Facultad de Ciencias

MÉTODOS AUTOMÁTICOS PARA LA CUANTIFICACIÓN DE LA CALIDAD DE LA VOZ

(Automatic methods for voice quality quantification)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN DATA SCIENCE

Autor: Julia Fábrega Torrano

Director: Steven Van Vaerenbergh

Índice

Agradecimientos

En primer lugar, me gustaría agradecer a Steven, mi tutor del TFM, por su apoyo para sacar este trabajo adelante, especialmente por su disponibilidad y su ayuda a la hora de afrontar nuevas tareas en el proyecto.

También me gustaría agradecer a mi familia, por preocuparse y estar conmigo en cada paso que he dado este año y a mis amigos, por estar conmigo en los buenos momentos y apoyarme en los no tan buenos.

Resumen

El objetivo de este trabajo es estudiar la cuantificación de la calidad acústica de la voz mediante métodos automáticos, tanto en voz hablada como en canto, con el fin de evaluar la mejoría en la calidad de la voz después de un tratamiento experimental específico. Para ello, se dispone de un dataset limitado de grabaciones de voces habladas y de canto, tanto de antes del experimento como de después, junto con una serie de etiquetas de valoraciones subjetivas proporcionadas por expertos.

Después de estudiar varios métodos relacionados con el Acoustic Voice Quality Index, que es una métrica estandarizada para medir la calidad de una voz, el estudio se centra en métricas que cuantifican la presencia de armónicos y la calidad de vibratos. Los resultados del estudio indican que las valoraciones subjetivas de los expertos no se pueden modelar mediante técnicas automáticas, posiblemente debido a la gran variedad en etiquetas y el número bajo de grabaciones. Tampoco se observa una mejoría en la calidad de la voz después del experimento según las métricas estudiadas.

Palabras clave: Análisis acústico de la voz, AVQI, calidad de la voz, HNR, vibrato.

Abstract

The aim of this work is to study the quantification of the acoustic quality of the voice by means of automatic methods, both in spoken and singing voice, in order to evaluate the improvement in the acoustic quality of the voice after a specific experimental treatment. For this purpose, a limited dataset of spoken and singing voice recordings is available, both before the experiment and after, along with a series of subjective rating labels provided by experts.

After studying various methods related to the Acoustic Voice Quality Index, which is a standardized metric for measuring voice quality, the study focuses on metrics that quantify he presence of harmonics and the quality of vibrates. The results of the study indicate that the subjective ratings of the experts cannot be modeled by automatic techniques, possibly due to the large variety in labels and the low number of recordings. There is also no improvement in voice quality after the experiment according to the metrics studied.

Keywords: Acoustic voice analysis, AVQI, voice quality, HNR, vibrato.

1. Introducción

En el mundo médico los tratamientos de la voz están en constante evolución e innovación. Un ejemplo de innovación en este campo sería un estudio experimental que busca mejorar la elasticidad de las cuerdas vocales mediante una experiencia de realidad virtual inmersiva [1], estudio del cual se han obtenido los datos para realizar este trabajo. Esta constante evolución genera la necesidad buscar métodos que cuantifiquen de forma objetiva la mejoría en la calidad de la voz. De esta necesidad nace el presente Trabajo de Fin de Máster que es una colaboración con el Centro de Foniatría y Logopedia de Santander.

El objetivo de este trabajo es estudiar la cuantificación de la calidad de la voz para poder aplicar métodos de Machine Learning que clasifiquen automáticamente las voces según su calidad.

Para realizar esta tarea se cuenta con unas grabaciones de voz de unas personas que han sido sometidas a un experimento de realidad virtual que tenía la finalidad de disminuir la rigidez y eutonía vocal. Para cada sujeto se cuenta con grabaciones de su voz antes y después del experimento. Se observaron cambios en la percepción de los sujetos: aumento de la sensación de flexibilidad vocal, el aumento de la sensación de resonancia (la diferencia más notable) y la ganancia de una voz más artística, más redonda, con más armónicos.

En este trabajo se estudia si esa mejora de sensaciones de las personas al ser sometidas al experimento va acompañada de un aumento de la calidad de la voz y en caso afirmativo, obtener los indicadores objetivos que lo avalen.

Con este fin, se debe encontrar una medida o un conjunto de parámetros que indiquen de forma objetiva la calidad de la voz. Para poder encontrar estos indicadores, se ha llevado a cabo el denominado análisis acústico de la voz que es un método objetivo de evaluación de la voz que se utiliza con fines terapéuticos. Una ventaja de este método es que es un método no invasivo, es decir, que no hay ninguna forma de dañar la integridad física del paciente. Para llevar a cabo el análisis de la voz se toman señales acústicas de voz hablada o cantada y señales de vocal sostenida. Las voces se graban con micrófonos profesionales para conseguir unas condiciones que permitan efectuar un análisis objetivo de la voz. Posteriormente, las grabaciones se tratan con programas especializados en el tratamiento de señales [2].

En la grabación de las voces es muy importante evitar el ruido ambiental para que no se alteren los resultados. Además, se tiene en cuenta que la fonación de una persona puede variar de unos momento a otros, pero esta variación es aceptable siempre y cuando se mantenga dentro de ciertos límites. Dos grabaciones de voz de un mismo sujeto en las mismas condiciones deberían ser lo suficientemente similares como para poder confiar en ellas [3].

Adicionalmente, se considera interesante realizar este análisis objetivo porque podría servir para completar otros métodos como el diagnóstico perceptual de la voz o la exploración laringoestroboscópica. El diagnóstico perceptual es un método subjetivo en el que un especialista puntúa la voz aplicando escala. Una fuente de subjetividad de este método sería la experiencia del especialista en la escucha de voces con distintos desórdenes. Un ejemplo de escala es la de GRBAS en la que se puntúan 5 parámetros para evaluar la calidad de la voces: grade indica el grado de afectación general y global de la voz; roughhness refleja el grado de ronquera; breathiness muestra el grado de voz soplada que ocurre cuando se pierde aire a través de las cuerdas vocales, asthenics indica el grado de fatiga de la voz y strain refleja el grado de tensión o dureza de la voz [4, 5]. Por otro lado, la exploración laringoestroboscópica es aquella en la cual el médico visualiza las cuerdas vocales en movimiento y evalúa su funcionamiento. Las cuerdas vocales vibran a una velocidad que no puede ser percibida por el ojo

humano, por lo que se hace uso de luz estroboscópica que ilumina las cuerdas vocales con breves destellos de luz a una frecuencia similar a la de vibración de las cuerdas vocales. Las cuerdas vocales se graban vibrando y se valora el cierre glótico, la periodicidad de los ciclos vocales o la simetría de las cuerdas vocales.

Para solventar los problemas de subjetividad y poder complementar los métodos explicados, se propone el análisis acústico de la voz. Es un método objetivo cuyo resultado permanece inalterado para una misma muestra de voz debido a que el resultado se obtiene a partir de una serie de parámetros obtenidos por un algoritmo, en vez de depender de la percepción que pueda tener el especialista de la voz.

Para el análisis acústico existen muchos programas. En este trabajo se ha utilizado un programa que se llama Praat. Praat es un programa de distribución libre que se suele utilizar para el análisis acústico en el ámbito médico cuyos resultados presentan alta fiabilidad [6]. Para utilizar Praat se ha hecho uso de una librería de Python llamada Parselmouth. Parselmouth es capaz de acceder al código interno de Praat, es decir, que los algoritmos y salidas son iguales que en Praat [7].

En lo tocante a la ciencia de datos, existe una amplia literatura en los métodos de Machine Learning aplicados a grabaciones de voz. La gran mayoría se centra en los siguientes campos:

1. El reconocimiento de voz para el entendimiento del habla. Por ejemplo, en el artículo [8] se da una visión general de los métodos *end to end* en el campo del reconocimiento automático de la voz. Dentro de estos métodos se encuentran el encoder-decoder basado en mecanismos de atención que hacen uso, por ejemplo, de Transformers o las Redes Neuronales Recurrentes transductoras (RNN-t).
2. La identificación de una persona a partir de su voz. En particular, en 2021 se entrenaron distintos métodos de aprendizaje automático para la detección de impostores mediante el habla. Los métodos que mejores resultados dieron fueron el random forest y el clasifficador Naive Bayes [9].
3. La síntesis de voz para generar audio a partir de texto escrito. Por ejemplo, en 2017 se propuso un sistema de Deep Learning para sintetizar voz a partir de texto en tiempo real [10].

A pesar del avance que está ocurriendo en este campo, la literatura acerca de la cuantificación de la calidad de la voz con métodos automáticos es muy escasa. En la mayoría de artículos encontrados en la literatura que tratan la calidad de la voz hacen uso de métricas clásicas que han sido diseñadas como la HNR, el jitter o el shimmer para analizar la calidad de la voz (véase Sección 3.2. índice de Calidad Acústica de la Voz). En cambio, en escasos trabajos se mencionan las características aprendidas por métodos de Machine Learning debido a la escasez de trabajos en los que se aplican estos métodos.

A continuación se mencionan algunos artículos sobre la detección de patologías de la voz mediante métodos de ML, las cuales están directamente relacionadas con la calidad de esta. En 2017 se llevó a cabo una investigación preliminar para la detección de patologías de la voz haciendo uso de Redes Neuronales Profundas (DNN) en las cuales se combinaban capas convolucionales con capas Long-Short-Term-Memory (LSTM) sobre las señales de audio en crudo [11]. Se obtuvieron resultados prometedores porque eran comparables a otros estudios que utilizaron otras metodologías. En 2020 se llevó a cabo una comparación de 45 estudios que realizaba una detección de patologías sobre datos de audio en el que se concluyó que el Support Vector Machine (SVM) es el algoritmo más utilizado para la detección de trastornos de la voz y que los investigadores se suelen centrar más en técnicas supervisadas que en técnicas no supervisadas [12]. Por otro lado, se utilizan Redes Neuronales Convolucionales (CNN) como la que se propuso en un estudio que lograba una alta precisión

en la detección de patologías en la voz [13]. En 2021 se propuso una Red Neuronal Recurrente que utilizaba como características el tono, el centroide espectral, el flujo espectral y la energía, entre otras, con la que se obtenían unos resultados muy parecidos a la CNN del anterior artículo [14]. Por otro lado, cabe señalar un artículo [15] en el que se basa en una métrica llamada Mean Opinion Score (MOS) que se utilizaba inicialmente para medir la calidad de las llamadas tradicionales de voz y actualmente se aplica a las llamadas a través de internet. Esta métrica se basa en puntuaciones subjetivas de personas, comprendidas en una escala del 1 al 5. En este artículo se utilizan CNN para predecir la calidad de las llamadas según esta métrica.

En este trabajo, no se ha podido finalmente aplicar estos métodos porque en primer lugar, en muchos casos el código utilizado no se ha publicado y en segundo lugar y más importante, el dataset con el que se cuenta es demasiado pequeño como para entrenar por ejemplo una CNN, además de que las etiquetas con las que se cuenta eran demasiado diversas (véase Sección 3.1. Descripción de los datos)

2. Teoría de la voz

La teoría que mejor explica la producción de la voz es la mioelástica-aerodinámica. Según esta teoría, dos mecanismos diferentes operan para la fonación. Por un lado, la elasticidad de los músculos, los ligamentos y la mucosa y por otro lado, los fenómenos aerodinámicos en los que interviene la presión subglótica, que es una fuerza que separa las cuerdas vocales, y la velocidad de flujo, que genera una presión negativa entre las cuerdas vocales provocando su unión.

La voz se produce en la respiración. Esta consiste en la inspiración, fase en la que los pulmones se llenan de aire y en la espiración, fase en la que el aire se expulsa de manera controlada. En la respiración las cuerdas vocales se encuentran abiertas (figura 2), pero cuando se produce una fonación estas se cierran inconscientemente y comienza el ciclo vibratorio. Esto hace que la presión subglótica (la presión bajo la glotis) aumente y provoque la separación de las cuerdas vocales. En este punto, la elasticidad actúa invirtiendo el sentido del desplazamiento hasta recuperar la posición inicial: se reduce el espacio de las cuerdas vocales, aumenta progresivamente la presión subglótica y aumenta la velocidad del aire que provoca, según el principio de Bernoulli, la disminución de la presión entre las cuerdas vocales. Esta disminución de la presión conduce a la oclusión completa de las cuerdas vocales, dando lugar al comienzo de un nuevo ciclo vocal. El número de repeticiones por segundo de este ciclo corresponde a la frecuencia fundamental o tono. El sonido de la voz de cada persona depende de la forma y tamaño de las cuerdas vocales y de las cavidades resonantes (garganta nariz y boca), además de la velocidad de vibración de las cuerdas vocales [16, 17].

La voz es un sonido y como tal, se propaga en forma de ondas mecánicas longitudinales producidas por la vibración de un cuerpo que se transmiten a través de un medio elástico como el aire. El sonido se puede propagar únicamente por medios materiales cuyas moléculas puedan vibrar. En el caso de la voz, las vibraciones llegan al oído y son transmitidas al cerebro donde se interpretan.

Cuando un cuerpo vibra, normalmente lo hace a varias frecuencias y en consecuencia, la onda de sonido que produce se transmite también a varias frecuencias. La mayoría de los sonidos de la naturaleza están constituidos por la combinación de múltiples ondas de distintas frecuencias. La frecuencia más baja se denomina fundamental y corresponde al primer armónico. El resto de frecuencias son múltiplos enteros de esta y corresponden a los demás armónicos.

Los sonidos en la realidad están compuestos, tanto por armónicos, como por no armónicos. Los sonidos puramente armónicos son periódicos, es decir, que la forma de onda se repite pasado cierto

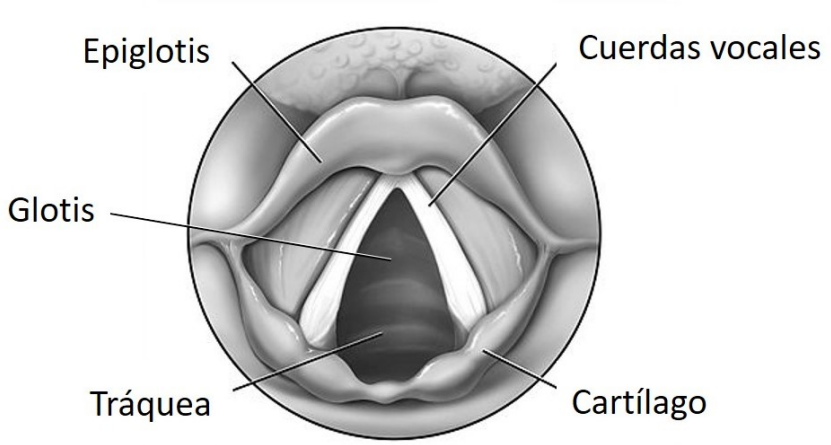


Figura 2: Ilustración de las cuerdas vocales abiertas. Autor: Alan Hoofring.

tiempo T.

Una fonación sostenida de una vocal es un sonido aproximadamente periódico, y este tipo de sonidos son la base de varias de las pruebas en el análisis acústico de la voz. Por otro lado, una frase hablada es un sonido complejo cuyas características cambian rápidamente en el tiempo y en su totalidad es un sonido aperiódico. Estos hechos se pueden observar en los siguientes oscilogramas. La figura 3 representa la variación de la presión del sonido en función del tiempo para una muestra de voz habla continua. Como se puede ver, hay espacios en blanco correspondientes a los silencios. La figura 4 representa lo mismo para una muestra de vocal sostenida en la cual no se ven espacios en blanco porque no hay silencios.

Para poder apreciar la periodicidad de la fonación de vocal sostenida, se han representado las figuras 6 y 8 para ventanas de tiempo menores. Se han representado, para las mismas ventanas de tiempo, las gráficas correspondientes a la muestra de voz de habla continua. Estas se representan en las figuras 5 y 7. Se aprecia cómo para el habla continua no se tiene una onda periódica.

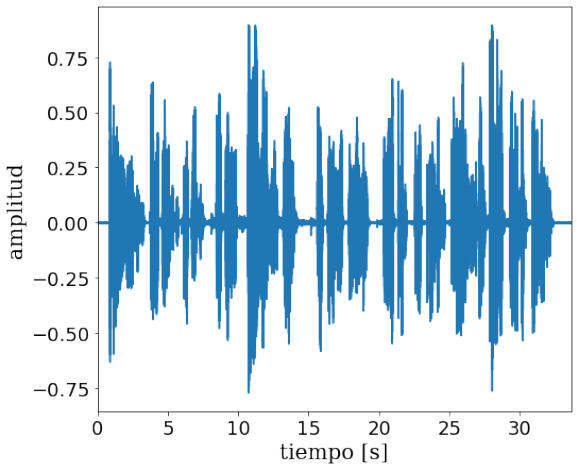


Figura 3: Oscilograma de una muestra de voz de habla continua.

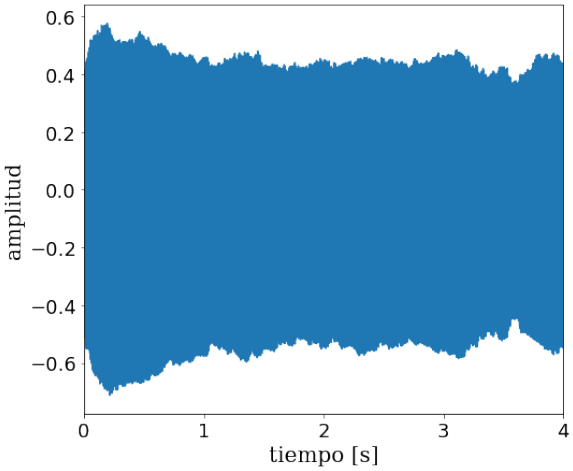


Figura 4: Oscilograma de una muestra de voz de “a” sostenida.

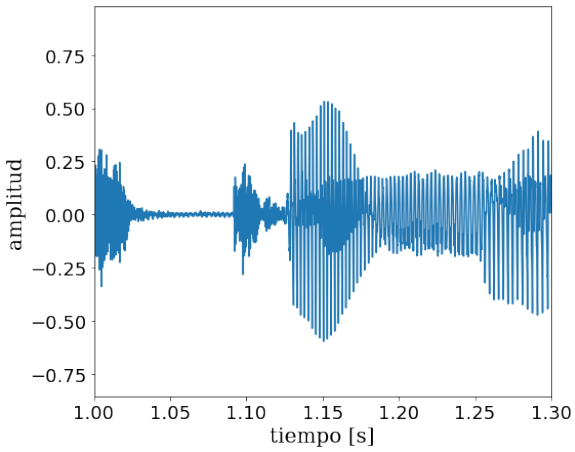


Figura 5: Ventana de 0.3 segundos del oscilograma de la figura 3.

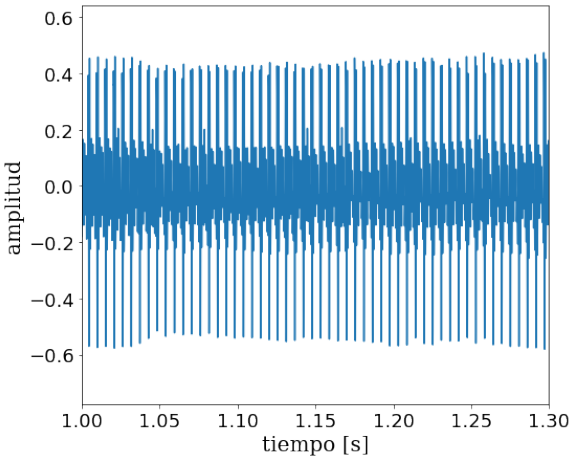


Figura 6: Ventana de 0.3 segundos del oscilograma de la figura 4.

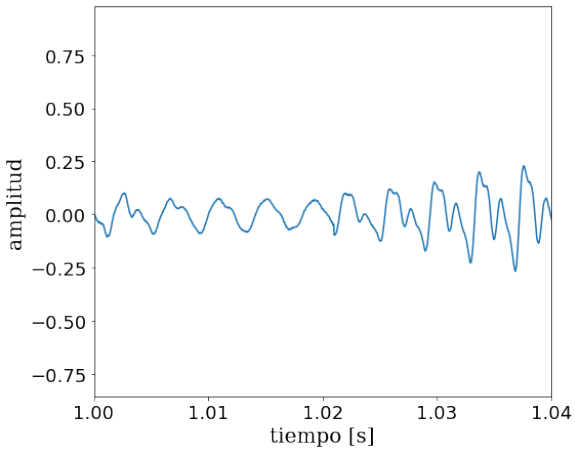


Figura 7: Ventana de 0.04 segundos del oscilograma de la figura 3.

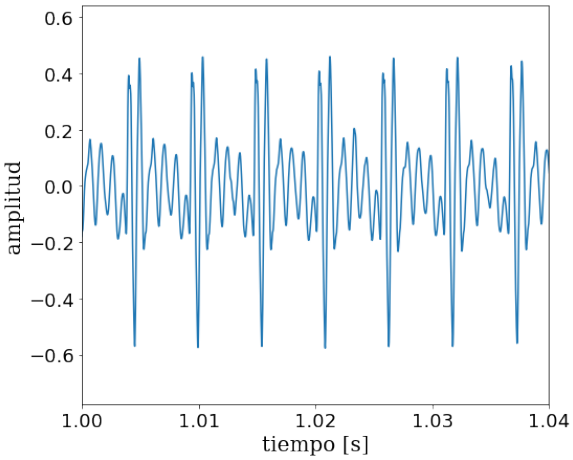


Figura 8: Ventana de 0.04 segundos del oscilograma de la figura 4.

En este trabajo se han estudiado grabaciones de voces cantadas. En algunas de ellas, los cantantes hacen vibrato que es un recurso que utilizan para aportar belleza a la voz. Este método consiste en una modulación periódica sinusoidal de la frecuencia fundamental de fonación. Además de la modulación de la frecuencia fundamental, paralelamente varían las frecuencias de los demás armónicos [18, 19].

3. Materiales y Metodología

3.1. Descripción de los datos

Para llevar a cabo este trabajo, se cuenta con dos muestras de datos. Estas muestras contienen grabaciones de voz de varios sujetos que han sido sometidos a un experimento de realidad virtual que tenía la finalidad de buscar la viabilidad de reducir o eliminar la rigidez del sistema fonatorio [1].

El experimento consistió en una serie de fases. En primer lugar se grababa a la persona cantando un fragmento y realizaba un cuestionario para conocer su percepción sobre ciertos parámetros: rigidez, esfuerzo vocal, volumen, potencia, proyección, resonancia, afinación, limpieza, vibrato y nasalidad. Posteriormente, la persona cantaba otro fragmento diferente en posición normal, después sobre una plataforma inestable y finalmente combinando la plataforma inestable con realidad virtual que simulaba una montaña rusa. Finalmente, se volvía a grabar el mismo fragmento del principio y se realizaba el mismo cuestionario sobre la percepción. De esta forma, se tenía una grabación del mismo fragmento antes y después de haber sido sometido a realidad virtual.

Las grabaciones se llevaron a cabo en la sala de grabación de Estudio Ovni localizado en Asturias, con un micrófono AudioTechnica A4033 y con el programa de grabación Pro Tools. Están guardadas en formato .wav.

Como se ha mencionado, para realizar este estudio se cuenta con dos muestras de datos. Una de ellas contiene voces de 34 sujetos (20 mujeres y 14 hombres). Para cada sujeto se tienen 2 audios grabados antes del experimento, uno de ellos de vocal sostenida y otro de habla continua y 2 audios grabados después del experimento, también de vocal sostenida y de habla continua. Los audios de vocal sostenida corresponden a la letra "a" y los audios de habla continua son tanto de voz cantada (31 audios), como hablada (3 audios). En los audios de habla continua, cada persona dice la misma oración o canta el mismo fragmento de canción en las grabaciones anteriores y posteriores al experimento. En la tabla 1 se muestra una clasificación de los audios de habla continua. También se indica si el audio cantado corresponde a canto gutural porque resulta de interés para el análisis.

sexo / tipo	Hablado	Cantado
Mujer	1, 3, 10	2, 4, 5, 6, 8, 12, 14, 16, 18, 19, 20, 22, 23, 25, 27, 28, 30 (gutural)
Hombre		7, 9, 11, 13, 15, 17, 21, 24, 26, 29 (gutural), 31 (gutural), 32 (gutural), 33, 34

Tabla 1: Clasificación de los audios de habla continua, identificados con un número, correspondientes a la primera muestra de audios. Se clasifican en función del sexo de la persona y del tipo de audio, según es cantado o hablado.

Para la mayoría de las grabaciones de habla continua se dispone de un documento con una serie de etiquetas que incluyen la valoración subjetiva de una experto. Se han recogido en la tabla 2. Como se puede ver, existe una gran cantidad de etiquetas diferentes que hace que difícilmente sean utilizadas para poder generar una modelo de Machine Learning. Además, son etiquetas comparativas en el sentido de que utilizan términos como "más" o "mejor" que indican un cambio con respecto a la situación anterior al experimento, pero no indican algo objetivo y bien definido que hubiera sido valorado sin conocer la grabación anterior. Además cada etiqueta solamente se refiere a un fragmento concreto de la grabación y no a toda en su conjunto.

Archivo	Etiquetas Pre	Etiquetas post
4	voz forzada	Voz con mejor cierre
5	sonido más apretado y algo forzado	Voz más ágil

Archivo	Etiquetas Pre	Etiquetas post
6	afinación imprecisa	mejor afinado y voz más segura
7	voz más apretada, más inestable	Voz más firme
9	sonido más apretado y descontrolado	sonido más libre, mejor control en el sonido rajado y mejor definido
11	sonido algo apretado, forzado, no define bien la nota	sonido más libre, más ligero, afinación más precisa
12	Voz algo inestable y algo aireada	Voz más segura, más precisa, mejor cierre glótico
13	Voz inestable, golpes de aire, algo empujada	Voz mejor controlada, mejor cierre glótico y menos esfuerzo
14	voz inestable, poco precisa	mejor estabilidad, afinación y control
17	sonido pequeño, inestable	sonido con más cuerpo, más presencia y más preciso.
18	voz pequeña, tímida, con ruido	sonido con más cuerpo, más proyectado, más preciso, libre y limpio
19	voz más inestable, vibrato algo descontrolado	voz más estable, con más cuerpo y menos vibrato no controlado
20	voz inestable, no proyectada	voz más grande, más precisa, con más cuerpo
22	voz suave, insegura	mejor cierre glótico, más descontrolada
23	voz más pequeña, mejor controlada	voz más grade y más descontrolada
24	voz inestable, vibrato incontrolado, mejor cierre	voz más aireada, menos controlada
26	voz fuerte, descontrolada, peor afinada	voz más precisa, más controlada
27	afinación menos precisa, voz con menos cuerpo	voz con más cuerpo, mayor seguridad y mejor afinación
28	voz más inestable, menos flexible	voz más segura, más precisa y con más cuerpo
30	sonido gutural peor definido	sonido gutural mejor definido, sonido más redondo
31	rajado más inestable, posiciones menos definidas	sonido mejor definido, voz más flexible
32	colocación y rajado menos precisos, sonido más inestable	rajado más preciso, mejor ejecutado, más libre
33	voz rígida, sonido algo cerrado cerrado	sonido más abierto, sonido más grande y flexible
34	sonido más apretado, más rígido, más pequeño	sonido más grande, voz más flexible y mayor precisión en afinación

Tabla 2: Etiquetas dadas por un experto sobre las voces de habla continua grabadas antes y después del experimento. Están identificadas con un número.

La otra muestra cuenta con grabaciones de 3 mujeres que cantan haciendo vibrato. Para cada mujer se tiene un audio cantado antes del experimento y otro después de este. Para estas grabaciones no se dispone de etiquetas.

3.2. Índice de Calidad Acústica de la Voz

El Índice de Calidad Acústica de la Voz, en inglés Acoustic Voice Quality Index (AVQI) es una medida que se desarrolló para superar la validez limitada que tenía utilizar un solo parámetro para cuantificar la calidad general de la voz [20]. Es una medida basada en seis parámetros acústicos que sirven para cuantificar la calidad general de la voz, siendo el pico cepstral de mayor prominencia en el cepstrum suavizado (CPPS), el principal parámetro.

En la literatura científica hay varios estudios que usan esta métrica para cuantificar la calidad de las voces. Por ejemplo, en 2019 se publicó un estudio [21] en el que se dividieron las muestras de voz en dos grupos, la primera sin disfonía y la segunda con ligera disfonía. Para clasificar las voces intervinieron tres oyentes y se utilizó el primer parámetro de la escala de GRBAS, es decir, G que indicaba el grado de afectación general de la voz (G = 0: normal o ausencia de ronquera, G = 1: voz ligeramente ronca, G = 2: moderadamente ronca, G = 3: severamente ronca). Solamente eran determinadas como normofónicas aquellas voces que obtenían una puntuación unánime de los tres oyentes de G = 0 que indicaba ausencia de ronquera. El otro grupo se dividió en tres subgrupos según el nivel de disfonía en función del número de oyentes que puntuaron con G = 1 (1 = mayoría de normofonía, 2 = mayoría de ligera disfonía, 3 = unánime disfonía leve).

En este estudio se utilizó la versión v.02.06 del AVQI para evaluar las voces [20]. No se encontraron efectos significativos debidos al género, lo que parece sugerir que las diferencias anatómicas vocales entre hombres y mujeres no afectan al AVQI. Por otro lado, a pesar de obtener valores mayores de AVQI para aquellos sujetos sin disfonía, esta diferencia no era suficientemente significativa para diferenciar el grupo de voces normofónicas del grupo de voces ligeramente disfónicas. Se encontró solamente una diferencia significativa para el AVQI entre el grupo con G = 0 y el segundo subgrupo, es decir, el grupo para el cual 2 oyentes habían votado con G = 1 para esas voces. Se concluía que era necesaria una muestra mayor de voces para tener un mayor poder estadístico y además implementar en un futura la versión novedosa de AVQI (v.03.01).

En este trabajo se ha calculado el AVQI por un lado con la versión 02.06 que corresponde a la ecuación 1 y la versión 03.01 [22] indicada en la ecuación 2. Estas ecuaciones se han obtenido mediante un proceso iterativo de regresiones lineales múltiples para obtener un modelo que representara la mejor combinación de los predictores acústicos [20].

$$\begin{aligned} AVQI_1 =&[(3{,}295 - (0{,}111 \cdot CPPS) - (0{,}073 \cdot HNR) - (0{,}213 \cdot Shim) \\ &+ (2{,}789 \cdot shdB) - (0{,}032 \cdot slope) + (0{,}077 \cdot tilt)) \cdot 2{,}208] + 1{,}797 \end{aligned} \tag{1}$$

$$\begin{aligned} AVQI_2 =&(4{,}152 - (0{,}177 \cdot CPPS) - (0{,}006 \cdot HNR) - (0{,}037 \cdot Shim) \\ &+ (0{,}941 \cdot shdB) + (0{,}01 \cdot slope) + (0{,}093 \cdot tilt)) \cdot 2{,}8902 \end{aligned} \tag{2}$$

Los estudios de la literatura se han llevado a cabo para grabaciones de voces habladas y de vocal sostenida. En este trabajo se cuenta con grabaciones se vocal sostenida, pero se cuenta con grabaciones de voces cantadas en su mayoría (se tienen solo 3 grabaciones de voz hablada). Por lo que se va a estudiar si se obtienen los resultados válidos con este tipo de grabaciones.

A continuación, se explican las variables obtenidas por Parselmouth que se utilizan para calcular el AVQI, además de otras que también han sido obtenidos para su posterior análisis.

Perturbación de la frecuencia

La perturbación de la frecuencia o **jitter** se define como los cambios abruptos y rápidos en periodos adyacentes de la frecuencia fundamental de la onda, es decir, es la perturbación de la frecuencia a corto plazo [23]. Una forma de expresar el jitter es el jitter medio absoluto o **jitta** que se calcula como el promedio de las diferencias en duración de periodos consecutivos. Si una fonación consta de N periodos, el jitta se calcula de la siguiente forma:

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - T_{i+1}\right| \tag{3}$$

donde T_i es la duración del i-ésimo periodo y T_{i+1} la duración del $i+1$-ésimo periodo. Esta medida se expresa en microsegundos.

El **jitter (local)** representa el promedio de la diferencia absoluta entre dos periodos consecutivos dividido por el periodo promedio [24]. La ecuación para obtenerlo es la siguiente:

$$jitt = \frac{jitta}{\frac{1}{N}\sum_{i=1}^{N} T_i} \cdot 100 \tag{4}$$

Por otro lado, el **jitter (rap)** representa el promedio de la perturbación [24], es decir, el promedio de la diferencia absoluta entre un periodo y el promedio del periodo con sus dos vecinos, dividido entre el promedio del periodo.

$$jitter(rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|T_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \cdot 100 \tag{5}$$

Finalmente, el **jitter (ppq5)** representa el ratio de la perturbación de de 5 períodos [24], es decir, el promedio de la diferencia absoluta entre un periodo y el promedio del periodo de los cuatro vecinos más cercanos, es decir, dos periodos anteriores y dos periodos posteriores, dividido por el promedio del periodo. Se calcula de la siguiente forma:

$$jitter(ppq5) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-2}\left|T_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \cdot 100 \tag{6}$$

Cuanto mayor es el jitter, mayor variabilidad tiene la frecuencia fundamental.

Perturbación de la intensidad

La perturbación de la intensidad o **shimmer** hace referencia a las variaciones de intensidad entre periodos consecutivos [23]. Una forma de expresarlo es el **shimmer (local)** que representa el promedio de la diferencia absoluta entre las amplitudes de dos periodos consecutivos, dividido por el promedio de las amplitudes [24]. Para una fonación consta de N periodos se calcula:

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - A_{i+1}\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \cdot 100 \tag{7}$$

donde A_i es la intensidad del i-ésimo periodo y A_{i+1} la intensidad del $i+1$-ésimo periodo.

Por otro lado, el **shimmer (local, dB)** representa el promedio del logaritmo en base 10 de la diferencia entre dos periodos consecutivos [24]:

$$shdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 \cdot \log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{8}$$

El shdB se expresa en decibelios.

Por otro lado, el **Shimmer (apq3)** representa el cociente de la perturbación de la amplitud en tres periodos [24], en otras palabras, el promedio de la diferencia absoluta entre la amplitud de un periodo y la media de las amplitudes de sus dos vecinos, dividido entre el promedio de la amplitud:

$$shimer(rap) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A_i - \left(\frac{1}{3}\sum_{n=i-1}^{i+1}A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \cdot 100 \tag{9}$$

Por último, el **Shimmer (apq5)** representa el ratio de la perturbación de la amplitud de 5 períodos [24], en otras palabras, el promedio de la diferencia absoluta entre la amplitud de un periodo y el promedio de la amplitud de esta y sus cuatro vecinos más cercanos dividido por el promedio de la amplitud. La ecuación para calcular lo es la siguiente:

$$shimmer(apq) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-2}\left|A_i - \left(\frac{1}{5}\sum_{n=i-2}^{i+2}A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \cdot 100 \tag{10}$$

Cuanto mayor sea el shimmer, mayor variabilidad presenta la intensidad.

La energía del ruido glótico

La relación ruido a armónico o **noise-to-harmonics ratio** (NHR) mide la energía total emitida por la voz y la energía que va contenida en los armónicos [23]. En este trabajo se hace uso de la relación inversa llamada relación armónico a ruido o **harmonics-to-noise-ratio** (HNR) debido a que es la variable que se utiliza para calcular el AVQI.

La energía total E_t de un sonido se calcula como la suma de la energía armónica denotada como E_a y la energía del ruido denotada como E_r:

$$E_t = E_a + E_r \tag{11}$$

La energía armónica es la suma de las contribuciones de los armónicos a la energía y la energía del ruido se obtiene como las contribuciones del ruido a la energía . Si la voz no presentase ruido, es decir, que solamente tuviera los armónicos, la energía total sería igual a la energía armónica.

La HNR es una medida instantánea que se puede obtener para distintos instantes de tiempo. Para el cálculo del AVQI se obtiene el promedio de esta medida. La HNR se obtiene como el cociente de la energía armónica aportada por las frecuencias entre 70 Hz y 4500 Hz y la energía de ruido aportada por las frecuencias entre 1500 Hz y 4500 Hz. Entonces la HNR queda como:

$$HNR = \frac{E_{a(70-4500)}}{E_{r(1500-4500)}} \tag{12}$$

Cuanto mayor es la HNR, implica que la voz tiene menor ruido y por lo tanto tiene una mayor calidad.

El pico cepstral de mayor prominenia en el cepstrum suavizado

Una forma de estudiar el sonido es mediante el espectro que se obtiene al aplicar la transformada rápida de fourier, en inglés Fast Fourier Transform (FFT), a la señal de sonido, que descompone la onda en sus frecuencias individuales. En el espectro se representa la frecuencia en el eje horizontal y la intensidad correspondiente cada frecuencia en el eje vertical, como se puede observar en la figura 9. Cada sinusoide que compone la onda de sonido está representada por una línea en el espectro.

Aplicando la transformada inversa de Fourier al logaritmo del espectro se obtiene el **cepstrum**. El pico cepstral de mayor prominenia en el cepstrum suavizado (CPPS) es la distancia entre el primer pico rahmónico y el punto de igual frecuencia de la recta de regresión a través del cepstrum suavizado. Cuanto más periódica es la señal de voz, es más armónica y presenta un pico espectral más prominente [5].

Slope

La siguiente variable es el **Slope** que es la pendiente de la línea de mínimos cuadrados de mejor ajuste que va desde los armónicos más graves a los agudos.

Tilt

El **tilt** es el cociente entre energía proporcionada por los armónicos graves y la energía proporcionada por los armónicos agudos.

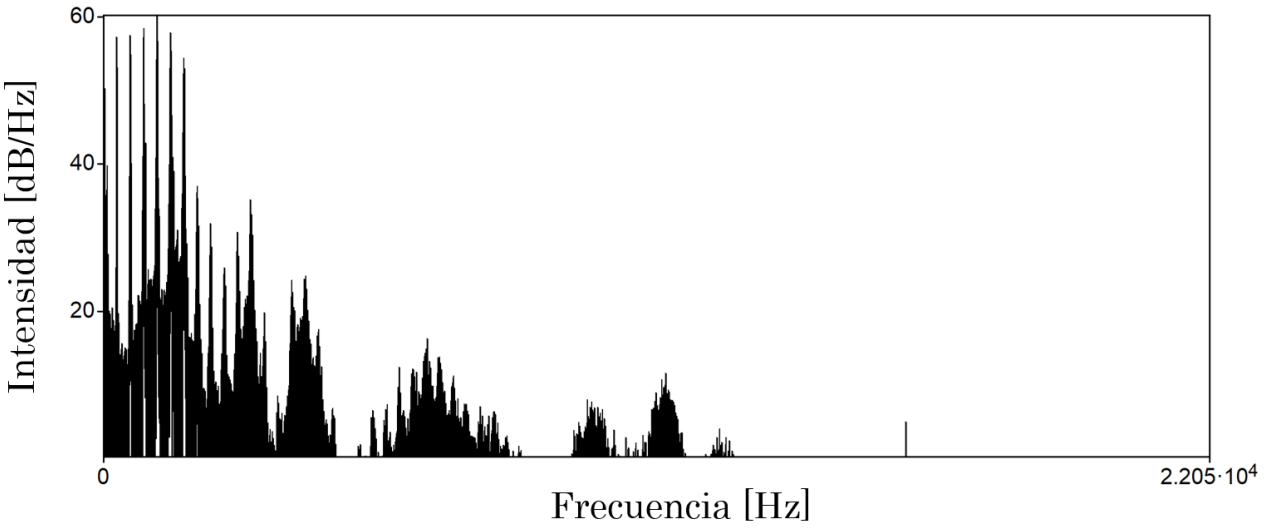


Figura 9: Espectro correspondiente a la grabación de "a" sostenida del sujeto 22 tras el experimento. Se representa la intensidad en decibelios frente a la frecuencia en hercios.

Tansformación de los datos para obtener el AVQI

Para poder aplicar el AVQI, se ha hecho uso de muestras de voz, tanto de habla continua, como de vocal sostenida, debido a la utilidad de combinar ambos tipos de voces. En la mayoría de las clínicas de voz, se hace uso únicamente de grabaciones de vocal sostenida debido que es más sencilla su producción y tiene menor influencia de las variedades dialectales. Pero se ha comprobado que es mejor combinar estas grabaciones con las de habla continua porque las grabaciones de vocal sostenida no representan la voz diaria de la persona [20, 25].

El AVQI se ha obtenido, por un lado, para los audios de vocal sostenida y de habla continua por separado y, por otro lado, se ha obtenido para una combinación de ambos, en el que se concatena el fragmento de 3 segundos de vocal sostenida con el audio de habla continua.

Para obtener el AVQI de las muestras de vocal sostenida, se han extraído los 3 segundos centrales de los audios correspondientes. Para obtener el AVQI de las muestras de habla continua se han eliminado los segmentos sin voz. Una vez hecho este procesamiento de las muestras, para cada sujeto se concatenaron los 3 segundos vocal sostenida grabados antes del experimento con la muestra de habla continua también grabada antes del experimento. Se repitió este procedimiento para las grabaciones posteriores al experimento. En la figura 10 se muestra un ejemplo del resultado de la concatenación. En la gráfica de arriba se muestra el oscilograma que representa la variación de la presión del sonido en función del tiempo y en la gráfica de abajo se muestra el espectrograma que es una representación de la frecuencia frente al tiempo. Cada punto del diagrama representa la intensidad de una frecuencia específica en un instante temporal, donde los colores más oscuros representan una mayor intensidad y los colores más claros una menor.

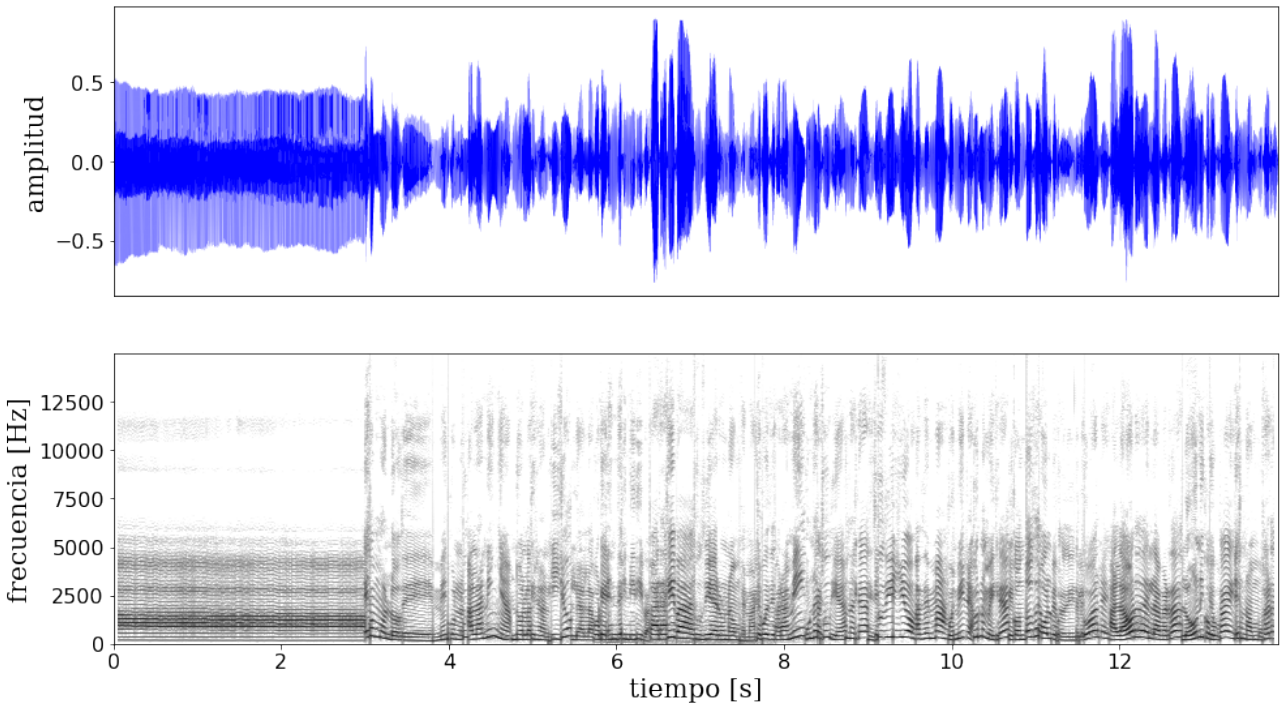


Figura 10: Oscilograma y espectrograma de la muestra de voz concatenada del sujeto 1 anterior al experimento. El área de la izquierda corresponde a los 3 segundos de la "a" sostenida. El área de la derecha corresponde a la grabación de habla continua sin silencios.

3.3. Distribución de HNR

La relación armónico a ruido es una medida que cuantifica la cantidad de ruido en la voz. Es una medida instantánea que se calcula durante ciertos intervalos de tiempo. En general suele ser interesante tomar la media de la HNR y el máximo, aunque en este caso se ha estudiado la distribución de los valores puntuales que se obtienen de la HNR. Se estudia la distribución completa porque interesa conocer si hay más intervalos de HNR alta para las grabaciones posteriores al experimento, lo cual indicaría una voz con mayor tasa de armónicos que los expertos describen como una voz con mayor riqueza de sonido.

En este caso, la HNR se ha medido durante un intervalo de duración de 0.01 segundos. Para estudiar la distribución de HNR se ha hecho uso del algoritmo denominado kernel density estimation (KDE) que obtiene una gaussiana por punto haciendo uso de todas las muestras. Posteriormente, se suman todas las gaussianas y se forma una única gaussiana.

Un estimador de la densidad es un algoritmo que modela la distribución de densidad de un conjunto de datos. Un estimador común es el histograma. El histograma divide los datos en bins discretos cuya altura es proporcional al número de puntos que cae dentro de cada uno. El histograma se puede normalizar, de forma que refleje la densidad de probabilidad.

Uno de los problemas que tiene el histograma es que el tamaño del bin y la posición de estos puede dar lugar a representaciones que tengan características diferentes. Un ejemplo de esto se observa en la figura 11. Para el mismo conjunto de datos, el histograma de la izquierda muestra una distribución bimodal y el histograma de la derecha muestra una distribución unimodal.

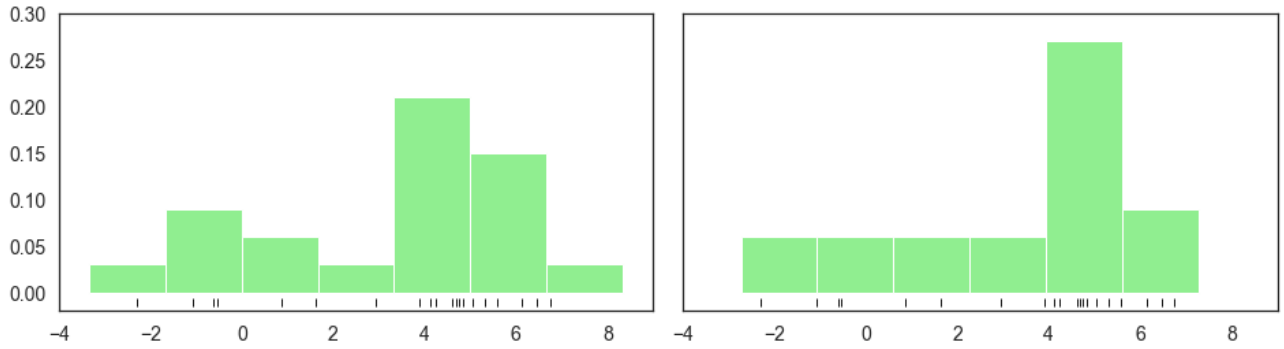


Figura 11: Dos histogramas normalizados para el mismo conjunto de datos cuya diferencia se encuentra en la posición de los bins respecto del eje x. El conjunto de datos utilizado es el siguiente: [1.62, -0.61, -0.53, -1.07, 0.87, -2.3 , 6.74, 4.24, 5.32, 4.75, 6.46, 2.94, 4.68, 4.62, 6.13, 3.9 , 4.83, 4.12, 5.04, 5.58].

El problema de los bins es que no refleja la densidad real de los puntos cercanos. En cambio, si cada punto aporta una distribución gaussiana, se obtiene una idea mucho más precisa de la distribución y el resultado varía mucho menos en función de las diferencias de muestreo [26]. Esta distribución es una KDE cuyo kernel es gaussiano. En la figura 12 se muestran unas distribuciones de este tipo.

Uno de los parámetros que hay que ajustar es el ancho de banda que controla la compensación entre el sesgo y la varianza. Un ancho de banda pequeño conduce a una distribución poco suave, es decir, con alta varianza. Un ejemplo de esta situación se muestra en la primera gráfica de la figura 12. En cambio, si se escoge un ancho de banda grande, la distribución de densidad es suave y, en consecuencia, tiene alto sesgo. Esto se puede observar en la tercera gráfica de la misma figura. La gráfica del centro es la que mejor representa la distribución de los datos.

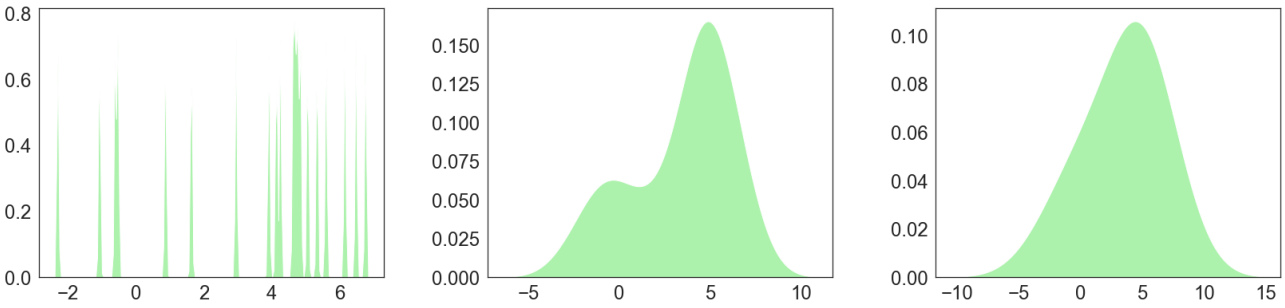


Figura 12: Gráficas de KDE correspondientes a los datos utilizados para representar los histogramas de la figura 11. Los anchos de banda de las distribuciones de izquierda a derecha son 0.01, 0.5 y 1.

Se ha obtenido la estimación de densidad de kernel de la HNR para los audios de habla continua y de vocal sostenida. En cada figura se ha representado la distribución de la HNR correspondiente a la grabación anterior al experimento y la distribución correspondiente a la grabación posterior al experimento para poder hacer una comparación de ambas.

3.4. Vibratos

Por último, se ha estudiado el vibrato como medida de la calidad de la voz cantada. Para estudiarlo, se tienen 6 audios correspondientes a 3 mujeres cantando en los cuales hacen vibrato con la voz. Para

cada mujer corresponden 2 audios, uno pre experimento y otro post experimento. Para estas muestras de voz continua se extrajeron de nuevo los segmentos sin voz. Se muestra en las figuras 13 y 14 un ejemplo de cómo queda el espectrograma de un audio antes y después de extraer los fragmentos sin sonido.

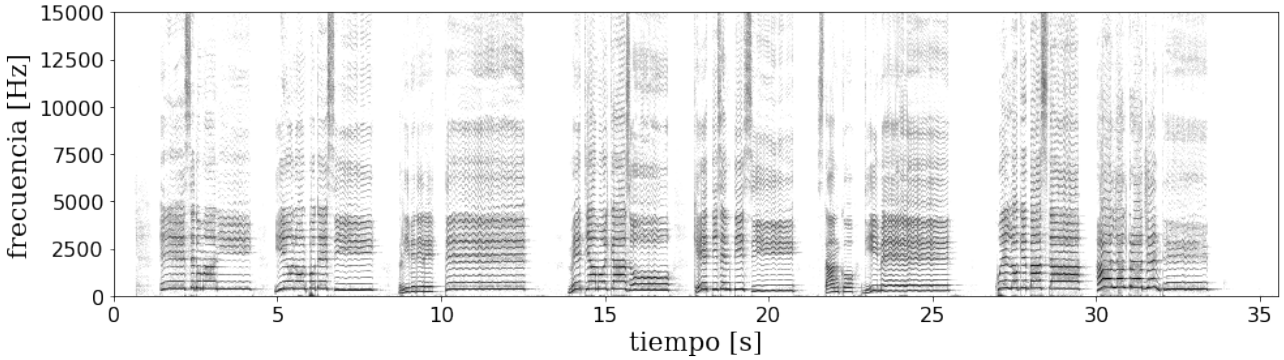


Figura 13: Espectrograma de la muestra de voz cantada de la mujer 1 anterior al experimento.

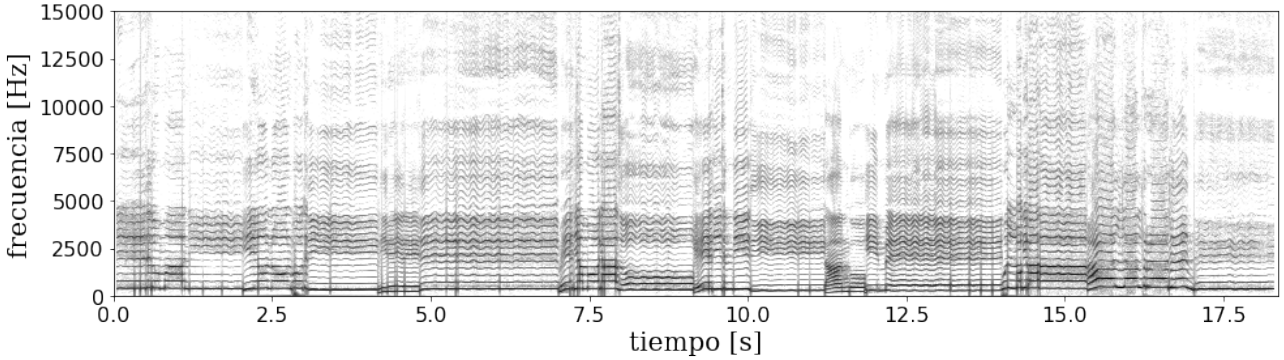


Figura 14: Mismo espectrograma que el de la figura 13 sin segmentos en los que no hay voz.

Un vibrato perfecto estaría representado en el espectrograma por una sinusoide perfecta de la frecuencia fundamental. En el espectro se vería un gran pico correspondiente a la frecuencia de modulación y magnitudes muy bajas para las demás frecuencias. Se conoce que la frecuencia de modulación suele estar entre 4 y 8 Hz.

Para obtener una métrica de calidad del vibrato se tiene en cuenta que el vibrato no es perfecto, es decir, que puede haber ligeros cambios en la frecuencia dominante de la señal del vibrato.

En el presente documento se propone una métrica para analizar la calidad del vibrato que consiste en la relación de la magnitud del pico del espectro (siempre y cuando este entre 4 y 8 Hz) y la suma de todas las demás amplitudes. Se obtendría algo similar al HNR que mide la tasa de armónicos frente a la tasa de ruido en una señal. Pero aquí no hay armónicos por lo que se compara la amplitud del pico principal frente a todas las demás componentes.

Cualquier señal puede descomponerse en un número de frecuencias discretas, o un espectro de frecuencias en un rango continuo. Esto se denomina densidad espectral (S(f)) y se representa en el espectro. De esta forma, la métrica de calidad que se propone para el vibrato es la siguiente:

$$calidad_v = \frac{S_v}{S_t - S_v} \tag{13}$$

Donde el numerador S_v corresponde a la integral de la densidad espectral para el rango de 4 Hz a 8 Hz:

$$S_t = \int_{f=0}^{f=f_{max}} S(f)\, df \tag{14}$$

y S_t corresponde a la integral sobre todas las frecuencias:

$$S_v = \int_{f=4}^{f=8} S(f)\, df \tag{15}$$

Para obtener un estudio de los vibratos de todos los audios, se cuantifica la métrica para cada segmento de vibrato en todas las grabaciones. Esto requiere etiquetar (manualmente) los tiempos de cada vibrato. Como alternativa, se parte de las grabaciones enteras en ventanas cortas y se cuantifica el posible vibrato en cada una de esas ventanas. El tono tiene valores en 0 que corresponden a ausencia de voz. Esos fragmentos no deberían afectar al cálculo, por lo que se sustituyen por sus valores interpolados. Es una medida sencilla pero evita tener que hacer un procesado más detallado.

Para obtener el espectro de una señal aplicando la FFT esta debe tener media 0. Por lo que se ha restado la media a cada fragmento y se ha obtenido la FFT como se explica en [19]. Finalmente se ha obtenido la métrica de evaluación del vibrato para cada fragmento y se ha representado en un histograma el conjunto de estas métricas para cada audio.

4. Resultados y discusión

4.1. Análisis exploratorio de los datos

En primer lugar, se ha llevado a cabo un análisis exploratorio de los datos con el fin de obtener una idea general de los datos con los que se cuenta.

Para empezar, se ha llevado a cabo la representación de oscilogramas y espectrogramas con la intención de ver si visualmente se encuentra alguna diferencia entre los audios pre y post-experimento para cada paciente. Un ejemplo se muestra en las siguientes figuras que corresponden los audios de habla continua de uno de los sujetos. En las figuras 15 y 16 se muestran los dos oscilogramas junto con los valores de HNR y en las figuras 17 y 18 se han representado los dos espectrogramas junto con la intensidad para los audios de habla continua. Estas mismas gráficas se han representado para los audios de vocal sostenida y se muestran en las figuras 19, 20, 21 y 22.

En el habla continua se hacen silencios al hablar y se emiten distintos sonidos. En los momentos de silencio la amplitud del oscilograma y la HNR decaen a 0. En estos momentos el espectrograma muestra tonos más claros de color y la intensidad decae. En cambio, para la vocal sostenida, la persona no realiza silencios y pronuncia la misma vocal en todo el audio. Al no haber momentos de silencio, la amplitud del oscilograma y la HNR no decaen a 0 durante momentos del orden de las décimas de segundo, como ocurre en el habla continua. En el espectrograma tampoco se producen tonos tan claros correspondientes a lo silencios y la intensidad se mantiene prácticamente constante.

Por otra parte, se han comparado las gráficas correspondientes a los audios anteriores al experimento con las gráficas correspondientes a los audios posteriores y aparentemente no se observan diferencias, tanto para los audios de vocal sostenida como para los de habla continua.

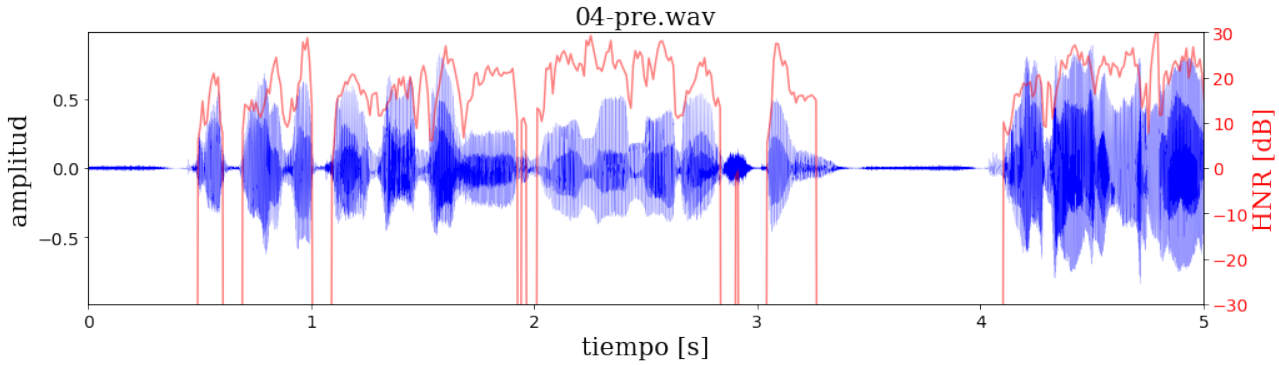


Figura 15: Oscilograma y HNR para cada instante de tiempo del audio pre-experimento de habla continua del sujeto 4. Se representan los 5 primeros segundos de audio.

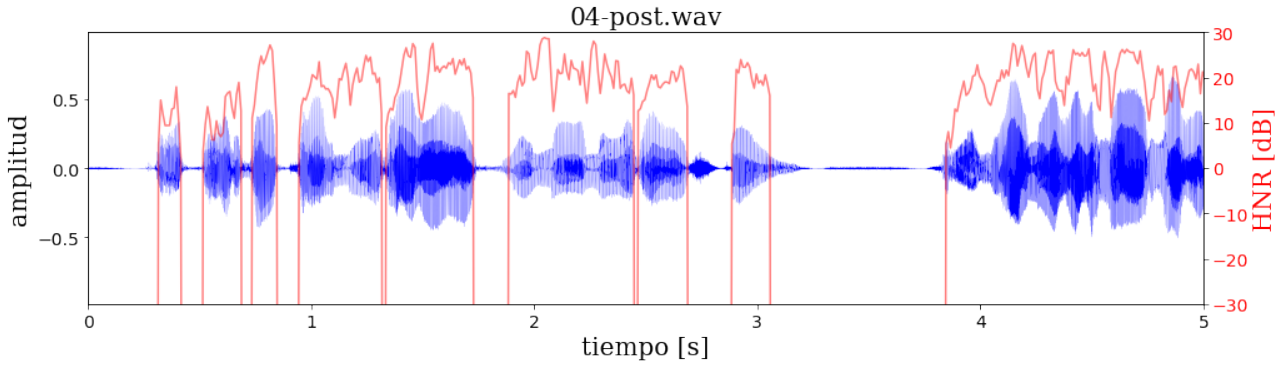


Figura 16: Oscilograma y HNR para cada instante de tiempo del audio post-experimento de habla continua del sujeto 4. Se representan los 5 primeros segundos de audio.

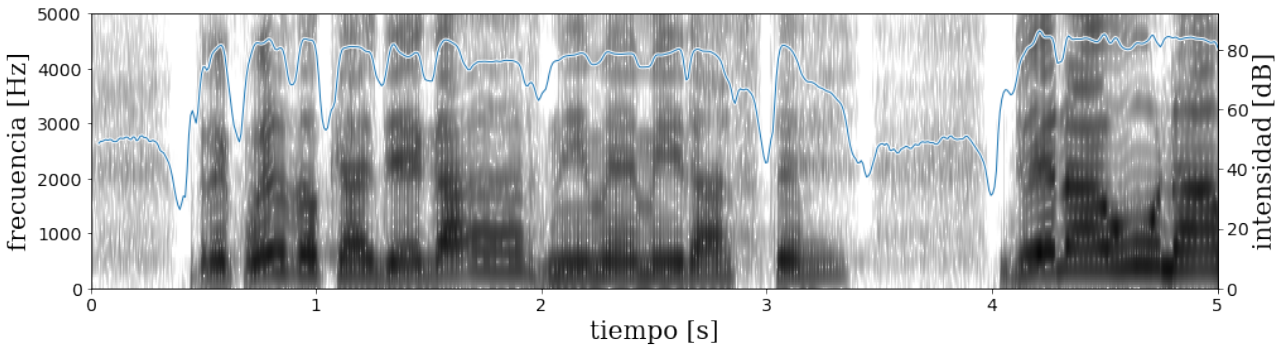


Figura 17: Espectrograma e intensidad del audio pre-experimento de habla continua del sujeto 4. Se representan los 5 primeros segundos de audio.

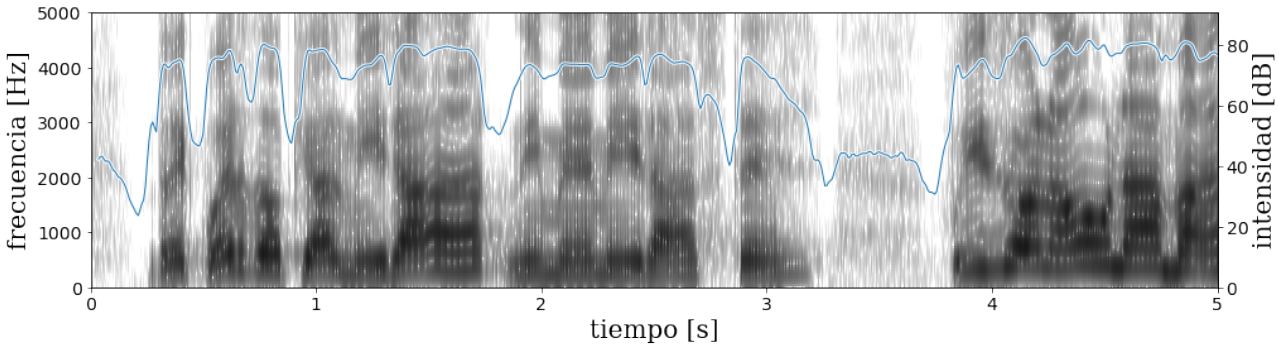


Figura 18: Espectrograma y tono del audio post-experimento de habla continua del sujeto 4. Se representan los 5 primeros segundos de audio.

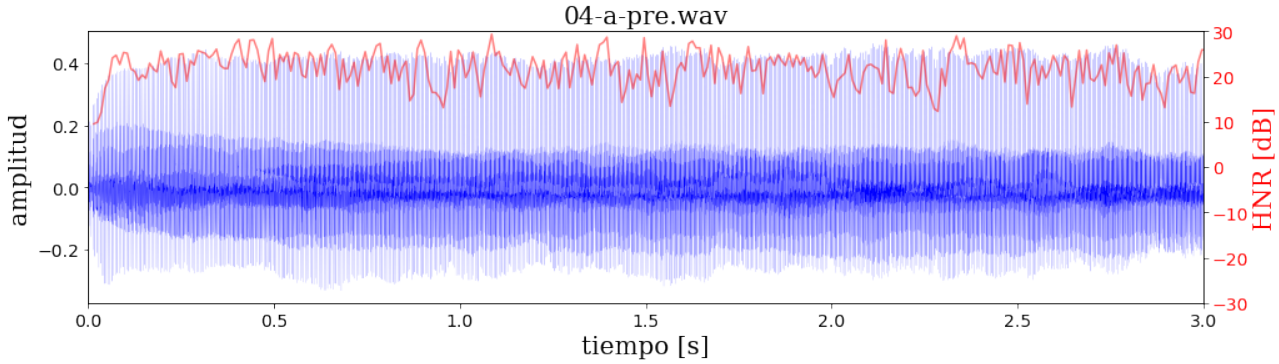


Figura 19: Oscilograma y HNR para cada instante de tiempo del audio pre-experimento de vocal sostenida del sujeto 4. Se representan los 3 primeros segundos de audio.

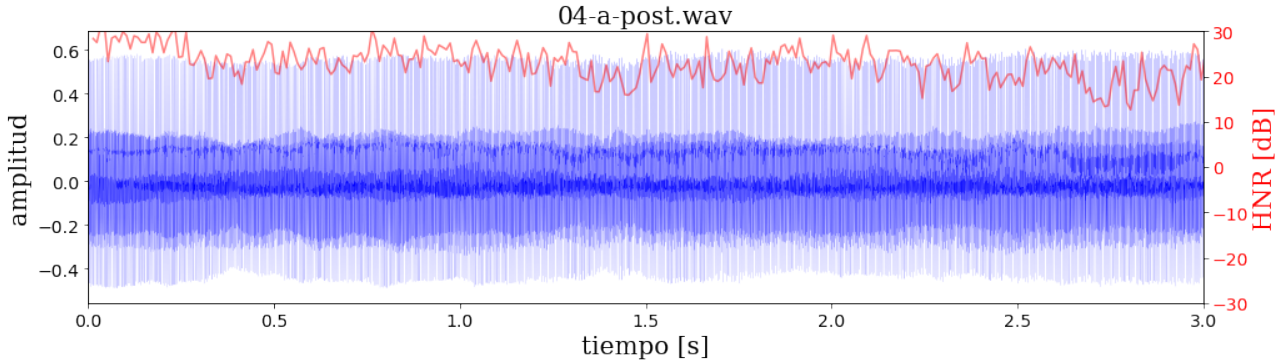


Figura 20: Oscilograma y HNR para cada instante de tiempo del audio post-experimento de vocal sostenida del sujeto 4. Se representan los 3 primeros segundos de audio.

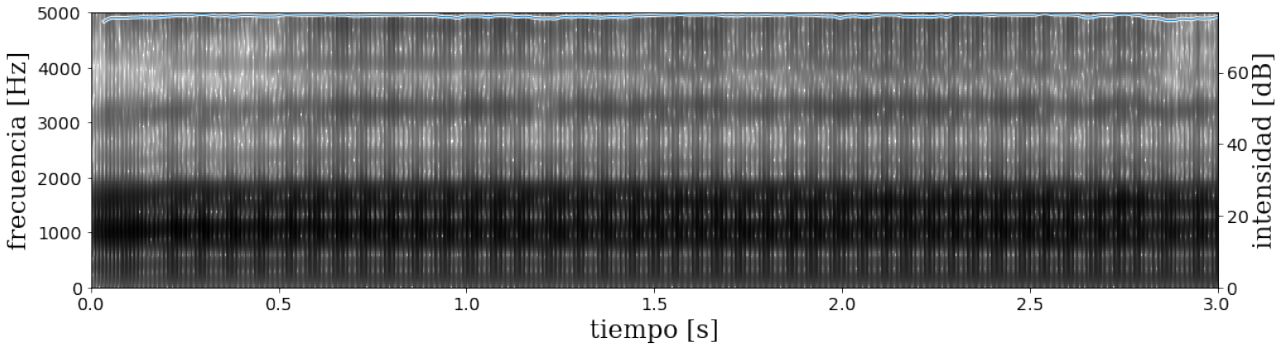


Figura 21: Espectrograma e intensidad del audio pre-experimento de vocal sostenida del sujeto 4. Se representan los 3 primeros segundos de audio.

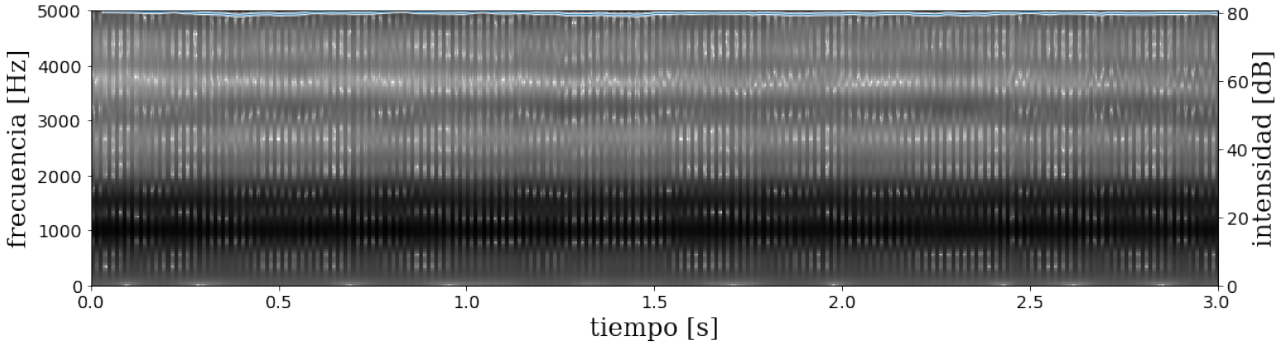


Figura 22: Espectrograma y tono del audio post-experimento de vocal sostenida del sujeto 4. Se representan los 3 primeros segundos de audio.

Para continuar con la exploración inicial de los datos, se han calculado los parámetros clásicos de análisis acústico de la voz descritos en la sección 3. Materiales y Metodología, algunos de ellos utilizados para el cálculo del AVQI, con el objetivo de encontrar algún parámetro que permita diferenciar las voces pre y post-experimento. En las figuras 23 y 46 del apéndice se ha representado en cada gráfica un parámetro con el valor post-experimento frente al valor pre-experimento, correspondiendo cada punto a un sujeto.

Si los puntos quedan sobre la línea quiere decir que el parámetro tiene el mismo valor pre y post experimento. Si el punto queda por debajo de la recta significa que tras el experimento el parámetro disminuye, en cambio, si queda por encima mejora. Estas gráficas se han representado para ver si se observa una disminución o aumento general de ciertos parámetros porque esto significaría que habría algunos que permiten diferenciar entre una mejor o peor calidad de la voz, teniendo una mejor calidad asociada a los valores post-experimento y una peor calidad a los valores pre-experimento. Como se puede observar no hay ningún caso en el que todos o la mayoría de puntos se encuentren por encima o por debajo de la recta.

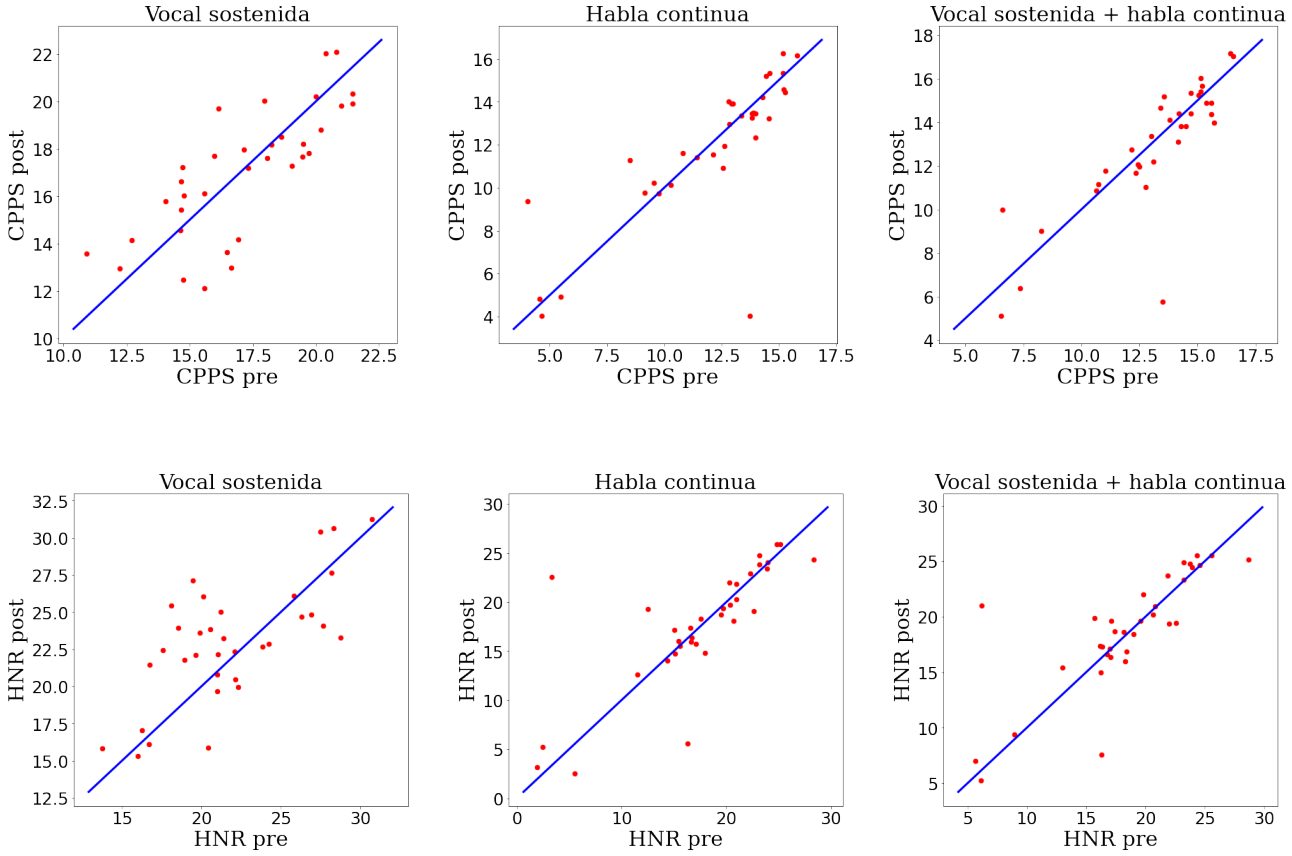
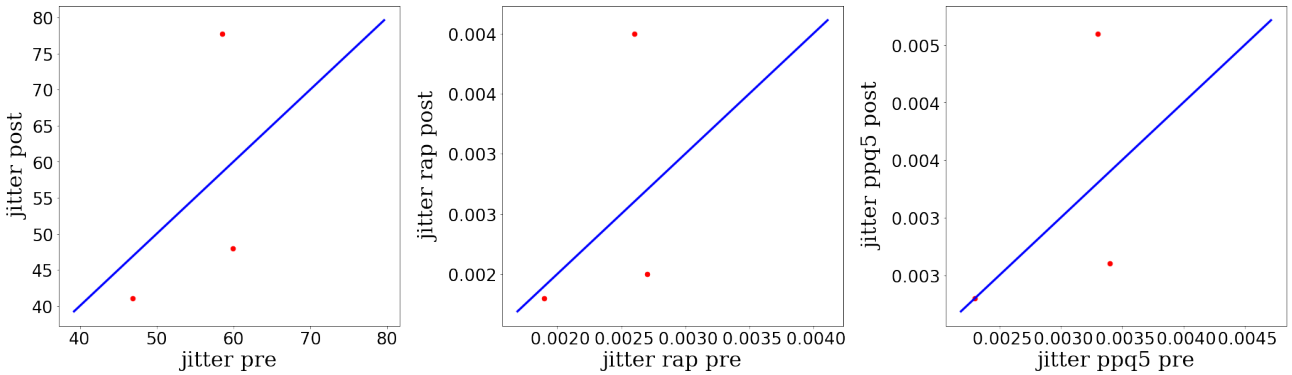


Figura 23: Gráficas en las que se representa cada parámetro post-experimento frente al parámetro pre-experimento para las grabaciones de vocal sostenida, habla continua y la combinación de ambas. Cada punto rojo corresponde a un paciente. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.

Estas gráficas también se han creado para las grabaciones de las mujeres que cantan con vibrato. Se muestran en las figuras 24 y 47 del apéndice. La única gráfica que muestra unos parámetros cuyos valores aumentan o disminuyen para todos los puntos sería la correspondiente al HNR. Al tener únicamente 3 puntos, este resultado tiene poco poder estadístico y por tanto podría ser fruto del azar.



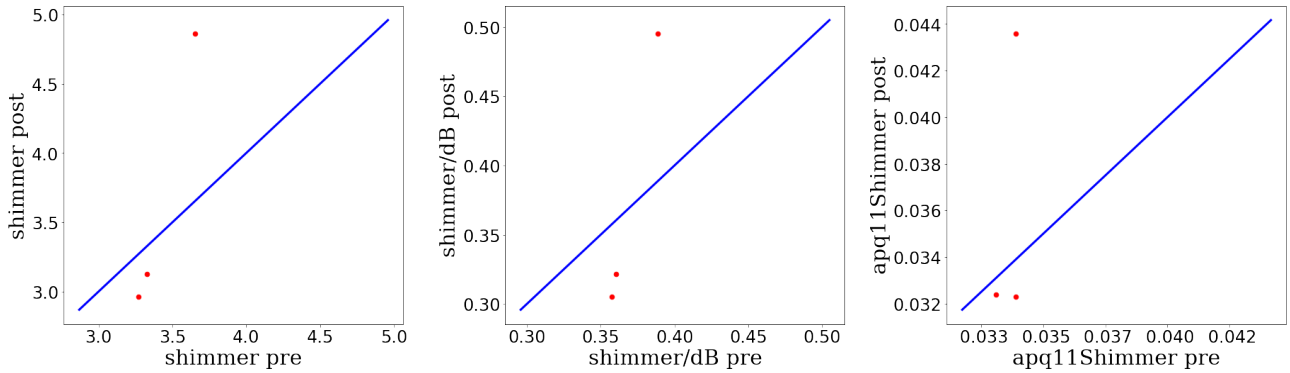
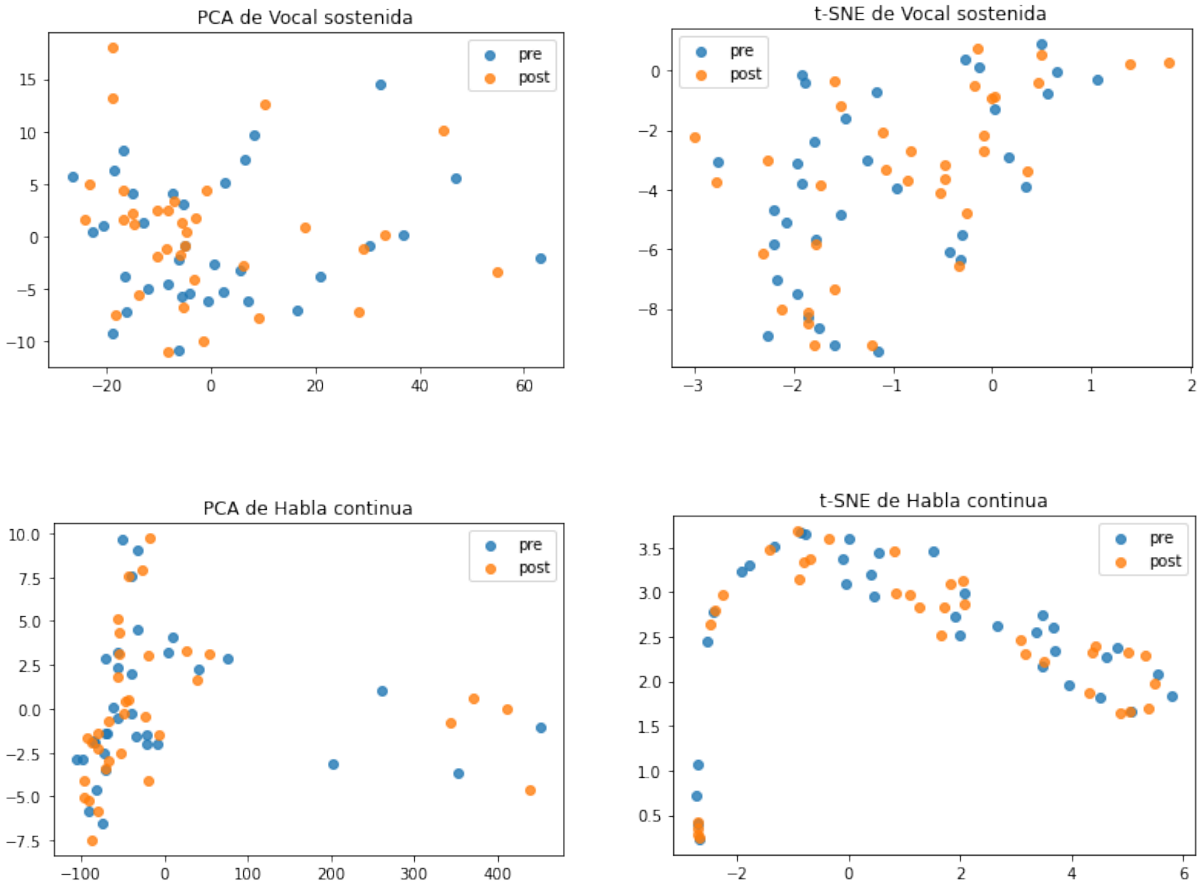


Figura 24: Gráficas en las que se representa cada parámetro post-experimento frente al parámetro pre-experimento para las grabaciones de voz de las tres mujeres que cantan con vibrato. Cada punto rojo corresponde a un paciente. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.

Por otro lado, para la primera muestra de grabaciones, se ha llevado a cabo el análisis de componentes principales (PCA) y el t-distributed Stochastic Neighbor Embedding (t-SNE) que son algoritmos para la reducción de la dimensionalidad. Para ello se han utilizado todos los parámetros (slope, HNR, CPPS, tilt, jitt, jitter (rap), jitter (ppq5), shim, shdB y shimmer (apq11)). En la figura 25 se observa la representación sobre las 2 primeras componentes principales y en la figura 26 se representa t-SNE para dos componentes. No se observa ningún tipo agrupación de puntos en función de pre y post experimento.



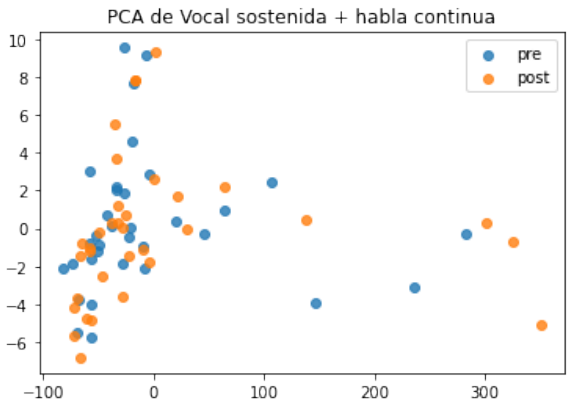


Figura 25: Representación de las 2 primeras componentes principales obtenidas a partir de los parámetros de las grabaciones de vocal sostenida, habla continua y la combinación de ambas. Cada punto corresponde a un paciente: en azul pre-experimento y en naranja post-experimento.

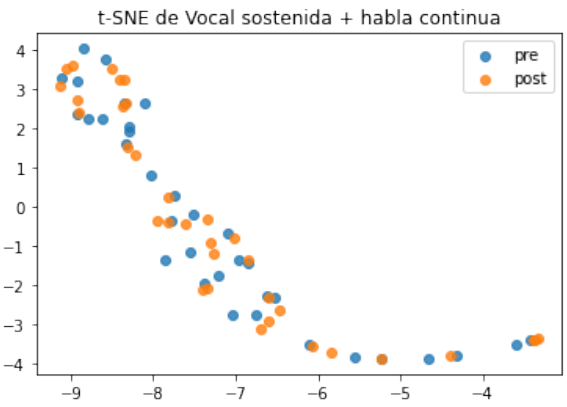


Figura 26: t-SNE a partir de los parámetros llevada a cabo para las grabaciones de vocal sostenida, habla continua y la combinación de ambas. Cada punto corresponde a un paciente: en azul pre-experimento y en naranja post-experimento.

La reducción de la dimensionalidad también se llevado a cabo para las grabaciones de las mujeres que cantan con vibrato. Las gráficas correspondientes se muestran en las figuras 27 y 28. Al igual que en las gráficas correspondientes a la primera muestra, no se observa ningún tipo agrupación de puntos en función de pre y post experimento.

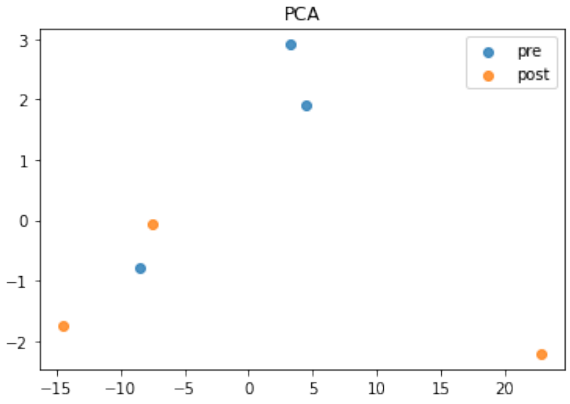


Figura 27: Representación de las 2 primeras componentes principales obtenidas a partir de los parámetros correspondientes a las grabaciones de las tres mujeres que cantan con vibrato. En azul los puntos pre-experimento y en naranja los puntos post-experimento.

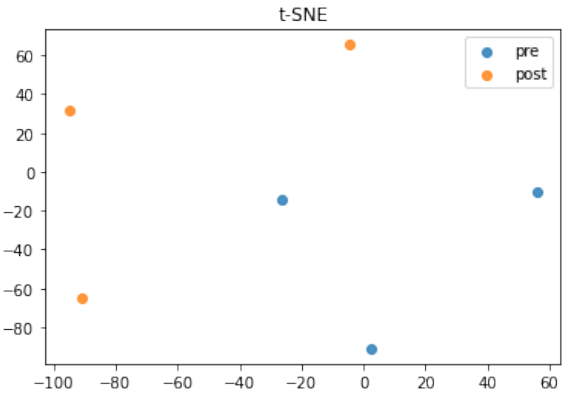


Figura 28: t-SNE a partir de los parámetros llevada a cabo a partir de los parámetros correspondientes a las grabaciones de las tres mujeres que cantan con vibrato. En azul los puntos pre-experimento y en naranja los puntos post-experimento.

Finalmente, se ha utilizado como clasificador un árbol de decisión para clasificar cada audio en uno de los dos grupos: antes del experimento o después de este. La profundidad profundidad máxima de los árboles es de 2 para que sea un modelo sencillo. Se ha obtenido un accuracy para el test de 0.39, resultado que indica la mala capacidad de predecir. También se ha empleado un Random Forest cuyos

árboles tienen una profundidad máxima de 2. Con este clasificador se ha obtenido una accuracy de 0.48 para el test, por lo que se podría decir que la clasificación que realiza el modelo es aleatoria.

4.2. Índice de Calidad Acústica de la Voz

Para analizar la utilidad del AVQI como métrica que sirva para clasificar la voz, este se ha calculado para los tres casos que se han mencionado en el apartado de metodología: vocal sostenida, habla continua y combinación de ambas. Posteriormente, se ha representado para cada paciente los valores correspondientes al audio posterior al experimento frente a los valores correspondientes al audio anterior al experimento. En la figura 29 se muestran las gráficas correspondientes a la primera muestra de datos.

Al igual que en la exploración inicial de los datos en la cual se buscaba algún parámetro que permitiera clasificar la voz, tampoco se observa ninguna gráfica en la que todos o la mayoría de puntos se encuentren por encima o por debajo de la recta. Esto ocurre tanto como para la primera muestra de voces, como para la muestra de las mujeres que cantan con vibrato, como se observa en la figura 30.

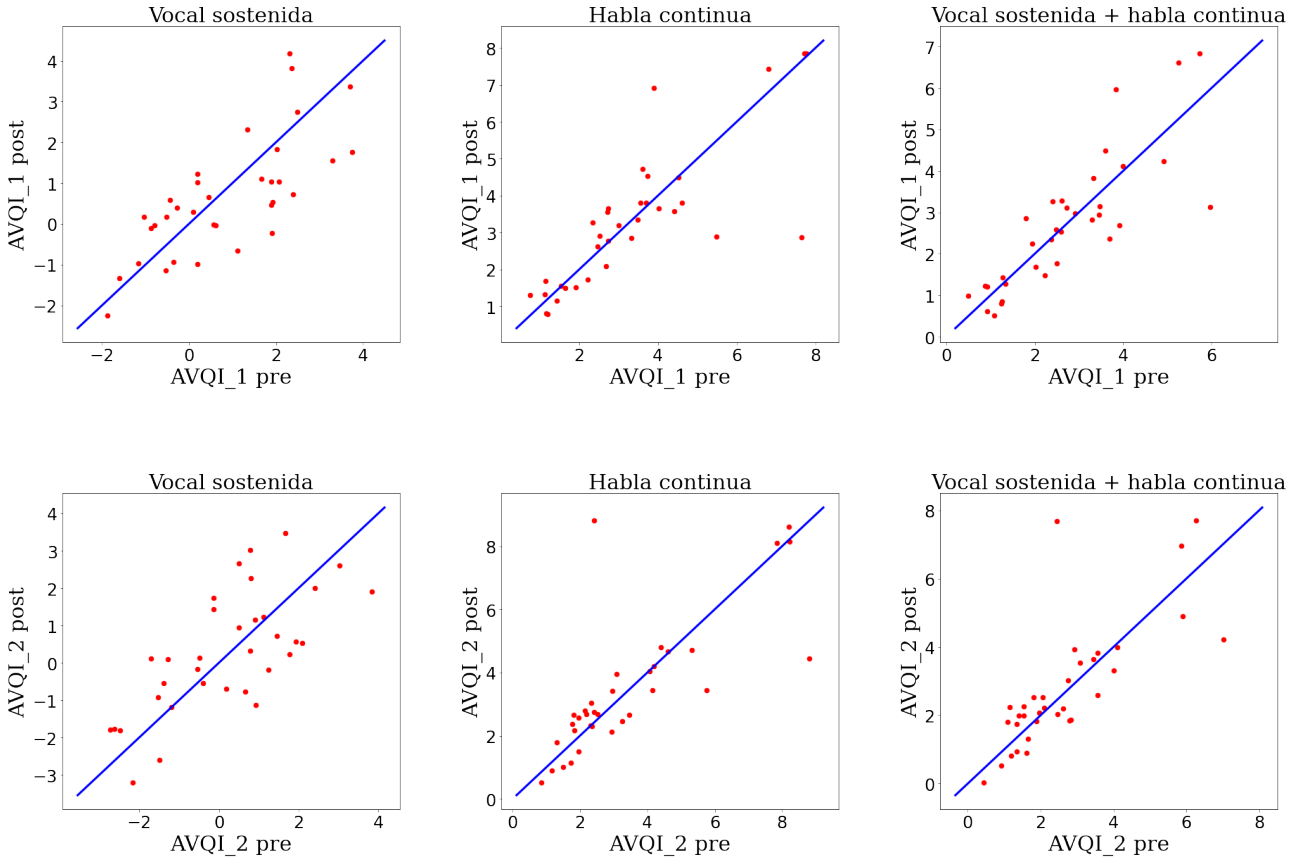


Figura 29: Gráficas en las que se representa el AVQI post-experimento frente al AVQI pre-experimento para las grabaciones de vocal sostenida, habla continua y la combinación de ambas. AVQI_1 corresponde a la ecuación 1 y AVQI_2 corresponde a la ecuación 2. Cada punto rojo corresponde a un paciente. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.

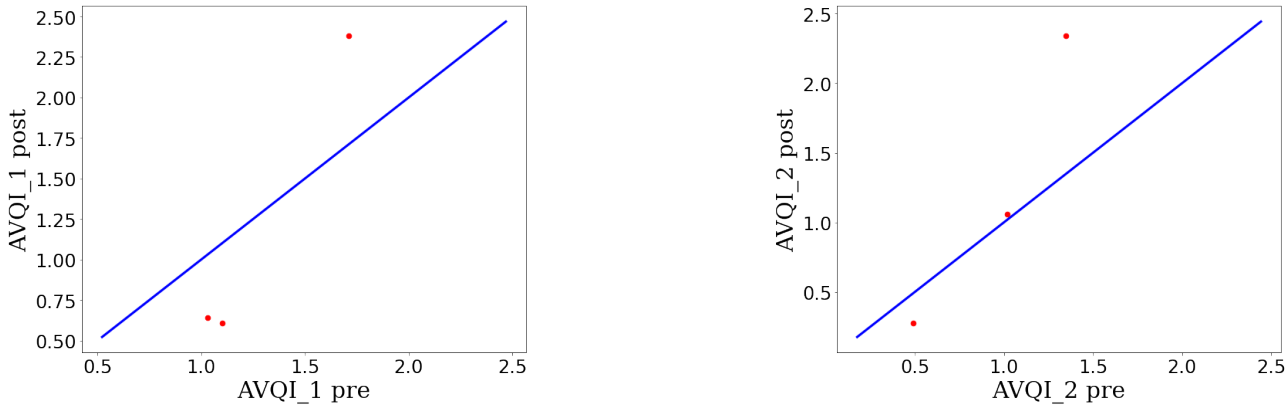


Figura 30: Gráficas en las que se representa el AVQI post-experimento frente AVQI pre-experimento para grabaciones de voz de las tres mujeres que cantan con vibrato. AVQI_1 corresponde a la ecuación 1 y AVQI_2 corresponde a la ecuación 2. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.

4.3. Distribución de HNR

A continuación se muestran los resultados correspondientes a la representación de la distribución de la HNR.

Cuanto más a la derecha se encuentra la distribución, la proporción de armónicos frente a ruido es mayor que si la distribución se encuentra más hacia la izquierda. Se espera que las voces grabadas posteriormente al experimento para mejorar la calidad de la voz, tengan una distribución hacia la derecha o que la cola derecha de la distribución se encuentre por encima de la cola derecha de la distribución pre experimento y que la cola izquierda de la distribución post experimento se encuentre por debajo de la cola izquierda de la distribución pre experimento. Esto se puede observar en la figura 31 donde la cola derecha de la distribución post experimento (de color rojo) se encuentra por encima de la cola derecha de la distribución pre experimento (de color azul) y la cola izquierda post experimento se encuentra por debajo de la cola izquierda pre experimento. Además la distribución post-experimento se encuentra desplazada hacia la derecha.

En la mayoría de los casos de habla continua, se obtiene una distribución similar para las grabaciones pre y post experimento, como se puede observar en la figura 32. En consecuencia, esta distribución no permite discernir entre una voz que ha sido sometida a experimento y otra que no.

En cambio para los audios de vocal sostenida se obtienen resultados muy diversos. Por un lado, se encuentran distribuciones similares como se muestra en la figura 33. Y por otro lado se observan distribuciones desplazadas una respecto a la otra como es el caso de las figuras 34 y 35. En estas figuras se obtienen resultados opuestos, ya que la primera de ellas indica que el audio post-experimento tiene valores mayores de HNR y la segunda indica lo contrario. Analizando estos resultados, estas distribuciones tampoco sirven para discernir entre pre y post-experimento.

En la figura 36 quedan reflejados los resultados que se han explicado. Para los audios de habla continua, para los cuales no se observa demasiada variación de la distribución de HNR entre pre y post-experimento, se tiene una diferencia cercana a 0 en la mayoría de audios. En cambio, para los audios de vocal sostenida, se han obtenido tanto gráficas similares, como gráficas una desplazada respecto a la otra. Por esta razón se observan tanto puntos cercanos a 0, como puntos con valores mayores por encima y por debajo del 0.

Figura 31: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de vocal sostenida correspondiente al sujeto 32. El ancho de banda de las distribuciones es 0.2.

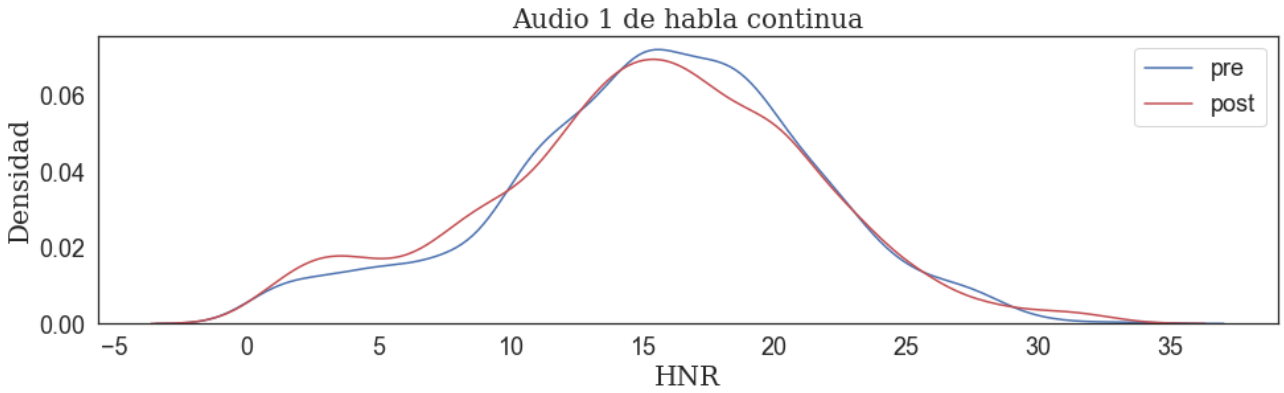


Figura 32: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de habla continua correspondiente al sujeto 1. El ancho de banda de las distribuciones es 0.2.

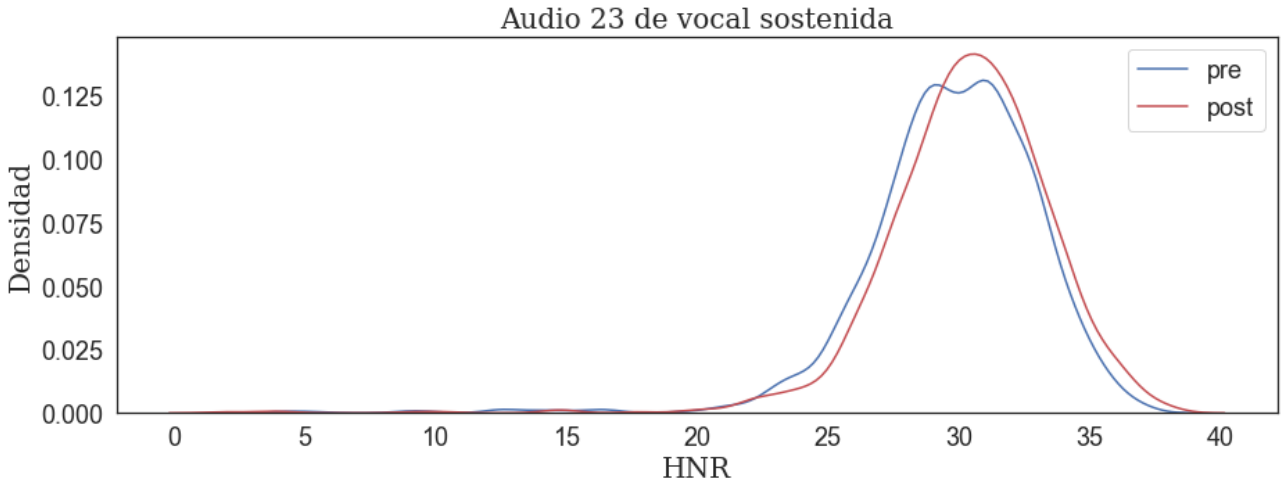


Figura 33: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de vocal sostenida correspondiente al sujeto 23. El ancho de banda de las distribuciones es 0.2.

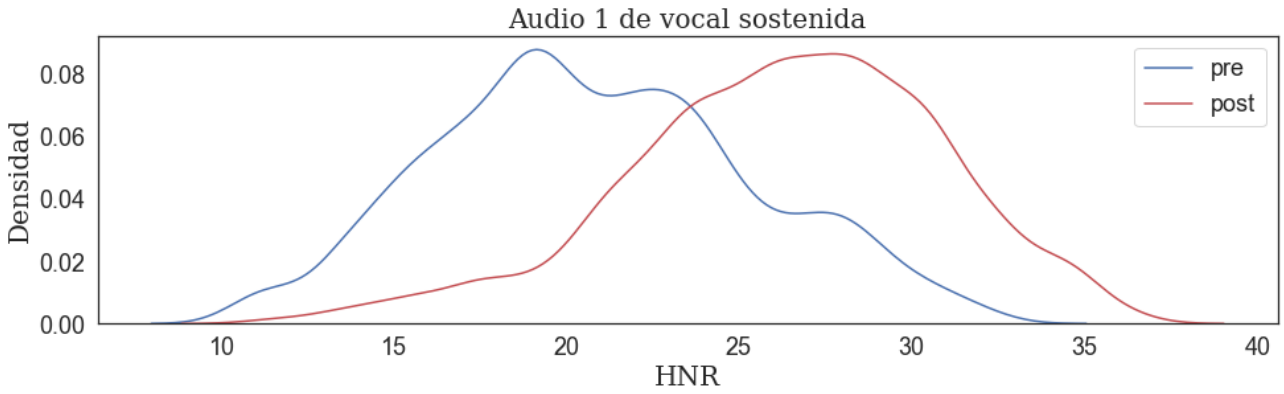


Figura 34: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de vocal sostenida correspondiente al sujeto 1. El ancho de banda de las distribuciones es 0.2.

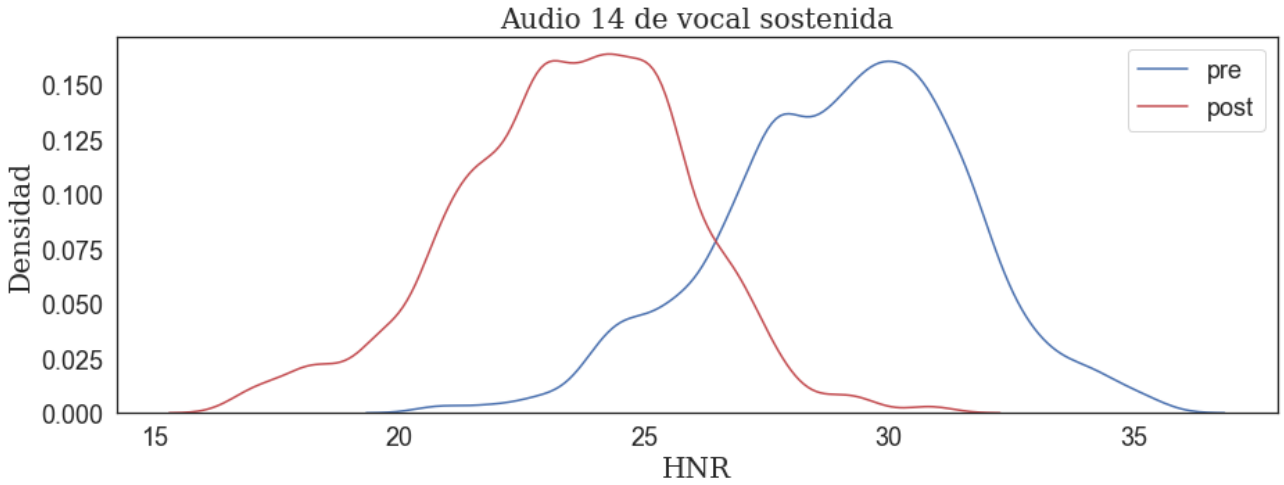


Figura 35: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de vocal sostenida correspondiente al sujeto 14. El ancho de banda de las distribuciones es 0.2.

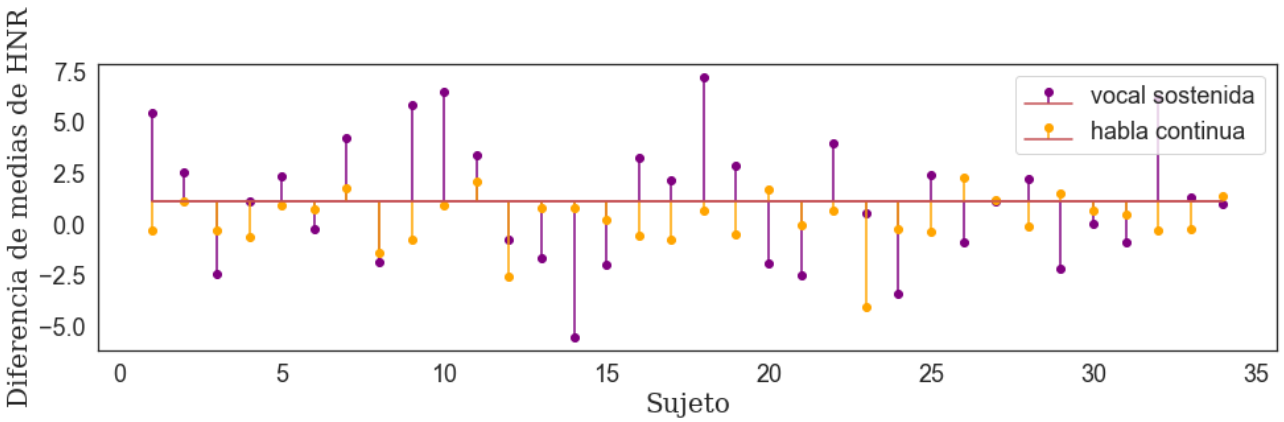


Figura 36: Diferencia entre la media de HNR del audio post experimento y la media de HNR del audio pre experimento para cada sujeto y diferenciado en vocal sostenida (morado) y habla continua (naranja).

Como curiosidad cabe señalar que para las voces guturales, se obtiene una distribución asimétrica a la izquierda como se muestra en las figuras 37 y 38. Esto quiere decir que estas grabaciones presentan más momentos con una proporción de ruido mayor que el resto de grabaciones. Este resultado podría servir para un futuro estudio de clasificación de tipo de canto en función de la distribución de la distribución de laHNR.

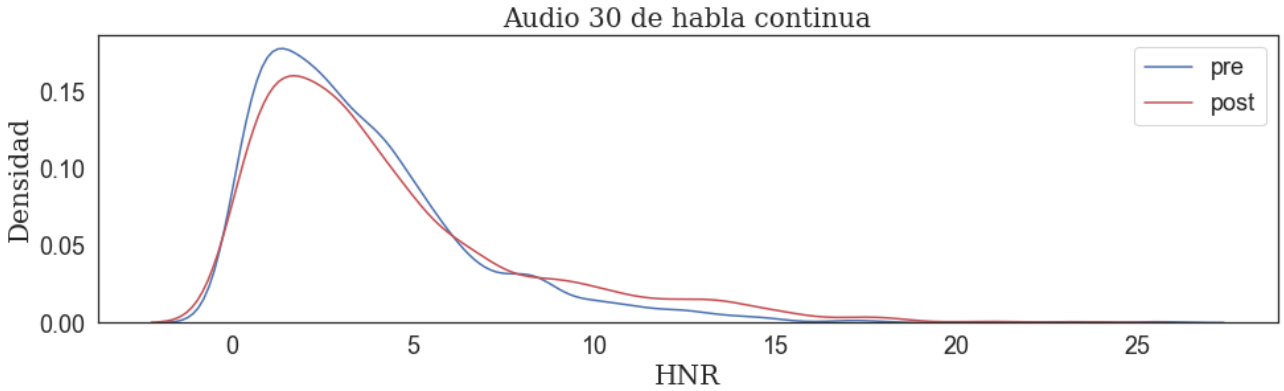


Figura 37: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de habla continua correspondiente al sujeto 30. El ancho de banda de las distribuciones es 0.2.

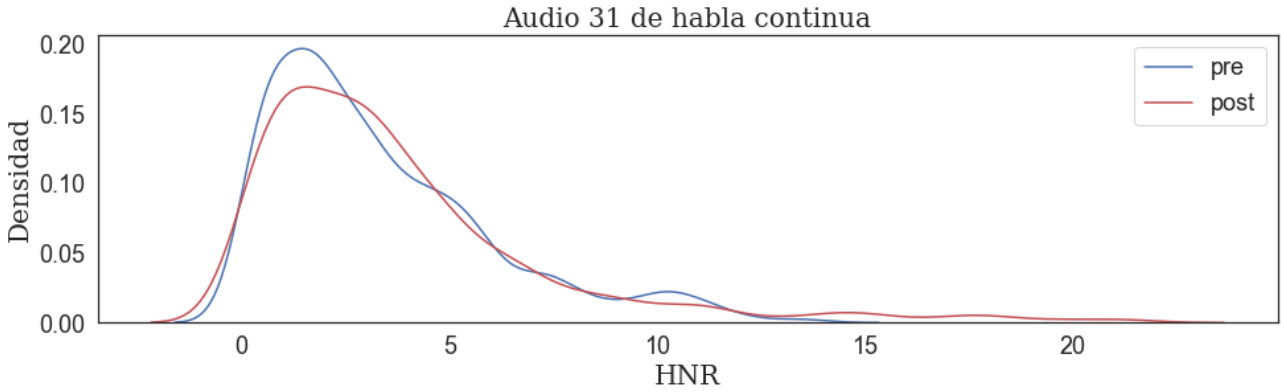


Figura 38: KDE de la HNR para el audio post-experimento (rojo) y pre-experimento (azul) de habla continua correspondiente al sujeto 31. El ancho de banda de las distribuciones es 0.2.

4.4. Vibratos

La última métrica de la calidad de la voz son los vibratos. En primer lugar, se han representado una serie de gráficas para poder identificar los vibratos. En la figura 39 se ha representado la frecuencia fundamental frente al tiempo de uno de los audios post experimento. Se pueden visualizar varias zonas con vibratos bien visibles, que serían aquellas zonas en las que la frecuencia fundamental varía en poco tiempo. Esto se observa en detalle en la figura 40.

La frecuencia de modulación suele estar entre 4 y 8 Hz y en la figura se observan unos 5 picos por segundo por lo que la frecuencia de modulación debe estar en torno a 5 Hz. En la figura 41 se ha obtenido el espectro aplicando la transformada de Fourier de ese intervalo de tiempo y se puede observar un pico en torno a 5 Hz. Las otras frecuencias indican perturbaciones de la señal sinusoidal perfecta.

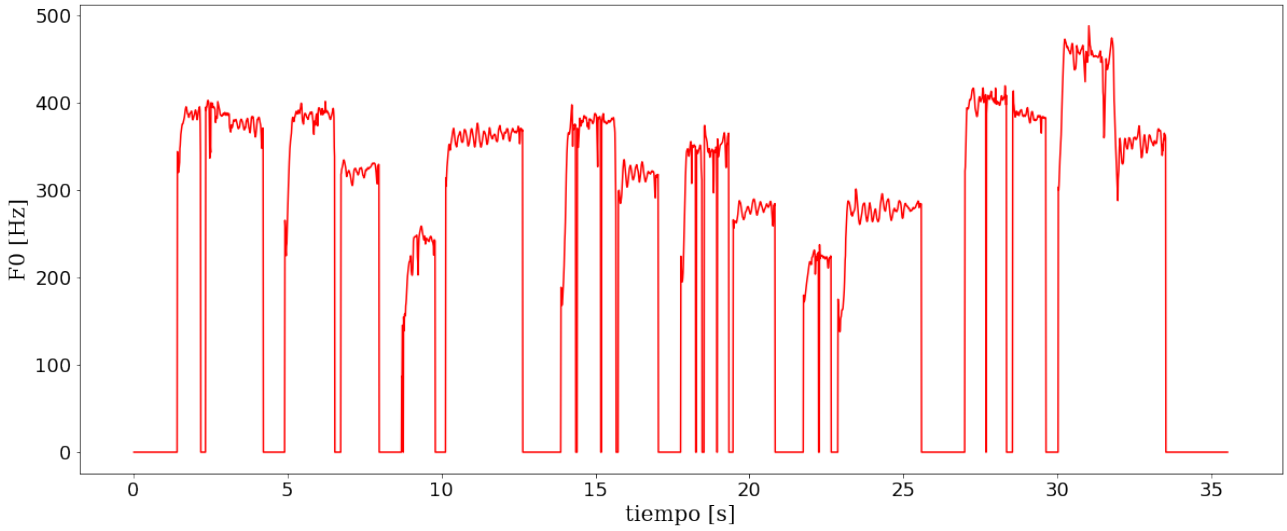


Figura 39: Frecuencia fundamental frente al tiempo del audio grabado tras experimento de la primera mujer que canta con vibrato.

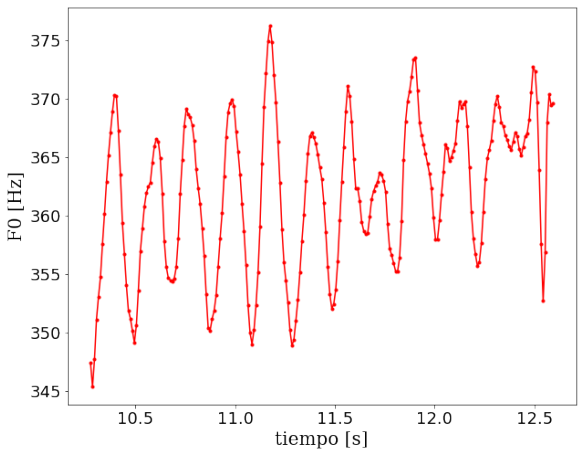


Figura 40: Ampliación de la gráfica de la figura 39.

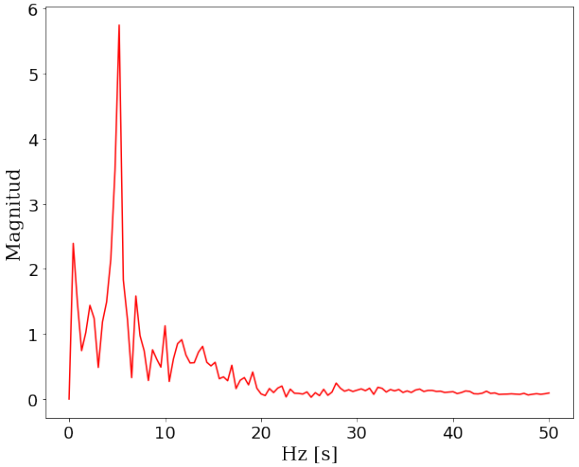


Figura 41: Espectro obtenido a partir del fragmento representado en la figura 40.

Repitiendo el proceso para la grabación anterior al experimento, se obtiene la gráfica 42 y gráfica la ampliada para el mismo tramo de canción que en el representado en la figura 40. Se puede apreciar en la figura 43 cómo la variación de la frecuencia fundamental es menos limpia en ese tramo que la obtenida para la grabación anterior al experimento. En la figura 44 se observa un pico sobre 4 Hz, que es una frecuencia que cae dentro del rango habitual del vibrato. Pero fuera de ese rango hay muchas componentes en el espectro, lo cual indica que está lejos de ser un vibrato perfecto (que se vería como un único pico en el espectro y magnitudes muy bajas para las demás frecuencias). Las frecuencias cercanas al espectro tienen magnitud alta y corresponden a ligeros cambios en la frecuencia dominante de la señal del vibrato. Estas se toman en cuenta con la métrica de calidad expuesta en la en la ecuación 13.

Para el fragmento de audio post experimento se obtiene una métrica de 1.33 y para el otro fragmento pre experimento se obtiene 0.50, es decir, la métrica indica un vibrato más perfecto para el fragmento de la señal post experimento.

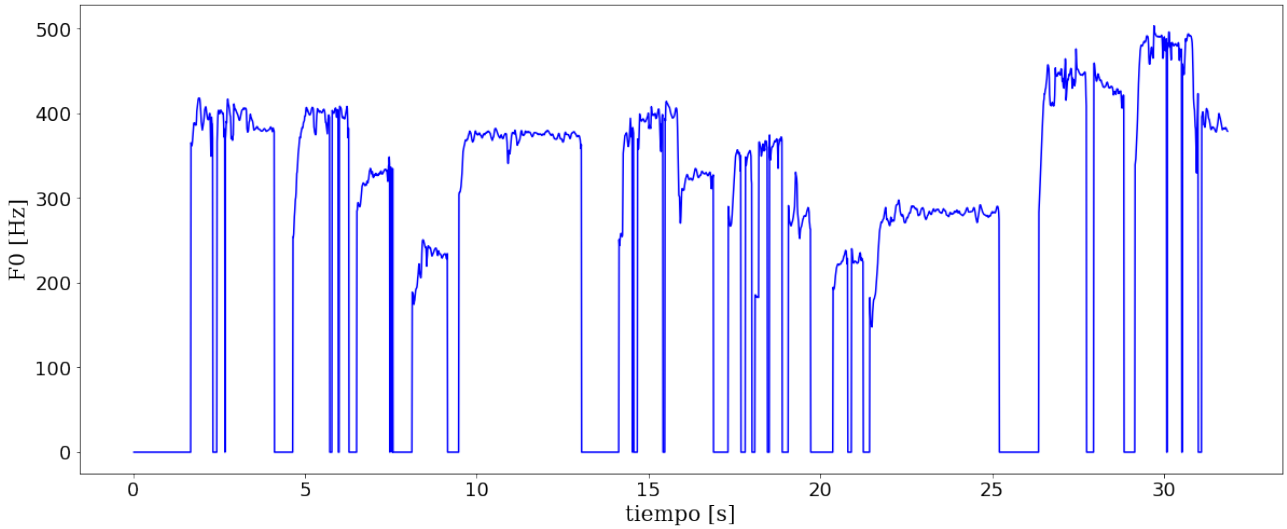


Figura 42: Frecuencia fundamental frente al tiempo del audio grabado antes experimento de la primera mujer que canta con vibrato.

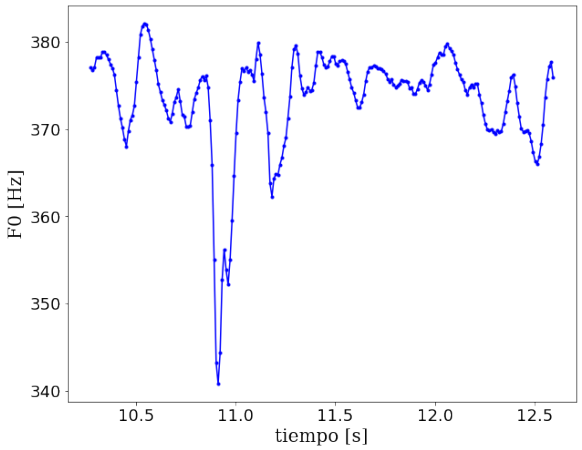


Figura 43: Ampliación de la gráfica de la figura 39.

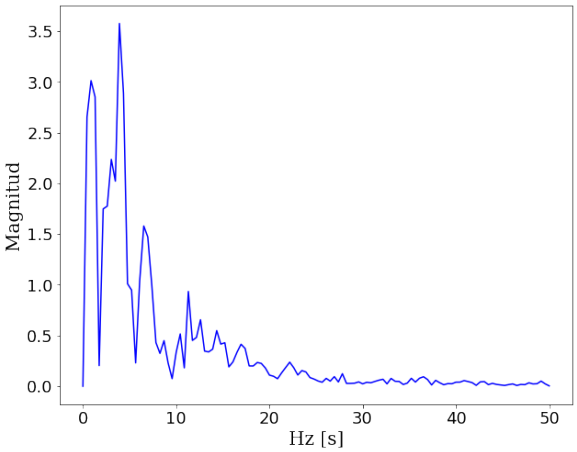


Figura 44: Espectro obtenido a partir del fragmento representado en la figura 40.

Finalmente, se ha calculado la métrica de calidad para todos los fragmentos de los audios y se han representado en histogramas. Estos histogramas se muestran en la figura 45. Se ha limitado el eje x por la izquierda a 1 porque se desea comparar el número de métricas cuya magnitud es igual o mayor a 1. La métrica de calidad mayor a 1 destaca la magnitud del rango de 4 a 8 Hz frente al resto de frecuencias y por lo tanto, destaca la aportación del vibrato.

Se observa que la primera cantante pasa de tener pocos momentos con calidad igual o mayor a 1 para la grabación pre-experimento a gran cantidad posterior al experimento, llegando incluso a una ventana con un vibrato para el cual la métrica da cerca de 5. Para las grabaciones de las otras cantantes no se nota ninguna mejora similar.

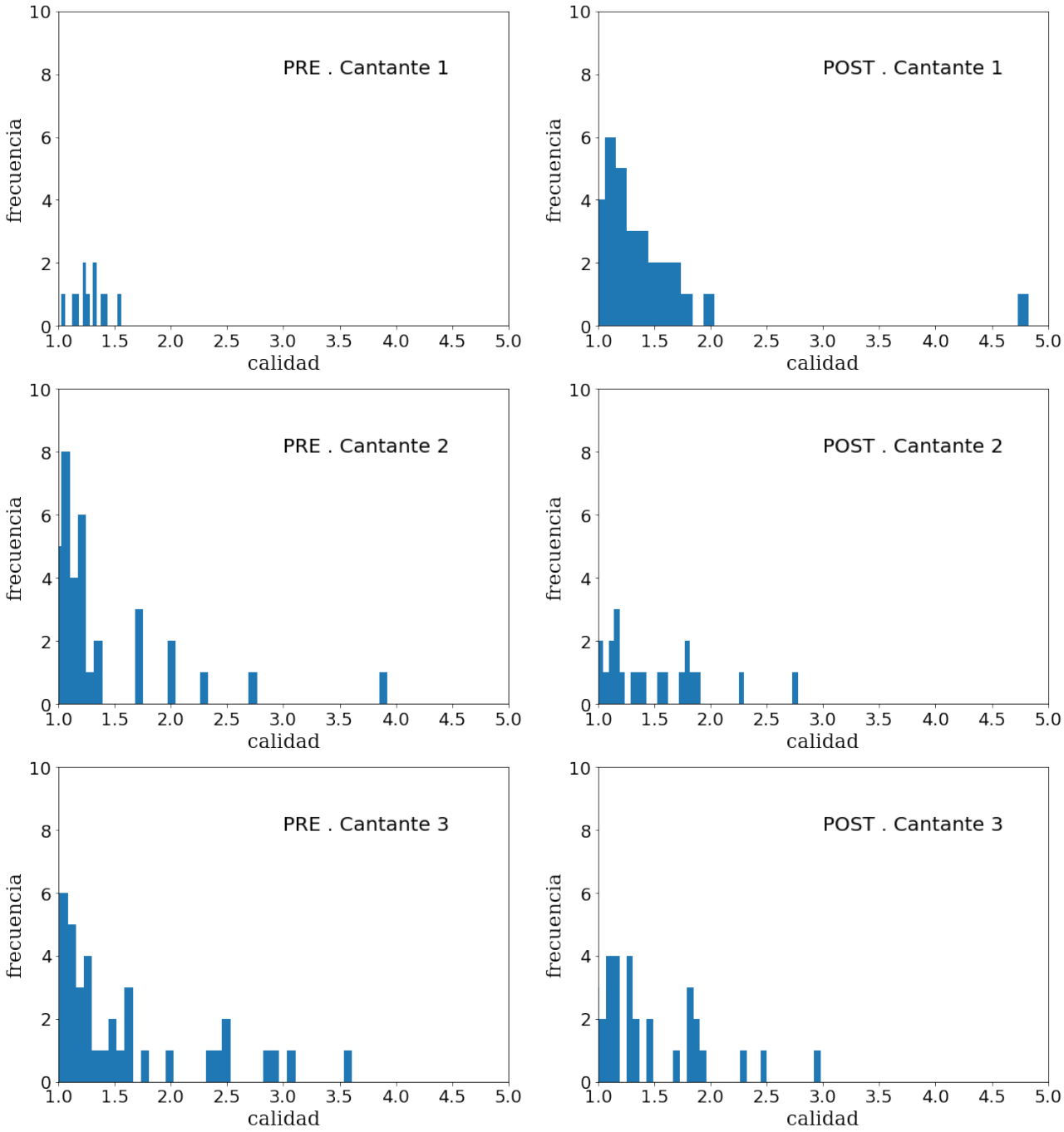


Figura 45: Histogramas de la métrica de calidad para los fragmentos de audio. Se indica a qué cantante pertenece y si es pre o post-experimento.

5. Conclusiones

Se han estudiado métodos para cuantificar la calidad de la voz. Para ello se han investigado varias métricas, algunas ya existentes en la literatura y otras se han propuesto en el marco de este trabajo. En primer, lugar se ha estudiado el AVQI que es una métrica basada en una regresión lineal múltiple de seis parámetros que se utiliza para evaluar el grado de calidad de voces habladas y de vocal sostenida. En segundo lugar, se ha estudiado la distribución de la HNR que indica la cantidad de ruido en la voz y por lo tanto, está relacionada con la calidad de esta. Finalmente, se dispone de unas grabaciones de canto con vibrato y se propone una métrica para evaluar la calidad de los vibratos. Analizando los resultados, no se observa una clara mejoría de la calidad de la voz en las señales grabadas tras el experimento.

Por otro lado, cabe señalar que el AVQI, cuando se aplica en la literatura para voces de habla continua, se hace para voces habladas y no cantadas. Algunos parámetros como el jitter y shimmer, evalúan los cambios rápidos de frecuencia y amplitud. Podría ocurrir que estos parámetros pierdan algo de sentido para evaluar la calidad de la voz cantada porque hay muchos más cambios en frecuencia y volumen en esta voz.

También cabe señalar que los métodos de Machine Learning sobre la detección de patologías de la voz que se incluyen en los papers mencionados en la introducción no se han podido utilizar porque el dataset con el que se cuenta es demasiado pequeño como para aplicar métodos como CNN y las etiquetas que se tienen son demasiado diversas.

Para futuras investigaciones se debería continuar el estudio con un dataset más grande para que los resultados tengan un poder estadístico mayor. Además, al contar con más datos se podrían probar técnicas más actuales y potentes como redes neuronales. Por otro lado, se debería contar con un número de etiquetas menor y que cada audio cuente con un valor correspondiente a cada etiqueta.

Referencias

[1] R. M. Suárez Alvarez y A. Borragán Torre. *Uso de la realidad virtual como medio para reducir o eliminar la rigidez del sistema fonatorio.* 2021.

[2] Y. G Droguett. "Aplicaciones clínicas del análisis acústico de la voz". En: *Revista de otorrinolaringología y cirugía de cabeza y cuello* 77.4 (dic. de 2017), págs. 474-483. DOI: 10.4067/s0718-48162017000400474. URL: https://doi.org/10.4067/s0718-48162017000400474.

[3] J. González, T. Cervera y J. L Miralles. "Análisis acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales". En: *Acta otorrinolaringológica espanola* 53.4 (ene. de 2002), págs. 256-268. DOI: 10.1016/s0001-6519(02)78309-x. URL: https://doi.org/10.1016/s0001-6519(02)78309-x.

[4] Y. De Los Ángeles et al. "Uso de la Escala GRABS en la evaluación perceptual de la voz de pacientes disfónicos". En: *Revista Cubana de Tecnología de la Salud* 6.4 (2015), págs. 78-87. ISSN: 2218-6719.

[5] J. D. Hernández, N. M. L. Gómez y A. J. Álvarez. "Precisión diagnóstica del pico cepstral de mayor prominencia en el cepstrum suavizado (CPPS) en la detección de la disfonía en español". En: *Loquens : revista española de ciencias del habla* 6.1 (feb. de 2019), pág. 058. DOI: 10.3989/loquens.2019.058. URL: https://doi.org/10.3989/loquens.2019.058.

[6] P. Boersma y D Weenink. *Praat: doing phonetics by computer [Computer program].* Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/. 2021.

[7] Y. Jadoul, B. Thompson y B. de Boer. "Introducing Parselmouth: A Python interface to Praat". En: *Journal of Phonetics* 71 (2018), págs. 1-15. DOI: 10.1016/j.wocn.2018.07.001. URL: https://doi.org/10.1016/j.wocn.2018.07.001.

[8] J. Li. *Recent Advances in End-to-End Automatic Speech Recognition*. 2021. DOI: 10.48550/ARXIV.2111.01690. URL: https://arxiv.org/abs/2111.01690.

[9] A. T. Ali, H. S. Abdullah y M. N. Fadhil. "Voice recognition system using machine learning techniques". En: *Materials Today: Proceedings* (abr. de 2021). DOI: 10.1016/j.matpr.2021.04.075. URL: https://doi.org/10.1016/j.matpr.2021.04.075.

[10] S. O. Arik et al. *Deep Voice: Real-time Neural Text-to-Speech*. 2017. DOI: 10.48550/ARXIV.1702.07825. URL: https://arxiv.org/abs/1702.07825.

[11] P. Harar et al. "Voice Pathology Detection Using Deep Learning: a Preliminary Study". En: *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*. IEEE, jul. de 2017. DOI: 10.1109/iwobi.2017.7985525. URL: https://doi.org/10.1109/iwobi.2017.7985525.

[12] S. A. Syed, M. Rashid y S. Hussain. "Meta-analysis of voice disorders databases and applied machine learning techniques". En: *Mathematical Biosciences and Engineering* 17.6 (2020), págs. 7958-7979. DOI: 10.3934/mbe.2020404. URL: https://doi.org/10.3934/mbe.2020404.

[13] M. A. Mohammed et al. "Voice Pathology Detection and Classification Using Convolutional Neural Network Model". En: *Applied Sciences* 10.11 (mayo de 2020), pág. 3723. DOI: 10.3390/app10113723. URL: https://doi.org/10.3390/app10113723.

[14] S. A. Syed, S. Rashid M.and Hussain y H. Zahid. "Comparative Analysis of CNN and RNN for Voice Pathology Detection". En: *BioMed Research International* 2021 (abr. de 2021). Ed. por Wen Si, págs. 1-8. DOI: 10.1155/2021/6635964. URL: https://doi.org/10.1155/2021/6635964.

[15] K. Doshi. *Audio Deep Learning Made Simple: Sound Classification, Step-by-Step*. Mar. de 2021. URL: https://towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by-step-cebc936bbe5.

[16] J. C. Montero y J. Herrera. "Trastornos de la voz". En: *Farmacia Profesional* 17.3 (2003), págs. 56-65. ISSN: 0213-9324.

[17] K. Y. Leon y Andino M. S. "Mecanismos morfofuncionales en la producción de la voz". En: (2021). Ed. por UCE.

[18] M. Marques-Girbau et al. "Vibrato de la voz cantada. Caracterización acústica y bases fisiológicas". En: *Revista de Medicina de la Universidad de Navarra* 50.3 (2006), págs. 65-72. ISSN: 0556-6177. URL: https://hdl.handle.net/10171/35902.

[19] E. Prame. "Vibrato extent and intonation in professional Western lyric singing". En: *The Journal of the Acoustical Society of America* 102.1 (jul. de 1997), págs. 616-621. DOI: 10.1121/1.419735. URL: https://doi.org/10.1121/1.419735.

[20] Y. Maryn, M. De Bodt y N. Roy. "Toward Improved Ecological Validity in the Acoustic Measurement of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels". En: *Journal of Voice* 24.5 (sep. de 2010), págs. 540-555. DOI: 10.1016/j.jvoice.2008.12.014. URL: https://doi.org/10.1016/j.jvoice.2008.12.014.

[21] C. Batthyany et al. "A Case of Specificity: How Does the Acoustic Voice Quality Index Perform in Normophonic Subjects?" En: *Applied Sciences* 9.12 (jun. de 2019), pág. 2527. DOI: 10.3390/app9122527. URL: https://doi.org/10.3390/app9122527.

[22] B. Barsties e Y. Maryn. "External Validation of the Acoustic Voice Quality Index Version 03.01 With Extended Representativity". En: *Annals of Otology, Rhinology &amp Laryngology* 125.7 (mar. de 2016), págs. 571-583. DOI: 10.1177/0003489416636131. URL: https://doi.org/10.1177/0003489416636131.

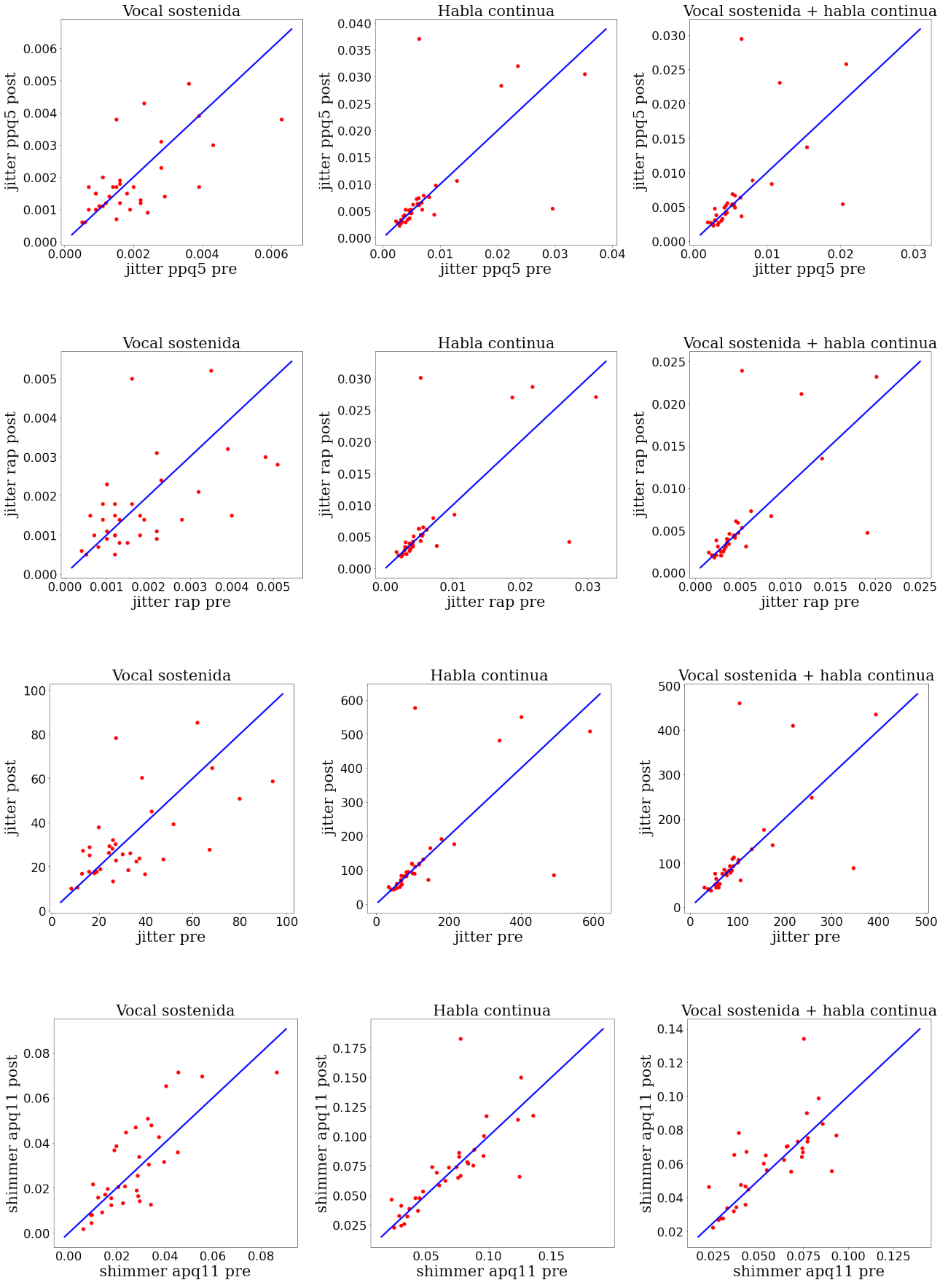
[23] S. Van Vaerenbergh. *Análisis Acústico de la Voz*. Oct. de 2016.

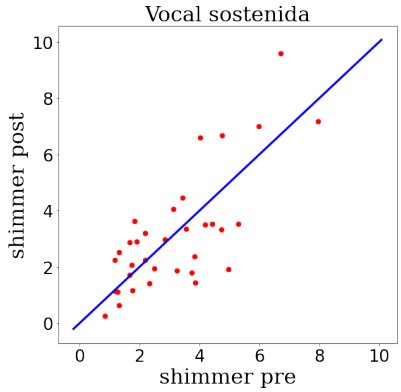
[24] J. P. Teixeira, C. Oliveira y Lopes C. "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters". En: *Procedia Technology* 9 (2013), págs. 1112-1122. DOI: 10.1016/j.protcy.2013.12.124. URL: https://doi.org/10.1016/j.protcy.2013.12.124.

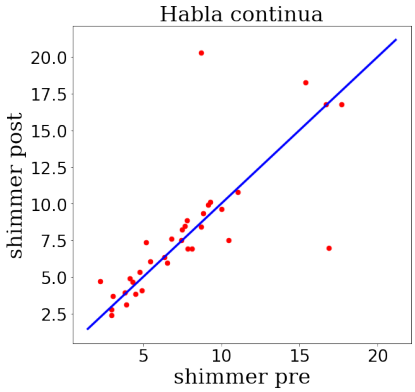
[25] J. Delgado et al. "Análisis acústico de la voz: medidas temporales, espectrales y cepstrales en la voz normal con el Praat en una muestra de hablantes de español". En: *Revista de Investigación en Logopedia* 7.2 (2017), págs. 108-127. ISSN: 2174-5218.

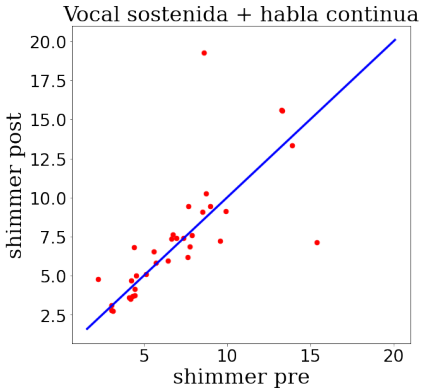
[26] *Density estimation*. https://scikit-learn.org/stable/modules/density.html. Sep. de 2015.

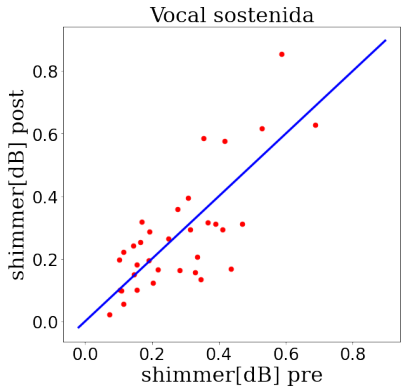
Apéndice

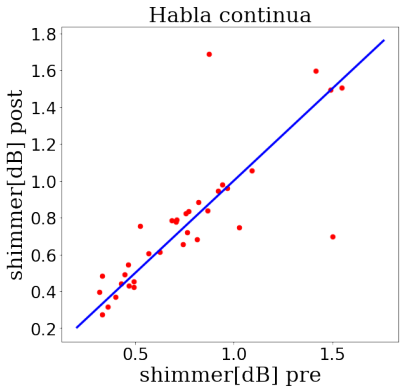

Vocal sostenida
Habla continua
Vocal sostenida + habla continua
jitter ppq5 post
jitter ppq5 pre
jitter rap post
jitter rap pre
jitter post
jitter pre
shimmer apq11 post
shimmer apq11 pre

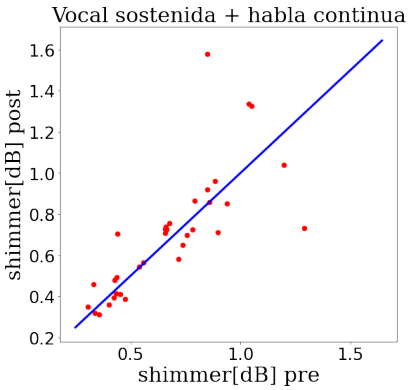

Vocal sostenida
shimmer post
shimmer pre

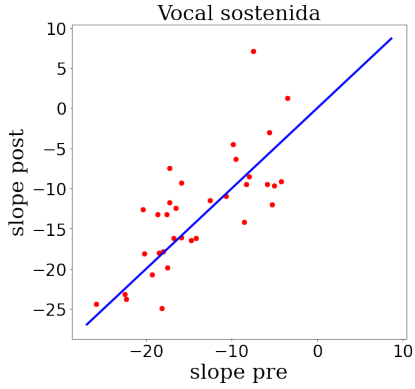

Habla continua
shimmer post
shimmer pre

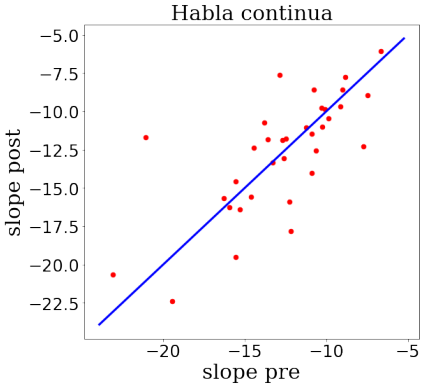

Vocal sostenida + habla continua
shimmer post
shimmer pre

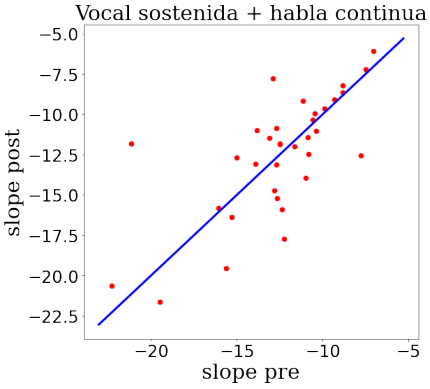

Vocal sostenida
shimmer[dB] post
shimmer[dB] pre


Habla continua
shimmer[dB] post
shimmer[dB] pre


Vocal sostenida + habla continua
shimmer[dB] post
shimmer[dB] pre


Vocal sostenida
slope post
slope pre


Habla continua
slope post
slope pre


Vocal sostenida + habla continua
slope post
slope pre

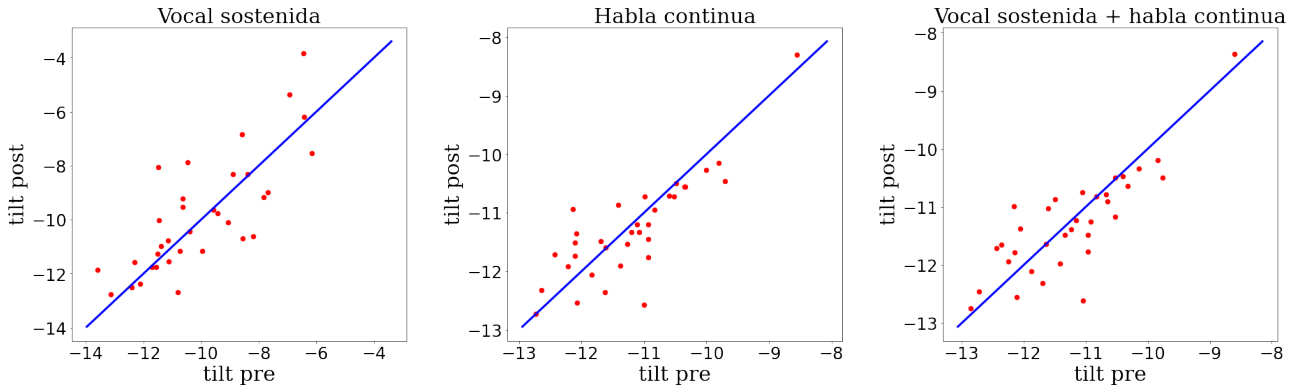


Figura 46: Gráficas en las que se representa cada parámetro post-experimento frente al parámetro pre-experimento para las grabaciones de vocal sostenida, habla continua y la combinación de ambas. Cada punto rojo corresponde a un paciente. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.

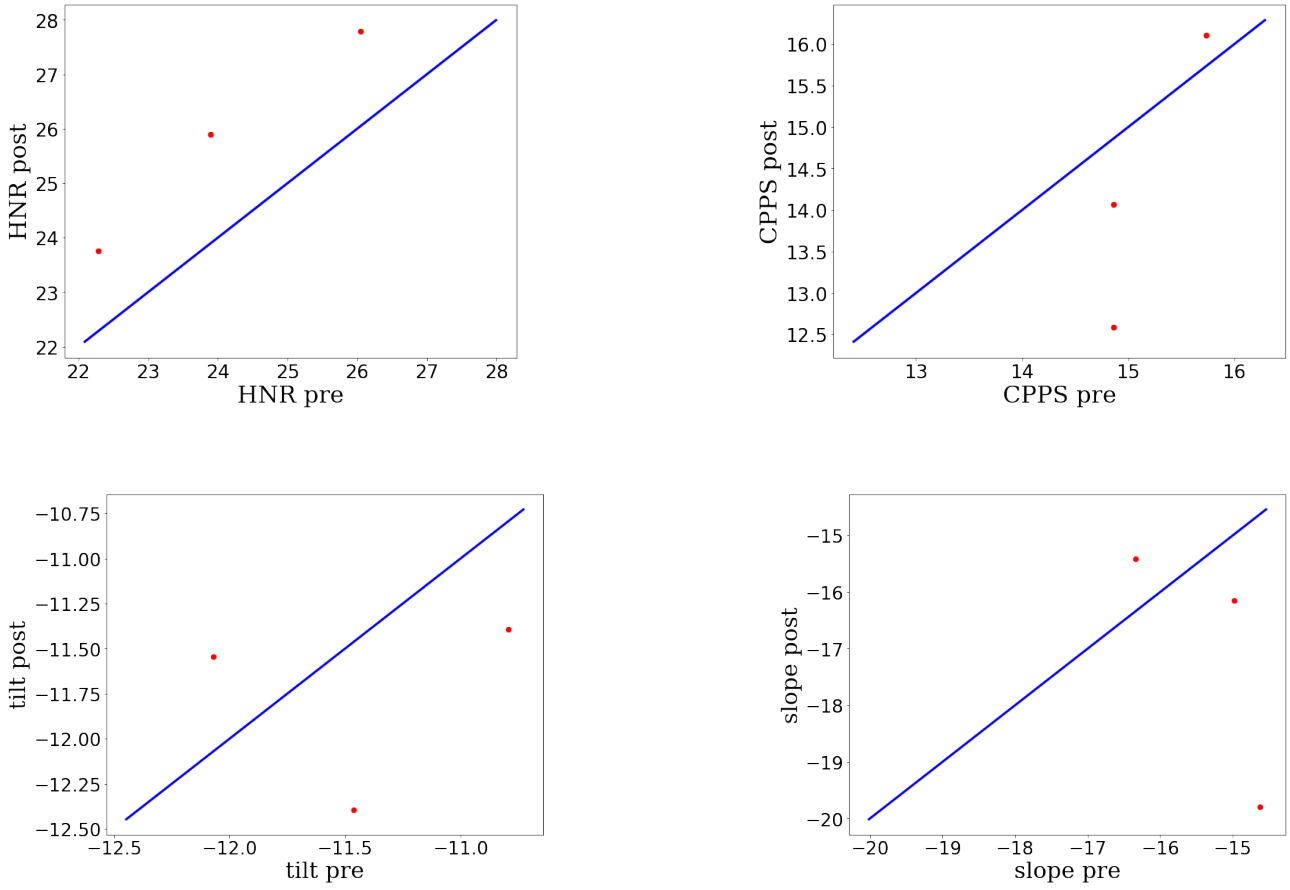


Figura 47: Gráficas en las que se representa cada parámetro post-experimento frente al parámetro pre-experimento para las grabaciones de voz de las tres mujeres que cantan con vibrato. Cada punto rojo corresponde a un paciente. Las rectas azules tienen pendiente 1 y ordenada en el origen 0.