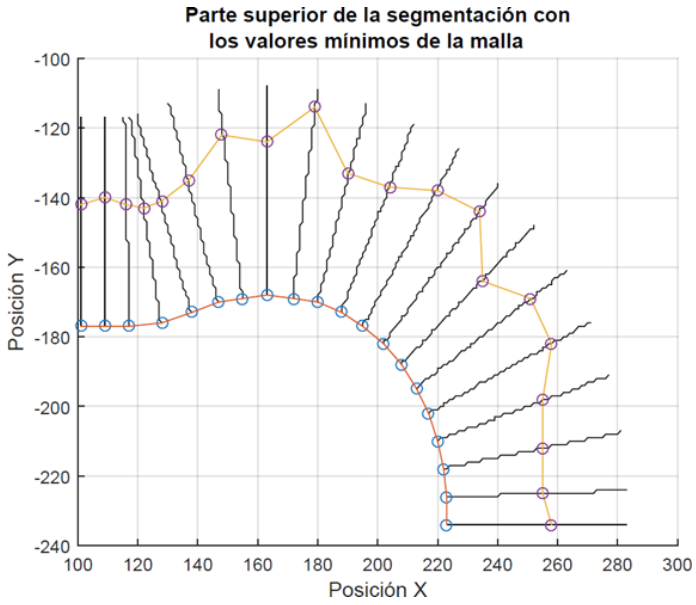


Figura 13. *Parte superior de la segmentación (Amarillo) con los valores mínimos de la malla (Rojo). Los segmentos del 7 al 10 corresponden a la sección del velo y los segmentos 5 y 6 a la referencia sobre la faríngе.*

3.2 Representación de la información acústica

3.2.1 Remoción de Silencios. Cada conjunto de datos (por hablante) cuenta con un poco menos de cien vídeos, cada uno con cinco frases separadas por espacios de silencios. Por tanto, se procedió a eliminar los silencios a partir de la señal de audio, para así determinar las fronteras temporales respectivas en los videos MRI. Este proceso se puede realizar mediante algoritmos de detección de voz (*Voice Activity Detector*, VAD). Aunque existen diversos algoritmos de VAD, en el presente trabajo se utiliza el algoritmo Google VAD-WebRTC en su versión webrtcvad 2.0.10 debido a su buen desempeño.

Los archivos de audio deben estar muestreados a 8, 16 o 32 kHz por lo tanto los audios de la base de datos se convirtieron a una frecuencia de 32 kHz. Los archivos de audio se cargaron al

algoritmo de VAD que arrojó valores de tiempo seguidos de unos o ceros, donde 1 corresponde a presencia de voz y 0 corresponde a momentos de silencio. El paso siguiente fue almacenar los datos de tiempo arrojados por el algoritmo en una matriz organizada en dos columnas de tal manera que la primera columna indicase el tiempo donde se termina un silencio y la segunda el tiempo donde comienza el siguiente silencio, en otras palabras, el tiempo donde hay presencia de sonido.

Finalmente, se creó un algoritmo cuyas entradas fueron el vector de datos que contenía los rangos donde existía sonido y el archivo de audio original, la función del algoritmo fue identificar dichos valores, relacionarlos con el tiempo del audio original y suprimir los tiempos y las señales de audio fuera de estos rangos, para finalmente unir todos los espacios de tiempo donde el sujeto estaba hablando. Este procedimiento fue realizado para cada archivo, obteniendo un audio de más corta duración. Este proceso se realizó de manera similar para los archivos de vídeo, identificando y suprimiendo los frames donde no había presencia de sonido a través del tiempo. Finalmente se verificó que los nuevos archivos de audio y vídeo tuvieran la misma duración para fines de sincronización.

3.2.2 Coeficientes Cepstrales en las frecuencias Mel (MFCC). Los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC) son una técnica de obtención de parámetros de voz a través de componentes relevantes del habla basados en la forma en que el sistema auditivo humano percibe la voz. Las características obtenidas por esta técnica son similares a la variación conocida del ancho de banda del oído humano con la frecuencia [17]. Los MFCC determinan propiedades locales de la señal de voz asociadas a las características propias de cada hablante [25], por esta razón son ampliamente usados en sistemas de reconocimiento de voz. Los coeficientes MFCC han mostrado mejores resultados frente a otras técnicas de extracción de parámetros de la voz [26] [27] en

especial en lo referente a la discriminación de fonemas [28], además de presentar mayor robustez ante la presencia de ruido y calidad en la información representativa que entrega [29] [30].

Los MFCC son el resultado de la transformada del coseno, del logaritmo real del espectro de energía de tiempo corto expresado en la escala de frecuencia de Mel [26]. La percepción humana del contenido frecuencial de sonidos de la voz no es lineal, razón por la cual para cada tono con una frecuencia f [Hz], existe un tono medido en la escala de Mel [27]. La escala de Mel es lineal por debajo de 1 kHz y logarítmica por encima este valor. El cálculo de esta frecuencia en Hertz se ilustra en la Ecuación 2.2.

$$Mel = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.2)$$

El proceso general para el cálculo de los MFCC se muestra en la Figura 14.

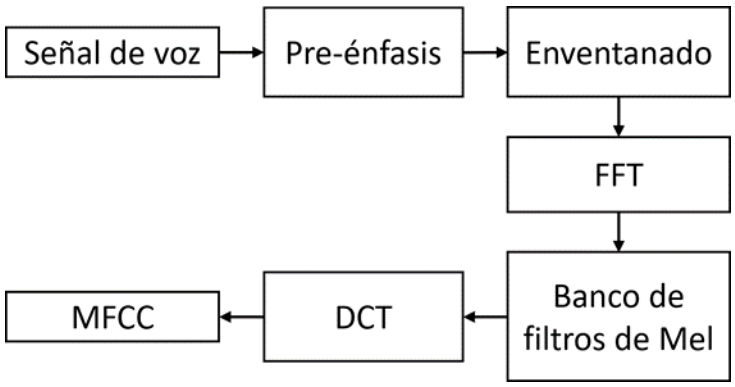


Figura 14. *Diagrama de bloques de MFCC*

El primer paso para hallar los coeficientes cespectrales en la escala de Mel es el pre énfasis de la señal de audio, que consiste en la aplicación de un filtro FIR de primer orden, como se muestra en la

Ecuación 2.3.

$$p(z) = 1 - \alpha z^{-1} \tag{2.3}$$

Este paso se realiza con el fin de incrementar la magnitud de la cantidad de energía que está en las frecuencias altas de la señal y hacer, de este modo, que el parámetro sea detectable en etapas siguientes. Posterior a esto se realiza un procedimiento de enventanado que consiste en segmentar las muestras de voz en un cuadro pequeño, con una longitud dentro de los 20 a 40[*ms*]. La señal de voz se divide en frames de N muestras espaciados por M (M<N). Los frames obtenidos por el paso anterior presentan cierta superposición, haciendo necesario usar el enventanado para guarda el cambio entre frames que se encuentra en el final de cada uno. Para esto se utiliza una ventana de Hamming, de forma tal que se va multiplicando por cada frame, dando como resultado los frames de ventana.

Cuando se tiene la señal enventanada, se aplica a cada tramo la FFT, para convertir cada frame al dominio de la frecuencia, tal como se muestra en la Ecuación 2.4.

$$S_k = \sum_{n=0}^{N-1} s\,(n) \cdot exp\left(\frac{-j2\pi nk}{N}\right), k = 0,1,\ldots,N-1 \tag{2.4}$$

Luego del cálculo de la FFT se obtiene la amplitud del espectro de la señal $|S(K)|$ a la cual se le aplica el banco de filtros de acuerdo con la escala de Mel (H_i), y se calcula el logaritmo de la energía de salida, tal como se ilustra en la ecuación 2.5.

$$S_i = log_{10}\left(\sum_{k=0}^{N-1} |\,S(k)| \cdot H_i(k)\right), i = 0,1,\ldots,M \tag{2.5}$$

Donde M que representa el número de filtros del banco de filtros de Mel.

Por último, para convertir el espectro del registro de Mel en el dominio del tiempo se aplica la Transformada Discreta de Coseno (DCT), dando como resultado los Coeficientes Cespectrales en la escala de Mel, tal como se ilustra en ecuación 2.6.

$$MFCC(r) = \sqrt{\frac{2}{M}} \sum_{i=0}^{M-1} S_{i+1} \cdot cos\left(\frac{r(i+0.5)\pi}{M}\right), r = 0,1,\ldots,R-1 \qquad (2.6)$$

Donde $R \leq M$, siendo R el número de coeficientes que se desea obtener.

En este proyecto la extracción de los coeficientes MFCC se realizó con la implementación del código disponible en [31]. Para el cálculo de los coeficientes a partir de la señal de voz, el primer paso consistió en enfatizar la señal mediante un filtro FIR de primer orden con coeficiente de preacentuación (α). A la señal de voz enfatizada se le aplicó la transformada de Fourier de corta duración (FFT) especificando la duración del frame, su desplazamiento y la función de ventana de análisis. Seguidamente se computó la magnitud del espectro y se realizó el diseño del banco de filtros con M filtros triangulares que se encuentran espaciados de manera uniforme en la escala de mel y que se exponen en [32], entre los límites de frecuencia anteriormente mencionados. El banco de filtros fue aplicado a los valores de la magnitud del espectro para encontrar energías del banco de filtros (FBEs). Finalmente, se aplicó a las FBE comprimidas la transformada discreta de coseno, para producir los 13 coeficientes cespectrales definidos para este trabajo.

3.2.3 Obtención y Adecuación de los MFCC. El software utilizado fue HTK MFCC de

Matlab diseñado por Kamil Wojcicki [31], un software diseñado para la extracción de los coeficientes cepstrales de Mel (MFCC) a partir de una señal de voz, que permite realizar una configuración de parámetros como el periodo de análisis T_s, la cantidad de coeficientes N_c, el coeficiente de preénfasis α o el rango de frecuencias sobre el cual se genera el banco de filtros [LF, HF], para adaptarse a las necesidades del usuario. El periodo de análisis Ts se obtuvo a partir de la Ecuación 2.7,

$$T_s = \frac{du}{nframes}[s] \qquad (2.7)$$

Donde «du» corresponde a la duración del vídeo en segundos y «nframes» al número de frames del vídeo.

La configuración del rango de los filtros fue realizada definiendo las variables LF (Low Frequency) y HF (High Frequency) que indican el límite de frecuencia inferior y superior respectivamente, con ayuda de la FFT de las señales de voz para hombres y mujeres (ver Figura 15).

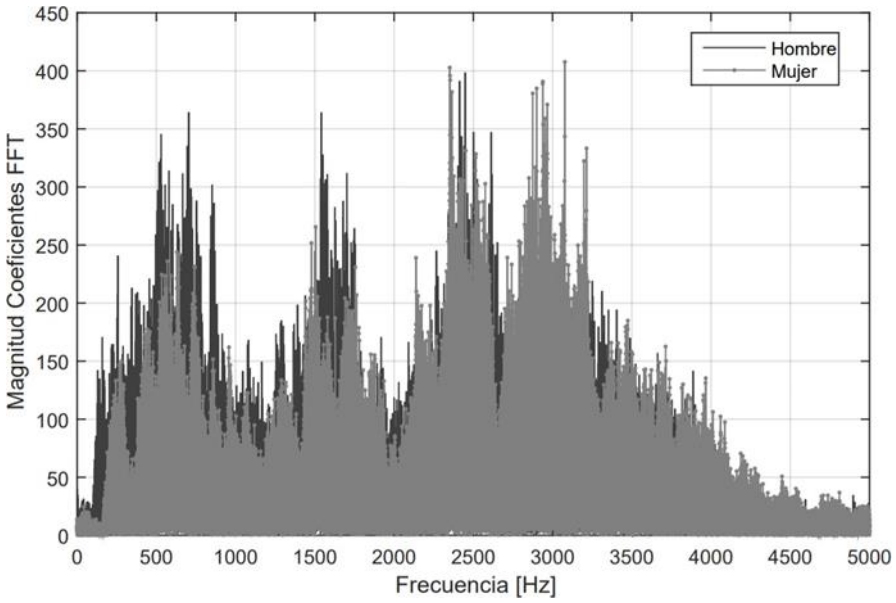


Figura 15. *Espectros de frecuencia promedio de la voz para hombres y mujeres.*

Los armónicos asociados a la voz incluyen frecuencias que van de los 300 hasta los 3000 [Hz] y algunos fonemas de tipo fricativo pueden alcanzar frecuencias superiores a los 4000 [Hz]. Estos rangos de frecuencia permitieron definir los limites superior e inferior como parámetros para la función de MFCC's como $LF = 50[Hz]$ y $HF = 3700[Hz]$, de forma que se pudiera incluir la mayor parte del ancho de banda asociado tanto a hombres como a mujeres por igual.

Por otro lado, al realizar un acercamiento a un rango de frecuencias de 0 a 500 [Hz], como se puede observar en la Figura 16, se puede evidenciar la existencia de una mayor cantidad de componentes a frecuencias cercanas a los 100 [Hz], en el caso de los hombres, y 250 [Hz] en el caso de las mujeres.

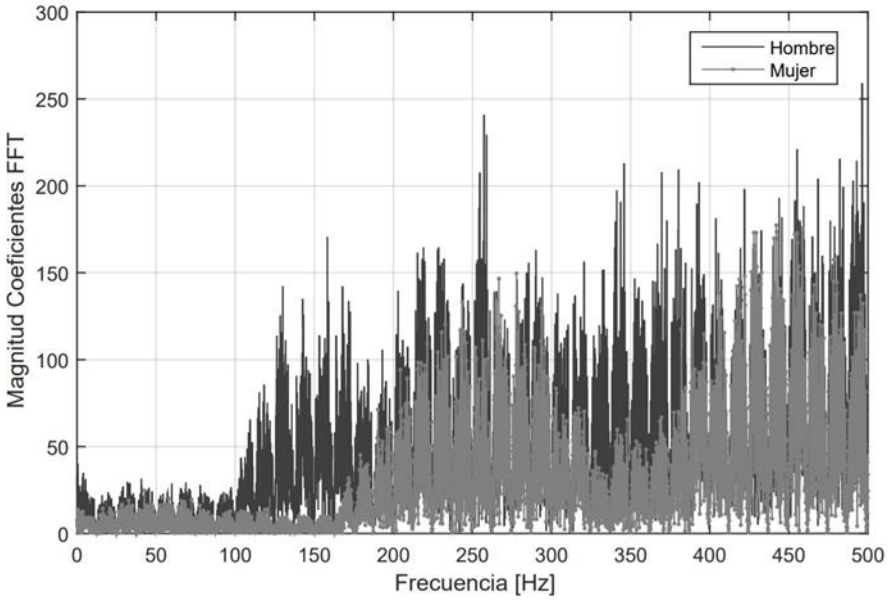


Figura 16. *Espectros de frecuencia promedio de la voz para hombres y mujeres entre 0 y 500 [Hz]*.

Se realizó un procedimiento un procedimiento de normalización de la influencia de la longitud del tracto vocal en el espectro mediante un proceso de transformación de frecuencias. Para realizar este proceso se hace necesaria la selección de un coeficiente de preénfasis, el cual se encuentra asociado a las características del espectro de frecuencias de un hablante particular, que pueden ser adaptadas a partir de la longitud del tracto vocal de la persona, respecto a uno de los hablantes cuya longitud del tracto puede ser utilizada como referencia. En este sentido, se hizo necesario obtener los valores estimados de dichas longitudes para cada uno de los hablantes.

La obtención de las longitudes del tracto vocal fue realizada a partir de medición de la longitud del tracto vocal sobre las mismas secuencias de videos MRI. Se utilizaron los valores obtenidos en [33], obteniendo una cantidad *L* de valores de longitud. De este conjunto de datos se utilizó la mediana como el valor aproximado de la longitud del tracto vocal del hablante, en busca de evitar el impacto de los datos atípicos al usar un valor promedio. Los valores de longitud obtenidos se pueden observar en la Tabla 2. para cada uno de los hablantes de la base de datos.

Tabla 2.

Longitudes del tracto vocal de cada hablante.

Hablante	Longitud [cm]
F1	13.13
F2	15.58
F3	14.26
F4	14.29
F5	14.61
M1	16.19
M2	15.71
M3	15.85
M4	15.41
M5	15.31

La obtención del coeficiente de normalización α fue realizada por medio de la selección de uno de los hablantes como referencia. En este caso se seleccionó el primer hablante masculino (M1), cuya longitud del tracto vocal sirvió como base para definir el resto de valores de α a partir de la Ecuación 2.8.

$$\alpha = \frac{L_i}{L_{M1}} \qquad (2.8)$$

Donde L_i es la longitud del tracto vocal de cada hablante i y L_{M1} es la longitud del tracto vocal del hablante M1.

La Tabla 3 recopila cada uno de los valores de α de acuerdo al procedimiento mencionado.

Tabla 3.

Coeficientes de preénfasis usado para cada hablante.

Hablante	α
F1	0.81
F2	0.96
F3	0.88
F4	0.88
F5	0.90
M1	1.00
M2	0.97
M3	0.97
M4	0.95
M5	1.07

Como se ilustra en la Figura 17, el objetivo se centró en obtener 13 coeficientes para cada frame del vídeo, que adicionalmente presentaran un grado de apertura asociado a los 4 puntos sobre el velo, mencionados anteriormente.

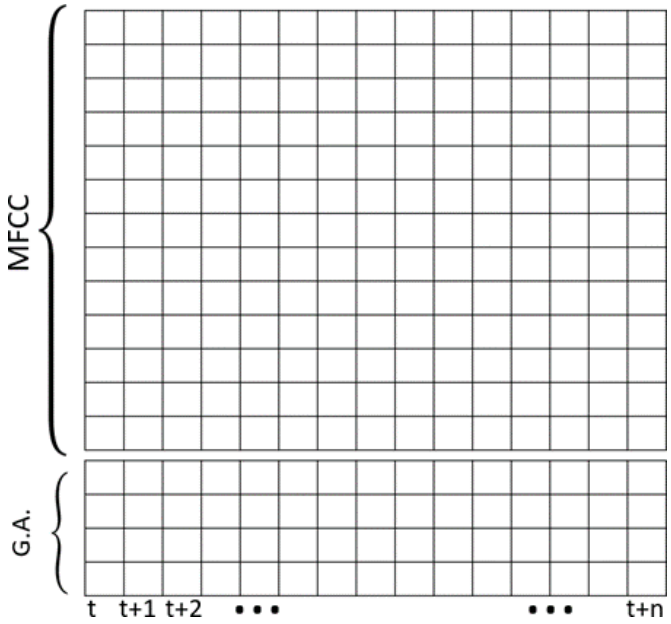


Figura 17. *MFCC y grado de apertura velofaríngea, para n espacios de tiempo.*

Si bien gran parte de le envolvente compleja de un fonema en un espacio de tiempo *t* se encuentra distribuida en los 13 coeficientes MFCC, el movimiento más lento de la articulación en las partes del tracto vocal, aumenta la probabilidad de existencia de información de un fonema particular distribuida en espacios de tiempo anteriores y posteriores a dicho instante, razón por la cual se propuso adicionar a cada espacio de tiempo *t*, los 13 coeficientes asociados a los espacios $t-1$ y $t+1$, obteniendo una matriz con dimensiones $39 \times n$, para *n* espacios de tiempo.

3.3 Preprocesamiento de los datos

3.3.1 Análisis de hablantes. A pesar de que el procedimiento realizado fue el mismo para cada uno de los hablantes, la existencia de ciertas diferencias en los datos extraídos a través de cada uno de los pasos realizados, evidenció la necesidad de corregir o descartar algunas de las muestras extraídas.

Partiendo de la segmentación del tracto vocal de cada uno de los hablantes, se pudo notar en general una buena segmentación, con excepción de algunos frames muy particulares que pudieron ser corregidos manualmente. Sin embargo, se observó que pese a la utilización de métodos de corrección de imágenes para la eliminación de ruido, la gran mayoría de los vídeos del hablante M5 mostraron errores de segmentación en frames en los que existía cierto grado de distorsión en la imagen (Ver Figura 18), presentando ciertas líneas de intensidad circulares a lo largo de cada frame.

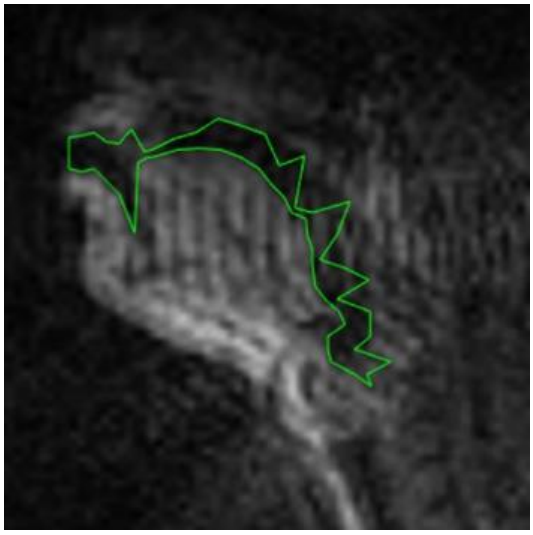


Figura 18. *Errores de segmentación producto de distorsiónes externas.*

Si bien los datos de segmentación podían ser corregidos con ayuda del software, la calidad de las imágenes de este hablante en particular no fue lo bastante buena como para obtener un resultado útil sin realizar las correcciones a la gran mayoría de frames de forma manual, razón por la cual se descartó el hablante directamente.

Una vez realizada la extracción del grado de apertura velofaríngea del resto de hablantes, se procedió a realizar una revisión de los datos segmentados. En principio, uno de los fenómenos observados en los datos, estuvo relacionado con la existencia de valores negativos en el grado de

apertura. Este hecho puede ser explicado por la naturaleza elástica del velo del paladar que, ante una apertura y cierre como la observada en la Figura 19, se curva por encima de la distancia máxima de referencia sobre la faríngе causando que el diferencial entre la distancia de referencia y cada uno de los puntos de medición sobre el velo, presente un valor negativo.

Figura 19. *Elasticidad del velo del paladar*

A pesar de que la existencia de estos valores puede ser considerada como un cierre del velo, no se pudo definir un valor fijo de 0 en el grado de apertura, ya que la gran cantidad de valores nulos implicaría una tendencia a este valor al entrenar la red neuronal, teniendo como consecuencia que el modelo de la red «aprendiera» a modelar 0's.

Al examinar los datos de manera más detallada, se pudo notar la existencia de cierto nivel de error en la gran mayoría de las frases, que no permitió realizar una visualización suficientemente clara de las aperturas y cierres del velo, para los hablantes F1 y F5. Un ejemplo de este fenómeno se puede visualizar en la Figura 20, donde se muestra una comparación entre los datos de los hablantes M2 y F5 al pronunciar la frase «Coconut cream pie makes a nice dessert».

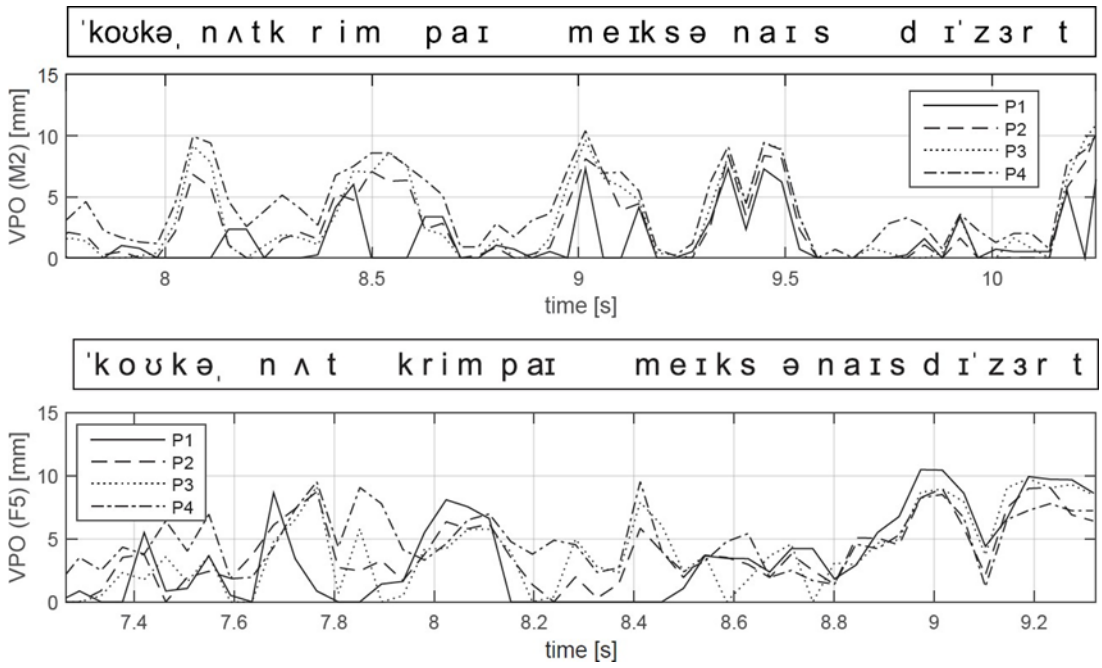


Figura 20. *Comparación entre el grado de apertura de los hablantes M2 y F5 al pronunciar la frase «Coconut cream pie makes a nice dessert».*

Tal como se puede observar, los datos asociados al hablante M2 presentan 4 aperturas y cierres muy bien definidos para las 4 consonantes nasales presentes en la oración, además de una apertura al final de la frase, probablemente relacionada con el descanso del velo. Por otro lado, el hablante F5 no mostró aperturas y cierres tan bien definidos, que, si bien en algunas ocasiones parece existir una apertura, como la observada alrededor de los 8 [s] para la consonante nasal /m/ asociada a la palabra «cream», son resultados con un nivel de error lo suficientemente alto como para reducir la calidad de los datos al entrenar la red. Una situación similar se pudo visualizar con el hablante F1, por lo cual se decidió descartar estos dos hablantes.

Si bien a cada hablante le fue aplicado el mismo procedimiento, existen ciertas diferencias en cuanto a las características de los datos extraídos, características que fueron estudiadas para optimizar el procedimiento a realizar. Una forma de analizar los datos fue mediante la medición y descripción de

los datos de cada uno de los puntos sobre el velo, por medio de las medidas de tendencia central y de dispersión.

Para el proceso realizado, se tomó como ejemplo ilustrativo los datos asociados al hablante F2 cuyas características pudieron ser descritas por medio de los valores observados en la Tabla 4, a partir de los cuales se pudo visualizar elementos como la media de los datos, que si bien toman valores muy pequeños respecto a las mediciones máximas, no son tan cercanos a 0 como se esperaría. Por otro lado, las mediciones mínimas permitieron evidenciar la existencia de valores negativos que fue mencionada anteriormente.

Tabla 4.

Resumen de valores estadísticos para los datos del hablante F2.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	4732	4732	4732	4732
Medición máxima [mm]	24.91	22.97	22.52	18.94
Medición mínima [mm]	-21.60	-20.07	-18.34	-20.92
Moda [mm]	-4.67	3.13	2.45	1.15
Mediana [mm]	-3.06	1.51	2.78	1.75
Media [mm]	-2.51	1.91	2.88	1.84
Desviación estándar [mm]	4.61	4.27	3.85	4.36
Varianza	21.25	18.29	14.87	19.03

Como adición de las medidas extraídas, la Figura 21 muestra el histograma de los datos de cada punto de medición, el cual mostró una distribución de tipo bimodal de los datos con una mayor cantidad de datos distribuidos cerca de 0 [mm] con ciertas diferencias que concuerdan con los valores obtenidos de la media y la moda, además de un conjunto más pequeño de datos cerca de 10 [mm], que están

relacionados con los espacios de tiempo en los cuales el velo se encuentra abierto. Todo este procedimiento, fue realizado de igual manera para cada uno de los hablantes, a partir del cual se extrajeron los datos que se pueden observar en la Tabla 2.5 y Tabla 2.6, sin embargo, si se desea realizar una revisión más detallada de los datos, es posible encontrar esta información en el Apéndice B.

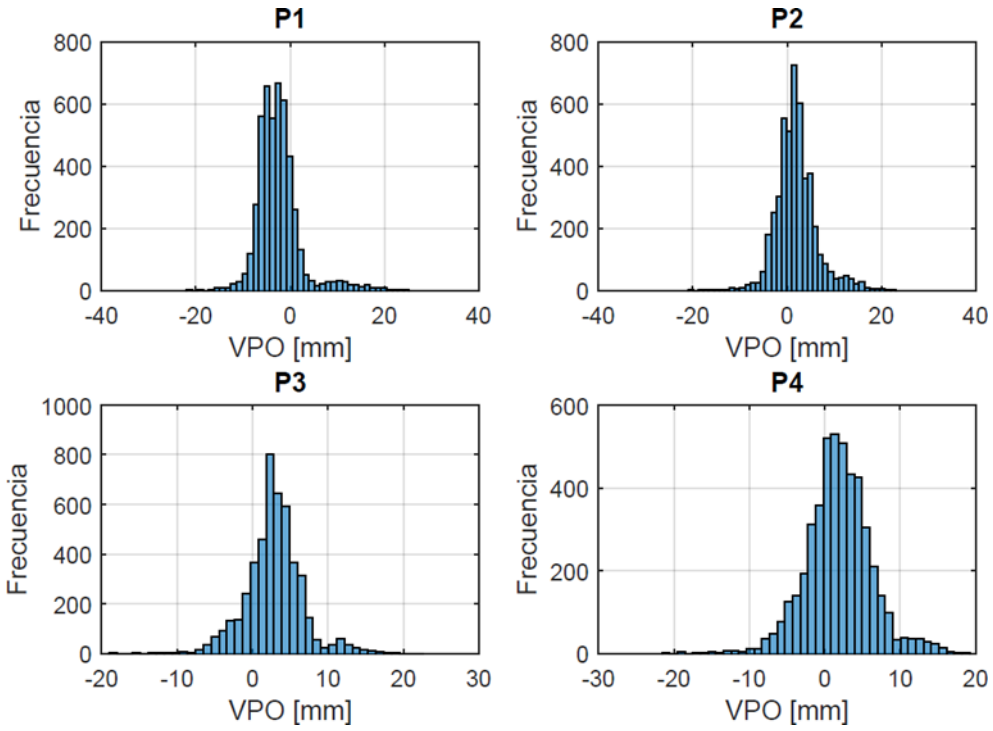


Figura 21. *Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.*

Si bien se realizó una revisión completa de los datos para todos los hablantes sobre los 4 puntos de medida, es muy importante tener en cuenta el grado de correlación que existe entre cada uno de estos, ya que en muchos instantes de tiempo existe cierto nivel de ruido que provoca aperturas o cierres para uno o varios puntos, que no se encuentran realmente correlacionados. En este sentido, lo que se esperó fue que ante una apertura y cierre velofaríngea, existiera un claro aumento seguido de una

reducción continua en el valor del VPO para entre 3 y 4 puntos en el mismo espacio de tiempo. Sin embargo, el hecho de tener 4 puntos implicó un grado de complejidad un tanto mayor, sobre el cual la estimación de alguno de estos podría ser afectada por el ruido presente en los demás. Por esta razón, se propuso realizar una reducción de la cantidad de puntos sobre el velo, en busca de reducir la probabilidad de un error de correlación entre los datos mencionados. Una forma de realizar esta separación fue midiendo el nivel de correlación lineal de Pearson, entre cada uno de los puntos de medición. Estos valores permitirían seleccionar 2 de los puntos que presentasen mayor correlación, de manera que pudieran ser utilizados como elementos principales para el posterior entrenamiento de la red.

Para realizar la extracción de los valores de correlación, se realizó una unión de todos los datos a utilizar, es decir de todos los hablantes, en una única matriz que serviría para obtener la correlación promedio entre los distintos puntos de medición. En este caso se utilizaron datos asociados a los hablantes F2, F3, F4, M1, M2, M3 y M4. Cada uno de los valores de correlación se encuentran en la Tabla 5.

Tabla 5.

Correlación promedio para la unión de todos los hablantes.

	P1	**P2**	**P3**	**P4**
P1	-	-	-	-
P2	0.56	-	-	-
P3	0.41	0.76	-	-
P4	0.42	0.69	0.88	-

A pesar de que la correlación obtenida entre los puntos 3 y 4 presentó un valor más alto, se pudo

observar que los datos del punto 4, al estar tan cercanos al eje de movimiento del velo, no presentaron valores de VPO cercanos al valor esperado del punto de medición 1. Por esta razón se propuso trabajar con los puntos 2 y 3 que presentaron el segundo grado de correlación más alto, mostrando aperturas en milímetros mucho más cercanas al valor esperado. Por otro lado, la baja correlación respecto al punto de medición 1 puede ser explicada por la gran cantidad de espacios de tiempo en los cuales la segmentación en este punto quedo por encima del velo, debido a que el punto se encontraba muy cercano al borde.

Finalmente, la Tabla 6 y la Tabla 7 muestran una recopilación de los valores estadísticos para todos los datos utilizados en el experimento, para mujeres y hombres respectivamente, que servirían para tener una descripción más detallada del mismo. En este caso solo se agregaron datos asociados a los puntos de medición seleccionados (2 y 3).

Tabla 6.

Recopilación de valores estadísticos para hablantes mujeres.

	F2		F3		F4	
Punto de Medición	**2**	**3**	**2**	**3**	**2**	**3**
Número de mediciones	4732	4732	3831	3831	4990	4990
Medición máxima [mm]	22.97	22.52	15.77	14.91	22.56	21.47
Medición mínima [mm]	-20.07	-18.34	-15.11	-14.31	-19.29	-13.84
Moda [mm]	3.31	2.45	-1.46	3.28	-2.05	-2.21
Mediana [mm]	1.51	2.78	1.90	2.23	-1.02	-0.70
Media [mm]	1.91	2.88	2.22	2.76	-0.07	0.73
Desviación estándar [mm]	4.27	3.85	5.43	4.34	5.24	5.44
Varianza	18.29	14.87	29.48	18.90	27.52	29.62

Tabla 7.

Recopilación de valores estadísticos para hablantes hombres.

	M1		M2		M3		M4	
Punto de Medición	**2**	**3**	**2**	**3**	**2**	**3**	**2**	**3**
Número de mediciones	3743	3743	5236	5236	4693	4693	5179	5179
Medición máxima [mm]	22.60	32.18	14.45	14.01	21.52	21.22	19.52	18.92
Medición mínima [mm]	-22.62	-29.67	-9.38	-10.64	-26.11	-30.68	-16.32	-11.89
Moda [mm]	-0.58	-4.55	-1.06	0.39	2.73	-3.52	4.30	4.05
Mediana [mm]	-0.12	-2.27	0.04	0.36	2.20	-2.18	5.49	5.65
Media [mm]	0.63	-1.26	0.96	1.21	1.84	-1.61	5.36	5.78
Desviación estándar [mm]	4.51	5.10	3.44	4.04	4.93	5.07	3.94	3.84
Varianza	20.38	26.06	11.85	16.38	24.29	25.79	15.56	14.80

3.3.2 Segmentación manual. El velo del paladar es un articulador muy importante a la hora de generar fonemas del tipo nasal; sin embargo, durante el proceso de habla continua, no es esta condición la única que provoca la apertura velofaríngea. En particular, esta apertura se puede dar en otros casos tales como en aquellos instantes de relajación del proceso del habla o de respiración, entre otros [34]. Por tanto, el análisis se centra en la vecindad temporal de los fonemas nasales. Es importante recordar el fenómeno de coarticulación, el cual, debido a la parsimonia de los articuladores, la influencia de estos puede extenderse un poco más allá del inicio y fin de cada fonema. En el presente trabajo se toman tres fonemas: el fonema inmediatamente anterior, el fonema nasal y el fonema siguiente.

Para realizar esta tarea se diseñó, por parte de los autores, una interfaz gráfica de usuario

en Matlab que permitiese seleccionar los espacios de tiempo correspondientes y guardarlos en una matriz, para ser usados posteriormente (ver Figura 22).

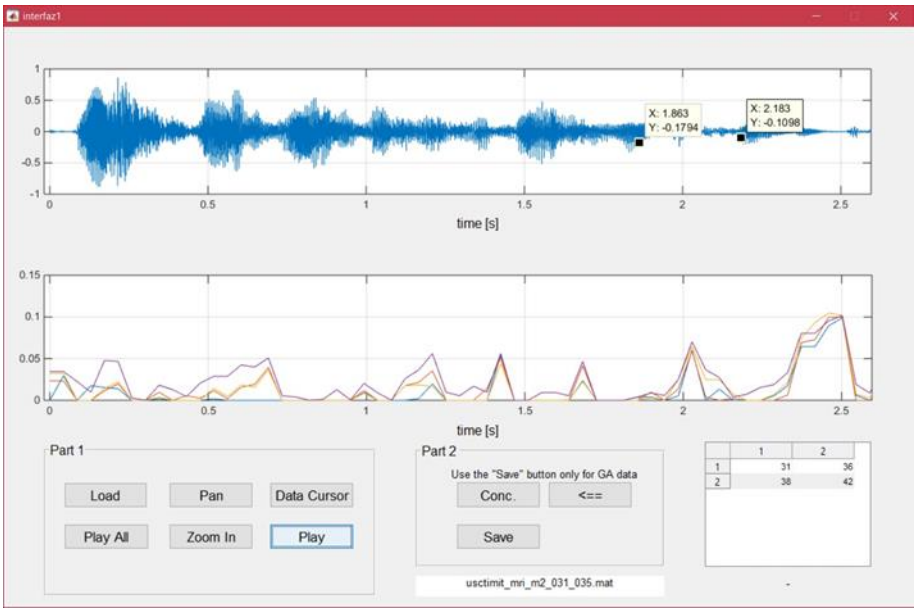


Figura 22. *Interfaz gráfica de usuario para la selección de fonemas nasales en cada vídeo.*

La separación fue realizada como se observa en la Figura 23, donde se muestra el VPO obtenido en la frase *"Flying standby can be practical if you want to save money"*, para el hablante M2. Como se puede observar, la sección marcada en rojo muestra la nasal /n/ de la palabra "want" entre 9,7 y 10,2 [s], donde se observa una apertura velofaríngea cercana a los 14 [mm]. La idea fue tomar los fonemas anterior y posterior al fonema nasal, con el fin de observar la apertura y cierre del velo.

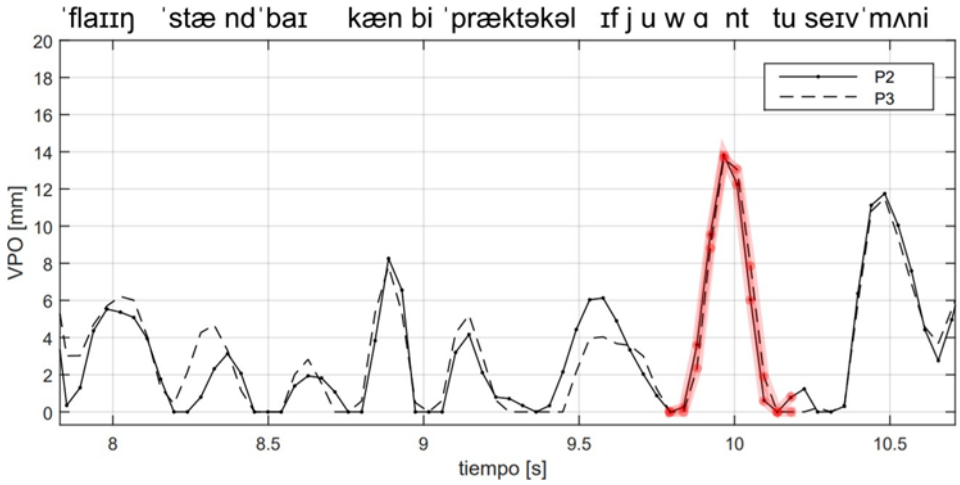


Figura 23. *VPO de la frase "Flying standby can be practical if you want to save money" para el hablante M2.*

3.3.3 Detección de fallas de sincronización entre audios y vídeos MRI. Durante el proceso de segmentación se pudo observar la presencia de retrasos del audio respecto al video, en algunas grabaciones. En la Figura 24 se muestra un ejemplo de este fenómeno; en donde, la primera fila corresponde la secuencia de imágenes MRI, la segunda es el sonograma y la tercera es la estimación de la apertura velofaríngea. Todas estas correspondientes a al rango de tiempos 8,92 a 9,51 [s] dentro del registro 3, del hablante F4, al pronunciar la frase "The museum hires musicians every evening". Se puede observar que, aunque no se está emitiendo voz, el hablante F4 está ya articulando. Además, nótese la apertura del velo en instantes de tiempo que no corresponden con el fonema.

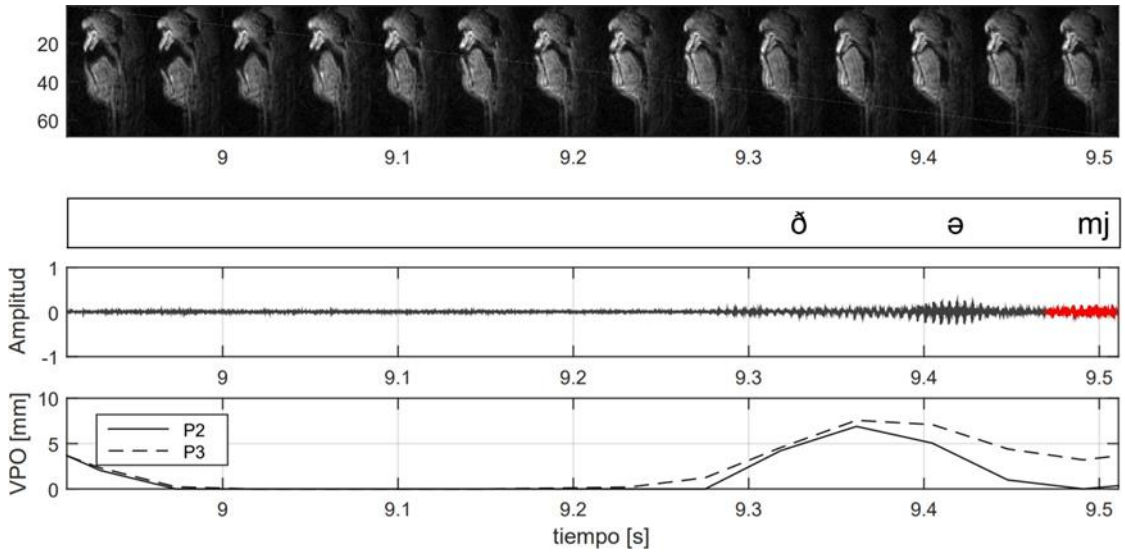


Figura 24. *Mala sincronización del vídeo y la voz para algunos vídeos de la base de datos (la sección en rojo indica la región donde debería estar la consonante nasal).*

Si bien el fenómeno pudo ser observado en varios de los vídeos de algunos de los hablantes, en muchos casos no se presentó en todo el vídeo. Por lo general el fenómeno ocurrió en instantes aleatorios para cada vídeo, por lo cual algunas secciones de un vídeo en particular no mostraron dicho problema, permitiendo su uso para la selección de nasales.

3.3.4 Filtrado de la información articulatoria. En [35] [36] se reporta que las señales articulatorias son del tipo pasa-bajas. En particular, en [36], basados en la base de datos MOCHA-TIMIT, se obtiene que en promedio el 95 % de la energía está por debajo de los 6.5 Hz. Por lo tanto, un valor de 23.18 FPS resultó adecuado para el muestreo del movimiento del velo del paladar. Aún así, se observaron cambios abruptos en la posición del velo entre fotogramas consecutivos y por consiguiente, cambios bruscos en la estimación del VPO. Este fenómeno podría estar causado por características propias de los elementos de grabación, tales como resolución y ruido. A esto hay que sumarle la presencia de errores de segmentación en uno o varios instantes particulares.

Un ejemplo de este fenómeno se observa en la Figura 25, donde se puede ver un error de segmentación en un instante de tiempo donde se observa el VPO de una parte de la frase *"Swing your arm as high as you can"*. En este caso se observó el instante de pronunciación de las palabras "*your arm as"*, donde se alcanzó un VPO cercano a los 10 [mm].

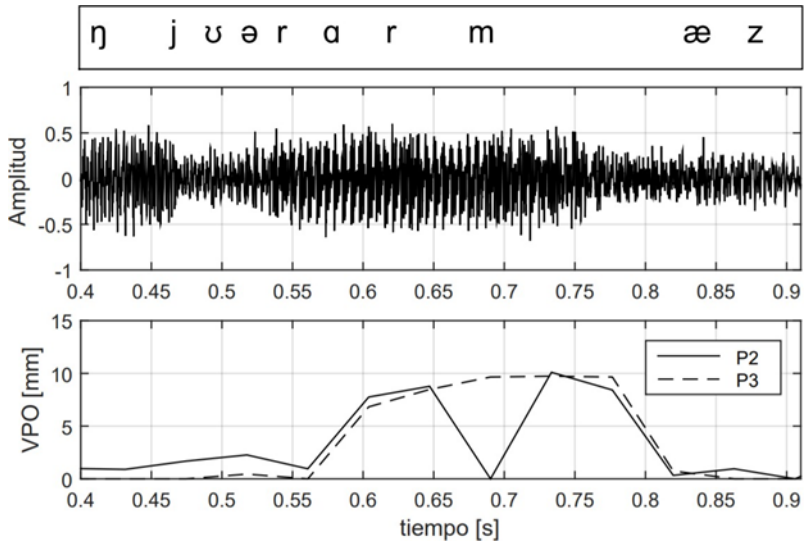


Figura 25. *VPO de las palabras "your arm as" de la frase "Swing your arm as high as you can".*

Una forma de reducir este tipo de errores fue mediante un proceso de filtrado paso-bajo de los datos a aproximadamente 6.5 [Hz], ver Figura 26.

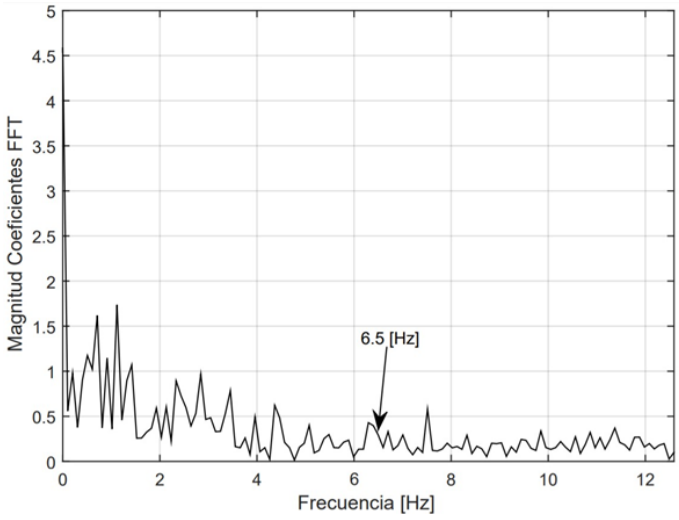


Figura 26. *FFT del VPO para uno de los hablantes. La frecuencia de corte para el filtro paso-bajo fue seleccionada en 6.5 [Hz]*.

De esta forma se obtiene un suavizado de la señal que se puede observar en la Figura 27, para el mismo espacio de tiempo. Razón por la cual se realizó la aplicación de este filtrado a todos los datos de VPO medidos.

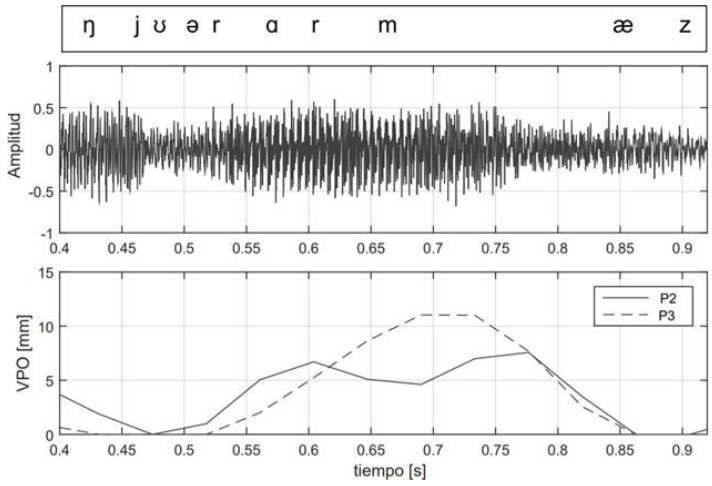


Figura 27. *VPO de las palabras "your arm as" de la frase "Swing your arm as high as you can"* luego de filtrar.

3.3.5 Errores de grabación en la voz. Una de las posibles fuentes de error es la existencia de eco, ruido de ambiente o ruido ocasionado por algún tipo de distorsión cercano, al momento de la creación de la base de datos. Si bien las distorsiones externas no pueden ser trabajadas fácilmente, no son situaciones que aparezcan con mucha frecuencia, razón por la cual solo se descartaron los muy pocos espacios de tiempo donde existe algún tipo de distorsión como el ruido de la máquina de resonancia magnética o perdida de información de grabación por el micrófono de fibra óptica.

3.3.6 Presencia de ruido en imágenes MRI. La presencia de ruido es una característica inherente a las imágenes y videos de resonancia magnética y el análisis de su incidencia sobre la información de interés supone un área de estudio de gran relevancia [37] [38] [39], por lo que es importante conocer en qué medida el ruido afecta la percepción de la imagen y la precisión de las mediciones que son un factor trascendental en el desarrollo del presente proyecto.

Para el análisis del ruido se midió físicamente la sensibilidad de una imagen respecto al ruido que esta presenta [40] [41]. El objetivo de este análisis no fue medir la calidad o el rendimiento de las imágenes con respecto al ruido sino demostrar la presencia de este en los frames de los videos.

La principal dificultad experimentada en el análisis de ruido se encuentra en que la base de datos no cuenta con una imagen de referencia que permita comparar las demás para realizar una estimación del ruido, puesto que se asumió que todas las imágenes al ser tomadas con el mismo equipo y bajo las mismas condiciones se vieron afectadas (En mayor o menor medida) por la presencia de ruido.

El análisis de ruido en imágenes se realizó a través de la función noiselevel [42] que permitió determinar el nivel de ruido presente en una sola imagen en función del cálculo preciso de la desviación estándar en términos de un único fotograma a través del análisis de componentes principales PCA, sin necesidad de la aplicación de aproximaciones basadas en filtros o enfoques

estadísticos. La función noiselevel selecciona parches de bajo rango descartando componentes de alta frecuencia en la imagen a través de las características representadas en los gradientes de cada parche, luego estima el nivel de ruido de los parches usando análisis de componentes principales PCA tal como se expone en [43] [44]. El resultado consiste en un valor de desviación estándar calculado directamente de la imagen en función de los parches y del análisis de componentes principales que permiten establecer el nivel de ruido presente en cada fotograma MRI en términos de la variación de intensidad aleatoria de los pixeles.

En la Figura 28 se muestra el grafico del nivel de ruido en función de cada frame para el hablante femenino número 4 en el registro 3.

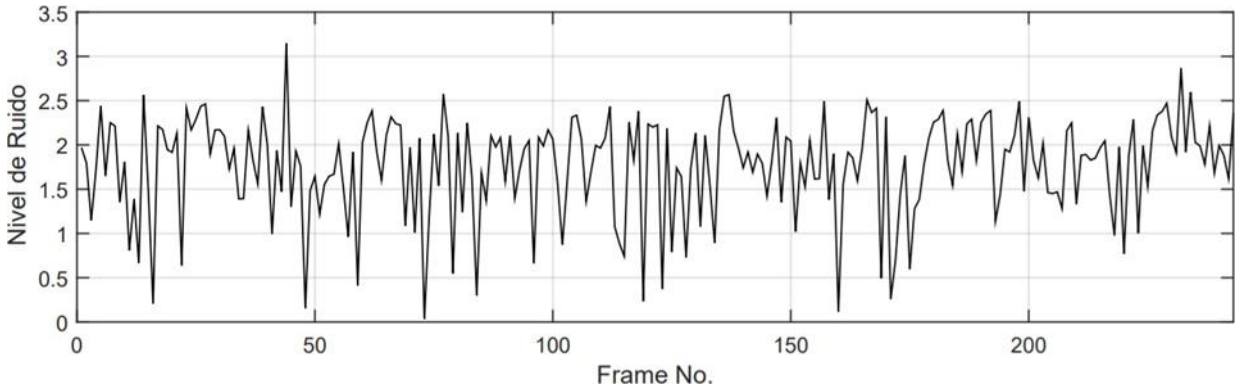


Figura 28. *Nivel de ruido presente en las imágenes MRI del registro 3 para el hablante femenino*

Se observa en la Figura 28 que el *frame* número 44 del hablante femenino 4 tiene un alto nivel de ruido, mientras que el *frame* número 49 muestra un nivel de ruido bajo con respecto al resto de del grupo.

En la Figura 29 se muestran los *frames* 44 y 49 respectivamente. Se observa tal como el análisis de ruido mostró que el *frame* 44 (Figura 29-a) tiene una alta presencia de ruido distribuido por toda la imagen. El *frame* número 49 (Figura 29-b) muestra por su parte una menor presencia de

ruido, en este caso se resaltan mejor los límites de intensidad para observar de manera más clara la estructura física y anatómica del tracto vocal del hablante. Un punto de comparación contundente corresponde al fondo de la imagen, cuyos valores de intensidad deben estar cerca del color negro, en el *frame* 49 se observa un fondo negro uniforme mientras que el fondo del *frame* 44 toma valores en un rango más amplio alcanzando incluso valores de intensidad cercanos a los que ilustran el cráneo y el cuello, disminuyendo la precisión en los límites y contornos de la imagen.

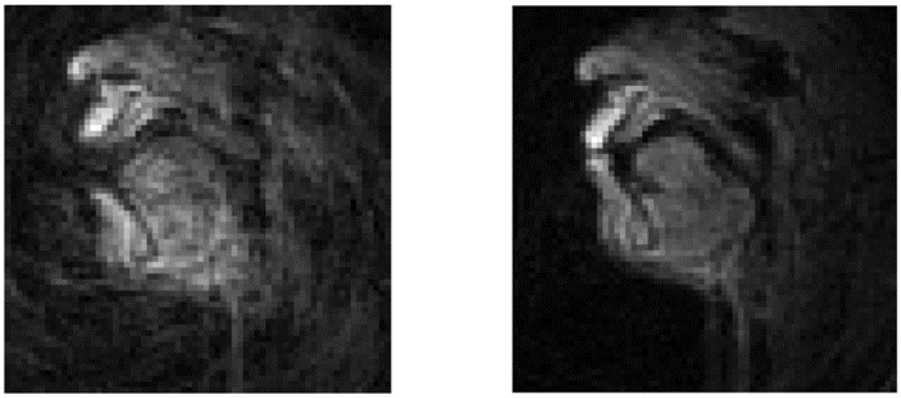


Figura 29. *Comparación del nivel de ruido presente en los fotogramas 44 (a) y 49 (b) del registro 3 para F4.*

El velo del paladar en el *frame* 44 no se observa con claridad y debido a los altos niveles de ruido son asociadas a él características que distorsionan su fisionomía, se observa un velo rígido y considerablemente más robusto con respecto a los demás *frames*, además su posición y movimiento tienen un grado de incertidumbre más alto. Para el *frame* 49 se observa claramente el contorno del velo, así como son ilustradas de manera más precisa sus características morfológicas naturales.

En la Figura 30 se ilustran los histogramas de los *frames* 44 (Figura 30-a) y 49 (Figura 30-b) respectivamente, la distribución de intensidad para el *frame* 44 muestra una tendencia hacia los

valores de intensidad medios como los diferentes tonos de gris, la presencia de ruido altera las características de intensidad de los pixeles, para este caso la desviación estándar se aleja del valor promedio de intensidad por lo que se encuentra una gran cantidad de valores atípicos en la imagen. Para el *frame* 49 la distribución es más compacta, los valores de intensidad negros son más frecuentes debido a la gran cantidad de pixeles oscuros y similares entre ellos, el valor de La desviación estándar es baja y toma valores cercanos al valor promedio de intensidad de toda la imagen por lo que se observa una imagen más nítida.

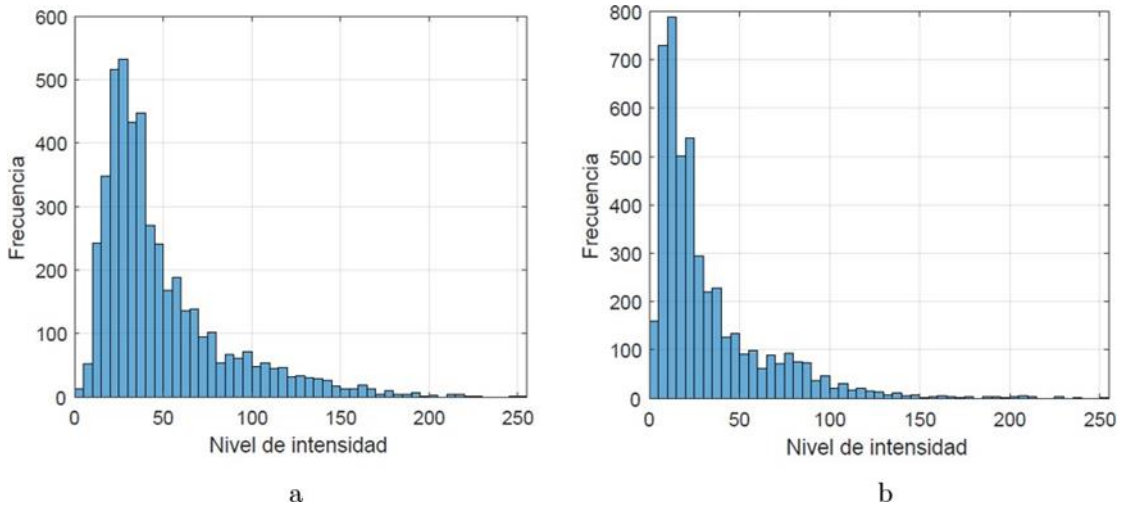


Figura 30. *Histogramas de los fotogramas 44 (a) y 49 (b).*

El nivel de ruido expresado en términos de la desviación estándar calculada a partir del análisis de componentes principales fue para todos los hablantes en promedio de 1.76. La tasa de ruido estableció un factor de incertidumbre directo para el desarrollo del proyecto ya que el grado de apertura se extrajo a partir de las imágenes de resonancia magnética. El velo del paladar es un órgano delgado y pequeño en comparación con el tracto vocal del hablante, estas características ligadas a la presencia del ruido anteriormente descrito generaron errores en gran cantidad de casos

de segmentación debido a que las características físicas del velo se ven distorsionadas. Un ejemplo de cómo la presencia de ruido puede afectar la percepción e interpretación de las imágenes de resonancia magnética se ilustra en la Figura 31 donde se muestra el *frame* número 132 del registro 3 del hablante femenino 4 (F4), los niveles de intensidad alrededor del área de interés son heterogéneos a tal punto que se observa el velo como un órgano desprendido de la cresta alveolar del hablante.

Figura 31. *Ejemplo de un fotograma afectado por el ruido y la calidad de la imagen.*

3.4 Mapeo acústico-articulatorio

Ya que para realizar la estimación de la función de mapeo, se hace necesario realizar una unión de todos los datos, se desarrolló una función sencilla que toma la información extraída de cada uno de los audios y vídeos, por hablante, concatenándolos tal como se puede observar en la Figura 32.

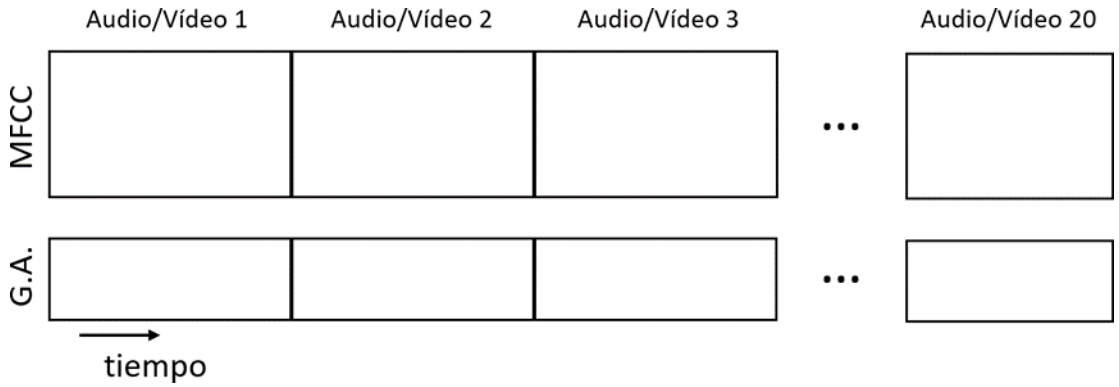


Figura 32. *Unión de datos realizada para cada hablante.*

Cada conjunto de datos, asociado a un hablante particular fue guardado por separado, de forma tal que los datos pudieran ser utilizados para el entrenamiento de la red neuronal en cualquier orden deseado. No obstante, las diferentes mallas de referencia usadas por hablante, generaron en muchas ocasiones niveles de offset diferentes para cada uno, dificultando el entrenamiento de la red neuronal. Por tanto, se realizó una normalización por medio de la Ecuación 2.9, buscando llevar el valor promedio de los datos hasta 0 y la varianza hasta 1 para todos los hablantes por separado, y así poder comparar los datos de cada hablante con el resto.

Los datos de distintos hablantes fueron concatenados teniendo en cuenta el orden de los segmentos de medición sobre el velo, mostrados el gráfico del lado izquierdo en la Figura 33, pasando a ser los 2 conjuntos de datos que se pueden observar en el gráfico del lado derecho.

$$z = \frac{x - \mu}{\sigma} \qquad (2.9)$$

Donde z es la representación en unidades estandarizadas de x, μ es la media y σ es la desviación típica. Cabe mencionar que los valores de μ y σ fueron extraídos antes de realizar la segmentación

manual por consonantes nasales.

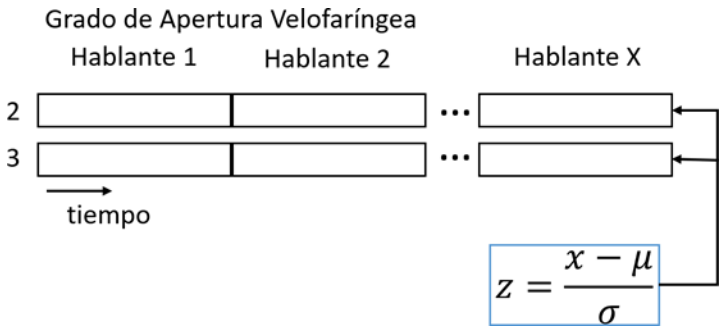


Figura 33. *Orden de los 2 vectores de datos estandarizados para X hablantes.*

3.4.1 Estimación de la función de mapeo, mediante el uso de Redes Neuronales. En los últimos años el estudio de las redes neuronales artificiales ha tenido una gran importancia como tecnología en el campo de la inteligencia artificial, para el reconocimiento de patrones y aprendizaje automático [45]. Pueden definirse como un modelo simplificado inspirado en el modo como el cerebro humano procesa la información, compuesto por un gran número de elementos in- terconectados que emulan una versión abstracta de neuronas [46]. Las unidades de procesamiento se conectan con ponderaciones y se encuentran organizadas en capas: una capa de entrada, una o varias capas ocultas y un capa de salida. Los datos de entrada que se encuentran en la primera capa se propagan por cada neurona hacia la capa siguiente, para finalmente enviar un resultado desde la capa de salida. En el proceso de entrenamiento la red examina los registros y genera una predicción para cada uno, modificando las características cuando la predicción no es correcta, proceso que se repite varias veces, obteniendo cada vez mejores resultados en sus predicciones. Una vez finalizado el entrenamiento, es posible hacer uso de datos diferentes, que permitan obtener una validación de la red [47].

El entrenamiento de la red o proceso de aprendizaje puede ser supervisado o no supervisado. Para el desarrollo del proyecto, se hizo uso del aprendizaje supervisado, donde la red se entrena partiendo

de un conjunto de patrones de entrenamiento compuesto por datos de entrada y salida. Por tanto, el objetivo del algoritmo de aprendizaje consiste en realizar ajustes para que los datos estimados y los datos objetivo sean lo más cercanos posible, encontrando un modelo de los procesos llevados a cabo, para generar una salida determinada. En el aprendizaje supervisado, el patrón de salida cumple el papel de supervisor de la red [48].

Una herramienta sumamente útil para la creación de redes neuronales en MATLAB es Deep Learning Toolbox, que permite el diseño y la implementación de redes neuronales profundas con algoritmos, modelos preentrenados y aplicaciones. Este paquete permite el uso de redes neuronales convolucionales, redes de memoria a largo o corto plazo para la clasificación y regresión de imágenes, series de tiempo y datos de texto. La interfaz cuenta con gráficos que facilitan la visualización de las activaciones, edición de arquitecturas de red y el monitoreo del progreso de la capacitación, facilitando el entrenamiento para conjuntos de datos grandes o pequeños [49].

Como se mencionó anteriormente, para obtener una función que permita obtener el mapeo desde el dominio de los parámetros acústicos hacia el de los parámetros articulatorios, se hizo uso del entrenamiento supervisado de redes neuronales, tomando como datos de entrada los MFCC y como datos objetivo el grado de apertura velofaríngea. En este caso, el uso de una red neuronal prealimentada (feed-forward backdrop) representaría una opción viable, permitiendo obtener los datos del dominio articulatorio solo a partir de los datos de entrada. La Figura 34 muestra un ejemplo de este tipo de red con una capa oculta cuya función de activación es de tipo sigmoidal y una capa de salida con función de activación de tipo lineal.

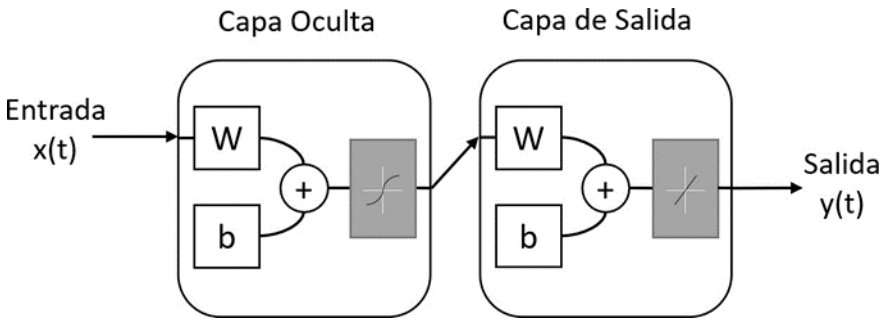


Figura 34. *Red neuronal prealimentada (Feed-forward backdrop).*

La herramienta de MATLAB que permite hacer uso de este tipo de red es «nftool» [50], la cual permite asignar para cada conjunto de entradas numéricas, un conjunto de objetivos numéricos. Además de evaluar el desempeño de la red, por medio del error cuadrático medio y el coeficiente de correlación lineal de Pearson.

Figura 35. *Neural fitting app («nftool»).*

La Toolbox permite seleccionar las variables para entrenamiento, definiendo para cada conjunto de

n de muestras de entrada, *n* muestras objetivo, como se puede observar en Figura 36, además de definir el porcentaje de los datos que será destinado a entrenamiento, validación y prueba, ver Figura 37.

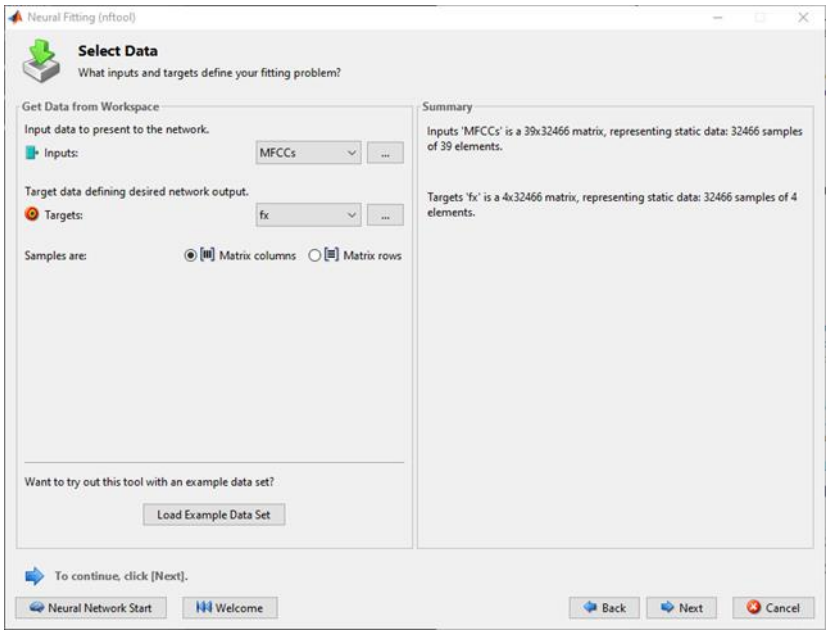


Figura 36. *Selección de variables de entrada y objetivo.*

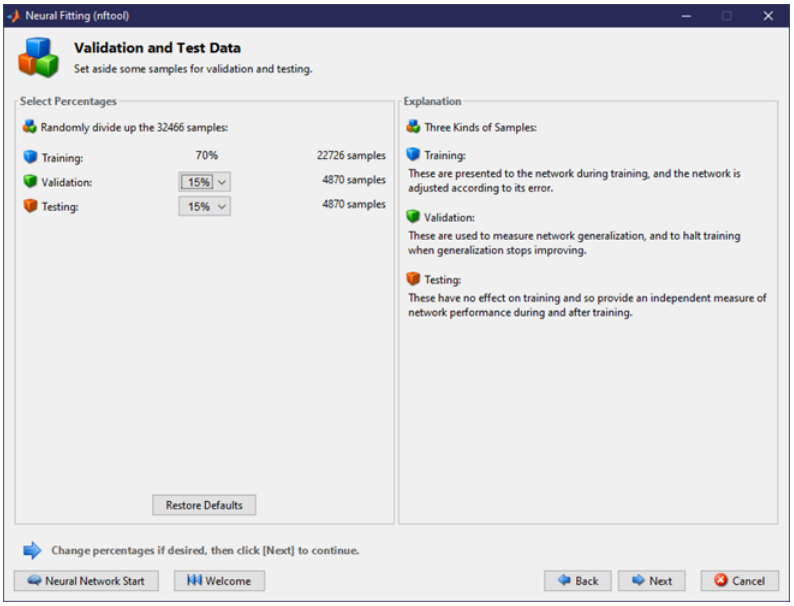


Figura 37. *Porcentajes destinados a entrenamiento, validación y prueba.*

A continuación, el software permite definir el número de neuronas a utilizar en la capa oculta Figura 38. Cabe mencionar que para el desarrollo del proyecto se tomó como base la utilización de 10 neuronas en solo una capa oculta. Esto sería modificado posteriormente con el fin de mejorar los resultados obtenidos al entrenar la red.

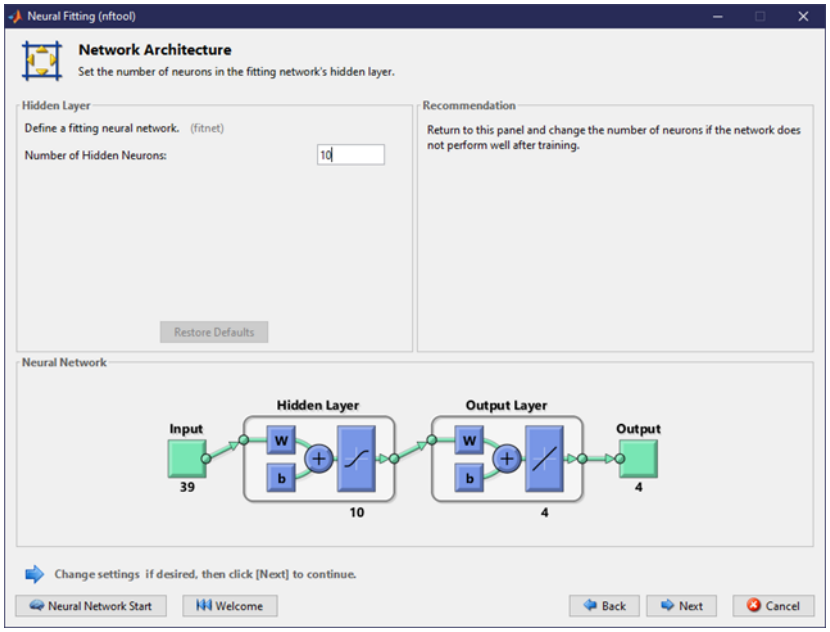


Figura 38. *Definición del número de neuronas en la capa oculta y el retardo.*

Finalmente, la herramienta permite seleccionar entre 3 algoritmos de entrenamiento: Levenberg-Marquardt, Bayesian Regularization y Scaled Conjugate Gradient. Cada uno de estos, presenta ciertas ventajas respecto a los demás, para resolver distintos tipos de problemas. Sin embargo, y sin entrar mucho en detalles, se ha utilizado el «Levenberg-Marquardt» por su velocidad a coste de un consumo ligeramente mayor de memoria RAM, obteniendo muy buenos resultados en una menor cantidad de iteraciones.

3.4.1 Entrenamiento de la Red Neuronal. La Toolbox posee una interfaz que, al realizar

el entrenamiento de la red, permite observar en tiempo real el MSE, la reducción del gradiente y el coeficiente de correlación de Pearson para cada iteración, haciendo uso de los datos de entrenamiento, validación y prueba previamente definidos, ver Figura 39.

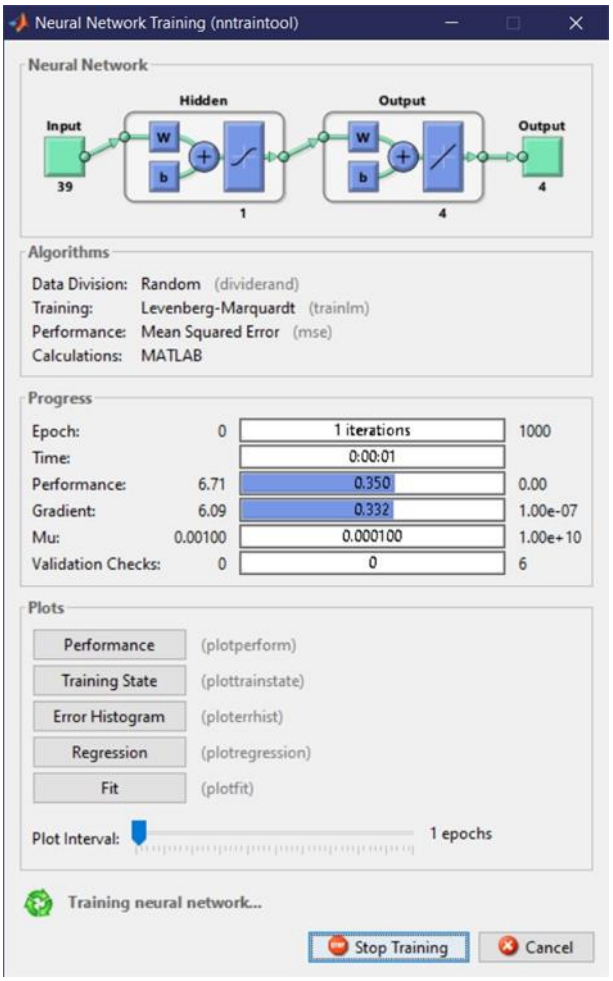


Figura 39. *Interfaz de entrenamiento de la red.*

Ya que toda la configuración del software para el entrenamiento de la red, también puede ser realizada por medio de funciones en la línea de comandos de MATLAB, se decidió trabajar con el código disponible en [51], con el fin de obtener resultados para distintos experimentos de forma rápida.

Para la obtención de resultados se realizaron dos tipos de experimentos. El primer tipo se centró en el entrenamiento intra-hablante, haciendo uso de aproximadamente el 70% de los datos para entrenamiento y el 30% restante para validación, que serviría para medir el grado de eficiencia que podía ser obtenido con cada uno de los hablantes en cuestión. Por otro lado, el segundo experimento se centró en el entrenamiento de la red con 7 diferentes configuraciones, seleccionadas respecto a un hablante escogido para validación, tal como se observa en la Figura 40.

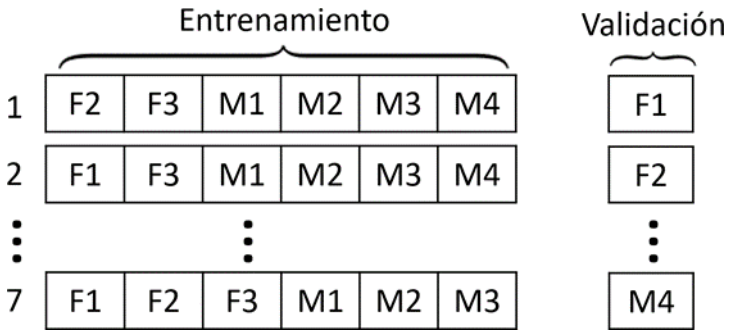


Figura 40. *7 configuraciones de hablantes para entrenamiento y validación.*

Finalmente, cada uno de los resultados de validación estimados a partir de la red le fue aplicado un filtrado paso-bajo, variando la frecuencia de corte en el rango de 1 a 11 [Hz], en busca de obtener para cada una de las configuraciones una curva asociada al error cuadrático medio (Ecuación 2.10) y una asociada coeficiente de correlación lineal de Pearson (Ecuación 2.11) en cada frecuencia.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y'_i - Y_i)^2 \qquad (2.10)$$

Donde Y'_i son los valores estimados y Y_i los medidos, con una cantidad n de muestras.

$$R = \frac{\sigma_{YY'}}{\sigma_Y \sigma_{Y'}} \qquad (2.11)$$

Donde σ_{YYJ} es la covarianza entre los valores medidos y estimados, σ_Y es la desviación estándar de los valores medidos y σ_{YJ} la desviación estándar de los valores estimados.

4. Pruebas y análisis de resultados

En el estado del arte existen variadas técnicas que permiten el modelado de fenómenos biológicos a partir de datos, lo cual se implica tener en cuenta modelos no-lineales. En el presente trabajo el objetivo primordial es indagar si a partir del espectro de la señal de voz es posible inferir la posición relativa del velo del paladar, y observar hasta qué punto se puede estimar este. Hemos seleccionado la técnica de redes neuronales por su grado de madurez, facilidad y su capacidad para representar fenómenos no lineales. El método de modelado por redes neuronales fue descrito en la sección 2.4.1, para el cual el dominio acústico se representa mediante coeficientes MFCC, y, la dinámica de movimiento del velo del paladar representado por el grado de apertura extraído a partir de las imágenes MRI.

A groso modo, se realizan dos tipos de experimentos: intra-hablantes e inter-hablantes. En el experimento intra-hablantes se busca determinar si existen posibles relaciones entre el velo del paladar y el espectro suavizado del tracto vocal, por hablante, asumiendo que esta relación puede variar de persona a persona debido a que son fenómenos de tipo biológico. De otra parte está el experimento inter-hablantes, el cual busca determinar alguna relación entre el velo y la acústica, asumiendo que existen elementos comunes en la señal de voz de persona a persona, que permiten inferir el movimiento relativo del velo del paladar. Este grado relativo lo representamos mediante el grado de

apertura velofaríngea expuesto en la sección 2.1.2.

El velo del paladar es un articulador crítico para la producción de sonidos del tipo nasal, mientras que para otros sonidos su función es secundaria. Las características morfológicas y biológicas de cada persona inciden directamente en el modo en que se articulan los fonemas, especialmente si la función del articulador es secundaria. Estas características motivaron a la realización de un experimento para cada hablante de manera que se pudiera desarrollar un modelo con estructura estándar que permitiera resaltar las propiedades que fortalecen o debilitan la relación entre el grado de apertura del velo y las propiedades de la voz.

En particular, en el presente trabajo se utilizaron redes neuronales de tipo feedforward, la cual corresponde a una configuración simple y adecuada para problemas de mapeo. En Matlab, para obtener un modelo basado en ANN se realizan los siguientes pasos: 1) Establecer el tipo de red neuronal que se va a usar; 2) Ingresar los datos de entrenamiento y establecer la variable dependiente (Grado de apertura) y la variable predictora (MFCC); y, 3) Determinar el número de capas ocultas y neuronas por capa. A modo de criterio de desempeño de los modelos se utiliza el error cuadrático medio (*Mean Square Error*, MSE) y la correlación de Pearson, calculados sobre datos de validación no involucrados en la etapa de entrenamiento.

4.1 Experimento intrahablante

El experimento intra hablante consistió en la realización de un modelo de red neuronal para cada hablante de modo que se lograra observar el comportamiento de los datos de cada persona considerando las características propias que esta posee. La distribución de muestras usadas para los procesos de entrenamiento y validación se muestran en la Tabla 8.

Tabla 8.

Disposición de muestras para entrenamiento y validación de los hablantes.

Hablante	F2	F3	F4	M1	M2	M3	M4
No. muestras de entrenamiento	1060	605	790	580	1245	1158	1270
No. muestras de validación	455	260	339	248	533	497	545
Total	1515	865	1129	828	1778	1655	1815

Para este análisis se utilizó una ANN de una sola capa intermedia y se tuvo cuidado de seleccionar un valor no demasiado grande de cantidad de neuronas en esta capa a fin de no aumentar la complejidad del modelo inadecuadamente. El rango óptimo de neuronas para cada experimento intra hablante osciló en un valor entre uno y cinco.

Se realizó una serie de modelos para cada hablante modificando el número de neuronas entre uno y cinco y se tomó aquel que mejores resultados obtuvo en términos de correlación y error cuadrático medio, finalmente con el objetivo de estandarizar el experimento, se seleccionó como valor estándar para todos los hablantes la moda entre estos valores, la cual fue de 3 neuronas. Con esta estructura general se realizaron los siete experimentos correspondientes a cada hablante.

Los resultados de correlación y MSE entre los valores medidos y estimados de cada hablantese muestran en la Tabla 9.

Tabla 9.

Resultados de correlación y error cuadrático medio para cada hablante.

Hablante	F2	F3	F4	M1	M2	M3	M4
Correlación P2	0.18	0.26	0.21	0.38	0.42	0.5	0.25
Correlación P3	0.3	0.23	0.32	0.39	0.41	0.3	0.28

Tabla 9. Continuación

MSE [mm]	5.46	5.5	3.55	4.1	4	2.73	9.4
Frecuecia filtrado [Hz]	4.8	3.4	4.2	3.2	4.2	5.1	5.2

Los valores de error cuadrático medio MSE y correlación de Pearson fueron observados en función de la frecuencia de corte para un filtro pasa bajas de manera que se determinase una frecuencia optima donde los resultados estimados tuvieran el mejor desempeño. El valor de frecuencia de filtrado según [35] debe estar comprendido entre 4.5 y 5 [Hz]. Este valor de fue determinado de manera individual para cada hablante.

El experimento intrahablante posibilitó la visualización de características particulares de cada hablante a través de modelos independientes enfocados en cada persona de la base de datos, al seleccionar un valor de neuronas estándar para la realización de todos los experimentos se buscó generar un entorno de análisis homogéneo que permitiese determinar las características de los hablantes que mejor se ajustan a los modelos propuestos. Este planteamiento permitió observar cómo entre los hablantes femeninos el mejor resultado obtenido fue a partir de los datos de F4, sobre los cuales el modelo identificó en mayor medida tanto aperturas como cierres del velo y cuyas estimaciones estuvieron más cerca del valor medido, respecto a los hablantes F2 y F3, evidenciando una relación fuerte entre el VPO y los MFCC. Por otro lado, la baja correlación y un grado de error considerablemente mayor en los hablantes F2 y F3, reveló la pobre relación existente entre los MFCC y el VPO de estos hablantes en particular.

Para el caso de los hablantes masculinos, M3 y M2 obtuvieron respectivamente los resultados con mejor eficiencia para el experimento intrahablante en términos de correlación y MSE, acertando en la identificación tanto de aperturas como cierres y obteniendo mejores estimaciones que los demás hablantes.

Los resultados de validación del hablante M2 se muestran en la Figura 41 donde se observa el comportamiento de los valores estimados con respecto a los medios para los dos puntos de referencia del velo, se observa que el modelo logra estimar de manera acertada las aperturas del velo presentes durante las articulaciones del tipo nasal, así como sus correspondientes cierres cuando se está articulando un sonido vocal, no obstante, los valores de estimación de las distancias de apertura tuvieron un de error considerable, cuando la apertura es del orden de 5 mm el modelo logra estimar con alta precisión la distancia, arrojando valores cercanos a los medidos, sin embargo, para aperturas superiores a 5 [mm] la diferencia entre los valores estimados y medidos tiende a ser más alta. Este fenómeno fue observado en todos los hablantes y se vio reflejado en el cálculo de error cuadrático medio.

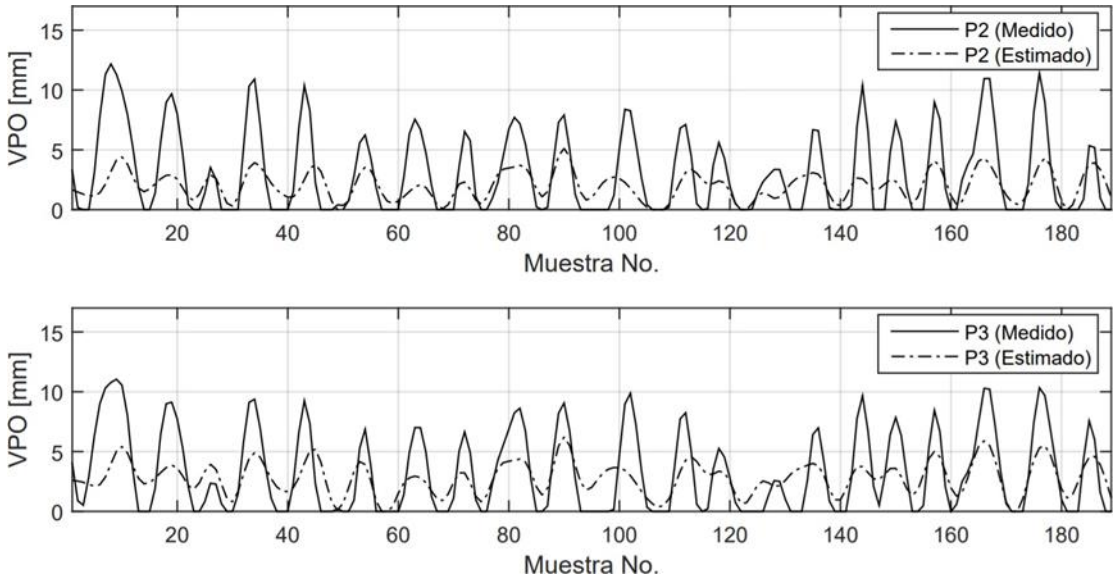


Figura 41. *Datos medidos y estimados para el grado de apertura del hablante F2*

Los gráficos de valores medidos y estimados para los demás hablantes del experimento se muestran en el Apéndice C.

En la Figura 42 se ilustran las gráficas de MSE y correlación de Pearson en función de la frecuencia de corte para el hablante M2, notase que este hablante presentó una correlación similar para los dos puntos de referencia del velo, evidenciando una fuerte relación entre el VPO y los MFCC además de obtener el mejor desempeño para la frecuencia de filtrado de 4 [Hz].

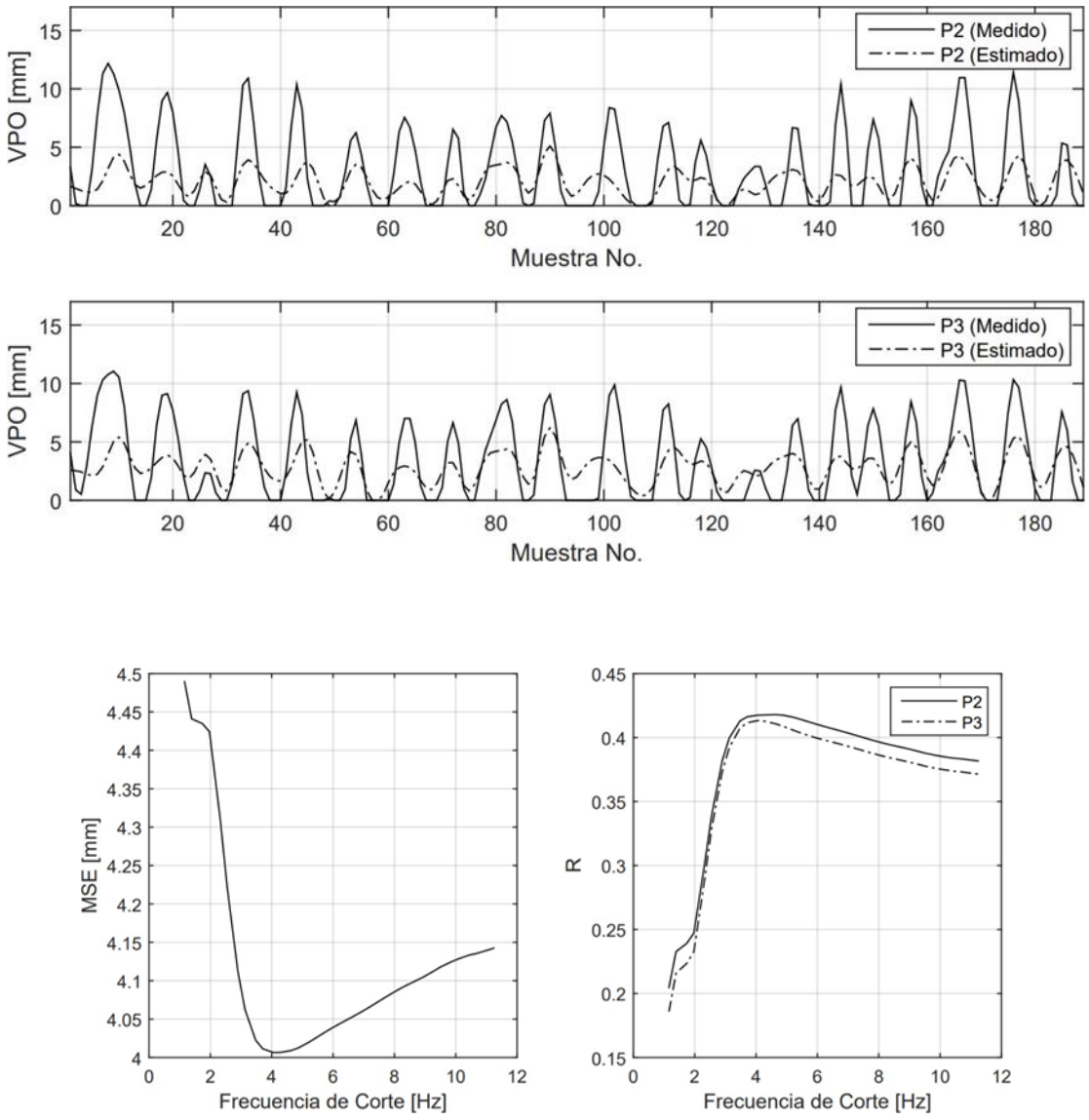


Figura 42. *MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante M2.*

Los gráficos de Error cuadradico medio MSE y correlación de Pearson en función de la frecuencia de corte del filtro para cada hablante se muestran en el apéndice D.

Por su parte el hablante M4 experimentó un fenómeno muy particular donde el MSE fue cercano a 9.3 [mm], a pesar de la existencia de una correlación similar a la de otros hablantes. Una observación más detenida del gráfico «*VPO vs Muestras*» permitió observar la existencia de aperturas y cierres del velo que concordaron con los espacios en los que ocurrió el fenómeno en valores medidos. Sin embargo, se observó un error considerablemente alto ya que la diferencia existente entre datos medidos y estimados fue bastante alta. Finalmente, M1 mostró un comportamiento más homogéneo estimando aperturas parciales cuando se tenían cierres totales en los valores medidos, además de la presencia de un fenómeno de desfase donde las aperturas en los valores estimados se producían con leve anterioridad respecto a las de los valores medidos.

En términos generales la relación entre el grado de apertura y la acústica de la voz se destacó en los hablantes M3, M2, M1 y F4 que obtuvieron mejores resultados de correlación y error cuadrático medio tal como se observó en las gráficas de estimación (Apéndice C), mientras que los hablantes F2, F3 y M4 tuvieron un desempeño bajo que evidenció una relación más débil, presentando un error de estimación alto con respecto a los valores medidos.

Los mejores resultados entre género fueron obtenidos por los hablantes masculinos, en los que se lograron estimaciones de apertura más cercanas a los valores medidos. Esta tendencia puede ser generada por la diferencia entre las dimensiones del tracto vocal para hombres y mujeres. El tracto vocal de los hombres es en general más grande [33], situación que mejoró la calidad de las imágenes MRI por la mejor relación entre proporciones, resolución, longitud física y ruido, mientras que el tracto vocal de las mujeres al tener menores dimensiones disminuye su robustez ante el ruido y la resolución, generando así un mayor error en el proceso de segmentación.

El rango de frecuencia para el filtrado a la salida de los datos para el experimento intra hablan- te estuvo comprendido entre 3.2 y 5.2 [Hz] (Tabla 3.2). Es de resaltar que para cada hablante se determinó el valor de frecuencia para el filtrado con mejores resultados de correlación y error cuadrático medio. La variación entre los valores de frecuencia fue de 2 [Hz], la varianza en la frecuencia óptima tomada para cada hablante del experimento fue de 0.6233 y la desviación estándar de 0.7895 [Hz]. En [35] se establece el valor de filtrado para hablantes en un rango de 4.5 a 5 [Hz], obtenido a partir del estudio de únicamente dos hablantes, en este experimento se trabajó con siete hablantes y se observó una tendencia más cercana al rango entre 3.7 y 4.3 [Hz] para los 4 hablantes con mejores resultados. En este caso, el promedio de los experimentos realizados entre-hablantes permitió observar una tendencia a 4 [Hz], valor ubicado dentro del rango previamente mencionado.

4.2 Experimento interhablante

Considerando cada uno de los experimentos realizados por hablante y tomando como base las características de los datos de cada uno de estos, se realizó una serie de experimentos para los 7 hablantes en cuestión, realizando una validación con cada uno. La ejecución de cada uno de los experimentos fue realizada teniendo en cuenta la cantidad de 3 neuronas definidas de acuerdo a la variabilidad entre aperturas y cierres estimadas, ya que si bien los datos estimados con 1 o 2 neuronas fueron similares a los obtenidos con 3 o 4, en términos de MSE y correlación, no mostraron diferencias tan marcadas entre aperturas y cierres.

En términos generales la idea fue visualizar el comportamiento del MSE y la correlación lineal de Pearson, a medida que se modificó la frecuencia de corte del filtro paso-bajo aplicado a los datos articulatorios estimados y validando con solo 1 hablante en cada caso, para luego obtener un

promedio de los datos de cada variable (MSE y correlación) de los 7 experimentos (ver Figura 43).

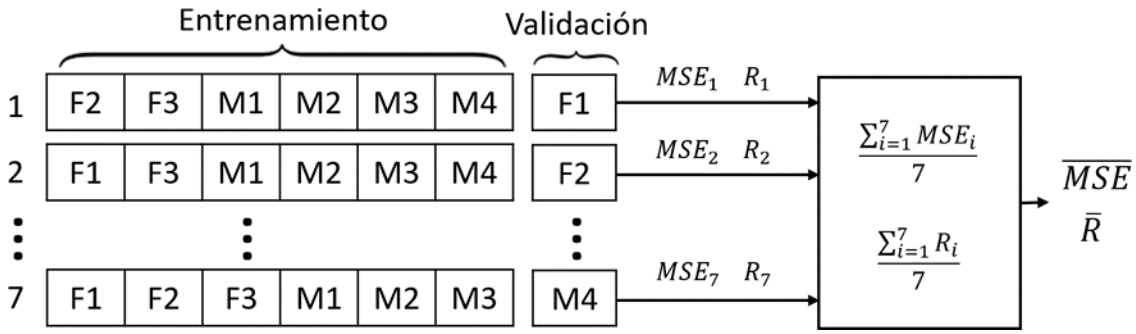


Figura 43. *Diagrama que representa el proceso realizado. Cada elemento MSE_i corresponde a una matriz fila de 1 35 y cada R_i es una matriz de 2x 35. Cada uno con 35 muestras a cada frecuencia de corte de 1 hasta 11 [Hz].*

La Tabla 103 recopila la cantidad de muestras y registros nasales usados en cada experimento, a partir del hablante de validación para un total de 9587 muestras con 1220 registros nasales.

Tabla 10.

Recopilación de valores para entrenamiento entre hablantes con un total de 9587 muestras con 1220 registros nasales.

Hablante de Validación	**F2**	**F3**	**F4**	**M1**	**M2**	**M3**	**M4**
No. muestras entrenamiento	8072	8722	8458	8759	7809	7932	7770
No. muestras validación	1515	865	1129	828	1778	1655	1817
Registros nasales entrenamiento	1014	1112	1094	1098	1011	1006	985
Registros nasales validación	206	108	126	122	209	214	235

En principio, los resultados obtenidos con 1 neurona, sugirieron un modelo lineal para realizar el

mapeo acústico-articulatorio, razón por la cual se propuso realizar inicialmente los experimentos por medio de una regresión lineal múltiple, obteniendo gráficos MSE [mm] vs Frecuencia de corte [Hz] y R vs Frecuencia de corte [Hz], que permitieron visualizar el comportamiento de estas variables al cambiar la frecuencia en un rango de 1 a 11 [Hz]. El promedio obtenido a partir de los 7 experimentos, mostró los resultados presentes en la Figura 44, donde se observó un mejor desempeño en una frecuencia de corte cercana a los 4 [Hz], con una disminución en el MSE hasta aproximadamente 5.01 [mm] correspondientes a alrededor de 1.7 píxeles, además de un aumento del coeficiente de correlación para ambos puntos de medición cercano a 0.2.

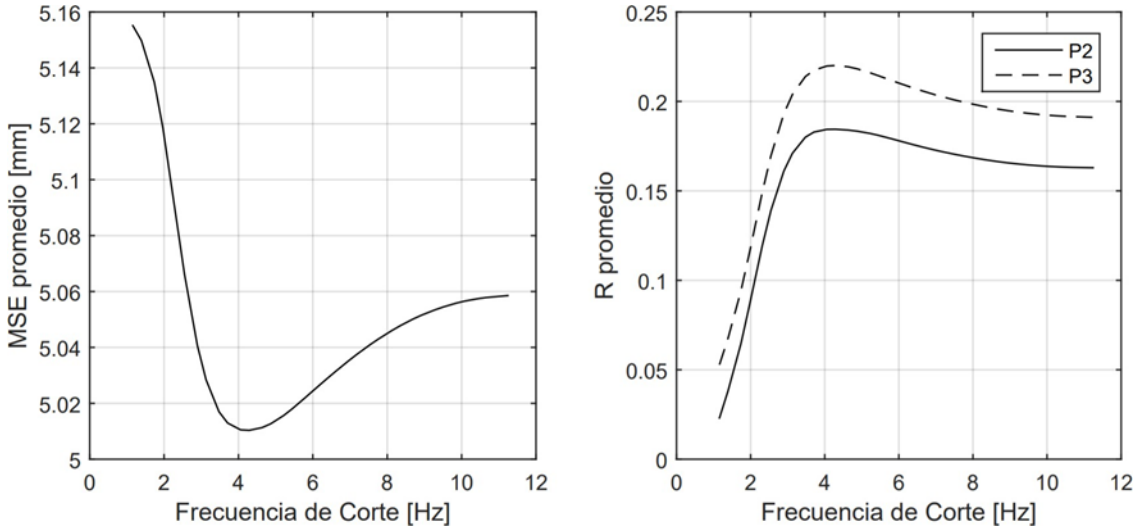


Figura 44. *Resultados obtenidos por medio de la regresión lineal múltiple. Izquierda: MSE promedio [mm] vs Frecuencia de corte [Hz]. Derecha: R promedio vs Frecuencia de corte [Hz].*

Por su parte, los resultados obtenidos por medio de la red neuronal no mostraron diferencias considerables respecto a los obtenidos por medio de la regresión lineal (ver Figura 45). No obstante, el uso de un modelo no lineal para la realización del mapeo, permitió visualizar un mayor contraste entre las aperturas y cierres velofaríngeas. Un ejemplo de esto se puede observar en la Figura 46, donde se

ilustra una comparativa entre los datos obtenidos por medio de ANN (Red Neuronal Artificial) y RLM (Regresión Lineal Múltiple), respecto a los valores medidos en P2 del hablante M2. Se observó que la estimación obtenida por medio de ANN mostró valores más dinámicos en términos de aperturas y cierres del velo del paladar a pesar de no alcanzar en términos generales los valores máximos del VPO medido.

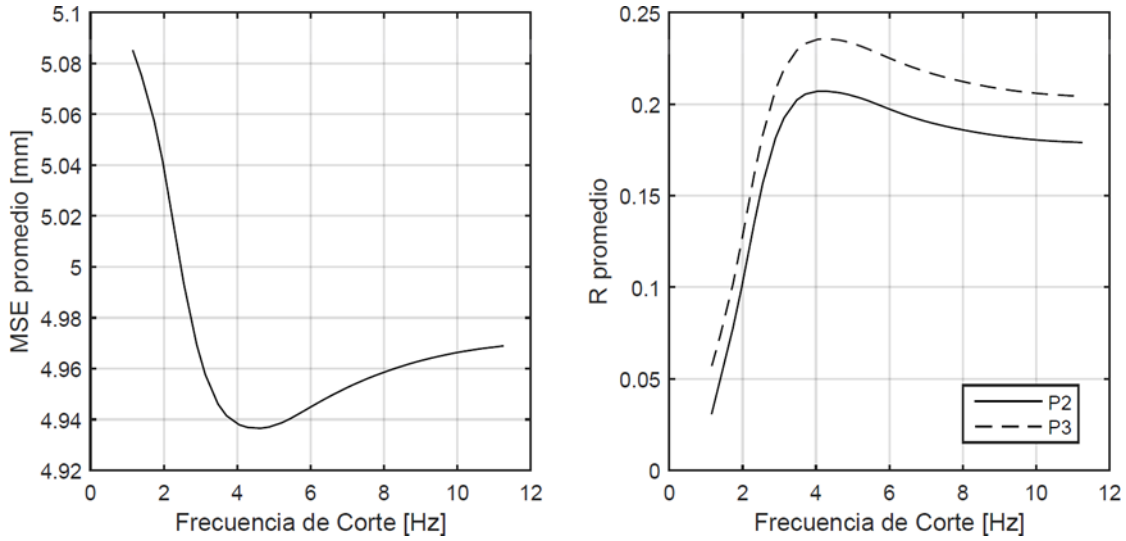


Figura 45. *Resultados obtenidos por medio de la red neuronal artificial con 3 neuronas en la capa oculta. Izquierda: MSE promedio [mm] vs Frecuencia de corte [Hz]. Derecha: R promedio vs Frecuencia de corte [Hz].*

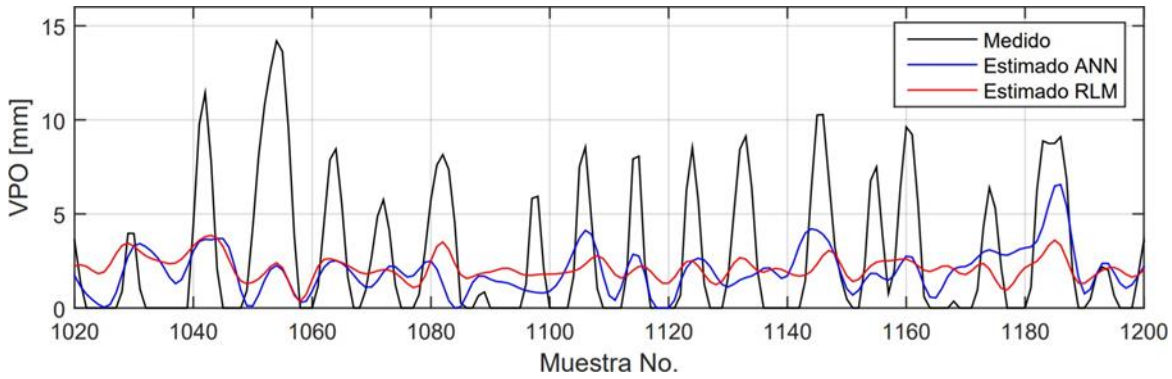


Figura 46. *Comparación del VPO medido y estimado con ANN (3 neuronas) y RLM para el punto P2 del hablante M2.*

A pesar que el resultado máximo de correlación obtenido para el experimento promedio fue cercano a 0.2, así como el MSE más bajo fue cercano a 4.94 [mm], los resultados experimentales obtenidos con algunos hablantes en particular mostraron resultados notoriamente superiores, como en el caso del hablante M3 cuyo MSE fue cercano a 3.5, con un R aproximadamente igual a 0.37 para P2 y a 0.45 para P3. La Tabla 11 recopila los resultados obtenidos para los 7 experimentos realizados con la ANN.

Tabla 11.

Recopilación de resultados de validación, con mayor desempeño para cada hablante, en la frecuencia de corte de 4 [Hz] del filtro paso-bajo.

Hablante de Validación	**F2**	**F3**	**F4**	**M1**	**M2**	**M3**	**M4**	**Media**
$MSE_{mín}$**[mm]**	5.61	5.56	3.76	2.43	4.53	3.49	9.25	**4.94**
$R_{máx}$**(P2)**	0.11	0.16	0.14	0.23	0.31	0.37	0.10	**0.20**
$R_{máx}$**(P3)**	0.09	0.13	0.21	0.17	0.37	0.45	0.16	**0.23**

La obtención de resultados más dinámicos por medio del uso de la ANN con 3 neuronas, mostró la necesidad de usar un modelo no lineal para la obtención del mapeo, ya que los resultados obtenidos mostraron no solo una detección de aperturas, sino también de cierres del velo, que fue más acertada respecto a los datos medidos para el hablante en cuestión.

5. Conclusiones

En el presente trabajo se analizó la relación entre el grado de apertura velifaríngea y el espectro suavizado de la señal acústica representando mediante los parámetros MFCC. Se probaron dos tipos de modelos: regresión lineal múltiple y redes neuronales artificiales. A modo de criterio de desempeño se utilizó el error cuadrático medio y la correlación lineal de Pearson, donde, desde el punto de vista inter-hablante se obtuvo resultados de MSE de 4.94 [mm], con correlación igual a 0.2 para P2 y 0.24 para P3 en el caso de la red neuronal; y, en el caso de regresión lineal múltiple se obtuvo un MSE de 5.02 [mm] y valores de correlación de 0.18 para P2 y 0.23 para P3. Si bien los resultados obtenidos a través de los dos métodos de modelado mostraron valores similares en términos de correlación y MSE, la comparación entre los valores estimados por ambos métodos evidenció un mejor desempeño por parte del modelo de red neuronal debido a que este fue más dinámico, mejorando las estimaciones en términos de identificación tanto de aperturas como cierres del VPO. Los resultados obtenidos apoyan la hipótesis de la posibilidad de estimación del grado de apertura velofaríngea a partir de la señal acústica. De otra parte, el entrenamiento de la red neuronal permitió evidenciar la capacidad de la red neuronal para detectar sonidos de tipo nasal a partir de parámetros MFCC, logrando detectar la apertura del velo del paladar sobre las posiciones de referencia definidas. Se observó que las posiciones dominantes de apertura y cierre del velo del paladar están acordes con

los sonidos generados (articulaciones vocales y nasales). Aunque la red neuronal no logra obtener resultados muy precisos del grado de apertura velofaríngea medida en milímetros, es muy bueno para detectar su apertura.

Durante las pruebas se encontró que es posible estimar pendientes de apertura y cierre, siendo el algoritmo acertado en la mayoría de circunstancias; sin embargo, la distancia de apertura sigue siendo aún un valor impreciso debido a las diferentes fuentes de error asociadas a los datos originales. En particular se observó que para aperturas del velo comprendidas entre 0 y 5 milímetros el algoritmo logra estimar un valor cercano al medido, en contraste con esta tendencia, las aperturas del velo superiores a 10 milímetros suponen un error de estimación más elevado y son los casos donde la estimación de la distancia de apertura tiene grandes falencias.

Referencias

1. Sepulveda-Sepulveda A., Castellanos-Dominguez G. Assessment of the Relation Between Low- Frequency Features and Velum Opening by Using Real Articulatory Data., Ronzhin A., Potapova R., Németh G. (eds) Speech and Computer. SPECOM 2016. Lecture Notes in Computer Science, vol 9811. Springer, Cham. 2016.
2. USC-TIMIT: A database of multimodal speech production data.
3. S. Narayanan, et. al. Real-time magnetic resonance imaging and electromagnetic articulo graphy database for speech production research., Signal Analysis and Interpretation Laboratory, Department of Linguistics, University of Southern California, Los Angeles, California. 2014.
4. Proctor Michael, Bone Danny, Nassos Katsamanis, Narayanan Shrikanth. Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis., Viterbi School of Engineering, University of Southern California, USA, 2010.
5. Paulino Del Pino, Iván Granadillo, Mario Miranda, Carlos Jiménez, José A. Díaz. Diseño de un sistema de medición de parámetros característicos y de calidad de señales de voz., Universidad de Carabobo, Facultad de Ingeniería, Escuela de Ingeniería Eléctrica, Departamento de Electrónica y Comunicaciones. 2008.
6. Stevens, K. S., *Acoustic Phonetics*; MIT Press: 2000.
7. María José Albalá. Análisis y síntesis de las consonantes nasales., Revista de Filología Española, vol. LXXII, N.º 1/2, 1992.
8. Sepúlveda Sepúlveda, A.; Delgado-Trejos, E.; Murillo Rendon, S. y Castellanos-

Domínguez, G. Tecnológicas **2009**, 223-237.

9. Matias Zañartu Aplicaciones del análisis acústico en los estudios de la voz humana., Unidad de Acústica - Escuela de Fonoaudiología, Universidad Mayor, Santiago, Chile, 2003.
10. María Florencia Assaneo Modela del sistema vocal humano y su aplicación a estudios de percepción y producción del habla., Departamento de Física, Facultad de ciencias exactas y naturales, Universidad de Buenos Aires, Argentina, 2014.
11. Federico Miyara La voz humana., Argentina, 2004.
12. Marco Guzman Acústica del tracto vocal., Escuela de fonoaudiología, Universidad de Chile, Santiago de Chile, Chile, 2004.
13. Carlos Maya León. Sistema telefónico con síntesis de voz y detección de tonos., Facultad de ingeniería eléctrica, Universidad Nacional Autónoma de México. 2010.
14. Héctor Arturo Kairuz Hernández Detección y marca de los cambios espectrales notorios en voces patológicas., Centro de estudios de electrónica y tecnología de la información, Facultad de ingeniería eléctrica, Universidad Central Marta Abreu de las villas, 2009.
15. José Maestre, Christian Serje Técnicas para reconocimiento automático de locutores., Fa- cultad de ingeniería eléctrica y electrónica, Universidad Tecnológica de Bolívar, Cartagena, 2007.
16. Ignacio Cárdenas Pinto, Humberto Ceballos Cartes, Shang-Hsueh Lee, Wladiir Pavez Rocha, Claudio Terrisse Gutiérrez Estudio acústico de la variación interlocutor en sujetos hablantes nativos del español de Santiago de Chile., Facultad de medicina, Escuela de Fonoaudiología, Universidad de Chile, 2010.
17. Lindasalwa Muda, Mumtaj Begam, I. Elamvazuthi Voice Recognition Algorithms using

Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques., Journal of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617, 2010.

18. Raúl Carranza, Nicolas Alessandroni Bentancor Dicción, Fonética y técnica vocal: Él entrenamiento vocal en clave interdisciplinaria., Grupo de investigación en técnica vocal, Facultad de bellas artes, Universidad nacional de la plata, 1999.
19. John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, Victor Zue TIMIT Acoustic-Phonetic Continuous Speech Corpus., Linguistic Data Consortium, The Trustees of the University of Pennsylvania. 1993.
20. Marie K. Huffman, Rena A. Krakow. Phonetics and phonology. Nasals, Nasalization, and the Velum, volume 5, Academic Press INC, A division of Harcourt Brace & Company, San Diego, CA, U.S.A. 2014.
21. Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, Shrikanth Narayanan. A Study of Emotional Speech Articulation using a Fast Magnetic Resonance Imaging Technique., Speech Analysis and Interpretation Laboratory (SAIL), Viterbi School of Engineering, University of Southern California, Los Angeles, California, U.S.A. 2005.
22. Jangwon Kim, Naveen Kumar, Sungbok Lee, Shrikanth Narayanan. Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data, University of Southern California, Los Angeles, CA, U.S.A. 2014.
23. Asterios Toutios, Shrikanth S. Narayanan. Advances in real-time magnetic resonance imaging of the vocal tract for speech science and technology research., SIP (2016), vol. 5, e6, page 1 of 12, 2016.

24. Chunlei Liu, Roland Bammer, Michael E. Moseley. Parallel Imaging Reconstruction for Arbitrary Trajectories Using k-Space Sparse Matrices (kSPA), Magnetic Resonance in Medicine 58:1171–1181, 2007.
25. Carmen Rincón Llorente Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz., Escuela técnica superior de ingeníerons de telecomunicación, Universidad politécnica de Madrid, 2007.
26. Vibha Tiwari MFCC and its applications in speaker recognition, International Journal on Emerging Technologies 1(1): 19-22(2010), 2010.
27. Chandar Kumar, Faizan ur Rehman, Shubash Kumar, Atif Mehmood, Ghulam Shabir Analy- sis of MFCC and BFCC in a Speaker Identification System., 2018 International Conference on Computing, Mathematics and Engineering Technologies, 2018.
28. Grazina Korvel, Bozena Kostek Examining Feature Vector for Phoneme Recognition, 2017 IEEE International Symposium on Signal Processing and Information Technology, 2017.
29. Olga L. Ramos, Diego A. Rojas, Leonardo A. Góngora Reconocimiento de patrones de habla usando MFCC y RNA., Visión electrónica, algo más que un sólido, Vol. 10, No. 1, 1-, Enero- Junio 2016, 2016.
30. Guillermo Arturo Martínez Mascorro, Gualberto Aguilar Torres Sistema para identificación de hablantes robusto a cambios en la voz., Ingenius. No. 8, (Julio/Diciembre). pp. 45-53. ISSN:1390-650X, 2012.
31. Kamil Wojcicki Mel frequency cepstral coefficient feature extraction that closely matches that of HTK's HCopy, HTK MFCC MATLAB, MathWorks, 2011.
32. Slaney, M. Auditory toolbox: A Matlab Toolbox for Auditory Modeling Work., Tech.

rep. 1998.

33. Julián O. Reyes M., Paula A. Vásquez S., Franklin A. Sepúlveda S. Análisis de la relación existente entre la longitud del tracto vocal obtenida a partir de imágenes por resonancia magnética, y parámetros acústicos de la voz., Facultad de Ingenierías Fisico-Mecánicas. Es- cuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones. Universidad Industrial de Santander. Colombia. 2018.
34. K. Matsuo, H. Metani, K.A. Mays, and J.B. Palmer. Effects of Respiration on Soft Palate Mo- vement in Feeding, Dept. of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, Baltimore, MD, USA. 2010.
35. Tomoki Toda, Alan W. Black, Keiichi Tokuda. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model., Speech Communication 50, 215–227, 2008.
36. Prasanta Kumar Ghosha and Shrikanth Narayanan. A generalized smoothness criterion for acoustic-to-articulatory inversion., Department of Electrical Engineering, Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, California 90089. 2010.
37. Munendra Singh, Ashish Verma, Neeraj Sharma. Multi-objective noise estimator for the applications of de-noising andsegmentation of MRI data.), School of Biomedical Engineering, Indian Institute of Technology (Banaras Hindu University), Varanasi, India, 2018.
38. José V. Manjón, Pierrick Coupé, Antonio Buades. MRI noise estimation and denoising using non-local PCA.), nstituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universidad Politécnica de Valencia,

Camino de Vera s/n, 46022 Valencia, Spain, 2015.

39. Alexander Hewer, Stefanie Wuhrer, Ingmar Steiner, Korin Richmond A multilinear tongue model derived from speech related MRI data of the human vocal tract.), Computer Speech & Language 51 (2018) 68-92, International Speech Communication Association (ISCA), 2018.
40. Chris D. Constantinides, Ergin Atalar, Elliot R. McVeigh Measurement of signal intensities in the presence of noise in MR images.), Med. Phys. 12,232-233 (1984). Erratum in 13, 544, 1986.
41. Chris D. Constantinides, Ergin Atalar, Elliot R. McVeigh Signal-to-Noise Measurements in Magnitude Images from NMR Phased Arrays.), MRM 38:852-857, 1997.
42. Xinhao Liu, Masayuki Tanaka, Masatoshi Okutomi. Noise Level Estimation from a Single Image., MATLAB, MathWorks, accedido en 10-02-2019 a traves de: https://www.mathworks.com/matlabcentral/fileexchange/36921-noise-level-esti mation-from- a-single-image, 2015.
43. Xinhao Liu, Masayuki Tanaka, Masatoshi Okutomi Noise level estimation using weak textured patches of a single noisy image, IEEE International Conference on Image Processing (ICIP), 2012.
44. Xinhao Liu, Masayuki Tanaka, Masatoshi Okutomi Single-image Noise Level Estimation for Blind Denoising), IEEE Transactions on Image Processing, Vol.22, No.12, pp.5226-5237), 2012.
45. Schmidhuber, J. Deep learning in neural networks: An overview., The Swiss AI Lab IDSIA, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, University of Lugano

and SUPSI, Galleria 2, 6928 Manno-Lugano, Switzerland, 2014.

46. Matich, J. Redes neuronales: conceptos básicos y aplicaciones., Universidad Tecnológica Nacional – Facultad Regional Rosario, 2001.

47. IBM. El modelo de redes neuronales., https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/components/neuralnet/neuralnet_model.html, 2018.

48. Salas, R. Redes neuronales artificiales., Universidad de Valparaíso. Departamento de computación, 2004.

49. The MathWorks, Inc. Deep Learning Toolbox., https://www.mathworks.com/products/deep-learning.html, 2018.

50. The MathWorks, Inc. Fit Data with a Shallow Neural Network., https://www.mathworks.com/help/deeplearning/gs/fit-data-with-a-neural-network.html, 2018.

51. The MathWorks, Inc. Shallow Neural Network Time-Series Prediction and Modeling., https://www.mathworks.com/help/deeplearning/gs/neural-network-time-series- prediction-and-modeling.html, 2018.

Apéndices

Apéndice A. Frases de la base de datos

Frases de la base de datos USC-TIMIT articuladas por los hablantes para el desarrollo del proyecto.

Numero	Frase
1	This was easy for us.
2	Jane may earn more money by working hard.
3	She is thinner than I am.
4	Bright sunshine shimmers on the ocean.
5	Nothing is as offensive as innocence.
6	Why yell or worry over silly items.
7	Where were you while we were away?
8	Are your grades higher or lower than Nancy's.
9	He will allow a rare lie.
10	Will robin wear a yellow lily.
11	Swing your arm as high as you can.
12	Before Thursday's exam review every formula.
13	The museum hires musicians every evening.
14	A roll of wire lay near the wall.
15	Carl lives in a lively home.
16	Alimony harms a divorced man's wealth.
17	She wore warm fleecy woolen overalls.
18	Alfalfa is healthy for you.
19	When all else fails use force.
20	Although always alone we survive.
21	Only lawyers love millionaires.
22	Did dad do academic bidding?
23	Help greg to pick a peck of potatoes.

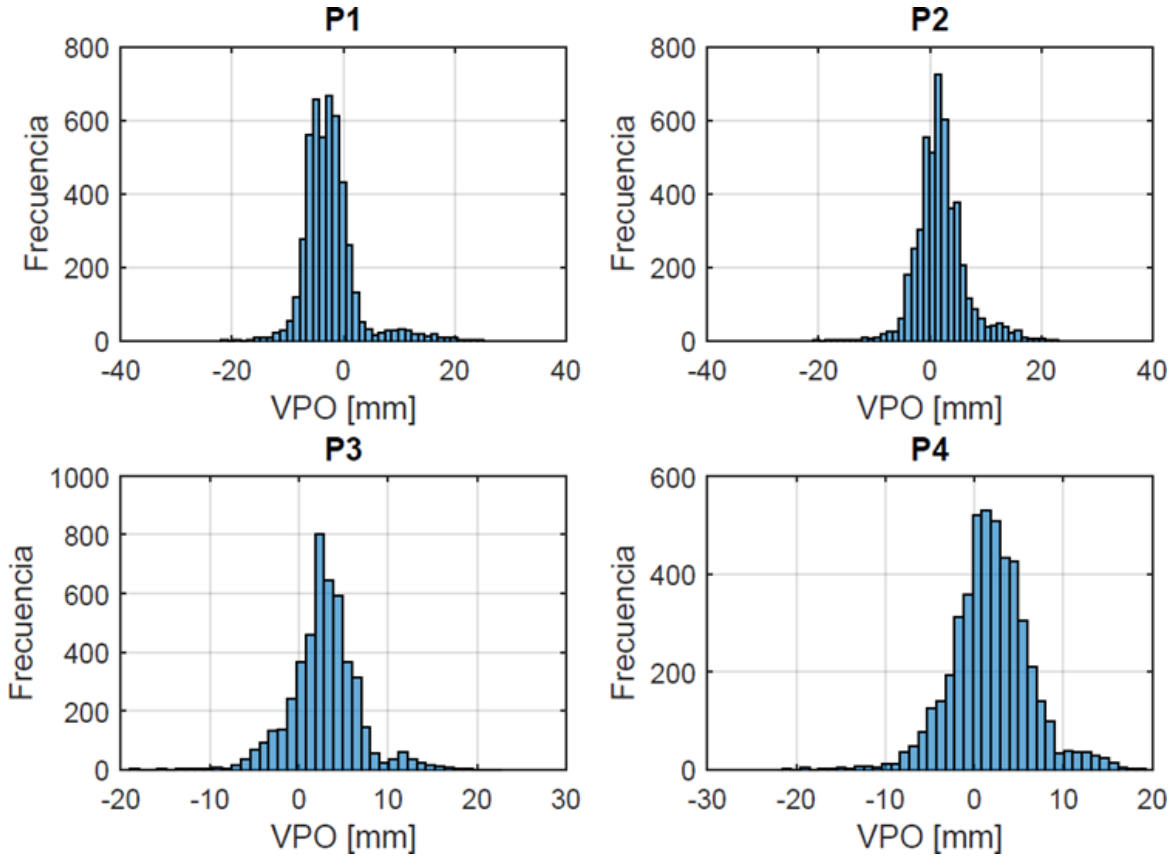
24	A good attitude is unbeatable.
25	Coconut cream pie makes a nice dessert.
26	Don't do Charlie's dirty dishes.
27	Help celebrate your brother's success.
28	Only the most accomplished artists obtain popularity.
29	Critical equipment needs proper maintenance.
30	Young people participate in athletic activities.
31	Barb's gold bracelet was a graduation present.
32	Stimulating discussions keep students' attention.
33	Etiquette mandates compliance with existing regulations.
34	Biblical scholars argue history.
35	Elderly people are often excluded.
36	Basketball can be an entertaining sport.
37	Addition and subtraction are learned skills.
38	Grandmother outgrew her upbringing in petticoats.
39	At twilight on the twelfth day we'll have chablis.
40	Catastrophic economic cutbacks neglect the poor.
41	Ambidextrous pickpockets accomplish more.
42	Even a simple vocabulary contains symbols.
43	The eastern coast is a place for pure pleasure and excitement.
44	The lack of heat compounded the tenant's grievances.
45	Academic aptitude guarantees your diploma.
46	The prowler wore a ski mask for disguise.
47	We experience distress and frustration obtaining our degrees.
48	The singer's finger had a splinter.
49	The legislature met to judge the state of public education.
50	Chocolate and roses never fail as a romantic gift.
51	Any contributions will be greatly appreciated.

52	Continental drift is a geological theory.
53	Regular attendance is seldom required.
54	Challenge each general's intelligence.
55	We got drenched from the uninterrupted rain.
56	Last year's gas shortage caused steep price increases.
57	Upgrade your status to reflect your wealth.
58	Eat your raisins outdoors on the porch steps.
59	Porcupines resemble sea urchins.
60	Spring street is straight ahead.
61	Cliff's display was misplaced on the screen.
62	An official deadline cannot be postponed.
63	Fill that canteen with fresh spring water.
64	Gently place Jim's foam sculpture in the box.
65	Bagpipes and bongos are musical instruments.
66	Doctors prescribe drugs too freely.
67	Will you please describe the idiotic predicament.
68	It's impossible to deal with bureaucracy.
69	Good service should be rewarded by big tips.
70	My instructions desperately need updating.
71	Cooperation along with understanding alleviate dispute.
72	Cement is measured in cubic yards.
73	Primitive tribes have an upbeat attitude.
74	Flying standby can be practical if you want to save money.
75	It's hard to tell an original from a forgery.
76	'The thinker' is a famous sculpture.
77	The misprint provoked an immediate disclaimer.
78	A large household needs lots of appliances.
79	Cut a small corner off each edge.

80	Iguanas and alligators are tropical reptiles
81	Masquerade parties tax one's imagination.
82	Penguins live near the icy antarctic.
83	Guess the question from the answer.
84	Medieval society was based on hierarchies.
85	Project development was proceeding too slowly.
86	Kindergarten children decorate their classrooms for all holidays.
87	Special task forces rescue hostages from kidnappers.
88	Call an ambulance for medical assistance.
89	He stole a dime from a beggar.
90	You must explicitly delete files.
91	A huge tapestry hung in her hallway.
92	Birthday parties have cupcakes and ice cream.
93	His scalp was blistered from today's hot sun.
94	She slipped and sprained her ankle on the steep slope.
95	The best way to learn is to solve extra problems.
96	His sudden departure shocked the cast.
97	Tugboats are capable of hauling huge loads.
98	A muscular abdomen is good for your back.
99	The cartoon features a muskrat and a tadpole.
100	The emblem depicts the acropolis all aglow.

Apéndice B. Resumen de valores estadísticos de la base de datos.

Hablante F2



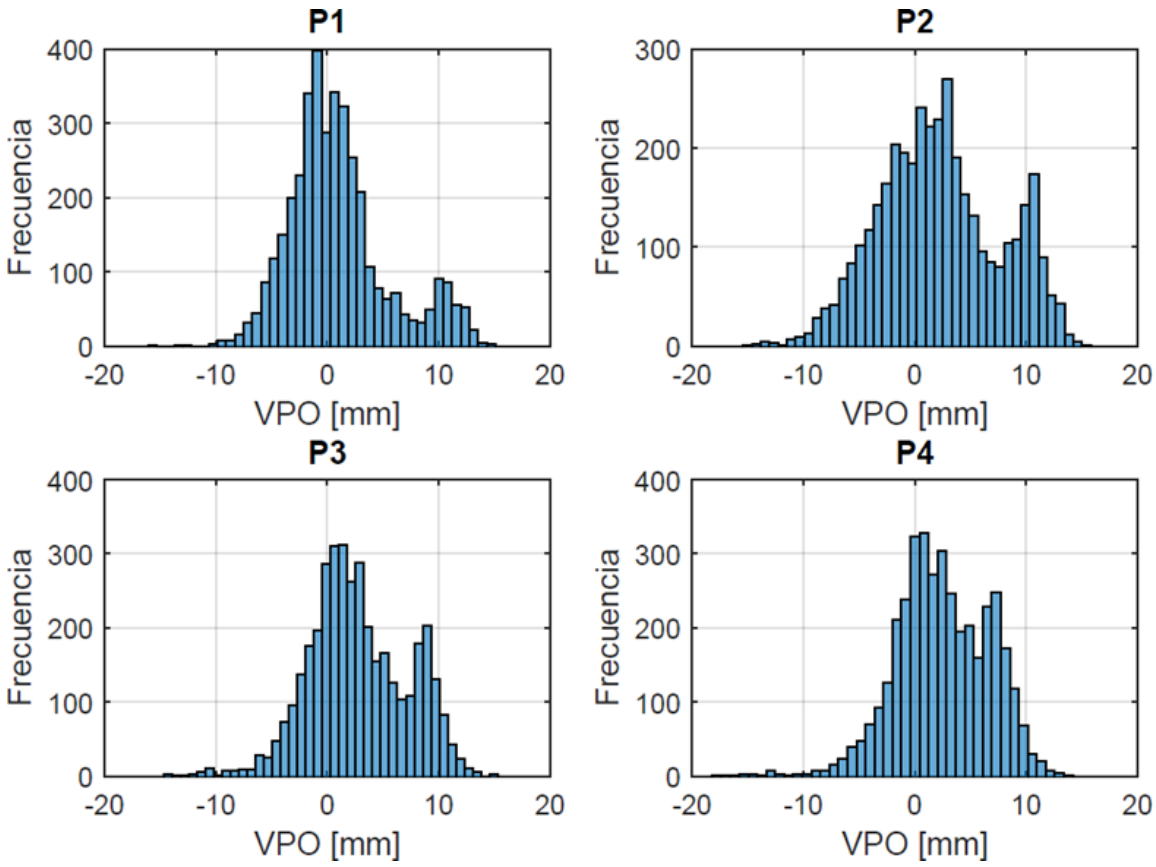
Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante F2.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	4732	4732	4732	4732
Medición máxima [mm]	24.91	22.97	22.52	18.94
Medición mínima [mm]	-21.60	-20.07	-18.34	-20.92
Moda [mm]	-4.67	3.13	2.45	1.15

Punto de Medición	**1**	**2**	**3**	**4**
Mediana [mm]	-3.06	1.51	2.78	1.75
Media [mm]	-2.51	1.91	2.88	1.84
Desviación estándar [mm]	4.61	4.27	3.85	4.36
Varianza	21.25	18.29	14.87	19.03

Hablante F3

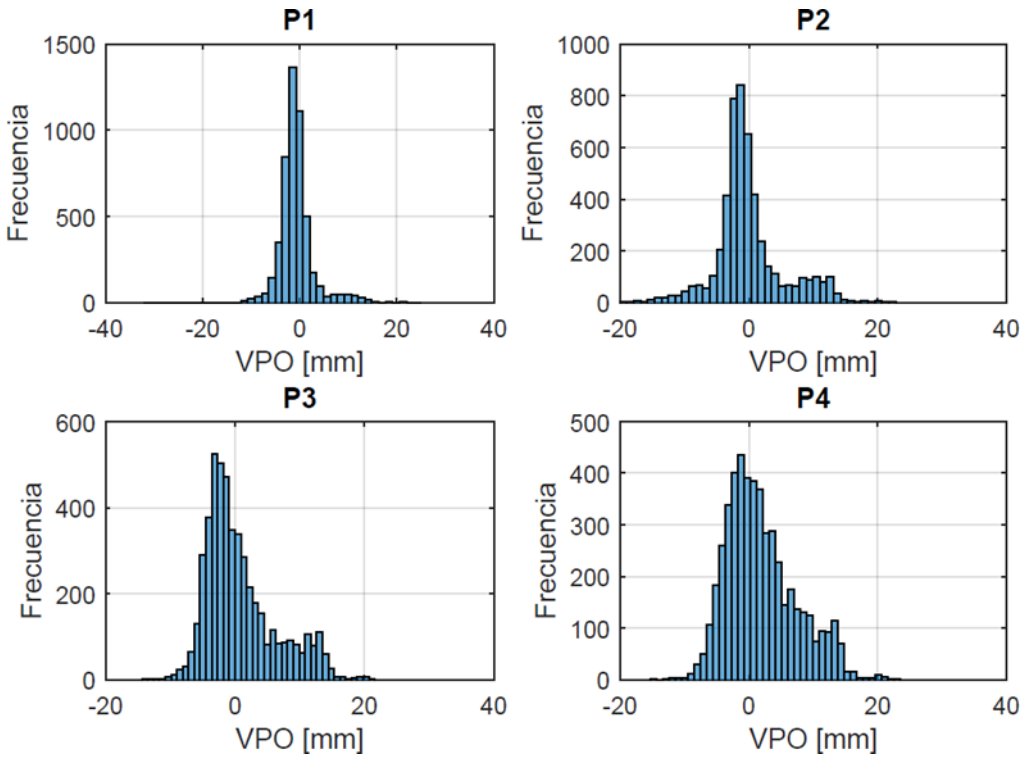


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante F3.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	3831	3831	3831	3831
Medición máxima [mm]	15.08	15.77	14.91	13.81
Medición mínima [mm]	-15.57	-15.11	-14.32	-17.61
Moda [mm]	-0.53	-1.46	3.28	7.57
Mediana [mm]	0.20	1.90	2.22	2.34
Media [mm]	0.98	2.22	2.76	2.49
Desviación estándar [mm]	4.49	5.43	4.34	4.16
Varianza	20.20	29.48	18.90	17.36

Hablante F4

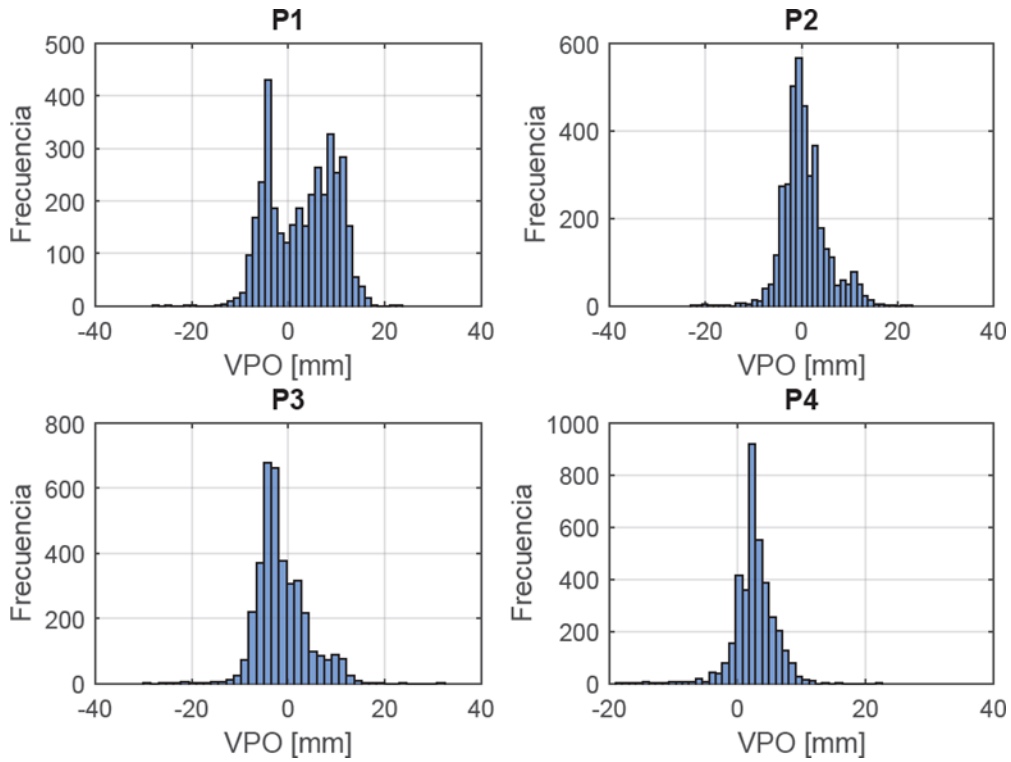


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante F4.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	4990	4990	4990	4990
Medición máxima [mm]	24.48	22.56	21.47	23.42
Medición mínima [mm]	-31.35	-19.29	-13.84	-14.47
Moda [mm]	-1.39	-2.05	-2.21	-2.65
Mediana [mm]	-1.09	-1.02	-0.70	0.93
Media [mm]	-0.72	-0.07	0.73	1.91
Desviación estándar [mm]	3.52	5.24	5.44	5.43
Varianza	12.40	27.52	29.62	29.54

Hablante M1

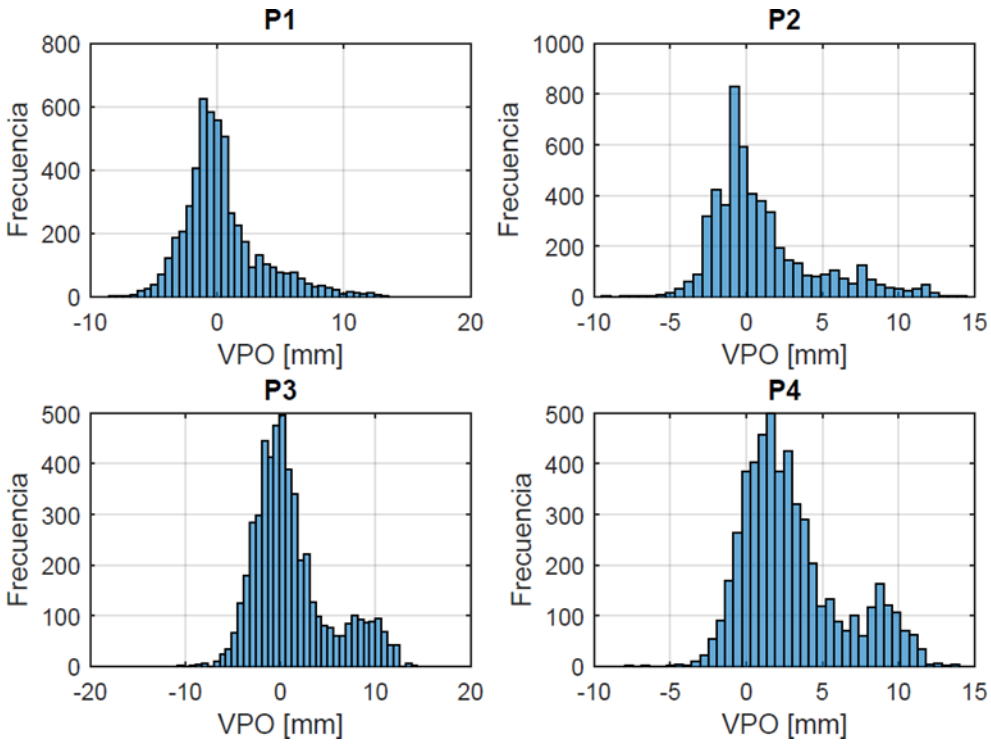


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante M1.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	3743	3743	3743	3743
Medición máxima [mm]	23.57	22.60	32.18	22.21
Medición mínima [mm]	-27.69	-22.62	-29.67	-18.58
Moda [mm]	-4.46	-0.58	-4.55	2.67
Mediana [mm]	3.73	-0.12	-2.27	2.67
Media [mm]	2.99	0.64	-1.26	2.62
Desviación estándar [mm]	6.72	4.51	5.10	3.26
Varianza	45.18	20.38	26.06	10.64

Hablante M2

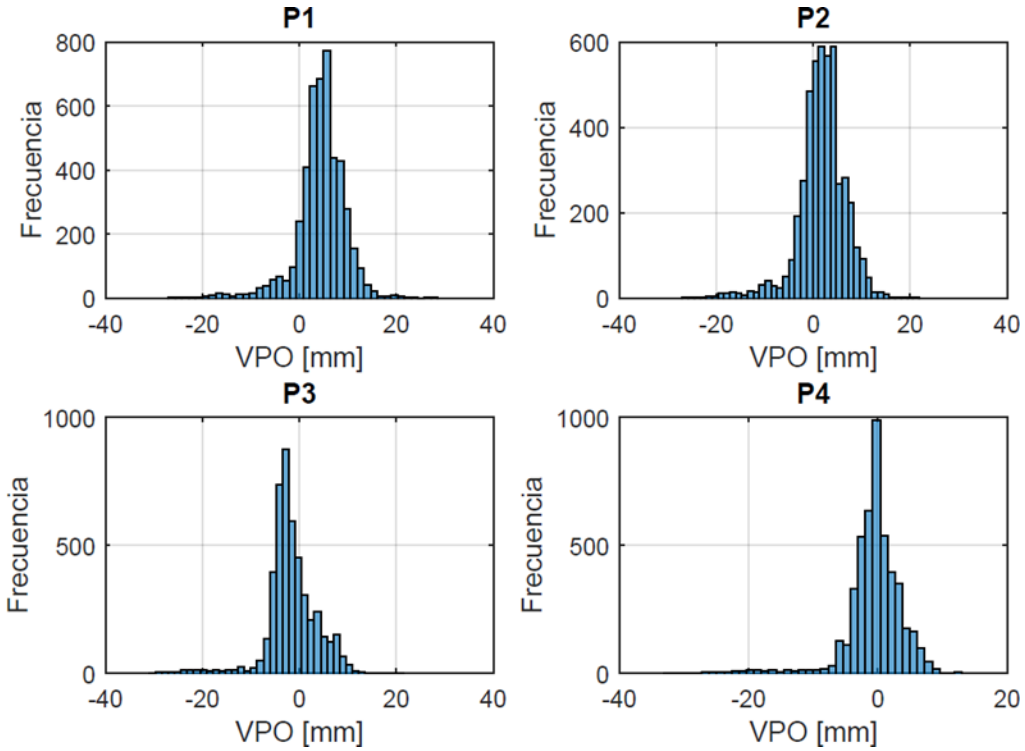


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante M2.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	5236	5236	5236	5236
Medición máxima [mm]	13.23	14.45	14.01	13.85
Medición mínima [mm]	-8.04	-9.39	-10.64	-7.87
Moda [mm]	-0.58	-1.06	0.39	2.52
Mediana [mm]	-0.07	0.03	0.36	2.23
Media [mm]	0.35	0.96	1.21	3.02
Desviación estándar [mm]	3.03	3.44	4.05	3.26
Varianza	9.21	11.85	16.38	10.67

Hablante M3

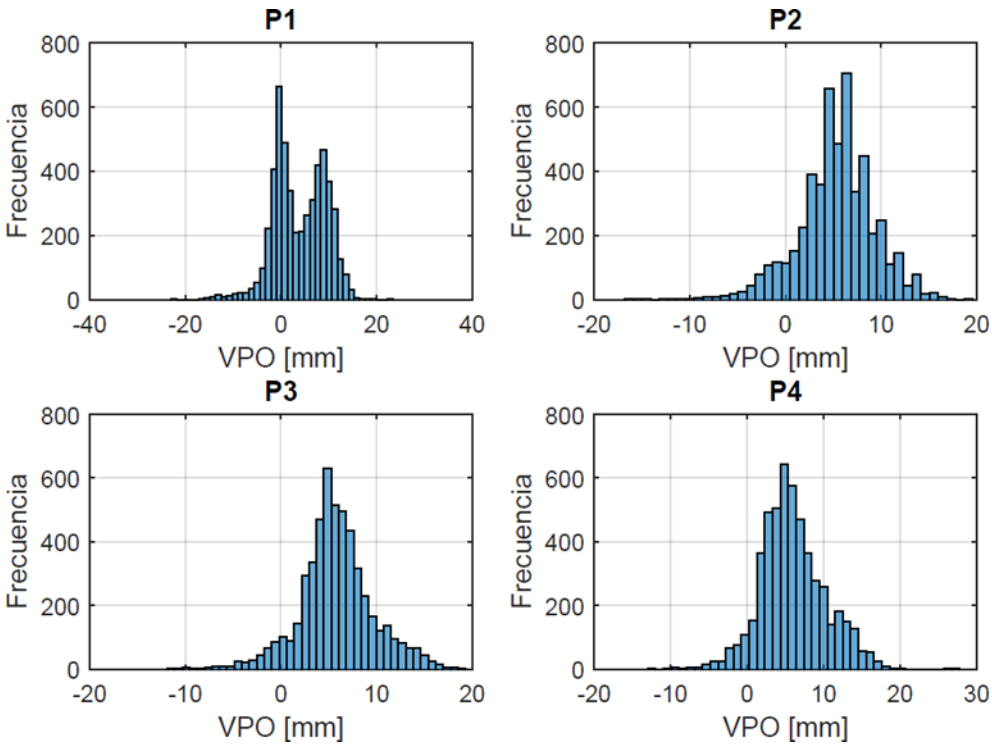


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

Resumen de valores estadísticos para los datos del hablante M3.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	4693	4693	4693	4693
Medición máxima [mm]	28.24	21.52	21.22	12.61
Medición mínima [mm]	-26.62	-26.11	-30.68	-32.15
Moda [mm]	3.76	2.73	-3.52	-0.73
Mediana [mm]	4.53	2.20	-2.18	-0.45
Media [mm]	4.43	1.84	-1.61	-0.52
Desviación estándar [mm]	5.01	4.93	5.07	4.28
Varianza	25.13	24.29	25.79	18.34

Hablante M4

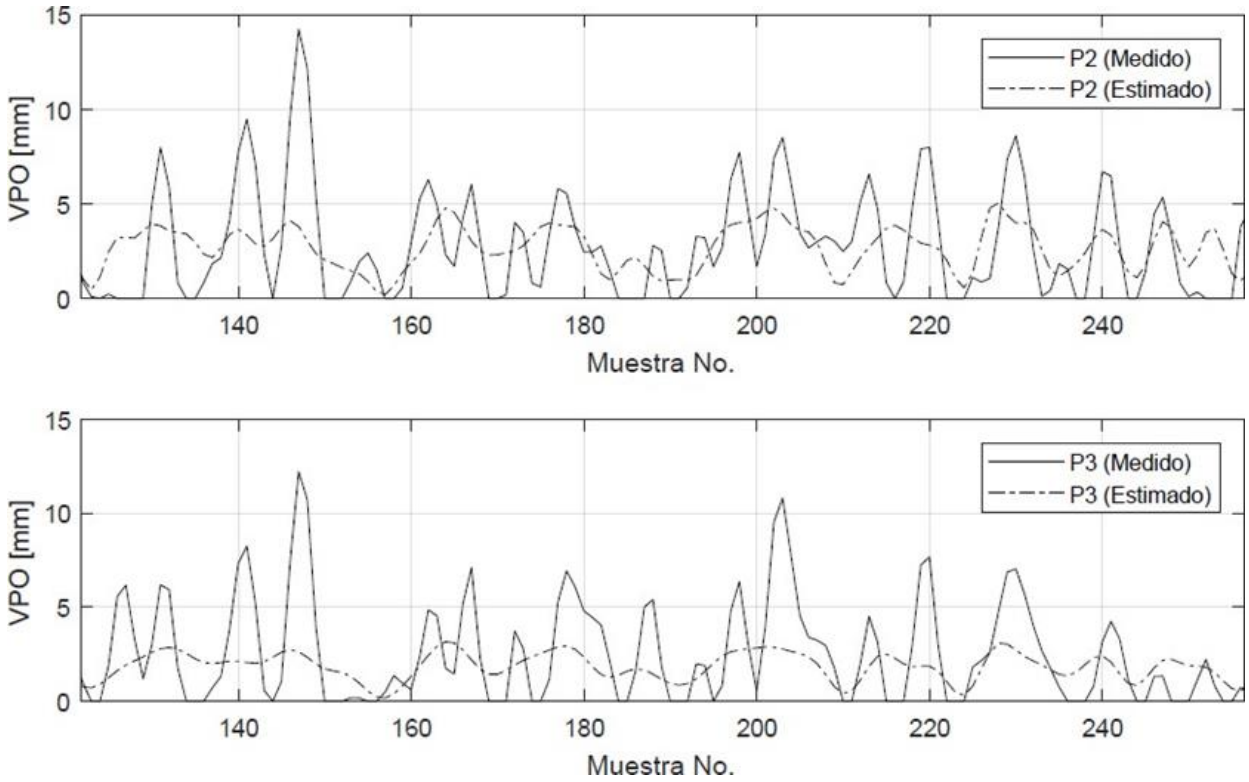


Histograma de frecuencias del VPO [mm] para los 4 puntos sobre el velo del paladar.

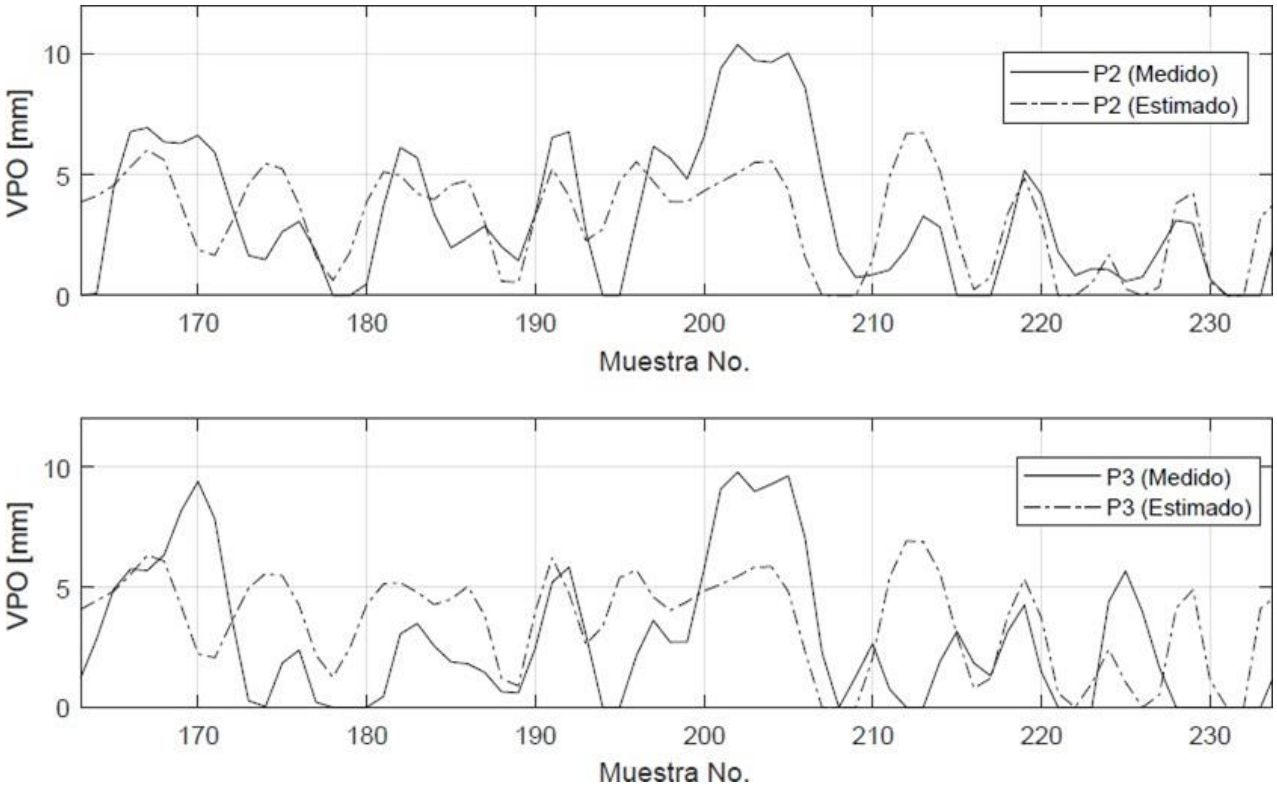
Resumen de valores estadísticos para los datos del hablante M4.

Punto de Medición	**1**	**2**	**3**	**4**
Número de mediciones	5179	5179	5179	5179
Medición máxima [mm]	23.04	19.52	18.92	27.63
Medición mínima [mm]	-22.08	-16.32	-11.89	-12.99
Moda [mm]	0.04	4.30	4.05	3.83
Mediana [mm]	3.63	5.49	5.65	5.56
Media [mm]	3.85	3.36	5.78	5.95
Desviación estándar [mm]	5.22	3.94	3.84	4.18
Varianza	27.27	15.56	14.81	17.50

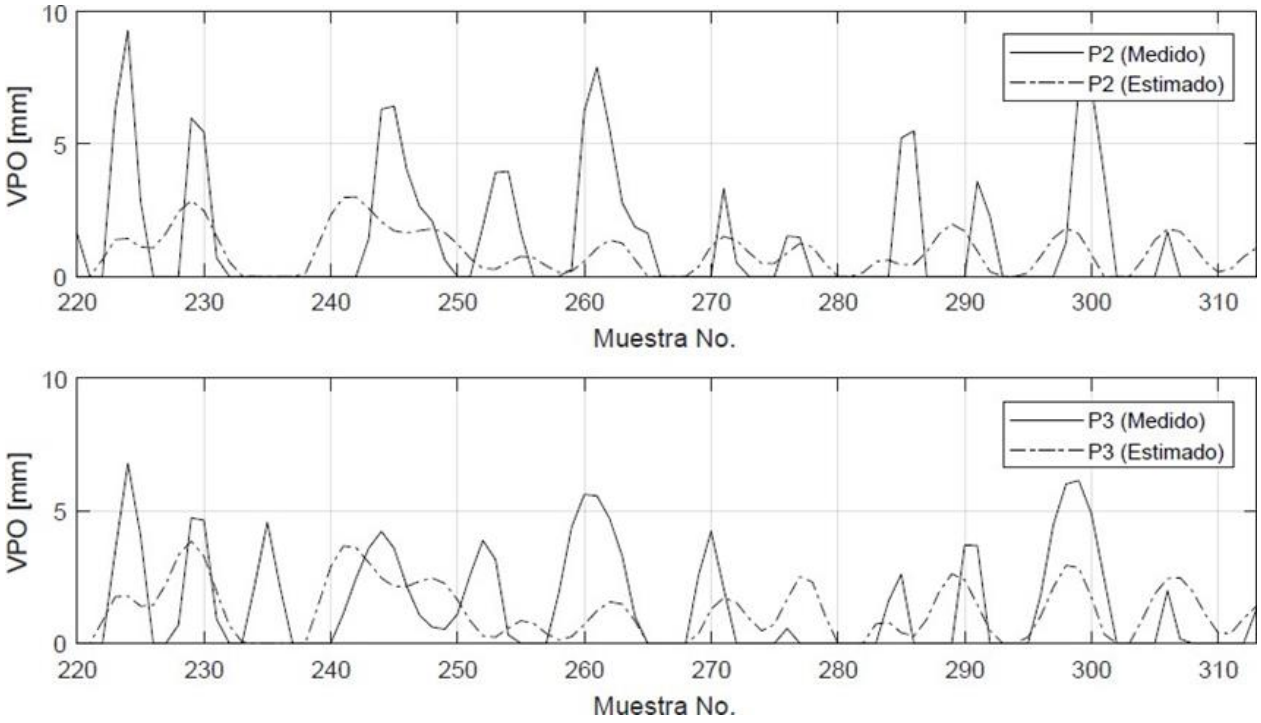
Apéndice C. Datos medidos y estimados para el grado de apertura de los hablantes.



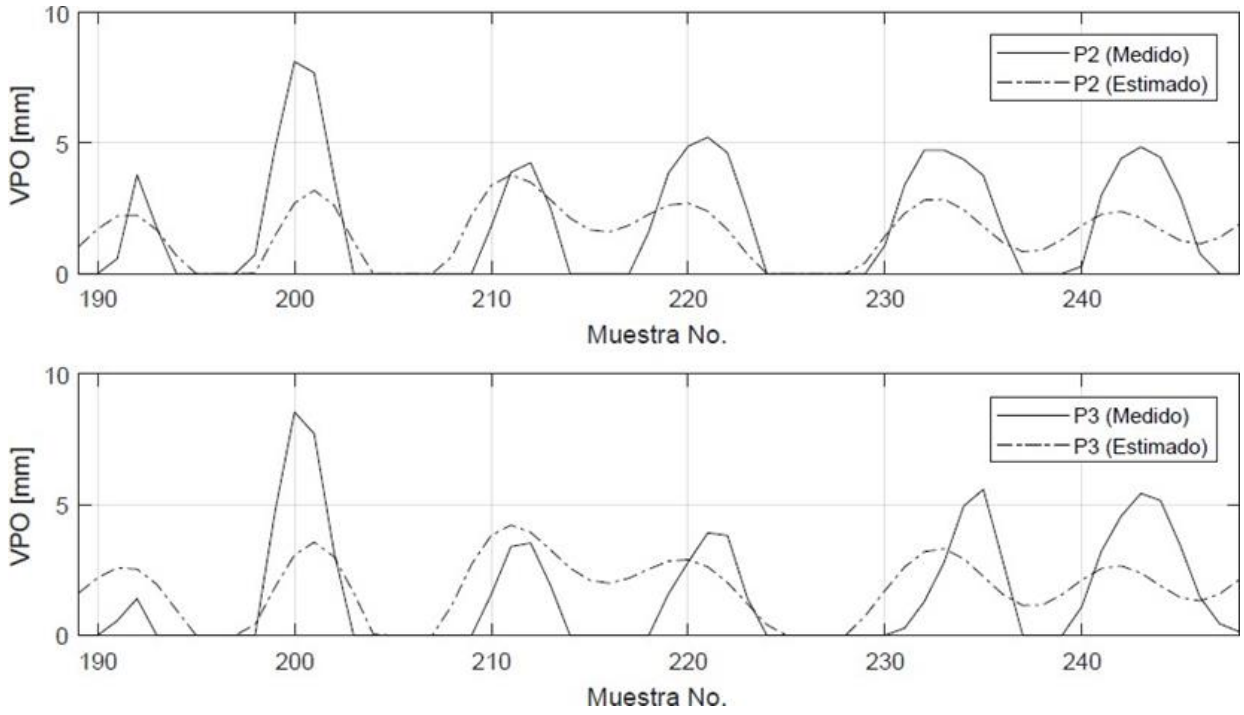
Valores medidos y estimados del grado de apertura del hablante F2 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.



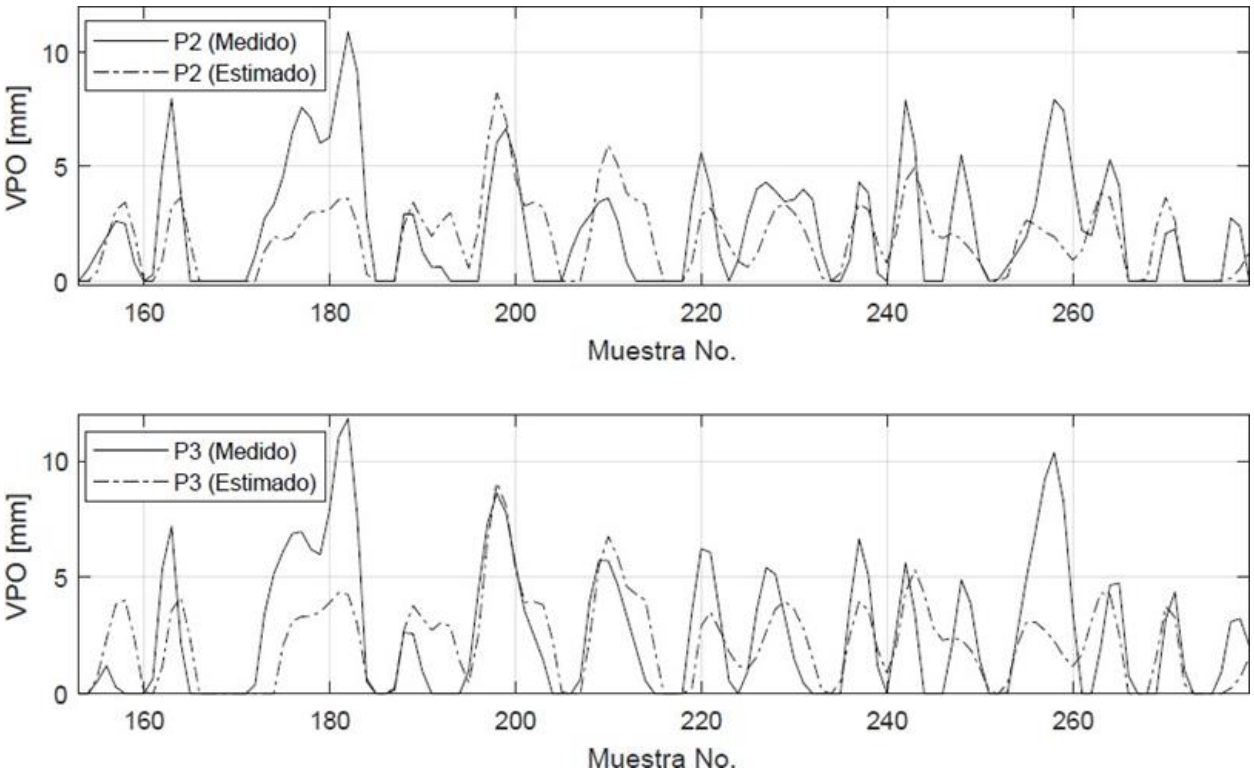
Valores medidos y estimados del grado de apertura del hablante F3 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.



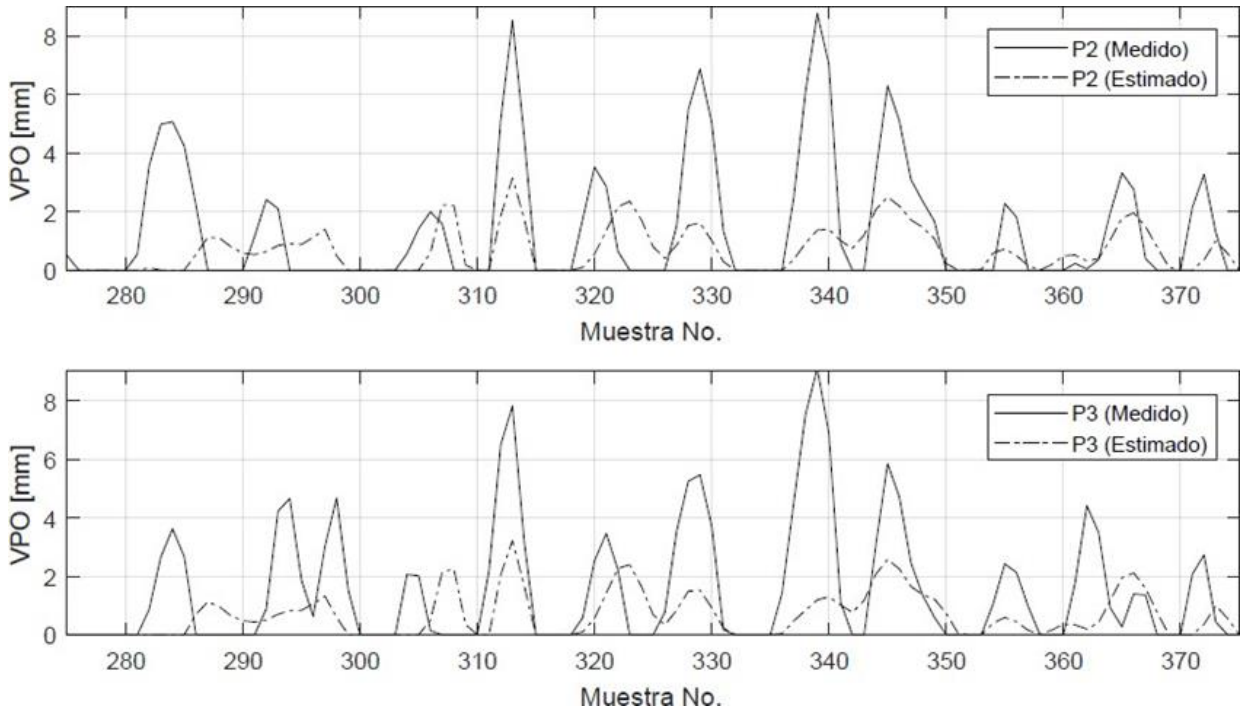
Valores medidos y estimados del grado de apertura del hablante F4 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.



Valores medidos y estimados del grado de apertura del hablante M1 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.

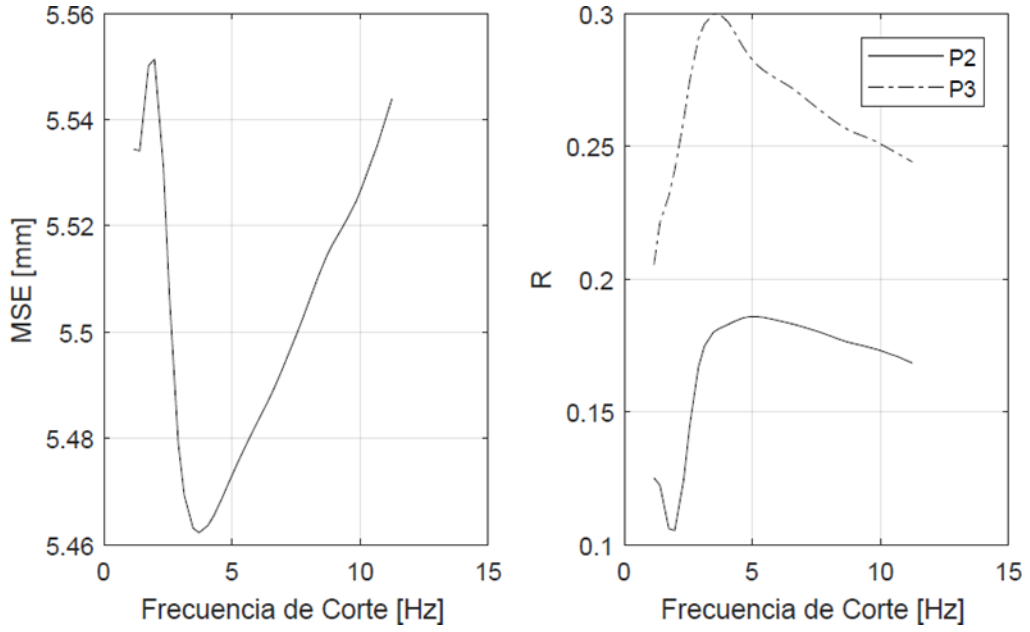


Valores medidos y estimados del grado de apertura del hablante M3 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.

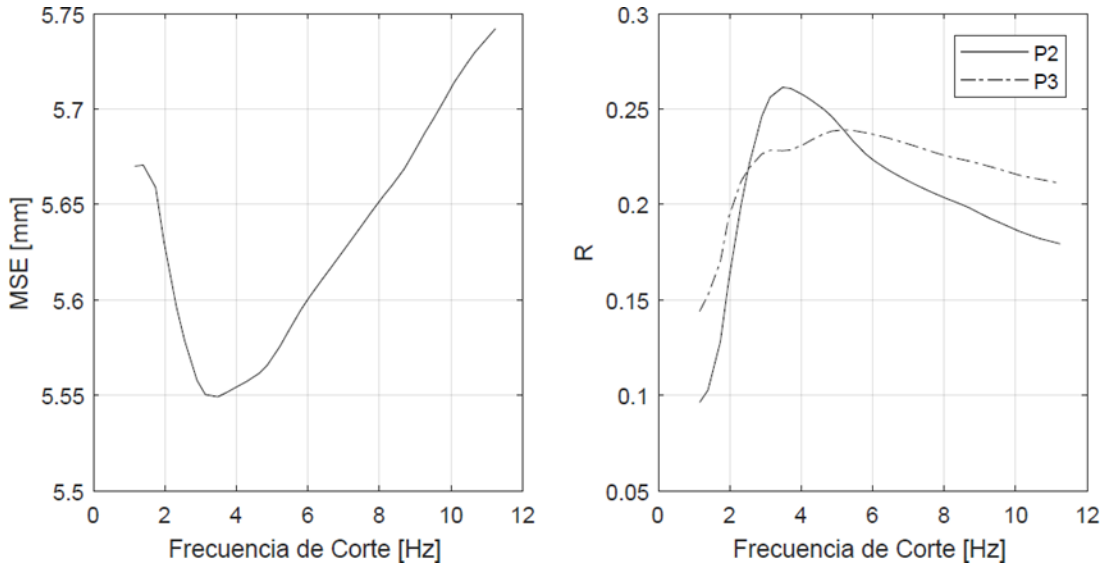


Valores medidos y estimados del grado de apertura del hablante M4 para los puntos de referencia 2 **(Superior)** y 3 **(Inferior)** del velo.

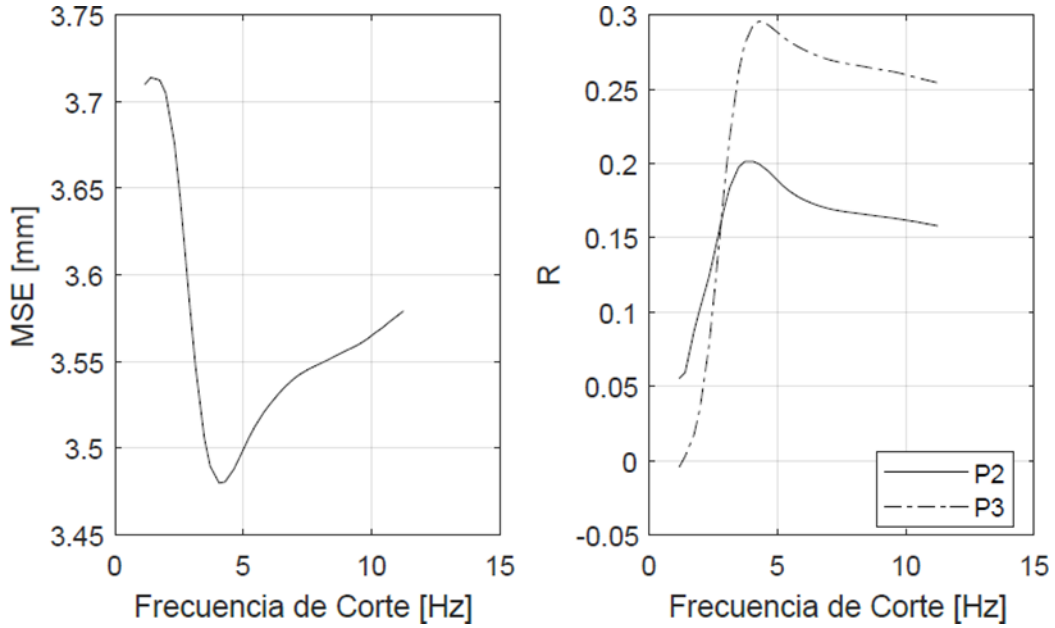
Apéndice D. MSE y R versus frecuencia de corte de filtro para datos estimados.



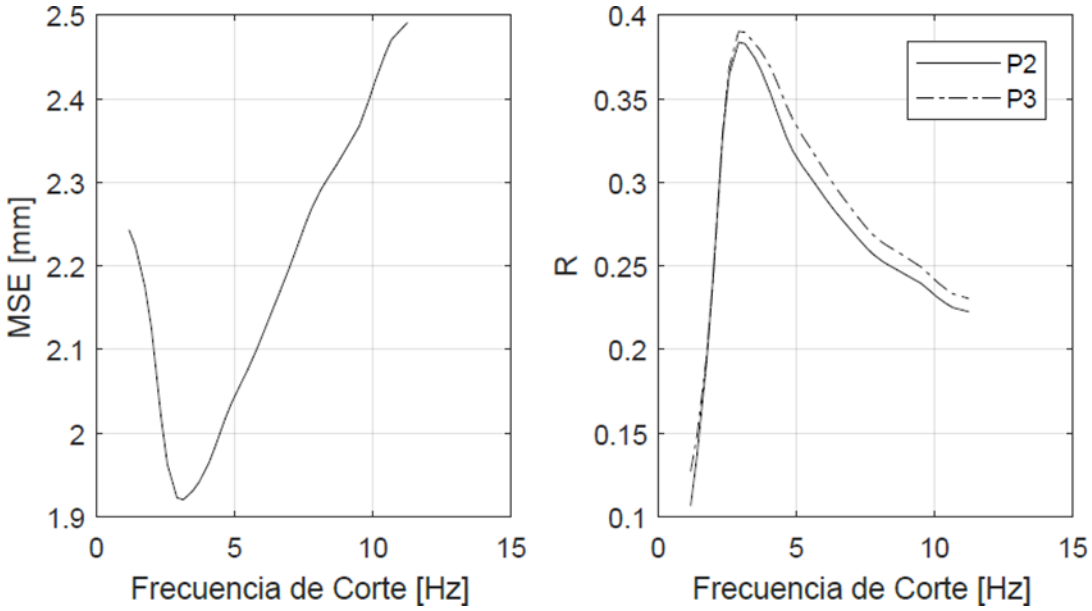
MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante F2.



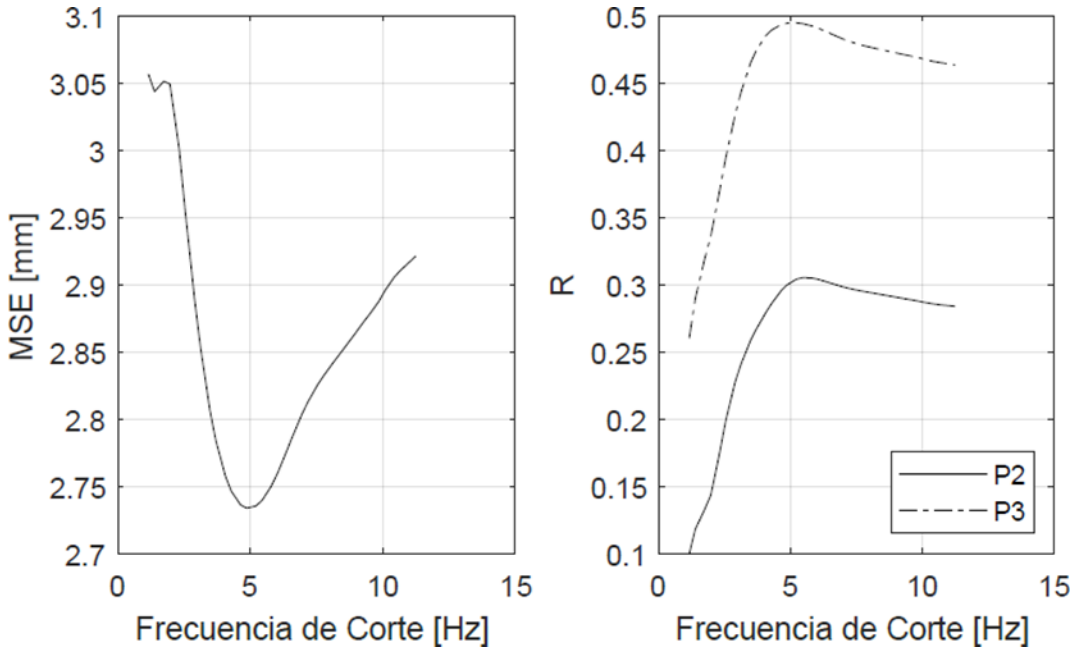
MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante F3.



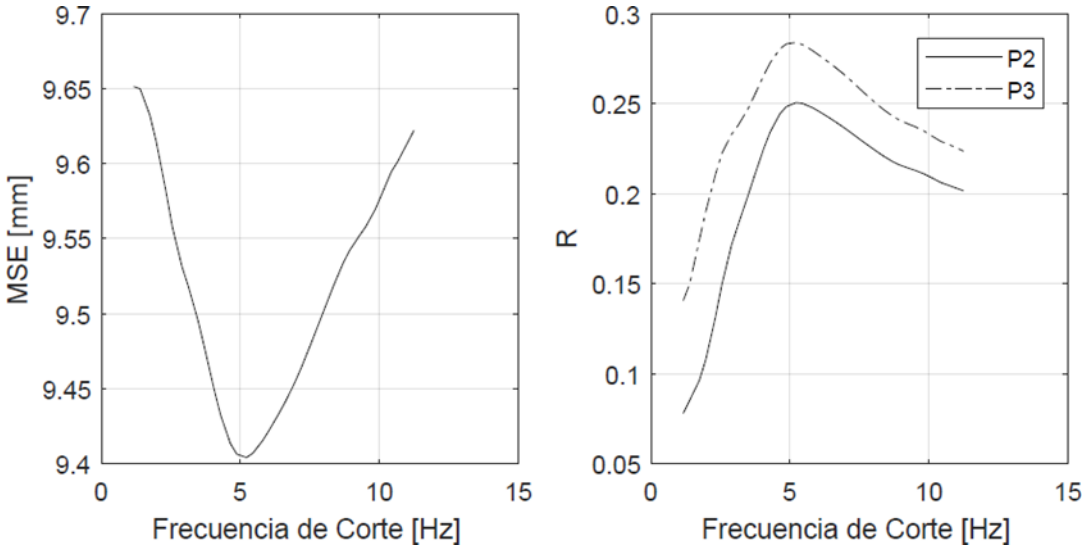
MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante F4.



MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante M1.



MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante M3



MSE [mm] vs frecuencia de corte [Hz] y R vs frecuencia de corte [Hz], para el hablante M4.