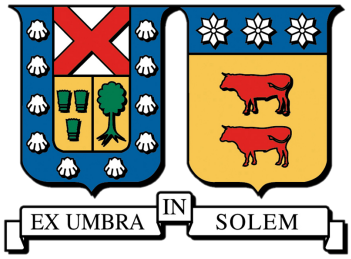
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
VALPARAÍSO - CHILE



"RESÍNTESIS DE VOZ UTILIZANDO HERRAMIENTAS DE PREDICCIÓN LINEAL PARA APLICACIONES DE NEUROCIENCIA"

FELIPE ANDRÉS IGNACIO RODRÍGUEZ MANSILLA

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRÓNICO

PROFESOR GUÍA: MATÍAS ZAÑARTU
PROFESOR CORREFERENTE: CHRISTIAN CASTRO

DICIEMBRE - 2020

AGRADECIMIENTOS

Quisiera comenzar agradeciendo al profesor Matías Zañartu, por haberme guiado, no solo en la realización de este trabajo, sino que también durante mi paso por la universidad, siempre con buena disposición, rigurosidad, y motivándome a mejorar. También al equipo de VPLab, poder participar del grupo ha sido un gran aporte para mí, y me ha dado una perspectiva mucho más completa del estudio de la voz.

Agradezco a mi familia, en especial a mis padres, por su amor, paciencia, y apoyo incondicional en cada aspecto de mi vida. A mi madre por seguirme en mis locuras, y fomentar mi lado creativo. A mi padre, por traspasarme su pensamiento lógico, y por mostrarme el mundo de la ingeniería. A mi abuela, a quien dedico este trabajo, quisiera agradecerle por su inmenso cariño, sus enseñanzas, y por creer siempre en mí. A mi hermano Carlos, por tantas conversaciones, consejos, y ayuda.

Finalmente, dar las gracias a mis amigos, Eduardo, Moses, Vale, Nacho, Nico, Javier, Cony, Zoji, e Isa, por apañarme en este proceso, y por todos los buenos momentos. También a quienes prestaron sus voces para mis experimentos, y a todas las personas que respondieron la encuesta de percepción auditiva.

En memoria de Marie Clavel

Resumen

En este trabajo de título, se hace una revisión bibliográfica de diversos experimentos de feedback auditivo alterado. Se analizan las técnicas y procedimientos utilizados para introducir las alteraciones, y los resultados obtenidos a partir de estos, tomando como referencia el modelo DIVA, y su contribución al entendimiento de los mecanismos cerebrales que están detrás de la producción del habla.

En base a los estudios analizados, se propone una metodología para alterar la calidad vocal de un hablante, a través de técnicas de procesamiento digital de señales. Esta metodología, busca ser un primer acercamiento para realizar un nuevo experimento de feedback auditivo, el cual, serviría para comprender cómo reacciona un hablante ante cambios artificiales en la calidad vocal.

La metodología propuesta consta de 3 partes principales, una separación de las señales en base al modelo Fuente-filtro, una introducción de las perturbaciones, y un proceso de resíntesis basado en herramientas de predicción lineal. Esta metodología ha sido implementada en MATLAB, y puesta a prueba con diferentes métodos de separación y perturbaciones.

Una vez implementada la metodología, se procede a aplicarla a voces de prueba, utilizando para esto, grabaciones de vocales sostenidas y del texto foneticamente balanceado "El Abuelo" realizadas por un grupo de hablantes nativos de español chileno. Se introducen diferentes tipos de perturbaciones, y se realiza un análisis acústico de la calidad vocal antes y después de las perturbaciones, también una encuesta de percepción auditiva de las perturbaciones.

Finalmente, se discuten los resultados, la factibilidad de implementar la metodología en una tarjeta de procesamiento de señales, y el trabajo futuro que habría que hacer para realizar el experimento de feedback auditivo.

Abstract

In this thesis work, a bibliographic review on altered auditory feedback feedback experiments is carried out, focusing on the methods used for these experiments and the results obtained from them. The DIVA model has been taken as a reference for understanding the underlying neuronal mechanisms for voice production.

Based on the reviewed articles, a new methodology is proposed, which has as purpose to alter the vocal quality of a speaker using digital signal processing techniques. This methodology, aims to be a first approach to develop a new altered auditory feedback experiment, which would be useful to understand how a speaker would react to these artificial changes in the vocal quality.

The proposed method, is composed of 3 main parts, a decomposition of the signals using the source-filter model, an alteration, and a resynthesis process based on Linear Prediction methods. This methodology has been implemented in MATLAB, and tested with different separation methods and perturbations.

Once the methodology was implemented, it was applied to a set of voices, using recordings of sustained vowels, and the phonetically balanced text "El Abuelo" done by a group of native spanish speakers from Chile. Different types of alterations were tested and the vocal quality was measured before and after the perturbations using acoustic measures. Also, a survey was carried out to measure the auditory perception of the alterations.

Finally, there is a discussion about the results, the feasibility of implementing the proposed method in a DSP board, and the future work needed to perform the experiment.

Índice general

Capítulo 1

Introducción

1.1 Motivación y problema a resolver

La voz es una herramienta imprescindible para el desarrollo de diversas actividades humanas, y es la manera más efectiva que tenemos de comunicarnos, y expresarnos. Para que una persona hable, su cerebro debe realizar de forma coordinada, complejos procesos cognitivos, motores y sensoriales. Fallas en estos procesos, pueden producir patologías del habla, las cuales incluyen desde problemas leves en la fonación o articulación, falta de fluidez, y problemas para pronunciar ciertos sonidos, hasta la pérdida total de la capacidad de hablar. Cuando se producen patologías del habla, quienes las padecen, ven afectada su capacidad de comunicarse con los demás, lo cual puede generar problemas sociales, laborales y psicológicos [1]. Se estima que, en Estados Unidos, 1 de cada 13 personas padece algún tipo de trastorno del habla durante el año [2].

Comprender los mecanismos neuronales que subyacen a la producción de la voz, permite explicar la causa y el funcionamiento de algunas de estas patologías, y proponer soluciones o tratamientos más efectivos para las personas que las padecen. Para estudiar los mecanismos cerebrales que hacen posible el lenguaje hablado, se han realizado experimentos, donde se mide la actividad cerebral de la persona, mientras realiza tareas relativas al habla. Estos experimentos se llevan a cabo, empleando técnicas como el Electroencefalograma (EEG), y la Imagen por Resonancia Magnética Funcional (fMRI), y permiten analizar las conexiones que existen entre diferentes zonas de nuestro cerebro, y cómo estas se coordinan para producir los movimientos necesarios para que una persona pueda hablar. Gracias a estos experimentos, ha sido posible formular DIVA, un modelo matemático computacional que simula un grupo de estructuras neuronales encargadas del control motor del habla. DIVA ha servido como herramienta para el estudio de patologías originadas en fallos neuronales, y también como base teórica para realizar nuevos experimentos relacionados. De los diferentes mecanismos presentes

en el modelo DIVA, en este trabajo, se pondrá el énfasis en el mecanismo de feedback auditivo, el cual funciona como un lazo cerrado de control que, utilizando los sonidos captados por nuestros oídos, nos permite monitorear y corregir, la manera en que producimos los sonidos al hablar. Una manera de estudiar los mecanismos de feedback auditivo, es alterar, de manera artificial, la forma en que un hablante percibe su propia voz, para luego medir la respuesta del hablante ante esta alteración. Esto se conoce como experimentos de feedback auditivo alterado. Los primeros experimentos de este tipo, demostraron que las personas cambian su forma de hablar en base a como se escuchan a sí mismas. A medida que ha avanzado la tecnología, ha sido posible proponer experimentos cada vez más complejos, que han permitido comprender e identificar las estructuras detrás de los procesos de feedback auditivo. En la actualidad, las tarjetas de procesamiento de señales digitales, hacen posible formular experimentos en donde se perturba alguna característica específica de la voz en tiempo real, y con esto, se ha observado que el mecanismo de feedback auditivo es capaz de diferenciar perturbaciones muy sutiles, y alterar la forma en que se produce la voz intentando corregir la perturbación escuchada. Este trabajo se centra en explorar métodos de procesamiento de señales, que permitan alterar la calidad vocal de una señal de voz, en el contexto de los experimentos de feedback auditivo. Estos métodos deberán ser simples y eficientes en términos de cómputo, y deberán producir una señal que suene natural. No se ha realizado a la fecha un experimento con perturbaciones de este tipo, y de llevarse a cabo, podría entregar información valiosa acerca de cómo percibimos nuestra propia voz, y de como funcionan los mecanismos neuronales encargados de la fonación.

1.2 Hipótesis

Este trabajo intenta responder las siguientes preguntas: **¿Es posible generar una metodología simple para introducir perturbaciones a la fuente de una señal de voz? ¿Se puede diseñar e implementar en MATLAB un algoritmo simple y eficiente para esta tarea?**

Basándose en estas preguntas la hipótesis principal de este trabajo es: **"El diseño e implementación de un algoritmo simple de procesamiento de señales, capaz de introducir perturbaciones que afecten solamente a la señal que proviene de la glotis**.

1.3 Objetivos

1.3.1 Objetivos generales

Explorar técnicas de procesamiento de señales sencillas y eficientes, que permitan alterar la calidad vocal de un hablante, en el contexto de los experimentos de feedback auditivo.

1.3.2 Objetivos específicos

- Diseñar una metodología de procesamiento de señales que introduzca cambios a la calidad vocal de manera simple y eficiente.
- Implementar la metodología diseñada en MATLAB, y realizar pruebas con diferentes hablantes, utilizando vocales sostenidas, y frases.
- Medir el efecto de la metodología aplicada, de manera instrumental y perceptual.

Capítulo 2

Estado del arte

2.1 Fonación, voz y habla

En esta sección, se busca explicar cómo ocurre el proceso de fonación, que da origen a la voz, y al habla. Para esto se realiza una breve explicación de la anatomía de la laringe, y de los pliegues vocales. Luego se explican los procesos físicos que forman parte del fenómeno fonatorio.

2.1.1 Fonación

La fonación, es el proceso mediante el cual se producen los sonidos que componen la voz, este ocurre en la laringe, e involucra diversas estructuras anatómicas. La fonación, comienza con los pulmones empujando aire a través de la tráquea, este aire va a dar a la laringe, donde pasa a través de los pliegues vocales. Los pliegues vocales pueden adoptar diferentes posiciones, y dependiendo de la manera en que estos se encuentren, vibrarán de diferentes formas, dando origen a los tipos de fonación. El sonido generado a través de la fonación, pasa por la cavidad bucal, y continúa siendo modificado a través de movimientos articulatorios. Los procesos conjuntos de fonación y articulación, son los que dan origen al habla.

2.1.2 La Laringe

La laringe o caja laríngea es un órgano tubular, que se encuentra ubicado en la parte anterior del cuello. Está compuesta por estructuras óseas, cartilaginosas, y musculares, que albergan los pliegues vocales. La caja laríngea cumple diversas funciones, ya que forma parte del sistema respiratorio, y también del aparato fonador. En este sentido, la laringe es capaz de regular los flujos de aire desde y hacia los pulmones, participando en la respiración, fonación y deglución.

Los músculos intrínsecos de la laringe, son el interaritenoideo, el aritenoideo transversal, el aritenoideo oblicuo, los cricoaritenoideos, los cricotiroideos, y los tiroaritenoideos. Estos músculos conectan y desplazan los cartílagos que funcionan como soporte estructural de la laringe, y la conectan con otras estructuras del cuerpo. Los cartílagos de la laringe son el cricoides, el tiroides, los aritenoides, y la epiglotis (Ver Figura 2.1).

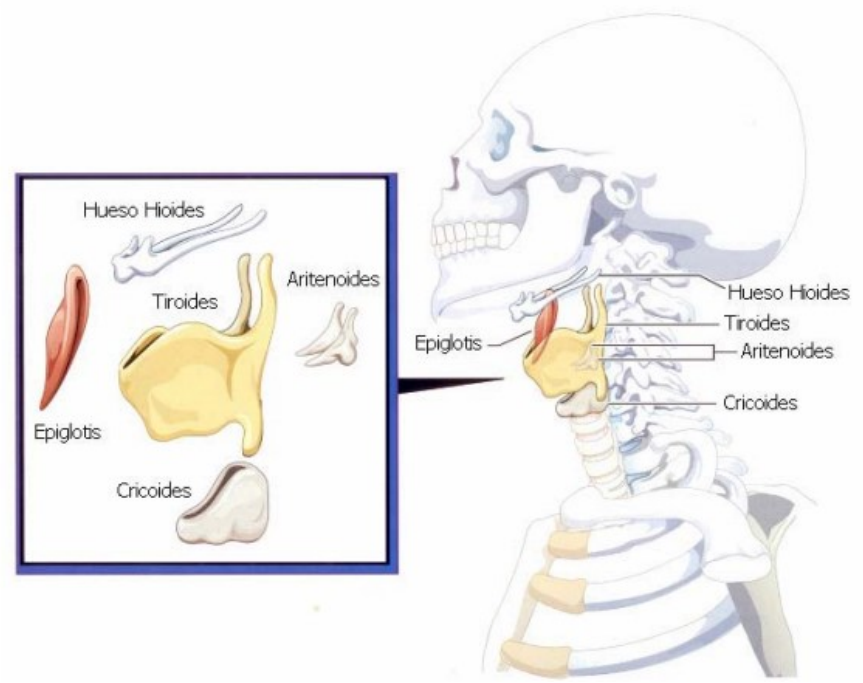


Figura 2.1: Estructura de la Laringe. Imagen de [79]

2.1.3 Los Pliegues Vocales

Los pliegues vocales, más conocidos como cuerdas vocales, son 2 pliegues ubicados dentro de la estructura laríngea, en el extremo superior de la tráquea y son una parte imprescindible de la producción de la voz. También participan de otras funciones como la respiratoria, ya que regulan el paso del aire desde y hacia los pulmones. En este trabajo se pone el énfasis en la función fonatoria de los pliegues vocales.

Están compuestos por diversas capas, desde la más superficial a la más interna encontramos: el epitelio, la lámina propia (dividida a su vez en superficial, intermedia y profunda) y el músculo vocal que correspomde al tiroaritenoideo. Desde un punto de vista mecánico, estas capas se pueden agrupar en tres secciones: la mucosa (formada por el epitelio y la capa superficial de la lámina propia), el ligamento

(compuesto por las capas intermedia y profunda de la lámina propia) y el músculo vocal (Ver Figura 2.2).

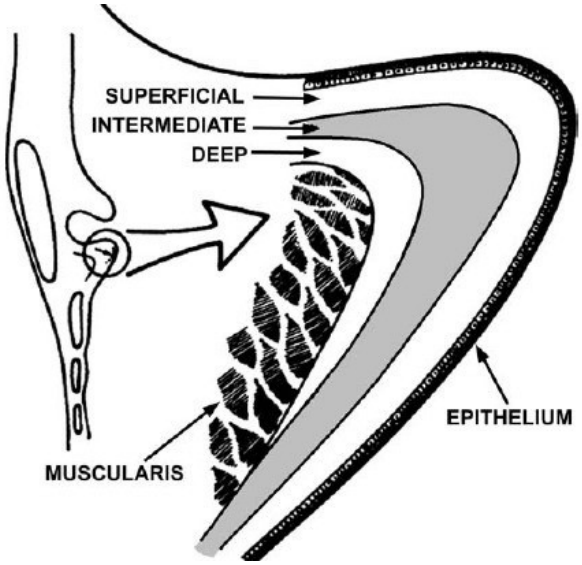


Figura 2.2: Estructura de los Pliegues Vocales. Imagen de [80]

Las capas de los pliegues vocales tienen diferente composición histológica que las dota de distintas propiedades visco elásticas que afectan a la vibración. Cualquier cambio en la composición de las cuerdas vocales, ya sea por la edad, cambios hormonales o patologías, tiene una repercusión en la calidad de la voz.

2.1.4 La Glotis

La glotis, o apertura glotal, es el espacio que está delimitado por los pliegues vocales, y por los cartílagos aritenoides. Gracias a la acción de los músculos de la laringe, es posible variar el área glotal, lo que nos permite respirar, aguantar la respiración, cerrar el acceso a los pulmones durante la ingesta de alimentos, y también, producir sonidos a través de la fonación (Ver Figura 2.3). Dependiendo de la manera en que coloquemos nuestra abertura glotal, podemos generar diferentes tipos de fonación, que dan origen a diferentes tipos de voz, como la voz modal, la voz susurrada, la voz chirriante, y la voz respirada (Ver Figura 2.4).

Figura 2.3: Abducción y Aducción de la glotis. Imagen de [79]

2.1.5 El fenómeno Fonatorio

Ahora que se han mencionado las diferentes partes que intervienen en la fonación, se procederá a explicar el fenómeno fonatorio, para una voz modal.

La dinámica de las cuerdas vocales puede dividirse en 2 partes, posicionamiento y vibración. El posicionamiento, ocurre gracias a movimientos de desplazamiento de los cartílagos por la acción de fuerzas musculares. La vibración, ocurre por pequeñas y rápidas deformaciones que se producen de manera oscilatoria por la acción de los cambios de presión del aire en la laringe.

Para que haya fonación, primero debe ocurrir la inspiración, durante esta se abducen los pliegues vocales, dejando la glotis abierta, y permitiendo el paso del aire desde el exterior hacia los pulmones. Luego las cuerdas se desplazan para iniciar la fonación.

1. Con el aire dentro de los pulmores, comienza el proceso de espiración, durante el cual los músculos aductores aproximan los pliegues vocales, haciendo contacto en la línea media. Esto junto con la espiración, genera un aumento de la presión subglótica.

2. Eventualmente, esta presión acumulada, se hace mayor que la fuerza que mantiene unidos los pliegues vocales, y estos se abren momentáneamente para dejar salir el aire.

3. La salida del aire, produce una baja en la presión subglótica, esto combinado con las propiedades elásticas de los pliegues vocales, hace que estos vuelvan a cerrarse.

4. Este cierre aumenta nuevamente la presión sobre los pliegues vocales, lo que hace que vuelvan a abrirse, comenzando nuevamente el ciclo.

Este ciclo se mantiene mientras exista una presión de aire impulsado por el diafragma. Estos cambios de presión periódicos, se propagan por la cavidad bucal, y al salir por la boca, dan origen al sonido que escuchamos.

Este ciclo periódico termina cuando la presión ejercida por los pulmones no es suficiente para separar los pliegues vocales, y puede reiniciarse con una nueva inspiración.

En personas sanas, la frecuencia a la que vibran los pliegues vocales es relativamente constante, y corresponde a la frecuencia fundamental de la voz. Esta frecuencia depende de diversos factores, como la edad, el sexo, el tamaño de la laringe, entre otros. Para hombres, la frecuencia fundamendal suele encontrarse en el rango de los 100 a los 140 [Hz]. Para las mujeres, ronda los 180 a 220 [Hz], y en el caso de las niñas y niños, esta suele ser más alta, rondando los 260 a 280 [Hz] [81]. Las personas que entrenan la voz, pueden lograr amplios rangos de frecuencia fundamental, abarcando desde 60 a 1500 [Hz] en casos excepcionales.

Para que la fonación ocurra, se requieren una serie de condiciones, sin las cuales, el mecanismo de producción del ciclo vibratorio no sería factible: es necesario que la presión de aire sea suficientemente fuerte como para separar los pliegues vocales; una glotis estrecha y un cuerpo muscular elástico; así como una mucosa suficientemente laxa, húmeda y libre de fijación al plano medio, con el fin de que sea capaz de ondular y desplazarse por una mínima presión negativa. La alteración de cualquiera de estas circunstancias puede afectar a la dinámica de la cuerda vocal, generando alteraciones de la voz. De todas ellas, la que más trascendencia tiene es la variación de las características físicas de la mucosa, pues la pérdida de ésta o su fijación al plano medio son incorregibles, causando una disfonía permanente.

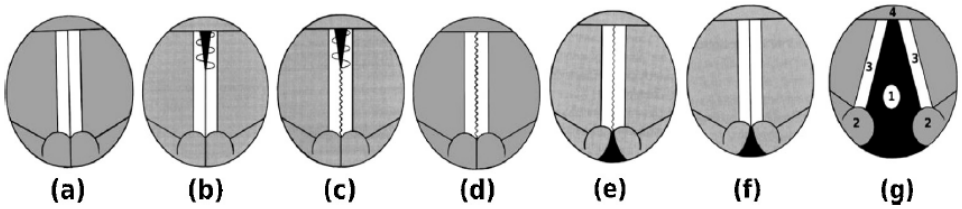


Figura 2.4: Configuraciones de la glotis que dan origen a diferentes tipos de fonación: (a)Glotis Cerrada, (b)Chirrido, (c)Voz chirriante, (d)Voz modal, (e)Voz respirada, (f)Susurro, (g)Partes de la glotis (1: Glotis,2: Cartílago Aritenoides, 3: Pliegues Vocales, 4: Epiglotis). Imagen de [46]

2.1.6 Diferencia entre Voz y Habla

La voz o vocalización, es el resultado del proceso de fonación que ocurre en la laringe cuando los pulmones empujan un flujo de aire que hace vibrar los pliegues vocales. Si bien la voz es una parte fundamental del habla, esta también se da en otras actividades, como la risa, el llanto, el canto, gemidos, entre otros. Incluso, otras especies de mamíferos tienen un aparato fonador que les permite vocalizar, pero no son capaces de hablar. El habla, es un proceso a través del cual los seres humanos son capaces de modificar y moldear el tono producido por la voz, para crear sonidos específicos y decodificables. Estos sonidos, nos permiten expresar pensamientos, sentimientos, e ideas de manera oral, y son la base de la comunicación humana. A diferencia de la producción de la voz, el habla requiere de una coordinación precisa de movimientos musculares, que ocurren en la cabeza, el cuello, el pecho y el abdomen, y es una habilidad que se adquiere de manera gradual en la infancia temprana, y que requiere años de práctica, en donde los niños aprenden complejos mecanismos motores que permiten producir un habla comprensible para otros seres humanos [82].

2.2 Modelo DIVA

El modelo DIVA (Directions Into Velocities of Articulators)[4] es una red neuronal adaptativa, que describe las interacciones sensorio-motoras involucradas en el control articulatorio del habla. Este modelo fue creado buscando mejorar la comprensión de los mecanismos neuronales que están detrás del control del la producción del habla, y es fruto de un gran número de estudios multidisciplinarios que combinan diferentes áreas de la medicina y la ingeniería.

Hasta la fecha, DIVA se ha utilizado para guiar diversos experimentos, y realizar predicciones y comparaciones entre los datos calculados a partir de simulaciones, y los obtenidos empíricamente. Los resultados de los experimentos han sido utilizados para realizar mejoras y ajustes al modelo en base a lo observado en sujetos de prueba. Una de las características más importantes del modelo, es que cada uno de sus bloques ha sido asociado a una zona específica del cerebro, es decir, la función que realiza cada bloque en el modelo, es equivalente a la acción de una determinada región del cerebro de una persona. Las ubicaciones de cada bloque, están descritas de acuerdo al atlas del MNI (Montreal Neurological Institute)[5]. A continuación, se explicará a grandes rasgos el funcionamiento del modelo, y algunas de sus partes.

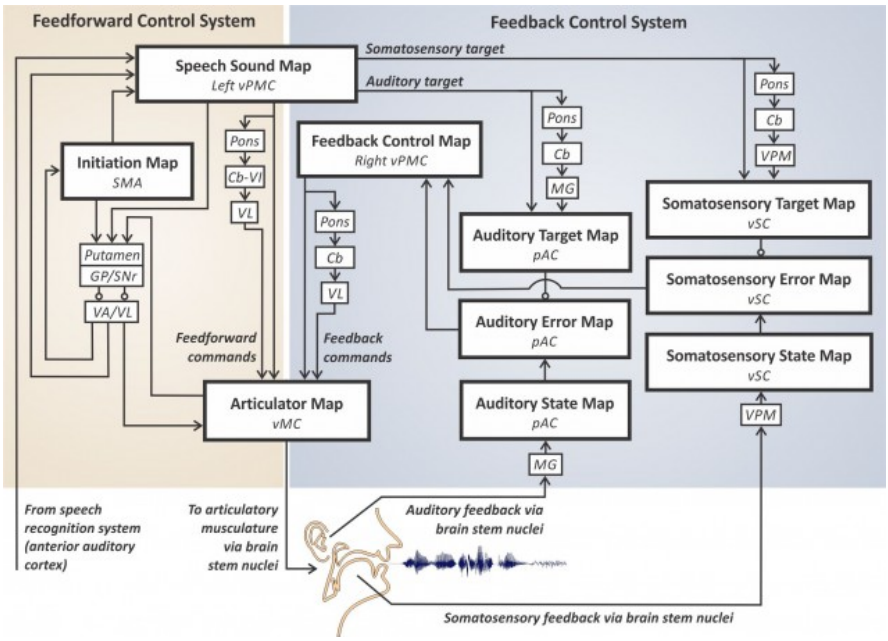


Figura 2.5: Diagrama de bloques del Modelo DIVA

2.2.1 Funcionamiento del Modelo

El modelo DIVA es una red neuronal, que una vez entrenada, recibe como entrada una señal de audio de voz, y en base a esta, genera una secuencia de posiciones articulatorias que controlan los movimientos de un tracto vocal simulado. Pueden verse en el modelo 2 subsistemas, "feedforward" y "feedback", los cuales trabajan en conjunto para generar y ajustar la secuencia de posiciones articulatorias. El modelo DIVA está compuesto por diferentes “mapas”, que representan a un conjunto de neuronas, que están ubicadas en un lugar del cerebro. En la figura 2.5 cada bloque representa un "mapa", y tiene asignado un lugar en base al atlas MNI. Las flechas del modelo representan transformaciones de una representación neuronal a otra. La producción del habla comienza con la activación de un mapa de sonidos del habla (Speech Sound Map), en el cual se encuentran representados todos los sonidos presentes en el idioma hablado por el hablante. Luego la representación es proyectada a un mapa articulatorio (Articulator Map), en el cual están representados todos los movimientos articulatorios necesarios para producir los sonidos presentes en el mapa de sonidos. Los mecanismos de feedback pueden dividirse en dos partes, un feedback somato-sensorial, y un feedback auditivo. Así como cada sonido de un idioma está asociado en el modelo a un determinado movimiento articulatorio, está también asociado a un mapa de objetivos auditivos (Auditory Target Map), y a un mapa de objetivos somatosensoriales (Somatosensory Target Map). Los “mapas objetivo” representan la percepción que tenemos sobre nosotros mismos cuando hablamos. Explicado de otra forma, cada vez que pronunciamos un sonido, espera-

mos a que este suene, y que se sienta de una determinada manera. Lo que hace el sistema de feedback, es generar mapas con el estado auditivo y somatosensorial (Auditory State Map y Somatosensory State Map respectivamente) en base a la información recopilada por el sistema auditivo y somatosensorial. Comparando los "mapas objetivo", con los "mapas de estado", es posible generar "mapas de error". En caso de existir una discrepancia entre lo esperado y lo ocurrido, existen otros bloques del modelo encargados de modificar el mapa articulatorio con el fin de corregir los movimientos articulatorios que no generen los resultados esperados. Este trabajo se enfocará en los diferentes experimentos que se han realizado para estudiar el mecanismo de feedback auditivo, y en proponer un nuevo experimento.

2.2.2 Feedback Auditivo

El feedback auditivo, corresponde a la percepción que tenemos de nuestra propia voz mientras hablamos. Esta señal llega a nuestro aparato auditivo a través de 2 caminos, uno de ellos corresponde a lo que se conoce como el tono lateral (sidetone), que es la señal sonora que se propaga por el aire y entra por nuestro oídos. La otra fuente de feedback auditivo, corresponde a la conducción ósea (bone conduction), la cual se debe a la propagación de las ondas mecánicas por nuestra cabeza, principalmente a través de los huesos del cráneo. Ambas señales se mezclan y entran juntas a través de nuestro sistema auditivo.

En el modelo DIVA, el feedback auditivo se procesa en un mecanismo neuronal, que se encarga de comparar los sonidos que producimos al hablar (Mapa Auditivo Actual), con los sonidos que esperábamos producir (Mapa Auditivo Objetivo). En base a las diferencias que ocurran entre ambos, nuestro cerebro actúa sobre la manera en que producimos la voz (Mapa Articulatorio). Se ha observado que este mecanismo, es fundamental para el aprendizaje del habla, ya que cuando los niños pequeños balbucean, lo que hacen es establecer relaciones entre movimientos articulatorios, y el sonido que estos movimientos producen [7], basándose en esta idea, es que DIVA entrena y ajusta los pesos de su red neuronal durante la fase de entrenamiento.

Además de su rol durante el aprendizaje, se ha podido observar que el feedback auditivo se encarga de ajustar nuestra manera de hablar a lo largo de nuestra vida. Esto se ha comprobado mediante personas que pierden la audición ya sabiendo hablar. Estas personas cuentan con un mapa de sonidos del habla, que está relacionado con el mapa de velocidades y posiciones articulatorias, por lo cual, pueden hablar, aún cuando no pueden escucharse, pero se ha visto que a medida que pasa el tiempo, su capacidad de hablar se va deteriorando, comienzan a presentar problemas con ciertos sonidos, pierden inteligibilidad y prosodia [8, 9]. De acuerdo a estas observaciones se entiende que el feedback auditivo no solo nos ayuda a generar los mapas de sonidos y articulatorios, sino que también, los

mantiene vigentes a lo largo de nuestras vidas, y los va adaptando a los cambios que ocurran en nuestra fisiología.

Además de servirnos durante el aprendizaje del habla, y para mantenerla a largo plazo, se ha observado que el feedback auditivo también tiene un rol en el corto plazo, que se encarga de monitorear los sonidos que generamos. Para estudiar el mecanismo de feedback auditivo, se ha utilizado un esquema de experimento, que consiste en alterar artificialmente la manera en que un hablante se escucha a sí mismo. Esto se realiza alterando el sidetone, ya que, por el momento, no es posible alterar el bone conduction. Con estos experimentos, se ha podido observar que los hablantes producen cambios inmediatos y largo plazo en su manera de hablar cuando se les altera el feedback auditivo. A medida que ha avanzado la tecnología, ha sido posible plantear experimentos cada vez más complejos, que ponen a prueba estos mecanismos de feedback, y permiten extraer información valiosa sobre los sistemas neuronales detrás de la producción del habla.

Estos experimentos también han sido realizados en conjunto con otras técnicas que permiten medir la actividad cerebral, como el EEG y el fMRI, los cuales han permitido asociar los mecanismos de feedback auditivo a zonas específicas del cerebro [10, 11].

2.3 Experimentos de feedback auditivo alterado

A continuación, se presenta el diagrama general que se utiliza para realizar este tipo de experimentos (Ver Figura 2.6), a grandes rasgos, una persona realiza alguna tarea relacionada con el habla, como leer frases, sostener una vocal, o comunicarse con otra persona. El audio es capturado por un micrófono, y a través del algún sistema, este audio es perturbado artificialmente, y devuelto a la persona a través de un par de auriculares con aislación, de manera tal, que la persona no se escuche al hablar, sino que escuche solamente el audio perturbado. Siempre hay un sistema de grabación de las señales, que captura la señal de voz emitida por el sujeto de pruebas, y la señal alterada que es entregada a través de los auriculares. Estos experimentos, se realizan en reiteradas ocasiones, con diferentes condiciones y sujetos de prueba.

Luego del experimento es posible analizar la voz de cada persona, antes, después y durante la perturbación, y así observar las adaptaciones que hace la persona a su propia voz una vez que esta es perturbada. Por lo general estos experimentos se realizan en una cámara con aislación acústica, con hablantes que no presenten patologías de la voz o audición, y luego del experimento se procede a aplicar una batería de preguntas para determinar cómo los sujetos de prueba percibieron el experimento, y si detectaron las perturbaciones.

Con esta metodología, es posible descifrar la manera en que actúa el mecanismo de

feedback auditivo. Este tipo de experimentos han permitido entender la dinámica de este mecanismo, y han aportado a generar mejores explicaciones y tratamientos para algunas patologías del habla. Además de las mediciones acústicas, estos experimentos pueden realizarse en conjunto con otras técnicas de medición como el EEG y el fMRI, para estudiar qué zonas de nuestro cerebro se activan durante el experimento, cuánto tardan en activarse, y en qué orden lo hacen.

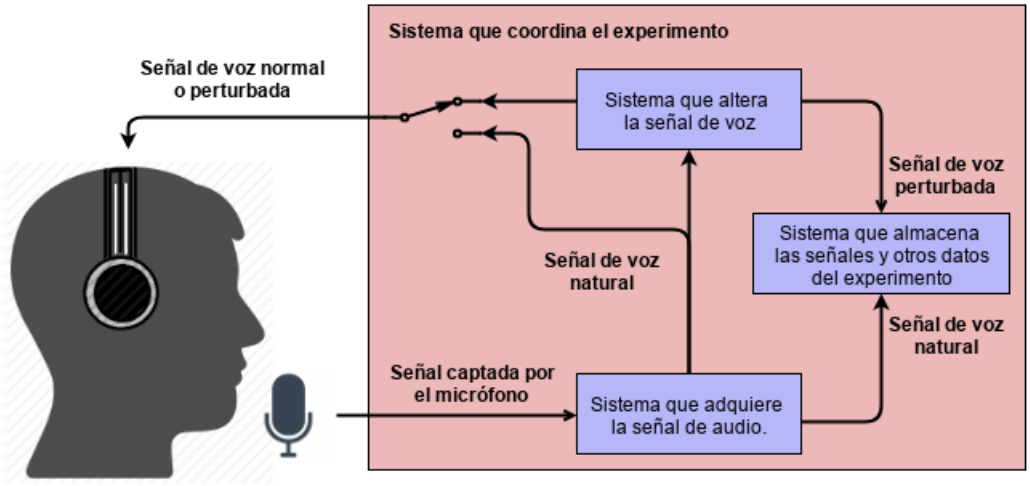


Figura 2.6: Esquema general de experimentos de feedback auditivo

Se procederá a describir algunos de los experimentos que se han realizado hasta la fecha, estos son presentados en orden cronológico, y se entrega una breve explicación de las herramientas tecnológicas que han sido utilizadas para su realización, y también de cuáles son los aprendizajes que se ha adquirido a partir de cada uno de ellos. Muchas de las perturbaciones presentadas han sido repetidas en estudios posteriores, aplicando variaciones, y aprovechando nuevas tecnologías. En este trabajo se busca describir el método utilizado en la primera vez que se tiene registro de cada perturbación, para entender los avances tecnológicos que la hicieron posible en su momento.

2.3.1 El reflejo de Lombard

Los resultados del primer experimento de perturbación del feedback auditivo, fueron publicados por Etienne Lombard en 1911. En este experimento, Lombard utilizó un aparato que generaba un fuerte ruido directo en el oído, y lo aplicó a un paciente que mantenía una conversación. Se pudo observar que, al aparecer el ruido, el sujeto de prueba tendía a hablar más fuerte, y al eliminar el ruido, volvía a disminuir el volumen al hablar, hoy en día esto se conoce como el “reflejo de

Lombard" [3]. Los resultados de este experimento, contribuyeron enormemente a diversas áreas de estudio, como la producción del habla, los efectos de la pérdida de la audición, y la relación entre habla y audición. Cabe mencionar, que el aparato utilizado en el experimento, fue diseñado para cancelar la audición de un oído, para poder realizar pruebas en el otro oído. Este aparato funcionaba de manera mecánica, con un pequeño martillo que golpeaba una membrana, generando un ruido que era introducido en el oído a través de un espéculo [12]. Este experimento fue replicado por otros investigadores, de forma mononaural y binaural solicitando a los sujetos de prueba que realizaran diferentes tareas como conversar con otros sujetos, leer frases, gritar frases, entre otros. En todas las pruebas, se pudo observar que hubo un aumento en el volumen de la voz en presencia de ruido, y que este fue mayor, en los casos en que el sujeto intentaba comunicarse con otro. De esta manera, se concluyó que, al hablar, las personas intentan mantener una relación señal-ruido favorable para la comunicación. El reflejo de Lombard también ha podido observarse en algunas especies de aves, que cambian su forma de cantar en presencia del ruido de las ciudades [6].

2.3.2 Feedback Auditivo Retrasado (DAF)

Con el desarrollo de la electrónica, fue posible realizar experimentos más complejos. En 1951, se desarrolló un sistema de grabación y reproducción utilizando una cinta magnética [13], la cual funcionaba en un bucle, con 2 cabezales, uno conectado a un micrófono, encargado de grabar una señal de audio en la cinta, y el otro, conectado a unos audífonos, encargado de reproducirla. Ajustando la distancia entre ambos cabezales, era posible generar retardos auditivos entre 0 y 350$[ms]$. Se realizaron pruebas con 22 hablantes, los cuales leyeron frases de 5 sílabas, y fueron expuestos a un feedback retrasado de entre 0 y 300$[ms]$. Se pudo observar, que retardos pequeños, hasta 180$[ms]$, generaron que los hablantes alargaran las palabras, y por ende, hablaran un poco más lento, para los delays mayores, los hablantes tendieron a acelerar la lectura de las frases. También se pudo observar que los hablantes tendieron a hablar más fuerte en presencia del feedback retrasado. En algunos casos, se presentó un bloqueo del habla, y también algunos sujetos comenzaron a tartamudear.

Este experimento de feedback auditivo ha sido replicado en una gran cantidad de estudios posteriores, con tecnologías más precisas, y probando con diferentes retardos. En general se ha podido observar que retardos de menos de 30 ms, no son percibidos por los sujetos de prueba, y que entre los 50 y 200ms, los sujetos muestran diferentes reacciones. Personas sanas comienzan a tartamudear, presentan disfluencias o paran de hablar [18, 19]. En general se observa frustración e incluso, se observó que un retardo de 175 ms induce un estado de estrés mental [14].

Al realizar este experimento con personas tartamudas, se ha podido observar que,

en algunos casos, es posible hayar un tiempo de retardo específico para una persona tartamuda, que produce una disminución en la tartamudez [15, 16], en base a estos descubrimientos, se han desarrollado dispositivos electrónicos, que generan un pequeño retardo auditivo, y han dado buenos resultados en el tratamiento de personas tartamudas [17].

2.3.3 Feedback Auditivo con alteraciones en frecuencia (FAF)

Con la aparición de la electrónica digital, fue posible proponer experimentos, en los que se realizaran corrimientos en frecuencia. En el primero de estos experimentos, se pidió a un grupo de personas que intentaran mantener vocales sostenidas, imitando una voz que les fue mostrada previamente. En algunas de las iteraciones, el feedback auditivo fue alterado, aumentando o disminuyendo la frecuencia de la voz en un 10 %. Se pudo observar que cuando la frecuencia del feedback fue aumentada o disminuida, los hablantes tendieron a disminuir o aumentar la frecuencia de su voz respectivamente, buscando ajustar la frecuencia escuchada con la esperada. La mayoría de los sujetos no notaron el cambio en el feedback auditivo [20]. En el mismo estudio, se realizó otro experimento, que obtuvo similares resultados.

Para este primer experimento, se utilizó un "Lexicon Varispeech II", el cual, de manera electrónica realizó un corrimiento en la frecuencia de la señal de voz ingresada. Para esto, el aparato contaba con 3 etapas esenciales, que son las mismas que tienen las tarjetas de procesamiento modernas que se utilizan hoy en día. Un conversor A/D, una etapa de procesamiento digital, y un conversor D/A. Para generar los corrimientos en frecuencia, el Varispeech es capaz de hacer estiramientos o compresiones en el tiempo, las cuales se traducen en corrimientos en frecuencia. Su circuito interno, es capaz de elegir qué muestras descartar, y también en qué lugares colocar silencios de forma que la señal no quede con saltos abruptos [21]. Luego de este experimento, se realizaron diversas pruebas similares, donde se pudo observar que en presencia de alteraciones en frecuencia, los hablantes tardan entre 100 y 150ms en intentar compensar la alteración, esta respuesta se conoce como "reflejo de corrimiento de frecuencia" (pitch-shift reflex) [22]. Una pregunta que surgió de estos experimentos, es si la frecuencia fundamental se ajusta en base a una frecuencia objetivo fija, o si se ajusta en base a la frecuencia anterior que se escuchaba antes de introducir la perturbación. Buscando respuesta a esta pregunta, se hizo un experimento con 2 tipos de perturbaciones, una que se introdujo en la mitad de la vocalización (condición de encendido), y otra que fue introducida antes de que la persona comenzara a vocalizar, y luego fue removida en medio de la vocalización (condición de apagado) [23, 24]. En estos experimentos se pudo ver que los hablantes responden de manera muy similar a las condiciones de encendido y apagado en medio de una vocalización. Para las perturbaciones introducidas previamente a la vocalización, se pudo observar una respuesta más grande. Se logró

concluir que para regular la frecuencia existen 2 mecanismos diferentes, uno que se encarga de comparar la frecuencia al inicio de una vocalización con la frecuencia esperada. El otro mecanismo se encarga de comparar la frecuencia actual con la frecuencia más reciente para mantener una frecuencia fundamental estable.

Al igual que el DAF, FAF también ha sido probado en sujetos tartamudos, y se ha podido ver que tanto el aumento como la disminución de la frecuencia fundamental del feedback auditivo, disminuyen la tasa de errores que comenten los sujetos de prueba [25].

2.3.4 Corrimiento de formantes en voz susurrada

Los formantes de la voz, corresponden a los máximos locales o global del espectro de esta, es decir las partes en que los armónicos alcanzan una mayor amplitud, y vienen dados por las resonancias del tracto vocal, que dan forma a los armónicos que provienen de la glotis [26]. La ubicación exacta de un formante, puede ser obtenida de 2 maneras, una opción es considerar el armónico que es más aumentado por una resonancia, la otra, es estimar la envolvente del espectro de la señal de voz, y tomar sus máximos locales. El primer método busca la posición del formante a partir del sonido generado, y el segundo, la estima analizando el sistema que genera el sonido. Los formantes vienen dados principalmente por la posición del tracto vocal, y se ha observado que los más relevantes son los 2 primeros (F1 y F2). En base a estos, es posible identificar las vocales de un idioma y generar un mapa de vocales (Ver Figuras 2.7 y 2.8) [27]. El estudio de los formantes y la producción de vocales, motivó la idea de realizar un experimento de feedback auditivo en que se haga un corrimiento de los formantes en tiempo real, lo que hace que un sujeto pruebas perciba una vocal diferente a la que intenta decir.

En el primer experimento con corrimiento de formantes, se pidió a un grupo de hablantes, que susurraran palabras que contenían una determinada vocal, en algunos casos, los hablantes escuchaban un feedback sin alteraciones, en otros, escucharon solo ruido, es decir, no tuvieron feedback, y en otras ocasiones escucharon un feedback con los formantes desplazados [29, 30]. El desplazamiento de los formantes utilizado en este estudio, se diseñó tomando en cuenta las rutas entre vocales que se pueden observar en un plano de formantes F1 y F2. Se buscó realizar transformaciones de los formantes, que generaran los corrimientos necesarios para ir entre vocales adyacentes. Se pudo demostrar que los sujetos de prueba, en mayor o menor medida, realizaron alguna acción de compensación ante la alteración del feedback auditivo. Además, al ser expuestos en repetidas ocasiones al feedback alterado, se pudo observar que al realizar el experimento con los sujetos sin feedback (usando ruido de enmascaramiento), estos mantuvieron la compensación, es decir, hubo una adaptación a largo plazo de los mecanismos que regulan la producción del habla. Ninguno de los sujetos de prueba detectó las alteraciones introducidas

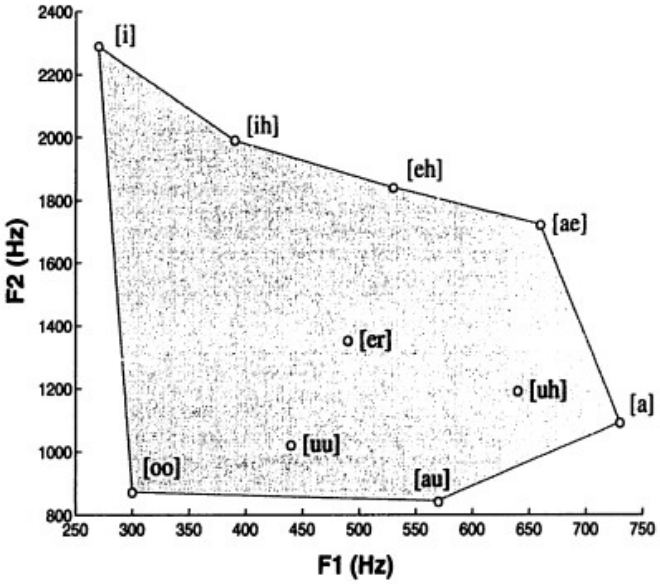


Figura 2.7: Plano de Vocales del inglés. Imagen tomada de "Control Methods used in a study of the vowels" [27]

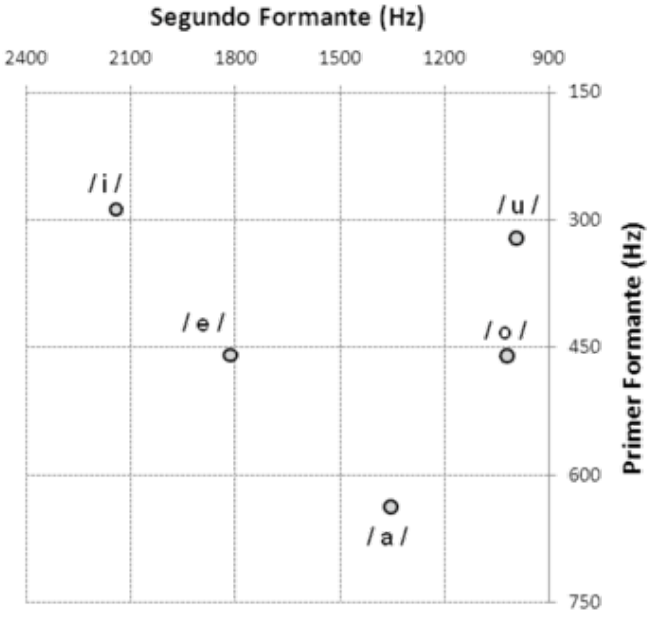


Figura 2.8: Plano de vocales del español. Generado a partir de la información entregada en "A comparative acoustic study of English and Spanish vowels"[28]

al feedback, pero a pesar de eso, todos compensaron en mayor o menor medida las alteraciones, es decir, la compensación fue involuntaria. La adquisición de los datos fue hecha a 8[kHz], debido a que una mayor tasa de muestreo hubiese requerido más tiempo de procesamiento. De acuerdo a estudios anteriores, el 4to formante en hablantes hombres, está por debajo de los 4[kHz], por lo cual, es suficiente para realizar el experimento con sujetos de prueba hombres. Para poder realizar este experimento, se implementó un algoritmo de procesamiento de señales, en una tarjeta "Ariel DSP-96". El algoritmo podría separarse 3 etapas principales, primero, una estimación de los formantes (F1, F2, F3 y F4), luego la alteración de los formantes (F1, F2 y F3), y finalmente una síntesis de la voz con los datos alterados. La señal fue dividida en frames de 8[ms] (64 muestras), los cuales se almacenan en una "ventana de análisis" de 128 muestras, la cual se multiplica por una ventana hamming, y luego se le calcula la FFT. Una vez que ingresa el siguiente frame de 64 muestras, la "ventana de análisis" se actualiza, eliminando el frame más antiguo, y manteniendo el anterior. Para estimar los formantes, el algoritmo aplica diferentes operaciones de suavizado del espectro, y finalmente ubica los peaks. Debido a la manera en que se realiza la FFT, los formantes solo pueden tomar 64 valores diferentes, entre 0 y 4[kHz], por lo cual, para realizar la alteración de los formantes de manera rápida, se utilizó una "lookup table", que se encargaba de buscar la combinación de F1, F2 y F3 obtenida, y reemplazarla por la combinación alterada definida previamente.

Finalmente, el algoritmo, se encargaba de sintetizar una voz, utilizando los formantes alterados. Para esto, primero se generaba la respuesta a impulso del tracto vocal, usando resonancias ajustadas a cada uno de los formantes. La fuente de la glotis, fue recreada utilizando una señal aleatoria. Utilizar una "lookup table" para las alteraciones de los formantes, y una señal aleatoria como flujo de la glotis, permitieron un procesamiento muy rápido de la señal, logrando que el feedback generado tuviera un retardo temporal de tan solo 16[ms] (8[ms] de muestreo y 8[ms] de procesamiento), lo cual está dentro del rango en que los hablantes no notan diferencias.

Utilizar una voz susurrada, ayudó a minimizar la percepción del feedback auditivo sin alterar debido al "bone-conduction", y también permitió realizar la síntesis de la voz, usando una señal aleatoria como fuente de la glotis. Con una fonación normal, esto no habría sido posible, ya que los hablantes hubiesen percibido que el feedback no correspondía con su fonación.

2.3.5 Corrimiento de Formantes en voz modal

El primer experimento de corrimiento de formantes, despertó gran interés de otros investigadores por realizar experimentos con condiciones similares. En 2006, se presentaron los resultados de experimentos en que se hacía un corrimiento del

primer formante en voces modales. En estos experimentos se buscó desplazar un solo formante, y se evaluó la respuesta de los sujetos de prueba. Se pudo observar nuevamente, que ante la alteración del feedback auditivo, existe una respuesta de los sujetos de prueba, intentando desplazar el formante en la dirección contraria al corrimiento. Además, se halló que, a mayor agudeza auditiva de los sujetos, mayor el tamaño de la respuesta ante la perturbación. Finalmente, se simuló el experimento utilizando DIVA, y se pudo observar que el modelo fue capaz de responder en varios aspectos, de manera similar a los sujetos de prueba, lo cual, valida el modelo, y sirvió para proponer algunos ajustes a este [31].

Para realizar estos experimentos, se utilizó una tarjeta "Texas Instruments C6701", en la cual se programó el experimento de corrimiento del primer formante. El algoritmo utilizado (Ver Figura 2.9), realiza un análisis con LPC, y calculando las raíces del polinomio obtenido, encuentra el primer formante. Con la posición de F1, el algoritmo calcula la posición del nuevo F1, en base al corrimiento que se desea realizar. Luego, se generan un "cero" encargado de eliminar el "polo" correspondiente al primer formante, y un "polo" encargado de generar un primer formante desplazado. Finalmente, se procede a filtrar la señal utilizando un filtro IIR. Este filtro se obtiene con el "cero" del primer formante y el "polo" del formante desplazado. Al igual que en el experimento anterior, la señal fue muestreada a 8[kHz], y el procesamiento tardó 128 muestras, lo que equivale a 18[ms].

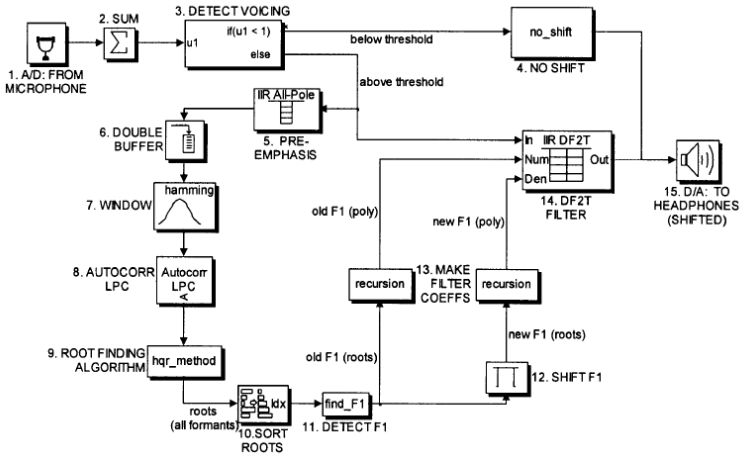


Figura 2.9: Algoritmo de Corrimiento del primer formante

Con un algoritmo similar, se introdujo un corrimiento de formantes en la mitad de una vocalización, y se pudo observar que los sujetos de prueba realizan una acción compensatoria, alrededor de 300[ms] después, de manera similar a la que ocurre ante cambios en la frecuencia fundamental [32].

2.3.6 Otros Experimentos realizados

Además de los ya mencionados, se han realizado otros experimentos en donde se perturba el feedback auditivo de diferentes maneras, se pudo ver que aumentos o disminuciones en la amplitud del feedback auditivo, generan disminuciones o aumentos en la intensidad de la voz respectivamente [34, 35]. Se experimentó también filtrando el feedback auditivo con pasa-altos y pasa-bajos, y se pudo observar que bajo ciertas condiciones los hablantes tienden a producir una menor nasalización. Esto también pudo observarse en sujetos que presentan hipernasalidad, por lo que se cree que esta alteración del feedback podría servir para tratar dicha enfermedad [33]. Otra línea de investigación que ha sido explorada es la perturbación de consonantes, se ha podido observar que los sujetos de prueba son capaces de detectar este tipo de alteraciones, pero aún no existen resultados claros acerca de cómo los hablantes responden a ellas [36].

2.4 Otras Tecnologías relativas a los experimentos de feedback auditivo

Además de las mediciones acústicas realizadas en los experimentos de feedback auditivo alterado, es posible obtener mediciones de la actividad cerebral si se combina la alteración del feedback con otras técnicas de medición. A continuación, se describen la Imagen por Resonancia Magnética Funcional (fMRI), y la Electroencefalografía (EEG), dos técnicas que han sido de gran utilidad para comprender los mecanismos de feedback auditivo.

2.4.1 fMRI

La Imagen por Resonancia Magnética (MRI), es una técnica no invasiva, que comunmente se utiliza para reverlar la estructura anatómica de alguna región del cuerpo de una persona. Para obtener este tipo de imágenes, se genera un campo magnético muy potente, el cual atraviesa el cuerpo del sujeto de pruebas. Este campo es perturbado por las propiedades magnéticas de las moléculas que componen el cuerpo de una persona. Analizando las perturbaciones que ocurren en el campo magnético aplicado, es posible generar una imagen tridimensional, estimando de qué están compuestos los tejidos que componen las estructuras anatómicas de la persona estudiada.

La Imagen por Resonancia Magnética funcional (fMRI), es un tipo de MRI, que hace uso de las propiedades de la hemoglobina, y de las diferencias magnéticas de esta molécula en sus diferentes estados de oxigenación para detectar cómo se comporta la irrigación sanguinea en el cerebro. Cuando realizamos cualquier

actividad, hay ciertas áreas del cerebro que se ven involucradas en la ejecución la tarea a realizar. Las neuronas que actúan en este proceso, requieren energía, lo cual se traduce en un mayor flujo de sangre a esa zona, y por lo tanto un flujo de hemoglobina. Al obtener estas imágenes en un periodo de tiempo, mientras la persona realiza una acción o piensa en algo, es posible obtener el estado de cada voxel (pixel volumétrico) de la imagen durante la ejecución de la acción. Con esta información, es posible estimar qué zonas del cerebro requirieron más sangre, y a partir de eso, inferir que zonas estuvieron activas durante la ejecución de la tarea [69].

Al realizar una fMRI, se busca alcanzar un equilibrio entre la resolución espacial y temporal de la medición, ya que una mayor resolución espacial requiere un mayor tiempo para obtener y procesar las muestras. Por lo general, una fMRI, se realiza con voxels de 3.4x3.4x4.0 mm^3, que son actualizados cada 60$[ms]$. En actualizar la imagen completa del cerebro, el método tarda aproximadamente 2$[s]$[70].

Desde la aparición de esta técnica, se han realizado una gran cantidad de estudios, en dónde se solicita a los sujetos de prueba que realicen una acción, mientras se mide qué zonas del cerebro se activan. Solicitando a los hablantes realizar tareas relativas al habla, ha sido posible identificar las zonas involucradas en la producción de la voz [71].

Se han realizado algunos experimentos de feedback auditivo alterado, mientras se hacen mediciones con fMRI, se mencionarán algunos a modo de ejemplo.

En uno de estos estudios, se hicieron pruebas con un feedback retardado en 200[ms], y también con un corrimiento en frecuencia de media octava hacia arriba (Ver Figura 2.10). Se pudo observar que hubo una mayor activación del giro temporal superior, y del cerebelo en presencia de perturbaciones [72].

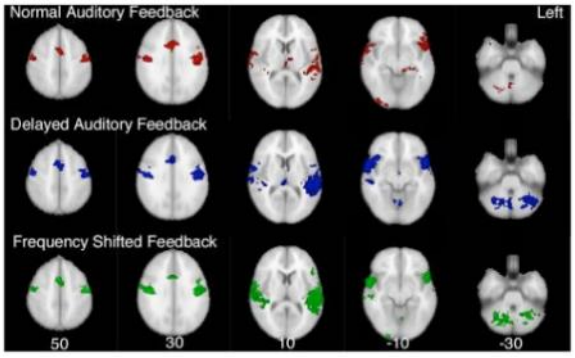


Figura 2.10: Resultados de estudio realizado con fMRI y alteraciones del feedback auditivo. Imagen de [72]

En otro estudio, se realizaron pruebas con una alteración del feedback donde el hablante escuchaba su propia voz perturbada, una voz de otra persona, y también

una voz sintetizada con un sonido no natural. Se pudo observar una mayor respuesta cuando los sujetos de prueba escucharon su propia voz perturbada, en comparación con las voces que los sujetos no identificaban como propias [73].

2.4.2 EEG

La electroencefalografía (EEG), es una técnica para medir la actividad cerebral, a través de la colocación de pequeños electrodos en el cuero cabelludo. Estos electrodos, se encargan medir las ondas electromagénicas que ocurren como consecuencia de la actividad neuronal. Analizando las señales medidas, es posible obtener información sobre lo que está ocurriendo en el cerebro de una persona o animal. Esta técnica puede utilizarse para el diagnóstico de la epilepsia, trastornos del sueño, tumores, e incluso para confirmar la muerte cerebral en pacientes en coma [74].

Debido a que el cerebro está compuesto por millones de neuronas, que están activas en todo momento, lo que detectan los electrodos, es la actividad de grupos de neuronas, que se encuentran en la corteza cerebral. Utilizando grandes cantidades de electrodos distribuidos, ha sido posible asociar estas mediciones a zonas del cerebro, pero en general la resolición espacial de esta técnica es bastante baja. En contraste con esto, las medicione eléctricas pueden realizarse con gran velocidad, porlo que esta técnica presenta una alta resolución temporal, del orden de los milisegundos.

Debido a la gran cantidad de señales eléctricas presentes en el cerebro, las mediciones obtenidas son bastante ruidosas, y cualquier movimiento de los sujetos de prueba, altera bastante las señales. Una manera de lidiar con el ruido, es trabajar las mediciones en el dominio de la frecuencia, esto ha permitido clasificar los tipos de ondas según la banda en que se encuentran. Otra manera de experimentar con EEG, es presentar un estímulo a los sujetos de prueba, y medir la reacción eléctrica del cerebro posterior al estímulo. Para mitigar los efectos del ruido, se aplica el estímulo en repetidas ocasiones, y se promedian los resultados de la encefalografía alineándolos con la ocurrencia del estímulo. La actividad cerebral que puede medirse en respuesta a un estímulo determinado, se conoce como Potencial Relacionado con Evento (ERP).

Tomando en cuenta los electrodos en dónde aparece el ERP, es posible inferir el la zona del cerebro en que ocurre la respuesta. Analizando la forma de onda del ERP, es posible determinar cuánto demora el cerebro en responder al estímulo, con qué intensidad lo hace, y asociar un tipo de respuesta, a un estímulo en específico.

Se ha podido observar que introducir alteraciones en el feedback auditivo, genera ERPs, esto ha servido para medir el tiempo que tarda el cerebro en detectar y responder a una alteración del feedback. Se presentarán algunos estudios a modo de ejemplo.

Se ha podido observar que cuando se aplican alteraciones en frecuencia (FAF), es posible medir los ERPs N1, P1 y P2 [77, 78].

Con el fin de estudiar el rol del feedback auditivo en la regulación de la frecuencia fundamental, se realizó una serie de experimentos, en que se aplicaba un corrimiento a la frecuencia fundamental durante un periodo de 200$[ms]$ en la mitad de una vocalización. Para un primer grupo, solo se aplicó el corrimiento en la mitad de la vocalización, y para el 2do grupo, se aplicó además un corrimiento inicial, que estuvo presente durante toda la vocalización. Se pudo ver que en todos los casos, al aplicar la perturbación en medio de la vocalización, los sujetos realizaron una corrección de la frecuencia fundamental en dirección contraria al corrimiento. Al analizar los resultados del EEG, se pudo ver que para el primer grupo hubo grandes ERPs en el momento en que se introdujo la perturbación. Pero inesperadamente, para los sujetos del 2do grupo que tuvieron el feedback alterado desde un principio, no se pudo observar ERPs con la misma amplitud [75].

También se ha podido ver que alteraciones en el feedback, provocan desincronizaciones en los ritmos mu [76].

2.5 Calidad Vocal

En esta sección, se indagará en el concepto de calidad vocal, los diferentes enfoques que existen en la actualidad, y finalmente, los métodos que se han desarrollado para poder medir la calidad vocal. Se analizarán las opciones existentes, las ventajas y desventajas que presenta cada una, y la factibilidad de aplicarlas en este trabajo.

La voz es una señal clave para diversas actividades humanas, y de ella puede obtenerse gran cantidad de información sobre el hablante, y lo que este desea expresar. Debido a que la voz tiene múltiples usos, su calidad es difícil de evaluar de manera objetiva, ya que una voz puede ser mejor o peor, dependiendo de la tarea a realizar, de los objetivos de quien produce la voz, y del contexto comunicativo. En la actualidad no existe un consenso sobre la definición de calidad vocal, y por lo mismo, tampoco hay acuerdo en cómo medirla [49]. Se han utilizado diferentes enfoques para evaluar la calidad vocal, que pueden dividirse en 2 grandes categorías, la evaluación perceptual, y la evaluación instrumental. Ambas han sido utilizadas en el diagnóstico, y seguimiento de tratamientos de las patologías del habla, así como también para investigaciones científicas relativas a la voz. A continuación, se presentarán estos enfoques, y se analizarán los métodos más utilizados en la actualidad.

2.5.1 Evaluación Perceptual

La calidad vocal podría pensarse como la manera en que un oyente interactúa con una señal de voz, de forma que el oyente toma ventaja de la información acústica disponible, para alcanzar algún objetivo perceptual. Los oyentes, por lo general cuando escuchan una voz, ponen su atención en diferentes aspectos de ella, dependiendo de su propósito, experiencia y contexto. Los aspectos más importantes de la voz para evaluar su calidad dependerán de la tarea a realizar, el tipo de estímulo, y del ambiente en que ocurre la comunicación. En este enfoque, la evaluación depende en parte del oyente, sus habilidades, objetivos y experiencias previas. Para realizar esta evaluación, la voz del hablante es presentada a uno o varios oyentes, de manera presencial, o por medio de grabaciones de audio. Para este tipo de evaluación, el oyente debe estar familiarizado con las características de la voz que se desean evaluar, incluso existen profesionales que se han especializado en este tipo de evaluaciones, y son capaces de distinguir sutilezas que el común de las personas pasa por alto. Una desventaja de este método es que requiere personal calificado para su realización, y sus resultados pueden tener variaciones para diferentes oyentes. Para llevar a cabo la evaluación perceptual, se han propuesto protocolos estandarizados, por medio de los cuales, los evaluadores pueden dar una puntuación a una determinada característica de la voz, y de esta manera, entregar una medida de la calidad vocal. Entre los protocolos existentes, destacan el GRBAS y al CAPE V, como los más mencionados en publicaciones recientes [50]. Un dilema que se presenta a la hora de formular estos protocolos, es elegir cuáles características son las más relevantes para medir la calidad vocal, cómo se definen estas características, y cómo las interpreta cada oyente.

- **GRBAS:** Escala de la Sociedad Japonesa de Logopedia y Foniatría [51], es una escala de 5 parámetros (GRBAS: severidad, ruido, aire, debilidad, tensión), que se evalúan con 4 categorías de severidad (de 0 a 3), siendo 0 "sin compromiso" y 3 "severo". Esta escala es fácil de aplicar, y permite una evaluación rápida, pero sus resultados no ofrecen un gran nivel de detalle.
- **CAPE V:** Escala presentada por la Asociación Americana de Patólogos de Habla y Lenguaje (ASHA), incluye 6 características a evaluar, y 2 características en blanco que pueden ser agregadas por el examinador de ser necesario. Las características se evalúan en una escala de 0 a 100, en base a una serie de tareas que debe realizar el hablante, que vienen definidas dentro de la evaluación. Los parámetros evaluados por esta escala son similares a los de GRBAS, a excepción de "debilidad", la cual es reemplazada por "tono" y "volumen"[52].

En comparación, la escala GRBAS, es más sencilla que CAPE-V, lo cual hace que su aplicación sea rápida, y fácil de realizar, incluso para evaluadores sin experiencia.

Esto hace que esta escala se haya extendido por todo el mundo. La escala CAPE-V, rescata algunos aspectos de GRBAS, bucando ser más exhaustiva y precisa en la evaluación. Para esto, entrega una escala visual, con 100 valores en vez de las 4 opciones de GRBAS. También agrega una lista de tareas que debe realizar el sujeto a evaluar durante el test, lo cual estandariza las mediciones. Finalmente, la opción de agregar parámetros a la escala CAPE-V, permite que el evaluador deje una nota con detalles específicos del paciente que podrían ser relevantes para el diagnóstico y seguimiento de las patologías.

2.5.2 Evaluación Instrumental

La forma instrumental de evaluar la voz, se basa en mediciones realizadas a través de aparatos. Las mediciones más comunes que pueden servir para este fin, son grabaciones de voz a través de micrófonos, imágenes endoscópicas de la laringe, y mediciones de flujos de aire. Para evaluar la calidad de la voz, se han propuesto diversos protocolos, entre los cuales destaca el presentado por la Sociedad Americana de Lenguaje y Audición (ASHA) [58]. Debido a que este trabajo se basa en señales acústicas y feedback auditivo, se pondrá el foco en el análisis acústico de la calidad vocal, dejando de lado otros métodos de evaluación instrumental.

El análisis acústico, se realiza con algoritmos matemáticos que extraen información de la señal de voz a través del procesamiento de señales, por esta razón, no necesitan de un especialista para realizarse, y ofrecen la posibilidad de ser replicados por cualquier persona utilizando herramientas computacionales. Este método es prácticamente instantáneo, y solo requiere de un micrófono y un computador, lo cual facilita las cosas para el paciente y también para el evaluador. A pesar de las ventajas que supone el análisis acústico, se presenta nuevamente la interrogante de cómo elegir las características más relevantes de la voz a la hora de medir su calidad, y cómo se logra calcular indicadores que estén asociados a esas características. Se han propuesto diversas medidas acústicas de calidad vocal, y aunque no ha sido posible encontrar una relación directa entre la calidad de la voz y una medida en particular, si se ha podido observar relaciones entre estos parámetros, y algunas patologías. Aún no existe un consenso sobre qué parámetros son los más adecuados para llevar a cabo un análisis acústico, y la comunidad científica continúa proponiendo nuevos enfoques, y probando parámetros o variaciones de los ya existentes. A continuación, se presentan algunos de estos parámetros, y los resultados que se han obtenido en estudios de patologías de la voz.

- f_0 : Es la frecuencia fundamental de la señal de voz, y se define como la cantidad de veces que una onda generada por las cuerdas vocales se repite en un segundo. Está asociada al número de ciclos de apertura/cierre de la glotis. La voz de una persona, por lo general se encuentra alrededor de una

determinada frecuencia, la cual guarda relación con la edad, el género, y también se ve afectada por el estado de ánimo de la persona, el momento del día, y el uso que se esté dando a la voz. En hombres, por lo general f_0 se encuentra entre los 50 y los 250 [Hz], para las mujeres, entre 120 y 500 [Hz][67]. En personas sanas, la frecuencia fundamental no tiene grandes variaciones, pero existen patologías que hacen que f_0 sea inestable, esto se ha podido observar para el Parkinson [66], y también para disfunciones del cerebelo [68]. Para poder medir las variaciones de f_0, se ha propuesto el parámetro jitter.

- **Jitter:** está definido como las variaciones que ocurren a f_0 de un ciclo a otro, una voz normal siempre tendrá variaciones entre ciclos, pero dentro de un rango acotado. El jitter se ve afectado principalmente por la falta de control de la vibración de las cuerdas vocales, y en muchos casos, al existir una patología de la voz, existe un mayor porcentaje de jitter. La mayoría de los investigadores consideran que un jitter normal está entre un 0.5 y un 1 % para una fonación sostenida en adultos jóvenes [54].

- **Shimmer:** el shimmer es una medida de la irregularidad porcentual de los ciclos de la glotis, y se calcula comparando la diferencia de amplitud entre ciclos adyacentes, para personas sanas, el nivel de shimmer debería estar por debajo del 3 % en personas sanas [55].

- **HNR:** La relación entre los armónicos y el ruido (Harmonics to Noise Ratio), es la evaluación de la razón entre las componentes periódicas y no periódicas de un segmento de voz [56]. Un mayor HNR implica que el habla se genera de manera efectiva, aprovechando el aire que viene de los pulmones para transformarlo en vibración de los pliegues vocales, un HNR bajo denota astenia y disfonía [54]. Se ha podido observar que las personas mayores presentan un HNR más bajo que personas jóvenes [57].

- **CPP:** La prominencia del peak cepstral (Cepstral Peak Prominence), es una medida de calidad vocal que a pesar de que no existe una definición exacta de qué es lo que mide, ha demostrado ser una medida de gran utilidad para la evaluación clínica de la voz [59, 61]. Ha sido utilizado para diversas investigaciones, obteniendo resultados satisfactorios en la medición de la disfonía [60], también se ha podido observar que pacientes con parálisis de cuerda vocal unilateral (UVFP) presentan un CPP menor a lo habitual [64]. Resultados similares fueron obtenidos para el caso de nódulos en los pliegues vocales [65]. Una de las ventajas que presenta CPP en comparación con otras medidas de calidad vocal, es que no depende de la intensidad de la señal, es decir, no se ve afectada por como hable la persona, ni por la cercanía que tenga con el micrófono. Para calcular el CPP, se debe realizar una regresión lineal del "cepstrum" de la señal de voz, y tomar la distancia desde el máximo

peak cepstral, hasta el valor de la recta de la regresión justo bajo el peak. Este peak, corresponde al periodo fundamental de la señal de voz. Señales de voz más periódicas, presentarán un mayor CPP, que voces irregulares.

- **Decaimiento Espectral:** El decaimiento espectral, busca medir la relación que existe entre los armónicos de una señal, una mayor o menor pendiente, implica que la energía de la señal, está más concentrada en los armónicos de frecuencias más altas o bajas respectivamente. Se ha podido observar, que cuando se pronuncian sílabas acentuadas, existe un mayor esfuerzo de la voz, y esto hace que aumente la amplitud de los armónicos en altas frecuencias, y por ende aumente el decaimiento espectral [62]. Algunos autores, han encontrado una relación entre un mayor decaimiento espectral y una voz respirada [63].

  A pesar de las relaciones encontradas entre algunos de estos parámetros, y patologías de la voz, no es sencillo establecer un método de diagnóstico basado solo en análisis acústico, ya que cada patología altera varios parámetros, y los resultados obtenidos pueden variar bastante entre hablantes. Por esta razón es que se han intentado establecer métodos que utilicen diferentes parámetros en su evaluación.

2.6 Un nuevo Experimento

Los experimentos de feedback auditivo son una manera no invasiva de estudiar los mecanismos neuronales que están detrás de la producción de la voz, y han aportado información valiosa para comprender y tratar algunos trastornos del habla. Con los avances tecnológicos, principalmente en las áreas de la electrónica, y procesamiento de señales, ha sido posible implementar experimentos específicos, que ponen a prueba los mecanismos neuronales, con el fin de comprender las capacidades del feedback auditivo, y la función que este cumple en la producción y control de la voz. El análisis acústico nos entrega el resultado de las correcciones que intenta realizar nuestro cerebro, y nos da una idea de cómo este opera para detectar y corregir las perturbaciones. Con ayuda del EEG, es posible analizar la respuesta cerebral en una escala temporal, y así ver cuánto demora una persona en contestar a las perturbaciones y con qué intensidad lo hace. Gracias al fMRI, ha sido posible contar con una apreciación espacial de la respuesta cerebral, observando que zonas específicas del cerebro se activan durante los experimentos.

En base a todos los experimentos estudiados, y a los avances tecnológicos disponibles, este trabajo busca proponer una metodología de procesamiento de señales, capaz de alterar la calidad vocal de una persona, es decir, hacer creer al sujeto de prueba, que algo está funcionando de manera anormal en su voz a nivel de los pliegues vocales. Esto con el fin de ver cómo reacciona ante esta perturbación, y

qué tipo de acciones correctivas aplica. Se cree que un experimento de este tipo, generaría una respuesta neuronal que podría medirse con alguna de las técnicas mencionadas, y también una corrección en la producción de la voz que se reflejaría en un análisis acústico de la señal.

Para esto, es necesario comprender cómo ocurre la producción de la voz, cómo se modela matemáticamente este proceso, y qué algoritmos de procesamiento de señales podrían ser útiles para alterar su calidad vocal.

Capítulo 3

Metodología

En este trabajo se propone realizar un experimento de feedback auditivo, en el que se altere la calidad vocal de los sujetos de prueba en tiempo real. Como primer acercamiento, se ha diseñado un algoritmo de procesamiento de señales, capaz de introducir este tipo de perturbaciones. Este algoritmo ha sido implementado en MATLAB, y testeado utilizando señales de voz grabadas. Adicionalmente, se plantea un método cuantitativo para evaluar la calidad vocal de las voces, antes y después de introducir las perturbaciones. Se propone un esquema de procesamiento de señales (Ver Figura 3.1), el cual consta de 3 etapas principales, la descomposición de la señal, la introducción de las perturbaciones, y la re-síntesis de la señal perturbada. Estas etapas serán explicadas en detalle a continuación:

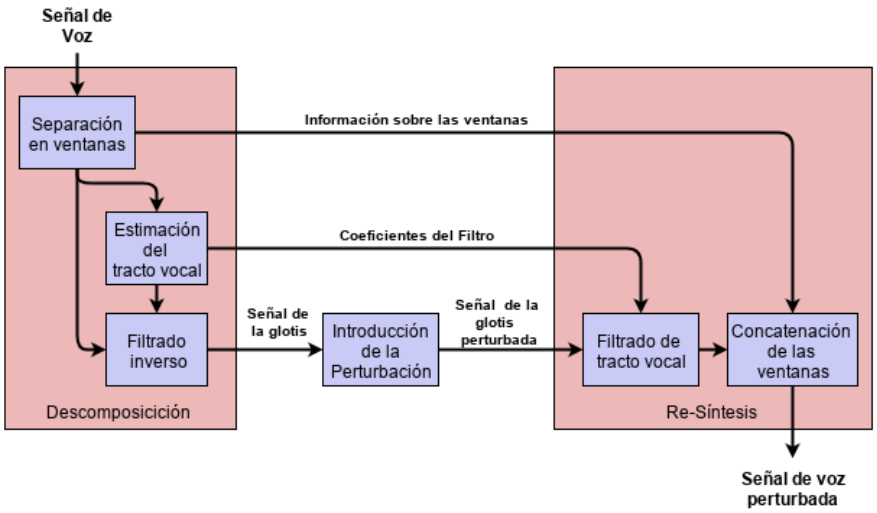


Figura 3.1: Diagrama general del Algoritmo para la introducción de las perturbaciones

3.1 Descomposición de la Señal

En esta sección, se presenta el concepto de "flujo de la glotis", se estudia la forma en que este se produce, y algunos métodos matemáticos que se han planteado para su estimación.

3.1.1 Flujo de la glotis

Cuando vocalizamos, los pliegues vocales, a través de sus vibraciones, dan forma al flujo de aire que viene desde los pulmones, la señal obtenida se conoce como flujo de la glotis. El flujo de la glotis da sus características principales a la voz de una persona.

Luego de pasar por la glotis, el aire debe atravesar la cavidad bucal, que da forma al flujo de la glotis, este proceso, se conoce como articulación. Finalmente, el flujo de aire sale irradiado a través de los labios, generando así la voz humana. Los cambios de volumen de aire producidos por la glotis y el tracto vocal son los que constituyen la voz de una persona, se propagan por el aire, y son los que podemos escuchar o grabar utilizando un micrófono. El flujo de la glotis, puede entregarnos información útil para diversas aplicaciones como reconocimiento de hablantes, transformación de la voz, diagnósticos médicos, entre otros, lo que la hace una señal de gran interés, pero, por el lugar en donde se encuentra la glotis, este flujo es difícil de medir. Existen aparatos que permiten obtener el flujo de la glotis realizando mediciones directamente en la laringe [39], pero son invasivos, requieren equipamiento avanzado, y los mismos aparatos de medición impiden que la persona hable de manera natural, lo cual hace que la señal medida no sea exactamente la que habría sin los aparatos de medición, por esta razón, no existen mediciones exactas de esta señal. Se han desarrollado gran cantidad de métodos matemáticos, que buscan estimar el flujo de glotis, a partir de mediciones de audio, lo cual presenta grandes ventajas, al ser métodos no invasivos, que pueden aplicarse sin la necesidad de contar con equipamiento especializado ni personal calificado. A continuación, se explicará el modelo Fuente-Filtro, y su aplicación a la estimación de la señal de la glotis a partir de una señal de audio, luego se procederá a analizar 3 métodos de estimación de la señal de la glotis, y se presentarán los resultados obtenidos para diferentes voces.

3.1.2 Modelo Fuente-Filtro

Este trabajo se basa en el modelo Fuente-Filtro (Source-Filter Model of Speech Production), el cual asume que la fonación y la articulación son procesos independientes, y nos permite analizarlos por separado. Este modelo, busca describir matemáticamente los procesos de fonación y articulación. Para esto, define la

fuente de la glotis, como el volumen de aire que se obtiene luego de pasar por los pliegues vocales. El filtro, corresponde a la manera en que el tracto vocal actúa sobre el flujo de la glotis. Mientras hablamos, estamos constantemente cambiando nuestra fonación (fuente) y nuestra articulación (filtro), pero para ventanas cortas, este proceso puede asumirse como un sistema lineal e invariante en el tiempo, en donde la voz producida corresponde a la convolución entre la fuente de la glotis, y el filtro del tracto vocal (Ver Figura 3.2). Cabe mencionar que este modelo no considera la interacción que pueda existir entre la glotis y la cavidad bucal.

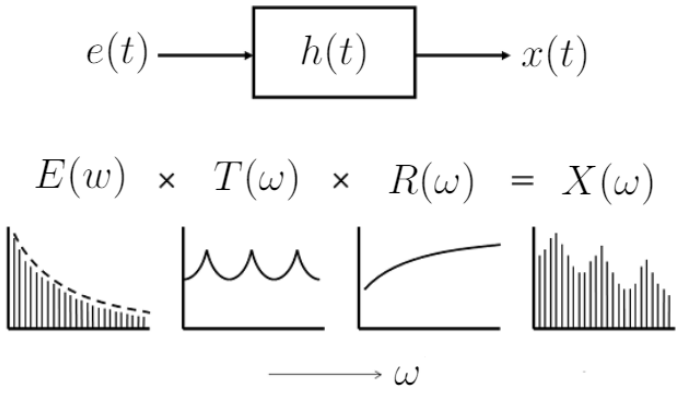


Figura 3.2: Modelo Fuente-Filtro

Consideremos que $x(t)$ corresponde a la voz de una persona, la cual es resultado de la convolución entre la señal que proviene de la glotis $e(t)$ y el filtro de tracto vocal $h(t)$. Esto también puede verse como una multiplicación en el dominio de la frecuencia, obteniéndose:

$$X(\omega) = E(\omega)H(\omega)$$

El filtro puede dividirse en 2 partes, una correspondiente al tracto vocal $T(\omega)$, y la otra correspondiente a la radiación de los labios $R(\omega)$.

$$H(\omega) = T(\omega)R(\omega)$$

A partir de este modelo, es posible estimar $E(\omega)$ como:

$$E(\omega) = \frac{X(\omega)}{H(\omega)}$$

Se han propuesto algoritmos para estimar $e(t)$ a partir de una medición de audio de $x(t)$, la cual puede ser realizada con un micrófono de manera sencilla y no invasiva. Los métodos propuestos pueden separarse en 2 grandes grupos: Métodos de Filtrado Inverso (GIF), y métodos de Descomposición de Fase Mixta [38]. En

este trabajo, se analizarán algunos métodos de Filtrado Inverso, los cuales, realizan una estimación de $h(t)$, y en base a ella, una deconvolución con $x(t)$, para obtener $e(t)$.

3.1.3 Métodos de Filtrado Inverso de la Glotis (GIF)

Todos los métodos de filtrado inverso de la glotis se basan en el mismo esquema de funcionamiento (Ver Figura 3.3). Primero toman ventanas de la señal de entrada, estiman la contribución del tracto vocal, realizan un filtrado inverso, y finalmente concatenan las ventanas. La principal diferencia entre los diferentes métodos que existen, es la manera en que estiman el tracto vocal. Otra diferencia a tomar en cuenta, es la manera en que los métodos realizan la descomposición en ventanas. Si las ventanas tienen un largo en función del largo del periodo de la glotis, estamos en presencia de un método sincrónico, en caso de que el método estime el tracto vocal sin tener en consideración la duración de los ciclos, se hablará de que el método es asincrónico. Los métodos sincrónicos han mostrado tener un mejor resultado, pero implican un mayor procesamiento de la señal para realizar la sincronización.

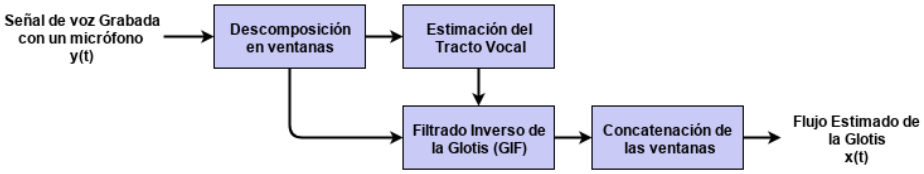


Figura 3.3: Diagrama general de algoritmos de filtrado inverso

3.1.3.1 Flujo esperado de la glotis

Para poder evaluar la estimación realizada por los métodos que se presentarán a continuación, es fundamental conocer la forma de onda que se espera obtener. Como no es posible conocer el flujo real a través de la glotis, es difícil hacer una evaluación cuantitativa del funcionamiento de los métodos de estimación. Para probar el funcionamiento de estos algoritmos, se ha intentado utilizar voces generadas con modelos computacionales, donde se conoce de antemano el flujo de la glotis, y se compara con el obtenido utilizando el algoritmo, pero en muchos casos, el sintetizador de voz también funciona basado en el modelo fuente-filtro, lo cual genera buenos resultados, pero no garantiza que eso ocurra para voces reales. Es por esto que se evaluarán los métodos de manera cualitativa, tomando como referencia el modelo de Liljencrants-Fant (LF) [40], y considerando que más que obtener la señal exacta del flujo, se espera poder obtener una señal que pueda ser perturbada, y luego utilizada para re sintetizar la señal de voz. De

acuerdo al modelo LF, la forma de onda esperada para un ciclo, consta de una fase de apertura, en donde el flujo aumenta, hasta llegar a un máximo, para luego comenzar a cerrarse, llegando a un estado de cierre (Ver Figura 3.4). Este modelo matemático, permite ajustar una curva con la forma de la señal de la glotis utilizando 5 parámetros, pero se debe tener en cuenta, que el flujo de la glotis es único para cada persona, por lo cual este modelo solo sirve como aproximación general.

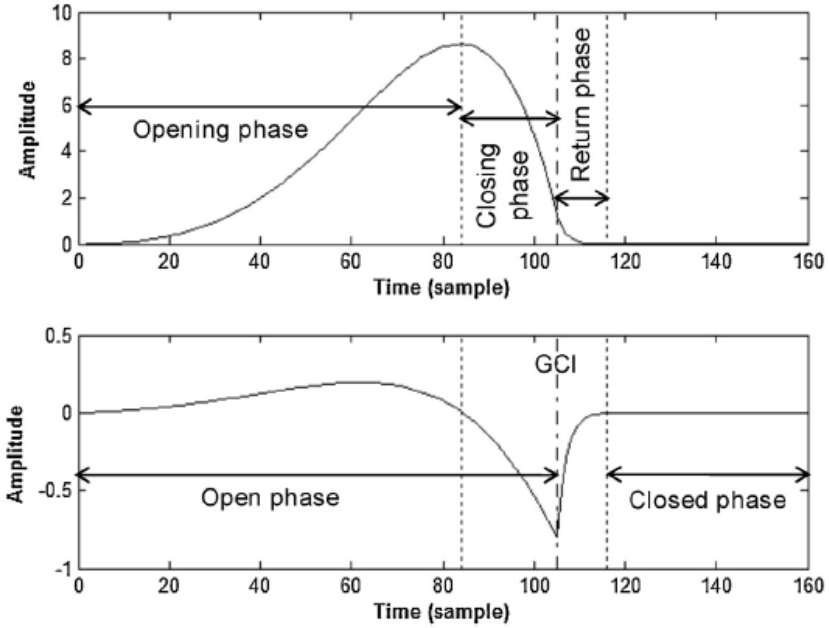


Figura 3.4: Forma de onda para la señal de la glotis: En la parte de arriba de la figura, puede verse el flujo de la glotis, y abajo, la derivada de este flujo. Notar también que se muestran las fases de apertura, y clausura, así como también el cierre de la glotis (GCI).

3.1.3.2 LPC

Para utilizar el modelo Fuente-Filtro en la descomposicion de señales de voz, es necesario estimar un filtro que sea capaz de modelar el tracto vocal. En este trabajo, se han implementado 3 métodos para esta tarea, los cuales se basan en LPC (Codificación Predictiva Lineal).

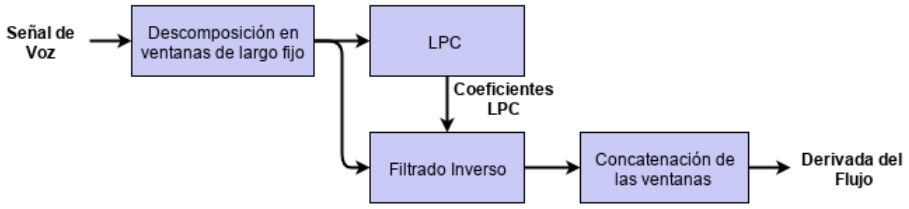


Figura 3.5: Diagrama de Bloques del sistema LPC.

LPC es una herramienta muy utilizada en el procesamiento de audio, y específicamente en el procesamiento de señales de voz. A continuación, se procederá a desarrollar matemáticamente los fundamentos de este método [43, 44].

Si consideramos una señal discreta de voz $x[n]$, utilizando el modelo fuente-filtro (Ver Figura 3.2), podemos decir que dicha señal corresponde a la convolución de la señal de la glotis $e[n]$ con un filtro del tracto vocal $h[n]$.

$$x[n] = h[n] * e[n] \tag{3.1}$$

como ya contamos con $x[n]$, es necesario encontrar alguna manera de estimar el filtro $h[n]$, para esto, supondremos que $h[n]$ es un filtro que solo contiene polos. Esto significa que el valor actual de $x[n]$, dependerá del valor actual de $e[n]$, y de algunas muestras anteriores $x[n-1], x[n-2]... \; x[n-p]$.

$$x[n] = \sum_{k=1}^{p} a_k x[n-k] + e[n] \tag{3.2}$$

donde "p" corresponde a la cantidad de muestras anteriores que se tomarán en cuenta, conocido como el orden del LPC. Aplicando la transformada Z a (3.2), se tiene lo siguiente:

$$X(z) = \sum_{k=1}^{p} a_k z^{-k} X(z) + E(z) \tag{3.3}$$

Retomando la idea del Modelo Fuente-Filtro, podemos observar la ecuación (3.3) como una función de transferencia:

$$\frac{X(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = H(z) \tag{3.4}$$

donde $H(z)$ corresponde al filtro que modela el tracto vocal. De esta manera, el problema se reduce a encontrar un grupo de coeficientes a_k con $k = 1, 2,...p$. Esto

se conoce como autoregresión, ya que se busca estimar los valores futuros de una señal, tomando en cuenta un conjunto de valores pasados.

Si contamos con una señal con N muestras, donde $N >> p$, es posible generar un conjunto de N ecuaciones para a_k, es decir a_k queda sobredeterminado.

Si definimos las matrices $\hat{x_i}$ y a como:

$$\hat{x_i} = \begin{bmatrix} x(n-1+i) & x(n-2+i) & ... & x(n-p+i) \end{bmatrix}$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ ... \\ a_p \end{bmatrix}$$

se pueden plantear las siguientes N ecuaciones:

$$\begin{aligned} x[n] &= \hat{x_0} \cdot a + e[n] \\ x[n+1] &= \hat{x_i} \cdot a + e[n+1] \\ &... \\ x[n+N] &= \hat{x_N} \cdot a + e[n+N] \end{aligned}$$

agrupando términos, es posible definir nuevas matrices:

$$b = \begin{bmatrix} x[n] \\ x[n+1] \\ ... \\ x[n+N] \end{bmatrix} \quad e = \begin{bmatrix} e[n] \\ e[n+1] \\ ... \\ e[n+N] \end{bmatrix} \quad A = \begin{bmatrix} \hat{x_0} \\ \hat{x_1} \\ ... \\ \hat{x_N} \end{bmatrix}$$

y con ellas, escribir las N ecuaciones en forma matricial:

$$e = b - A \cdot a \tag{3.5}$$

La ecuación (3.5) tiene la forma de una regresión lineal, donde $A \cdot a$ corresponde a una estimación de b, y e corresponde al error o residuo de dicha estimación. Debido a que el problema está sobredeterminado, no será posible hayar una solución única, pero sí es posible encontrar una solución que minimice el error. Existen diferentes maneras de minimizar el error, una solución típica para este problema, es utilizar mínimos cuadrados, y a través de ellos, buscar el conjunto de valores a_k que minimicen E.

$$E = \sum_{n=1}^{N} e^2[n] \tag{3.6}$$

Ahora que contamos con la función que se debe minimizar, es posible calcular las derivadas parciales de los a_k, e igualarlas a 0 para generar un sistema de p ecuaciones. Resolviendo este sistema, es posible minimizar el error.

$$\frac{\partial E}{\partial a_k} = 0 \quad k \in 1, 2, ... \ p \tag{3.7}$$

Teniendo los coeficientes a_k, es posible calcular $e[n]$ realizando un filtrado inverso. Elegir el orden a utilizar para una aproximación con LPC es un problema abierto, que dependerá de diversos factores, entre ellos la naturaleza de la señal analizada, y la frecuencia con que fue muestreada. Para la voz humana, se sabe que es posible reconocer una vocal a partir de sus 3 primeros formantes, y para que estos se vean reflejados en la aproximación con LPC, este debe ser como mínimo de orden 6.

3.1.3.3 Método de la Autocorrelación:

Cuando hablamos, nuestro tracto vocal varía constantemente para generar diferentes sonidos, por lo cual, no tiene sentido estimar el tracto vocal para periodos de tiempo largos. Considerando que la voz humana tiene una frecuencia fundamental en el rango $[85 - 300][Hz]$, y que para poder hacer un análisis LPC es necesario tomar como mínimo un ciclo completo de la glotis, es posible tomar ventanas de $10 - 30[ms]$, y aplicar el método a ellas, para luego concatenar los resultados obtenidos de las diferentes ventanas (Ver Figura 3.5).

Para esto se aplica el método a segmentos de largo N, y se asume que $x_n[m]$ es 0 fuera del intervalo $0 \leq m \leq N - 1$, lo que puede expresarse como:

$$x_n[m] = x[m + n]w[m] \tag{3.8}$$

donde $w[m]$ corresponde a algún tipo de ventana, cuyos valores fuera del intervalo $0 \leq m \leq N - 1$ son iguales a 0 (se suele utilizar alguna que disminuya los valores de los extremos, típicamente ventana Hamming).

De esta manera, si $x_n[m]$ es distinto a 0 solo para $0 \leq m \leq N - 1$, entonces el error $e_n[m]$, de un predictor de orden p será distinto a 0, en el intervalo $0 \leq m \leq N-1+p$, y E_n puede expresarse como:

$$E_n = \sum_{m=0}^{N+p-1} e_n^2[m] \tag{3.9}$$

Tomando en cuenta que $x_n[m] = 0$ fuera del intervalo $0 \leq m \leq N-1$, es posible calcular los $\phi[i,k]$:

$$\phi[i,k] = \sum_{m=0}^{N+p-1} x_n[m]x_n[m+i-k] \qquad 1 \leq i \leq p \quad 0 \leq k \leq p \tag{3.10}$$

se puede ver que $\phi[i,k]$ está en función de $i-k$, pudiendo escribirse de la siguiente forma

$$\phi[i,k] = R_n[i-k] \qquad 1 \leq i \leq p \quad 0 \leq k \leq p \tag{3.11}$$

donde $R_n[i-k]$ corresponde a la autocorrelación de tiempo reducido de $x_n[m]$ evaluada en $i-k$, donde

$$R_n[i-k] = \sum_{m=0}^{N-1-k} x_n[m]x_n[m+k] \tag{3.12}$$

como $R_n[k]$ es par, entonces $\phi_n[i,k] = R_n[|i-k|]$ para $1 \leq i \leq p\, 0 \leq k \leq p$, y de esta forma, las ecuaciones quedan expresadas como:

$$\sum_{k=1}^{p} a_k \phi_n[i-k] = \phi_n[i,0] \qquad 1 \leq i \leq p \tag{3.13}$$

$$\sum_{k=1}^{p} a_k R_n[|i-k|] = R_n[i] \qquad 1 \leq i \leq p \tag{3.14}$$

con lo que el mínimo error cuadrático quedaría como:

$$E_n = \phi_n[0,0] - \sum_{k=1}^{p} a_k \phi_n[0,k] = R_n[0] - \sum_{k=1}^{p} a_k R_n[k] \tag{3.15}$$

de forma matricial se obtiene:

$$\begin{bmatrix} R_n[0] & R_n[1] & ... & R_n[p-1] \\ R_n[1] & R_n[0] & ... & R_n[p-2] \\ ... & ... & ... & ... \\ R_n[p-1] & R_n[p-2] & ... & R_n[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ ... \\ a_p \end{bmatrix} = \begin{bmatrix} R_n[1] \\ R_n[2] \\ ... \\ R_n[p] \end{bmatrix}$$

Finalmente el problema se reduce a encontrar la inversa de la matriz de los R_n, la cual es una matriz Toeplitz (simétrica y con todos los elementos de la diagonal iguales), para la cual existen algoritmos más eficientes que una simple inversión de la matriz. El método más utilizado para el cálculo eficiente de los coeficientes a_k es el algoritmo de Levinson-Durbin[41], cuyo costo computacional es de $\theta(n^2)$.

3.1.3.4 Sincronización

Los algoritmos PSIAIF y QCP son métodos sincrónicos, por lo cual, se procederá a explicar el algoritmo de sincronización utilizado para la implementación de estos. Cuando hablamos de sincronismo en señales periódicas, por lo general nos basta con identificar la frecuencia fundamental de la señal de interés para la sincronización. En el caso del flujo de la glotis, a pesar de ser una señal con una cierta periodicidad, puede variar mucho entre ciclos, tanto en forma de onda como en duración, dependiendo del hablante, el tipo de fonación, patologías que pueda presentar. Por esta razón, para realizar la sincronización, se identifican ciertos instantes del ciclo de la glotis en la señal. Con estos instantes es posible seleccionar ventanas que contengan un número entero de ciclos. Si bien los métodos sincrónicos han presentado mejores resultados que los asincrónicos, el resultado final depende en gran medida de la calidad de la sincronización. Errores de sincronización podrían significar desde pequeñas fallas, hasta resultados completamente erróneos en la estimación de la señal de la glotis.

3.1.3.5 Instantes de apertura y cierre de la glotis

Para realizar la sincronización con la glotis, lo que se hace el método utilizado es buscar 2 instantes específicos en los ciclos de esta: Los instantes de apertura y de cierre. En el instante de máxima apertura (GOI), es donde ocurre el mayor flujo de aire a través de la glotis, y justo luego de este, la glotis comienza a cerrarse. El instante de cierre (GCI), es el punto en que los pliegues vocales se juntan, lo cual puede observarse como un mínimo en la derivada de la señal de la glotis.

Para poder detectar los instantes de apertura y cierre, se ha utilizado el método que se muestra a continuación (Ver Figura 3.6) [45]:

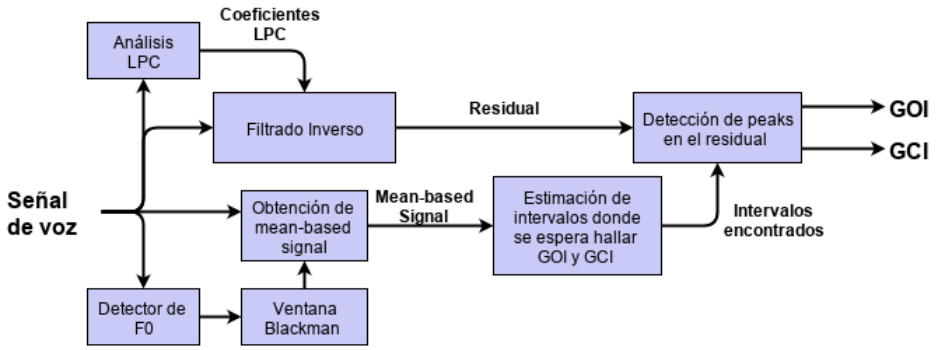


Figura 3.6: Diagrama de Bloques método de detección de instantes de apertura y cierre de la glotis.

El método obtiene la "mean-based signal" y a partir de ella, realiza una estimación de los tramos dónde podrían encontrarse los instantes de apertura (GOI) y de cierre (GCI) de la glotis. Luego, realiza un filtrado inverso utilizando LPC, y utilizando los tramos seleccionados y el residual del LPC es capaz de seleccionar los puntos donde ocurren los instantes GCI y GOI. Algo fundamental para este método es la elección del largo de la ventana blackman, los autores recomiendan que este sea 1.75 veces T_0, por lo cual es importante para este método contar con una buena estimación de f_0.

3.1.3.6 Algoritmo PSIAIF (Pitch Synchronous Iterative Adaptive Inverse Filtering)

PSIAIF[47] se basa en el método IAIF, el cual, a su vez, se basa en AIF. De manera general, PSIAIF consta de 2 partes principales, una primera etapa de sincronización, la cual divide la señal en segmentos de un ciclo de duración, que van desde el instante de máxima apertura de la glotis al próximo instante de máxima apertura. La segunda etapa, consiste en la aplicación del método IAIF, a cada segmento por separado, para luego concatenar los segmentos filtrados (Ver Figura 3.7).

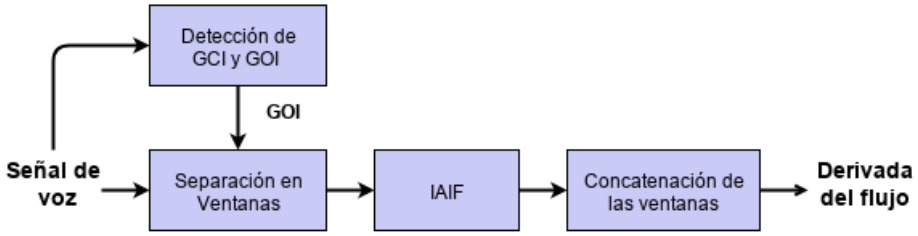


Figura 3.7: Diagrama de Bloques método PSIAIF.

IAIF: El método IAIF [37] busca estimar la función de transferencia del tracto vocal utilizando LPC. Es iterativo adaptativo, porque realiza una primera estimación de baja precisión, y la utiliza para obtener una segunda estimación más precisa. El funcionamiento de este puede explicarse a través de los siguientes pasos (Ver Figura 3.8).

1. Estimar a grandes rasgos la envolvente de la señal de la glotis utilizando un LPC de primer orden (Bloque 1).

2. Quitar la contribución estimada de la glotis, utilizando filtrado inverso (Bloque 2).

3. Estimar a grandes rasgos la envolvente del tracto vocal utilizando LPC (Bloque 3).
4. Eliminar la contribución estimada del tracto vocal (Bloque 4).
5. Estimar de manera más precisa la envolvente de la señal de la glotis (Bloque 6).
6. Eliminar la contribución aproximada de la señal de la glotis (Bloque 7).
7. Estimar de manera precisa el envolvente del tracto vocal (Bloque 8).
8. Realizar un filtrado inverso para obtener la señal de la glotis (Bloque 9).

Luego de realizar un filtrado inverso, se obtiene la derivada del flujo de la glotis, los bloques de integración (Bloques 5 y 10), se ocupan para obtener el flujo de la glotis a partir de su derivada. Este documento se enfocará en obtener la derivada del flujo, por lo cual, el bloque 10, no ha sido utilizado en la implementación de este algoritmo.

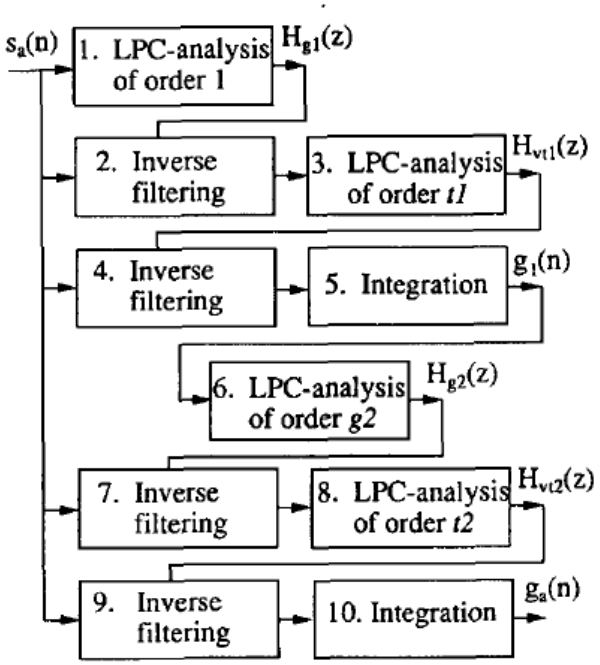


Figura 3.8: Diagrama IAIF. Imagen de [47].

3.1.3.7 Algoritmo QCP

Al igual que el algoritmo PSIAIF, el método QCP [48] también corresponde a un método sincrónico, por lo cual, consta de una etapa de sincronización, luego de

esta, estima un modelo para el tracto vocal utilizando Predicción Lineal Ponderada (WLP), y luego realiza el filtrado inverso. Los métodos "closed phase", se basan en la idea de que cuando la glotis se encuentra cerrada, el acoplamiento de la señal de la glotis con la señal de voz es mínimo. De esta manera es posible estimar un modelo del tracto vocal que no esté tan influido por la glotis. El método QCP [42], utiliza WLP, con lo que da mayor importancia a las muestras que se encuentran en la parte en que la glotis se encuentra cerrada. Para esto define una función de pesos (AME), la cual se aplica a la hora de minimizar el error cuadrático en la aplicación del LPC, esto hace que las muestras que se encuentran en la fase cerrada de la glotis tengan una mayor ponderación a la hora de estimar el tracto vocal (Ver Figura 3.9). La estimación del tracto vocal se realiza tomando varios periodos de la señal de voz, lo cual hace que la estimación del tracto vocal sea más robusta a errores en la estimación de los instantes de apertura y cierre de la glotis.

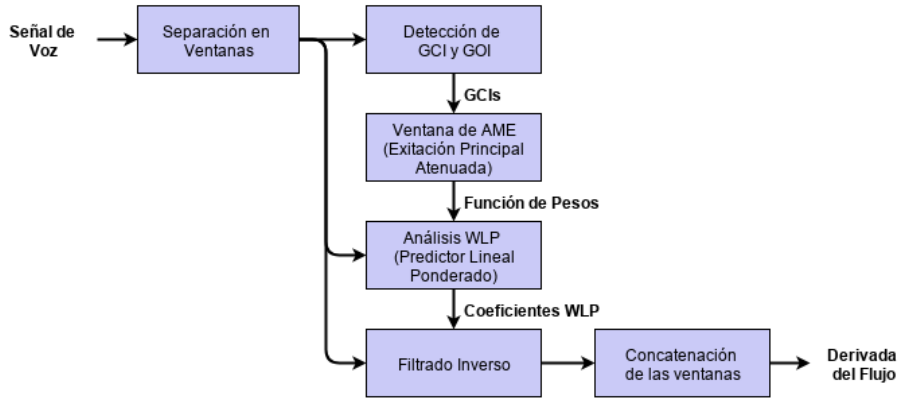


Figura 3.9: Diagrama de bloques del método QCP.

3.1.3.8 WLP

La Predicción Lineal Ponderada, es una variación del clásico método LP, que intenta estimar el valor de una señal $x[n]$ basándose en sus valores anteriores (Ecuación 3.2). La diferencia con LPC, ocurre en el momento de calcular el error de la predicción, ya que se impone una ponderación temporal W_n al cuadrado del residuo.

$$E = \sum_{n=n_1}^{n_2} \left(x[n] - \sum_{i=1}^{p} a_k x[n-i] \right)^2 W_n \tag{3.16}$$

donde $n_1 = 1$, $n_2 = N + p$ (considerando que se calcula con el método de la autocorrelación), y N corresponde al largo del cuadro que se está procesando. Al igual que en LPC, el problema consiste en encontrar los parámetros a_k que

minimicen E (3.16). De manera similar a LPC, esto puede hacerse obteniendo las derivadas de E e igualando a 0.

$$\frac{\partial E}{\partial a_k} = 0 \quad k \in 1, 2, ... \ p$$

Luego de esto, es posible plantear algunas matrices, y hacer un desarrollo muy similar al de LPC para obtener los a_k. Los inconvenientes que presenta este método, son que la matriz resultante no es de tipo Toeplitz, lo cual hace que no sea posible aplicar el Algoritmo de Levinson-Durbin, y tampoco es posible garantizar que el filtro obtenido sea estable.

3.1.3.9 Función de Pesos W_n

Para el método QCP es esencial elegir correctamente la función de pesos W_n. La idea de esta, es disminuir la contribución de las muestras que se encuentran en el cierre de la glotis (GCI). Para esto, se genera una función llamada "Exitación Principal Atenuada" (AME en inglés). La cual para cada ciclo de la glotis, tiene un valor 1 en todo momento a excepción de la vecindad del cierre de la glotis, donde W_n se atenúa, llegando a ser una constante d cercana a 0 ($d = 10^-5$ en este trabajo). Los instantes en que comienza y termina la atenuación, y la pendiente de la rampa, están definidos como parámetros del algoritmo, y es posible ajustarlos en base a la voz que se esté procesando (Ver Figura 3.10).

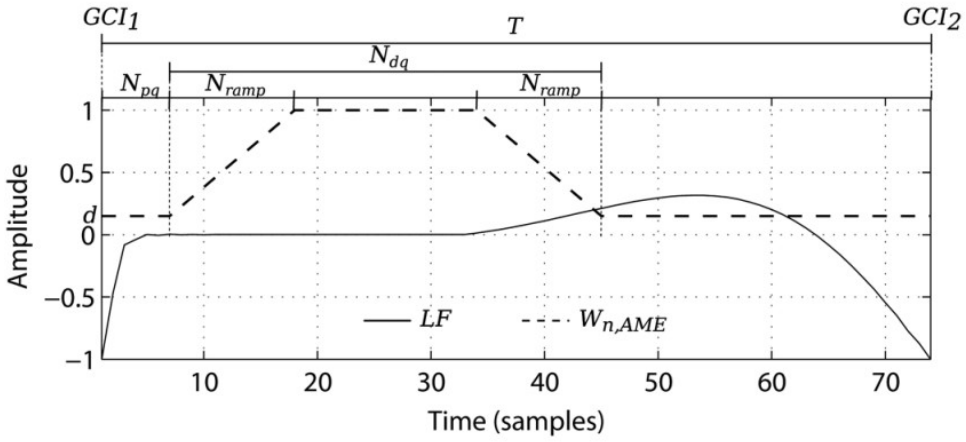


Figura 3.10: Derivada del flujo de la glotis en base al modelo LF, con su respectiva función de pesos AME. Figura extraida de [48].

3.2 Perturbaciones

Introducir perturbaciones a la fuente de la glotis, no es una tarea sencilla, ya que estas deben cumplir con varios requerimientos para poder ser implementadas en un experimento de feedback auditivo y se debe buscar que el sonido logrado sea natural.

En este trabajo se exploran algunas perturbaciones sencillas que podrían aplicarse en un experimento de este tipo. Cada perturbación tiene parámetros que permiten ajustar diferentes aspectos de esta. Las perturbaciones también podrían ser aplicadas en serie para conseguir perturbaciones más complejas.

Las perturbaciones han sido implementadas en MATLAB de manera tal el primer argumento de la función corresponde al residuo que se desea perturbar, y luego se tienen otros argumentos que dependerán de los parámetros de cada perturbación. La salida corresponde al residuo perturbado, el cual tiene el mismo largo que la señal de entrada.

Se entrega una descripción matemática de cada perturbación, en dónde se define el residuo como $x[n]$, y el residuo perturbado como $x_p[n]$, donde $n = 1, 2, ..., L$. La frecuencia de muestreo corresponde a f_s, y el periodo de muestreo se define como $t_s = \frac{1}{f_s}$.

Los gráficos presentados corresponden a una vocal "a" sostenida, se muestra el residuo antes y después de introducir la perturbación, en el dominio del tiempo y de la frecuencia.

3.2.1 Introducción de ruido

La perturbación más sencilla consiste en agregar ruido a la señal de la glotis. Para esto se ha implementado la función

$$[residuo\_pert, noise] = pert\_noise(residuo, fs, amp, range);$$

que tiene como entradas, el residuo que será perturbado ($residuo$), la frecuencia de muestreo (fs), la amplitud del ruido que se utilizará (amp) y el rango de frecuencias que contendrá este ruido ($range$). Este rango corresponde a un vector con 2 elementos que deben estar en $[1, \frac{fs}{2}]$ (Ver Figura 3.11).

Se genera un vector $r[n]$ compuesto por números aleatorios, uniformemente distribuidos entre 0 y 1, utilizando el comando $rand(L, 1)$. Luego se procede a filtrar dicho vector con un filtro pasabandas de orden mínimo que cumpla con una caída de 60[dB] en la banda de corte, obteniéndose $r_f[n]$. Finalmente el residuo perturbado corresponde a lo siguiente:

$$x_p[n] = x[n] + amp(2r_f[n] - 1) \tag{3.17}$$

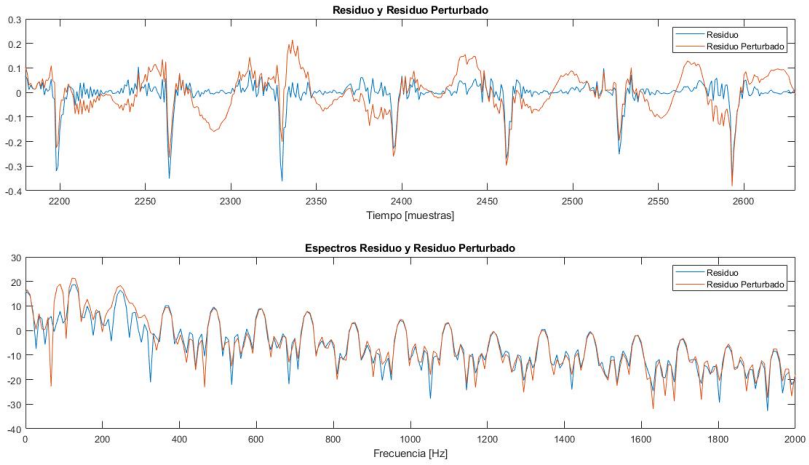


Figura 3.11: Residuo y residuo perturbado con ruido de amplitud 0.3 , y rango de frecuencia [50 300] [Hz]. Puede verse que el espectro de ambas señales es casi idéntico de los 300[Hz] en adelante.

3.2.2 Filtros

Se han implementado también funciones para aplicar filtros pasaaltos y pasabajos a la señal. Se ha utilizado la función de MATLAB *'butter'* la cual genera los coeficientes para un filtro de tipo butterworth (Ver Figuras 3.12 y 3.13).

El filtro butterworth busca tener una respuesta lo más plana posible en la banda de paso, esto es una ventaja para esta aplicación ya que no debería distorsionar demasiado los armónicos de interés. Tomando en cuenta la importancia de que la señal sea procesada en el menor tiempo posible, se generan filtros de segundo orden.

$$residuo\_pert = pert\_pasabajos(residuo, f_s, f_c)$$
$$residuo\_pert = pert\_pasaaltos(residuo, f_s, f_c)$$

estas tienen como entrada, el residuo que será perturbado ($x[n]$), la frecuencia de muestreo (f_s), y la frecuencia de corte desde la que se aplicarán los filtros (f_c).

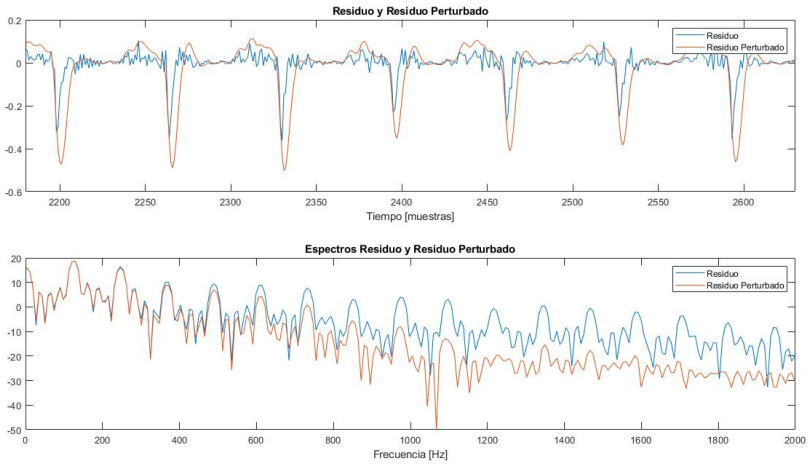


Figura 3.12: Residuo y residuo perturbado con filtro pasabajos, frecuencia de corte 500[Hz].

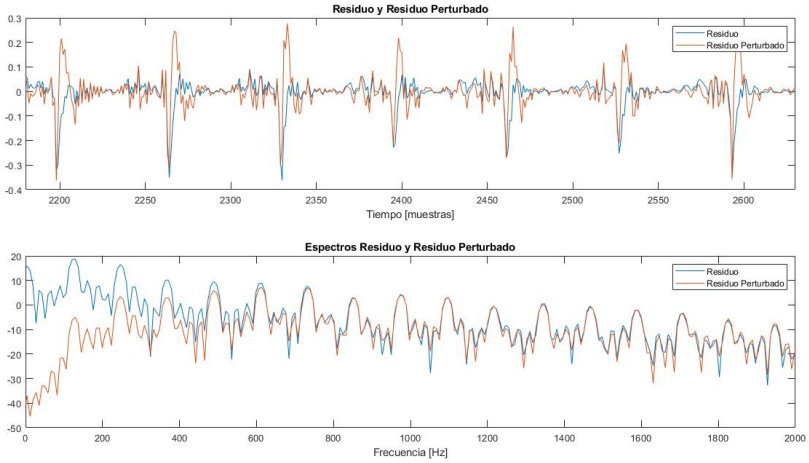


Figura 3.13: Residuo y residuo perturbado con filtro pasaaltos, frecuencia de corte 500[Hz].

3.2.3 Efecto Tremor

Se ha implementado una función que aplica un efecto "tremor" al residuo multiplicándolo por una señal sinusoidal. Para esto se han utilizando los conceptos de amplitud modulada, donde la portadora corresponde al residuo, y el mensaje es una sinusoidal (Ver Figura 3.14).

$$residuo\_pert = pert\_tremor(residuo, f_s, amp, f)$$

esta función tiene como entrada el residuo que será perturbado ($residuo$), la frecuencia de muestreo (f_s), la amplitud de la sinusoidal (amp) y su frecuencia fundamental (f). Matemáticamente, se define lo siguiente:

$$x_p[n] = x[n] \cdot (1 + amp \cdot \sin \frac{2\pi f n}{f_s}) \tag{3.18}$$

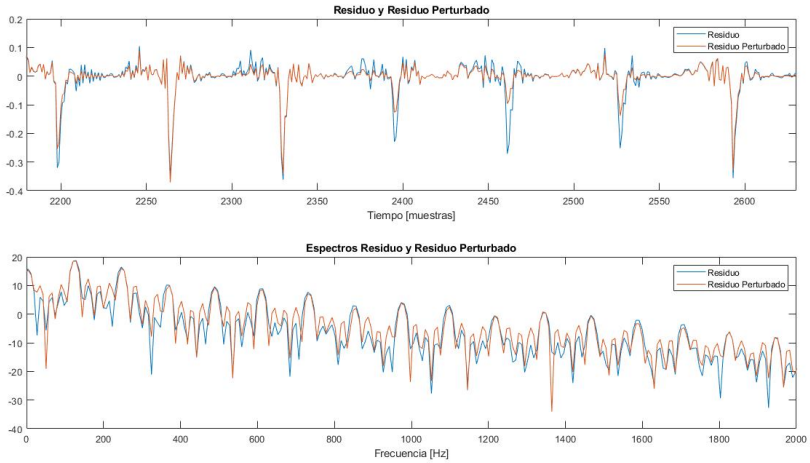


Figura 3.14: Residuo y residuo perturbado con tremor de frecuencia 100[Hz] y 0.5 de amplitud.

3.2.4 Narrow Band FM

Basado en la estructura de Narrow Band FM, se ha implementado una perturbación, en la cual, el residuo se comporta como la portadora, y el mensaje corresponde a una sinusoidal de frecuencia f_c (Ver Figura 3.16).

$$residuo\_pert = pert\_NBFM(residuo, f_s, A, f, k_f)$$

La idea detrás de este método es alterar el espectro del residuo de la menera que un mensaje altera a la portadora en la modulación Narrow Band FM (Ver Figura 3.15).

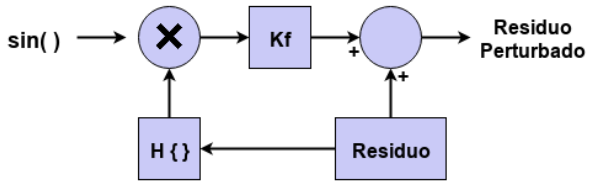


Figura 3.15: Diagrama de Bloques Perturbación NBFM.

matemáticamente, la señal puede expresarse de la siguiente forma:

$$x_p[n] = -A\left(-x[n] + k_f\hat{x}[n]sin(\frac{2\pi fn}{f_s})\right);$$

donde $\hat{x}[n]$ corresponde a la Transformada de Hilbert de $x[n]$.

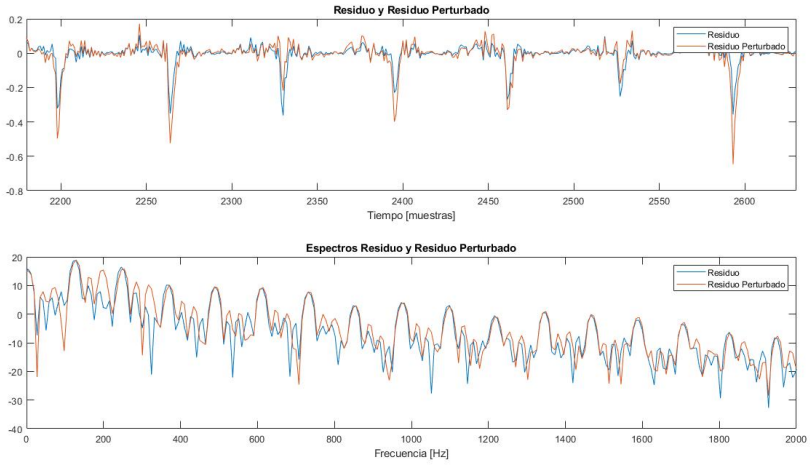


Figura 3.16: Residuo y residuo perturbado con NBFM de frecuencia 200[Hz], kf=0.5 y A=1.

3.2.5 Alteración de forma de los ciclos

Finalmente, se ha diseñado una perturbación que busca alterar la forma de cada ciclo de la glotis de manera independiente. Para no tener problemas a la hora de unir los diferentes ciclos, se ha generado una ventana que mantiene los extremos de estos para facilitar la concatenación.

$$[residuo\_pert, n\_vector] = pert\_shape(residuo, fs, goi\_ins, vector)$$

esta función tiene como entrada el residuo que será perturbado (*residuo*), la frecuencia de muestreo (f_s), los instantes de máxima apertura de la glotis (*goi_ins*), y un vector de coeficientes (*vector*), que determinará el tipo de alteración aplicada a cada ciclo.

La ventana utilizada está definida por tramos en función de la longitud de cada ciclo de la glotis L_g, y de un factor b, y siendo $a_0 = 053836$ y $a_1 = 0,46164$ (Ver Figura 3.17).

$$v[n] = \begin{cases} 1 & si \quad n \leq 4 \\ 1 + b(a_0 - a_1 \cos(\frac{2\pi n}{L_g - 9})) & si \quad 4 < n < L_g - 3 \\ 1 & si \quad n \geq L_g - 3 \end{cases} \tag{3.19}$$

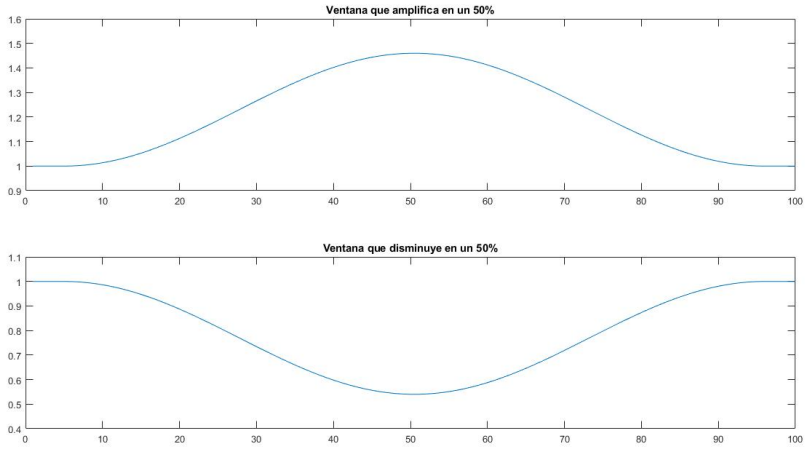


Figura 3.17: Ejemplos de Ventana que amplifica (b=0.5) y disminuye el ciclo (b=-0.5).

El vector de coeficientes "*vector*" contiene los factores b que se aplicarán a las ventanas que multiplicarán cada ciclo de la glotis. En caso de que "*vector*" tenga menos coeficientes que el número de ciclos del residuo a perturbar, se repite nuevamente la secuencia de coeficientes hasta completar el número de ciclos necesarios. En caso de que tenga más coeficientes que el número de ciclos, se trunca para obtener el valor necesario. Por ejemplo, al utilizar un $vector = [05 - 05]$, se tiene que la perturbación amplifica un ciclo, y disminuye el siguiente (Ver Figura 3.18).

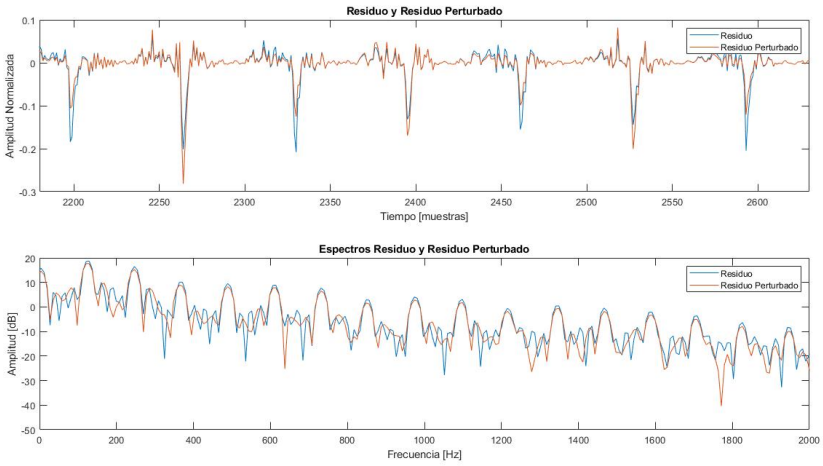


Figura 3.18: Residuo y residuo perturbado con un vector de [0.5 -0.5].

3.3 Resíntesis

La resíntesis de las señales perturbadas es un proceso que dependerá en gran medida del método que se haya utilizado para la descomposición. Debido a que los 3 métodos utilizados en la descomposición de las señales, consisten en realizar un filtrado inverso para eliminar el efecto del tracto vocal, para realizar la síntesis se utiliza un filtrado de tracto vocal. Para poder realizar este proceso correctamente, es necesario almacenar la información de los diferentes filtros utilizados en la descomposición, para luego invertir su efecto en el proceso de síntesis. Otro detalle a tomar en cuenta es el uso de algún tipo de ventana para los frames utilizados en la descomposición, de ser así, se debe buscar anular el efecto de la ventana en la síntesis.

Para realizar la resíntesis de las señales, se han implementado en MATLAB 3 funciones, las cuales son complementarias a las implementadas para la descomposición, y reciben como entrada las salidas de esta.

$$sintesis = sintesis\_LPC(residuo\_LPC, LPC\_M, LPC\_ins)$$

$$sintesis = sintesis\_IAIF(residuo\_IAIF, IAIF\_M, goi\_ins)$$

$$sintesis = sintesis\_QCP(residuo\_QCP, QCP\_M, QCP\_ins)$$

Las 3 funciones tienen como entrada el residuo obtenido de la respectiva descomposición ($residuo$), una matriz que contiene los coeficientes de los filtros

utilizados en la descomposición ($LPC\_M, IAIF\_M, QCP\_M$) y finalmente los instantes entre los que se tomaron las ventanas para la descomposición ($LPC\_ins, goi\_ins, QCP\_ins$). La salida de todas ellas es la señal de voz que ha sido sintetizada ($sintesis$).

3.4 Perturbación de Habla

El método planteado hasta ahora permite introducir perturbaciones a segmentos que contengan sonidos vocales, esto da buenos resultados para pruebas con vocales sostenidas, pero no es posible aplicar este método a un segmento de habla, ya que el lenguaje natural contiene diversos sonidos, que se intercalan con silencios para formar frases.

Para poder aplicar perturbaciones como las diseñadas, será necesario que la metodología que introduce las perturbaciones, sea capaz de discernir entre los espacios que contienen habla, y los que no, y además de esto, deberá ser capaz de separar los sonidos vocales, de los demás sonidos que no incluyen vibración de las cuerdas vocales.

Para separar los instantes en que hay fonación modal del resto de la señal, se ha implementado la función

$$[y, inicio\_ins, termino\_ins] = get\_un\_voiced(x, fs)$$

la cual recibe como entrada un segmento de voz natural (x), y la frecuencia de muestreo a la que fue grabada (fs), y entrega como salida un segmento del mismo largo de la señal de entrada (y), compuesto por 1s y 0s para los segmentos en que hay fonación, y no hay fonación respectivamente. También entrega $inicio\_ins$ y $termino_ins$, que corresponden a los puntos en que comienzan y terminan las partes con voz respectivamente.

Esta función trabaja con una ventana de 240 muestras, la cual se va actualizando agregando 80 muestras nuevas, y eliminando las 80 muestras más antiguas. La función calcula la frecuencia fundamental y el valor RMS de cada ventana, y en base a los valores obtenidos, es capaz de clasificar si el segmento de 80 muestras analizado corresponde a un tramo en donde hay voz modal o no (Ver Figura 3.19).

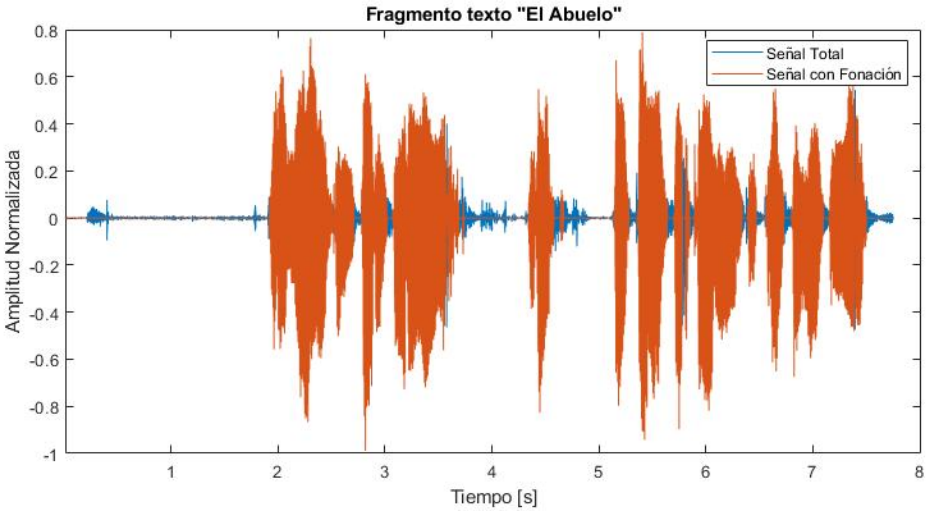


Figura 3.19: Segmento del texto "El Abuelo", en azul, la señal completa, en rojo, solo la parte que contiene voz modal.

Luego de obtener las partes del texto con voz, se aplica la metodología para introducir pertubaciones a cada uno de los segmentos de la misma manera que se perturbaron las vocales sostenidas, y luego se procede a concatenar los segmentos con voz perturbados, y los sin voz sin perturbar para formar una señal de lenguaje natural con la calidad vocal alterada.

3.5 Evaluación de las perturbaciones

Aplicar perturbaciones a la fuente de la glotis, tiene una repercusión directa en la señal de voz, estos cambios podrían ser o no detectados por los sujetos de prueba, de manera consciente o inconsciente, incluso podría ocurrir que algunos sujetos lo noten, y otros no. Para poder evaluar la reacción de las personas ante las perturbaciones, es fundamental tener una manera de evaluar las perturbaciones en sí mismas, para que sea posible establecer márgenes entre los cuáles las perturbaciones son razonables, y qué efecto tiene aumentar o disminuir los parámetros de una perturbación. Para esto, se ha optado por 2 maneras diferentes, la evaluación instrumental a través de medidas acústicas de la calidad vocal, y la evaluación perceptual a través de un experimento realizado con un grupo de personas que dieron su opinión respecto a las perturbaciones.

- **Evaluación Acústica:** La evaluación acústica de la calidad vocal se realiza a través de parámetros que se calculan matemáticamente a partir de las señales de audio. Para medir la calidad vocal de las señales antes y después de ser perturbadas, se ha optado por utilizar 3 parámetros: CPP, HNR y

Decaimiento Espectral. Cada perturbación cuenta con parámetros que pueden ajustarse, los cuales modulan algunas características de cada perturbación. Para ver el efecto que tienen estos parámetros en la calidad vocal, se graficarán los valores de CPP, HNR, y Decaimiento Espectral a medida que se desplazan dichos parámetros.

- **Evaluación Perceptual:** La evaluación perceptual, como su nombre lo indica, se trata de como las personas perciben una voz al escucharla. Este método de evaluación puede llegar a ser altamente subjetivo, ya que cada persona percibe las voces de manera diferente, y cada voz tiene atributos únicos. La manera más fidedigna de realizar una evaluación perceptual, es recurrir a especialistas entrenados, capaces de aplicar un test estarizado para medir cualidades de la voz. Realizar una evaluación profesional, está fuera del alcance de este trabajo, por lo cual, se optó por realizar una evaluación perceptual con personas comunes, sin experiencia en el tema, que pudieran calificar voces normales y perturbadas en una escala de 1 a 5. Debido a la situación de pandemia, este test fue realizado de manera online, los detalles de este, serán mencionados en la sección Pruebas a Realizar.

3.6 Pruebas a realizar

Para evaluar la metodología propuesta, se ha implementado el algoritmo en el software MATLAB, y se ha puesto a prueba con diferentes señales obtenidas a partir de hablantes hombres y mujeres. Las diferentes pruebas realizadas se describen a continuación.

3.6.1 Señales de prueba

Se solicita a 5 hablantes, realizar 2 tareas, la primera es decir una vocal “a”, sostenida durante 3 segundos. La segunda, consiste en la lectura en voz alta, y de forma pausada de un fragmento de texto llamado “El Abuelo”.

Los sujetos de prueba corresponden a adultos jóvenes, hablantes nativos de español chileno, residentes de la ciudad de Valparaíso. Ninguno de ellos ha sido diagnosticado con alguna patología de la voz o audición.

La vocal sostenida y la lectura del texto “El Abuelo”, fueron grabadas con un teléfono celular, a una distancia de entre 3 y 5 [cm] de los hablantes, en un lugar silencioso. Las señales fueron muestreadas a 8[kHz], y almacenadas en formato “wav”.

3.6.2 Pruebas estimación del flujo de la glotis y resíntesis

Se realizan pruebas con los 3 métodos mencionados: LPC, PSIAIF y QCP. Para estas pruebas, se utilizan segmentos de las grabaciones de vocales sostenidas de 4 hablantes. Se analizan los resultados tomando como referencia el modelo LF, y también se observa la estimación del espectro generada por cada método.

Para la prueba de resíntesis de las señales, se hará la descomposición de las vocales sostenidas, y luego se sintetizarán sin introducir perturbaciones, para luego comparar la señal sintetizada con la original.

3.6.3 Pruebas con medidas acústicas de calidad vocal

Se medirá la calidad vocal de las señales, antes y después de introducir las perturbaciones, utilizando 3 medidas de calidad vocal: CPP, HNR y Decaimiento Espectral. Debido a que las perturbaciones pueden modularse variando sus parámetros, es importante medir de alguna manera el efecto que tienen algunos parámetros en la calidad vocal. Para esto se ha optado por realizar un barrido con los parámetros dentro de un rango, y en base a este barrido, graficar los resultados obtenidos con las medidas de calidad vocal, para cada uno de los hablantes (Ver Figura 3.20).

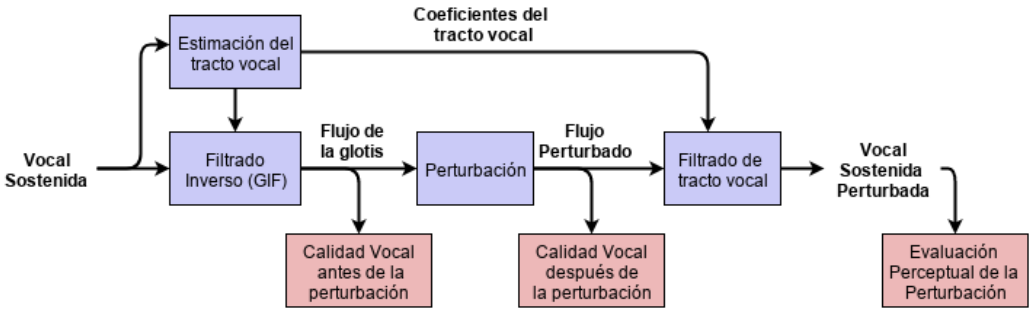


Figura 3.20: Diagrama de Medición de la Calidad Vocal.

3.6.4 Prueba Perceptual

Para la prueba perceptual, se realizó una encuesta a través de un sitio web especializado [83]. Este sitio permite añadir archivos de audio a las encuestas, lo cual lo hace ideal para realizar una evaluación perceptual. La encuesta realizada, contó de 3 partes que se describen a continuación:

- **Preguntas Generales:** En la primera parte, se solicitaron los siguientes datos personales: nombre, edad, género, lengua materna y presencia de problemas de audición.
- **Percepción de Vocales Sostenidas:** En la segunda parte se pedía evaluar las 20 vocales sostenidas utilizando una escala de 1 a 5, donde 1 correspondía a una voz natural, y 5 a una voz muy perturbada. De las 20 vocales, 3 correspondían a voces sin alteraciones, que debían servir como grupo de control, y las otras 17, correspondían a vocalizaciones con alguna de las 5 perturbaciones diseñadas (Noise, Tremor, Filtrado, NBFM o Shape). Las vocalizaciones se presentaron en orden aleatorio, intercalando tipos de perturbación, y hablantes.
- **Percepción de Frases:** En la tercera parte, se hizo algo similar a la segunda parte, pero esta vez con frases tomadas del texto "El Abuelo". Se dejaron 2 frases sin perturbar, que servirían como control, y se presentaron 8 frases perturbadas, de manera aleatoria.

Esta encuesta fue distribuida de manera online, y podía ser contestada desde un computador, teléfono celular, o tablet. Se solicitó a las personas que realizaran la evaluación en un lugar silencioso, utilizando audífonos, y se les indicó que podían escuchar las voces las veces que consideraran necesario.

La escala utilizada para evaluar las perturbaciones es subjetiva, y cada persona debió establecer sus propios parámetros de evaluación de las perturbaciones (Ver Figura 3.21).

Se intentó que entre las vocalizaciones y frases escuchadas, hubiesen diferentes variaciones e intensidades de cada perturbación, y también que aparecieran diferentes hablantes. Se buscó hacer una encuesta dinámica, que no tomara más de 10 minutos, un tiempo prudente para que una persona mantenga su concentración.

Finalmente, cabe mencionar que en la introducción de la encuesta, se mencionó brevemente que se enmarcaba en un trabajo de título de la Universidad Técnica Federico Santa María, y que los resultados solo serían utilizados con fines académicos y se presentarían los datos respetando el anonimato.

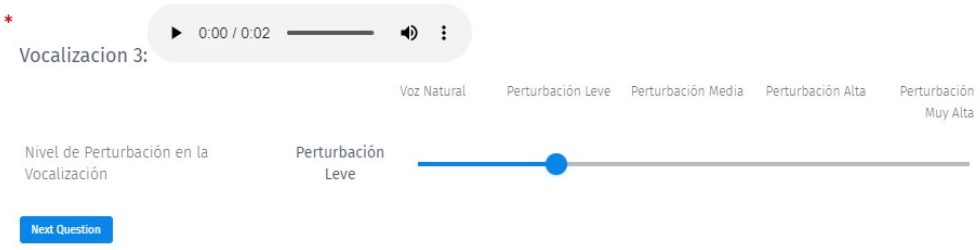


Figura 3.21: Pregunta tipo de la encuesta de Percepción Auditiva.

3.6.5 Prueba de tiempo de cómputo

Como primer acercamiento a un experimento en tiempo real, es importante tener una noción del tiempo de cómputo que requiere el algoritmo planteado. Para poder realizar la medición temporal, se utilizan los comandos de MATLAB *tic* y *toc*, en las pruebas de descomposición, introducción de perturbaciones, y de resíntesis, tanto en vocales sostenidas como en frases, para diferentes hablantes.

Se debe tener en cuenta que el computador utilizado corresponde a un "Intel(R) Core (TM) i5-6200U CPU 2.30GHz" con 8GB de memoria RAM. El software utilizado corresponde a "MATLAB R2019b".

Capítulo 4

Resultados

4.1 Resultados de la Estimación del flujo de la glotis

Se presentan los resultados obtenidos utilizando los diferentes métodos de estimación del flujo de la glotis.

4.1.1 LPC

Se realizaron 3 estimaciones con LPC de grado 8, 10 y 12 (Ver Figura 4.1). Es posible apreciar que las estimaciones de orden 10 y 12 son bastante similares, y la de grado 8 no da tan buenos resultados como las otras 2.

Se realiza el filtrado inverso utilizando las 3 estimaciones (Ver Figura 4.2). En el primer gráfico es posible observar un segmento de 600 muestras de la señal de voz analizada, la cual corresponde a una vocal sostenida "a", producida por un hombre. También pueden verse en el primer gráfico, las "x" en color rojo, que corresponden a los puntos entre los que se seleccionaron las ventanas para realizar el LPC (cada 160 muestras). Al observar las señales filtradas, podemos ver que es posible obtener algo bastante similar al flujo de la glotis, descrito por el modelo LF. Un detalle a tener en cuenta, es que señal obtenida presenta peaks extraños en donde termina una ventana e inicia la siguiente. Comparando la estimación espectral que realizan los diferentes LPC, puede verse que el LPC de orden 8 no es capaz de captar el tercer formante de la señal, y que los espectros de orden 10 y 12 son bastante similares.

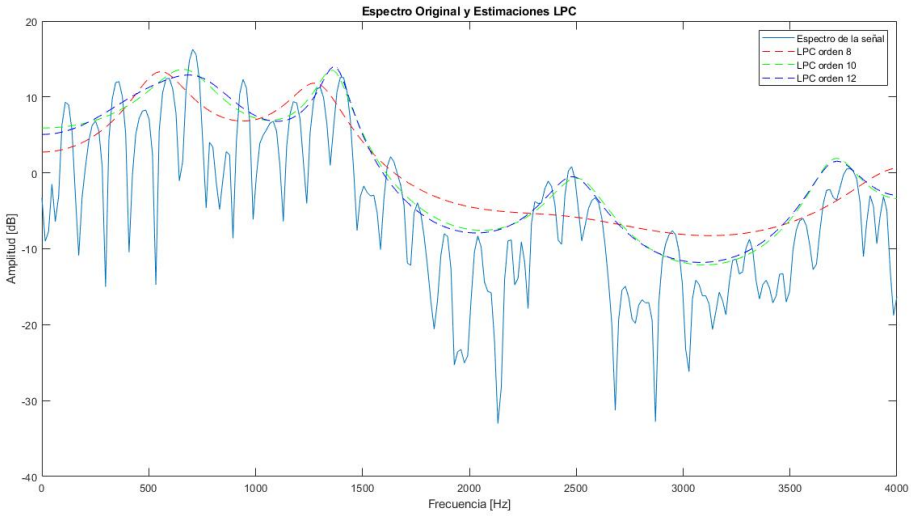


Figura 4.1: Diagrama de Medición de la Calidad Vocal.

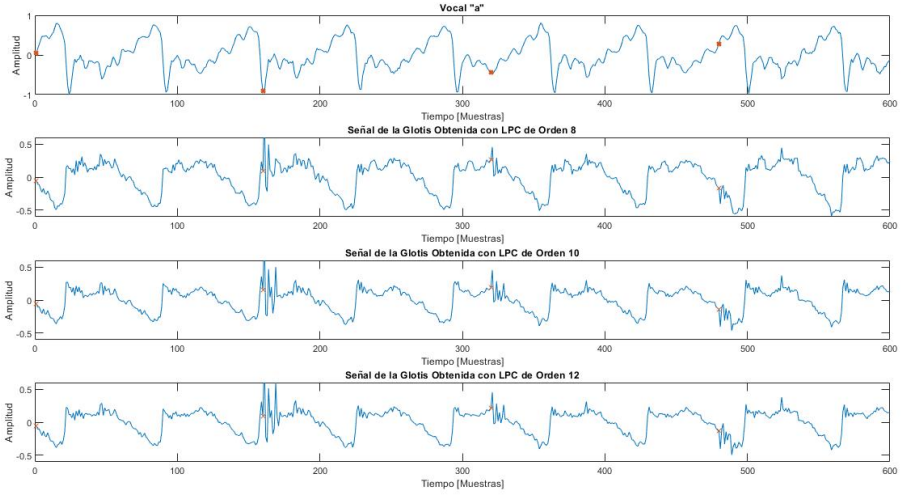


Figura 4.2: Diagrama de Medición de la Calidad Vocal.

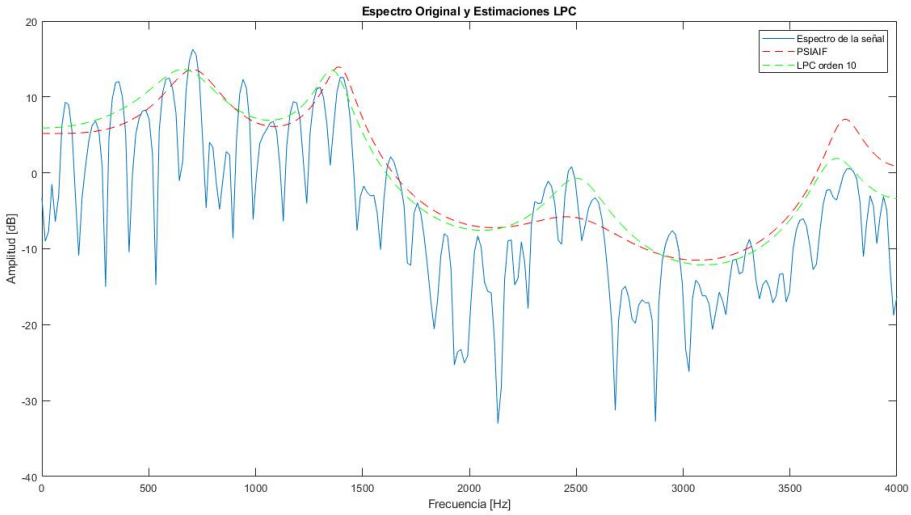


Figura 4.3: Diagrama de Medición de la Calidad Vocal.

4.1.2 PSIAIF

Para probar el método PSIAIF, se utiliza el mismo segmento de voz que fue utilizado con el filtrado LPC, los resultados se presentan a continuación.

Al observar el comportamiento del filtro obtenido, puede verse que este corresponde a la envolvente del espectro, de acuerdo a lo esperado (Ver Figura 4.3).

Pueden verse en la figura, los instantes estimados de máxima apertura de la glotis, que separan los diferentes segmentos, así como también la señal obtenida para la derivada del flujo de la glotis. Puede apreciarse que los valores obtenidos para el primer y el último segmento, no coinciden con lo esperado, esto se debe a que el algoritmo está diseñado para funcionar con ventanas que contengan un ciclo completo de la glotis, de máximo a máximo. A pesar de esto, se logra un resultado satisfactorio para los segmentos que toman un ciclo completo (Ver Figura 4.4))

4.1.3 QCP

Se puede ver que el espectro estimado corresponde con lo esperado (Ver Figura 4.5). Para la señal en el tiempo, se ha graficado también la función de pesos, que se asigna a la señal de acuerdo a los instantes de apertura y clausura de la glotis. Esta función da una mayor ponderación a los segmentos en que la glotis se encuentra cerrada.

A continuación, se presentarán los resultados obtenidos para diferentes hablantes, se analizará el funcionamiento de los métodos para los diferentes casos, y se verán las ventajas y desventajas de cada uno.

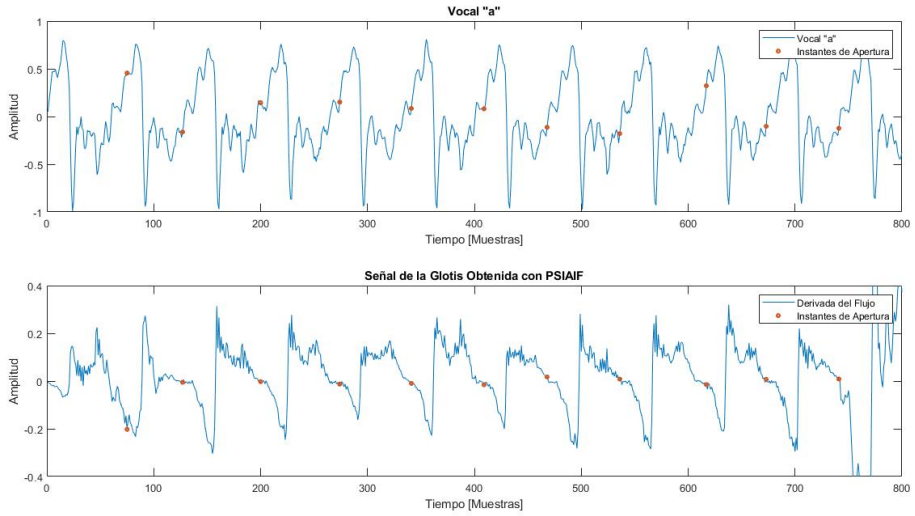


Figura 4.4: Diagrama de Medición de la Calidad Vocal.

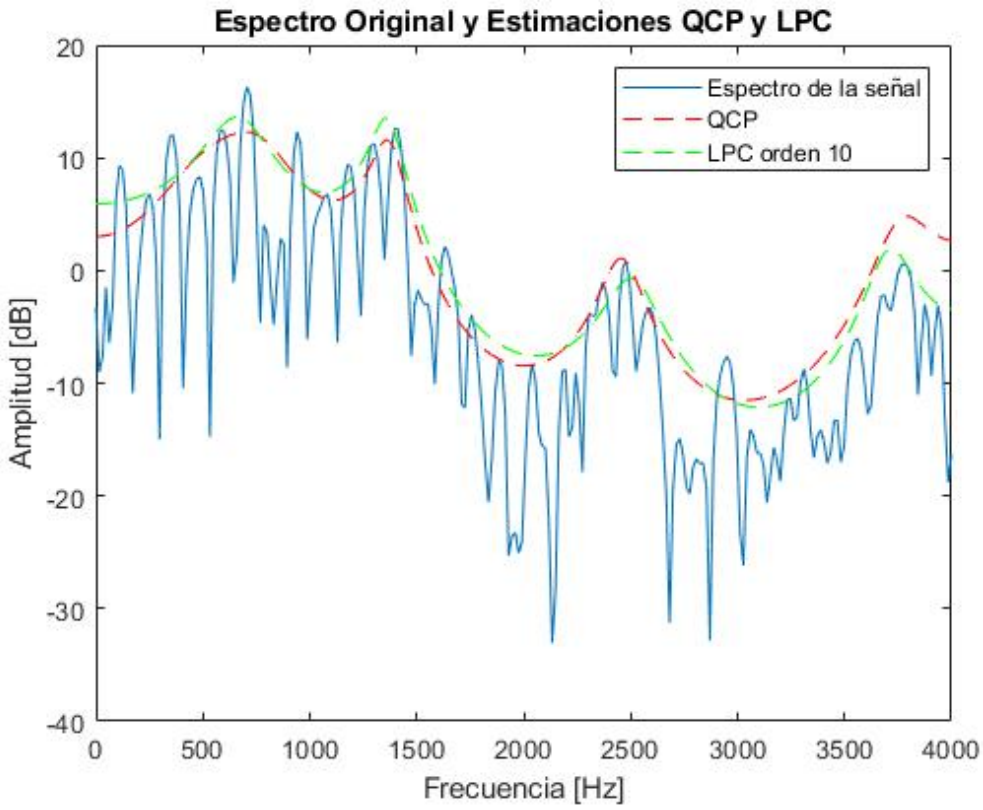


Figura 4.5: Diagrama de Medición de la Calidad Vocal.

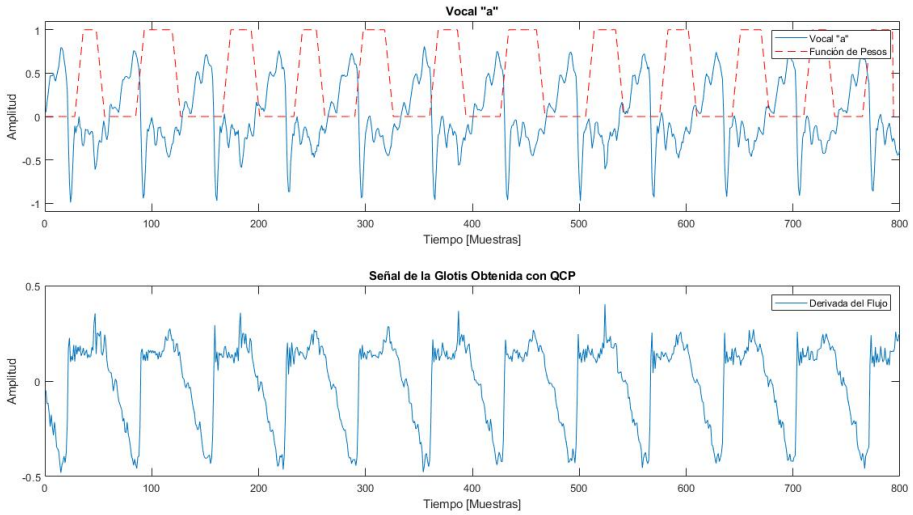


Figura 4.6: Resultados obtenidos para el método QCP. Puede verse la función de pesos (AME) en rojo.

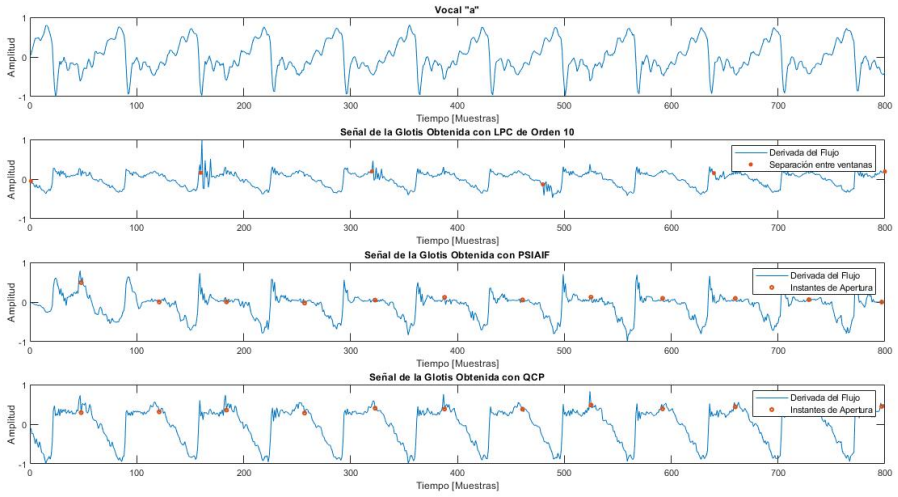


Figura 4.7: Resultados para Hablante 1: Hombre, 26 años, f_0=118[Hz] .

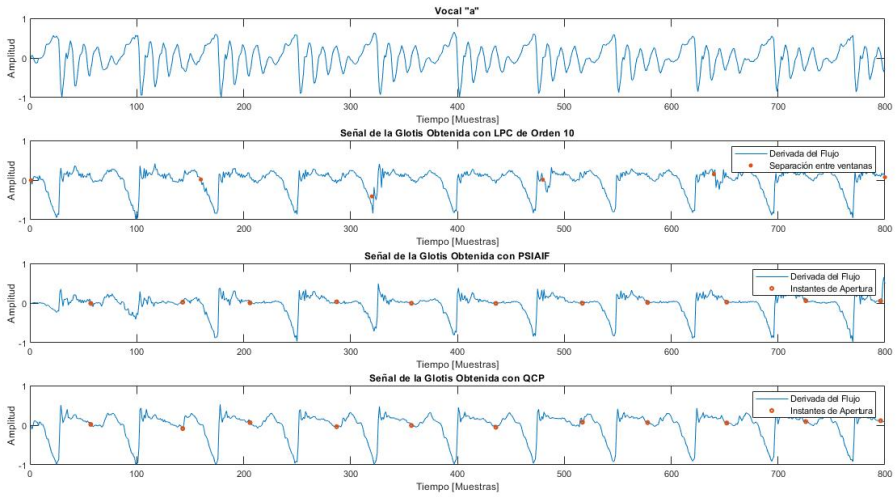


Figura 4.8: Resultados para Hablante 2: Hombre, 27 años, f_0=108[Hz] .

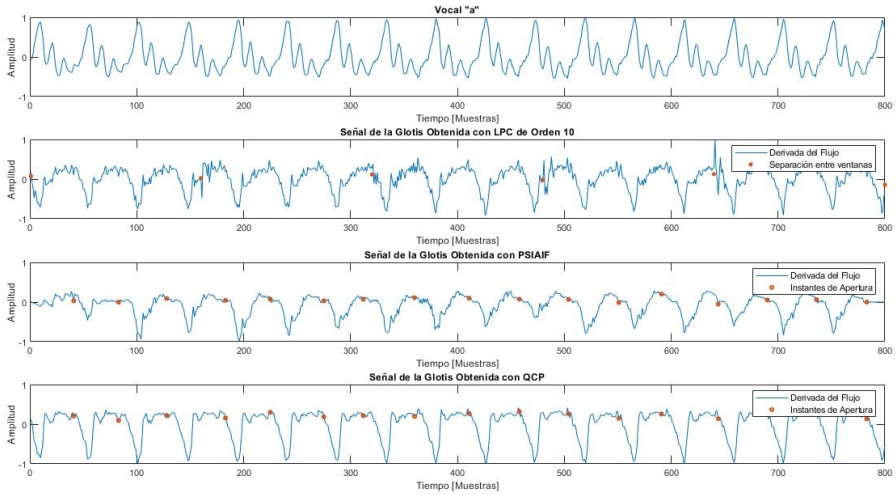


Figura 4.9: Resultados para Hablante 3: Mujer, 32 años, f_0=174[Hz] .

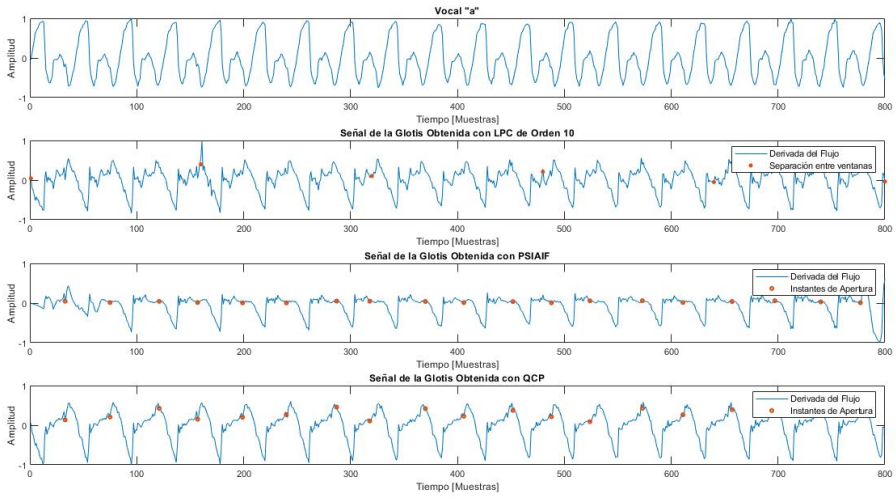


Figura 4.10: Resultados para Hablante 4: Mujer, 23 años, f_0=195[Hz] .

Al observar los resultados obtenidos para cada hablante (Ver Figuras 4.7, 4.8, 4.9, 4.10), es posible notar que el resultado para el método QCP parece ser el mejor para los 4, seguido del método PSIAIF, y luego por el LPC de grado 10. Esto era de esperarse, ya que el método QCP es el más actual de los 3, y también el que tiene un mayor costo de procesamiento. También se debe considerar que, para realizar la estimación del tracto vocal, toma varios periodos de la glotis, por lo cual es más robusto que los otros 2. Para el caso de PSIAIF, es posible ver que los resultados son buenos cuando se toma un ciclo completo de la glotis, pero para los primeros y últimos ciclos de cada ventana, es común que presente fallas, debido a que la estimación del tracto vocal se realiza con ciclos incompletos de la glotis. Por esta razón, este método es mucho más sensible a posibles errores de sincronización. Finalmente, el filtrado LPC de orden 10, presenta los peores resultados para los 3 hablantes, esto era de esperarse debido a que el método toma segmentos de voz en puntos cualquiera de la señal, sin realizar ningún tipo de análisis previo, por lo cual, presenta algunos peaks no deseados en los inicios o términos de cada ventana. Esto podría corregirse agregando algún mecanismo de sincronización al método. A pesar de todo, el método logra obtener una forma de onda dentro de lo esperado. Tomando en cuenta las características del experimento que se desea realizar, y la rapidez con la que debe realizarse el procesamiento para que no sea detectado por los sujetos de prueba, tomar el algoritmo QCP no parece una buena opción, ya que necesita varios periodos de la glotis para ser aplicado. Por otra parte, LPC, a pesar de su bajo costo computacional, no entrega resultados satisfactorios, y presenta perturbaciones en los cambios de ventana. Por lo que finalmente se opta por utilizar el algoritmo PSIAIF.

4.2 Resultados de la síntesis de vocales sostenidas

Para observar los resultados de la síntesis se descomponen las señales y luego son sintetizadas, a continuación se muestran los resultados obtenidos con diferentes métodos y hablantes, el error de la síntesis se calcula como el valor absoluto de la diferencia entre la señal original y la señal sintetizada (Ver Figuras 4.11, 4.12, 4.13).

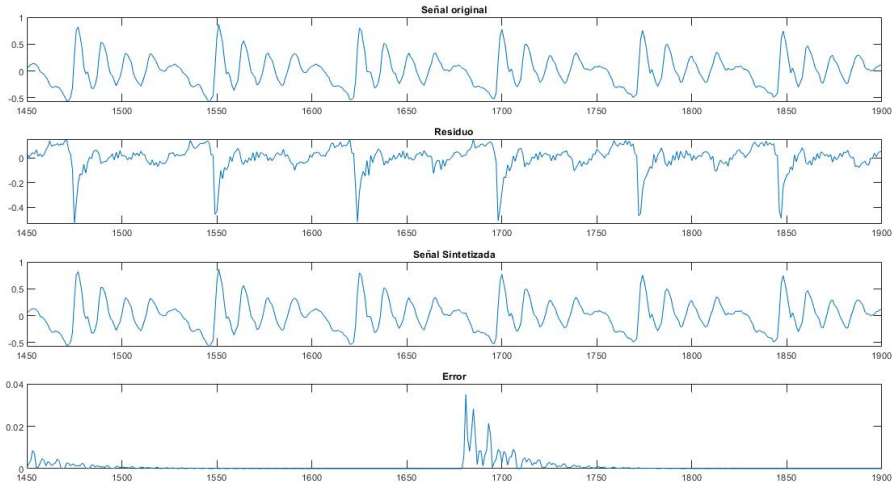


Figura 4.11: Resultados prueba de síntesis LPC .

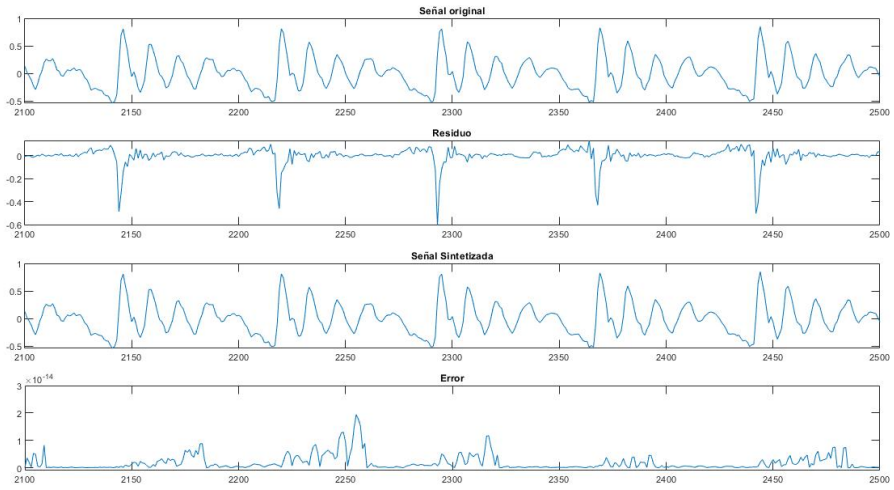


Figura 4.12: Resultados prueba de síntesis PSIAIF

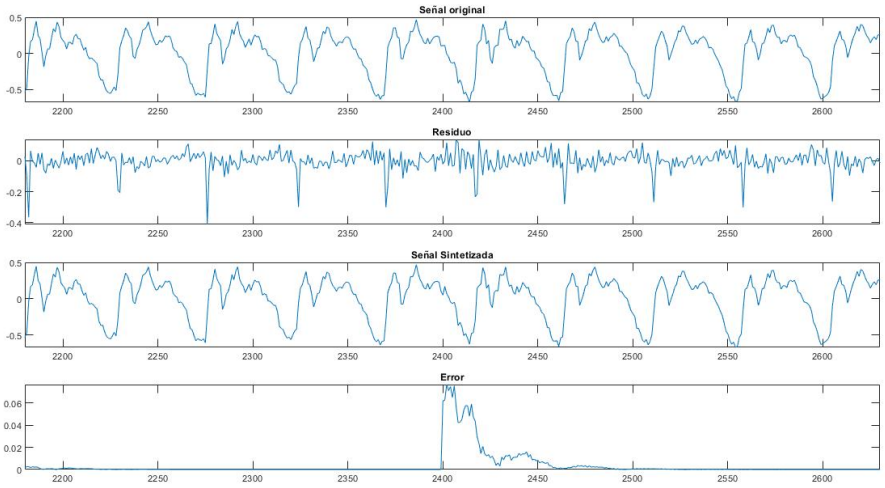


Figura 4.13: Resultados prueba de síntesis QCP.

Es posible apreciar que para los 3 casos, la señal sintetizada es, a simple vista, idéntica a la original. Al observar la señal de error, es posible ver que existe una diferencia entre ambas señales, más acentuada en la parte en que se concatenan las ventanas, pero el orden de magnitud de esta no es significativo, por lo cual, puede concluirse que se logra la síntesis para vocales sostenidas sin perturbaciones con el método propuesto.

4.3 Resultados de la introducción de perturbaciones en vocales sostenidas

Para medir el efecto que tienen las perturbaciones en las señales de voz, se han calculado 3 medidas acústicas de calidad vocal, a medida que se varían los parámetros de las perturbaciones. En base a los resultados obtenidos, se han generado gráficos, que permiten visualizar el efecto que tiene un determinado parámetro en las medidas de calidad vocal.

4.3.1 Introducción de ruido

Utilizando la función *pert_noise*, se ha generado un vector aleatorio de ruido en un rango entre 200 y 1200 [Hz], con una amplitud entre 0 y 1. Para los diferentes hablantes, se ha podido observar lo siguiente (Ver Figuras 4.14 y 4.15).

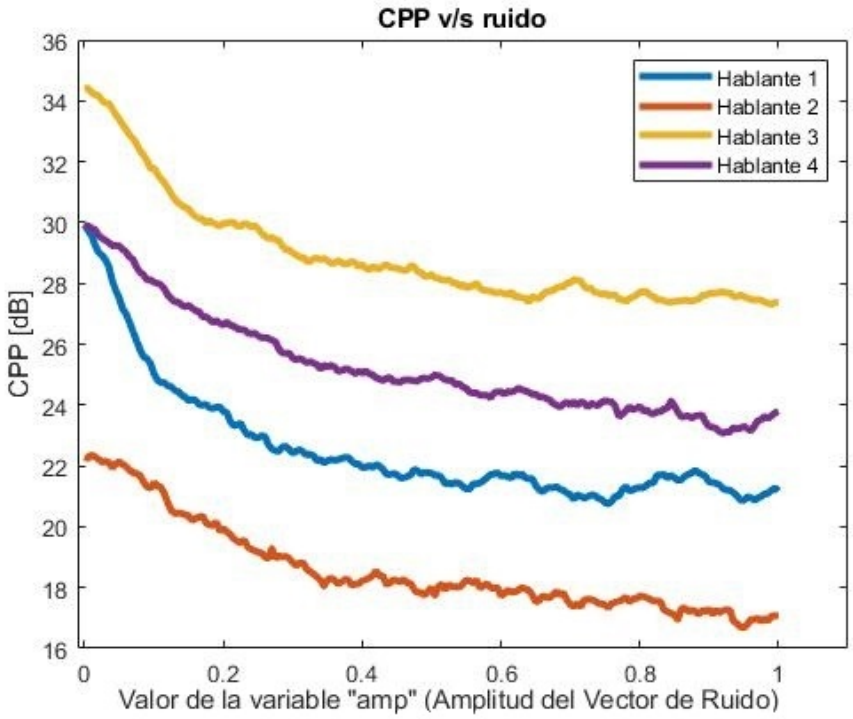


Figura 4.14: CPP en función de la amplitud del ruido introducido

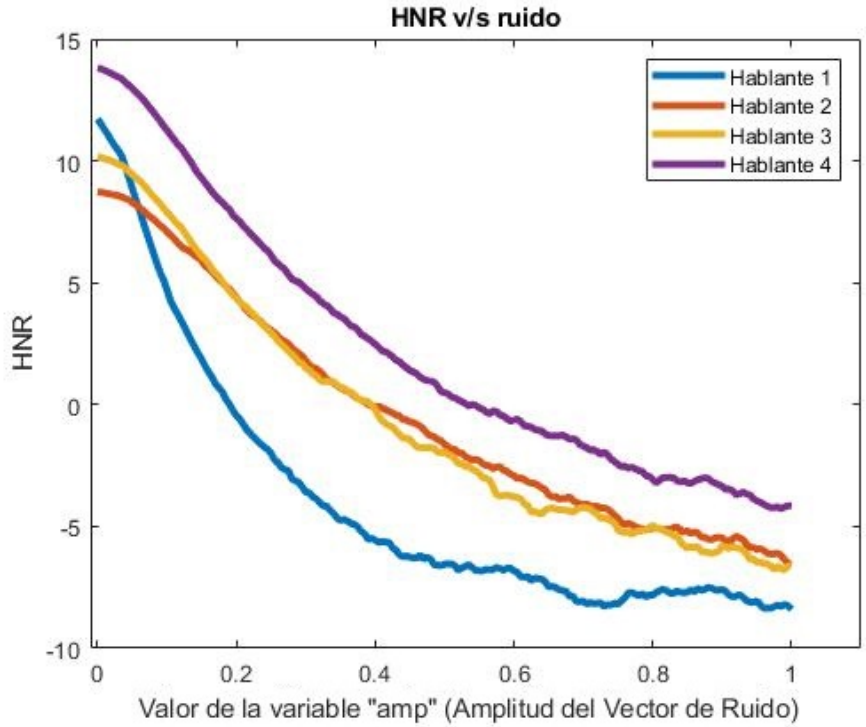


Figura 4.15: HNR en función de la amplitud del ruido introducido

En ambos casos, es posible ver que a medida que aumenta el nivel de ruido en las señales, las medidas comienzan a disminuir, de manera más abrupta en un principio, y pasado un cierto punto, tienden a disminuir más lentamente. Esto era algo totalmente esperable para el caso del HNR, que como su nombre indica, es la proporción entre los armónicos y el ruido. Para la medida CPP, este comportamiento se observa, pero de una manera menos directa, y con una mayor oscilación en la gráfica. Cabe mencionar que los 4 hablantes tienen un comportamiento bastante similar para ambas medidas.

4.3.2 Filtrado

Utilizando las funciones *pert_pasabajos* y *pert_pasaaltos*, se filtran los residuos, y se mide la calidad vocal variando la frecuencia de corte de los fitros.

4.3.2.1 HNR

Es posible ver que cuando solo se dejan pasar frecuencias por debajo de la frecuencia fundamental, el HNR, tiene un valor máximo, esto se debe a que la señal no tiene ninguno de sus armónicos. No tiene sentido aplicar esta perturbación, ya que

practicamente no hay energía por debajo de la frecuencia fundamental. Cuando la frecuencia de corte, permite que pase solo la frecuencia fundamental, el HNR alcanza un máximo global (notar que esto ocurre a diferencia frecuencias de corte para cada hablante). A medida que aparecen los próximos armónicos, el HNR va disminuyendo (Ver Figura 4.16).

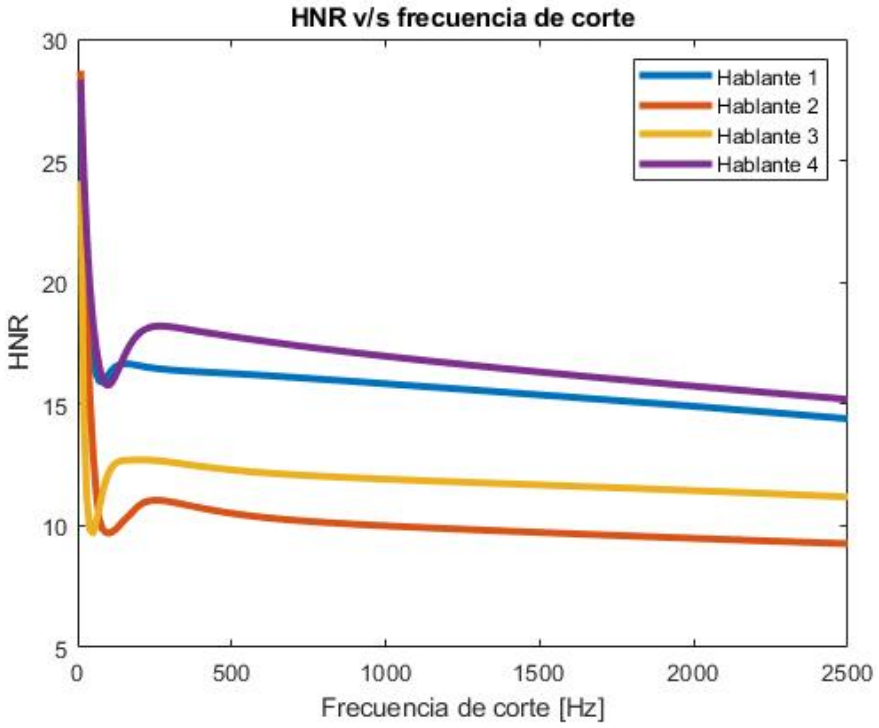


Figura 4.16: HNR en función de la frecuencia de corte de un filtro pasabajos

Al observar la gráfica del HNR para el filtrado con un pasaaltos, se puede observar que al tener una frecuencia de corte muy baja, estamos dejando pasar casi toda la señal, y el valor del HNR es muy similar al valor del HNR para el caso pasabajos con frecuencia de corte alta. Se puede observar un máximo en la señal previo a la frecuencia fundamental de cada hablante, esto ocurre porque el filtro deja pasar todos los armónicos y elimina todo el ruido que pueda haber en frecuencias más bajas. Pasado la frecuencia fundamental, el HNR comienza a disminuir, ya que se van eliminando los primeros armónicos de la señal (Ver Figura 4.17).

4.3.2.2 CPP

Al observar los gráficos de CPP en función de la frecuencia de corte, no es posible establecer ninguna relación. En general se puede observar que no existe una gran variación de este parámetro para las diferentes frecuencias (Ver Figuras 4.18 y 4.19).

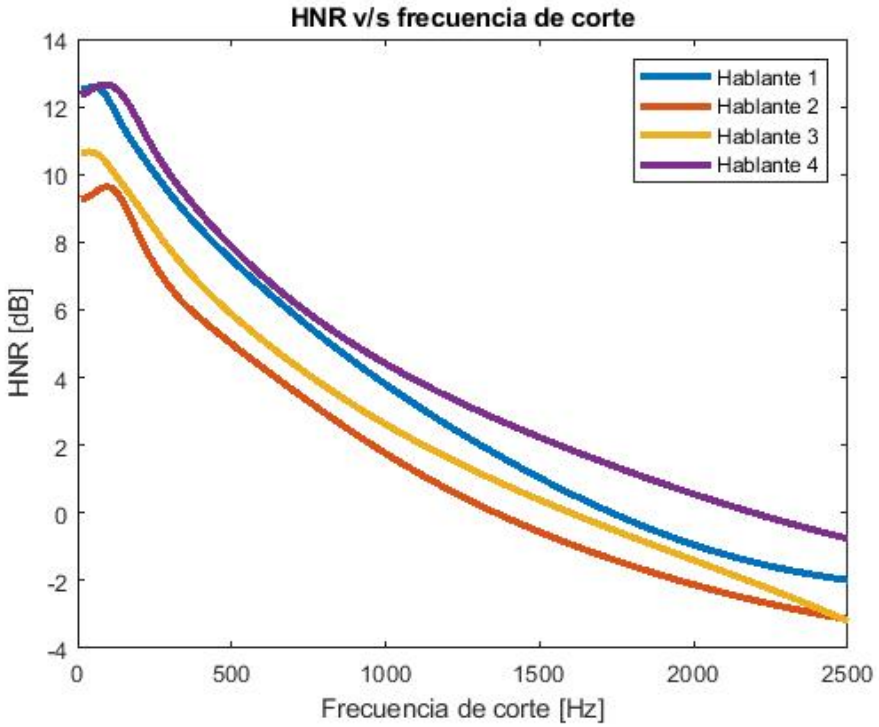


Figura 4.17: HNR en función de la frecuencia de corte de un filtro pasaaltos

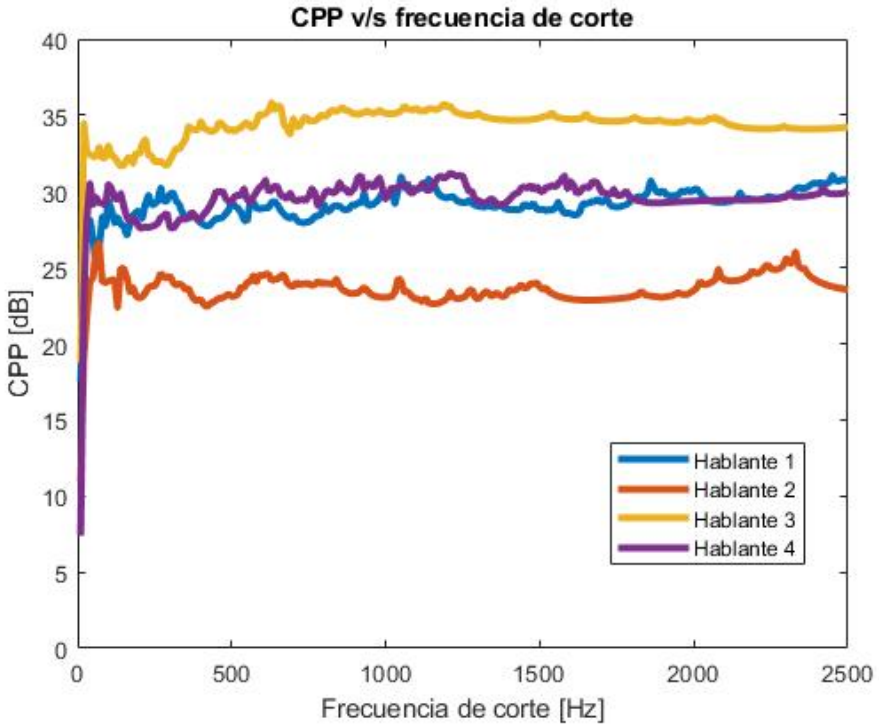


Figura 4.18: CPP en función de la frecuencia de corte de un filtro pasabajos

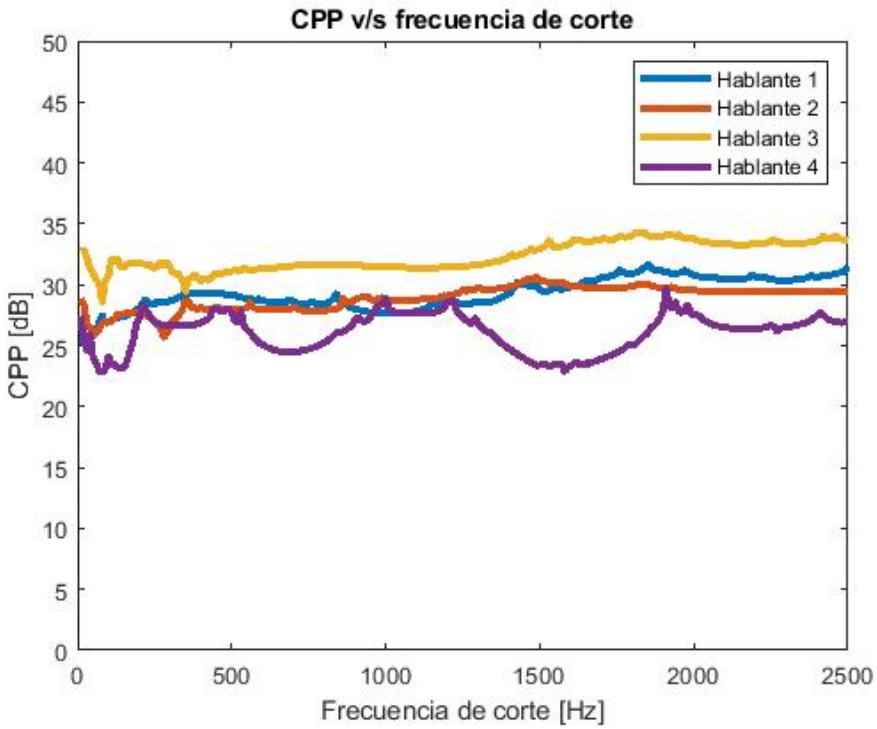


Figura 4.19: CPP en función de la frecuencia de corte de un filtro pasaaltos

4.3.2.3 Inclinación Espectral

Al observar la inclinación espectral, se puede observar que para el caso del filtro pasabajos, esta presenta grandes variaciones para los primeros 2 armónicos de las señales, y tiende a aumentar a medida que se van agregando armónicos, pero sin llegar a ser un valor positivo, esto se debe a que hay una mayor cantidad de energía en los primeros armónicos de las señales (Ver Figura 4.20). Para el filtro pasaaltos, la inclinación va aumentando de manera constante, y pasados los 500[Hz] pasa a ser una pendiente positiva (Ver Figura 4.21).

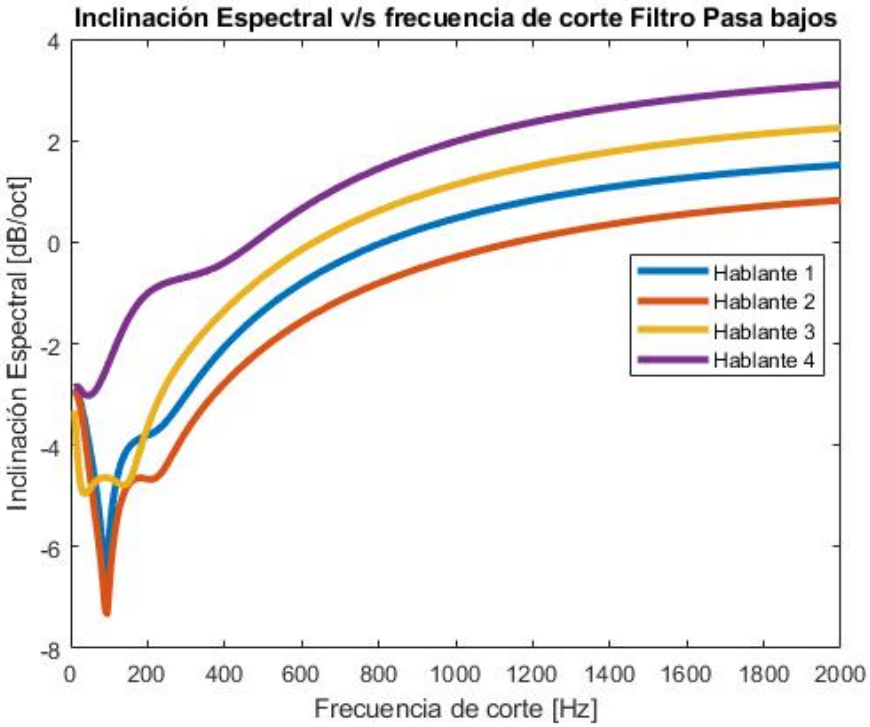


Figura 4.20: Inclinación Espectral en función de la frecuencia de corte de un filtro pasabajos

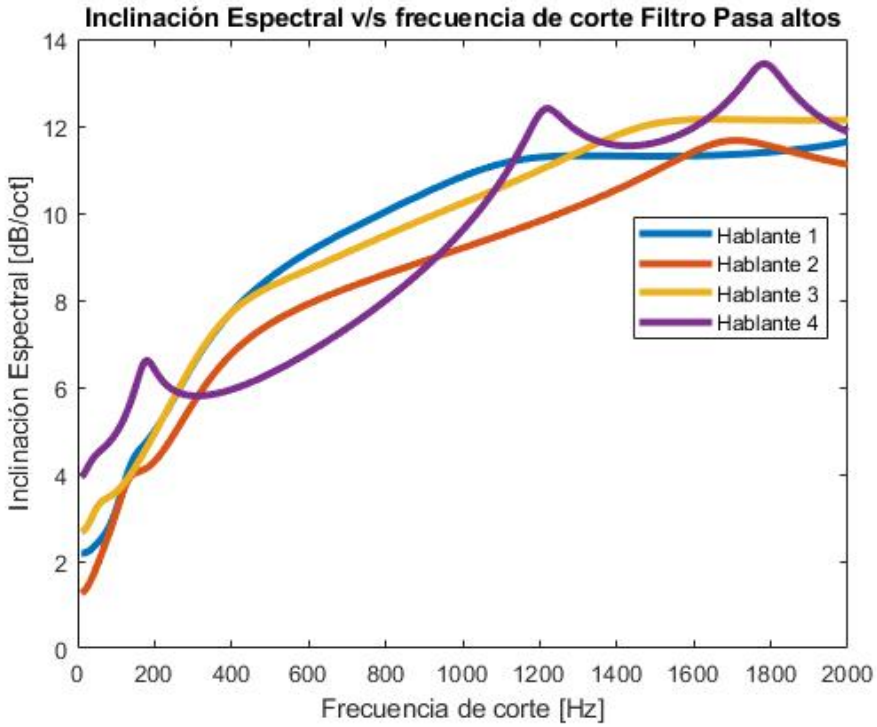


Figura 4.21: Inclinación Espectral en función de la frecuencia de corte de un filtro pasaaltos

4.3.3 Efecto Tremor y NBFM

Para los efectos tremor y NBFM, se ha analizado qué ocurre con las medidas de calidad vocal al variar la frecuencia de cada efecto. Se ha podido observar algunas similitudes entre los resultados obtenidos para los efectos tremor y NBFM, las cuales se describen a continuación.

4.3.3.1 HNR

Al observar el comportamiento del HNR en función de la frecuencia elegida para las perturbaciones, se puede ver que cuando la frecuencia coincide con la frecuencia fundamental de la señal, o algún armónico de esta, el HNR alcanza un valor máximo. Esto ocurre para las 4 voces, tanto para el efecto tremor, como para el efecto NBFM. Es de esperarse que tremor y NBFM tengan un comportamiento parecido, debido a la similitud espectral de ambos métodos. Un detalle que llama la atención, son los máximos locales que aparecen justo entre 2 armónicos para los hablantes 1 y 3 (Ver Figuras 4.22 y 4.23).

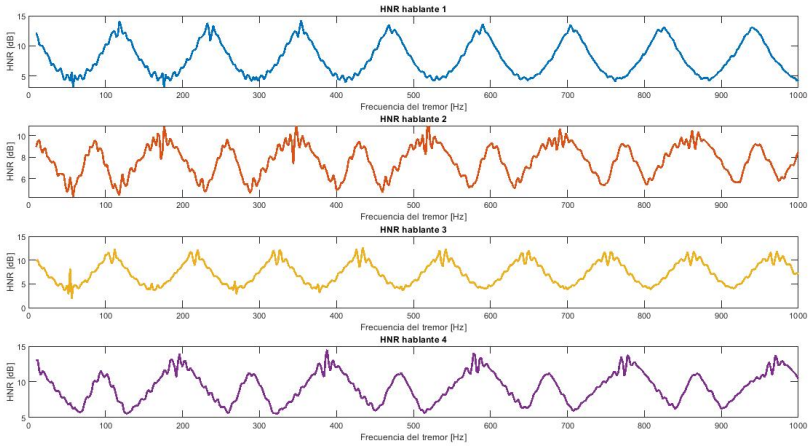


Figura 4.22: HNR en función de la frecuencia del efecto tremor

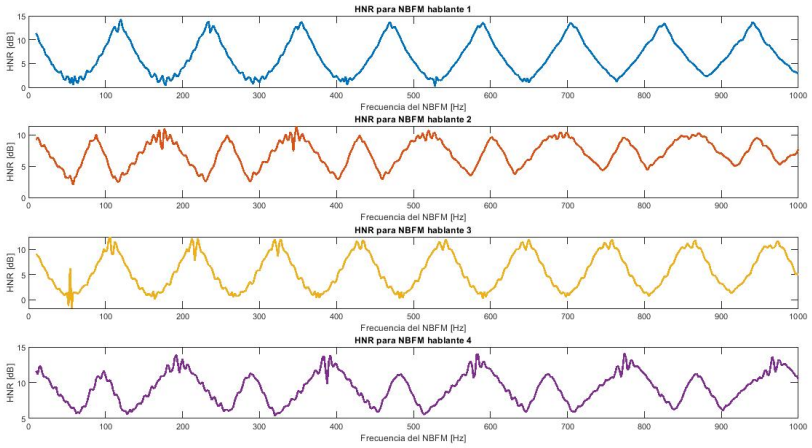


Figura 4.23: HNR en función de la frecuencia del efecto NBFM

4.3.3.2 CPP

Al observar qué ocurre con el CPP, se puede ver que este, de manera similar al HNR, también presenta un comportamiento periódico, que está relacionado con la frecuencia fundamental de cada hablante. Esto es bastante evidente para los primeros 3 hablantes, pero se vuelve menos claro para el hablante 4 (Ver Figuras 4.24 y 4.25).

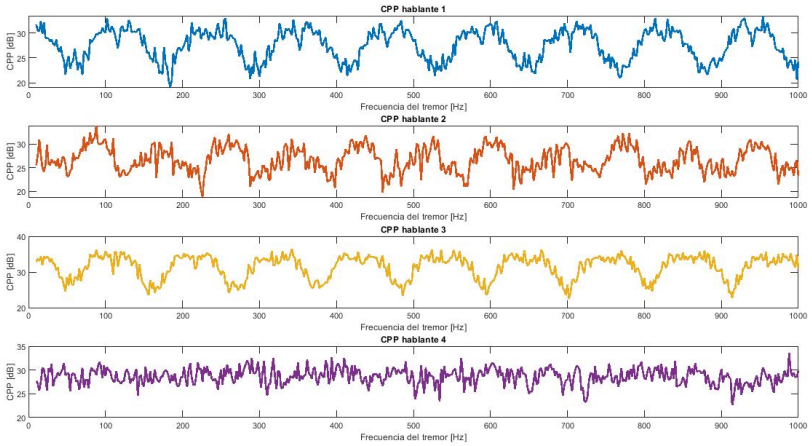


Figura 4.24: CPP en función de la frecuencia del efecto tremor

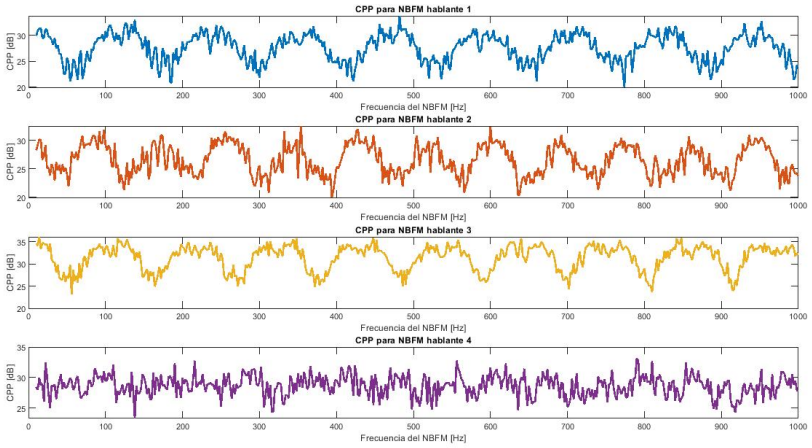


Figura 4.25: CPP en función de la frecuencia del efecto NBFM

4.3.3.3 Inclinación Espectral

Finalmente, al observar la inclinación espectral, también puede notarse un cierto grado de periodicidad relacionado con la frecuencia fundamental, pero bastante menor a lo encontrado para HNR y CPP. Los cambios en la inclinación son sutiles, y no alcanza a ocurrir un cambio de signo (Ver Figuras 4.26 y 4.27).

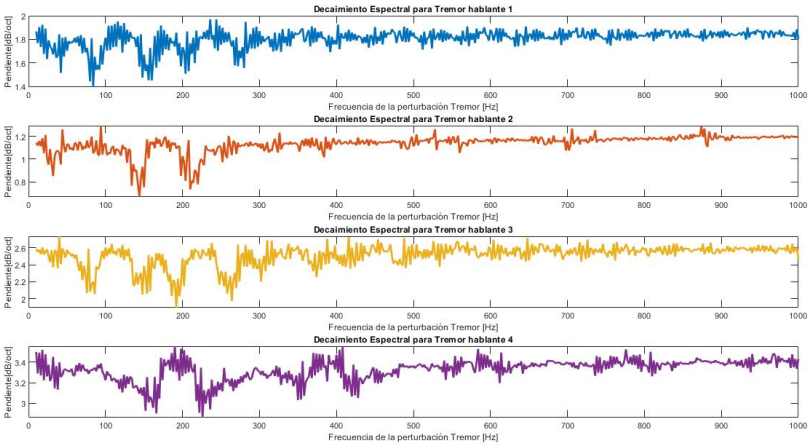


Figura 4.26: Inclinación Espectral en función de la frecuencia del efecto tremor

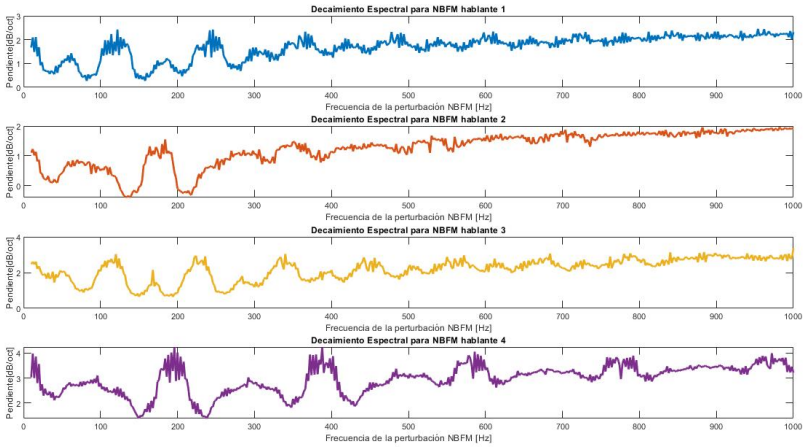


Figura 4.27: Inclinación Espectral en función de la frecuencia del efecto NBFM

4.4 Resultados Encuesta

La encuesta de percepción auditiva fue completada por 30 personas (15 hombres y 15 mujeres), a las cuales les tomó un promedio de 9 minutos. Todos los sujetos declararon que su lengua materna correspondía al español chileno, y 2 personas dijeron presentar algún tipo de discapacidad auditiva. El promedio de edad fue de 35 años. Se presentan los detalles de cada segmento de audio (hablante, perturbación y parámetros) y el puntaje promedio asignado por quienes respondieron la encuesta (Ver Cuadros 4.1 y 4.7).

	Vocales Sostenidas				
N°	**H**	f_0	**Pert.**	**Parámetros**	**P**
1	H1	117	-	-	1,63
2	H1	117	P.bajos	fc=1000	3,7
3	H1	117	P.bajos	fc=500	3,5
4	H3	105	-	-	1,53
5	H3	105	P.altos	fc=500	3,1
6	H3	105	P.altos	fc=1000	2,97
7	H2	174	-	-	1,93
8	H2	174	Tremor	amp= 0,1 f=50	2,17
9	H2	174	Tremor	amp= 0,2 f=50	3,23
10	H2	174	Tremor	amp= 0,1 f=174	1,9
11	H2	174	Tremor	amp= 0,1 f=200	2,73
12	H3	105	NBFM	kf=0.1 fc = 50	1,87
13	H3	105	NBFM	kf=0.2 fc = 109	2,27
14	H3	105	NBFM	kf=0.2 fc = 125	3,57
15	H3	105	Shape	[0,1 -0,1]	1,77
16	H3	105	Shape	[0,3 -0,3]	3,83
17	H1	117	Noise	[1000 1200] a=0.02	2,87
18	H1	117	Noise	[100 800] a=0.02	3,1
19	H1	117	Noise	[1000 1200] a=0.05	3,1
20	H1	117	Noise	[100 800] a=0.05	4

Cuadro 4.1: Resultados generales de la encuesta para Vocales Sostenidas (H: Hablante, P: Puntaje promedio obtenido para cada vocalización)

Algo que llama la atención, es que las voces sin perturbar (1,4,7), en algunos casos obtuvieron puntajes bastante cercanos al nivel de "Perturbación Leve", incluso algunas voces perturbadas, obtuvieron puntajes más bajos que voces naturales (10,12,15). Esto puede deberse a diferentes causas, se debe tomar en cuenta que las grabaciones fueron realizadas a 8[kHz], con un teléfono celular, y en un ambiente no controlado, todo esto afecta la calidad de las señales. Otro factor a tener en cuenta, es que una vocal sostenida, no es una vocalización que las personas hagan normalmente, por lo cual no estamos tan acostumbrados a escuchar ese tipo de sonidos prolongados. Finalmente, cada persona realizó el experimento en un lugar diferente, con sus propios dispositivos electrónicos, lo cual constituye otro factor que podría alterar la percepción de los audios.

En las vocalizaciones generadas, se utilizó un solo hablante para cada perturbación, lo cual permite comparar los datos obtenidos con la voz sin perturbar, y obsevar que ocurrió al variar los parámetros para una misma vocalización.

A continuación, se presentarán los resultados agrupados por perturbación, para facilitar su análisis.

4.4.1 Filtrados

	Filtrado		
	Sin filtrar	f_c **= 500**	f_c **= 1000**
Pasabajos	1,63	3,5	3,7
Pasaaltos	1,53	3,1	2,97

Cuadro 4.2: Resultados para las perturbaciones correspondientes a filtros

Para ambos filtros, el resultado a penas tuvo variaciones al cambiar la frecuencia de corte. Se observa que la evaluación general es que están entre una perturbación media a alta. El filtro pasaaltos fue calificado ligeramente menos perturbado.

4.4.2 Tremor

	Tremor		
	Sin Tremor	**amp = 0,1**	**amp = 0,2**
f=50	1,93	2,17	3,23
f=174	1,93	1,9	-
f=200	1,93	2,73	-

Cuadro 4.3: Resultados para la perturbación Tremor

Para la perturbación tremor, es interesante destacar que la frecuencia fundamental de la vocalización es de 174[Hz], y que al introducir esta frecuencia de perturbación, el resultado fue incluso más bajo que para la señal sin perturbar. Un resultado similar, pudo observarse en los gráficos de CPP y HNR presentado en la sección anterior. Para las otras 2 perturbaciones con 0,1 de amplitud, parece ser que una frecuencia más alta es más detectable que una frecuencia baja.

Otro punto a tener en cuenta es que al aumentar la amplitud, la vocalización fue percibida considerablemente más perturbada.

4.4.3 NBFM

	NBFM		
	Sin NBFM	k_f **= 0,1**	k_f **= 0,2**
f**=50**	1,53	2,17	3,23
f**=109**	1,53	1,9	-
f**=125**	1,53	2,73	-

Cuadro 4.4: Resultados para la perturbación NBFM

Para NBFM se obtuvieron resultados muy similares a Tremor, al utilizar la frecuencia fundamental de la vocalización natural, la perturbación fue percibida como más leve que las otras 2, pero con un aumento respecto a la voz sin perturbar. También se observa que al aumentar el factor k_f aumenta considerablemente el puntaje.

4.4.4 Noise

	Noise		
	Sin Noise	**amp=0,02**	**amp=0,05**
[100 - 800]	1,63	3,1	4
[1000 - 1200]	1,63	2,87	3,1

Cuadro 4.5: Resultados para la perturbación Noise

La perturbación Noise, en todos los casos fue bastante notoria en comparación con la voz sin perturbar. El ruido en el rango $[1000 - 1200]$ obtuvo un resultado ligeramente menor, lo cual podría atribuirse a que corresponde a una banda más angosta, o a que se encuentra en un rango de frecuencias muy por encima de los modos fundamentales de la voz.

4.4.5 Shape

	Shape		
	Sin Shape	**amp=0,1**	**amp=0,3**
[1 , -1]	1,53	1,77	3,83

Cuadro 4.6: Resultados para la perturbación Shape

Para la perturbación Shape, se obtuvo un resultado para $amp = 0,1$, a penas por encima de la voz natural. Al aumentar la amplitud, se observó una mayor percepción de la perturbación.

4.4.6 Frases de "El Abuelo"

	Frases				
N°	**H**	f_0	**Pert.**	**Parámetros**	**P**
21	H1	117	-	-	1,43
22	H1	117	P.altos	f_c=500	2,53
23	H1	117	P.bajos	f_c=1000	3,1
24	H4	180	-	-	1,53
25	H4	180	Tremor	amp=0.2 f=10	3,17
26	H4	180	Tremor	amp=0.2 f=100	3,5
27	H4	180	Shape	0.05 -0.05	3,23
28	H4	180	NBFM	f_c=10 kf=0.3	3
29	H1	117	P.altos	f_c=1000	3,47
30	H1	117	P.bajos	f_c=500	3,23

Cuadro 4.7: Resultados generales de la encuesta para Frases (H: Hablante, P: Puntaje promedio obtenido para cada vocalización)

Los resultados obtenidos para la evaluación de frases, muestran que los oyentes asignaron puntajes inferiores a las voces naturales en comparación con las perturbadas, con una diferencia más marcada que en el experimento de vocales sostenidas. Las frases perturbadas, fueron descubiertas por los oyentes sin problemas en todos los audios. Por una parte, estamos mucho más acostumbrados a escuchar frases que vocales sostenidas en nuestro lenguaje cotidiano, por lo que es posible distinguir distorsiones más sutiles. Otro factor que podría haber influido, es que es mucho más complejo perturbar frases, ya que estas contienen una amplia gama de sonidos, silencios, y cambios abruptos, que en algunos casos, el algoritmo no es capaz de procesar correctamente.

Al aplicar el algoritmo de perturbación para habla, este es capaz de diferenciar de manera correcta los silencios de las partes con habla. Lo cual funciona bastante bien para eliminar el ruido de ambiente que ocurre en los instantes en que la persona no está hablando, pero al momento de diferenciar entre voz con contenido vocal y sin contenido vocal, el algoritmo es capaz de hacerlo, pero en algunos casos, llega a tomar partes de algunas consonantes sin vocalización que están adyacentes a partes con vocalización.

Al introducir perturbaciones, el algoritmo es capaz de introducirlas a cada segmento identificado como voz modal, y también de realizar una resíntesis, pero en algunas ocasiones, dependiendo de la perturbación aplicada, se pueden tener problemas al concatenar los segmentos del texto, lo cual genera ruidos artificiales que pueden llegar a ser bastante evidentes.

Se han obtenido mejores resultados para una lectura pausada del texto, y con voz

clara. También es importante elegir las perturbaciones correctas para cada hablante de acuerdo a las característica de su voz, de manera que el texto resintetizado tenga un sonido natural.

4.5 Resultados tiempo de procesamiento

Usando los comandos de MATLAB *tic* y *toc*, se ha medido el tiempo que toman diferentes procesamientos, para segmentos de audio de duración variable, obteniendo la siguiente (Ver Cuadro 4.8). Se entrega el tiempo que tomó procesar 1 segundo de la vocalización de cada hablante, primero solo para la obtención del residuo, luego solo para la síntesis, y el tiempo que tomó el proceso completo para cada una de las perturbaciones. H1, H2, H3 y H4, corresponden a los 4 hablantes que realizaron la vocal "a" sostenida. Además se han calculado los tiempos para los hablantes H1 y H4, los cuales leyeron el texto "El Abuelo".

	Tiempo para 1 [s] de señal			
Cómputo	**H1**	**H2**	**H3**	**H4**
Obtención del residuo	0,211	0,260	0,245	0,363
Resíntesis de la señal	0,017	0,032	0,020	0,025
NBFM	0,259	0,320	0,293	0,416
Tremor	0,229	0,302	0,272	0,400
Pasa-altos	0,235	0,331	0,325	0,395
Pasa-bajos	0,235	0,320	0,293	0,395
Noise	0,765	1,060	1,397	1,499
Shape	0,267	0,320	0,315	0,416
El Abuelo	0,32	-	-	0,45

Cuadro 4.8: Tabla tiempos que tarde el método en procesar 1 [s] de señal

Observando la tabla, es posible notar que la obtención del residuo es lo que consume la mayor cantidad de tiempo, siendo entre un 70 % y un 90 %. La introducción de la perturbación y la resíntesis de las señales consumen aproximadamente el 10 % del tiempo cada una. Esto se cumple para todas las perturbaciones, excepto para la perturbación "noise" que tarda bastante en generar los vectores de ruido aleatorio de la perturbación.

Otra cosa que puede apreciarse en la tabla es que H2 y H4, tienen tiempos más largos que H1 y H3, esto se debe a que H2 y H4 tienen frecuencias fundamentales más altas, lo que hace que él método a utilizar deba procesar por separado más ciclos de la glotis.

Para los segmentos del texto "El Abuelo", en los 2 casos utilizados, se tuvo un mayor tiempo de procesamiento, esto se debe a todo el pre-procesamiento que hay que

aplicarle a la señal antes de aplicar la metodología para introducir perturbaciones. A pesar de este aumento en el tiempo, sigue tomando menos de 1[s] procesar 8000 muestras de la señal.

Finalmente, es un buen indicio que el tiempo en todas las perturbaciones excepto "noise" sea menor a un segundo, ya que esto quiere decir que es factible realizar el procesamiento en tiempo real.

Capítulo 5

Discusión y Conclusiones

En esta sección, se procederá a comentar los resultados obtenidos, en relación a los objetivos planteados. También se comentarán las lecciones aprendidas, y el trabajo futuro que podría hacerse en base a los descubrimientos e interrogantes que surgieron durante este trabajo.

5.1 Respecto a los Experimentos de feedback auditivo

Desde que Lombard realizó el primer experimento de feedback auditivo, las herramientas de la ingeniería, han sido fundamentales para el desarrollo de esta área del conocimiento. Los diferentes descubrimientos que se han tenido, han venido emparejados con la proliferación de nuevas tecnologías, especialmente en el área de la electrónica.

En la actualidad, contamos con grandes avances en el área del procesamiento de señales digitales, los cuales hacen posible la manipulación de señales con bajos tiempos de latencia. Esto permite la realización de experimentos altamente específicos, que pongan a prueba los mecanismos de feedback auditivo de maneras que nunca antes habían sido posibles.

Considerando los tiempos de cómputo obtenidos en MATLAB, y tomando en cuenta que el computador utilizado no está especializado para este tipo de tareas, se cree que es muy posible implementar este experimento en tiempo real, y obtener menores tiempos de latencia.

Tomando en cuenta los resultados de la encuesta de percepción auditiva, es posible notar que las personas son capaces de distinguir muchas de las perturbaciones conscientemente, pero en algunos casos, a pesar de que las voces estaban interveni-

das, muchos oyentes no lo notaron, lo cual nos indica que es posible ajustar las perturbaciones para que la voz siga pareciendo natural.

Estos 2 hallazgos, hacen que parezca bastante factible a nivel de procesamiento de señales, llevar a cabo un experimento en donde se aplique la metodología propuesta.

5.2 Cumplimiento de los Objetivos

Tomando como referencia el modelo fuente-filtro, se ha diseñado e implementado en MATLAB, una metodología para descomponer y resintetizar señales de voz, la cual utiliza métodos de filtrado inverso basados en LPC.

Se logró implementar la descomposición y resíntesis de señales con 3 métodos de filtrado diferentes, los cuáles fueron comparados considerando sus ventajas y desventajas. De los 3 métodos, PSIAIF, se ubica como un buen candidato para realizar un experimento en tiempo real, debido a la facilidad que ofrece para su implementación, su bajo costo de procesamiento, y el hecho de que utiliza ventanas de un ciclo de la glotis de duración, lo cual no introduce tanto retardo en la adquisición de los datos. El método QCP, permite obtener mejores resultados en la descomposición, pero requiere un mayor tiempo de procesamiento de las señales, y también ventanas de trabajo que incluyan varios ciclos de la glotis, lo que lo hace menos adecuado para un experimento en tiempo real.

Los 3 métodos de descomposición de las señales, permiten una resíntesis fácil de llevar a cabo, que consiste en filtrar y concatenar las ventanas de la señal. Este proceso es muy sencillo cuando no se introducen perturbaciones, pero presenta algunas dificultades dependiendo de la perturbación aplicada que se traducen en ruidos no naturales en la señal resintetizada. Este problema se ha resuelto realizando algunas mejoras a la función de resíntesis, que lograron un sonido más natural.

Las perturbaciones propuestas, son bastante sencillas, lo cual significa un bajo costo de procesamiento, y por ende, un retardo temporal pequeño, esto ayudaría en la implementación del experimento en tiempo real. Además, las perturbaciones funcionan en base a diversos parámetros, que permiten ajustarlas, y también combinarlas para producir perturbaciones más complejas. El funcionamiento modular de la metodología propuesta, permite diseñar nuevas perturbaciones y agregarlas al algoritmo sin mayores complicaciones. Cabe mencionar que las perturbaciones propuestas no responden a un estudio del funcionamiento de las cuerdas vocales, ni de los mecanismos de control motor de estas, por lo que no se cuenta con una hipótesis del impacto que estas tendrían en un experimento de feedback auditivo. Se recomienda realizar un estudio más profundo de estos temas a la hora de proponer nuevas perturbaciones.

Finalmente, se han utilizado 3 medidas acústicas de la calidad vocal, y una encuesta de percepción auditiva para evaluar el impacto de las perturbaciones en las voces. Se ha podido observar que las perturbaciones son capaces de alterar la calidad vocal, y que estas alteraciones, además de estar relacionadas con los parámetros específicos de cada perturbación, también guardan relación con las características propias de cada voz. Otra cosa que pudo observarse, es que existe una relación entre el comportamiento de las medidas de calidad vocal, y la percepción que tuvieron los oyentes sobre las voces.

De acuerdo a los resultados de las evaluaciones acústicas y perceptuales de la calidad vocal, parece una buena idea, generar perturbaciones que guarden relación con la voz a perturbar, y que se adapten a las características de cada voz, en este sentido, la forma en que opera la perturbación Shape, parece adecuada, ya que altera cada ciclo en función de su duración individual.

En base a lo mencionado, puede concluirse que se tiene un buen nivel de logro de los objetivos planteados al comienzo de este trabajo, pero aún hace falta mejorar diversos aspectos de esta metodología, para poder implementarla en un experimento.

5.3 Trabajo Futuro

Para continuar con el desarrollo de la metodología, se recomienda realizar un estudio minucioso del funcionamiento de las cuerdas vocales, tanto sanas como patológicas, con el fin de poder proponer perturbaciones que tengan relación con algún aspecto del funcionamiento de los pliegues vocales que pudiera activar algún mecanismo compensatorio específico.

También se recomienda mejorar algunos aspectos técnicos de la adquisición de las señales de prueba, con mejores micrófonos, y en ambientes con ruido controlado, que aseguren que la señal medida corresponda únicamente a voz. Por otro lado, sería interesante realizar una evaluación perceptual de las perturbaciones con fonoaudiólogos expertos, utilizando audífonos de buena calidad, para tener una apreciación técnica más detallada.

Finalmente, sería necesario hacer un estudio técnico de las tarjetas de procesamiento de señales disponibles en la actualidad y de sus capacidades para la implementación de un experimento de este tipo. También es necesario trabajar en mejorar algunos aspectos del algoritmo propuesto, que permitan obtener mejores resultados, y una mayor eficiencia en los tiempos de procesamiento.

Bibliografía

[1] Dietrich M, Verdolini Abbott K, Gartner-Schmidt J, Rosen CA. The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice. Journal of voice: official journal of the Voice Foundation. 2008;22:472–488. [PubMed] [Google Scholar]

[2] Bhattacharyya, N. (2014). The prevalence of voice problems among adults in the United States. The Laryngoscope, 124(10), 2359–2362. https://doi.org/10.1002/lary.24740

[3] Lane, H., Tranel, B. (1971). The Lombard Sign and the Role of Hearing in Speech. Journal of Speech and Hearing Research, 14(4), 677–709.

[4] Tourville, J. A., Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. Language and Cognitive Processes, 26(7), 952–981.

[5] Mazziotta J, Toga A, Evans A, Fox P, Lancaster J, Zilles K, et al. A four-dimensional probabilistic atlas of the human brain. Journal of the American Medical Informatics Association. 2001;8(5):401–430. [PMC free article] [PubMed] [Google Scholar]

[6] Slabbekoorn, H., Peet, M. (2003). Birds sing at a higher pitch in urban noise. Nature, 424(6946), 267.

[7] Oller, D.K., and Eilers, R.E. (1988). "The role of audition in infant babbling," Child Dev. 59, 441-449.

[8] Lane, H., and Webster, J.W. (1991). "Speech deterioration in postlingually deafened adults," J. Acoust Soc. Am. 89, 859-866.

[9] Waldstein, R.S. (1990). "Effects of postlingual deafness on speech production - Implications for the role of auditory feedback." J. Acoust. Soc. Am. 88, 2099-2114.

[10] Toyomura A, Koyama S, Miyamaoto T, Terao A, Omori T, Murohashi H. Neural correlates of auditory feedback control in human. Neuroscience. 2007;146(2):499–503. [PubMed]

[11] Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. Neuroimage. 2008;39(3):1429–1443. [PMC free article] [PubMed]

[12] Bárány, R. (1908). Noise Apparatus for the Detection of Unilateral Deafness. The Journal of Laryngology, Rhinology, and Otology, 23(7), 363–364.

[13] Black, J. W. (1951). The Effect Of Delayed Side-Tone Upon Vocal Rate And Intensity. Journal of Speech and Hearing Disorders, 16(1), 56–60.

[14] Badian, M., Appel, E., Palm, D., Rupp, W., Sittig, W., Taeuber, K. (1979). Standardized mental stress in healthy volunteers induced by delayed auditory feedback (DAF). European Journal of Clinical Pharmacology, 16(3), 171–176.

[15] Van Borsel, J., Reunes, G., Van den Bergh, N. (2003). Delayed auditory feedback in the treatment of stuttering: clients as consumers. International Journal of Language Communication Disorders, 38(2), 119–129.

[16] Soderberg, G. A. (1969). Delayed Auditory Feedback and the Speech of Stutterers: A Review of Studies. Journal of Speech and Hearing Disorders, 34(1), 20–29.

[17] SpeechEasy - Equipos:
https://speecheasy.com/devices/

[18] Stuart, A., Kalinowski, J., Rastatter, M. P., Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. The Journal of the Acoustical Society of America, 111(5), 2237.

[19] Yates, A. J. (1963). Delayed auditory feedback. Psychological Bulletin, 60(3), 213–232.

[20] Elman, J. L. (1981). Effects of frequency-shifted feedback on the pitch of vocal productions. The Journal of the Acoustical Society of America, 70(1), 45–50.

[21] Artículo sobre Lexicon Varispeech:
https://valhalladsp.com/2010/05/06/the-first-digital-pitch-shifter-lexicon-varispeech

[22] Burnett, T. A., Freedland, M. B., Larson, C. R., Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. The Journal of the Acoustical Society of America, 103(6), 3153–3161.

[23] Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., and Hain, T. C. (2001). Comparison of voice F0 responses to pitch-shift onset and offset conditions. J. Acoust. Soc. Am. 110, 2845–2848.

[24] Hawco, C. S., and Jones, J. A. (2009). Control of vocalization at utterance onset and mid-utterance: different mechanisms for different goals. Brain Res. 1276, 131–139

[25] Natke, U., Grosser, J., Kalveram, K. T. (2001). Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons. Journal of Fluency Disorders, 26(3), 227–241.

[26] Titze, I. R., Baken, R. J., Bozeman, K. W., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R. D., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noé, H., Scherer, R. C., Smith, J. R., Story, B. H., ... Wolfe, J. (2015). Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization. The Journal of the Acoustical Society of America, 137(5), 3005–3007.

[27] Peterson, G. E., Barney, H. L. (1951). Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America, 23(1), 148.

[28] Bradlow, A. R. (1995). A comparative acoustic study of English and Spanish vowels. Journal of the Acoustical Society of America, 97(3), 1916–1924. https://doi.org/10.1121/1.412064

[29] Houde, J. F. (1997). Sensorimotor Adaptation in Speech Production. Thesis (Ph. D.) –Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences.

[30] Houde, J. F. (1998). Sensorimotor Adaptation in Speech Production. Science, 279(5354), 1213–1216.

[31] Villacorta, V. M. (2006). Sensorimotor Adaptation to Perturbations of Vowel Acoustics and its Relation to Perception. Thesis (Ph. D.) –Massachusetts Institute of Technology, Harvard-MIT Division of Health Sciences and Technology.

[32] Purcell, D. W., Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. The Journal of the Acoustical Society of America, 119(4), 2288–2297. doi:10.1121/1.2173514

[33] Garber, S. R., Moller, K. T. (1979). The Effects of Feedback Filtering on Nasalization in Normal and Hypernasal Speakers. Journal of Speech, Language, and Hearing Research, 22(2), 321–333.

[34] Siegel, G. M., Pick, H. L. (1974). Auditory feedback in the regulation of voice. The Journal of the Acoustical Society of America, 56(5), 1618–1624. doi:10.1121/1.1903486

[35] Bauer, J. J., Mittal, J., Larson, C. R., Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. The Journal of the Acoustical Society of America, 119(4), 2363–2371.

[36] Klein, E., Brunner, J., Hoole, P. (2019). The relevance of auditory feedback for consonant production: The case of fricatives. Journal of Phonetics, 77, 100931. doi:10.1016/j.wocn.2019.100931

[37] Alku, P., Vilkman, E., Laine, U.K.: Analysis of glottal waveform in different phonation types using the new IAIF-method. In: Proc. 12th Int. Congress Phonetic Sciences, vol. 4, pp. 362–365 (1991).

[38] A comparative study of glottal source estimation techniques Thomas Drugman, Baris Bozkurt, Thierry Dutoit.

[39] Priyanko Mitra, "Glottography for the Diagnosis of Vocal Disorders"

[40] Fant, G., Liljencrants, J., Lin, Q., 1985a. A four-parameter model of glottal flow. STL-QPSR 26 (4), 1–13.

[41] F. Itakura, "Minimum Prediction Residual Principle Applied To SpeechRecognition", IEEE Trans. Acoust. , Speech and Signal Process. , Vol. ASSP-23, Feb. 1975, pp. 67-72.

[42] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," Speech Commun., vol. 12, no. 1, pp. 69–81, 1993.

[43] Benesty, J., Sondhi, M. M., Huang, Y., & Greenberg, S. (2009). Springer Handbook of Speech Processing. The Journal of the Acoustical Society of America, 126(4), 121–134.

[44] Hyung-Suk Kim. Linear Predictive Coding is All-Pole Resonance Modeling. Center for Computer Research in Music and Acoustics, Stanford University.

[45] Drugman, T., Dutoit, T., 2009. Glottal closure and opening instant detection from speech signals. Proc. Interspeech.

[46] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana (2014). Glottal source processing: From analysis to applications. Computer Speech & Language, 28(5), pp. 1117–1138.

[47] P. Alku (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering Speech Communication, 11(2-3), pp. 109–118.

[48] Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo Alku (2014). “Quasi Closed Phase Glottal Inverse Filtering Analysis With Weighted Linear Prediction”. IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 3

[49] DEFINING AND MEASURING VOICE QUALITY. Jody Kreiman, Diana Vanlancker-Sidtis, & Bruce Gerratt. Division of Head and Neck Surgery, School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA.

[50] Angel Gordillo, L. F. (2018). Hitos de la evaluación perceptual auditiva de la voz: ¿hay evidencia? Areté issn-l:1657-2513, 18 (2), 65-74. Obtenido de: https://revistas.iberoamericana.edu.co/index.php/arete/ article/view/1413

[51] Hirano M. Clinical examination of voice. New York: Springer Verlag, 1981:81-4.

[52] Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. American Journal of Speech Language Pathology, 18, 124-132.

[53] Perceptual Assessment of Voice Quality: Past, Present, and Future. Jody Kreiman and Bruce R. Gerratt. Department of Head and Neck Surgery, University of California School of Medicine Los Angeles, CA

[54] Teixeira, J.P., Oliveira, C., Lopes, C. 2013. Vocal Acoustic Analysis–Jitter, Shimmer and HNR Parameters. Procedia Technology, 9, 1112-1122.]

[55] Teixeira, J. P.; Ferreira, D.; Carneiro, S. Análise acústica vocal - determinação do Jitter e Shimmer para diagnóstico de patalogias da fala. In 6º Congresso Luso-Moçambicano de Engenharia. Maputo, Moçambique, 2011.

[56] Murphy, P. and Akande, O. Cepstrum-Based Estimation of the Harmonics-to noise Ratio for Synthesized and Human Voice Signals. In Nonlinear Analyses and Algorithms for Speech Processing. Barcelona, LNAI 3817, Springer, 2005.

[57] Ferrand C. Harmonics-to-noise ratio: an index of vocal aging. J Voice. 2002;16:480–487.

[58] Patel RR, Awan SN, Barkmeier-Kraemer J, et al. Recommended protocols for instrumental assessment of voice: American speech- language-hearing association expert panel to develop a protocol for instrumental assessment of vocal function. Am J Speech-Language Pathol. 2018;27:887–905.

[59] Fraile R, Godino-Llorente JI. Cepstral peak prominence: a comprehensive analysis. Biomed Signal Process Control. 2014;14:42–54. https://doi.org/10.1016/j.bspc.2014.07.001.

[60] Heman-Ackah YD, Heuer RJ, Michael DD, et al. Cepstral peak prominence: a more reliable measure of dysphonia. Ann Otol Rhinol Laryngol. 2003;112:324–333.

[61] Ferrer Riesgo, C. A., Nöth, E. (2020). What Makes the Cepstral Peak Prominence Different to Other Acoustic Correlates of Vocal Quality? Journal of Voice, 34(5), 806.e1-806.e6.

[62] Campbell, N., Beckman, M. E. (1997). Accent, stress, and spectral tilt. The Journal of the Acoustical Society of America, 101(5), 3195.

[63] Hillenbrand, J., Houde, R. A. (1996). Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. Journal of Speech, Language, and Hearing Research, 39(2), 311–321.

[64] Balasubramanium, R. K., Bhat, J. S., Fahim, S., Raju, R. (2011). Cepstral Analysis of Voice in Unilateral Adductor Vocal Fold Palsy. Journal of Voice, 25(3), 326–329.

[65] Radish Kumar, B., Bhat, J. S., Prasad, N. (2010). Cepstral Analysis of Voice in Persons With Vocal Nodules. Journal of Voice, 24(6), 651–653.

[66] Bowen LK, Hands GL, Pradhan S, Stepp CE. Effects of Parkinson's disease on fundamental frequency variability in running speech. Journal of medical speech-language pathology. 2013;21(3):235

[67] Deller JR, Proakis JG, Hansen JHL. Discrete-time processing of speech signals. New York: Maxwell McMillan, 1993.

[68] Ackermann H, Ziegler W. Acoustic analysis of vocal instability in cerebellar dysfunctions. Ann Otol Rhinol Laryngol 1994;103(2):98–104.

[69] European Commission, DG Research and Innovation. Functional Magnetic Resonance Imaging: Understanding the technique and addressing its ethical concerns with a future perspective.

[70] Glover G.H. Overview of functional magnetic resonance imaging. Neurosurg. Clin. N. Am. 2011;22:133–139. [PMC free article] [PubMed] [Google Scholar]

[71] Jason W. Bohland; Frank H. Guenther (2006). An fMRI investigation of syllable sequence production. , 32(2), 821–841. doi:10.1016/j.neuroimage.2006.04.173

[72] Watkins K., Patel N., Davis S., Howell P. Brain activity during altered auditory feedback: an FMRI study in healthy adolescents. Neuroimage. 2005;26(Supp 1):304. [PMC free article] [PubMed] [Google Scholar]

[73] Fu, C. H.Y. (2005). An fMRI Study of Verbal Self-monitoring: Neural Correlates of Auditory Verbal Feedback. Cerebral Cortex, 16(7), 969–977. doi:10.1093/cercor/bhj039

[74] EEG - Mayo Clinic: https://www.mayoclinic.org/es-es/tests-procedures/eeg/about/pac-20393875

[75] Scheerer, N. E., Jones, J. A. (2018). The Role of Auditory Feedback at Vocalization Onset and Mid-Utterance. Frontiers in Psychology.

[76] Kittilstved, T., Reilly, K. J., Harkrider, A. W., Casenhiser, D., Thornton, D., Jenson, D. E., Hedinger, T., Bowers, A. L., Saltuklaroglu, T. (2018). The Effects of Fluency Enhancing Conditions on Sensorimotor Control of Speech in Typically Fluent Speakers: An EEG Mu Rhythm Study. Frontiers in Human Neuroscience.

[77] Oleg Korzyukov; Laura Karvelis; Roozbeh Behroozmand; Charles R. Larson (2012). ERP correlates of auditory processing during automatic correction of unexpected perturbations in voice auditory feedback. , 83(1), 0–78.

[78] Scheerer, Nichole E.; Jones, Jeffery A. (2017). Detecting our own vocal errors: An event-related study of the thresholds for perceiving and compensating for vocal pitch errors. Neuropsychologia, S0028393217304724–. doi:10.1016/j.neuropsychologia.2017.12.007

[79] Hixon, T. J., Weismer, G. y Hoit, J. D., Preclinical speech science, Plural Publishing, 2008.

[80] Gray, SD. Cellular physiology of the vocal folds. The Otolaryngologic Clinics of North America. V.30. No. 4. pp 679-98. Aug 2000.

[81] Sorenson, D. N. (1989). A fundamental frequency investigation of children ages 6–10 years old. Journal of Communication Disorders, 22(2), 115–123.

[82] National Institute of Deafness and other Communication Disorders (NIDCD). "What Is Voice? What Is Speech? What Is Language?" https://www.nidcd.nih.gov/health/what-is-voice-speech-language

[83] Encuesta de Percepción Auditiva- Felipe Rodríguez https://www.questionpro.com/t/AR2ByZkcsQ