UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Aportes al análisis de perturbaciones desde el aprendizaje maquinal y el análisis tiempo-frecuencia

Juan Manuel Miramont

FICH
FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE MÉTODOS COMPUTACIONALES

sinc(*i*)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES, SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2021**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

APORTES AL ANÁLISIS DE PERTURBACIONES DESDE EL APRENDIZAJE MAQUINAL Y EL ANÁLISIS TIEMPO-FRECUENCIA

Juan Manuel Miramont

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERIA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2021

Secretaría de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje “El Pozo”, S3000, Santa Fe, Argentina

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigaciones en Señales, Sistemas e Inteligencia Computacional

APORTES AL ANÁLISIS DE PERTURBACIONES DESDE EL APRENDIZAJE MAQUINAL Y EL ANÁLISIS TIEMPO- FRECUENCIA

Juan Manuel Miramont

Lugar de Trabajo:

LSyDNL
Laboratorio de Señales y Dinámicas No Lineales
Facultad de Ingeniería
Universidad Nacional de Entre Ríos

Director:
Gastón Schlotthauer — IBB, FI-UNER / CONICET

Co-director:
César Martínez — sinc(i), FICH-UNL

Jurado Evaluador:
Humberto Maximiliano Torres — LIS, INIGEM, CONICET
Sebastián Rodrigo Vanrell — sinc(i), FICH-UNL / CONICET
Juan Carlos Gómez — CIFASIS , UNR / CONICET

2021

ACTA DE EVALUACIÓN DE TESIS DE DOCTORADO

En la sede de la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, a los diez días del mes de septiembre del año dos mil veintiuno, se reúnen en forma virtual los miembros del Jurado designado para la evaluación de la Tesis de Doctorado en Ingeniería titulada ***"Aportes al análisis de perturbaciones desde el aprendizaje maquinal y el análisis tiempo-frecuencia"***, desarrollada por el Bioing. Juan Manuel MIRAMONT, DNI Nº 37.279.835. Ellos son: Dr. Juan Carlos Gómez, Dr. Humberto Torres y Dr. Sebastián Vanrell.

La Presentación oral y defensa de la Tesis se efectúa bajo la modalidad virtual según lo establecido por resolución de Rector N° 529/20 y resolución del Consejo Directivo N° 015/20.

Luego de escuchar la Defensa Pública y de evaluar la Tesis, el Jurado resuelve:

La presentación fue muy clara y organizada, así como el documento de tesis presentado. El tesista respondió con solvencia las preguntas del jurado. Los resultados alcanzados han sido notablemente superiores a los del estado del arte y se han comunicado en publicaciones reconocidas internacionalmente y de buen factor de impacto.
Los resultados son promisorios en el sentido que pueden tener transferencia a la práctica clínica fonoaudiológica.

Por lo tanto, el Jurado, por unanimidad, resuelve calificar la tesis con nota 10 (Diez) Sobresaliente.

Sin más, se da por finalizado el Acto Académico con la firma de los miembros del Jurado al pie de la presente. ----------

Dr. Juan Carlos Gómez	Dr. Humberto Torres	Dr. Sebastián Vanrell

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS * SANTA FE *

Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 10 de Septiembre de 2021.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Aportes al análisis de perturbaciones desde el aprendizaje maquinal y el análisis tiempo-frecuencia"***, desarrollada por el Bioing. Juan Manuel MIRAMONT, en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Juan Carlos Gómez	Dr. Humberto Torres	Dr. Sebastián Vanrell

Santa Fe, 10 de Septiembre de 2021.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

Dr. César Martínez Codirector de Tesis	Dr. Gastón Schlotthauer Director de Tesis

FACULTAD DE INGENIERIA Y CIENCIAS HIDRICAS * SANTA FE * R.A.
Dr. JOSE LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

Declaración del Autor

Esta Tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería, mención Inteligencia Computacional, Señales y Sistemas, ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el reglamento de la mencionada Biblioteca.

Citaciones breves de esta Tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de ese manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra.

Dedicado a mis padres,
Patricia y Guillermo;
y a mis abuelos.

"El hombre se compone de lo que tiene «y de lo que le falta». Si usa sus dotes intelectuales en largo y desesperado esfuerzo no es simplemente porque las tiene, sino, al revés, porque se encuentra menesteroso de algo que le falta y a fin de conseguirlo moviliza, claro está, los medios que posee. El error radicalísimo de todas las teorías del conocimiento ha sido no advertir la inicial incongruencia que existe entre la necesidad que el hombre tiene de conocer y las «facultades» que cuenta para ello. Sólo Platón entrevió que la raíz del conocer, diríamos, su sustancia misma, está precisamente en la insuficiencia de las dotes humanas, que está en el hecho terrible de que el hombre «no sabe». Ni Dios ni la bestia tiene esta condición. Dios sabe todo y por eso no conoce. La bestia no sabe nada y por eso tampoco conoce. Pero el hombre es la insuficiencia viviente, el hombre necesita saber, percibe desesperadamente que ignora. Esto es lo que conviene analizar. ¿Por qué al hombre le duele su ignorancia, cómo podía dolerle un miembro que nunca hubiese tenido?"

José Ortega y Gasset, ¿Qué es filosofía?.

Agradecimientos

Quisiera agradecer a mi director, Gastón Schlotthauer, por darme la oportunidad de satisfacer mi curiosidad, por corregirme cuando me equivoqué y por aconsejarme cuando necesité alguna certeza. A mi codirector, César Martínez, por acompañar este proyecto y hacerlo posible. A Marcelo Colominas por compartir y discutir conmigo sus ideas y mostrarme nuevos horizontes, y a Gabriel Alzamendi, por estar siempre disponible para una consulta.

Agradezco especialmente a Juan Felipe Restrepo, Juliana Codino y María Cristina Jackson Menaldi por su indispensable colaboración.

Agradezco a todos los miembros del Laboratorio de Señales y Dinámicas no Lineales y de la cátedra de Funciones de Variable Compleja, que me abrieron las puertas a la docencia en la Universidad Pública. Particularmente a Ariel Stassi y Victoria Peterson, por sumergirnos juntos en la aventura de la virtualidad... y no ahogarnos.

A Ramiro Casal por transitar conmigo el *Camino del Becario*, y hacerlo más ameno mientras nos quejábamos del sistema. A Meri Cuaranta, por el generosísimo gesto de su parte de compartirme el caloventor en invierno.

A Marcos, Carla, Juan y Franco por la enorme paciencia y comprensión. A la *Biofamily*, por acompañarme todos estos años entrerrianos.

Finalmente, quisiera agradecer a mi familia. A mis padres, a quienes les debo el privilegio de poder seguir el camino que hoy recorro. A mis hermanas, por este club de tres que formamos. Y a mi abuela, por sus manos colmadas de ternura.

Índice general

Índice de figuras

Índice de tablas

Resumen

Las perturbaciones son pequeñas variaciones ciclo a ciclo en diferentes parámetros de la señal que están siempre presentes en las señales de voz correspondientes a vocales sostenidas. Dado que existe evidencia de que diversas patologías del aparato fonador, y de otros sistemas asociados, afectan la magnitud de las perturbaciones, resulta de interés cuantificarlas. Este es uno de los objetivos de un área conocida como *análisis de perturbaciones*, que se vale de *medidas de perturbación* para estimar la magnitud de estas variaciones de corto plazo. Dichas medidas se basan en dos hipótesis fundamentales:1) la señal debe ser aproximadamente periódica y 2) el parámetro estudiado debe permanecer constante durante la duración de cada ciclo. El objetivo de esta tesis doctoral es mejorar la aplicación del análisis de perturbaciones, para lo cual se estudiaron dos situaciones problemáticas, cada una de ellas relacionadas con las hipótesis mencionadas.

En primer lugar, se abordaron las dificultades de la clasificación de señales en tres tipos (1, 2 y 3) de acuerdo a la periodicidad de la señal. Esta clasificación fue ideada para evitar que las medidas de perturbación sean empleadas sobre señales que no cumplan con la primera hipótesis mencionada anteriormente. Según la definición de cada tipo, sólo el tipo 1 abarca a aquellas señales aproximadamente periódicas, por lo que las medidas de perturbación deberían aplicarse únicamente sobre señales en esta categoría. La clasificación en tres tipos, también llamada *tipificación*, es ampliamente utilizada en la clínica y en la investigación de la salud vocal. No obstante, la tipificación es una tarea subjetiva, ya que se basa en la percepción de los especialistas clínicos, lo que genera cierta variación interprofesional.

En segundo lugar, se estudió la incapacidad del jitter relativo para estimar niveles altos de perturbación del periodo fundamental. El jitter relativo es una de las medidas de perturbación más utilizadas en el ámbito clínico, y consiste en la razón entre la perturbación promedio y la duración promedio de los ciclos de la señal. Con ese fin, requiere la estimación previa de una sucesión con la duración de cada uno de dichos ciclos, también llamada serie de periodos. Se ha observado que esta medida subestima el verdadero valor de jitter relativo cuando éste supera un umbral de entre $5\,\%$ y $8\,\%$.

Con el fin de aumentar la objetividad de la clasificación de señales en tres tipos, se propuso la caracterización de un sistema capaz de clasificar de manera automática las señales de voz mediante una estrategia de reconocimiento de patrones, basada en la extracción de características empleadas en la práctica clínica y máquinas de vectores de soporte lineales. Se etiquetaron más de 1200 señales provenientes de las bases de datos *Massachusetts Eye and Ear Infirmary* (MEEI) y *Saarbruecken Voice Database* (SVD) con la colaboración de dos especialistas en la tipificación. Luego se emplearon estos conjuntos en experimentos de clasificación intra e inter bases de datos. Los resultados obtenidos muestran que el enfoque propuesto para distinguir automáticamente entre los tres tipos de señales superan el estado del arte, obteniéndose exactitudes promedio de $87.06\,\%$ y $83.36\,\%$ para MEEI y para SVD, respectivamente.

Para mejorar la estimación de jitter relativo se propuso una nueva técnica basada en la variación total de una estimación del periodo fundamental instantáneo, en lugar de la serie de periodos. Dicha estimación se obtuvo mediante los operadores de *synchrosqueezing*, un método de posprocesamiento utilizado normalmente para aumentar la concentración de los coeficientes de la transformada de Fourier de tiempo corto. Los experimentos llevados a cabo en esta tesis con señales de voz sintéticas, con valor de jitter relativo conocido, demostraron que el método propuesto es más robusto frente al ruido y frente a la presencia de jitter relativo de hasta $15\,\%$, en comparación con PRAAT, un *software* muy utilizado en la clínica. Una prueba preliminar realizada sobre voces reales muestra que el algoritmo propuesto posee un desempeño comparable a PRAAT para voces con bajos niveles de perturbación.

Estos resultados permiten establecer nuevos hitos en la aplicación de medidas de perturbación. Estudios prospectivos tendrán como objetivo acercar los desarrollos de esta tesis doctoral a su aplicación práctica en la clínica.

Abstract

Perturbations are small, random, cycle-to-cycle deviations of the signal's parameters that are always present in signals corresponding to sustained vowels. There is evidence that a number of diseases affecting the phonatory system, and other associated systems, have an effect on these perturbations, which is why its *quantification* has become a topic of increasing interest. This is the aim of an area known as *perturbation analysis*, that uses *perturbation measures* to determine the magnitude of those short-range fluctuations. These measures are based on two fundamental hypothesis: 1) Signals must be nearly periodic, and 2) the parameter under study should be constant within the duration of each cycle. The objective of this thesis is to improve the application of perturbation analysis by focusing on two problematic situations related with each of the previously mentioned hypothesis.

Firstly, the difficulties with the classification of voice signals in three types (1,2 and 3) will be addressed. This classification scheme was created to avoid using perturbation measures on signals that are not nearly periodic. Given that only the type 1 comprises nearly periodic signals, perturbation measures can only be applied to signals within this category. Classification in three types, also termed *signal typing*, is widely used in the clinic. However, signal typing is a rather subjective task, based solely on the perception of the clinicians, which in turn produces a high interprofessional variation in the type assigned.

Secondly, the failure of relative jitter to estimate higher levels of perturbation of the fundamental period was studied. Relative jitter is one of the most used perturbation measures, and is defined as the ratio between the average period perturbation and the average cycle duration. Therefore, a sequence comprising the duration of each cycle, also called period sequence, must be found prior the computation of this parameter. It has been found that relative jitter computed this way underestimates the actual magnitude of the perturbation when this is above a threshold around $5\,\%$ and $8\,\%$.

In order to increase the objectivity of the three-type classification scheme, an automatic classification system based on clinically relevant features and support vector machines was proposed. Over 1200 signals from the *Massachusetts Eye and Ear Infirmary* (MEEI) dataset and the *Saarbruecken Voice Database* (SVD) were labeled in close collaboration with two expert clinicians with previous experience in signal typing. After this, intra and inter dataset experiments were conducted. An accuracy of $87.06\,\%$ was found for MEEI, whereas $83.36\,\%$ was found for SVD. These results show that the performance of the here proposed approach is better than the state-of-the-art.

A novel technique for relative jitter estimation was also proposed, based on the total variation of an estimation of the instantaneous fundamental period instead of the sequence of periods. Synchrosqueezing operators, commonly used for sharpening the short-time Fourier transform, were computed in order to get such estimation. The here reported results from numerical experiments using synthetic voice signals with known jitter magnitude show that the proposed method is more robust to noise and higher levels of relative jitter, up to $15\,\%$, than PRAAT, a widely used software. A preliminary test using actual voice signals was also conducted, the result of which shows that the novel technique has a comparable performance to that of PRAAT for voices with low levels of perturbation.

All these findings set new milestones in the application of perturbation measures. Future studies will try to close the gap between what has been developed in this doctoral thesis and its application in the clinical practice.

Capítulo 1

Introducción

La idea de que el cuerpo humano es un autómata que se rige por las leyes de la mecánica era un concepto predominante en el estudio de la anatomía durante el siglo XVII. René Descartes, gran racionalista e impulsor de esta concepción, que incluso realizó él mismo numerosas disecciones, defendía la afirmación de que el cuerpo humano era semejante a un mecanismo de relojería y que debía regirse por leyes matemáticas o físicas [1].

Desde aquel entonces se ha descubierto que la *máquina* humana es mucho más compleja que un reloj, y menos precisa en sus movimientos. Las interacciones neuronales, el reclutamiento de las fibras musculares y muchos otros procesos fisiológicos se rigen por comportamientos no lineales [2] que dificultan la posibilidad de que dos procesos se repitan de manera perfecta. En consecuencia, dos movimientos no pueden ser perfectamente iguales, y las razones por las que esto sucede son aún difíciles de explicar y modelar [3].

No obstante, es aceptado que ciertas variaciones en los distintos ritmos del cuerpo son fisiológicas, no patológicas, y en consecuencia su presencia es indicativa de un funcionamiento normal [4]. Por otro lado, cuando algunas variables fisiológicas sufren mayores fluctuaciones, y se alejan de las variaciones normalmente presentes, pueden ser indicativo de patologías. Este es el caso, por ejemplo, de la voz. En ella persisten pequeños cambios de corto plazo llamados perturbaciones, cuyo análisis es de importancia para los profesionales de la salud vocal (fonoaudiólogos/as, otorrinolaringólogos/as, logopedistas, etc) [5].

La señal de voz, obtenida mediante la grabación de las alocuciones de un sujeto al emitir, por ejemplo, una vocal sostenida, es una herramienta muy útil para estudiar estas perturbaciones [5, 6]. Hoy en día, dicha señal puede obtenerse con una calidad adecuada para su estudio en instalaciones de bajo costo y representa un método de estudio nada invasivo. Para comprender la importancia del estudio de las perturbaciones de la voz, se realizará a continuación una breve descripción del proceso de la fonación y los diferentes actores involucrados en esta. Seguido de esto, se identificarán las problemáticas a abordar en este documento, para luego definir propuestas asociadas a estas problemáticas. Finalmente se establecerán los objetivos generales y particulares de esta tesis doctoral.

1.1. Breve introducción a la fonación

El aparato fonador es un complejo entramado de sistemas fisiológicos relacionados, cuyo objetivo es producir la voz humana y los sonidos que la caracterizan. Lo componen el sistema respiratorio, los distintos órganos de la fonación (laringe, cuerdas vocales, resonadores nasal, bucal y faríngeo) y los llamados órganos de la articulación (paladar, lengua, dientes, labios y glotis) [7][1]. El entendimiento de la producción de la voz humana no podría lograrse sin el estudio de cada uno

[1] Los aspectos referidos a la irrigación e inervación no serán tratados en este texto.

de los elementos nombrados, tanto en forma individual como combinada. En particular, merece especial atención la laringe, un órgano tubular ubicado en el cuello, compuesto principalmente por cartílagos, un hueso (el hioides), y numerosas estructuras musculares. Debido a que la laringe se mueve y cambia su forma de diversas maneras (puede descender varios centímetros durante el bostezo, se sella y se traslada en dirección anterior en la deglución) es lógico que esté compuesta principalmente por tejidos más flexibles que el hueso, como cartílagos y músculos [7, 8].

Tal como puede verse en la Figura 1.1, la laringe forma parte del conjunto de las vías respiratorias, que se extienden desde el tórax hasta el cuello y la cabeza. Las vías aéreas, como también se las conoce, son los conductos que transportan el aire durante la respiración y durante la fonación. Debido a que la laringe es el órgano central de la fonación, se hará foco en sus componentes y funcionamiento.

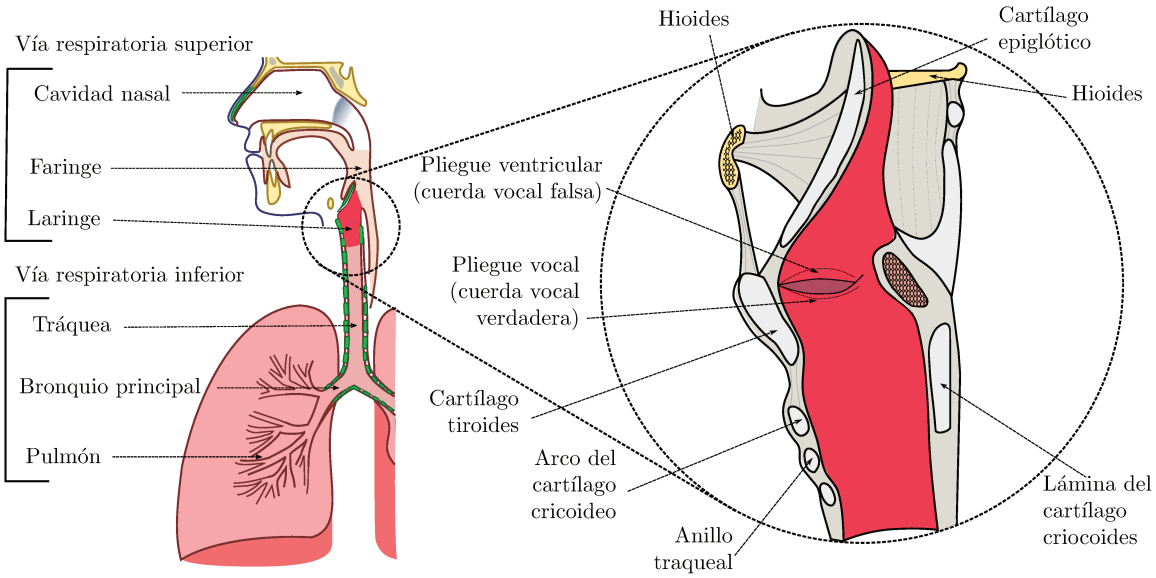


Figura 1.1: Vías aéreas superiores e inferiores, con detalle de la laringe (modificado de [9]).

1.1.1. La laringe

La laringe está situada en el cuello, por delante del esófago y la columna vertebral, rodeada de una cantidad de vasos sanguíneos, nervios y glándulas. Las estructuras cartilaginosas y óseas que forman la laringe, junto con los ligamentos y membranas que los unen, le dan soporte y forma a este órgano tubular. En la Figura 1.2 se expone otra vista más detallada de la laringe, donde pueden apreciarse las relaciones entre estas estructuras. A continuación se describirán los principales cartílagos y músculos de la laringe, responsables de la fonación.

Cartílagos de la laringe

Tiroides El cartílago tiroides está formado por dos placas o láminas que se encuentran unidas por el borde en su cara anterior, formando un ángulo de entre 90º y 120º (Figura 1.3). Los hombres adultos suelen tener un ángulo más agudo que mujeres y niños, coloquialmente conocido como *nuez de Adán*. En la Figuras 1.2 y 1.3 pueden apreciarse las astas superior e inferior de este cartílago. Las primeras se conectan via ligamentos con el hueso hioides, mientras las últimas se articulan con el cartílago cricoides [8].

Cricoides Este cartílago, que se ubica inmediatamente por debajo del cartílago tiroides, forma un anillo sólido que rodea completamente a la vía aérea de la laringe. Podría ser considerado como

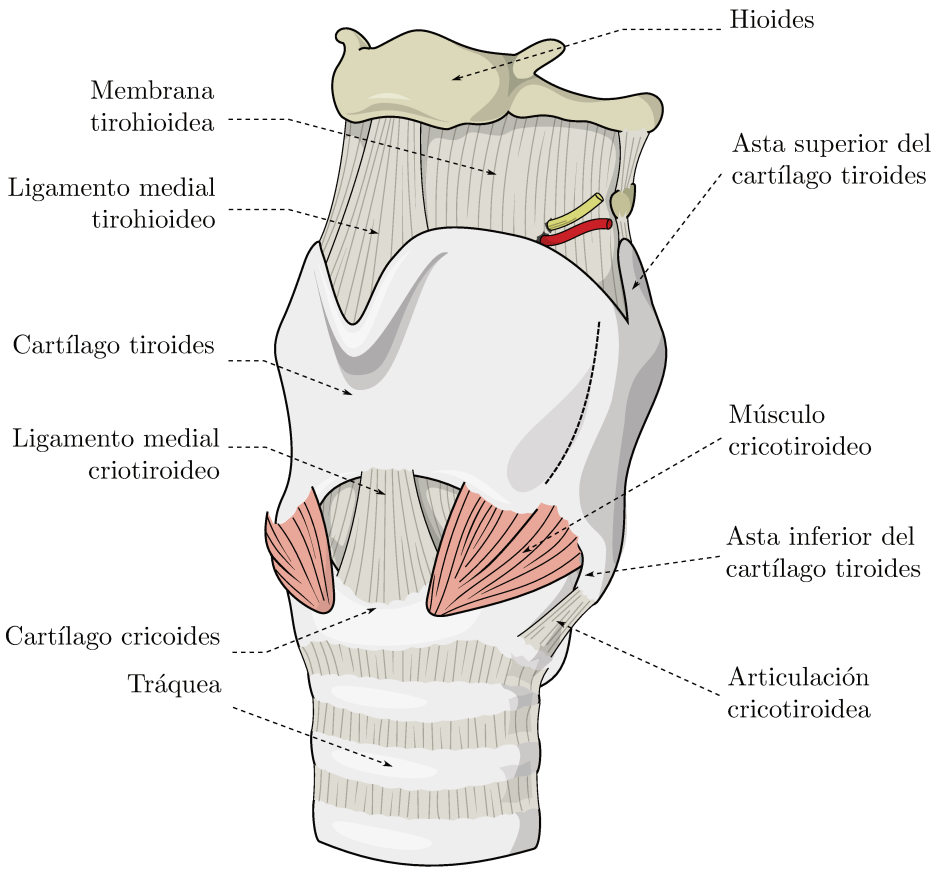


Figura 1.2: Vista anterolateral de la laringe, indicando los principales cartílagos, ligamentos y músculos intrínsecos (modificado de [10]).

el anillo traqueal más superior, pero difiere en su forma y en que constituye una estructura de anillo *cerrado*. El anillo que forma este cartílago es más ancho y alto posteriormente. En la zona posterior superior articula con los cartílagos aritenoides, mientras que en las zonas laterales se articula con el cartílago tiroides (ver Figura 1.2).

Aritenoides Los cartílagos aritenoides son dos piezas piramidales situadas sobre la zona alta y posterior del cartílago cricoides. En su zona inferior poseen dos proyecciones o *apófisis* llamadas *apófisis muscular* (posterolateral) y *apófisis vocal* (anteromedial). La apófisis vocal es de suma importancia pues allí se inserta el ligamento vocal, un componente clave de las cuerdas vocales. De esta manera, las apófisis vocales pueden, al cambiar de posición, realizar la *aducción* (juntar) o la *abducción* (separar) de las cuerdas vocales (Ver Figura 1.4).

Epiglótico Este cartílago se repliega sobre la apertura de la vía aérea en la laringe cuando se necesita cerrar la vía respiratoria herméticamente. Uniéndose mediante el ligamento tiroepiglótico con la superficie interna del cartílago tiroides, el cartílago epiglótico forma la pared anterior de la laringe.

Músculos de la laringe

Los músculos de la laringe pueden dividirse en dos grupos: intrínsecos y extrínsecos. Los intrínsecos conectan los cartílagos de la laringe, mientras que los extrínsecos unen la laringe con

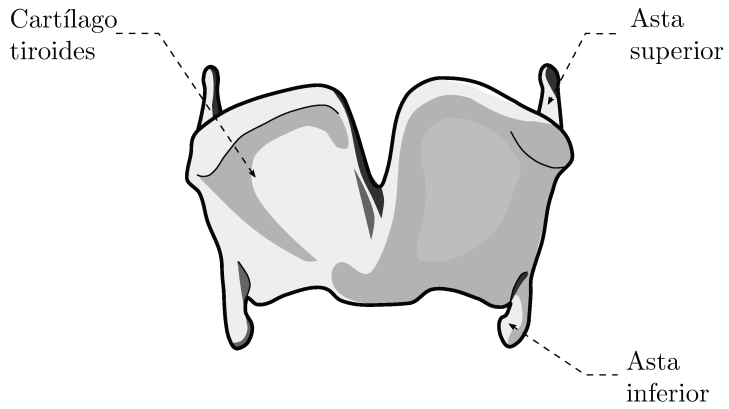


Figura 1.3: Vista anterior del cartílago tiroides.

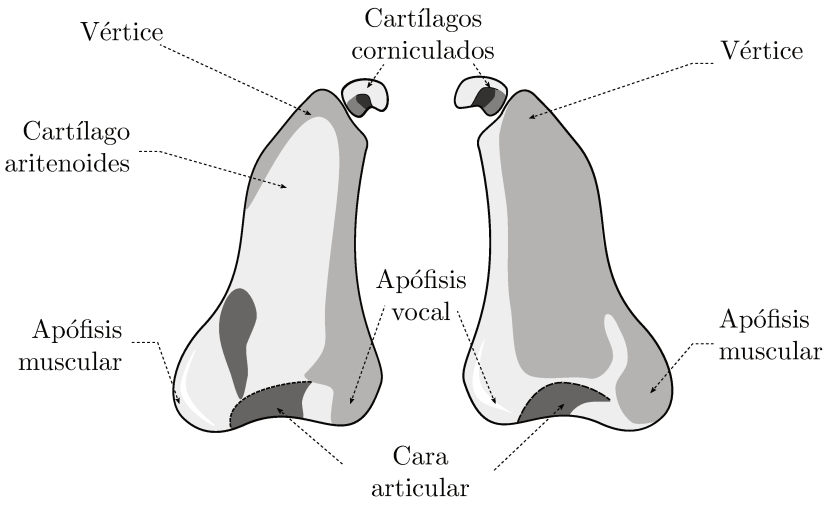


Figura 1.4: Vista posterior de los cartílagos aritenoides.

otras estructuras externas como el esternón y el hueso hioides. En particular, se describirán en este texto sólo los músculos intrínsecos, ya que son los más relacionados con la fonación [7].

Músculos Tiroaritenoideos Estos músculos pares se extienden desde el cartílago tiroideo hasta los cartílagos aritenoideos. Se encuentran divididos en dos haces: *vocal* y *muscular* (Ver Figura 1.5). Se cree que el haz muscular permite un acortamiento rápido de las cuerdas vocales, mientras que el haz vocal realiza un *ajuste fino* de la tensión en las fibras ubicadas más medialmente (hacia el centro del cuerpo). Cuando se contraen en forma conjunta, las cuerdas vocales se alejan, se acortan y se ensanchan.

Músculos Cricotiroideos Estos músculos pares también se dividen en dos haces, los cuales se originan en el arco anterior del cartílago cricoides. Sin embargo, el haz vertical se dirige hacia arriba para insertarse en el borde inferior del cartílago tiroides; mientras que el haz oblicuo se desvía en dirección posterior y se inserta en el asta inferior del cartílago tiroides. La importancia de este músculo radica en que mediante su contracción - relajación, controla el tono de la voz. Al elevar el arco cricoideo y descender el tiroides causa la elongación de las cuerdas vocales.

Músculos Cricoaritenoideos Laterales Al igual que los anteriores, también son un conjunto de músculos pares. Discurren desde los bordes superiores del arco cricoideo y se insertan en el

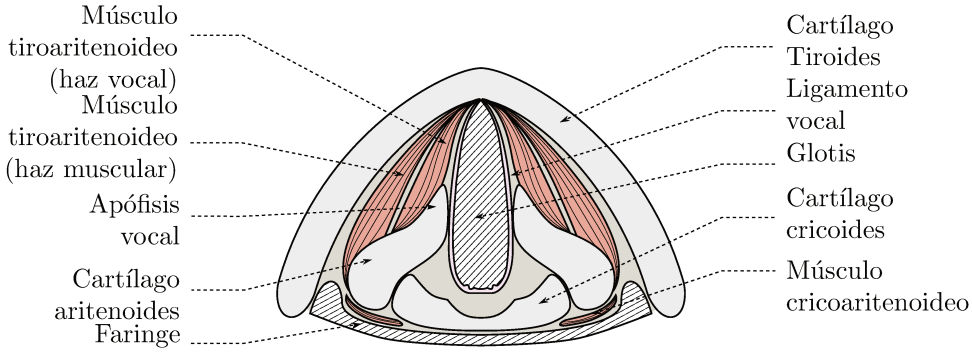


Figura 1.5: Vista superior de un corte transversal de la laringe a la altura del ventrículo laríngeo (entre los pliegues vocales superiores e inferiores).

aritenoides correspondiente. La función de estos músculos es causar la aducción de las cuerdas vocales, mediante el movimiento de los cartílagos aritenoides.

Músculos Cricoaritenoideos Posteriores Junto con los músculos cricoaritenoideos laterales, forman un par *agonista - antagonista*. En este caso, los músculos cricoaritenoideos posteriores son los responsables de la abducción de las cuerdas vocales. Su inserción de origen es la cara posterior del cartílago cricoides, y discurre de manera posteromedial hacia los aritenoides.

Músculo Interaritenoides Conecta ambos aritenoides. Está compuesto por dos haces, uno transverso que cubre enteramente la superficie posterior de los aritenoides y otro oblicuo. El primero se origina en borde lateral de un cartílago aritenoides y termina en el borde lateral del otro. El segundo se inserta en la apófisis muscular de un aritenoides y en el vértice del aritenoides opuesto. Este músculo tiene dos funciones, la aducción de las cuerdas vocales y sellar la zona posterior de la glotis.

Morfología de las cuerdas vocales

Consideremos ahora las estructuras principalmente responsables de la fonación: las denominadas cuerdas o, más correctamente, *pliegues* vocales. En la Figura 1.6 se muestra un corte coronal a lo largo de la laringe en la que puede observarse que la zona de la glotis es la de menor calibre de la vía aérea. Allí, se destacan los pliegues vestibulares (cuerdas vocales superiores o *falsas*), los pliegues vocales (cuerdas vocales inferiores o *verdaderas*) y, entre ellos, un espacio conocido como ventrículo laríngeo. Los pliegues vestibulares carecen de contenido muscular y no pueden contraerse. Por el contrario, los pliegues vocales están formados:

Medialmente, por el cono elástico laríngeo y el ligamento vocal.

Lateralmente, por el haz vocal del músculo tiroaritenoideo, o músculo vocal.

Superficialmente, por la mucosa.

Esta combinación de tejidos, forma una saliente horizontal vigorosa, elástica y, especialmente, *contráctil*.

Si se realiza un corte coronal, el pliegue vocal es triangular [2, 7, 8] (Ver Figuras 1.6 y 1.7). Se describe:

- Una base lateral, libre de mucosa, apoyada sobre la cara profunda del cartílago tiroides.
- Una cara superior, que forma el piso del ventrículo laríngeo.

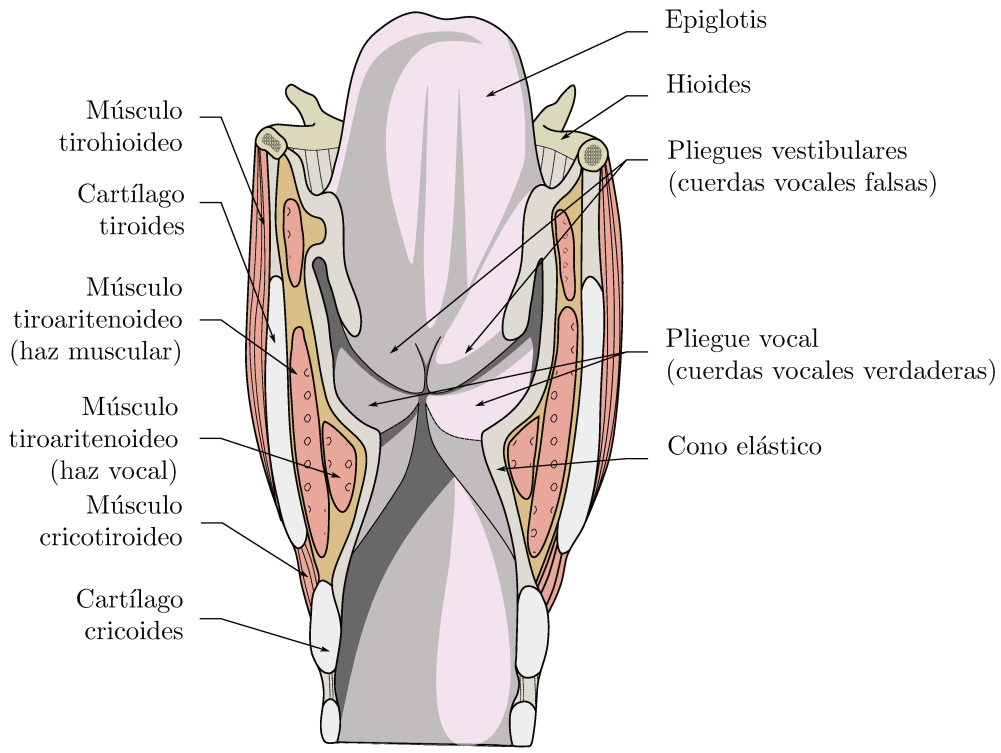


Figura 1.6: Vista posterior de un corte coronal de la laringe. Pueden observarse las diferencias entre los pliegues vestibulares y los pliegues vocales.

- Una cara inferior, que constituye el techo de la cavidad infraglótica.
- Un borde libre, que con el del pliegue del lado opuesto, limita a la hendidura glótica. La dimensión de este espacio es controlada por los pliegues vocales.

Los pliegues vocales se distinguen de los pliegues vestibulares por:

Su aspecto: Los pliegues vestibulares son delgados, recubiertos por un revestimiento rosado; los pliegues vocales son vigorosos, gruesos, móviles y blanquecinos.

Su dirección: Los pliegues vestibulares estan separados entre sí por atrás, mucho más que los pliegues vocales. Vistos desde arriba utilizando un laringoscopio, los pliegues vocales se ubican más mediales, dentro de la separación de los pliegues vestibulares.

Su estructura: Los pliegues vestibulares son esencialmente ligamentosos, mientras que los vocales son considerablemente musculares. Por esa razón, los pliegues vocales tienen una acción predominante en las funciones de la respiración y de la fonación.

Fisiología de los pliegues vocales y producción de la voz

La voz es producida por el pasaje del aire a través de la glotis, ubicada entre los pliegues vocales, hacia el espacio aéreo formado por la faringe, y las cavidades bucal y nasal. Debido a que el aire que atraviesa la laringe es movilizado por la mecánica respiratoria, es necesario resaltar que la interacción entre el sistema respiratorio y el tracto vocal es indisoluble y fundamental. Sin embargo, para no ahondar demasiado en estos contenidos, se dejará de lado la cuestión de la mecánica respiratoria en este texto, recomendándose consultar bibliografía especializada [11].

Cuando una persona desea hablar, el cerebro comanda los distintos órganos intervinientes en la producción de la voz, para modificar las propiedades acústicas del *tracto vocal* y de los estímulos sonoros provenientes de la laringe. El resultado final de esta interacción, es la producción

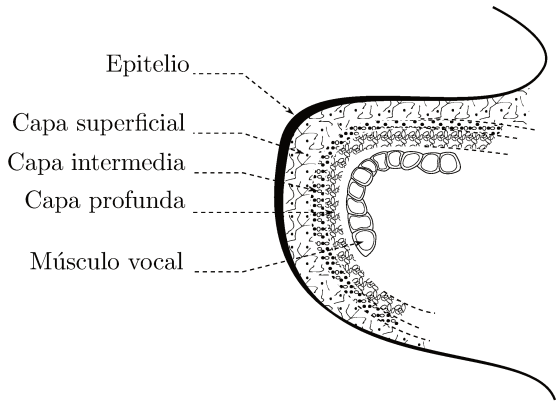


Figura 1.7: Esquema de las distintas capas que forman un pliegue vocal.

de patrones característicos de variación de la presión sonora que al capturarse con un transductor, como un micrófono, forman la señal de voz [12]. Los cambios producidos en el volumen torácico, generados fundamentalmente por la contracción del diafragma y de otros músculos respiratorios principales y secundarios, son los responsables de las diferencias de presión que producen un flujo inspiratorio o espiratorio de aire.

Al hablar, se reduce el volumen torácico y de los pulmones, aumentando la presión en las vías inferiores y formando un flujo de aire que circula desde los pulmones hacia el ambiente y que, al atravesar la glotis, causa la vibración de los pliegues vocales. Orquestados por el sistema nervioso, los pliegues vocales modifican su rigidez mediante la contracción de sus músculos componentes para variar sus propiedades viscoelásticas, resultando en modificaciones de la frecuencia de vibración [7, 13]. La vibración de los pliegues vocales no es similar a la de una cuerda o una membrana, sino que sigue un movimiento de rotación [2], esquematizado en la Figura 1.8. De esta manera, existen periodos de cierre y apertura de los pliegues vocales (Figura 1.8, paso 1 y 4 respectivamente), determinados por la existencia, o no, de contacto entre los pliegues. Una característica importante de la oscilación de los pliegues vocales es que la fase de cierre ocurre mucho más rápido que la fase de apertura. El flujo de aire a través de la glotis se detiene de forma abrupta, mientras que la circulación se reanuda de forma suave. Cuando las cuerdas vocales no se cierran de forma completa durante el ciclo vibratorio (esto puede darse por diversas patologías como nódulos, edema, parálisis unilateral, etc.), el aire atraviesa la glotis en forma continua y turbulenta, incluso en la fase de cierre. Esto genera lo que se conoce como ruido de aspiración o voz aérea [14].

Al producirse la abducción de los pliegues, un chorro de aire ingresa en las cavidades del aparato fonador como el flujo que brota de la punta de una aguja en una jeringa [5]. Las variaciones de presión que se producen por la inyección cíclica de este flujo son moduladas por las estructuras resonantes del aparato fonador, como la cavidad nasal, la boca y los senos paranasales, amplificando algunas frecuencias correspondientes a sus frecuencias de resonancia. La interacción con otros elementos articulantes como los dientes o los labios también modifica las variaciones de presión sonora que se trasmitirán al ambiente [5, 15].

El movimiento de los pliegues vocales no se repite dos veces de la misma forma, sino con pequeñas variaciones relacionadas con la forma en la que los pliegues convergen y divergen en sus extremos posterior, anterior, superior e inferior, así como también la interacción con las paredes de la laringe o el estado de la mucosa en el tracto vocal [16, 17]. Las propiedades viscoelásticas de los pliegues vocales se modifican con la contracción del tejido muscular que las compone, generando cambios en la frecuencia de vibración. Asimismo, el control nervioso de los músculos de los pliegues vocales y de la laringe tampoco opera de manera tal que las repeticiones de cada ciclo de vibración sean perfectamente iguales [16].

Como se mencionó anteriormente, la señal de la voz refleja estas fluctuaciones de corto plazo.

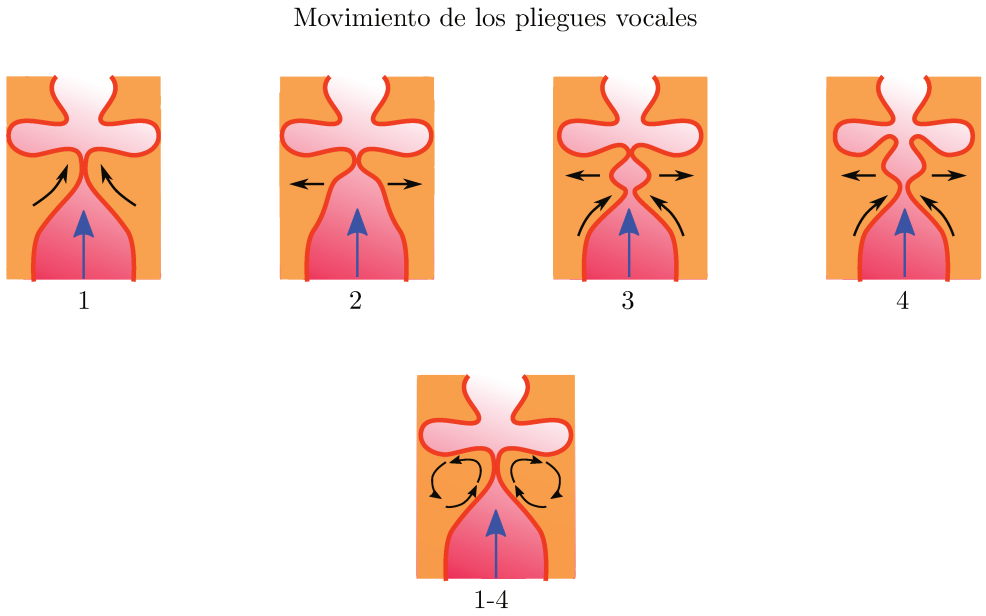


Figura 1.8: Movimiento de rotación de los pliegues vocales (modificado de [2]).

Lieberman [6] fue el primero en advertir que existen pequeñas diferencias entre los ciclos de una señal de voz supuestamente periódica y las nombró *perturbaciones* con la idea de que son pequeñas desviaciones respecto de un parámetro que "en promedio" se mantiene igual a lo largo de toda la fonación.

1.2. Motivación

Las señales obtenidas del cuerpo humano contienen información acerca de los sistemas fisiológicos que las producen, y la voz no es una excepción. Esta transporta información de los impulsos de presión producidos a nivel de la glotis así como el estado del tracto vocal, formado por una serie de cavidades de resonancia, como la boca o los senos paranasales. Incluso hay evidencia empírica sobre el efecto de enfermedades de base neurológica [18–32] o psicológica [33] en la señal de la voz y su efecto sobre las medidas de perturbación. Patologías que afectan otros sistemas, como el respiratorio [34] o digestivo-endocrino [35–37], también tienen efecto sobre la voz y sus perturbaciones. Adicionalmente, las perturbaciones pueden indicar el estado de los pliegues vocales al estar directamente relacionadas con el fenómeno vibratorio que da origen a la fonación [38, 39]. Esto justifica su investigación como una herramienta para describir el estado de salud del aparato fonador y otros sistemas asociados.

El análisis de las perturbaciones comprende el estudio de los pequeños cambios que se producen ciclo a ciclo en determinados parámetros de la señal de voz, fundamentalmente en su frecuencia fundamental, su amplitud y, en general, su forma de onda. El *jitter* y el *shimmer*, correspondientes a las perturbaciones de la frecuencia y de la amplitud máxima ciclo a ciclo respectivamente, son dos fenómenos estudiados dentro del análisis de perturbaciones mediante una variedad de medidas como el jitter y el shimmer relativo, perturbación promedio relativa (RAP), coeficiente de perturbación de periodos (PPQ), etc [5]. Los cambios en la forma de onda de un ciclo al siguiente son menos estudiados que las fluctuaciones de corto plazo en la frecuencia y la amplitud, y en consecuencia no existe una forma particular de denominarlos, aunque existen algunas medidas aplicables con ese fin [40, 91].

Para poder aplicar el análisis de perturbaciones debe ser posible definir adecuadamente los

ciclos de la señal. Esta suposición es normalmente referida como hipótesis de periodicidad [41]. De otra forma, la aplicación de las medidas englobadas dentro del análisis de perturbaciones no tiene sentido [41, 42]. A pesar de esto, estas medidas, particularmente el *jitter* y el *shimmer*, son utilizadas en la práctica clínica con asiduidad, incluso en casos que no satisfacen la hipótesis de periodicidad [42, 43]. Esto produce resultados que no reflejan el verdadero estado de salud de las cuerdas vocales, ni describen verdaderamente el fenómeno vibratorio subyacente a la señal [5, 38, 43]. Este uso inadecuado se debe principalmente a la alta disponibilidad que tienen estas medidas, ya que existe una variedad de programas de computadora que pueden calcularlas (aunque sus resultados pueden diferir bastante entre sí [43–48]). Algunos de estos *software*, como PRAAT [49], ofrecen un número diverso de parámetros sin contemplar la verificación de la condición de la señal a analizar. En otras palabras, las medidas para caracterizar perturbaciones se utilizan masivamente porque están al alcance de la mano de los especialistas, lo que no implica que necesariamente todas las señales satisfagan los requisitos para someterse al análisis de perturbaciones.

Para evitar esta situación y fomentar un uso consciente de las medidas de perturbación, Titze [42] propuso una clasificación en tres tipos, también llamada *tipificación*, que debe aplicarse antes del empleo de cualquier herramienta de análisis.

- Tipo 1: Señales aproximadamente periódicas que no muestran cambios cualitativos en el segmento de análisis, con frecuencias modulantes o subarmónicas cuyas energías sean de un orden de magnitud inferior a la de la frecuencia fundamental, o nulas (ver Figura 1.9a).

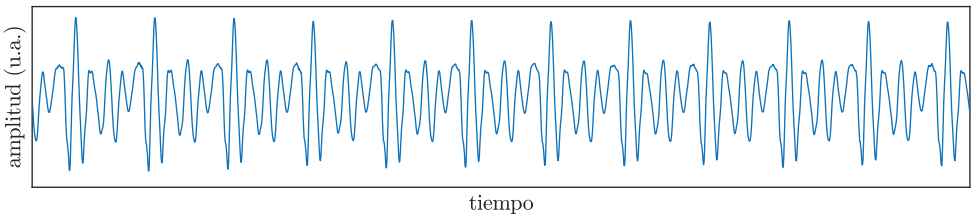
- Tipo 2: Señales con cambios cualitativos en el segmento de análisis, o señales con frecuencias subarmónicas y/o modulantes cuyas energías se aproximen en magnitud a la energía de la frecuencia fundamental. Puede, por lo tanto, no existir una única frecuencia fundamental obvia en el segmento analizado (ver Figura 1.9b).

- Tipo 3: Señales sin estructura periódica aparente (ver Figura 1.9c).

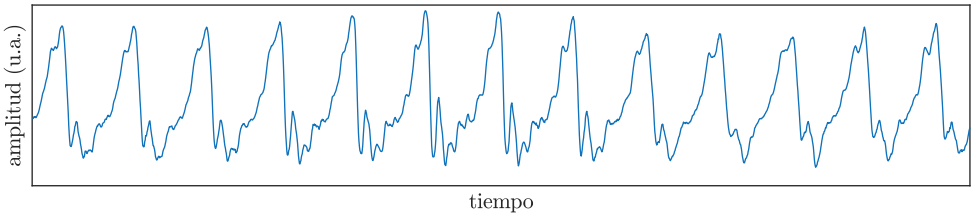
En esta clasificación, las señales tipo 1 pueden someterse a cualquier análisis, incluyendo el de perturbaciones. Para las señales tipo 2 se recomienda el uso de espectrogramas banda angosta o escalas perceptuales como GRBAS [40] o CAPEV [50] para su estudio. Por último, se sugiere que las señales tipo 3 sean evaluadas mediante escalas perceptuales. La clasificación fue ampliamente adoptada, y es un procedimiento habitual previo a la aplicación de cualquier medida de perturbación [51–58].

Sin embargo, la tipificación de señales de voz adolece de un defecto intrínseco: las definiciones de los tipos es vaga, y su interpretación es, de alguna manera, librada al profesional. Debido a que las definiciones no están ligadas a criterios cuantitativos, la clasificación se vuelve subjetiva, lo que genera cierta variación interprofesional. Por esta razón, ha despertado el interés en parte de la comunidad científica que estudia la salud de la voz en generar medidas que puedan distinguir entre los tres tipos propuestos. Estos nuevos descriptores se han basado en medidas de dinámicas no lineales, transformada de Fourier de tiempo corto o estadísticos de orden superior [59–66]. No obstante, su estudio se ha limitado a la descripción estadística de su capacidad para diferenciar a los distintos tipos, no abordándose su uso para la clasificación de las voces en forma directa, por ejemplo mediante un algoritmo de clasificación (una excepción a esto es [61]). A largo plazo, sería deseable la existencia de un sistema capaz de clasificar las señales de manera automática para asistir a los profesionales de la salud de la voz a la hora de tipificar las señales.

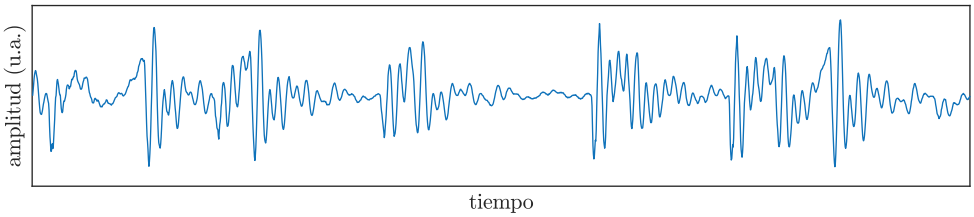
De entre todas las medidas de perturbación que existen, el *jitter relativo* es una de las más empleadas [67–70]. Consiste en el cociente entre la perturbación promedio del periodo fundamental y el periodo fundamental promedio. Tradicionalmente, la determinación de esta medida requiere, en primer lugar, la estimación de la serie de periodos. Ésta es una sucesión que comprende las duraciones de cada ciclo de la señal, y para hallarla es necesario la identificación de los puntos donde



(a) Tipo 1



(b) Tipo 2



(c) Tipo 3

Figura 1.9: Ejemplos de señales clasificadas en los tres tipos propuestos por Titze [42].

comienzan y terminan los ciclos. Estos puntos *fiduciarios* son usualmente determinados mediante técnicas basadas en:

1. Detección de cruces por cero.
2. Detección de picos (máximos locales).
3. Coincidencia de forma de onda.

Estos principios para identificar los puntos fiduciarios definen tres familias de algoritmos. La primera busca los cruces por cero en la dirección positiva o negativa de una señal, luego de la aplicación de un filtro pasabajos con una frecuencia de corte cercana a la frecuencia fundamental promedio. La segunda busca los máximos locales de la señal con la premisa de que el máximo global dentro de cada ciclo es un punto destacable y fácilmente detectable por su prominencia. La tercera familia utiliza un criterio de similitud entre la forma de onda de un ciclo y el siguiente para determinar la duración de cada ciclo [71, 72]. Tanto la detección de cruces por cero como la detección de picos son bastante sensibles a la relación señal a ruido, debido a que la diferencia en la posición específica de un punto en la señal puede deberse a la presencia del ruido y no a una diferencia intrínseca entre los ciclos dada por la vibración de los pliegues vocales [71]. El método de coincidencia de onda es, de los tres métodos aquí descritos, el más robusto frente al ruido, ya que utiliza la minimización del error cuadrático medio entre un ciclo y el ciclo posterior para identificar los puntos fiduciarios, eliminando parte de la influencia del ruido. No obstante, también utiliza una estimación del periodo promedio para establecer un área de búsqueda del punto fiduciario siguiente, lo que resulta en una incapacidad de detectar cambios rápidos en el periodo fundamental [71, 72].

Existe evidencia de que el cálculo de jitter relativo basado en la serie de periodos se vuelve insensible para mayores niveles de jitter, observándose una dinámica de saturación para valores de jitter relativo por encima de entre $5\,\%$ y $8\,\%$, según la frecuencia fundamental promedio

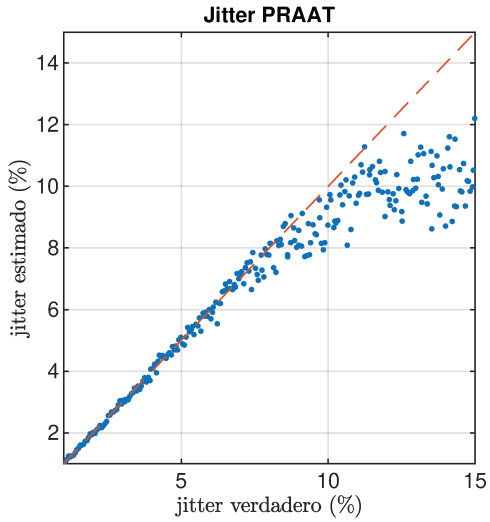


Figura 1.10: Estimación de jitter relativo a partir de la serie de periodos, calculado con PRAAT. Cada punto representa un valor medido sobre una señal sintética cuyo nivel de jitter verdadero es conocido (eje horizontal). La línea de trazos marca la recta identidad.

[67, 68, 70, 72]. Esto puede deberse, en parte, a los algoritmos utilizados para la identificación de los puntos fiduciarios, necesarios para la construcción de la serie de periodos. No obstante, incluso métodos de amplia difusión y basados en el criterio más robusto de coincidencia de forma de onda padecen de la limitación descrita. Este es el caso de PRAAT [49], un software libre y de código abierto de uso generalizado en la práctica clínica. La Figura 1.10 muestra cómo la estimación de jitter mediante PRAAT falla en la estimación de valores de jitter por arriba de, aproximadamente, $8\,\%$. Otra razón detrás de estas fallas es que el uso de la serie de periodos implica suponer que la frecuencia fundamental, o su recíproco, el periodo fundamental, se mantiene constante durante cada ciclo [41]. Adicionalmente, el error de estimación del jitter relativo es proporcional a la frecuencia fundamental promedio de la señal, dada una frecuencia de muestreo, lo que dificulta su estimación en voces agudas.

1.3. Problemática

Se abordarán las siguientes situaciones problemáticas:

Problemática 1: La clasificación de señales en los tres tipos propuestos por Titze [42], es subjetiva, promueve la variabilidad interprofesional y carece de criterios cuantitativos asociados a la tipificación.

Problemática 2: El jitter relativo basado en la secuencia de periodos subestima el verdadero valor de jitter para mayores niveles de perturbación y depende fuertemente de la frecuencia fundamental promedio de la señal de voz.

La Problemática 1 tiene similitudes con una amplia variedad de problemas biomédicos, en los que se observa la necesidad de reducir la variabilidad interprofesional generalmente asociada a criterios subjetivos. Para el caso particular de la voz, se ha propuesto la denominación de Sistemas de Análisis de la Condición de la Voz para referirse al uso de herramientas de la ingeniería con el propósito de aumentar la objetividad en las escalas perceptuales de la calidad de la voz [73, 74].

1.4. Propuestas

Se propone una investigación original que estudie las dificultades del análisis de perturbaciones identificadas anteriormente como situaciones problemáticas.

Para abordar la Problemática 1 se propone la implementación de un sistema de reconocimiento de patrones basado en: 1) La extracción de características relevantes para el problema de la tipificación. 2) El uso de un algoritmo de clasificación que permita la tipificación automática en base a los descriptores propuestos y que emule el criterio de profesionales altamente calificados. La hipótesis detrás de esta propuesta es:

> **Hipótesis 1:** Las características propuestas proveerían de indicadores cuantitativos sobre los que basar la clasificación. Luego, el entrenamiento de un modelo de clasificación que tome como entrada dichas características aportaría una tipificación más objetiva. Dado que el paradigma de aprendizaje es supervisado, la clasificación tomada como *gold standard* debería estar hecha por, al menos, un profesional de referencia en el cuidado de la voz.

Las voces utilizadas para esta propuesta provendrán de bases de datos de difundido uso en el área del procesamiento de la voz. Adicionalmente se intentará utilizar clasificadores lineales o árboles de decisión para conservar la interpretabilidad de la clasificación, una cualidad de interés en los problemas biomédicos.

Para la Problemática 2 se propone estudiar la estimación de jitter relativo a partir de técnicas más modernas de análisis de señales. Esto se sugiere con el propósito de obtener mejores aproximaciones de la frecuencia fundamental instantánea, y evitar utilizar la estimación de la serie de periodos, que constituye una aproximación de orden *cero*, o constante a trozos incapaz de reflejar variaciones rápidas del periodo instantáneo (ver Sección 6.2.1). Esta proposición se basa en la siguiente hipótesis:

> **Hipótesis 2:** El jitter relativo computado a partir de la secuencia de periodos heredaría de esa aproximación del periodo fundamental instantáneo la imposibilidad de reflejar mayores fluctuaciones en el periodo de la señal. Dado que un nivel de jitter elevado implica mayores variaciones del periodo, una nueva técnica para estimar el jitter relativo a partir de aproximaciones de orden superior permitiría aliviar la insensibilidad existente a niveles más altos de perturbación.

Para esta propuesta se pretende utilizar señales de voz sintéticas cuyo jitter relativo verdadero sea conocido. Con ese fin se implementarán modelos estocásticos de jitter y se sintetizarán voces utilizando el clásico modelo fuente filtro [15, 16]. Para obtener mejores estimaciones de la frecuencia fundamental instantánea se propone el uso de los operadores de *synchrosqueezing* de orden superior [75].

1.5. Objetivos

Teniendo en cuenta lo mencionado, se proponen los siguientes objetivos.

1.5.1. Objetivo general

- Realizar aportes para mejorar la aplicación del análisis de perturbaciones.

Se buscará dar cumplimiento a este objetivo general a través de, por un lado, la implementación de un sistema de análisis de la condición de la voz para mejorar la objetividad de la clasificación en tres tipos y, por otro lado, a través de nuevas medidas de perturbación. Considerando esto, se proponen los siguientes objetivos específicos.

1.5.2. Objetivos específicos

1. Identificar los criterios clínicos para la tipificación, basados en la audición del especialista y la inspección visual de las señales.
2. Traducir estos criterios a cantidades cuantificables a través del procesamiento digital de la señal de la voz y señales asociadas (electroglotograma, vibración en la piel del cuello, electromiograma, etc).
3. Definir nuevos criterios objetivos para la clasificación automática de voces.
4. Proponer nuevos estimadores de la perturbación del periodo fundamental que resulten más robustos frente a mayores niveles de perturbación y ruido que los utilizados actualmente.

1.6. Organización del documento

El presente documento se organiza de la siguiente manera. El Capítulo 2 se dedicará a la descripción de la señal de la voz y el análisis de perturbaciones. En particular se abordará la clasificación en tres tipos, su variantes y antecedentes, así como también se describirá la perturbación del periodo fundamental (jitter), sus orígenes, estimadores y modelos de síntesis. En el Capítulo 3 se realizará una introducción somera a las técnicas de análisis tiempo frecuencia empleadas más adelante en el documento. Allí se repasarán, en primer lugar, conceptos básicos como la transformada de Fourier de tiempo corto. En segundo lugar, se presentarán técnicas más modernas como el *synchrosqueezing* de orden superior y sus operadores. El Capítulo 4 versa sobre las técnicas de aprendizaje maquinal empleadas para el reconocimiento de patrones en el Capítulo 5, principalmente para la selección de características y las máquinas de vectores de soporte. En los Capítulos 5 y 6 se desarrollarán las propuestas indicadas anteriormente, cuyos resultados han sido total o parcialmente publicados en revistas con referato. En el Capítulo 5 se abordará la clasificación automática de señales en los tres tipos mediante máquinas de vectores de soporte, basada en características propuestas para tal fin. Por otro lado, el Capítulo 6 describe una nueva técnica más robusta para la estimación del jitter relativo a partir de estimaciones locales de orden superior de la frecuencia instantánea de la señal de voz y su derivada. Finalmente, en el Capítulo 7 se indicarán las principales conclusiones obtenidas a partir del desarrollo de esta tesis, junto con posibles trabajos futuros y líneas de investigación.

Capítulo 2

La señal de voz y sus perturbaciones

2.1. Introducción

El estudio de la salud vocal considera varias perspectivas para explorar el aparato fonador, así como también tratamientos para mejorar su estado. Algunas formas de estudiar las estructuras responsables de la fonación son directas, con distintos grados de invasividad. Por ejemplo la videolaringoestroboscopía o la medición de la presión sobre las cuerdas vocales, que tienen una invasividad media o alta [5]. Por otro lado, las escalas perceptivas o parámetros acústicos son formas no invasivas de estudiar la fonación. De hecho, las medidas perceptivas como la escala GRBAS o CAPEV [40, 50] son estándares en este campo [76]. Estas escalas resultan útiles por no ser invasivas pero, al mismo tiempo, sufren una notable subjetividad, ya que las conclusiones derivadas a partir de ellas dependen en gran medida del profesional de la salud vocal que las utiliza [77].

Otra forma no invasiva de estudiar la vibración de las cuerdas vocales consiste en utilizar señales provenientes del aparato fonador [78]. De todas ellas, la señal de voz es la más empleada debido a que su adquisición es sencilla y provee gran cantidad de información. Algunas técnicas permiten evaluar la vibración de los pliegues vocales a partir de la señal de voz. Por ejemplo, el filtrado inverso. Este consiste en utilizar un modelo matemático para la estimación de la función glótica, una magnitud muy difícil de estimar en la práctica pero harto importante para caracterizar la mecánica vibratoria de los pliegues vocales [15, 79]. Asimismo, el habla transmite mucha más información que sólo la relacionada al estado del aparato fonador. Debido a que un número de sistemas fisiológicos interactúan para lograr la fonación, la señal resultante codifica adicionalmente información sobre el sistema nervioso, las emociones o enfermedades del sistema respiratorio, entre otros [18–37]. La decodificación de esta información depende en gran medida del procesamiento realizado, por lo que el campo de extracción de información de las señales relacionadas al aparato fonador está en constante expansión [80]. En este documento nos abocaremos al estudio de la señal de la voz, que debe distinguirse de la señal *del habla*. La primera consiste en la emisión vocal de fonemas sonoros o vocales, mientras que la segunda se trata del registro completo de una alocución, ya sea espontánea o guionada, para estudiar aspectos como la articulación de los fonemas o la prosodia [15].

En este capítulo se presentará la señal de voz, comenzando con la descripción del conocido modelo fuente-filtro de la fonación. Luego se detallarán las perturbaciones que afectan la voz y, en particular, la problemática en torno a las fluctuaciones de corto plazo de la frecuencia fundamental. Finalmente se abordará la clasificación en tres tipos con el objetivo de asegurar la idoneidad de aquellas señales pasibles de ser parametrizadas mediante medidas de perturbación. Todos estos aspectos serán de relevancia en capítulos posteriores, principalmente en los Capítulos 5 y 6, donde se describirán nuevos aportes para el análisis de perturbaciones.

Fonemas	Vocales: /a/ /e/ /i/ /o/ /u/
	Fricativos: /f/ /s/ /j/ /y/
	Africados: /ch/
	Oclusivos: /b/ /d/ /g/ /p/ /t/ /k/
	Nasales: /n/ /m/ /ñ/
	Vibrantes: /r/ /rr/
	Laterales: /l/ /ll/

Tabla 2.1: Clasificación de fonemas según sus características acústicas.

2.2. Señal de voz

Tras ser producidas por el aparato fonador y emitidas al medio, las variaciones de presión sonora son capturadas por un *transductor*, que usualmente las transforma en diferencias de potencial eléctrico para ser almacenadas en algún soporte, previa digitalización del registro. El resultado de este procedimiento es la señal de voz digitalizada. Las señales son representaciones de fenómenos físicos, que transportan información acerca del sistema que las produjo. En general, dicha información se encuentra contenida o codificada en un patrón de variaciones de alguna magnitud [81]. En el caso de la voz, dicha magnitud es la variación de presión de aire en el medio y la información codificada refleja el estado del tracto vocal durante su producción y su excitación. Esta última puede darse por el paso intermitente de aire debido a la mecánica oscilatoria de los pliegues vocales (excitación del tracto con *pulsos glóticos)*, o por el avance de un flujo de aire que atraviesa la laringe sin intermitencia (los pliegues se mantienen alejados, permitiendo un flujo continuo de aire) [5, 12].

Las distintas combinaciones de excitación y configuraciones del tracto vocal, dan lugar a la producción de diferentes sonidos o *fonemas*, que son las unidades lingüísticas básicas del habla. Se los puede definir como el conjunto mínimo de unidades que permite decir cualquier palabra en un idioma determinado [12], y la morfología de la señal depende fuertemente del fonema pronunciado. Los fonemas se clasifican, según las características acústicas y gestos articulatorios que involucran, en *vocálicos* y *consonánticos* (ver Tabla 2.1). Se dejará de lado en este documento la descripción de los fonemas consonánticos, ya que las señales utilizadas en capítulos posteriores corresponden a la vocal $/a/$ sostenida.

La Figura 2.1 esquematiza el recorrido del flujo de aire proveniente de los pulmones, que atraviesa la laringe hasta las cavidades que constituyen el tracto vocal. En la articulación de vocales el tracto adopta una configuración abierta, similar a una serie de *tubos* de sección variable interconectados, y la fuente de excitación la conforman pulsos glóticos, cuya duración o *periodo glótico* es aproximadamente T_0, generados por la intermitencia en la apertura de los pliegues vocales. Estos pulsos estimulan el tracto vocal, que actúa como un sistema resonador capaz de modificar sus propiedades acústicas alterando la posición de la mandíbula, la lengua y los labios. De esta manera, el tracto vocal se comporta como un filtro acústico *adaptativo* [15].

Como se muestra en la Figura 2.1, las principales estructuras resonantes del tracto vocal son la cavidad faríngea, la cavidad oral y la cavidad nasal. Esta última puede acoplarse o desacoplarse de la fonación mediante la apertura o cierre de una estructura anatómica conocida como *velo*. Para los fonemas nasales, esta estructura permite el paso del aire hacia la cavidad nasal, haciéndola partícipe de la fonación. Por el contrario, para los fonemas no nasales, el velo se cierra por completo [15].

En un sujeto sano, la señal de voz correspondiente a una vocal sostenida es aproximadamente periódica, sus ciclos tienen una duración aproximada de un *periodo fundamental* T_0, y su morfología no sufre variaciones importantes ciclo a ciclo [5]. En la Figura 2.2a puede observarse una representación temporal de una señal de voz correspondiente a una vocal sostenida. En la Figura 2.2b se observa una estimación del espectro de amplitud de la misma señal, de donde es posible apreciar

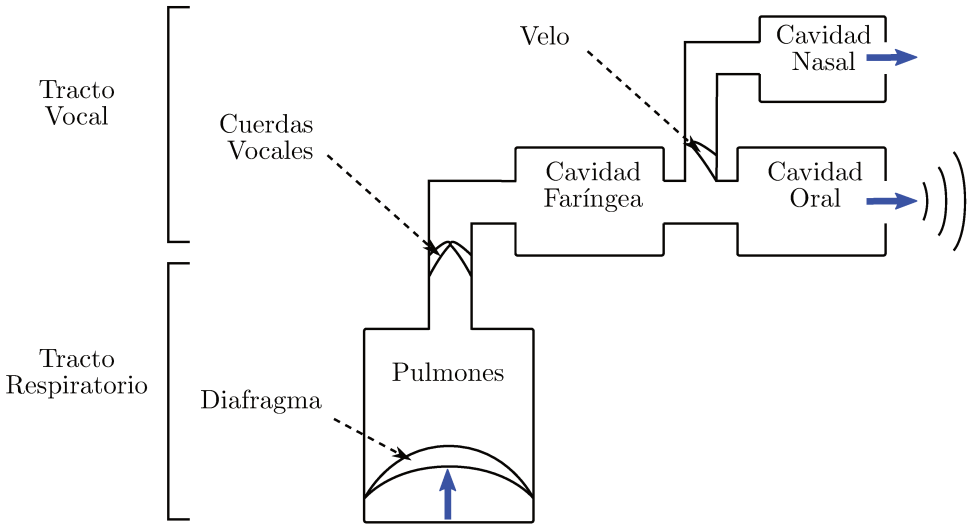
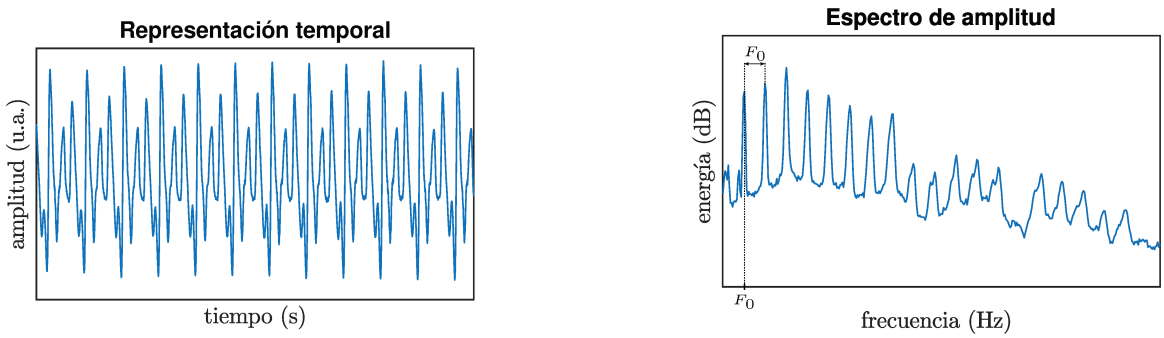


Figura 2.1: Esquema que muestra la fuente de excitación, constituida por elementos del sistema respiratorio y los pliegues vocales de la laringe; y el tracto vocal, con sus cavidades de resonancia (nasal, faríngea y oral).

que la representación frecuencial se compone de picos separados a una distancia $F_0 = 1/T_0$, es decir que son *armónicos* de la frecuencia fundamental. La presencia de estos armónicos es indicativa de la periodicidad de la señal en el tiempo. El cierre repentino de la glotis es la fuente de la mayor parte de la energía de la señal, y el origen de los armónicos de frecuencias más altas de la fuente glótica. Su amplitud depende de la intensidad de la fonación, y disminuye conforme aumenta la frecuencia [5]. Esto marca la *pendiente espectral*, es decir, el ritmo de decaimiento de la amplitud de los armónicos con la frecuencia. Al aumentar la intensidad vocal, el cierre de la glotis se produce aun más rápidamente que lo normal, resultando en armónicos con mayor amplitud y una pendiente espectral menos acentuada [5].

En la fonación aérea o con ruido de aspiración el cierre glótico es menos hermético, lo que incrementa el flujo turbulento a través de la glotis y aumenta el tiempo de cierre de los pliegues, lo que se corresponde en el espectro con la ausencia de armónicos en altas frecuencias (una pendiente espectral elevada). Dado que la turbulencia es aleatoria y aperiódica, y en consecuencia no posee un espectro con armónicos, existe en este caso un nivel de energía más elevado entre los picos del espectro que en la fonación normal [5, 82, 83].

Para estudiar la señal de voz, resulta útil contar con un modelo de ésta. Los primeros intentos de simular la producción de la voz dieron lugar a modelos mecánicos, y datan de fines del siglo XVIII. Hacia mediados del sigo XX, fueron creadas algunas máquinas eléctricas capaces de imitar la voz, y hacia fines de ese siglo, surgieron los primeros modelos implementados en la computadora [15]. La motivación detrás del modelado del habla va más allá de meramente reproducir la fonación, ya que también permite comprender en profundidad los mecanismos que producen la voz. Un conocido modelo para estudiar la relación entre el tracto vocal y la señal de la voz se denomina *Modelo Fuente-Filtro*, cuyo diagrama en bloques simplificado se muestra en la Figura 2.3 [12, 15, 84]. En él, se considera que la señal de la voz es la respuesta de un sistema (el tracto vocal), a distintas entradas que se corresponden con los tipos de excitación. Los pulsos glóticos son modelados como



(a) (b)

Figura 2.2: (a) Representación temporal. (b) Espectro de amplitud de la señal.

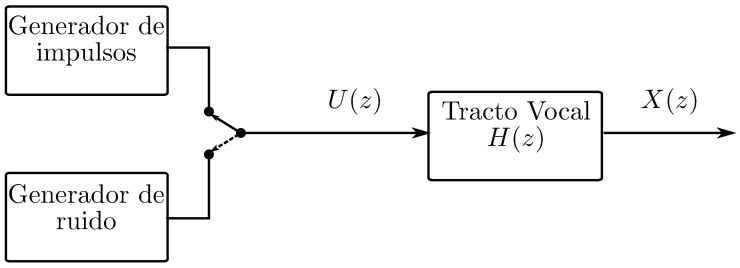


Figura 2.3: Diagrama en bloques del modelo Fuente - Filtro (Ver Ecuación (2.2)).

una señal pulsante o como un tren de impulsos, mientras que el paso de un flujo turbulento de aire circulando en forma continua a través de la laringe se modela como *ruido blanco*. De esta manera el tracto vocal es visto como un *filtro* que modifica la señal de entrada, proveniente de alguna *fuente*, y obtiene una respuesta a la salida en función de su *función de transferencia*. Poniendo el foco en el caso de los fonemas vocales, la fuente de excitación consistirá en una señal pulsante que excita las estructuras de resonancia del tracto vocal (ver Figura 2.1). La configuración de cada una de estas estructuras modificará la señal de la voz, y se reflejará en cambios en su morfología y en su espectro.

Considerando un marco de tiempo discreto y la Figura 2.3, el modelo Fuente - Filtro puede resumirse en la siguiente ecuación:

$$X(z) = U(z)H(z), \tag{2.1}$$

donde $X(z)$ y $U(z)$ son la transformada Z de la señal de la voz discreta $x[n]$ y de la fuente de excitación $u[n]$ respectivamente, y $H(z)$ es la función de transferencia del tracto vocal.

De la Ecuación (2.1), suponiendo que las transformadas de Fourier de la señal de la voz y de la fuente de excitación existen, se desprende la siguiente expresión:

$$\left|Y(e^{i2\pi f})\right| = \left|X(e^{i2\pi f})\right|\left|H(e^{i2\pi f})\right|, \tag{2.2}$$

donde $\left|Y(e^{i2\pi f})\right|$ y $\left|X(e^{i2\pi f})\right|$ son los módulos de las transformadas de Fourier de la señal de la voz y de la fuente de excitación correspondientemente; y $\left|H(e^{i2\pi f})\right|$ es la respuesta en frecuencia de módulo del tracto vocal. Luego, es posible observar que el espectro de la señal de la voz será igual al producto entre el espectro de la fuente y la respuesta en frecuencia del tracto.

Como puede observarse en la Figura 2.4, el espectro de una señal de voz correspondiente a un fonema vocal se encuentra modulado por la respuesta en frecuencia del tracto vocal, que

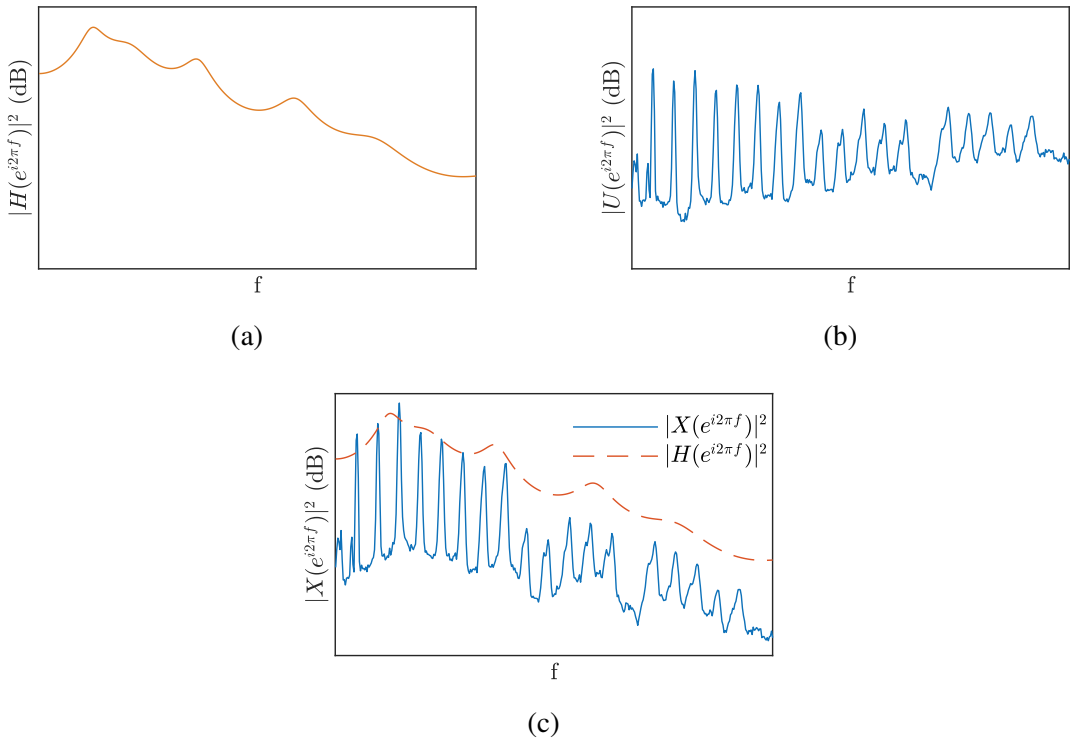


Figura 2.4: *a)* Respuesta en frecuencia del tracto vocal. *b)* Espectro de potencia de la fuente de excitación. *c)* Espectro de potencia de la señal de voz, junto con la respuesta en frecuencia del tracto superpuesta.

amplifica determinadas frecuencias mientras que atenúa otras según su envolvente (ver Figura 2.4a y línea de trazos en la Figura 2.4c). Esta envolvente posee máximos locales, correspondientes a las frecuencias de resonancia de las cavidades del tracto vocal, denominadas *formantes*. Cada una de las vocales se diferencia espectralmente por las frecuencias de las formantes en la respuesta del tracto vocal. Generalmente se utilizan de tres a cuatro formantes para describir a las diferentes vocales [15].

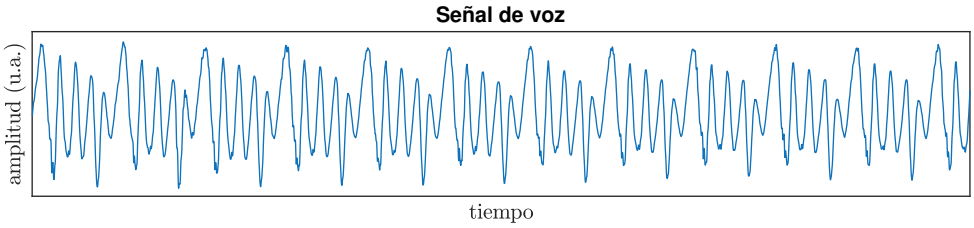
La obtención de $H(z)$ y de $u[n]$ a partir de la Ecuación (2.1) es el objetivo de un área particular del análisis de la voz conocida como *filtrado inverso*. A grandes rasgos, este área estudia formas de obtener la función glótica $u[n]$, a partir de la señal de voz $x[n]$, como una manera de estudiar de manera indirecta el estado de los pliegues vocales. Una forma de hacerlo consiste en estimar la función de transferencia del tracto vocal $H(z)$ para luego realizar el filtrado de $x[n]$ con $H^{-1}(z)$, de ahí el nombre de filtrado inverso. Es habitual caracterizar $H(z)$ mediante un modelo *todos polos* autorregresivo:

$$H(z) = \frac{G}{1 + \sum_{j=1}^{J} h[j] z^{-j}} \tag{2.3}$$

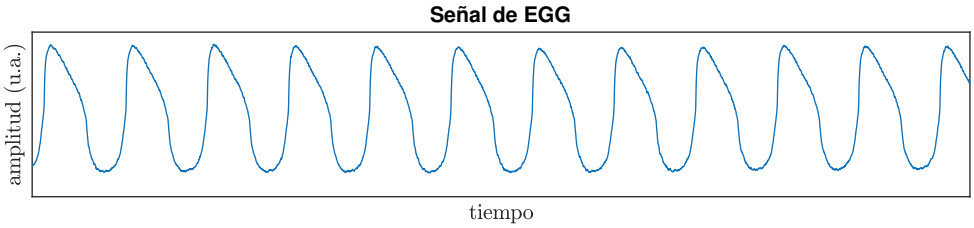
donde G es un factor de ganancia, J es el orden del modelo, $h[j]$ una secuencia de coeficientes del modelo. Normalmente $h[j]$ es estimada a partir de los coeficientes de predicción lineal *LPC* (del inglés *linear prediction coding/coefficients*) [85, 86]. Otra utilidad de conocer $H(z)$ es la de la *síntesis*, que constituiría el problema opuesto al del filtrado inverso.

Por simplicidad, se ha excluido el efecto de radiación de los labios, que puede modelarse como un tercer factor en el lado derecho de la Ecuación (2.1), usualmente mediante un filtro pasa altos:

$$R(z) = 1 - z_0 z^{-1}, \; z_0 < 1. \tag{2.4}$$



(a)



(b)

Figura 2.5: Señal de voz (arriba) y de electroglotografía (EGG, abajo), adquiridas simultáneamente. Puede apreciarse que la señal de EGG posee una morfología más sencilla.

2.3. Señal de electroglotografía

Otra señal proveniente del aparato fonador es la señal de electroglotografía, o electroglotograma (EGG). Los electroglotogramas se obtienen mediante impedanciometría ubicando un par de electrodos en el cuello, al nivel del cartílago tiroides [5, 87, 88]. Si bien algunos autores prefieren que la señal muestre el momento de cierre como un mínimo (proporcional a la impedancia estimada entre los electrodos), es común la inversión de la señal, de manera tal que el momento de mayor contacto entre los pliegues se muestre como un máximo local (proporcional a la intensidad de corriente) [89]. La Figura 2.5 muestra un ejemplo de esta señal adquirida simultáneamente con la señal de voz. En este caso, el EGG es proporcional al contacto entre las cuerdas vocales. Como puede apreciarse, su forma de onda es mucho más sencilla que la de la señal de voz.

Dada la simplicidad de su morfología, algunos parámetros relativos a la vibración de las cuerdas vocales pueden extraerse de esta señal con mayor facilidad que utilizando la señal de voz. Asimismo, es una señal auditivamente ininteligible, por lo que para algunas aplicaciones que requieran el resguardo de la privacidad del hablante podría ser de utilidad [78]. No obstante, su adquisición es aparatosa y requiere el equipo adecuado, que puede ser costoso. Asimismo, carece de utilidad para estudiar otros aspectos de la fonación relacionados con el tracto vocal, ya que esta señal está directamente relacionada con el fenómeno vibratorio de los pliegues vocales, y no porta información sobre las modulaciones producidas por las cavidades de resonancia o elementos articulantes del tracto [78].

2.4. Periodicidad y medidas de perturbación

Una función $x(t)$ es *periódica* si cumple para todo t con:

$$x(t) = x(t \pm kT),\ k \in \mathbb{Z}, \tag{2.5}$$

donde el mínimo valor de T que satisface esta expresión es llamado periodo *fundamental* de la función. No obstante el uso del término "periódica" para describir una dinámica *cíclica* es muy común en el área del procesamiento de la voz. Intentando no caer en el error que supondría llamar periódica a cualquier señal que exhiba cierto comportamiento repetitivo, se ha utilizado también el término *cuasiperiódica* [16, 42]. Sin embargo, este vocablo también tiene una definición precisa,

y se utiliza para denominar a aquellas series temporales obtenidas mediante la suma de dos series periódicas cuyos periodos no son conmensurables, es decir, la razón entre ambos periodos no es un número entero [16]. Para evitar caer en más trampas lingüísticas y de definiciones matemáticas, se propuso el término en inglés *nearly periodic*, que se traducirá aquí como *aproximadamente periódico*. De esta forma se hará referencia a señales cuyo comportamiento es en apariencia cíclico, pero que no cumplen, en rigor no *pueden* cumplir, con la definición matemática de periodicidad. Para el caso de estas señales, se hará referencia a una repetición de la forma de onda de la señal como *ciclo* en lugar de periodo, entendiendo que un ciclo es una repetición de la onda que puede estar levemente modificada, a la vez que su duración puede acortarse o alargarse en forma imperceptible [16, 42].

La diferencia entre una señal ideal, perfectamente periódica, y una aproximadamente periódica se encuentra en los cambios pequeños, microscópicos según algunos/as autores [42], que ocurren de un ciclo a otro. Dichos cambios son llamados *perturbaciones*, y pueden ocurrir en cualquier parámetro de la señal: frecuencia, amplitud o en su morfología. Otros cambios, como grandes desviaciones de la línea de base, que ocurren a lo largo de varios ciclos, son comúnmente llamados fluctuaciones. Las perturbaciones, sin embargo, pueden entenderse como desviaciones de un valor "central", que ocurren en el corto plazo. Esta idea permite asignarle a funciones aproximadamente periódicas un valor promedio de un parámetro a lo largo de todo el segmento de señal analizado, donde se hace la suposición implícita de la existencia de una distribución unimodal y simétrica respecto de dicho valor central de las perturbaciones [42].

Las perturbaciones más estudiadas son el jitter y el shimmer, que se corresponden con perturbaciones en la frecuencia fundamental y en la amplitud máxima de cada ciclo respectivamente [5]. Si bien no hay un nombre para las perturbaciones en forma de onda, se ha propuesto que la razón entre la energía de la componente armónica y la componente aperiódica de la señal es una medida de las perturbaciones de la morfología de la señal [40, 90, 91]. De todas estas, el jitter es probablemente la más estudiada [68–70]. Adicionalmente, la frecuencia fundamental es uno de los parámetros más estudiados de la voz y el habla, ya sea en la identificación de hablantes, en el modelado de la prosodia, u en otras aplicaciones [92]. Por esta razón, el jitter vocal es la perturbación más conocida y existen diversas formas de medirlo.

2.5. Jitter vocal

El jitter ha sido definido como un fenómeno consistente en la perturbación aleatoria de la frecuencia fundamental de la voz [6]. Esta definición del jitter como un fenómeno fisiológico no ha de confundirse con las distintas formas de medir esta perturbación, que se denominarán medidas de jitter. Una medida de jitter debe estimar la variación de la frecuencia fundamental en el corto plazo, por ejemplo desde un ciclo dado hasta el ciclo posterior.

Si bien el jitter es un fenómeno fisiológico, su origen no se encuentra completamente determinado. En particular se han identificado algunas fuentes de perturbación que explican la existencia del jitter, aunque pueden estar presentes individualmente o combinadas. Las fuentes de perturbación pueden ser [5]:

Neurogénicas: El estado contráctil de los músculos es controlado por descargas de motoneuronas que excitan las unidades motoras musculares. Las excitaciones musculares individuales duran una pequeña fracción de segundo, pero el tono muscular refleja la integración de cada una de esas pequeñas contracciones. Esta integración es imperfecta, produciendo pequeñas, y en apariencia aleatorias, desviaciones de un estado muscular en reposo. Estas variaciones son probablemente importantes contribuyentes al fenómeno del jitter, al operar sobre los músculos fonatorios.

Aerodinámicas: Los pulsos de aire que emergen de la glotis son eyectados a las cavidades del aparato fonador como flujos concentrados. Bajo algunas condiciones, esos pulsos de aire pueden, en forma impredecible, generar turbulencias. Este comportamiento aerodinámico errático es también una posible fuente de perturbaciones.

Mecánica: La alteración de las propiedades biomecánicas de los pliegues vocales, incluidas cambios en su masa y la estructura del tejido asociados a patologías, parecen ser una causa mayor de perturbaciones. Pero otros fenómenos mecánicos son capaces de causar alteraciones de corto plazo sobre la señal, como por ejemplo cambios rápidos en la impedancia de la vía aérea (asociado a la producción de vocales).

Estilístico: Por razones estéticas para cantantes u otros profesionales de la voz. Por ejemplo el *vibrato*.

La estimación de jitter está íntimamente relacionada con la estimación de la frecuencia fundamental. Los primeros métodos para estimar la frecuencia fundamental se basaban en la medición directa desde un osciloscopio, a partir del cual y mediante una gráfica de unos pocos ciclos, se estimaba una frecuencia fundamental promedio [5]. Estas formas fueron sustituidas por otras mas precisas y capaces de considerar cientos de ciclos. Eventualmente, con el advenimiento de la tecnología digital, surgieron algoritmos para estimar la frecuencia fundamental a partir de una serie de mediciones de la duración de cada ciclo de la señal. Estos métodos consisten en detectar sobre la señal una serie puntos fiduciarios, correspondientes con el principio y el fin de cada ciclo, como muestra la Figura 2.6. Luego, la diferencia temporal entre cada punto fiduciario determinará la duración de cada periodo.

Esta manera de estimar la frecuencia fundamental es la más básica posible, y es en cierta forma equivalente a un muestreo no uniforme de la frecuencia fundamental *instantánea*, esto es, una función que asigne a cada instante de tiempo un valor de frecuencia [41]. Es importante recordar que, para lograr detectar eficazmente los puntos fiduciarios se necesita, en primer lugar, que la señal sea aproximadamente periódica, de manera tal que los mentados ciclos existan y estén bien definidos. Como resultado se obtendrá una sucesión $\{T_j\}_{j\in\mathbb{N}^0}$, que se compone de la duración de cada ciclo de la señal en el segmento analizado, como se ilustra en la Figura 2.6. Esta sucesión también es conocida como serie de periodos, aunque un nombre más adecuado sería serie de duraciones de ciclos, a la luz de lo discutido anteriormente con respecto a la periodicidad.

2.5.1. Estimación de la serie de periodos

Como se mencionó en el Capítulo 1, existen tres familias de métodos comúnmente utilizados para encontrar los puntos fiduciarios de la señal, basadas en: la detección de cruces por cero (CC), la identificación de picos (DP) y la coincidencia de forma de onda (CF) (ver Capítulo 10 de [92]). La serie de periodos obtenida a partir de estos puntos puede tener diferencias significativas según el método utilizado para hallarlos. En efecto, es recomendable que cualquier medición de la duración de los ciclos sea acompañada de la descripción del algoritmo empleado para facilitar la comparación, así como de la sucesión de puntos fiduciarios utilizada originalmente para su cómputo [41]. Los tres métodos mencionados suelen beneficiarse del uso de interpolación para encontrar los puntos fiduciarios en forma más precisa [41, 91, 93]. Adicionalmente, contar con una estimación burda del periodo fundamental promedio $\bar{T}_0$, hallado mediante la autocorrelación o mediante estimación espectral, vuelve a estos métodos más robustos al ruido, ya que permite acotar la zona de búsqueda del punto fiduciario siguiente una vez encontrado el anterior. A continuación, se describirá cada uno de ellos.

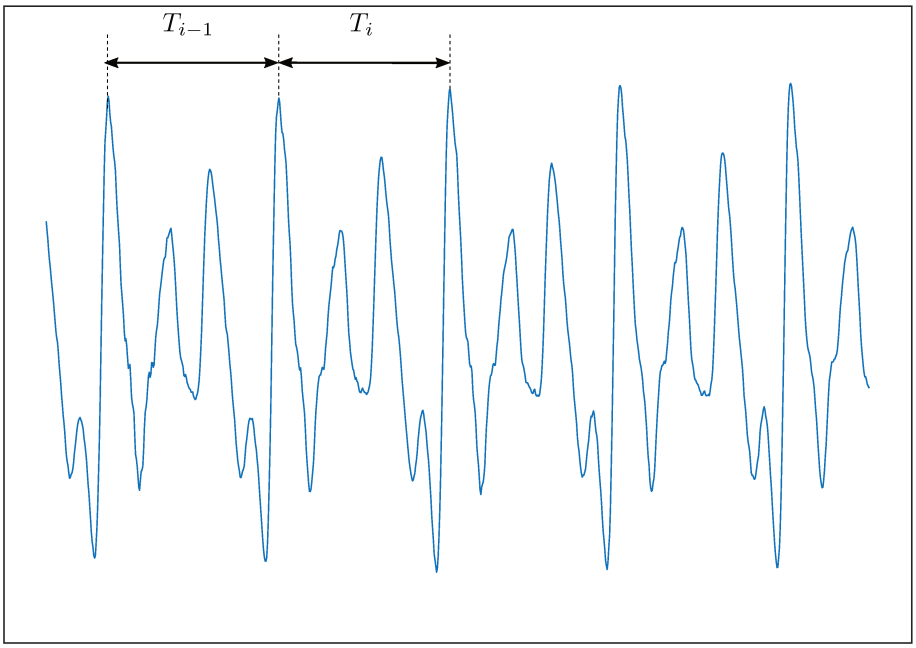


Figura 2.6: En el esquema puede apreciarse la forma de medir los ciclos T_i y T_{i-1} (para la estimación del jitter).

Método de coincidencia de forma de onda (CF)

Suponiendo una señal aproximadamente periódica, cada ciclo sufrirá ciertos cambios, pequeños idealmente, respecto al ciclo anterior y el ciclo siguiente. Considerando esta situación, este método utiliza una medida del parecido entre el ciclo anterior y el que sigue, considerando el ciclo siguiente como aquel comprendido entre el último punto fiduciario y el candidato actual. De esta forma, el punto fiduciario siguiente se busca en un intervalo dado, para cuyos puntos una función de parecido es calculada. Existen diversas opciones para la función de parecido, pero comúnmente el error cuadrático medio [94] es utilizado con este fin. Éste método es más robusto frente al ruido que los que se describen a continuación, aunque no es capaz de responder a cambios rápidos en el periodo de la señal [41, 72].

Método de cruces por cero (CC)

Consiste en buscar los cruces por cero de la señal en la dirección negativa o positiva. Dado que este método puede aplicarse sólo si hay dos cruces por cero por ciclo, debe aplicarse un filtrado pasabajos que conserve a la F_0 previo a la búsqueda de los cruces [91]. Este filtrado es de vital importancia cuando se utiliza con la señal de voz, ya que esta señal suele cruzar múltiples veces por cero en un mismo ciclo. También es posible tomar otro valor distinto de cero sobre el que identificar el cruce, que puede utilizarse junto con la detección del flanco descendente de picos más prominentes en la señal [91, 92].

Método de detección de picos (DP)

Esta técnica consiste en buscar los máximos locales de la función que, a su vez, sean máximos absolutos dentro del ciclo analizado, de allí el nombre del método. En ocasiones se prefiere el uso de los mínimos locales, dado que en muchas señales de voz son más pronunciados. De los tres métodos aquí descriptos para encontrar los puntos fiduciarios de segmentación, este método es el más afectado por el ruido en la señal de cualquier origen [41, 91]. Este defecto se ve exagerado

en el caso de señales de menor frecuencia fundamental o con picos más anchos, ya que pequeñas cantidades de ruido alteran la posición del máximo local fácilmente en estas situaciones [93]. Es por esto que, para mejorar su capacidad de detección, el uso de esta herramienta suele ser acompañado de alguna técnica de preprocesamiento como *clipping* central [92]. Opcionalmente, puede aplicarse un filtro pasabajos, no necesariamente tan exigente como el utilizado para CC, con el objetivo de reducir algo de ruido de alta frecuencia y mejorar la estimación de los máximos locales.

2.5.2. Medidas de jitter

El nivel de jitter puede medirse de diversas maneras. Se ha propuesto en la literatura un número de medidas para estimar el jitter, con diversas variaciones, pero que se basan en la secuencia de periodos $\{T_j\}_{j=1}^{M}$, donde M es el número de ciclos, para estimar el nivel de perturbación de la frecuencia. Algunas de ellas son:

Factor de perturbación Es el promedio del valor absoluto de las perturbaciones [95]:

$$FP = \frac{1}{M-1} \sum_{j=1}^{M-1} |T_{j+1} - T_j|. \tag{2.6}$$

Factor de perturbación direccional Esta medida de perturbación es diferente a todas las otras ya que descarta la magnitud de las perturbaciones pero conserva el *signo* [96]. Para calcularla, primero es necesario encontrar la serie de periodos $\{T_j\}_{j=1}^{M}$ y luego calcular la diferencia entre la duración de ciclos sucesivos, de igual manera que para el factor de perturbación, pero sin aplicar el valor absoluto. Luego, el factor de perturbación direccional se calcula como el cociente entre el número de veces que se produjo un cambio de signo en la perturbación y el número de perturbaciones $(M-1)$.

Razón de jitter Es la medida más sencilla para adaptar el valor de la perturbación promedio al valor del periodo fundamental promedio [5]:

$$RJ_{\times 1000} = \frac{\frac{1}{M-1} \sum_{j=1}^{M-1} |T_{j+1} - T_j|}{\frac{1}{M} \sum_{j=1}^{M} T_j} \times 1000, \tag{2.7}$$

donde el denominador es simplemente el promedio del periodo fundamental. Por lo tanto esta medida puede entenderse como el promedio de las perturbaciones normalizado por la longitud promedio de los ciclos. El factor "$\times 1000$" es el original utilizado en su definición, pero actualmente se prefiere expresar esta medida en porcentaje, dando lugar al jitter relativo porcentual, que se enuncia más adelante.

Índice de Variabilidad del periodo Este enfoque para cuantificación de las perturbaciones del periodo está basado en medidas de estadística descriptiva, puntualmente el coeficiente de variación [97]. Al aplicar este estadístico sobre la serie de periodos se obtiene:

$$IVP = \frac{\frac{1}{M-1} \sum_{j=1}^{M-1} (T_j - \bar{T}_0)^2}{\bar{T}_0^2} \times 1000, \tag{2.8}$$

donde $\bar{T}_0$ es el promedio de los elementos de la serie de periodos.

Jitter relativo porcentual También denominada simplemente jitter relativo, es una de las medidas de perturbación de frecuencia más difundida [16, 41, 49, 67, 67–69, 71, 72]. Es idéntica a la Razón de jitter pero expresada en porcentaje:

$$jitter_{\%} = \frac{\frac{1}{M-1}\sum_{j=1}^{M-1}|T_{j+1} - T_j|}{\frac{1}{M}\sum_{j=1}^{M} T_i} \times 100\,\%. \tag{2.9}$$

Esta medida de jitter será utilizada luego en el Capítulo 6. Es posible obtener una cota superior para el error cometido mediante el jitter relativo. La exactitud con la que se puede medir cualquier punto fiduciario es de $2/f_s$, donde f_s es la frecuencia de muestreo utilizada. Cuando la F_0 crece, la duración de los ciclos se acorta y por eso el número de muestras presentes por ciclo disminuye. Dado que el error de medición, $2/f_s$, permanece igual, el error relativo a la magnitud de la perturbación crece proporcionalmente a la frecuencia fundamental. El error relativo máximo depende entonces de F_0 y f_s, y se expresa en porcentaje como [5]:

$$error_{jitter} = 50\frac{F_0}{f_s}\,\%, \tag{2.10}$$

lo que identifica formalmente una de las dificultades típicas en la medición de jitter relativo: su estimación empeora con el aumento de la frecuencia fundamental.

Perturbación promedio relativa (RAP) y Cociente de perturbación de periodos (PPQ5): Ambas medidas, definidas como [49, 98]:

$$RAP = \frac{\frac{1}{M-2}\sum_{j=2}^{M-1}\left|\frac{T_{j-1}+T_j+T_{j+1}}{3} - T_j\right|}{\frac{1}{M}\sum_{j=1}^{M} T_j} \tag{2.11}$$

y

$$PPQ5 = \frac{\frac{1}{M-4}\sum_{j=3}^{M-2}\left|\frac{T_{j-2}+T_{j-1}+T_j+T_{j+1}+T_{j+2}}{5} - T_j\right|}{\frac{1}{M}\sum_{j=1}^{M} T_j}, \tag{2.12}$$

son similares al jitter relativo, pero en lugar de utilizar la diferencia entre la duración de ciclos sucesivos, hallan la diferencia entre la duración de cada ciclo y la duración promedio de sus vecinos más cercanos en la serie de periodos. Por ejemplo, para RAP se calcula el promedio entre las duraciones del ciclo anterior, el ciclo actual y el siguiente (tres ciclos). Para PPQ5 se utilizan el actual, sus dos antecesores inmediatos y los dos siguientes (5 ciclos) . Otras medidas como PPQ7, PPQ11, emplean un número mayor. Este procedimiento es equivalente a encontrar la diferencia entre la serie y un promedio móvil de un número variable de ciclos. Dado que un promedio móvil se comporta como un filtro pasabajos, estas medidas miden la desviación de la serie de periodos respecto de una versión suavizada de la misma.

2.5.3. Modelos de jitter

El modelado del jitter permite el estudio de este fenómeno para distinguirlo de otros tipos de fluctuaciones en la frecuencia fundamental, principalmente el trémolo o los microtemblores, al intentar comprender cómo se produce, introduciendo modificaciones en los modelos de vibración de las cuerdas vocales [16]. La utilidad de estos modelos se extiende a su vez a la síntesis de voces,

en donde una pequeña cantidad de jitter mejora la naturalidad percibida de las voces sintéticas [16, 99]. Al mismo tiempo, estos modelos tienen utilidad para la calibración de algoritmos de procesamiento, como los descriptos para la segmentación en ciclos. Esta última aplicación es la que se utilizará en este trabajo, especialmente en el Capitulo 6, donde voces sintetizadas con una cantidad de jitter conocida se emplearán para evaluar el desempeño de un método para la estimación de la perturbación de la frecuencia fundamental.

Para modelar el jitter, es necesario conocer sus propiedades. Se repasan algunas de ellas a continuación [16]:

1. La distribución de probabilidad de las perturbaciones en la duración de cada ciclo es aproximadamente Gaussiana.

2. Habitualmente el rango del jitter relativo está entre $0.1\,\%$ y $1\,\%$ (para voces sanas).

3. Las perturbaciones en la duración de ciclos adyacentes están correlacionadas positivamente. Esto implica que la perturbación actual depende de aquellas pasadas.

4. La fuente más probable de la correlación observada entre las perturbaciones es un fenómeno conocido como *microtemblores*, que consiste en una modulación de la frecuencia fundamental cuya frecuencia media oscila entre 7 y 10 Hz, aunque puede ser inferior o superior en muchos casos.

5. El jitter parece ser un fenómeno genuinamente estocástico.

6. En promedio, las perturbaciones incrementan con la duración de los periodos, es decir, para voces con baja frecuencia fundamental las perturbaciones son mayores.

7. El nivel de jitter aumenta en presencia de patologías laríngeas.

Con respecto a la propiedad 3, cabe destacar que el orden de la correlación es aproximadamente dos para hablantes sanos. Es decir, la perturbación actual depende del valor de, al menos, las dos perturbaciones pasadas. No obstante, en general, los órdenes se encuentran entre uno y nueve para la mayoría de los sujetos [16, 100]. La propiedad 5 se desprende de observar que al remover las correlaciones lineales de una serie de perturbaciones, el residuo restante es puramente estocástico [100]. Esto puede realizarse modelando una serie de perturbaciones mediante un modelo (lineal) autorregresivo y substrayendo la serie modelada de la original. Como resultado se obtiene un residuo que, en caso de no existir correlaciones no lineales será puramente estocástico, es decir, no correlacionado y con un espectro aproximadamente plano. Caso contrario, indicaría la presencia de correlaciones *no lineales*, posiblemente explicadas por una dinámica caótica. Este procedimiento es equivalente a pensar que las perturbaciones se obtienen al filtrar una serie de periodos mediante un sistema autorregresivo. Dado que un sistema lineal sólo puede tener una salida aleatoria cuando la entrada es estocástica, los resultados obtenidos en [100] justifican el modelado de la serie de periodos como una serie temporal estocástica filtrada por un sistema lineal. Asimismo, experimentos con un modelo fisiológico que vinculan las propiedades estadísticas de las microcontracciones musculares con desviaciones de la frecuencia instantánea de vibración [7] confirman indirectamente los resultados encontrados por [100]. Los microtemblores mencionados en la propiedad 4 se deben a un fenómeno de origen desconocido, pero estudiado cuantitativamente [101], y diferente del jitter, fundamentalmente en su frecuencia. Mientras que este último es una perturbación ciclo a ciclo, de corto plazo y frecuencia más alta, los microtemblores están caracterizados por vibraciones entre 1 y 15 Hz [102]. Para sujetos identificados como masculinos, la mediana de esta frecuencia suele ser de 6 Hz, y de 5 Hz para hablantes identificados como femeninos. A grandes rasgos, la frecuencia de modulación de los microtemblores determina el grado de correlación entre las perturbaciones [16].

A continuación se describirán dos modelos simples de jitter que reflejan un subconjunto de las propiedades expuestas anteriormente. El primer modelo intenta reflejar la naturaleza estocástica del jitter así como también la distribución Gaussiana encontrada para las perturbaciones. El segundo modelo, además de lo tenido en cuenta en el primero, permite modelar la correlación positiva entre las perturbaciones de periodos adyacentes mencionadas en las propiedades 3, 4 y 5. Como salida de ambos modelos se obtiene una secuencia de periodos cuyo jitter relativo porcentual es conocido. Más adelante, se utilizarán señales sintetizadas con dichos modelos, particularmente en el Capítulo 6, donde se emplearán para estimar el desempeño de un método novedoso para la medición del jitter.

Modelo 1: duración de periodos independiente e idénticamente distribuida

Si se considera la duración de cada periodo T_j como una variable aleatoria con distribución Gaussiana $\mathcal{N}(\bar{T}_0, \sigma_T^2)$ [103, 104], podemos hallar la distribución de probabilidad de

$$|\Delta T_j| = |T_{j+1} - T_j| \tag{2.13}$$

como:

$$\begin{cases} \mathcal{N}(0, 2\sigma_T^2) & \text{para } |\Delta T_j| \geq 0. \\ 0 & \text{para } |\Delta T_j| \leq 0. \end{cases} \tag{2.14}$$

En consecuencia, el valor esperado $E\{|\Delta T_j|\}$ está dado por:

$$E\{|\Delta T_j|\} = \frac{2}{\sqrt{\pi}}\sigma_T. \tag{2.15}$$

Teniendo en cuenta que el numerador de la Ecuación (2.9) es un estimador muestral del valor esperado de $|\Delta T_j|$, podemos reemplazar esta estimación por el valor esperado dado en la Ecuación (2.15) considerando que el proceso estocástico que origina $|\Delta T_j|$ es *ergódico*. Esto implica que los estadísticos obtenidos a partir de una realización pueden reemplazarse por sus equivalentes poblaciones. De esta forma, obtenemos:

$$jitter = \frac{2\sigma_T}{\sqrt{\pi}\bar{T}_0} \times 100\,\%, \tag{2.16}$$

de donde concluimos que es posible obtener una secuencia de periodos con un valor de jitter: $jitter = jitter_{\text{deseado}}$ generando una secuencia de periodos con una distribución Gaussiana de media $\bar{T}_0$ y

$$\sigma_T = jitter_{deseado}\frac{\sqrt{\pi}\bar{T}_0}{200}, \tag{2.17}$$

donde $\bar{T}_0$ es un parámetro de entrada junto con $jitter_{deseado}$ (en porcentaje).

Modelo 2: duración de periodos correlacionada

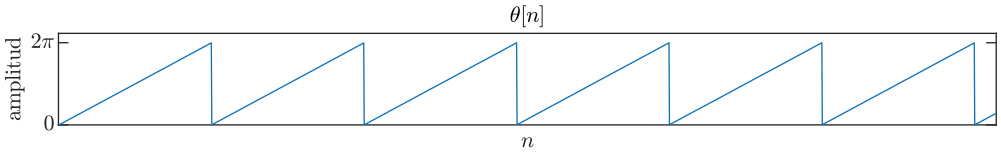
Para este modelo, una versión discreta de la fase de una señal de voz $\theta[n]$, donde n es el índice temporal, es modelada como [16]:

$$\theta[n] = \theta_0 + \frac{2\pi n}{\bar{T}_0 f_s} + 2\pi\sum_{j=1}^{n} y[j-1] \tag{2.18}$$

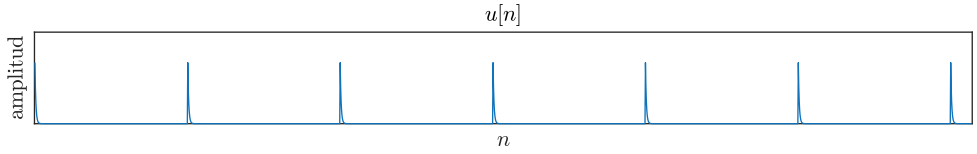
con

$$y[n] = a_1 y[n-1] + a_2 y[n-2] + a_0 be[n] \tag{2.19}$$

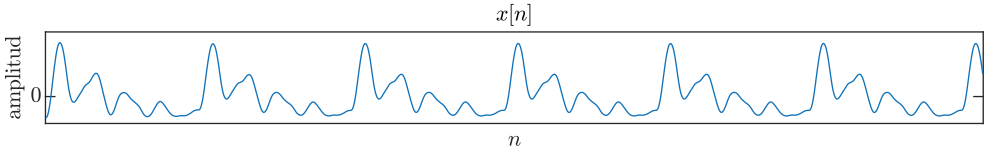
donde $0 < a_i < 1$, y f_s es la frecuencia de muestreo utilizada para la síntesis. El valor de b puede modificarse para producir voces con valores de jitter deseado, mientras que los valores de a_0, a_1 y



(a) Función de fase $\theta[n]$ discreta generada con el Modelo 2.



(b) Función glótica $u[n]$ construida a partir de $\theta[n]$.



(c) Señal de voz sintética $x[n]$.

Figura 2.7: Síntesis de voces con jitter conocido.

a_2 pueden elegirse para adecuarse a un valor de frecuencia y ancho de banda de los microtemblores. $e[n]$, en el tercer término, es una realización de un proceso aleatorio de dos puntos con media nula definido como

$$e[n] = \sqrt{f_s^{-1}}W, \text{ para cada } n \in \mathbb{Z}. \tag{2.20}$$

donde W puede valer 1 o -1 con probabilidad igual a 0.5 para cada valor. Por la tanto la Ecuación 2.19 se corresponde con un filtro autorregresivo de orden 2, en cuya entrada se aplica una secuencia estocástica.

La Ecuación 2.18 puede interpretarse de la siguiente manera. El primer término corresponde a un valor de fase inicial, que generalmente será igual a 0. El segundo término corresponde a un factor lineal que aumenta su valor con n a una tasa fija. El tercer término, se corresponde con la salida del filtro autorregresivo descripto. Por la influencia del segundo término, $\theta[n]$ es creciente en el largo plazo, con una gráfica similar a una rampa. Cada vez que $\theta[n]$ supera el valor 2π, un nuevo ciclo comienza en la señal de voz. El tercer término agrega un rizado aleatorio sobre dicha rampa, de manera tal que la cantidad de tiempo que tarda $\theta[n]$ en alcanzar un múltiplo de 2π varía cada vez, produciendo periodos de distinta duración correlacionados, satisfaciendo las propiedades descriptas más arriba.

Basándose en esto, cada periodo T_k está asociado con el número N_k de muestras necesarias para llevar a $\theta[n]$ desde $2(k-1)\pi$ hasta $2k\pi$, de manera tal que $\theta[N_k] = 2k\pi$. De esta forma, la secuencia de periodos obtenida mediante el Modelo 2, puede definirse como:

$$T_k = \bar{T}_0 - \bar{T}_0 \sum_{j=1}^{N_k} y[j-1]. \tag{2.21}$$

2.6. Clasificación de voces en tipos 1, 2 y 3

Al mencionar la frecuencia fundamental de una señal, F_0, se asume que es posible describir con este único valor la cantidad de veces por unidad de tiempo que se repite una determinada forma de onda de la señal. Como se vio anteriormente, resulta imposible para una señal real ser estrictamente periódica, por lo que al mismo tiempo no es posible que su frecuencia sea exactamente F_0. No

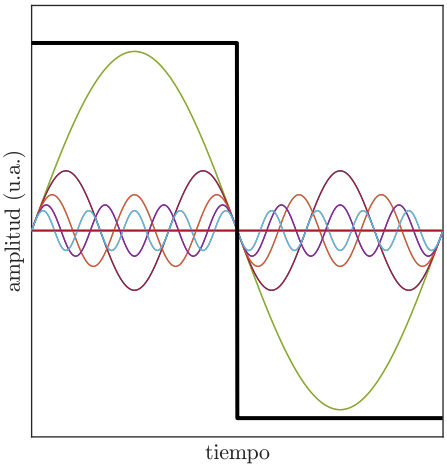


Figura 2.8: Señal cuadrada (en negro) obtenida, por ejemplo, mediante la apertura o cierre de un interruptor. En colores se observan las ondas senoidales obtenidas a partir de su serie de Fourier.

obstante, para señales aproximadamente periódicas, utilizar un valor de F_0 para describirlas no resulta del todo inadecuado, ya que es *aproximadamente* correcto. Podría decirse que si los periodos de una oscilación senoidal del tipo:

$$y(t) = \text{sen}(2\pi F_0 t) \tag{2.22}$$

coinciden aproximadamente con los ciclos de una señal, entonces puede describírsela mediante la frecuencia de $y(t)$, es decir, F_0. Este tipo de modelo senoidal se aplica implícitamente cada vez que se hace referencia a *la* frecuencia de la señal. Resulta conveniente recordar, entonces, que es el modelo el que tiene características de frecuencia, y no la señal en sí misma [41].

Aquí vale la pena hacer una corta digresión para resaltar la diferencia entre señal y modelo, que se discute también en otros textos [41, 105]. El caso descripto guarda cierta semejanza con lo que sucede al analizar una señal cuadrada obtenida mediante la medición de la corriente en un conductor, cerrando y abriendo un interruptor que permite la circulación eléctrica o la imposibilita. Al estudiar su serie de Fourier observaremos que, según este modelo, la interferencia destructiva de un número infinito de ondas senoidales genera un valor nulo en el momento de "apagado" (ver Figura 2.8). Aunque matemáticamente este análisis es correcto, sabemos que lo que sucede en realidad es que allí *no* hay circulación de corriente [105]. Por lo tanto, si un modelo no es utilizado conscientemente para estudiar la realidad, podría conducir a conclusiones equivocadas.

Si bien esta distinción es sutil, y en general no se hará diferencia alguna entre modelo y señal en este sentido, a la hora de discurrir sobre la periodicidad y sus perturbaciones adquiere una importancia mayor. Para una señal aperiódica, para la que no es posible "ajustar" una senoidal que coincida con sus ciclos, la idea de frecuencia se vuelve ambigua [41]. En consecuencia, aplicar una herramienta para analizar una señal aproximadamente periódica, cuando en realidad no lo es, puede resultar en conclusiones erradas sobre el fenómeno físico vibratorio subyacente a la señal, ya que los parámetros obtenidos mediante este análisis no se ajustarían a la realidad [38].

La posibilidad de contar con la tecnología digital para el cómputo de un número bastante grande de medidas de perturbación, como las mencionadas anteriormente, volvió este tipo de descriptores muy populares. No obstante, su aplicación en la práctica clínica sobre voces que no son aproximadamente periódicas llevó a cuestionar su utilidad [43], ya que el *software* utilizado con ese fin puede entregar valores para medidas de perturbación sin tener en cuenta su periodicidad. Con el fin de evitar la aplicación de medidas de perturbación sobre voces que no sean aproximadamente periódicas, se ideó una clasificación en tres tipos, propuesta por Titze [42], basada en la regularidad temporal. A continuación se detallará la clasificación y se definirán los tipos que la componen.

2.6.1. Tipificación de voces

En el sistema propuesto las voces se clasifican en tres tipos cuyas definiciones se transcriben a continuación [42]:

- Tipo 1: Señales aproximadamente periódicas que no muestran cambios cualitativos en el segmento de análisis, con frecuencias modulantes o subarmónicas cuyas energías sean de un orden de magnitud inferior a la de la frecuencia fundamental, o nulas.
- Tipo 2: Señales con cambios cualitativos en el segmento de análisis, o señales con frecuencias subarmónicas y/o modulantes cuyas energías se aproximen en magnitud a la energía de la frecuencia fundamental. No existe, por lo tanto, una única frecuencia fundamental obvia en el segmento analizado.
- Tipo 3: Señales sin estructura periódica aparente.

Algunas voces tipo 2 y tipo 3 pueden presentar también oscilaciones subarmónicas. Bajo circunstancias aún poco comprendidas, en ocasiones la fonación es caracterizada por un movimiento oscilatorio con diferente frecuencia entre los pliegues vocales, normalmente asociado a la presencia de pólipos o parálisis unilateral de las cuerdas vocales, lo que produce una alternación regular de dos ciclos ligeramente diferentes [5, 106]. Dado que esta situación produce una fuga de aire excesivo a través de la glotis, se asocia comúnmente con la percepción auditiva de voz ronca [106]. Asimismo, genera una situación en la que los puntos fiduciarios de inicio y fin de ciclo se vuelven indeterminados. Por ejemplo, en los paneles izquierdo superior e inferior de la Figura 2.9 se observa una señal con oscilaciones subarmónicas. Una partición en ciclos posible podría ser entre cada marca "+" y "o". Otra forma podría ser entre marcas "+", obviando los puntos "o". En ese último caso, los ciclos formados durarían el doble que en el primer caso (considerando tanto los puntos "+" como los puntos "o"). Por esta razón, este tipo de oscilación subarmónica es conocido como *duplicación de periodo*, bifurcación o diplofonía. Dada esta indeterminación, la aplicación de medidas de perturbación en estas señales es problemática [107]. Este patrón de vibración surge en presencia de una frecuencia con la mitad de la frecuencia fundamental como puede verse en el espectrograma de la Figura 2.9. En general, es posible la presencia de subarmónicos del orden F_0/K, y su presencia ha sido asociada a dinámicas *caóticas* [5, 42].

Algunos autores agregan algunos *subtipos* a cada uno de los tipos definidos, aunque su utilización no es tan difundida [5]:

- Tipo 1:
  - Subtipo A: Existe una variación menor aleatoria de F_0 o su forma de onda.
  - Subtipo B: Se produce una modulación monotónica (F_0 aumenta o decrece).
  - Subtipo C: Presenta una modulación leve en la amplitud.
- Tipo 2:
  - Subtipo B: Existen discontinuidades en la amplitud
  - Subtipo C: Presenta grandes modulaciones de la (F_0).
- Tipo 3:
  - Subtipo A: Sin estructura observable.
  - Subtipo B: Caótica.

Basándose en esta clasificación, se recomienda la utilización de análisis de perturbación únicamente para las voces tipo 1. Para las señales tipo 2 y 3 se recomienda el uso de técnicas *visuales*, como espectrogramas o retratos de fase. En el caso de voces Tipo 3 también son útiles las escalas perceptivas (como las escalas GRBAS o CAPEV [40, 50]) . En la práctica clínica fonoaudiológica,

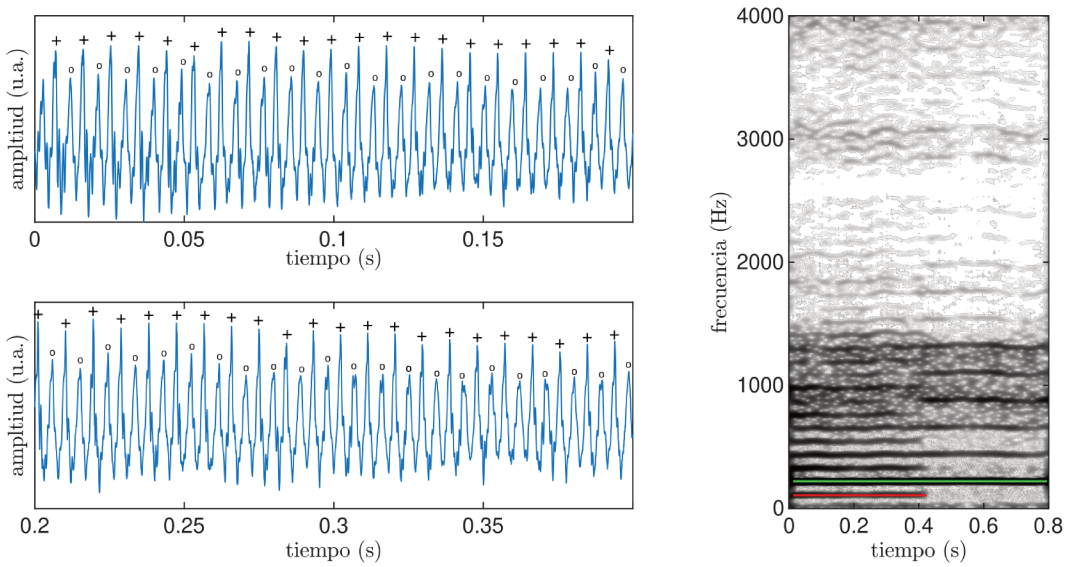


Figura 2.9: Representación temporal (izquierda) y espectrograma (derecha) de una señal tipo 2 con duplicación de periodo. Puede observarse en el espectrograma la coexistencia de la frecuencia fundamental (en verde) y un subarmónico (en rojo) hasta aproximadamente la mitad de la duración analizada.

la clasificación de voces en los tres tipos propuestos, también llamada *tipificación de voces*, se realiza mediante la inspección visual de la representación temporal de la señal, la evaluación de la impresión auditiva y un espectrograma, a partir del cual es posible observar fácilmente el contenido frecuencial en cada instante de tiempo (ver Capítulo 3). La tipificación de voces en los tres tipos se ha vuelto una práctica habitual y obligada previo al uso de cualquier medida de perturbación [51–58]. También se ha utilizado como una medida de la calidad vocal y se ha propuesto su uso como forma de monitoreo de la progresión de terapias sobre el aparato fonador [58, 65, 108].

2.6.2. Un cuarto tipo de voces

Una corriente importante de autores se ha dedicado a estudiar el uso de medidas provenientes del análisis de dinámicas no lineales, particularmente la *dimensión de correlación* [4], D_2, como parámetros relevantes para la caracterización de la voz [59]. En contraste con las medidas de perturbación, los parámetros utilizados en el análisis de dinámicas no lineales no requieren una estimación de la frecuencia fundamental [109]. Como desventaja, si el comportamiento de la voz es predominantemente estocástico entonces D_2 tiende a infinito, por lo que su estimación se vuelve imposible. D_2 tiene dimensión finita cuando el comportamiento de la voz es *caótico*, es decir, sigue una dinámica determinística pero impredecible a largo plazo debido fundamentalmente a dos condiciones: 1) No linealidad del sistema que la genera y 2) Sensibilidad a las condiciones iniciales [4]. No obstante, una señal caótica *no* es estocástica, y una diferencia fundamental entre ambas es una dimensión de correlación finita para el caso de señales caóticas.

Con el objetivo de distinguir aquellas señales para las que es posible calcular D_2, Sprecher y cols. [110] optaron por redefinir las voces tipo 3, separando en dos esta categoría. En primer lugar, el nuevo tipo 3 abarcaría sólo a aquellas señales de naturaleza *caótica*, con D_2 finita. En segundo lugar, se agrega un cuarto tipo (tipo 4) que incluiría únicamente a aquellas voces cuyo comportamiento es *estocástico*. Por lo tanto, esta nueva categorización plantea el problema, de larga data, de distinguir una dinámica caótica (tipo 3) de una estocástica (tipo 4), que para un observador humano puede resultar una tarea bastante difícil, si no imposible [16]. Por ejemplo,

una señal caótica con dimensión muy alta puede ser indistinguible de una señal estocástica para un observador utilizando los métodos habituales (percepción auditiva, espectrogramas, etc.).

Teniendo en cuenta lo anterior, la distinción planteada entre señales caóticas y estocásticas dentro de las señales tipo 3 se corresponde al interés de utilizar D_2 como parámetro. No tiene ningún efecto directo en términos del uso de medidas de perturbación, ya que no pueden emplearse ni en señales caóticas ni en aquellas predominantemente estocásticas de todas maneras. Sin perjuicio de la existencia del cuarto tipo de señales, se utilizará en el Capítulo 5 la clasificación original en tres tipos, detallada anteriormente.

2.6.3. Características propuestas para una clasificación automática

Las definiciones de los tipos de voz descriptos carecen de información cuantitativa que permita clasificar una señal de forma completamente unívoca. Por ejemplo, si bien se describe que las señales tipo 2 pueden tener modulaciones de amplitud o frecuencia, no se especifican valores para la magnitud de la modulación, o cómo calcularla. Tampoco se especifica qué ocurre cuando existe alguna modificación importante pero de corto plazo en la señal, fuera de la cual podría considerarse aproximadamente periódica. Aunque algunos de estos aspectos fueron considerados en los subtipos descriptos, no se sugiere una forma cuantitativa para abordar la clasificación en cada caso. Adicionalmente, no existe un consenso generalizado en cuanto a los valores que deberían tomar las variables de los espectrogramas utilizados para la clasificación. Si bien se suelen utilizar los valores sugeridos por Sprecher y cols. [110]:

- **Forma de la ventana:** Hamming
- **Ancho de ventana:** 50 ms
- **Paso de tiempo:** 0.002 s
- **Paso de frecuencia:** 5 Hz
- **Rango Dinámico:** 40 dB

no todos los autores los emplean y, en muchos casos, incluso no se reportan los valores utilizados. La Figura 2.10 muestra espectrogramas graficados con los valores indicados.

Lo anterior redunda en que la tipificación realizada por un/a especialista sea subjetiva, dependiente de su formación (fonoaudiólogos/as, otorrinolaringólogos/as, etc), su experiencia, su estado de ánimo y/o su comprensión de las herramientas utilizadas para la clasificación [73]. En consecuencia, se ha observado cierta variación interprofesional [63, 108]. Adicionalmente, la clasificación suele ser una tarea que requiere de un tiempo considerable para muchos especialistas. Por estas razones, existe interés en obtener medidas capaces de distinguir en forma cuantitativa entre los tres tipos de señales, lo que resultaría en una clasificación más objetiva. Este interés no es exclusivo del problema de tipificación de voces, de hecho es un problema tradicional de la ingeniería biomédica traducir criterios subjetivos, generalmente clínicos, en parámetros cuantitativos más objetivos para el diagnóstico, cribado o monitoreo. Particularmente, dado que el estudio de la salud de la voz se realiza predominantemente en forma perceptual, se han desarrollado sistemas capaces de aplicar una estrategia de reconocimiento de patrones para hacer más objetivas las apreciaciones perceptuales de la voz. Dichos sistemas de Análisis Automático de la Condición de la Voz (AVCA, del inglés *Automatic Voice Condition Analysis*) [73, 74], como se los conoce, utilizan descriptores cuantitativos de la calidad de la voz en el marco de una estrategia de aprendizaje supervisado para intentar emular el criterio de los especialistas mediante un algoritmo de clasificación automática.

Con el objetivo de generar parámetros capaces de describir en forma cuantitativa los distintos tipos de voces presentados anteriormente, se han propuesto un número de nuevos parámetros capaces de diferenciar en algún grado los distintos tipos de voz. A continuación se describen algunas de estas medidas.

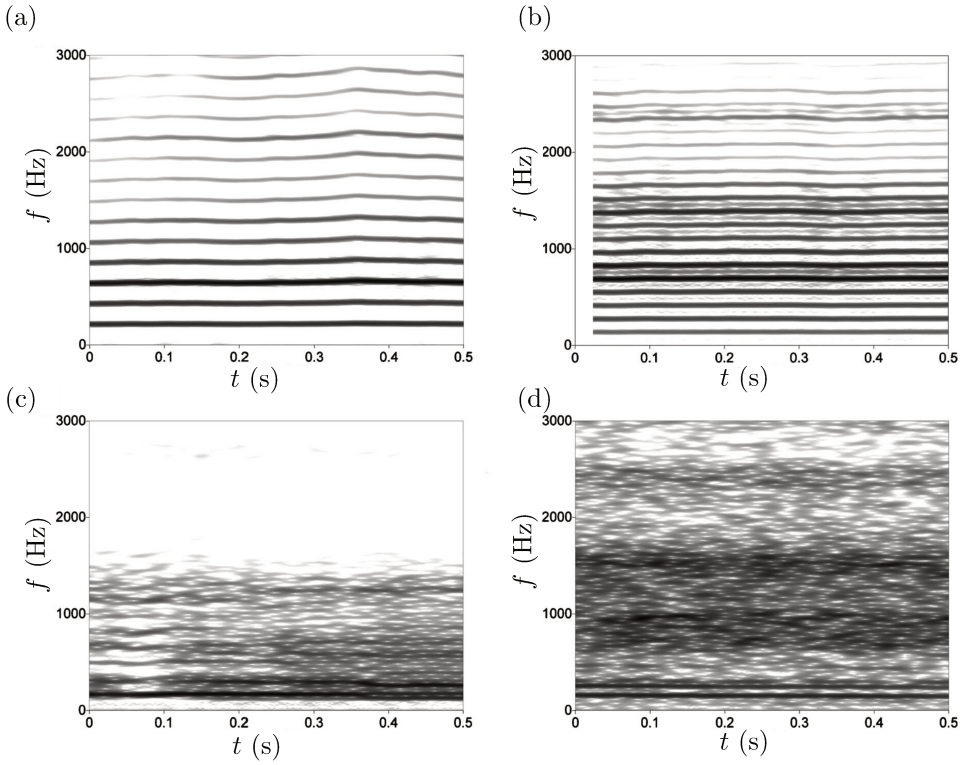


Figura 2.10: Ejemplos de espectrogramas para cada tipo de voz. a) Tipo 1, b) Tipo 2, c) y d) Tipo 3 (modificado de [110]).

En [59], los autores utilizaron la dimensión de correlación D_2 para 122 señales, empleando el método de Grassberger-Procaccia [111] que requiere la intervención de un operador altamente entrenado, para la obtención de D_2. Se observó que el valor de este parámetro incrementa con el tipo de la señal, siendo las señales tipo 1 aquellas con menor dimensión de correlación y las tipo 3 aquellas con mayor valor de D_2. Se encontraron diferencias estadísticamente significativas entre los valores que toma D_2 para cada tipo de señal, aunque los resultados para separar las voces 1 de las tipo 2 son equivalentes a la separación realizada con características ya conocidas, como medidas de jitter y shimmer.

En [60], se propuso una medida basada en la transformada de Fourier de tiempo corto (ver Capítulo 3) para cuantificar la proporción de ruido en la señal generado por turbulencias en el flujo aéreo, sobre la hipótesis de que cuanto más alto sea el índice del tipo de señal, mayor será el nivel de ruido presente en ella. Basados en esto, aplicaron una medida denominada Razón de Convergencia Espectral (SCR, del inglés *spectrum convergence ratio*) para 148 señales. Nuevamente, se reportaron diferencias estadísticamente significativas entre los valores que toma SCR para cada tipo de señal.

Volviendo a las medidas de dinámicas no lineales, en [62] se propone un método basado en el cálculo de coeficientes de Lyapunov, Tasa de Divergencia (ROD, del inglés *rate of divergence)*, que, de acuerdo a los resultados reportados, también muestra diferencias estadísticamente significativas para cada tipo de señal (se utilizaron 147 señales de [112] en este caso), excepto entre los tipo 2 y 3. Asimismo, el cálculo de coeficientes de Lyapunov no está exento de dificultades, por lo que la aplicabilidad real del método es cuestionable [4].

En [61] se propone la utilización de 3 estadísticos de orden superior, además de 4 parámetros

acústicos entre los que se encuentran medidas de jitter y shimmer, para diferenciar entre los distintos tipos de señales empleando árboles de decisión. La utilización de estos 7 parámetros lleva a una mejora en la identificación correcta de las voces tipo 1 respecto al uso de parámetros acústicos solamente, elevando de $63\,\%$ a $85.71\,\%$ el porcentaje de señales correctamente clasificadas de ese tipo. También mejora la identificación de voces tipo 2 en la misma medida. En contraposición, este método confunde voces tipo 3 con tipos 1 y 2. Los porcentajes de clasificación correcta fueron de $85.71\,\%$ para voces tipo 1 y 2, $62.86\,\%$ para tipo 3, y $94.29\,\%$ para tipo 4. Este trabajo constituye el primer antecedente del uso de una herramienta de reconocimiento de patrones, junto con características propuestas especialmente para el problema, para la tipificación automática de señales. No obstante, el uso de un conjunto de señales pequeño (140 señales) y no accesible públicamente hace difícil la reproducción del experimento y su replicación en nuevos conjuntos de datos, así como la comparación de resultados obtenidos sobre otras bases de datos.

Una nueva medida llamada Razón de la diferencia de energía no lineal (NEDR) [63], que busca caracterizar cómo varía con el tiempo la distribución de la energía en el espectro de la señal, fue presentada bajo la hipótesis de que las señales 1 presentan una distribución de dicha energía más estable a lo largo del tiempo, mientras que los tipos 2 y 3 tienen distribuciones progresivamente menos estables. Al igual que en trabajos anteriores, se utilizaron 135 señales provenientes de [112] y se hallaron diferencias estadísticamente significativas entre los valores de NEDR para todas las clases.

En [64] se propone el cálculo de la dimensión intrínseca D_i, que es análoga a la dimensión de correlación pero calculada para distintos segmentos o ventanas temporales en la señal, para diferenciar entre los distintos tipos. Bajo la hipótesis de que los diferentes segmentos de la señal pueden clasificarse individualmente en los cuatro tipos de voces, los autores definen cuatro cantidades llamadas Componente de Tipo de Voz, VTC_i (del inglés *Voice Type Component*) donde $i = 1, ..., 4$ es el tipo de voz correspondiente. VTC_i es la proporción de segmentos del tipo i presentes en la señal, y la agregación de los cuatro VTC_i forma lo que los autores definen como Perfil de Componentes de Tipo de Voz (VTCP, del inglés *voice type component profile*). Para definir el tipo correspondiente a cada segmento, los autores asumieron que el valor de D_i con mayor prevalencia entre los segmentos de la señal se correspondería con el tipo asignado a la señal completa por un especialista. Basados en esto, se definieron umbrales para D_i para cada tipo de señal. Los autores encontraron diferencias estadísticamente significativas entre los valores que toma cada VTC_i para cada tipo de señal. No obstante, a juzgar por los gráficos de caja y bigotes reportados, existe una importante superposición entre dichos valores. En [64] se utilizaron 135 voces provenientes de una base de datos pública [112].

En [65] los autores aplicaron un test de caos [113] sobre señales sintéticas con el fin de evaluar su uso para la tipificación. Las señales utilizadas consistían en una señal periódica (una senoidal de 180 Hz) con diferentes realizaciones de ruido blanco Gaussiano real. Para simular los distintos tipos de señal, los autores incrementaron progresivamente el nivel de ruido, de manera tal que las señales tipo 1 se correspondieran con el menor nivel de ruido mientras que las señales tipo 4 se correspondieran con el máximo nivel de ruido. De esta forma, los autores pretendieron modelar la complejidad creciente de los tipos de señal.

En [66], los autores retoman el enfoque utilizado en [64], consistente en clasificar segmentos de la señal en los distintos tipos, bajo la premisa de que una misma señal puede adoptar tipos diferentes a lo largo del tiempo. Aquí, utilizan el test de caos empleado en [65], pero sobre un conjunto de 135 señales reales provenientes de [112]. Aquí también se plantea la creación de un perfil de componentes de tipo de voz (VTCP), mediante cuatro proporciones definidas en [64], llamadas VTC_i que determinan la proporción del tipo i presente en una señal. Es importante remarcar que si bien las señales son clasificadas por especialistas en los distintos tipos, los segmentos no. La determinación de cada segmento se realiza en base a cuatro umbrales (uno para cada tipo) que se calcularon utilizando las señales sintéticas de [65], por lo que ninguna información proveniente

de los especialistas se utilizó para calcular dichos umbrales. Considerando los distintos VTC_i, se hallaron diferencias estadísticamente significativas entre los distintos tipos de señal, excepto entre los tipos 2 y 3.

En la totalidad de los trabajos revisados, las señales utilizadas fueron clasificadas por especialistas y se excluyeron del estudio aquellas señales para las que la clasificación no fuera unánime entre los observadores. Con excepción de [61], ninguno de los trabajos descritos anteriormente abordó la etapa de toma de decisión con respecto a la clasificación asignada. Esto es relevante porque el hecho de que exista una diferencia significativa en la media de los valores que toma una medida para distintas clases, no implica que esta medida sea útil para distinguir entre los distintos tipos de voz. Esto se debe a la superposición de las distribuciones de una determinada medida para cada clase. En el caso de [61], la clasificación se realizó utilizando un enfoque de reconocimiento de patrones, empleando un árbol de decisión como clasificador. Por otro lado, el número de señales utilizadas en los trabajos descriptos oscila entre 40 y 148, mayoritariamente provenientes de [112] excepto en [61], con proporciones aproximadamente iguales para cada tipo. En ninguno de los trabajos descriptos anteriormente se reporta cómo se realizó el etiquetado de las voces tipo 3 (caóticas) y de las tipo 4 (estocásticas) cuando se utilizaron los cuatro tipos de señales. Es decir, no se explica el detalle clave de cómo se diferenciaron dinámicas estocásticas y caóticas.

Con respecto a los modelos de señales sintéticas utilizadas en [65], debe decirse que el modelo es muy sencillo para representar la complejidad del problema de la clasificación, sin mencionar que no se hace una validación con los verdaderos tipos de señal. Asimismo el uso de ruido blanco Gaussiano para intentar modelar diferentes "niveles de caos" sobre una senoidal puede ser, desde algunas perspectivas, polémico, y en el peor de los casos un error conceptual de gravedad. El ruido utilizado es una realización de un proceso estocástico y, en consecuencia, no es, ni puede considerarse, caótico. Por otro lado el uso de una señal senoidal para modelar la periodicidad de la voz tiene sus inconvenientes. Por ejemplo, los picos de una señal senoidal son menos prominentes que los de una señal de voz (real o sintética), de manera tal que los algoritmos utilizados para calcular medidas de jitter o shimmer basados en la detección de máximos locales pueden sufrir una distorsión importante en presencia de ruido para este tipo de señales [41, 93].

Con respecto al enfoque de perfiles de voz, la cuantificación de la proporción de cada tipo de voz a lo largo de una señal, es necesario tener en cuenta que, si ningún especialista clasificó los segmentos dentro de cada voz, ni se utilizó la información de los especialistas para crear los umbrales que determinan los tipos de cada segmento a clasificar, entonces no es válido afirmar que un segmento pertenece a un tipo dado o a otro. Es decir, el hecho de que el tipo de señal de mayor prevalencia entre los segmentos analizados, según los umbrales utilizados, se corresponda con la clase asignada por el especialista, no implica que esos segmentos de mayor prevalencia sean necesariamente de esa clase. La conclusión debería deducirse en el sentido contrario: dada la clasificación de cada segmento por un especialista, entonces el tipo de los segmentos con mayor prevalencia debería corresponder con el tipo asignado a la señal como un todo.

2.7. Comentarios de final de capítulo

En este Capítulo se presentó la señal de voz y sus características espectrales, así como también el conocido modelo Fuente-Filtro. Adicionalmente, se introdujo la señal de EGG, también proveniente del aparato fonador. Luego se describió la problemática en torno a las perturbaciones de la voz, y los elementos más importantes del análisis de perturbaciones, como las medidas de jitter y la clasificación de voces en tres tipos previo a la aplicación de cualquier medida sobre las señales.

Todos estos conceptos revisados aquí serán de importancia en capítulos posteriores. En el Capítulo 5 se retomará la clasificación en tres tipos con el objetivo de estudiar un algoritmo capaz de tipificar las señales en forma automática. En el Capítulo 6 se pondrá el foco en el jitter relativo

y las dificultades que acarrea su estimación cuando las señales tienen perturbaciones importantes en el periodo instantáneo. Con ese fin, se dará uso a los modelos estocásticos de jitter presentados más arriba para sintetizar voces con niveles de jitter conocidos.

Capítulo 3

Análisis tiempo-frecuencia

3.1. Introducción

El análisis tiempo-frecuencia, y más recientemente tiempo-escala, es una piedra angular del procesamiento de señales. Aunque es un campo de larga tradición, el avance de nuevas ideas que se desprenden del análisis tiempo-frecuencia lo hace objeto, a la vez, de un interés renovado, lo que mantiene este área viva y en constante transformación. Ejemplos de estas conexiones entre el análisis tiempo-frecuencia y tiempo-escala con nuevos y emocionantes campos de estudio son la transformada *scattering* [114–116] o los diccionarios ralos convolucionales [117] y su relación con el aprendizaje profundo.

Un área de particular interés es la mejora de las representaciones tiempo-frecuencia para incrementar su intrepretabilidad, en particular del espectrograma (que se definirá formalmente más adelante en este capítulo). Esta idea en sí misma no es nada nueva. Las distribuciones de Wigner, por ejemplo, están mejor concentradas que el espectrograma pero con la presencia de valores espurios en el plano tiempo-frecuencia causados por términos cruzados en el análisis [105, 118]. En un intento por mejorar la concentración de la energía en el espectrograma y, a la vez, conservar sus propiedades más importantes, principalmente la positividad, se propuso en 1976 la relocalización de los coeficientes del espectrograma sobre el *centro de masa* o centroide de la representación [119]. El resultado obtenido logró su objetivo, aunque recién luego de 20 años fue redescubierto por Flandrin [120] por un lado, y Daubechies [121] por otro, en la década del 90. El primero, formalizando las ideas originales y expandiéndolas a otras clases de representaciones. La segunda, mediante la introducción de *synchrosqueezing* en el marco de la transformada ondita continua y su relación con el procesamiento auditivo. *Synchrosqueezing* consiste en la relocalización de los coeficientes de la transformada ondita continua únicamente en las escalas. Posteriormente la idea de *synchrosqueezing* fue generalizada para la transformada de Fourier de tiempo corto, además de la introducción de estimadores polinómicos locales de alto orden para la estimación de la frecuencia instantánea [75, 122, 123].

En este capítulo se revisarán algunos conceptos del análisis tiempo-frecuencia que serán de importancia posteriormente, principalmente en el Capítulo 6. Allí se utilizará la estimación de la frecuencia instantánea que proveen los operadores de *synchrosqueezing* en una nueva aplicación, consistente en estimar el jitter relativo descrito en el Capitulo 2. Lejos de pretender ser un material exhaustivo, se repasarán primero las definiciones de la transformada de Fourier y su versión de tiempo corto. Luego, se abordará el modelo de señal multicomponente y la estimación de las crestas asociadas a cada modo. Seguido de esto se presentará el método de reasignación y sus operadores de primer orden. Finalmente se describirá la transformación de *synchrosqueezing* y los operadores de alto orden. El/la lector/a interesado/a en detalles sobre estas técnicas y su implementación es referido/a la bibliografía especializada, principalmente [75, 105, 118].

3.2. Transformada de Fourier de tiempo corto

La transformada de Fourier (TF) de una señal $x(t)$ esta definida mediante la siguiente transformación integral:

$$\mathcal{F}\{x(t)\} = \hat{x}(f) := \int_{-\infty}^{+\infty} x(t)e^{-i2\pi ft}\, dt, \tag{3.1}$$

donde $\mathcal{F}$ es el operador transformada de Fourier, i es la unidad imaginaria, y f la frecuencia. Comúnmente, la gráfica de $|\hat{x}(t)|$ vs. f, llamada espectro de amplitud o magnitud de $x(t)$, permite observar el contenido frecuencial de la señal, aunque resulta imposible identificar cómo cambian en el tiempo dichas componentes frecuenciales. El espectro de magnitud de $x(t)$, entonces, muestra una especie de contenido frecuencial "promedio" [105]. Esta información resulta útil bajo la hipótesis de *estacionariedad*, es decir, considerando que los parámetros de la señal, como sus componentes frecuenciales, no cambian en el tiempo o que dichos cambios son despreciables [118].

Una extensión intuitiva de esta idea se basa en que, si se desea conocer el contenido frecuencial para distintos tiempos, se debería aplicar la TF de $x(t)$ en una vecindad de cada valor de t mediante una *ventana*, para luego *trasladar* esa ventana en el tiempo y calcular nuevamente la TF para cada traslación. De esta forma sería posible obtener una versión "local" en el tiempo de la TF, para cada ventana trasladada. Esta idea se ve limitada por el principio de incertidumbre, que determina que las TF sobre las ventanas trasladadas tendrán una menor resolución frecuencial que la TF sobre la señal completa, por contar con una duración menor [105]. Es posible comprender esta situación al considerar a la Ecuación (3.1) como el producto interno entre $x(t)$ y $e^{i2\pi ft}$. Dado que

$$e^{i2\pi ft} = \cos(2\pi ft) + i\,\mathrm{sen}(2\pi ft) \tag{3.2}$$

se compone de oscilaciones (en cuadratura) que existen para todo valor de t, su *concentración* en el tiempo es mínima. Sin embargo, esta exponencial compleja posee la máxima concentración en la frecuencia:

$$\mathcal{F}\left\{e^{-i2\pi qt}\right\} = \delta(f-q) \tag{3.3}$$

donde $\delta(f)$ es la función impulso o delta de Dirac. Por lo tanto el producto interno en la Ecuación (3.1) "detecta" oscilaciones con frecuencia f a lo largo de todo el dominio t.

En contraste con esta situación, es posible formalizar la intuición descripta más arriba respecto de una "ventana móvil" como la Transformada de Fourier de tiempo corto (TFTC) de $x(t)$ utilizando una ventana $g(t)$:

$$V_x^g(t,f) := \int_{-\infty}^{+\infty} x(\tau)g(\tau-t)e^{-i2\pi f\tau}d\tau \tag{3.4}$$

donde $g(t)$ es una función real, par, y que tiende a cero rápidamente hacia ambos lados del origen. En general, $g(t)$ será considerada una ventana Gaussiana:

$$g(t) = \frac{1}{\sigma}e^{-\frac{\pi}{\sigma^2}t^2} \tag{3.5}$$

donde σ es un parámetro que permite definir la duración *efectiva* de la ventana como $6\frac{\sigma}{\sqrt{2\pi}}$, fuera del cual podemos afirmar que $g(t) \approx 0$. El módulo al cuadrado de $V_x^g(t,f)$ es el ya conocido *espectrograma* de $x(t)$:

$$S_x^g(t,f) := \left|\int_{-\infty}^{+\infty} x(\tau)g(\tau-t)e^{-i2\pi f\tau}d\tau\right|^2, \tag{3.6}$$

que permite visualizar la evolución temporal de las componentes frecuenciales de la señal.

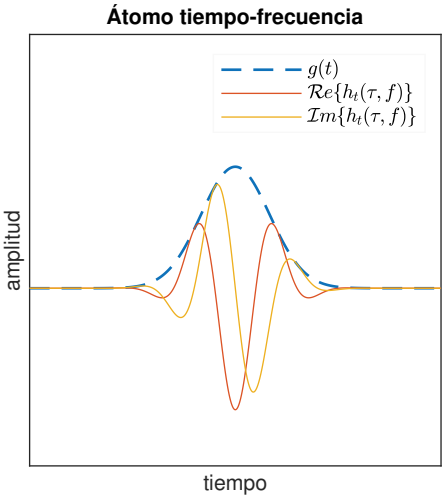


Figura 3.1: Representación de un átomo tiempo frecuencia. Se observa que las oscilaciones de la parte real e imaginaria, en cuadratura, se extinguen por efecto de la ventana $g(t)$.

A lo largo de este capítulo se utilizará, sin pérdida de generalidad, una versión *modificada* de la TFTC [75]:

$$V_x^g(t, f) := \int_{-\infty}^{+\infty} x(\tau) g(\tau - t) e^{-i2\pi f(\tau - t)} d\tau \tag{3.7}$$

en la que la única diferencia consiste en un cambio en la fase de $V_x^g(t, f)$, sin afectar el espectrograma resultante. Esta versión modificada tiene algunas ventajas, tanto para la demostración de algunas propiedades como para la interpretación. Asimismo, posibilita la interpretación de la Ecuación (3.7) como el producto interno entre $x(t)$ y un *átomo* tiempo-frecuencia desplazado t unidades:

$$h_t(\tau, f) = g(\tau - t) e^{i2\pi f(\tau - t)}. \tag{3.8}$$

Esto permite formalizar la noción de localidad que está detras de la TFTC, ya que que se trata de un producto interno entre la señal y oscilaciones concentradas en el tiempo por la ventana $g(t)$, fuera de la cual se extinguen, como se observa en la Figura 3.1 donde se muestra la parte real e imaginaria del átomo y la ventana $g(t)$ superpuesta.

La TFTC es además una transformación invertible. Una fórmula de reconstrucción a partir de la TFTC está dada por [122]:

$$x(t) = \frac{1}{g(0)} \int_{-\infty}^{+\infty} V_x^g(t, \omega) d\omega, \tag{3.9}$$

siempre que $g(0) \neq 0$.

3.3. Frecuencia instantánea

Considerando el caso de un tono puro, también llamado onda monocromática, dado por:

$$x(t) = a \cos(2\pi f_0 t), \tag{3.10}$$

resulta natural definir a como la amplitud de esta señal y a f_0 como su frecuencia, dado que f_0 es proporcional a la cantidad de ciclos de la señal que se suceden en una unidad de tiempo (donde 2π es la constante de proporcionalidad). No obstante, si se quisiera extender esta intuición al caso dado por funciones de amplitud y fase variables en el tiempo de la forma:

$$x(t) = a(t) \cos(2\pi \phi(t)), \tag{3.11}$$

donde $a(t)$ y $\phi(t)$ son las funciones de amplitud y fase respectivamente, se encontraría que no existe un único par de dichas funciones que satisfagan esta expresión [105]. En consecuencia no sería posible definir la amplitud y la frecuencia de forma univoca a partir de esta definición.

Una solución a este problema es considerar el caso de una *señal analítica*[1] dada por:

$$x_A(t) = x(t) + i\mathcal{H}\{x(t)\}, \tag{3.12}$$

donde $\mathcal{H}\{\cdot\}$ denota la transformada de Hilbert [86, 105]. Retomando el caso de una onda monocromática, es posible ver que la señal analítica puede expresarse como:

$$x_A(t) = a\cos(2\pi f_0 t) + ia\,\mathrm{sen}(2\pi f_0 t) = ae^{i2\pi f_0 t}, \tag{3.13}$$

y la frecuencia en este caso puede encontrarse como:

$$f_0 = \frac{1}{2\pi}\frac{d}{dt}\arg\{x_A(t)\}. \tag{3.14}$$

Considerando ahora el caso más general, en el que la amplitud y la frecuencia son dependientes del tiempo, la señal analítica puede expresarse, teniendo en cuenta ciertas consideraciones sobre $a(t)$ [105, 124], como:

$$x_A(t) = a(t)\cos(2\pi\phi(t)) + ia(t)\,\mathrm{sen}(2\pi\phi(t)) = a(t)e^{i2\pi\phi(t)}. \tag{3.15}$$

Luego, la amplitud y la frecuencia *instantáneas* pueden derivarse de esta expresión de forma similar al caso monocromático como:

$$a(t) = |x_A(t)| \tag{3.16}$$

y

$$f_0(t) = \frac{1}{2\pi}\frac{d}{dt}\arg\{x_A(t)\}. \tag{3.17}$$

En consecuencia, la frecuencia instantánea (FI) es considerada como la derivada primera de la función de fase $\phi(t)$ respecto a t. En base a esto pueden confeccionarse señales con frecuencia instantánea conocida mediante una elección conveniente de la función de fase $\phi(t)$. Ejemplos de estas señales son las denominadas *chirps*, siendo el *chirp* lineal el ejemplo más conocido, expresado como:

$$x(t) = \cos(2\pi\phi(t)) = \cos\left(2\pi(\alpha t^2 + \beta t)\right),\ \alpha, \beta \in \mathbb{R}. \tag{3.18}$$

donde se aprecia que su fase es un polinomio de orden 2. Luego, su FI estará dada por: $\phi'(t) = 2\alpha t + \beta$. Un ejemplo del módulo de la TFTC de esta señal y otros *chirps* puede observarse en la Figura 3.2.

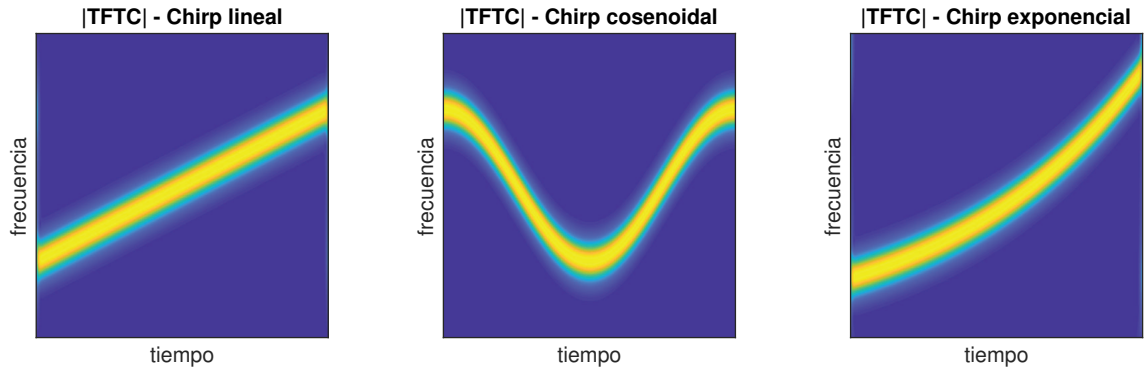
3.4. Señales multicomponente y crestas

El modelo de señal multicomponente considera que una señal $x(t)$ se encuentra definida como:

$$x(t) = \sum_{k=1}^{K} m_k(t) = \sum_{k=1}^{K} a_k(t)e^{i2\pi\phi_k(t)} \tag{3.19}$$

donde $m_k(t)$ es una función oscilante del tipo AM-FM [105], que llamaremos simplemente *modo*, y $a_k(t)$ y $\phi_k(t)$ son la amplitud y la fase instantáneas, respectivamente, del modo $m_k(t)$, que satisfacen lo siguiente:

[1] Para ver la relación que existe entre este nombre y las funciones analíticas de una variable compleja, el/la lector/a puede referirse a [86]



(a) (b) (c)

Figura 3.2: Módulo de la TFTC de tres *chirps* sintéticos con frecuencia instantánea conocida.

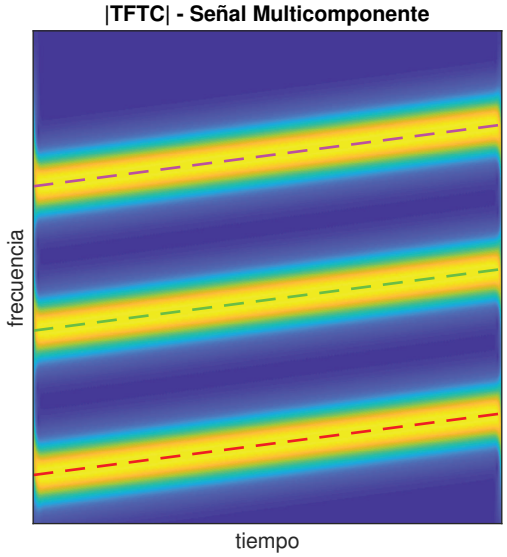


Figura 3.3: Señal multicomponente. Cada modo de la señal es un *chirp* lineal con igual pendiente. En línea de trazos de colores se ilustra la frecuencia instantánea de cada modo.

i) $a_k(t) > 0$.

ii) $\phi'_K(t) > ... > \phi'_k(t) > ... > \phi'_1(t) > 0$.

La primera de estas condiciones permite obtener la amplitud de un modo simplemente como $|m_k(t)|$, siguiendo el modelo de función analítica descrito anteriormente. La segunda condición establece que las componentes de la señal estarán "apiladas" en el plano tiempo-frecuencia y que las frecuencias instantáneas de cada componente no se intersecan. Esto es fundamental para que la FI de cada modo sea única para cada instante de tiempo. Basándose en esto, es posible afirmar que la frecuencia instantánea del modo $m_k(t)$ es $\phi'_k(t)$. Como ejemplo, la Figura 3.3 muestra una señal multicomponente sintética cuyos modos son tres *chirps* lineales de idéntica pendiente. En colores se marca la frecuencia instantánea de cada uno.

La TFTC de un modo $m(t)$ utilizando una ventana $g(t)$ puede ser aproximada por [75, 123]:

$$V_m^g(t, f) \approx m(t)\hat{g}(f - \phi'(t)) \tag{3.20}$$

siempre que $m(t)$ no tenga fuertes modulaciones (tanto $a'(t)$ como $\phi'(t)$ deben ser pequeñas). La Ecuación (3.20) determina que la energía del modo $m(t)$ se concentrará en torno a la FI y tiende

a dispersarse a medida que se aleja de la FI. Esto ilustra el hecho de que la TFTC de una función AM-FM existe en una franja del plano tiempo-frecuencia que está centrada en $\phi'(t)$, donde el módulo de la TFTC posee un máximo local denominado *cresta*, y cuyo ancho depende del soporte de $\hat{g}(f)$ [125, 126].

Un método sencillo para estimar la FI de un modo consiste, entonces, en detectar su cresta $r(t)$ correspondiente, que es exactamente la FI para el caso sin ruido [125, 127, 128]. No obstante, para casos más generales, considerar que:

$$r(t) \approx \phi'(t) \tag{3.21}$$

tampoco resulta una mala aproximación. Por esta razón, el estudio de las crestas y su detección es un área de exploración actual e íntimamente relacionada con la extracción de modos de una señal. Idealmente, un algoritmo extractor de crestas debería encontrar la solución $r(t)$ del siguiente problema de optimización no convexo [127, 129]:

$$\max_{r\in\Gamma} \int_{-\infty}^{+\infty} (|V_x^g(t, r(t))|^2 - \alpha[r'(t)]^2 - \beta[r''(t)]^2)dt \tag{3.22}$$

donde $r(t)$ pertenece al espacio Γ de funciones cuadrado integrables y diferenciables al menos dos veces, y $\alpha, \beta \in \mathbb{R}$. Este problema consiste en maximizar la energía de la cresta (primer término del integrando) a la vez que favorece la continuidad y la suavidad mediante las penalizaciones del segundo y tercer término, respectivamente. Si bien para resolverlo existen diferentes enfoques, se utilizará en este trabajo, especialmente en el Capítulo 6, una estrategia *voraz* para encontrar una solución no necesariamente óptima para el problema planteado en la Ecuación (3.22).

Con el fin de describir el algoritmo utilizado, consideremos una versión discreta de la TFTC, $V_x^g[n, k]$, y de la cresta, $r[n]$, donde $n = 0, 1, ...N - 1$ y $k = 0, 1, ..., K - 1$ son los índices temporal y frecuencial, respectivamente. El enfoque para hallar la cresta se basa en maximizar la energía sobre ella, $|V_x^g\Big[r[n], k\Big]|^2$, tomando un valor de n al azar y eligiendo los valores de k que maximicen $|V_x^g[r[n], k]|^2$ hacia la derecha primero y luego hacia la izquierda. Si bien los parámetros α y β de la Ecuación (3.22) se hacen 0 en este caso, se limita el "salto" que puede existir entre los valores de la cresta al restringir la búsqueda del siguiente valor $r[n + 1]$ en un intervalo acotado en torno al valor anterior, $r[n]$. Todo el procedimiento es repetido un número P de veces a determinar y luego se selecciona aquella cresta con mayor energía entre las halladas para cada repetición. El Algoritmo 1 detalla el método recién descrito para la detección de crestas.

Una vez detectada la cresta de un modo, es posible su *extracción* a partir de una "franja" sobre la TFTC entorno a $r(t)$. Una fórmula sencilla para la extracción del modo $m_k(t)$ una vez determinada su cresta $r_k(t)$ es la siguiente:

$$m_k(t) = \frac{1}{g(0)} \int_{\{f:|f-r_k(t)|<R\}} V_x^g(t, f)\, df, \tag{3.23}$$

donde $2R$ es el ancho de la franja centrada en $r(t)$. Debe observarse que para utilizar esta fórmula los modos deben estar separados entre sí una distancia mayor a R en el plano tiempo-frecuencia. De otra forma, la franja utilizada capturaría la influencia de otras componentes frecuenciales de la señal.

Algoritmo 1: Detección de crestas.

Entrada: La TFTC de una señal $x[n]$, $V_x^g[n,k]$, la cantidad de repeticiones de la búsqueda P y el máximo salto permitido J entre un valor de cresta y su valor anterior o posterior.

Salida: : $r[n]$ la cresta con mayor energía entre las P repeticiones.

para $p \in [1, P]$ **hacer**
- Elegir un valor $n_0 \in [0, N-1]$; `// Elegir un indice temporal al azar`
- $k_0 = \arg\max_k |V[n_0, k]|^2$; `// Conservar` k `del mayor coeficiente`
- $c_p[n_0] = k_0$; `//` $k0$ `es el primer elemento de la cresta`
- $E_p[n_0] = |V[n_0, k_0]|^2$; `// Para calcular la energía sobre la cresta`
- $I = [k_0 - J, k_0 + J]$; `//` `Intervalo de búsqueda,` J `limita el salto`
- $n = n_0$;
- **mientras** *n <N-1* **hacer**
  - $n = n + 1$; `// Primero buscar hacia la derecha`
  - $c_p[n] = \arg\max_{k \in I} |V[n, k]|^2$;
  - $E_p[n] = |V[n, c_p[n]]|^2$;
  - $I = [c_p[n] - J, c_p[n] + J]$;
- $n = n_0$;
- **mientras** *n >1* **hacer**
  - $n = n - 1$; `// Luego hacia la izquierda`
  - $c_p[n] = \arg\max_{k \in I} |V[n, k]|^2$;
  - $E_p[n] = |V[n, c_p[n]]|^2$;
  - $I = [c_p[n] - J, c_p[n] + J]$;

$p_{max} = \arg\max_p \sum_{q=0}^{N-1} E_p[q]$; `// Seleccionar cresta de mayor energía`

$r = c_{p_{max}}$;

3.5. *Synchrosqueezing*

Al observar el módulo de los coeficientes de la TFTC, como en la Figura 3.4, se aprecia que la energía de un modo $m(t)$ se concentra en la vecindad de la FI (línea roja en la Figura 3.4). Idealmente, una representación tiempo-frecuencia perfecta debería concentrar la energía sólo en $\phi'(t)$, es decir, sobre la línea roja. Con el objetivo de mejorar la representación, considerando por ejemplo las bondades de otras representaciones como la distribución de Wigner [105], se propuso una técnica de posprocesamiento conocida como *método de reasignación* (MR) [119, 120]. El método de reasignación se basa en el cálculo de dos operadores denominados *operadores de reasignación complejos*, definidos como:

$$\tilde{\omega}_x(t,f) := \frac{\partial_t V_x^g(t,f)}{i2\pi V_x^g(t,f)} := f - \frac{V_x^{g'}(t,f)}{i2\pi V_x^g(t,f)} \tag{3.24}$$

$$\tilde{\tau}_x(t,f) := t + \frac{V_x^{tg}(t,f)}{V_x^g(t,f)}, \tag{3.25}$$

donde $V_x^{tg}(t,f)$ y $V_x^{g'}(t,f)$ son las TFTC calculadas con las ventanas $tg(t)$ y $g'(t)$ respectivamente. Luego, el MR relocaliza los coeficientes del espectrograma de $x(t)$ de acuerdo a la aplicación:

$$(t,f) \to (\mathcal{R}e\left\{\tilde{\tau}_x(t,f)\right\}, \mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\}). \tag{3.26}$$

Esto resulta en una mejora de la representación en términos de interpretación, logrando una mayor concentración de la energía en las crestas, debido a que $\mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\}$ es una aproximación de la FI y $\mathcal{R}e\left\{\tilde{\tau}_x(t,f)\right\}$ una aproximación del retardo de grupo [105, 118]. Esto puede observarse a partir de un ejemplo adecuado. Si se considera el caso de un tono puro $x(t) = e^{i2\pi\phi_0 t}$, donde $\phi_0 \in \mathbb{R}$ es la FI (constante en este caso), cuya TFTC es:

$$V_x^g(t,f) = e^{i2\pi\phi_0 t}\hat{g}(f-\phi_0), \tag{3.27}$$

entonces el operador $\mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\}$ da como resultado:

$$\begin{aligned}\mathcal{R}e\left\{\frac{\partial_t V_x^g(t,f)}{i2\pi V_x^g(t,f)}\right\} &= \mathcal{R}e\left\{\frac{e^{i2\pi\phi_0 t}\hat{g}(f-\phi_0)i2\pi\phi_0}{2\pi e^{i2\pi\phi_0 t}\hat{g}(f-\phi_0)}\right\} \\ &= \mathcal{R}e\left\{\frac{V_x^g(t,f)i2\pi\phi_0}{i2\pi V_x^g(t,f)}\right\} \\ &= \phi_0.\end{aligned} \tag{3.28}$$

Este resultado indica que el operador $\mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\}$ será una aproximación perfecta de la FI para un tono puro, es decir, con una función de fase $\phi(t)$ lineal. Para funciones con fase no lineal, como un *chirp* lineal, con fase cuadrática, esta aproximación empieza a fallar progresivamente conforme aumenta el orden de la fase. Aún así, el método de reasignación es capaz de relocalizar en forma perfecta los coeficientes de un *chirp* lineal al utilizar la información adicional que proporciona el operador complejo de reasignación temporal $\tilde{\tau}_x(t,f)$. No obstante, este método posee una desventaja. Dado que rompe con la causalidad de la representación (se reasignan coeficientes en tiempos diferentes a los que se encontraban originalmente, por efecto del operador de reasignación temporal), el resultado es no invertible. Esto vuelve imposible el filtrado u otro tipo de procesamientos en el plano tiempo-frecuencia luego de aplicar el MR.

Un caso particular de reasignación que lidia con este problema es la transformación *synchrosqueezing* (SST, del inglés *synchrosqueezing transform*), originalmente presentada para la transformada ondita continua [121, 125, 126, 130] y luego para la TFTC (FSST, del inglés *Fourier-based*

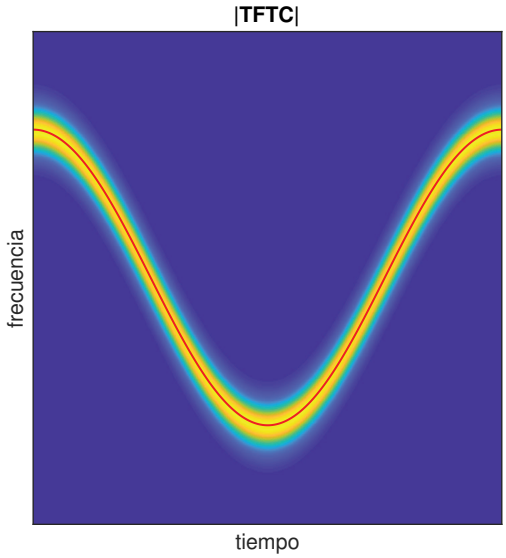


Figura 3.4: Módulo de la transformada de Fourier de tiempo corto de un chirp cosenoidal. La linea roja demarca la frecuencia instantánea, en torno a la cual se distribuyen los coeficientes con mayor energía.

synchrosqueezing transform) [122, 126]. FSST es un método basado en la fase que permite la concentración de la representación tiempo-frecuencia reubicando coeficientes de la TFTC únicamente en forma *vertical*, es decir en la frecuencia, según la aplicación:

$$(t, f) \to (t, \mathcal{R}e\left\{\tilde{\omega}_x(t, f)\right\}). \tag{3.29}$$

Formalmente la FSST de una señal $x(t)$ está definida como [123]:

$$U_x(t, \omega) = \frac{1}{g(0)} \int_{\{f:|V_x^g(t,f)|>\gamma\}} V_x^g(t, f)\delta(\omega - \mathcal{R}e\left\{\tilde{\omega}_x(t, f)\right\})df \tag{3.30}$$

donde $g(0) \neq 0$ y γ es un umbral para evitar la reasignación de coeficientes cercanos a 0.

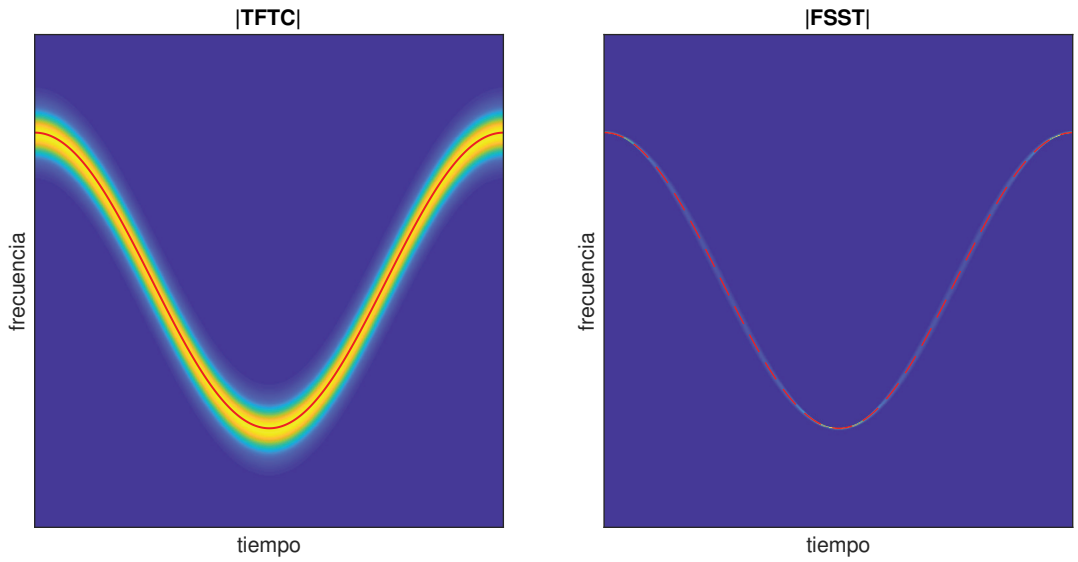
En contraste con el MR, la FSST permite invertir la representación, permitiendo la reconstrucción de modos y el filtrado utilizando una fórmula de reconstrucción como la siguiente:

$$x(t) = \int_{-\infty}^{+\infty} U_x(t, \omega)d\omega. \tag{3.31}$$

La Figura 3.5b muestra el resultado de aplicar la FSST a un *chirp* cosenoidal, mientras que el módulo de la TFTC se muestra en la Figura 3.5a. Como puede observarse, los coeficientes se han relocalizado mucho más cerca de la FI, llevando la representación a una forma más cercana a la ideal. Para lograr este resultado, FSST utiliza tres TFTC con diferentes ventanas: $g(t)$, $g'(t)$ y $tg(t)$.

Considerando la Ecuación (3.28), se advierte que, a diferencia del método de reasignación que utiliza ambos operadores, FSST sólo podrá reasignar en forma perfecta los coeficientes de la TFTC de un tono puro, ya que $\mathcal{R}e\left\{\tilde{\omega}_x(t, f)\right\}$ es una buena aproximación de la FI siendo $\mathcal{R}e\left\{\tilde{\omega}_x(t, f)\right\} = \phi'(t)$ para ese caso.

Para señales con fase cuadrática o de órdenes mayores, FSST posee menos capacidad de concentrar los coeficientes que el método de reasignación. En particular, dado que el operador $\tilde{\omega}_x(t, f)$ sólo permite una representación perfecta para una función idealmente de fase lineal, la reasignación utilizando este operador se denominará FSST *de primer orden*. La reasignación continúa siendo lo suficientemente buena para el caso de tonos ligeramente perturbados, pero comienza a fallar para modos con modulaciones más importantes que no pueden ser despreciadas.



(a) (b)

Figura 3.5: Módulos de la TFTC y de la FSST para un *chirp* cosenoidal. a) Módulo de la TFCT. b) Módulo de la FSST. Se observa una mejora en la concentración de los coeficientes en torno a la FI.

Con el objetivo de superar esta dificultad, resulta de interés observar qué ocurre para funciones con fase de orden superior. Considérese, por ejemplo, un *chirp* lineal. Para este caso, es posible demostrar que la FI está dada por la expresión [123]:

$$\phi'(t) = \mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\} + \phi''(t)(\mathcal{R}e\left\{\tilde{\tau}_x(t,f) - t\right\}) \tag{3.32}$$

de donde es puede verse que al término $\mathcal{R}e\left\{\tilde{\omega}_x(t,f)\right\}$, que según la Ecuación (3.28) es una aproximación perfecta de la FI para un tono puro, se le suma un segundo término que constituye un sesgo sobre la aproximación de primer orden.

3.5.1. *Synchrosqueezing* de segundo orden

Si se deseara *mejorar* la aproximación a la FI para el *chirp* lineal, se debería calcular el sesgo descrito en la Ecuación (3.32), que depende del operador temporal $\tilde{\tau}(t,f)$, que ya se conoce, y de $\phi''(t)$, hasta ahora desconocida. Si $\phi(t)$ es aproximable localmente con un polinomio de orden 2, puede demostrarse que una forma de estimar $\phi''(t,f)$ está dada por [75, 123]:

$$\begin{aligned}\phi''(t) &= \mathcal{R}e\left\{\frac{\partial_f\tilde{\omega}(t,f)}{\partial_f\tilde{\tau}(t,f)}\right\} \\ &= \mathcal{R}e\left\{\tilde{q}_x(t,f)\right\},\end{aligned} \tag{3.33}$$

donde ∂_f denota la derivada parcial con respecto a la frecuencia y $\tilde{q}_x(t,f)$ es el operador complejo de modulación de la señal $x(t)$, dado por:

$$\tilde{q}_x(t,f) = \frac{1}{2\pi i}\frac{\left(V_f^g(t,f)\right)^2 + V_f^g(t,f)V_f^{tg'}(t,f) - V_f^{g'}(t,f)V_f^{tg}(t,f)}{V_f^g(t,f)V_f^{t^2g}(t,f) - \left(V_f^{tg}(t,f)\right)^2}. \tag{3.34}$$

Luego, basándose en la Ecuación (3.32), puede definirse al operador de frecuencia instantánea complejo de segundo orden como:

$$\tilde{\omega}_x^{[2]}(t,f) = \begin{cases} \tilde{\omega}_x(t,f) + \tilde{q}_x(t,f)(t - \tilde{\tau}_x(t,f)) & \text{si } \partial_f\tilde{\tau}_x(t,f) \neq 0 \\ \tilde{\omega}_x(t,f) & \text{en otro caso} \end{cases} \tag{3.35}$$

donde el supraíndice entre corchetes indica el orden del operador. De la Ecuación (3.35) puede verse que el operador de segundo orden es igual al de primer orden si $\partial_f \tilde{\tau}_x(t,f)$ (que es el denominador de $\tilde{q}(t,f)$) es nulo, mientras que si $\partial_f \tilde{\tau}_x(t,f) \neq 0$ el operador de segundo orden es igual al de primer orden más un término de corrección que depende de la derivada segunda de la fase (recordar que $\mathcal{R}e\left\{\tilde{q}(t,f)\right\} = \phi''(t)$). En cierta forma, esto es semejante a una serie de Taylor, que utiliza derivadas de orden superior para mejorar la aproximación de una función en torno a un punto. Este concepto permite anticipar que, para construir operadores de mayor orden de la FI, deberán obtenerse estimadores de sus derivadas de órdenes cada vez más altos.

La transformación de *synchrosqueezing* de segundo orden (FSST2) será entonces idéntica a la Ecuación (3.30), empleando $\tilde{\omega}_x^{[2]}(t,f)$ en lugar de $\tilde{\omega}_x(t,f)$. Nótese que la estimación de $\tilde{q}(t,f)$ requiere el cómputo de las TFTC de $x(t)$ con las ventanas $tg'(t)$ y $t^2 g(t)$, además de las que se requieren para calcular los operadores $\tilde{\omega}_x(t,f)$ y $\tilde{\tau}_x(t,f)$. En total, se requieren calcular cinco TFTC con distintas ventanas para la aplicación de FSST2, frente a las dos que se requieren para FSST, aumentando el costo computacional.

3.5.2. *Synchrosqueezing* de orden superior

De manera similar al caso de segundo orden, si la función de fase $\phi(t)$ y el logaritmo de la amplitud de una señal puedan ser aproximados localmente con un polinomio de orden N [75], entonces es posible probar que la estimación local compleja de la FI de orden N está dada por $\tilde{\omega}_x^{[N]}(t,f)$:

$$\tilde{\omega}_x^{[N]}(t,f) = \begin{cases} \tilde{\omega}_x(t,f) + \sum_{k=2}^{N} \tilde{q}_x^{[k,N]}(t,f)(-\chi_{k,1}(t,f)) & \text{si } V_x^g(t,f) \neq 0 \text{ y} \\ & \partial_f \chi_{j,j-1}(t,f) \neq 0 \\ \tilde{\omega}_x(t,f) & \text{en otro caso.} \end{cases} \tag{3.36}$$

donde:

- $\tilde{q}_x^{[k,N]}(t,f)$ se denomina operador complejo de modulación de orden N. Estos operadores pueden utilizarse para encontrar una aproximación local de orden $N-k$ de la derivada k-ésima de $\phi(t)$ como:

$$\mathcal{R}e\left\{\tilde{q}_x^{[k,N]}(t,f)\right\} = \frac{\phi^{(k)}(t)}{(k-1)!}, \tag{3.37}$$

  donde el supraíndice $[k,N]$ indica con k el orden de la derivada aproximada, y con N el orden de la aproximación local de la fase.

- $\chi_{j,\ell}(t,f)$ y $\nu_j(t,f)$ son funciones auxiliares que permiten expresar $\tilde{q}_x^{[k,N]}(t,f)$ en términos de transformadas de Fourier de tiempo corto con diferentes ventanas. Primero $\chi_{k,1}(t,f)$ se define como:

$$\chi_{k,1}(t,f) = \frac{V_x^{t^{k-1}g}(t,f)}{V_x^g(t,f)}, \tag{3.38}$$

  luego las funciones $\chi_{k,j}(t,f)$ y $\nu_j(t,f)$ se definen en forma recursiva como:

$$\nu_j(t,f) = \frac{\partial_f \nu_{j-1}(t,f)}{\partial_f \chi_{j,j-1}(t,f)},$$

$$\chi_{k,j}(t,f) = \frac{\partial_f \chi_{k,j-1}(t,f)}{\partial_f \chi_{j,j-1}(t,f)},$$

  para $j = 2, ..., N$ y $k = j, ..., N$, siempre que $V_x^g(t,f) \neq 0$ y $\partial_f \chi_{j,j-1}(t,f) \neq 0$, siendo $\nu_1(t,f) = \tilde{\omega}(t,f)$.

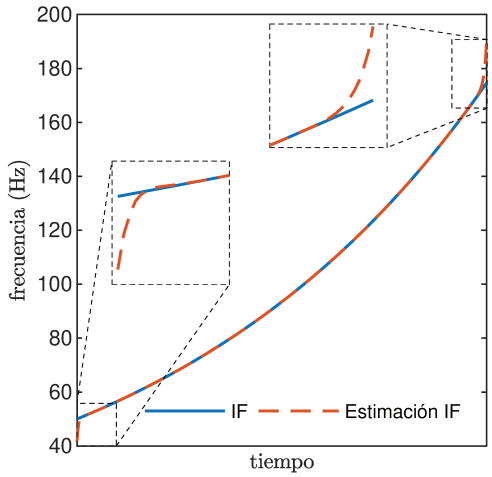


Figura 3.6: Estimación de la FI mediante la Ecuación (3.41). Puede apreciarse la diferencia en la estimación en el principio y fin del segmento analizado debido a los efectos de borde en la estimación de las TFTC.

- A partir de $\chi_{j,j}(t,f)$ y $\nu_j(t,f)$ los operadores complejos de modulación $\tilde{q}_x^{[k,N]}(t,f)$ se definen como [75]:

$$\tilde{q}_x^{[N,N]}(t,f) = \nu_N(t,f) \tag{3.39}$$

y

$$\tilde{q}_x^{[j,N]}(t,f) = \nu_j(t,f) - \sum_{k=j+1}^{N} \chi_{k,j}(t,f)\tilde{q}_x^{[k,N]}(t,f) \tag{3.40}$$

para for $j = N-1, N-2, ..., 2$.

De igual manera que en el caso de FSST2, la transformación de *synchrosqueezing* de orden superior (FSSTN) estará dada por la Ecuación (3.30), empleando $\tilde{\omega}_x^{[N]}(t,f)$ en lugar de $\tilde{\omega}_x(t,f)$.

3.6. Estimación de la frecuencia instantánea y su derivada

Tal como se enunció en la sección anterior, los operadores de *synchrosqueezing* pueden proveer localmente aproximaciones polinómicas de la frecuencia fundamental y sus derivadas. En el caso de una señal multicomponente, es posible obtener estimaciones de la FI (o sus derivadas) de cada modo presente en la señal. Para obtenerlas, es necesario en primer lugar determinar sus crestas. Luego, la obtención de las FI y su derivadas requiere la evaluación de los operadores en las crestas de cada modo. Suponiendo que $r_k(t)$ es la cresta del modo $m_k(t)$, las aproximaciones de orden superior de $\phi_k'(t)$ y su derivada $\phi_k''(t)$ pueden hallarse como:

$$\phi_k'(t) = \mathcal{R}e\left\{\tilde{\omega}_x^{[N]}(t, r_k(t))\right\} \tag{3.41}$$

y

$$\phi_k''(t) = \mathcal{R}e\left\{\tilde{q}_x^{[2,N]}(t, r_k(t))\right\}. \tag{3.42}$$

La Figura 3.6 muestra la FI de un chirp exponencial junto con su estimación de segundo orden, es decir mediante la Ecuación (3.41) para $N = 2$. Como puede observarse en la Figura, la estimación sufre una degradación en los extremos del intervalo de análisis. Esta se debe a los efectos de borde que ocurren al calcular las TFTC involucradas en el cómputo del operador, como consecuencia de que la ventana de análisis $g(t)$ descrita anteriormente no cabe completamente en la

señal hasta que haya transcurrido un número de muestras igual a su duración efectiva. Para aliviar estos artefactos, es posible descartar un número de muestras tanto del principio como del final de la señal analizada. Otra opción es extender la duración de la señal tanto al comienzo como al final, concatenando segmentos cuya duración supere al menos un ancho efectivo de la ventana de análisis [105, 118]. Dichos segmentos pueden contener ceros únicamente, o una versión reflejada de las primeras muestras de la señal y las últimas, concatenadas al principio y al final de la señal respectivamente. Otras formas más elaboradas consisten en reflejar las primeras (o últimas) muestras de la señal junto a alguna estrategia que permita asegurar la continuidad en la unión con los segmentos auxiliares concatenados [131].

3.7. Comentarios de final de capítulo

En este capítulo se han repasado las herramientas provenientes del análisis tiempo-frecuencia que se utilizarán más adelante en este documento. El concepto de frecuencia instantánea será de vital importancia en el Capítulo 6 donde se presentará una aplicación de los métodos descritos a la estimación del jitter relativo. El modelo de señal multicomponente permite representar una amplia clase de señales, entre ellas la señal de voz. Las componentes de esta señal, cuando es aproximadamente periódica, tienen como FI múltiplos enteros de la FI del primer modo, es decir, conforman una serie de armónicos en el plano-tiempo frecuencia. Usualmente, la cresta del primer modo posee mayor energía que aquellas de los modos superiores, y en consecuencia es el modo *dominante* [132]. En ocasiones esto puede no ser así, especialmente debido al efecto de modulación que produce la respuesta en frecuencia del tracto vocal sobre el espectro de la señal de voz. No obstante, en general, la frecuencia fundamental *instantánea* se corresponde con la FI del primer modo. Finalmente, es importante notar que los operadores de *synchrosqueezing* proveen aproximaciones polinómicas locales de la FI y su derivada para cada modo. Para obtener estas aproximaciones, cuando la señal es multicomponente, es necesario evaluar los operadores en la cresta del modo cuya FI (o alguna de sus derivadas) se desea estimar.

Capítulo 4

Herramientas de aprendizaje maquinal

4.1. Introducción

La inteligencia computacional es un área de investigación que estudia el diseño de agentes inteligentes. "Agente" es un término general que busca designar a *algo* que interactúa con su ambiente, ya sea un ser humano o una sociedad, o desde un termostato hasta un avión [133]. Un agente *inteligente*, por otro lado, es aquél que actúa con inteligencia, esto es: lo que hace es apropiado dado un contexto y un objetivo, su comportamiento es flexible a los cambios en el ambiente y en el objetivo, y es capaz de aprender de la experiencia [133]. El término inteligencia computacional es preferible en este contexto a inteligencia *artificial* porque hace explícita la hipótesis de que tal inteligencia, definida como se desee, se modelará *computacionalmente* [133].

La inteligencia computacional engloba otros campos de estudios que son importantes en su propia ley, particularmente el campo del aprendizaje maquinal (*machine learning* en inglés). Este último tiene su origen en la ciencia de la computación, e involucra tanto investigación básica como aplicada. Su objetivo es abordar la cuestión de cómo generar programas que le permitan a una computadora, en general, mejorar con la experiencia y a partir de un conjunto de datos [134]. Para ello, el aprendizaje maquinal se vale de resultados de otras áreas tales como la estadística, la computación, la matemática aplicada, la biología o la filosofía [134].

Paralelamente al aprendizaje maquinal, el campo del reconocimiento de patrones consiste en tomar datos crudos y procesarlos con el objetivo de tomar decisiones basadas en la categoría o tipo de patrón en los datos. Se relaciona con la inteligencia computacional a partir del hecho de que un agente necesita reconocer patrones en su ambiente, por ejemplo, para actuar en consecuencia. El reconocimiento de patrones, a diferencia del aprendizaje maquinal, surje dentro de la ingeniería, no la ciencia de la computación, pero pueden verse como dos facetas del mismo campo en muchas aplicaciones [135]. Por ejemplo en el reconocimiento de caracteres, secuencias de ADN, reconocimiento del habla o de hablantes, o visión artificial, donde los algoritmos deben "aprender" de los datos la manera de identificar ciertos patrones. Así, ambas áreas han evolucionado en el tiempo, aunque el aprendizaje maquinal es el concepto predominante a la hora de describir problemas de clasificación. Esto se ha vuelto especialmente cierto a partir del advenimiento del aprendizaje profundo (*deep learning* en inglés) y sus relaciones con otras áreas de investigación actuales como el análisis de grandes cantidades de datos (*big data* en inglés) o minería de datos (*data mining* en inglés). Como ejemplo de este devenir histórico, la Figura 4.1 muestra la relevancia de los términos *pattern recognition*, *machine learning*, y *deep learning* en las búsquedas de Google entre los años 2004 y 2021. Puede verse que si bien en algún momento "aprendizaje maquinal" y "reconocimiento de patrones" eran términos utilizados con la misma relevancia, con el correr de los años este último a caído en desuso. Por el contrario, el aprendizaje maquinal ha aumentado su relevancia, principalmente por la importancia que ha ganado el aprendizaje profundo en el último tiempo.

En este capítulo se presentarán algunos elementos de aprendizaje maquinal que se utilizarán

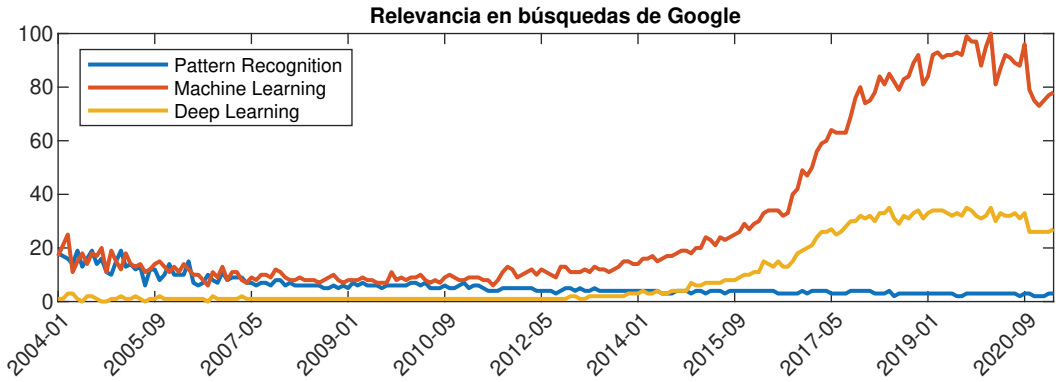


Figura 4.1: Relevancia en las búsquedas de Google de los términos "*pattern recognition*", "*machine learning*", "*deep learning*" (figura realizada con información de Google Trends).

posteriormente, principalmente en el Capítulo 5, para la clasificación de señales de voz. La clasificación es una tarea típica del reconocimiento de patrones, por lo que las técnicas descritas en este capítulo bien podrían ser consideradas herramientas de ese área. En el contexto de este capítulo, se denominarán *patrones* a vectores de características, que han sido extraídas convenientemente, generalmente en base al conocimiento del problema a tratar. La primera sección de este capítulo versa sobre la reducción de la dimensionalidad de los patrones mediante técnicas de selección de características. Seguido de esto se describirán las máquinas de vectores de soporte, un algoritmo de clasificación que utiliza el paradigma de *aprendizaje supervisado* para aprender a clasificar patrones en diferentes categorías. Bajo este paradigma, la información de la clase a la que efectivamente pertenecen los patrones es utilizada durante el entrenamiento, y es el clasificador el que "aprende" una regla de asignación de un vector de características a una clase determinada. Finalmente se detallarán algunos conceptos generales sobre la medición del desempeño de un algoritmo de clasificación. De aquí en más se denotarán en **negrita** a los a los vectores, por ejemplo $\mathbf{x}$, que serán considerados vectores columna a menos que se explicite lo contrario, y se denotará $\mathbf{x}^T$ a la traspuesta de estos elementos. Asimismo, se denotarán a las matrices en negrita y mayúscula, por ejemplo $\mathbf{C}$.

4.2. Selección de características

La reducción del número de características utilizadas para representar una señal en un problema de clasificación permite contrarrestar la influencia de la conocida *maldición de la dimensionalidad*, consistente en la necesidad de aumentar la cantidad de datos exponencialmente al aumentar la cantidad de descriptores [135, 136]. Otras ventajas pueden ser desde mejorar el entendimiento del problema (cuanto menor sea la dimensión de los datos, más fácil es analizar las relaciones entre ellos y su influencia en la clasificación) hasta reducir el tiempo que se emplea en su clasificación (importante para sistemas *en línea*, por ejemplo).

A grandes rasgos, existen dos alternativas para reducir el número de características. El primer enfoque se basa en reducir la dimensión de los datos mediante su proyección en un subespacio que conserve la mayor parte de la varianza de los datos originales, por ejemplo mediante análisis de componentes principales (PCA, del ingés *principal component analysis*). Sin embargo, la proyección de los datos en las direcciones de las componentes principales puede ignorar aquellas direcciones necesarias para distinguir entre clases [136]. Esto se debe a que PCA busca aquellas direcciones *eficientes para la representación*, que no necesariamente deben coincidir con aquellas que son *eficientes para la discriminación*. La segunda alternativa es la *selección* de características, donde se busca conservar sólo aquellas que maximicen el desempeño del clasificador, y descartar aquellas que no aporten información.

Los métodos de selección de características pueden dividirse en tres categorías: filtrado, envolventes o embebidos. Los métodos de filtrado suelen aplicarse como preprocesamiento, y consisten en la eliminación de atributos que no aporten información, o que aporten información redundante.

Los métodos envolventes, por otro lado, buscan seleccionar las características que maximicen el desempeño de un clasificador previamente elegido. La diferencia fundamental con los métodos de filtrado es que en este caso se tiene en cuenta el clasificador a utilizar, y la evaluación de las características a seleccionar se realiza entrenando y validando el modelo cada vez que se alteran las características utilizadas. En consecuencia son computacionalmente más costosos que los de filtrado [137–139]. Otra consecuencia de este tipo de métodos es que las características seleccionadas podrían no ser óptimas para el mismo problema utilizando otro clasificador, ya que fueron seleccionadas para maximizar el desempeño de un modelo en particular.

Por último, los métodos embebidos realizan la selección de características simultáneamente con el entrenamiento de un clasificador. La diferencia con los métodos envolventes es que la selección de características se expresa en la forma de un vector de pesos que pondera la relevancia de cada característica. Estos métodos son computacionalmente menos costosos que los envolventes, a la vez que han demostrado funcionar bien en una importante cantidad de casos, utilizando datos reales y artificiales [140].

A continuación se describirá el método de selección de características por vecinos más cercanos y el método de selección de características hacia adelante. El primero, corresponde a un método embebido, basado en la maximización del desempeño de un clasificador de K vecinos más cercanos. El segundo es un método de selección envolvente, para el que es necesario seleccionar un clasificador previamente.

4.2.1. Selección de características por vecinos más cercanos

La selección de características por vecinos más cercanos (SCVMC) busca maximizar el valor esperado de la exactitud de la clasificación mediante vecinos más cercanos, empleando un método de ascenso por gradiente [140]. Este problema de optimización es planteado utilizando un término de regularización que promueve la raleza en la dimensión de los datos empleados, de manera tal que el máximo es alcanzado a la vez que se obtiene un vector de pesos que pondera las características más importantes. A continuación se desarrollará la derivación del método para entender cómo son seleccionadas las características.

Sea $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_j, y_j), ..., (\mathbf{x}_N, y_N)\}$ un conjunto de datos, donde $\mathbf{x}_j$ es un vector de características de dimensión d, N el número total de patrones, $y_j \in \{1, ..., K\}$ es la etiqueta de clase correspondiente al j-ésimo patrón y K el número de clases del problema. El objetivo del método es encontrar un vector de pesos $\mathbf{w}$ que asigne un peso mayoritario a las características más relevantes para el problema al maximizar la exactitud de *leave-one-out* (se valida sobre un dato a la vez) de un clasificador basado en los vecinos más cercanos. Así, cada dato es clasificado como la clase mayoritaria de los más próximos en el espacio de características, y la exactitud es la proporción de los datos bien clasificados. Dicha función de exactitud no es diferenciable, depende de la moda del conjunto de etiquetas de los datos más cercanos, y, en consecuencia, no puede optimizarse mediante un método basado en derivadas tal como está definida [140].

Con el objetivo de franquear este obstáculo, es posible aproximar la función de exactitud para un clasificador de vecinos más cercanos mediante una distribución de probabilidad. La probabilidad de que para un patrón $\mathbf{x}_j$ se seleccione un patrón $\mathbf{x}_k$, como referencia para la clasificación, puede expresarse como [140]:

$$p_{jk} \begin{cases} \dfrac{\exp(-\mathcal{D}(\mathbf{x}_j, \mathbf{x}_k)/\sigma)}{\sum\limits_{i \neq k} \exp(-\mathcal{D}(\mathbf{x}_j, \mathbf{x}_k)/\sigma)}, & \text{si } i \neq k \\ 0, & \text{si } i = k \end{cases} \tag{4.1}$$

donde $\mathcal{D}(\mathbf{x}_j, \mathbf{x}_k)$ es una función de la distancia entre $\mathbf{x}_j$ y $\mathbf{x}_k$, y $\exp(-\mathcal{D}(\mathbf{x}_j, \mathbf{x}_k)/\sigma)$ una función que decrece con la distancia entre los datos, de manera tal que la probabilidad de que un patrón $\mathbf{x}_k$ sea elegido como referencia para determinar la clase de $\mathbf{x}_j$ disminuya cuanto más lejos se encuentre $\mathbf{x}_k$ de $\mathbf{x}_j$. El parámetro σ establece el ritmo al que decrece dicha probabilidad con la distancia. Para $\sigma \to 0$ sólo el patrón más cercano determinará la clase de $\mathbf{x}_j$, mientras que para $\sigma \to \infty$ todos los patrones tienen igual probabilidad de influenciar la clase de $\mathbf{x}_j$.

A partir de p_{jk}, es posible escribir la función exactitud buscada como:

$$A = \frac{1}{N}\sum_{j=1}^{N} p_j = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N} y_{jk}p_{jk}, \tag{4.2}$$

donde $y_{jk} = 1$ si $y_j = y_k$, y $y_{jk} = 0$ si $y_j \neq y_k$.

Definiendo ahora una función distancia $\mathcal{D}(\mathbf{x}_j, \mathbf{x}_k)$ como

$$\mathcal{D}_{\mathbf{w}}(\mathbf{x}_j, \mathbf{x}_k) = \sum_{\ell=1}^{d} w_\ell |x_{j\ell} - x_{k\ell}| \tag{4.3}$$

donde $x_{j\ell}$ es la ℓ-ésima componente del dato $\mathbf{x}_j$ y w_ℓ es el peso asociado a la ℓ-ésima característica, podemos redefinir a la función exactitud como:

$$A(\mathbf{w}) = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{N} y_{jk}p_{jk} - \lambda\sum_{\ell=1}^{d} w_\ell^2 \tag{4.4}$$

donde el último término de regularización promueve la raleza del vector $\mathbf{w}$ al maximizar $A(\mathbf{w})$. Finalmente, la derivada de $A(\mathbf{w})$ puede computarse como [140]:

$$\frac{\partial A(\mathbf{w})}{\partial w_\ell} = 2\left(\frac{1}{\sigma}\sum_{j=1}^{N}\left(p_j\sum_{j\neq i} p_{jk}|x_{j\ell} - x_{k\ell}| - \sum_{k}^{N} y_{jk}p_{jk}|x_{j\ell} - x_{k\ell}|\right) - \lambda\right) w_\ell, \tag{4.5}$$

a partir de la cual puede utilizarse el método de ascenso por gradiente para maximizar $A(\mathbf{w})$.

Como resultado final de este método se obtiene el vector $\mathbf{w}$, a partir de cuyos pesos pueden seleccionarse las características más relevantes, ya sea mediante un umbral, o bien mediante una selección posterior que requiera ordenar las características por orden de relevancia (por ejemplo un método envolvente).

4.2.2. Selección secuencial de características hacia adelante

Este método envolvente consiste en una búsqueda heurística de una combinación óptima de descriptores que aumente el desempeño de un clasificador previamente seleccionado. Dado que es un método voraz, no está asegurado que encuentre el óptimo global de este problema. Asimismo, la selección de características será adecuada para el clasificador elegido, no siendo posible afirmar que es una combinación óptima para cualquier algoritmo de clasificación [141].

La búsqueda secuencial hacia adelante requiere, en primer lugar, ordenar las características en función de su capacidad de discriminación en base a un criterio independiente (por ejemplo el test-t, el área bajo la curva ROC (del inglés *Receiver Operator Characteristic*) [143], o incluso el vector de pesos obtenido con el método explicado en la sección anterior). De esta forma, puede seleccionarse la *mejor* característica en términos del criterio elegido. Seguido de esto, la búsqueda consiste en estimar el desempeño de todos los pares de características formados por el mejor descriptor y cada uno de los restantes. Aquel par con el mejor desempeño es conservado, y a continuación se prueban todas las ternas conformadas por el par elegido anteriormente y cada una de las características restantes. Así, en cada paso, se agrega secuencialmente una característica. El algoritmo puede detenerse al encontrar una combinación de un número de descriptores fijado previamente, o puede continuar hasta que no queden descriptores para evaluar [141].

4.3. Máquinas de vectores de soporte

Las máquinas de vectores de soporte (SVM, del inglés *support vector machine*) son un conjunto de algoritmos surgidos a partir de la teoría estadística del aprendizaje, y cuyo objetivo es hallar un hiperplano óptimo de separación entre dos clases basado en sólo unos pocos datos, llamados *vectores de soporte* [135, 144, 145]. Como "óptimo" se considera aquí a aquel hiperplano que deje el mayor margen posible entre ambas clases. El margen se define como la mínima distancia entre la frontera de decisión y el dato más cercano a ella, como se muestra para un caso bidimensional en la Figura 4.2.

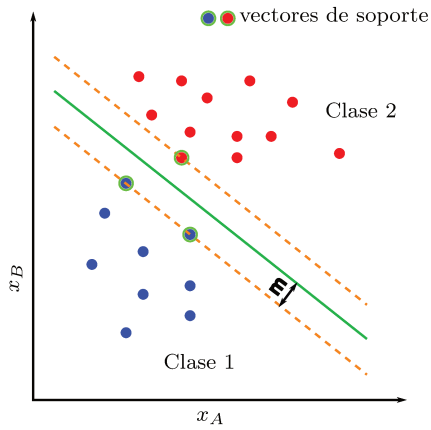


Figura 4.2: Gráfico de dispersión que muestra la recta óptima de separación entre dos clases hallada mediante SVM (en verde). Esta se encuentra en el punto medio entre las rectas paralelas (línea de trazos) que tocan las muestras más cercanas de cada clase, o vectores de soporte. Se indica con "m" el margen.

4.3.1. Formulación primal y dual

Considerando un conjunto de datos $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_j, y_j), ..., (\mathbf{x}_N, y_N)\}$ linealmente separable, donde $\mathbf{x}_j$ es un dato de dimensión d y $y_j \in \{-1, 1\}$, es posible formular un problema de optimización para encontrar el hiperplano de mayor margen. Considérese la función discriminante lineal

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0, \tag{4.6}$$

donde $\mathbf{w} \in \mathbb{R}^d$ es un vector de pesos (no confundir con el vector de pesos utilizado en la Sección 4.2.1) y w_0 es conocido como umbral o sesgo. Para distintos valores de $\mathbf{w}$ y w_0, la Ecuación 4.6 describe diferentes hiperplanos, siendo $g(\mathbf{x}) = 0$ la frontera de decisión. Las rectas en línea de trazos de la Figura 4.2 determinan los márgenes hacia cada lado del hiperplano e intersecan a los vectores soporte. Las ecuaciones que las describen son:

$$\mathbf{w}^T\mathbf{x} + w_0 = 1 \text{ Para } \mathbf{x} \text{ perteneciente a la clase 1 } (y = 1).) \tag{4.7}$$

$$\mathbf{w}^T\mathbf{x} + w_0 = -1 \text{ Para } \mathbf{x} \text{ perteneciente a la clase 2 } (y = -1).) \tag{4.8}$$

donde los valores 1 y -1 se han elegido por conveniencia, aunque sin pérdida de generalidad. Dado que es posible demostrar que la distancia de $\mathbf{x}_j$ al hiperplano de separación es [146]:

$$y_j \frac{g(\mathbf{x}_j)}{\|\mathbf{w}\|}, \tag{4.9}$$

donde $\|\mathbf{w}\| = \left(\sum_{j=1}^{d} w_j^2\right)^{1/2}$, se definirá al margen m, ver Figura 4.2, como:

$$m(\mathbf{w}) = \frac{2y^\star g(\mathbf{x}^\star)}{\|\mathbf{w}\|}, \tag{4.10}$$

donde $\mathbf{x}^\star$ es un vector de soporte y $y^\star$ es su etiqueta. Como $y^\star g(\mathbf{x}^\star) = 1$ en ese caso, el margen queda definido como:

$$m(\mathbf{w}) = \frac{2}{\|\mathbf{w}\|}. \tag{4.11}$$

Para poder encontrar la mejor región de separación será necesario maximizar el margen hasta las muestras más próximas de cada clase, de manera tal que ninguna muestra quede mal clasificada, lo que constituye una restricción a aplicar en el problema de maximización. Maximizar $m(\mathbf{w})$ es equivalente a minimizar $\|\mathbf{w}\|^2$, en consecuencia el problema a resolver estará dado por:

$$\arg\min_{\mathbf{w},w_0} \frac{1}{2}\|\mathbf{w}\|^2. \tag{4.12}$$

El producto $y_j[w^T\mathbf{x}_j + w_0]$ será mayor o igual a 1 para cualquier clase bien clasificada, por lo tanto, es posible formular las restricciones del problema como:

$$y_j[\mathbf{w}^T\mathbf{x}_j + w_0] \geq 1. \tag{4.13}$$

Utilizando estos elementos y aplicando multiplicadores de Lagrange, denotados con α_j, es posible expresar el Lagrangiano del problema (4.12) de la siguiente forma:

$$L(\mathbf{w}, \boldsymbol{\alpha}, w_0) = \frac{\mathbf{w}^T\mathbf{w}}{2} + \sum_{j=1}^{N} \alpha_j(1 - y_j(\mathbf{w}^T\mathbf{x}_j + w_0)), \tag{4.14}$$

cuya minimización se conoce como *formulación primal* de SVM. Derivando ahora $L(\mathbf{w}, \alpha)$ respecto a $\mathbf{w}$ e igualando a 0 resulta:

$$\mathbf{w} = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j. \tag{4.15}$$

Luego, derivando respecto a w_0 e igualando a 0 se obtiene:

$$\sum_{j=1}^{N} \alpha_j y_j = 0. \tag{4.16}$$

Es posible obtener una nueva formulación del problema reemplazando las Ecuaciones (4.15) y (4.16) en (4.14), resultando en el problema de maximización de la siguiente función objetivo, denominado *formulación dual*,

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N} \alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j\alpha_k y_j y_k \mathbf{x}_j^T\mathbf{x}_k, \tag{4.17}$$

sujeta a las restricciones $\alpha_j \geq 0$ y $\sum_{j=1}^{N} \alpha_j y_j = 0$. Tanto la formulación primal como la dual son problemas de optimización *convexos*, que pueden resolverse utilizando programación cuadrática [135, 147, 148]. Para clasificar un nuevo dato, $\mathbf{x}$, sólo se debe determinar el signo de $g(\mathbf{x})$. Reemplazando la Ecuación (4.15) en la Ecuación (4.6) se obtiene la siguiente expresión para evaluar la clase de $\mathbf{x}$:

$$g(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}^T\mathbf{x}_j + w_0. \tag{4.18}$$

Un detalle interesante puede destacarse de las Ecuaciones (4.17) y (4.18), y es que ambas expresiones dependen de un producto interno. Para la Ecuación (4.17), es el producto interno entre todos los datos, mientras que para (4.18) es entre el nuevo patrón y todos los utilizados para el entrenamiento. Esto podría parecer inconveniente, ya que si la cantidad de datos disponibles es muy grande (como en muchos problemas actuales), entonces tanto el entrenamiento como la predicción parecerían ser computacionalmente costosos. Como se verá más adelante, este no es el caso. En primer lugar, se concluirá que sólo es necesario "memorizar" los vectores de soporte, lo que constituye de hecho una ventaja importante de las máquinas de vectores de soporte frente a otros métodos [135, 146]. En segundo lugar, los productos internos en las Ecuaciones (4.36) y (4.18) pueden reemplazarse por una función de *kernel*, cuyo objetivo es transformar de manera implícita el espacio de características inicial en uno de más alta dimensión (potencialmente infinita) en el que buscar un hiperplano de separación entre las clases sea más fácil. Esto se conoce como el *truco del kernel*, al que se hará referencia más adelante.

Para comprender la importancia de los vectores de soporte, es conveniente recordar las condiciones de Karush-Kuhn-Tucker (KKT) de optimalidad para programación no lineal [148, 149] y luego ponerlas en contexto de los problemas de optimización abordados en la formulación de las SVM.

Condiciones KKT: Considerando el siguiente problema de optimización no lineal:

$$\min_{\mathbf{w}} = f(\mathbf{w}) \tag{4.19}$$

sujeto a :

$$\begin{array}{rcl} h_k(\mathbf{w}) & = & 0,\ k = 1, ..., \ell \\ v_j(\mathbf{w}) & \leq & 0,\ j = 1, ..., m \end{array} \tag{4.20}$$

El vector $\bar{\mathbf{w}} \in \mathbb{R}^d$ satisface las condiciones KKT si existe un par de vectores $\boldsymbol{\lambda} \in \mathbb{R}^\ell$ y $\boldsymbol{\alpha} \in \mathbb{R}^m$ tales que:

$$\nabla f(\bar{\mathbf{w}}) + \sum_{k=1}^{\ell} \lambda_k \nabla h_k(\bar{\mathbf{w}}) + \sum_{j=1}^{m} \alpha_j \nabla v_j(\bar{\mathbf{w}}) = 0 \tag{4.21}$$

y

$$h_k(\bar{\mathbf{w}}) = 0,\ k = 1, ..., \ell \tag{4.22}$$

$$v_j(\bar{\mathbf{w}}) \leq 0,\ j = 1, ..., m \tag{4.23}$$

$$\alpha_j v_j(\bar{\mathbf{w}}) = 0,\ j = 1, ..., m \tag{4.24}$$

$$\alpha_j \geq 0,\ j = 1, ..., m. \tag{4.25}$$

Las Ecuaciones (4.22) y (4.23) se conocen como condiciones de factibilidad primal, la Ecuación (4.24) como condición de complementariedad y finalmente la Ecuación (4.25) como condición de factibilidad dual.

Analizando las condiciones de KKT para el problema planteado en la Ecuación (4.12), es posible ver que en el óptimo debe cumplirse:

$$\alpha_j \geq 0 \tag{4.26}$$

$$y_j g(\mathbf{x}_j) - 1 \geq 0 \tag{4.27}$$

$$\alpha_j [y_j g(\mathbf{x}_j) - 1] = 0 \tag{4.28}$$

equivalentes a las condiciones de factibilidad dual, factibilidad primal y de complementariedad. Esta última implica que para cada dato, o bien $\alpha_j = 0$ o bien $y_j g(x_j) = 1$. He aquí uno de los

aspectos más relevantes de las SVM, y es que, considerando la Ecuación (4.18), aquellos datos para los que $\alpha_j = 0$, no juegan ningún papel en la predicción de nuevos datos. En contraste, aquellos puntos que satisfacen $y_j g(x_j) = 1$, que corresponden a los vectores de soporte, son los únicos datos relevantes para la clasificación y para la determinación del hiperplano. En consecuencia, sólo es necesario almacenar los vectores de soporte para la clasificación y no todo el conjunto de datos.

Queda pendiente aún la determinación de w_0. Es posible obtener el valor del umbral o sesgo a partir de cualquier vector de soporte, ya que $y^\star g(\mathbf{x}^\star) = 1$, y por ende puede despejarse el valor de w_0 de la Ecuación (4.18). No obstante, una solución práctica y más estable numéricamente consiste en promediar el valor de w_0 obtenido para cada vector de soporte:

$$w_0 = \frac{1}{N_\mathcal{S}} \sum_{j \in \mathcal{S}} \left(y_j - \sum_{k \in \mathcal{S}} \alpha_k y_k x_j^T x_k \right) \tag{4.29}$$

donde $\mathcal{S}$ es el conjunto de los vectores de soporte y $N_\mathcal{S}$ es su cardinalidad.

4.3.2. Margen suave

Como se indicó anteriormente, la solución planteada hasta aquí requiere que el problema sea linealmente separable. Sin embargo, este requisito, aunque práctico, no es realista. La mayoría de los problemas a los que se aplican las técnicas de reconocimiento de patrones no cumplen con esta suposición [135]. En consecuencia, es necesaria una formulación que permita la superposición de las clases para aplicar SVM a este tipo de problemas.

Una propuesta para lidiar con esta situación se conoce como SVM de *margen suave*, en contraste con el *margen duro* del caso linealmente separable, que permite cierta superposición de las clases y utiliza variables de holgura en el problema de optimización para relajar las restricciones de clasificación perfecta. Dichas restricciones se expresarán como:

$$y_j g(\mathbf{x}_j) \geq 1 - \xi_j \tag{4.30}$$

con una variable de holgura $\xi_j \geq 0$ por dato. Si $\xi_j = 0$, entonces el dato se encuentra correctamente clasificado. Por otro lado, si $0 < \xi_j < 1$, el dato se encuentra dentro del margen pero en el lado correcto del hiperplano. Por el contrario, si $\xi_j > 1$, el dato se encuentra en lado "equivocado" del plano, mal clasificado. Teniendo esto en cuenta, el problema a minimizar será:

$$\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{j=1}^{N} \xi_j, \tag{4.31}$$

donde el segundo término de la Ecuación (4.31) es una cota superior a la cantidad de datos mal clasificados. $C > 0$ permite controlar el compromiso entre el número de datos del lado incorrecto del hiperplano y el ancho del margen. De hecho, para $C \to \infty$ se obtiene el problema original planteado en la Ecuación (4.12), ya que en ese caso la sumatoria de las $\xi_j > 1$, es decir, datos mal clasificados, debería tender a cero para lograr la minimización. Considerando lo anterior, el Langrangiano para la formulación primal de margen suave será [135]:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{j=1}^{N} \xi_j - \sum_{j=1}^{N} \alpha_j [y_j g(\mathbf{x}_j) - 1 + \xi_j] - \sum_{j=1}^{N} \mu_j \xi_j \tag{4.32}$$

donde $\alpha_j \geq 0$ y $\mu_j \geq 0$ son los multiplicadores de Lagrange. De igual forma que para el caso linealmente separable, podemos encontrar una formulación dual del problema resolviendo:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x}_j \tag{4.33}$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{j=1}^{N} \alpha_j y_j = 0 \tag{4.34}$$

$$\frac{\partial L}{\partial \xi_j} = 0 \Rightarrow \alpha_j = C - \mu_j, \tag{4.35}$$

y reemplazando en la Ecuación (4.32) para obtener [135]:

$$L(\alpha) = \sum_{j=1}^{N} \alpha_j - \frac{1}{2} \sum_{j=1}^{N} \sum_{k=1}^{N} \alpha_j \alpha_k y_j y_k \mathbf{x}_j^T \mathbf{x}_k, \tag{4.36}$$

sujeto a :

$$0 \leq \alpha_j \leq C \tag{4.37}$$

$$\sum_{j=1}^{N} \alpha_j y_j = 0 \tag{4.38}$$

donde las restricciones (4.37) constituyen las *restricciones de caja*. Tanto la Ecuación (4.32) como (4.36) constituyen problemas de optimización que pueden resolverse utilizando programación cuadrática [135, 147, 148]. Adicionalmente, puede verse que la predicción de la clase de nuevos datos se realiza de la misma forma que en el caso de margen duro, utilizando la ecuación (4.18).

Para interpretar los diferentes resultados posibles teniendo en cuenta las variables α_j, ξ_j y μ_j, se considerarán ahora las condiciones de KKT para este caso:

$$\alpha_j \geq 0 \tag{4.39}$$

$$y_j g(\mathbf{x}_j) - 1 + \xi_j \geq 0 \tag{4.40}$$

$$\alpha_j [y_j g(\mathbf{x}_j) - 1 + \xi_j] = 0 \tag{4.41}$$

$$\mu_j \geq 0 \tag{4.42}$$

$$\xi_j \geq 0 \tag{4.43}$$

$$\mu_j \xi_j = 0. \tag{4.44}$$

Es posible ver que, como en el caso linealmente separable, debido a la Ecuación (4.41) un conjunto de los datos cumple con $\alpha_j = 0$, en cuyo caso no contribuyen con la predicción de la clase de nuevos patrones. El resto de los datos constituye los vectores de soporte, que deben satisfacer $y_j g(\mathbf{x}_j) = 1 - \xi_j$ con $\alpha_j \neq 0$. Si $\alpha_j < C$ entonces, por la Ecuación (4.35), $\mu_j > 0$ y a su vez, por la condición de complementariedad dada por la Ecuación (4.44), $\xi_j = 0$. Esto implica que, para los datos que cumplen con $\alpha_j < C$, la restricción de la Ecuación (4.30) es activa y en consecuencia se encuentran sobre los márgenes. Los puntos para los que $\alpha_j = C$ caen dentro de los márgenes y pueden estar tanto del lado correcto del hiperplano, bien clasificados, con $\xi_j < 1$, o mal clasificados con $\xi_j > 1$.

4.3.3. Funciones *kernel*

Hasta aquí, se trató únicamente el problema de buscar un hiperplano capaz de separar dos clases con el máximo margen entre ellas. Si bien se ha visto que esta solución resulta aplicable si existe cierto grado de superposición entre las clases, es posible intentar encontrar un nuevo espacio de características en el que los datos sean linealmente separables aplicando una transformación. Empleando una función $\phi(\mathbf{x})$, puede obtenerse un nuevo espacio cuya dimensión podría ser mucho mayor al espacio de características original, incluso infinita. Luego, es posible sacar ventaja de la formulación dual de SVM, reemplazando el producto interno entre los vectores $\mathbf{x}_j^T \mathbf{x}_k$ por $\phi(\mathbf{x}_j)^T \phi(\mathbf{x}_k)$ en las Ecuaciones (4.17), (4.18) y (4.36). Esto último resulta muy conveniente,

ya que de esta manera no es necesario conocer explícitamente la función $\phi(\mathbf{x})$ sino el producto interno $\phi(\mathbf{x}_j)^T\phi(\mathbf{x}_k)$. Con ese objetivo, una una función *kernel* $\mathcal{K}$ se define como [146]:

$$\mathcal{K}(\mathbf{p},\mathbf{q}) = \phi(\mathbf{p})^T\phi(\mathbf{q}), \forall \mathbf{p},\mathbf{q} \in \mathcal{X} \tag{4.45}$$

donde $\mathcal{X}$ es el espacio de características original y $\phi(\mathbf{x})$ es una función que aplica $\mathcal{X}$ a un nuevo espacio de características $\tilde{\mathcal{X}}$. Al reemplazar $\mathbf{x}_j^T\mathbf{x}_k$ por una función de *kernel* en, por ejemplo, la Ecuación 4.17 se obtiene:

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N}\alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_j\alpha_k y_j y_k \phi(\mathbf{x}_j)^T\phi(\mathbf{x}_k) \tag{4.46}$$

$$L(\boldsymbol{\alpha}) = \sum_{j=1}^{N}\alpha_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\alpha_j\alpha_k y_j y_k \mathcal{K}(\mathbf{x}_j,\mathbf{x}_k). \tag{4.47}$$

lo que se conoce comúnmente como *el truco del kernel*. Esta última expresión implica que no es necesario conocer la función $\phi(\mathbf{x})$, sino $\mathcal{K}(\mathbf{x}_j,\mathbf{x}_k)$.

Es posible visualizar la utilidad de la función *kernel* mediante un ejemplo. Las muestras de la Tabla 4.1, corresponden a la función XOR, un problema clásico para ilustrar un conjunto de datos que no es linealmente separable, graficadas en la Figura 4.3a). La siguiente transformación:

$$\phi(\mathbf{x}_j) = \begin{bmatrix} [x_{j1}]^2 \\ [x_{j2}]^2 \\ \sqrt{2}x_{j1}x_{j2} \end{bmatrix} \tag{4.48}$$

aplica los patrones del problema XOR a los puntos de un espacio tridimensional con coordenadas (0;1;0) (clase 1) y (1;0;0) (clase 2). Como se esquematiza en Figura 4.3, el problema en este nuevo espacio tridimensional $\tilde{\mathcal{X}}$ es, en efecto, linealmente separable. Esto ejemplifica claramente la ventaja de transformar el espacio de características original en uno de mayor dimensión: encontrar un hiperplano de separación en este nuevo espacio puede ser más sencillo.

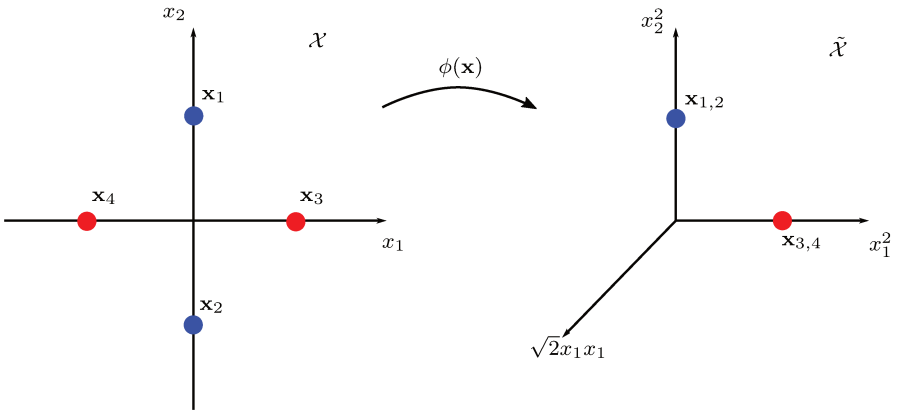


Figura 4.3: Transformación del espacio de características $\mathcal{X}$ en el espacio $\mathcal{X}$ (de mayor dimensión) mediante la función $\phi(\mathbf{x})$.

Sin embargo, no es necesario aplicar la transformación $\phi(\mathbf{x})$ a cada muestra, ya que sólo basta conocer la función de kernel para aplicar la Ecuación (4.47). Para el ejemplo dado, la función $\mathcal{K}(\mathbf{x}_j,\mathbf{x}_k)$ puede calcularse como:

$$\begin{aligned} \mathcal{K}(\mathbf{x}_j,\mathbf{x}_k) &= \phi(\mathbf{x}_j)^T\phi(\mathbf{x}_k) && (4.49)\\ &= \mathbf{x}_{i1}^2\mathbf{x}_{j1}^2 + 2\mathbf{x}_{i1}\mathbf{x}_{j1}\mathbf{x}_{i2}\mathbf{x}_{j2} + \mathbf{x}_{i2}^2\mathbf{x}_{j2}^2 \\ &= [\mathbf{x}_{i1}\mathbf{x}_{j1} + \mathbf{x}_{i2}\mathbf{x}_{j2}]^2 \\ &= [\mathbf{x}_j^T\mathbf{x}_k]^2. && (4.50) \end{aligned}$$

	$\mathbf{x}_{(1)}$	$\mathbf{x}_{(2)}$	Clase
$\mathbf{x}_1$	0	1	1
$\mathbf{x}_2$	0	-1	1
$\mathbf{x}_3$	1	0	2
$\mathbf{x}_4$	-1	0	2

Tabla 4.1: Muestras del problema XOR, no linealmente separable.

de manera tal que sólo se requiere conocer la matriz de Gram de los datos [146]. Desde el punto de vista computacional, esto es esencialmente la misma información necesaria para optimizar la Ecuación (4.36) ya que el único requisito adicional en este caso es conocer la función $\mathcal{K}(\mathbf{p}, \mathbf{q})$.

Otras funciones *kernel* utilizadas comúnmente pueden observarse en la Tabla 4.2.

Lineal	$\mathcal{K}(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{x}_j^T \mathbf{x}_k$
Polinómico	$\mathcal{K}(\mathbf{x}_j, \mathbf{x}_k) = (1 + \mathbf{x}_j^T \mathbf{x}_k)^p$
Gaussiano	$\mathcal{K}(\mathbf{x}_j, \mathbf{x}_k) = e^{-\gamma \left\| \mathbf{x}_j^T - \mathbf{x}_k \right\|^2}$

Tabla 4.2: Otras funciones *kernel* de uso habitual.

4.3.4. Estimación de probabilidad a posteriori

La probabilidad *a posteriori*, denotada como $p(y_j|\mathbf{x}_j)$, expresa la probabilidad de que un patrón sea de una determinada clase una vez conocidos todos los valores de sus características. Un número de algoritmos de clasificación se basan en la estimación de la densidad de esta probabilidad utilizando los datos de entrenamiento, particularmente aquellos basados en clasificación Bayesiana [135, 141]. Para proceder a clasificar un nuevo dato, estos modelos calculan la probabilidad a posteriori correspondiente a cada clase, y aquella categoría con el valor más alto es asignada al nuevo patrón.

Como se mostró anteriormente, las máquinas de vectores de soporte son algoritmos de clasificación netamente discriminativos. Es decir, no utilizan ninguna estimación de las probabilidades a posteriori a partir de los datos para la clasificación, sino que se basan en el signo de la función discriminante lineal para la predicción. No obstante, obtener una estimación de la probabilidad a posteriori puede ser útil en algunos casos, siendo deseable un clasificador que la estime de alguna forma [136, 150].

Una manera muy utilizada de calibrar el valor de $g(\mathbf{x})$ para estimar las probabilidades a posteriori para las SVM fue propuesta por Platt [151], y consiste en ajustar una sigmoidea al valor de $g(\mathbf{x})$ sobre los datos de entrenamiento en el marco de una validación cruzada. La función sigmoidea sugerida depende de dos parámetros A y B:

$$p_j = \frac{1}{1 + \exp(Ag(\mathbf{x}_j) + B)}. \tag{4.51}$$

Luego, para encontrar los valores de A y B se minimiza la entropía cruzada de p_j sobre los datos de entrenamiento:

$$\min_{A,B} - \sum_j \tilde{y}_j log(p_j) + (1 - \tilde{y}_j) log(1 - p_j) \tag{4.52}$$

donde

$$\tilde{y}_j = \frac{y_j + 1}{2}. \tag{4.53}$$

El método para minimizar la Ecuación (4.52) puede elegirse a conveniencia, aunque se proponen algunos enfoques en [151] y en [152].

4.3.5. Clasificación multiclase

Como se ha visto hasta aquí, las máquinas de vectores de soporte son clasificadores binarios. No obstante, en la práctica, es común la presencia de problemas en los que las clases involucradas son más de dos. En consecuencia, es necesario encontrar una estrategia para clasificar más de dos clases mediante el uso de SVM.

Un primer enfoque consiste en entrenar K clasificadores, en el que la k-ésima función discriminante $g_k(\mathbf{x})$ es entrenada considerando una clase como la clase positiva, y el resto de los datos pertenecientes a las $K-1$ clases restantes son considerados como parte de una única clase negativa. Esta estrategia es conocida como *uno contra todos* y padece ciertas desventajas. En primer lugar, el uso de las K funciones discriminantes puede dar lugar a resultados inconsistentes al clasificar un nuevo dato, por ejemplo asignándose a varias clases simultáneamente. Para evitar esto, es posible tomar el máximo valor entre las funciones $g_k(\mathbf{x})$ para todo k, aunque tampoco asegura buenos resultados ya que cada función $g_k(\mathbf{x})$ fue entrenada en un problema diferente y sus escalas no tienen porqué ser compatibles [135, 146]. En segundo lugar, en el caso de *uno contra todos*, las clases dentro del conjunto de entrenamiento quedan completamente desbalanceadas, ya que la clase negativa abarca a las muestras de todas las $K-1$ clases restantes.

Otro enfoque consiste en entrenar $K(K-1)/2$ modelos diferentes de SVM cubriendo todos los pares de clases posibles, de manera tal que luego sea posible clasificar un dato nuevo asignando la clase con mayor cantidad de "votos" de entre todos los modelos entrenados. Esta idea es comúnmente llamada *uno contra uno*. Es claro que para un valor de K muy grande, este enfoque requiere significativamente más tiempo tanto durante el entrenamiento como en la predicción de nuevos datos que en el caso *uno contra todos*. Adicionalmente, este enfoque también puede llevar a inconsistencias en la clase asignada, aunque algunos de estos problemas pueden aliviarse con enfoques más modernos como el uso de grafos acíclicos direccionados (DAGSVM) [135].

En general ambas estrategias pueden utilizarse en la práctica, aunque la aplicación de una u otra dependerá del caso [153], y puede considerarse a este factor como un hiperparámetro a ajustar para cada problema particular.

4.4. Evaluación del desempeño de un clasificador

Al estimar el desempeño de un clasificador, el verdadero interés se encuentra en estimar qué capacidad tiene el algoritmo de discriminar *nuevas* muestras, es decir, su capacidad de generalización [136, 141]. Conocer el error de generalización tiene dos utilidades muy relevantes. En primer lugar, permite conocer si el clasificador es efectivo para resolver el problema planteado. En segundo lugar, provee una medida para comparar contra otros modelos [136].

Dado que el conjunto de datos con el que se trabaja es limitado, debe tomarse una decisión sobre cómo estimar el error de generalización sin contar, en realidad, con "nuevos" datos. Si se entrenara el clasificador con todos los datos, sólo sería posible validar su desempeño con el mismo conjunto utilizado para el entrenamiento. En ese caso, la estimación del desempeño sería *optimista*, dado que es altamente probable que el clasificador se encuentre sesgado para clasificar correctamente los datos que ya conoce. Este enfoque, en consecuencia, queda descartado ya que el error obtenido tiene un importante sesgo. En la práctica, los conjuntos de entrenamiento y validación jamás deben compartir datos [141].

4.4.1. Validación Cruzada

Otra posibilidad consiste en generar una partición de los datos para luego entrenar el clasificador con una parte, y validar con otra. De esta forma puede entrenarse y evaluarse el clasificador

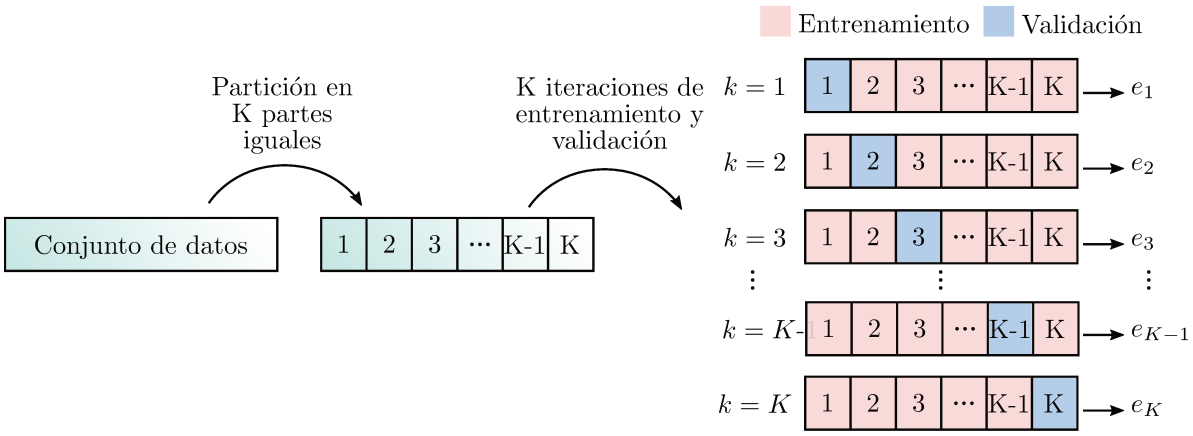


Figura 4.4: Esquema del procedimiento de validación cruzada. En el primer lugar se genera una partición en K conjuntos de datos de tamaño aproximadamente igual. Luego se realizan K entrenamientos y validaciones, cambiando la partición utilizada para entrenamiento y validación. Finalmente se promedian los errores obtenidos de cada iteración para estimar el error de generalización.

con conjuntos disjuntos de datos y así reducir el sesgo en la estimación del error de generalización [136]. El precio a pagar por esta ventaja es una reducción en el número de datos a utilizar para entrenar, aunque el entrenamiento suele ocupar entre $80\,\%$ y $90\,\%$ de los datos, y los datos restantes se reservan para la validación [141]. Si bien el sesgo es menor en este enfoque, aún presenta una desventaja. Dado que sólo hay un conjunto de validación, sólo se obtiene una única estimación del error. Por lo tanto no es posible estudiar la varianza de este estimador.

A partir de esta situación se observa la necesidad de obtener un número K de conjuntos de entrenamiento y validación, de manera tal que sea posible obtener una estimación de la media y de la varianza del error de generalización. Una estrategia para lograr este cometido se conoce como *validación cruzada* y consiste dividir la totalidad de patrones en K conjuntos disjuntos [136, 154]. Luego, el clasificador es entrenado y validado K veces, utilizando $(K-1)$ conjuntos para el entrenamiento y validando en el conjunto restante. La Figura 4.4 muestra un esquema de la partición de los datos en cada iteración, y su designación para entrenamiento o validación. Valores habituales de K suelen ser 5 o 10, aunque este valor depende de la cantidad de datos disponibles. Para cada iteración se obtiene el error de clasificación e_k y finalmente el error estimado del método es la media de los K errores obtenidos:

$$\bar{e}_{vc} = \frac{1}{K}\sum_{k=1}^{K} e_k. \tag{4.54}$$

De igual manera, puede estimarse la varianza del estimador como:

$$\mathrm{var}\{e_{vc}\} = \frac{1}{K-1}\sum_{k=1}^{K}(e_k - \bar{e}_{vc})^2. \tag{4.55}$$

4.4.2. Matriz de confusión

Las matrices de confusión o tablas de contingencia permiten resumir en una única tabla el desempeño de un clasificador [141, 143]. Para el caso de un clasificador binario, un ejemplo de esta matriz puede verse en la Figura 4.5. Los elementos de la diagonal de la matriz de confusión determinan la cantidad (o proporción) de patrones correctamente clasificados. Los restantes elementos de cada fila, indican la cantidad (o proporción) de los elementos de la clase correspondiente a esa

		Salida del Clasificador	
		P	N
Clase Verdadera	P	VP	FP
	N	FN	VN

$$\text{SENSIBILIDAD} = \frac{TP}{P}$$

$$\text{ESPECIFICIDAD} = \frac{TN}{N}$$

Figura 4.5: Matriz de confusión para la salida de un clasificador binario con clases "P" (positiva) y "N" (negativa). Se muestran las fórmulas para el cálculo de la sensibilidad y la especificidad a partir de la matriz de confusión, donde VP, FP, VN, FN se refiere a *verdaderos positivos*, *falsos positivos*, *verdaderos negativos* y *falsos negativos*, respectivamente.

fila clasificados, erróneamente, como otras clases [141]. Para el caso particular de dos clases, la matriz de confusión es la base de varias métricas comúnmente asociadas a un clasificador binario, como la sensibilidad o la especificidad (ver Figura 4.5) [143].

La utilidad principal de estas matrices consiste en su capacidad de describir rápidamente los resultados obtenidos, detallados por clase. Esto adquiere una relevancia aún mayor cuando el conjunto de datos no está balanceado, es decir, las clases no presentan la misma prevalencia en el conjunto de datos. Para comprender porqué esto es importante, es necesario recordar que ante un conjunto no balanceado, la exactitud del clasificador puede no ser un buen indicativo del desempeño. Por ejemplo, considérese un conjunto de 100 datos con dos clases, en el cual 90 patrones pertenecen a la clase "P", mientras que 10 pertenecen a la clase "N". En este caso, podríamos obtener una exactitud del $90\,\%$ solamente afirmando que todos los patrones pertenecen a la clase "P". Como es lógico, este resultado debe revisarse a la luz de la exactitud *por clase*, para poder ver que si bien para la clase "P" hay una exactitud del $100\,\%$, para la clase "N" es de $0\,\%$ y, en consecuencia, el resultado probablemente carezca de utilidad práctica [143].

Para el caso de un modelo de clasificación binario, la sensibilidad y la especificidad expresadas en porcentaje son equivalentes a las exactitudes por clase. En general, lo deseable es obtener un clasificador cuya exactitud sea lo más alta posible a la vez que la sensibilidad y especificidad tengan aproximadamente el mismo valor, aunque esto podría no ser necesario en algunos problemas particulares (como métodos de tamizado o *screening*, donde una sensibilidad mayor es aceptable) [141, 143].

En caso de tener $K > 2$ clases, la matriz de confusión $\mathbf{C}$ será de tamaño $K \times K$, y la exactitud de cada clase puede calcularse como el cociente entre cada elemento de la diagonal y la suma de los elementos de cada fila (siempre que la matriz contenga cantidades, y no proporciones, de los datos clasificados bajo cada clase) [141]:

$$\text{Exactitud}_j = \frac{c_{jj}}{\sum_{k=1}^{K} c_{jk}} \times 100\,\%, \tag{4.56}$$

donde Exactitud_j es la exactitud correspondiente a la clase j, c_{jk} son los elementos de la matriz de confusión $\mathbf{C} \in \mathbb{R}^{K \times K}$, y c_{jj} los elementos de la diagonal.

Luego, la exactitud general del clasificador puede expresarse como:

$$\text{Exactitud} = \frac{\text{tr}\{\mathbf{C}\}}{\sum_{j=1}^{K} \sum_{k=1}^{K} c_{jk}}, \tag{4.57}$$

donde $\text{tr}\{\mathbf{C}\}$ es la traza de la matriz de confusión.

La estimación de la matriz de confusión también puede obtenerse por validación cruzada, así como cualquier otro estadístico sobre la clasificación [154], construyendo la matriz para cada iteración de la validación y luego promediando las matrices obtenidas.

4.5. Comentarios de final de capítulo

En este capítulo se presentaron aquellas herramientas provenientes del aprendizaje maquinal que se emplearán en el Capítulo 5 para la clasificación de voces.

La selección de características es crucial para reducir la dimensionalidad de los datos de entrada. Ambos métodos presentados, selección por vecinos más cercanos y selección secuencial hacia adelante, serán utilizados en combinación en el Capítulo siguiente. El primero para obtener una valoración de la relevancia de las características individualmente en forma de un vector de pesos, mientras que el segundo método se utilizará para realizar una búsqueda voraz de la combinación óptima de características previamente ordenadas en base a dicho vector de ponderación. Según [140], el algoritmo de selección de características por vecinos más cercanos es insensible a los valores de σ y λ de la Ecuación (4.4), aunque en la práctica es común encontrar el valor del parámetro de regularización λ mediante validación cruzada con un pequeño subconjunto de los datos que luego no se utilizarán para el entrenamiento.

Si bien aquí se ha realizado una presentación superficial de las SVM, más detalles sobre su desarrollo teórico e implementación, u otras aplicaciones como regresión mediante SVM, pueden encontrarse en bibliografía especializada como [135, 145, 146, 155, 156].

Capítulo 5

Tipificación automática de voces

5.1. Introducción

Como se detalló en el Capítulo 2, las señales de voz son clasificadas por especialistas con el objetivo de evaluar su idoneidad para la aplicación de medidas de perturbación, que dependen fuertemente de que las señales a analizar sean aproximadamente periódicas.

La clasificación consiste en tres tipos (ver comentarios respecto a un cuarto tipo de señales en la Sección 2.6.2), donde el tipo 1 se corresponde con señales que efectivamente son aproximadamente periódicas, el tipo 2 con señales que presentan cierta periodicidad pero que se encuentran afectadas por frecuencias subarmónicas y modulantes, y finalmente el tipo 3 se corresponde con señales de periodicidad no evidente.

Para la determinación del tipo de una señal, los especialistas en el cuidado de la voz se apoyan mayoritariamente en el uso de espectrogramas de banda angosta [42, 110]. Esta información visual que obtienen de los espectrogramas es normalmente complementada con gráficos temporales de la señal así como también de la impresión auditiva que se obtiene al escucharla. Esto permite detectar diferentes tipos de ruido así como también cambios importantes en el tono percibido (*pitch*). Dado que los profesionales del cuidado de la voz utilizan esta información *perceptual*, la tipificación se ha vuelto una tarea subjetiva, que requiere abundante tiempo y está sujeta a cierta variación interprofesional [65, 108] ya que es afectada por aspectos del evaluador como su experiencia o su formación profesional (por ejemplo si es médico/a otorrinolaringólogo/a o fonoaudiólogo/a) [77].

Existe un número de investigaciones orientadas hacia la caracterización de sistemas basados en el reconocimiento de patrones que puedan utilizar medidas cuantitativas para asistir a los profesionales en tareas perceptuales, como la tipificación de voces. Un sistema de esta naturaleza podría ayudar a disminuir el sesgo existente en la clasificación así como también el tiempo invertido en esta tarea de preprocesamiento [61, 73]. No obstante, a pesar de la numerosa cantidad de medidas cuantitativas propuestas para determinar el tipo de voz [59–63, 65, 66], la tarea de clasificar automáticamente estas señales ha sido poco explorada en trabajos previos (a excepción de [61]). Asimismo, los parámetros sugeridos fueron validados sobre conjuntos pequeños de señales (entre 40 y 148), provenientes de un mismo corpus. Por lo que la verdadera capacidad de diferenciar entre los distintos tipos de señal de estos parámetros no se encuentra lo suficientemente investigada.

Basado en lo anterior, se propone un trabajo experimental con el siguiente par de objetivos. Primero, proponer un enfoque de reconocimiento de patrones para la clasificación automática de señales basado en características objetivas, también llamadas medidas en este contexto, y un algoritmo de clasificación. Para ello utilizaremos parámetros ampliamente difundidos en la práctica, medidas de dinámicas no lineales y dos nuevas características propuestas con el propósito de evaluar cambios en la forma de onda. En segundo lugar, validar el enfoque propuesto utilizando un número mayor de señales que todos los trabajos previos, provenientes de dos corpus conocidos, y clasificadas manualmente por dos expertas en el área.

La novedad de esta propuesta está basada en los siguientes puntos:

1. El uso de dos medidas objetivas nuevas, capaces de medir la variación de la forma de onda de la señal.

2. La evaluación de la probabilidad a posteriori de la clasificación como una medida de la confiabilidad en la clase asignada automáticamente.

3. El desempeño es estimado sobre un conjunto de señales mayor al de todos los trabajos previos, en experimentos con señales del mismo conjunto de datos y experimentos *cruzados*, en los que el entrenamiento y la validación se realizan sobre corpus diferentes.

4. En contraste con otros trabajos, se utilizaron medidas de dinámicas no lineales que no requieren la intervención del usuario, propuestas en [157].

5.2. Corpus de voces

Se utilizaron señales de voz patológicas y correspondientes al fonema /a/, provenientes de los corpus que se describen a continuación. La utilización de la vocal /a/ ofrece una configuración del tracto totalmente abierta, que no ocurre en /i/ y /u/ por la separación de las cavidades frontal y posterior de la boca, permitiendo el estudio completo de las cavidades del tracto vocal. Adicionalmente, existe evidencia de que las vocales /i/ y /u/ influencian la percepción de la calidad vocal por el grado de aproximación entre los pliegues vocales para estos fonemos vocales [73].

5.2.1. *Massachussets Eye and Ear Infirmary Voice Disorder Database*

Este corpus, denominado *Massachussets Eye and Ear Infirmary Voice Disorder Database* (MEEI), distribuido por Kay Elemetrics [112], ha sido y continúa siendo un conjunto de señales de amplísimo uso en el estudio de la señal de la voz, el habla y las patologías asociadas al aparato fonador. Está constituido por aproximadamente 750 señales, tanto de sujetos normofónicos como con fonación patológica. Las señales correspondientes a voces patológicas fueron adquiridas con una frecuencia de muestreo de 25 kHz, mientras que las de habla normal con una frecuencia de 50 kHz. Para ambas señales se utilizó una resolución de digitalización de 16 bits. Las señales de este corpus, además, se encuentran preprocesadas para asegurar que la parte estable de la alocución se encuentre presente.

5.2.2. *Saarbruecken Voice Database*

La base de datos *Saarbruecken Voice Database* (SVD) de acceso libre por internet [158], es producida y mantenida por el *Institut fur Phonetik* de la *Saarland University* y por la *Phoniatry Section* de la Caritas Clinic St. Theresia en Saarbrucken, Alemania. Contiene más de 2000 grabaciones de hablantes del idioma alemán con voces sanas y patológicas. La frecuencia de muestreo en este caso es de 50 kHz, tanto para voces sanas como patológicas, y 16 bits de resolución para la digitalización. A diferencia de MEEI, las señales de este corpus no están preprocesadas, por lo que se les aplicó un preprocesamiento consistente en: 1) Eliminación del *onset* y *offset* de la señal, de manera tal que sólo la parte estable de la fonación se encuentre presente en la señal (este trabajo se realizó manualmente); 2) Se le aplicó un pasabajos y submuestreo para disminuir la frecuencia de muestreo a 25 kHz, de manera tal que sus parámetros coincidan con aquellos del corpus MEEI.

	Tipo 1	Tipo 2	Tipo 3	Total
MEEI	185	335	129	649
SVD	175	330	108	613

Tabla 5.1: Distribución de los tipos de voces para cada base de datos.

5.2.3. Clasificación manual de las señales

Señales con interrupciones u otros problemas como la presencia de otras voces durante la grabación, o una duración menor a 800 ms, fueron descartadas. Posteriormente las señales fueron clasificadas por dos fonoaudiólogas con amplia experiencia clínica en la clasificación de señales en los tres tipos propuestos [159–161]. Sólo aquellas señales para las cuales ambas profesionales estuvieron de acuerdo en la clasificación fueron utilizadas, siguiendo la metodología de trabajos previos [59–63]. La Tabla 5.1 resume la distribución de los distintos tipos de señales para cada base de datos.

5.3. Características

La periodicidad de la señal de voz puede explicarse a grandes rasgos por dos propiedades: regularidad temporal y forma de onda [82]. La primera, a su vez, puede verse afectada por perturbaciones acústicas, como el jitter o el shimmer y el ruido aéreo. No obstante, ambas no son independientes una de otra en la práctica. Dado que son muchos los factores que pueden afectar estas propiedades, ha sido imposible, al día de hoy, encontrar un único parámetro capaz de distinguir objetivamente entre todos los tipos de voces. La razón detrás de esto puede encontrarse en el hecho de que, como muchas tareas basadas en la percepción, la tipificación es un problema multidimensional [73]. Teniendo esto en cuenta, se propone una *combinación* de características con el objetivo de reflejar las distintas influencias sobre la periodicidad de una señal.

5.3.1. Jitter y Shimmer

En primer lugar, para cuantificar la cantidad de jitter y shimmer, se utilizarán las medidas de jitter y shimmer relativas. A pesar de que se han propuesto otros estimadores del nivel de jitter y shimmer, estos dos parámetros se encuentran entre las medidas de perturbación más utilizadas. Recordando la definición de la Sección 2.5.2, el jitter relativo puede expresarse como:

$$jitter_{\%} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}|T_{i+1}-T_i|}{\frac{1}{M}\sum_{i=1}^{M}T_i} \times 100\,\%. \tag{5.1}$$

mientras que el shimmer relativo puede expresarse como:

$$shimmer_{\%} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}|A_{i+1}-A_i|}{\frac{1}{M}\sum_{i=1}^{M}A_i} \times 100\,\%. \tag{5.2}$$

donde $\{A\}_{i=1}^{M}$ se conoce como serie de amplitudes y, de manera análoga a la serie de periodos, comprende la máxima amplitud pico a pico de los M ciclos de la señal analizada.

Para calcular estas medidas se utilizó el *software* PRAAT [49], que utiliza un método de coincidencia en la forma de onda para estimar los puntos fiduciarios de inicio y fin de los ciclos de la

señal. Si la señal tiene ciclos poco definidos, por ejemplo por una baja relación señal a ruido, los puntos fiduciarios darán como resultado un valor de jitter o shimmer más elevado de lo normal. En casos extremos donde el algoritmo no puede segmentar la señal o no puede encontrar una estimación del periodo promedio, el valor devuelto por el *software* es "`indefinido`". En este caso se cambió estos valores por un valor extraordinario de $100\,\%$ con el fin de poder utilizar el vector de características en la clasificación. Cabe mencionar que no se utilizarán estas medidas aquí como una forma de estudiar la fisiología de la señal, si no como meros indicadores de las fluctuaciones en el periodo y en la forma de onda, obteniendose valores más altos de ambos descriptores a medida que el tipo de la señal aumenta de 1 a 3.

5.3.2. Razón armónicos/ruido

En segundo lugar, se utilizarán parámetros que describan la componente de ruido que afecta a la periodicidad. Un parámetro directamente relacionado con esto es la razón armónicos/ruido (HNR, del inglés *Harmonic to Noise Ratio*) [90] que cuantifica el cociente entre la energía de la componente armónica de la señal y la energía de la componente aperiódica, como el ruido aéreo. De esta forma, HNR tomará valores altos para señales aproximadamente periódicas y valores bajos para señales con pobre periodicidad, haciendo de esta medida una característica prometedora para la clasificación.

La razón entre la energía de la componente armónica y la energía de la componente ruidosa, *Harmonics to Noise Ratio* (HNR), es una medida propuesta por Yumoto y cols. [90] que busca indicar el grado de periodicidad de la señal. Se define como:

$$HNR = 10\log_{10}\left(\frac{\mathcal{S}}{\mathcal{W}}\right)\text{ dB}, \tag{5.3}$$

donde $\mathcal{S}$ es la energía correspondiente a la componente periódica de la señal, y $\mathcal{W}$ es la energía de la componente de ruido de la señal expresada en dB. Cuanto más bajo es el valor de HNR, mayor es la energía de la componente de ruido de la señal; y en consecuencia más pobre la calidad de la voz. Para el cálculo de HNR también se empleó el *software* PRAAT [49]. Existe evidencia [162] de que, en la práctica, no hay diferencias sustanciales entre el algoritmo para estimar HNR de PRAAT [163] y método original ideado descripto en [90].

El algoritmo de PRAAT no utiliza la información espectral para el cálculo del HNR. En su lugar, utiliza la función de autocorrelación de una señal continua $x(t)$, definida como:

$$r_x(\tau) = \int_{-\infty}^{+\infty} x(t)x(t-\tau)dt. \tag{5.4}$$

y define como período fundamental (T_0) al valor de τ correspondiente al máximo global de $r_x(\tau)$ (excluyendo $\tau = 0$). Luego, la energía de la señal analizada será igual al valor $r_x(0)$, y puede descomponerse como:

$$r_x(0) = r_p(0) + r_{ap}(0), \tag{5.5}$$

donde $r_p(0)$ y $r_{ap}(0)$ son la energía de la componente periódica y la energía de la componente aperiódica de la señal respectivamente.

Si se define la autocorrelación normalizada como:

$$\gamma_x(\tau) = \frac{r_x(\tau)}{r_x(0)}, \tag{5.6}$$

entonces, suponiendo que el ruido de la señal es ruido blanco (no correlacionado) aditivo, la energía de la componente periódica esta dada por:

$$\gamma_p(0) = \gamma_p(T_0) = \gamma_x(T_0). \tag{5.7}$$

Por lo tanto, la energía de la componente aperiódica es:

$$\gamma_{ap}(0) = 1 - \gamma_p(0) = 1 - \gamma_x(T_0). \tag{5.8}$$

Luego, el valor de HNR es calculado como [163]:

$$HNR = 10 \log \left(\frac{\gamma_p(0)}{\gamma_{ap}(0)} \right). \tag{5.9}$$

Reemplazando las Ecuaciones (5.7) y (5.8) en esta última expresión se obtiene:

$$HNR = 10 \log \left(\frac{\gamma_x(T_0)}{1 - \gamma_x(T_0)} \right). \tag{5.10}$$

5.3.3. Prominencia del pico cepstral

La prominencia del (primer) pico cepstral (CPP) es la amplitud (en dB) del primer máximo local del *cepstrum* de una señal, medida desde una aproximación lineal del nivel de ruido. La palabra cepstrum proviene de la inversión de las sílabas de la palabra en inglés *spectrum* (espectro). De forma similar, se han denominado otros parámetros del cepstrum tomando como base el nombre del parámetro análogo en el espectro frecuencial (ver Tabla 5.2). El cepstrum *real* [164, 165] de una señal discreta $x[n]$ se define como:

$$c[q] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left| X(e^{j\omega}) \right| e^{j\omega q} d\omega, \tag{5.11}$$

donde $X(e^{j\omega})$ es la transformada de Fourier de tiempo discreto de $x[n]$ y q es la *cuefrencia* (equivalente a un índice temporal).

cepstrum	*spectrum* (espectro)
quefrency (cuefrencia)	*frequency* (frecuencia)
rahmonic (ramónico)	*harmonic* (armónico)

Tabla 5.2: Equivalencias entre parámetros del *cepstrum* y del espectro. Entre paréntesis se consigna la "traducción" en español de estos términos que se utilizará en este documento.

La prominencia del primer pico cepstral será relevante para cuantificar la regularidad temporal de la señal [166], de hecho la cuefrencia de ese máximo local se corresponde con el periodo fundamental, así como el primer armónico del espectro tradicional se corresponde con la frecuencia fundamental. Asimismo, en [83, 167] se reporta que la amplitud del primer pico cepstral es directamente proporcional a la media geométrica del HNR. Esta medida cepstral también está altamente correlacionada con la impresión perceptual del ruido de aspiración en voces *aéreas*, con la disfonia general [82], y con los tres tipos de voces utilizados en este trabajo [168]. Dado que las impresiones perceptuales son muy importantes para los especialistas en el área a la hora de clasificar las señales, resulta lógico considerar este tipo de características. Una ventaja adicional de esta medida es que, en contraste con muchas medidas de perturbaciones, no necesita una segmentación ciclo a ciclo de la señal ni una estimación precisa de la frecuencia fundamental de la voz.

Para hallar el valor de CPP, primero es necesario calcular el cepstrum de la señal a analizar. En este caso se halló el cepstrum real utilizando la transformada discreta de Fourier:

$$c[q] = \sum_{k=0}^{N-1} \log \left| \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}nk} \right| e^{j\frac{2\pi}{N}qk}, \tag{5.12}$$

donde N es el número de muestras de la señal y q es el índice de cuefrencia discreto. Luego, debe localizarse el primer pico del cepstrum, o primer *ramónico*. Para ello se busca la cuefrencia q_{cpp} en la cual se alcanza un máximo local en la región donde se espera que se encuentre el periodo fundamental, considerando los periodos fisiológicamente posibles (correspondientes a frecuencias fundamentales entre 50 y 300 Hz).

Seguido de esto, se ajusta una recta al cepstrum, de manera tal que se obtenga una aproximación lineal del nivel de ruido que se denominará $c_\ell[n]$. De esta forma es posible establecer un "piso" desde el cual medir la prominencia del pico. Finalmente CPP es calculado como [166]:

$$CPP = c[q_{cpp}] - c_\ell[q_{cpp}]\ (dB). \tag{5.13}$$

5.3.4. Medidas de dinámicas no lineales

Estas medidas fueron usadas previamente para la tipificación de señales [59, 62, 64, 65, 110, 157], dado que describen la dinámica de una serie temporal, y del sistema que las produce, tomando diferentes valores para señales periódicas, caóticas o estocásticas. Las características de dinámicas no lineales (DNL) utilizadas en este trabajo son:

1. Dimensión de correlación (D_2)
2. Entropía de correlación (K_2)
3. Nivel de ruido.

En contraste con trabajos previos, estas características fueron calculadas empleando un algoritmo reciente, basado en la integral de correlación U [157, 169]. A diferencia de los métodos convencionales como [111], el método utilizado aquí no requiere la intervención de un usuario y en consecuencia es completamente automático. El nivel de ruido es una medida complementaria que mide el grado al que las hipótesis para la estimación de invariantes, como D_2 y K_2, se satisfacen. En otras palabras, cuando D_2 y K_2 se calculan junto al nivel de ruido, este último puede utilizarse como indicador de qué tan confiables son estas estimaciones. Sin embargo, cuando las señales se adquieren en condiciones controladas de manera tal que otras fuentes de ruido son disminuidas al máximo posible, el nivel de ruido puede emplearse como un indicador de una componente estocástica predominante asumiendo que su variación es causada principalmente por la propia dinámica de la señal y por parámetros de inmersión inadecuados [169]. Las señales periódicas tienen usualmente valores bajos de D_2, K_2 y nivel de ruido. En contraste, las señales caóticas y estocásticas suelen estar asociadas con valores más altos de estas características.

Los invariantes calculados, D_2 y K_2, dependen de dos parámetros: retardo de inmersión (τ) y dimensión de inmersión (m) [4]. Se calcularon dichas medidas utilizando distintos valores para los parámetros de inmersión, considerándolos diferentes descriptores durante el proceso de selección de características explicado más adelante. Los valores de m utilizados fueron $4, 6, 8, 10, 12$ y 14, mientras que los retardos considerados para τ fueron $25, 50, 80$ y 110. Esto produjo un total de 48 características entre entropías y dimensiones de correlación, más una característica adicional de Nivel de Ruido (independiente de m y τ).

5.3.5. Nuevas características propuestas

Hasta aquí se han presentado medidas que describen la regularidad temporal de la señal, pero no se ha abordado la evaluación de la forma de onda. Aunque el shimmer relativo describe cambios promedio en la forma de onda, sólo representa desvíos en la amplitud máxima, y no en la *forma* propiamente dicha. En consecuencia, esta medida es insensible a cambios más generales de la morfología de los ciclos. Con el objetivo de cuantificar estos cambios, se proponen dos nuevas

medidas que llamaremos Varianza Normalizada de la Componente Principal (VNCP) y su desvío estándar (DSVNCP). La primera consiste en el cálculo de la varianza explicada por la componente principal del conjunto de periodos de la señal. La segunda consiste en obtener el desvío estándar de una versión de VNCP de tiempo corto. Para ello, se calcula el valor de VNCP para ventanas de la señal, y luego se obtiene el desvío estándar de estas medidas. A continuación se detalla el cálculo de estos dos descriptores.

Varianza Normalizada de la Componente Principal

En primer lugar, es necesario segmentar la señal en sus periodos, como en el caso de jitter y shimmer relativo (ver Figura 5.1). Con el objetivo de tener una segmentación similar a esas medidas, también se utilizó PRAAT [49] para la identificación del principio y fin de cada ciclo.

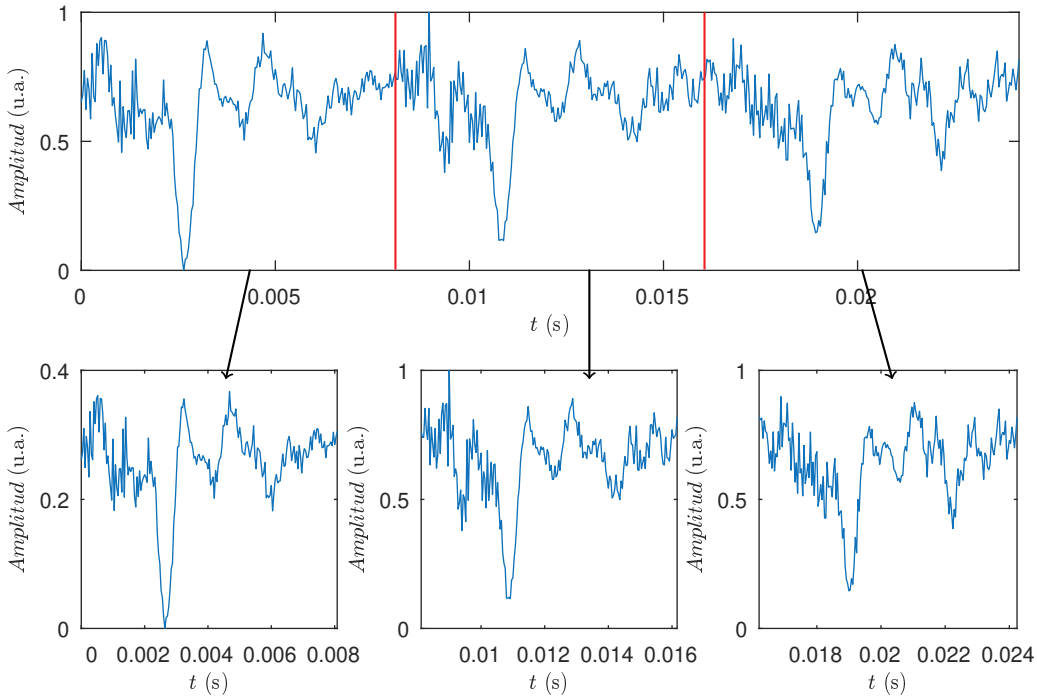


Figura 5.1: Segmentación de la señal en cada uno de los periodos.

El paso siguiente consiste en "estirar" cada segmento $b_i[n]$ a la longitud del de mayor duración, de manera tal que todos los segmentos tengan el mismo número de muestras. Con este objetivo, se buscó el periodo de mayor duración, cuyo número de muestras será N_{max}. Luego, para lograr que todos los segmentos tengan la misma duración, es necesario *aumentar* el número de muestras N_i de aquellos segmentos $b_i[n]$ para los que $N_i \leq N_{max}$. Esto es equivalente a *remuestrear* cada periodo con una frecuencia de muestreo f_s^i, mayor a la frecuencia de muestreo original f_s, tal que:

$$f_s^i = \frac{N_{max}}{N_i} f_s. \tag{5.14}$$

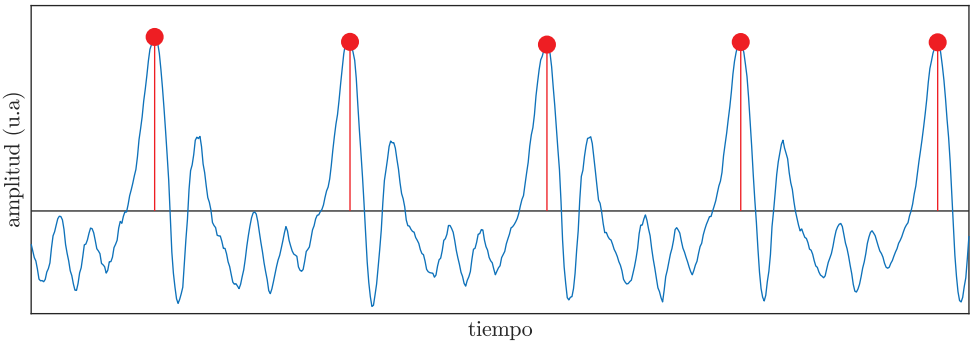
La implementación de este último paso se realizó mediante la interpolación lineal de cada segmento para una nueva cantidad de muestras equiespaciadas N_{max}.

Luego de que todos los segmentos tengan el mismo número de muestras, se realiza un análisis de componentes principales [135] para este conjunto de segmentos de señal. La primera componente principal podría considerarse como la forma de onda más representativa de la señal, como se puede observar en la Figura 5.2b, donde se superponen todos los ciclos con la misma duración y

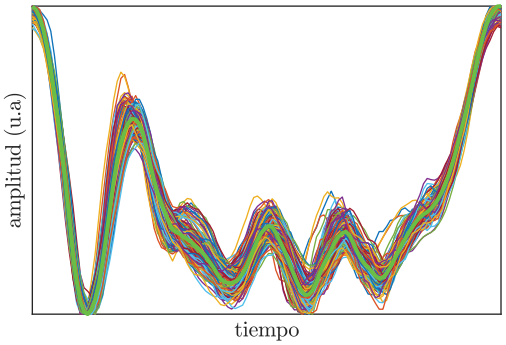
sobre ellos la primera componente principal. El porcentaje de la varianza explicado por esta componente será mas grande cuanto mayor sea la regularidad de la forma de onda la señal, ya que en ese caso los periodos se parecerán más entre sí. En consecuencia, el valor reportado finalmente es el porcentaje de la varianza total debido a la componente principal, calculado como:

$$\text{VNCP} = \frac{\lambda_1}{\sum_{i=1}^{P} \lambda_i} \times 100\,\%, \tag{5.15}$$

donde λ_1 es la varianza de la componente principal, λ_i son las varianzas de cada componente, y P el número de componentes calculadas, que coincide con M, la cantidad de ciclos, si se utilizan todas las componentes.



(a)



(b)

Figura 5.2: Segmentación de una señal y representación de la primera componente principal para el cálculo de VNCP. a) Señal con los puntos fiduciarios de segmentación superpuestos (en rojo). Luego cada ciclo de la señal, entre dos puntos fiduciarios adyacentes, es segmentado y sobremuestreado para el cálculo de VNCP. b) Superposición de todos los periodos de la señal sobremuestreados a la misma longitud y la primera componente principal (en verde y línea más gruesa). Puede apreciarse que la primera componente principal posee una forma de onda representativa del conjunto de ciclos de la señal.

Se espera que para las señales tipo 1, con formas de onda más similares periodo a periodo, la componente principal de los segmentos de onda sea responsable de un porcentaje de la varianza mayor al mismo porcentaje calculado para señales tipo 2 y tipo 3. Para señales tipo 3, los trozos de señal obtenidos mediante el algoritmo de segmentación no se corresponden con verdaderos ciclos de la señal, por lo que la varianza explicada cae a valores más pequeños en este caso que para los tipos 1 y 2. El Algoritmo 2 resume los pasos para calcular el valor de VNCP para una señal $x[n]$.

Nota: El uso del algoritmo para el cálculo de valores singulares λ_i (SVD, del inglés *singular values decomposition*) puede optimizar el paso 7 del algoritmo 2 al entregar los valores ordenados.

Algoritmo 2: Estimación de la variación normalizada de la componente principal (VNCP) para una señal $x[n]$.

Parámetros de entrada: Señal $x[n]$
Parámetros de salida : Valor de VNCP

1 Determinar los puntos fiduciarios de inicio y fin de cada ciclo de la señal $x[n]$, $0 < n < N$, donde N es la cantidad de muestras de la señal.
2 Segmentar la señal en cada ciclo $b_i[n],\ 1 < i < M$ donde M es el número de ciclos de la señal.
3 Calcular la cantidad de muestras de cada ciclo $N_i,\ 1 < i < M$.
4 Determinar la duración máxima $N_{max} = \max_i N_i,\ 1 < i < M$
5 Interpolar linealmente todos los ciclos de manera tal que todos tengan N_{max} muestras.
6 Armar una matriz $\mathbf{B} \in \mathbb{R}^{M \times N_{max}}$, cuyas filas se corresponden con los ciclos extraídos e interpolados.
7 Calcular la matriz de covarianza $S = \mathbf{B}^T\mathbf{B}$ y estimar sus autovalores λ_k.
8 Calcular la varianza normalizada de la componente principal como en la Ecuación (5.15)

Desvío Estándar de la Varianza Normalizada de la Componente Principal (DSVNCP)

Para calcular el valor de DSVNCP, primero debe dividirse la señal mediante ventanas de 100 ms sin superposición, de manera tal que para señales con frecuencia fundamental muy baja (como 50 Hz) haya un número de ciclos razonable en una ventana. Luego, se calcula el valor de VNCP para cada una de los trozos de la señal, utilizando el Algoritmo 2, y se computa el desvío estándar de los valores obtenidos para cada ventana. Llamando vncp_j al valor de VNCP para la i-ésima ventana, el valor de DSVNCP se calcula como:

$$\text{DSVNPC} = \sqrt{\frac{1}{\mathcal{M}-1}\sum_{i=1}^{\mathcal{M}}\left(\mathrm{vncp}_j - \overline{\mathrm{vncp}}\right)^2}, \tag{5.16}$$

donde $\mathcal{M}$ es el número de ventanas de la señal, y $\overline{\mathrm{vncp}}$ es el promedio de los valores de VNCP para todas las ventanas. DSVNCP se comporta de manera opuesta a VNCP, siendo más alta para señales tipo 3 y sucesivamente menor para los tipos 2 y 1.

5.4. Metodología

Los experimentos realizados para este trabajo se dividieron en tres partes. En primer lugar se estudiaron las características propuestas, realizando una selección de los descriptores y luego una evaluación estadística de su capacidad discriminativa. Para esta primera parte se utilizó un conjunto pequeño ($10\,\%$) del total de las señales disponibles (ambas bases de datos) no utilizado luego para el entrenamiento y validación. En segundo lugar se llevaron a cabo los experimentos de clasificación de voces mediante un clasificador automático y en base a las características seleccionadas en la primera parte del trabajo. Finalmente se abordó la evaluación de la probabilidad a posteriori como un indicador de la confianza en la clasificación automática.

Para reducir la cantidad de características utilizadas y al mismo tiempo seleccionar aquellas más relevantes para la clasificación, se utilizaron dos estrategias: SCVMC [140] y selección secuencial de características *hacia adelante* [141] (ver Secciones 4.2.1 y 4.2.2). El resultado de aplicar SCVMC es un vector de pesos que pondera las características, desde la más relevante hacia la menos útil para la clasificación. Esta información fue utilizada posteriormente para elegir la característica inicial de la selección secuencial.

Pesos SCVMC			
Carcterística	Peso	Característica	Peso
CPP	1.62	Shimmer	2.24×10^{-29}
DSVNCP	1.25	$K_2(4, 25)$	2.15×10^{-31}
VNCP	1.06	$K_2(4, 50)$	1.12×10^{-31}
HNR	0.89	$D_2(6, 80)$	2.85×10^{-32}
$D_2(10, 50)$	1.13×10^{-13}	$K_2(6, 25)$	3.32×10^{-34}
Nivel de Ruido	2.94×10^{-17}	$K_2(4, 110)$	4.46×10^{-36}
$D_2(10, 25)$	2.53×10^{-20}	$K_2(4, 80)$	6.91×10^{-37}
$D_2(8, 50)$	7.36×10^{-28}	$D_2(6, 110)$	1.74×10^{-38}
$D_2(6, 25)$	1.30×10^{-28}	$D_2(4, 50)$	2.99×10^{-39}
$D_2(8, 80)$	5.02×10^{-29}	$D_2(12, 25)$	2.15×10^{-39}

Tabla 5.3: Resultados de la selección de características por vecinos más cercanos. Las características se ordenaron de mayor a menor peso en el vector de ponderación obtenido mediante dicha técnica de selección. Esta tabla muestra las primeras 20 características y los pesos asignados.

Luego de la reducción de características, se estudió si existían diferencias estadísticamente significativas entre los valores que toman los descriptores seleccionados, para los distintos tipos de voces. Con ese fin, se utilizó el test de Kruskal-Wallis [142], equivalente no paramétrico del test ANOVA de una vía, cuya hipótesis nula es que todas las muestras provienen de una misma distribución. Si se falla en rechazar la hipótesis nula, entonces no hay evidencia de que existan diferencias estadísticamente significativas entre los valores que toma una característica para cada tipo de voz.

Por otro lado, si se rechaza dicha hipótesis, el test no permite conocer a qué clase corresponden las distribuciones que son distintas. Por esa razón, debe complementarse con *tests de comparaciones múltiples*, o *post hoc*, que permitan encontrar las clases entre las cuales se presentan diferencias estadísticamente significativas, de a un par a la vez (tipo 1 vs. tipo 2, tipo 1 vs. tipo 3 y tipo 2 vs. tipo 3). Para esta evaluación, se utilizó el test corregido de Bonferroni [142].

Con respecto a la segunda parte de los experimentos, el modelo empleado para la clasificación fue una SVM de kernel lineal (ver Sección 4.3). Los experimentos llevados a cabo incluyeron el entrenamiento y validación con la misma base datos (intra-corpus); y entrenamiento con un conjunto de señales y validación en el otro (inter-corpus). Todos los resultados se obtuvieron mediante validación cruzada de 10 iteraciones.

5.5. Resultados

5.5.1. Selección de características

El número total de características consideradas fue de 55: jitter$_{\%}$, shimmer$_{\%}$, HNR, CPP, VNPC, SDVNCP, Nivel de Ruido y 48 descriptores correspondientes a los parámetros de dinámicas no lineales para cada combinación de dimension y retardo de inmersión.

La Tabla 5.3 muestra las primeras 20 características ordenadas según los pesos asignados por SCVMC. Este vector de pesos se empleó para determinar la característica con mayor poder de discriminación individual, un dato de importancia para inicializar la selección de características hacia adelante. Este último, tal como se describió en la Sección 4.2.2, es un método voraz para encontrar una *combinación* de características que sea efectiva para la clasificación, comenzando

	CPP		DSVNCP		VNCP		HNR	
	MEEI	SVD	MEEI	SVD	MEEI	SVD	MEEI	SVD
K-W	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
1 vs 2	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
1 vs 3	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
2 vs 3	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
	Nivel de Ruido		$D_2(10, 25)$		Shimmer		$K_2(6, 25)$	
	MEEI	SVD	MEEI	SVD	MEEI	SVD	MEEI	SVD
K-W	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
1 vs 2	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P = .118$	$P < .001$
1 vs 3	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$
2 vs 3	$P < .001$	$P < .001$	$P = .7895$	$P < .001$	$P < .001$	$P < .001$	$P < .001$	$P < .001$

Tabla 5.4: Resultados del test de Kruskal-Wallis (K-W) y las comparaciones múltiples.

por la característica más relevante según algún método de ponderación, que en este caso es CPP según el vector de pesos obtenido a partir de SCVMC.

Como resultado de la selección secuencial hacia adelante, una combinación de 8 características fue seleccionada: CPP, DSVNPC, VNCP, HNR, Nivel de ruido, $D_2(10, 25)$, Shimmer y $K_2(6, 25)$. Se determinó que este era ún número suficiente de características ya que agregar más variables generó únicamente ganancias marginales en el desempeño (medido por la exactitud de la clasificación).

La diferencia entre las características elegidas y el órden de la Tabla 5.3 podría deberse a que la capacidad de discriminación de una combinación de características puede ser superior al de una característica individual. Adicionalmente, es importante recordar que el método de selección secuencial no asegura que la combinación de características encontradas sea óptima para la clasificación, en consecuencia sería posible que otro conjunto de descriptores resultara en un mejor desempeño. No obstante, no fue posible obtener mejores resultados al explorar otras combinaciones de características elegidas a voluntad. En consecuencia, se utilizará el conjunto de ocho características encontrado de aquí en adelante. Previo a su utilización, las características obtenidas fueron normalizadas substrayendo su media y haciendo su varianza unitaria.

5.5.2. Análisis estadístico de los descriptores

Se realizó un análisis estadístico de las características para evaluar su capacidad de discriminación entre los distintos tipos de señal. Dado que la distribución de valores no es normal para ninguna de las características, se utilizó un test no paramétrico de Kruskal-Wallis con significancia del 0.005. En caso de que este test fuera significativo, se emplearon tests de comparaciones múltiples (Bonferroni) para cada par de clases. La Tabla 5.4 muestra los resultados de este análisis en términos del valor P para cada test. De allí puede verse que los valores que toman casi todas las características poseen diferencias estadísticamente significativas para cada tipo de señal. Las únicas excepciones a esto son $D_2(10, 25)$ ($P = 0.7895$ para los tipos 2 y 3) y $K_2(6, 25)$ ($P = 0.118$ para los tipos 1 y 2), considerando la base de datos MEEI.

Con el propósito de evitar redundancia de información, las características no deberían estar fuertemente correlacionadas entre sí. Esto puede estudiarse al observar las matrices de correlación de las Tablas 5.5 y 5.6. Es posible ver que casi todos los pares de características muestran una baja correlación, exceptuando HNR con CPP, DSVNCP, VNCP y Nivel de Ruido (para MEEI); y HNR y CPP, Nivel de Ruido y Shimmer (para SVD). En todos estos casos la correlación entre

Matriz de correlación de características (para MEEI)								
	CPP	DSVNCP	VNCP	HNR	Nivel de ruido	D_2 (10,25)	Shimm.	K_2 (6,25)
CPP	1.00	-0.64	0.57	0.76	-0.55	-0.17	-0.37	-0.18
DSVNCP		1.00	-0.70	-0.78	0.63	0.19	0.45	0.23
VNCP			1.00	0.83	-0.81	-0.17	-0.74	-0.23
HNR				1.00	-0.78	-0.21	-0.56	-0.29
Niv. de R.					1.00	0.18	0.62	0.28
$D_2(10,25)$						1.00	0.03	0.10
Shimm.							1.00	0.11
$K_2(6,25)$								1.00

Tabla 5.5: Matriz de correlación de las características para MEEI.

Matriz de correlación de características (para SVD)								
	CPP	DSVNCP	VNCP	HNR	Nivel de ruido	D2 (10,25)	Shimm.	K2 (6,25)
CPP	1.00	-0.65	0.48	0.78	-0.54	-0.26	-0.53	-0.29
DSVNCP		1.00	-0.53	-0.74	0.57	0.30	0.56	0.27
VNCP			1.00	0.68	-0.70	-0.19	-0.64	-0.19
HNR				1.00	-0.79	-0.35	-0.75	-0.35
Niv. de R.					1.00	0.26	0.70	0.29
$D_2(10,25)$						1.00	0.25	0.11
Shimm.							1.00	0.18
$K_2(6,25)$								1.00

Tabla 5.6: Matriz de correlación de las características para SVD.

Clasificador	Exactitud
SVM *(kernel lineal)*	**82.96 % (1.87 %)**
Discriminante Lineal	82.81 % (2.37 %)
SVM *(kernel cuadrático)*	82.25 % (2.67 %)
KNN ($K = 17$)	82.09 % (2.30 %)
SVM *(kernel Gaussiano)*	81.85 % (3.42 %)
SVM *(kernel cúbico)*	79.87 % (3.10 %)
Árbol de clasificación	79.17 % (4.10 %)
Discriminante Cuadrático	73.69 % (5.34 %)

Tabla 5.7: Exactitud estimada para distintos clasificadores, expresada como media % (desvío estándar) %.

las características fue mayor a 0.75 en valor absoluto. No obstante, en general, las características seleccionadas parecen estar poco correlacionadas.

5.5.3. Clasificación

Como se mencionó anteriormente, la clasificación se realizó utilizando una SVM de kernel lineal. La selección de este clasificador se realizó en base a pruebas con varios algoritmos de clasificación en las que SVM con dicho kernel obtuvo el mejor desempeño, como puede verse en la Tabla 5.7. La estrategia para lidiar con este problema de tres clases fue del tipo *uno contra uno* (ver Sección 4.3.5), elegida por resultar en un mejor desempeño (coincidiendo con reportes previos [153]). Todos los resultados de aquí en más fueron obtenidos mediante validación cruzada de 10 iteraciones. Dado que el problema presenta clases desbalanceadas, se emplearon costos de clasificación proporcionales a la prevalencia de cada clase normalizada por la prevalencia de la clase mayoritaria (tipo 2).

Las Tablas 5.8 y 5.9 muestran las matrices de confusión del clasificador para experimentos intra-corpus (entrenando y validando sobre el mismo corpus) e inter-corpus (entrenando y validando en diferentes corpus), tanto para MEEI como para SVD. Puede verse en la Tabla 5.8 que la diferencia entre la exactitud general del algoritmo para el caso intra-corpus e inter-corpus es despreciable ($87.06\,\%$ y $86.53\,\%$ respectivamente) para MEEI. Lo mismo puede decirse en el caso de SVD ($83.36\,\%$ y $82.71\,\%$ para los mismos experimentos). A pesar de estas diferencias son poco significativas entre los distintos experimentos para ambos conjuntos de señales, debe observarse que, sin embargo, existen diferencias apreciables entre los porcentajes de clasificación correcta para los tipos 1 y 2. Por ejemplo, considerando la Tabla 5.8, este porcentaje es más alto para las voces tipo 1 que para las voces tipo 2 en el experimento intra-corpus ($90.84\,\%$ y $82.93\,\%$ para tipo 1 y tipo 2 respectivamente), que en el experimento inter-corpus ($81.95\,\%$ y $87.28\,\%$). Asimismo, al comparar las Tablas 5.8 y 5.9 puede observarse que el desempeño obtenido con MEEI es superior al obtenido con SVD.

La Tabla 5.10 muestra el desempeño considerando un único conjunto de datos obtenido al unir los dos corpus estudiados (MEEI + SVD). La exactitud general en este caso es del $82.96\,\%$, aunque el porcentaje de voces tipo 2 clasificadas correctamente es significativamente menor, alcanzando un $77.9\,\%$. Es de notar que el porcentaje de voces voces tipo 3 correctamente clasificadas supera el $90\,\%$ en todos los experimentos y en todos los conjuntos de datos. Adicionalmente, no existe confusión entre los tipos 1 y 3 en un ningún experimento.

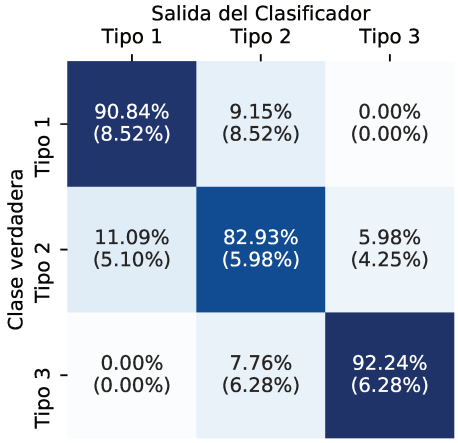
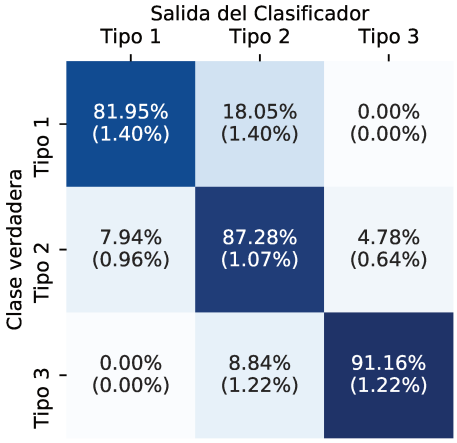
Validación: MEEI	
Entrenamiento: MEEI	Entrenamiento: SVD
Exactitud: 87.06 % (4.58 %)	Exactitud: 86.53 % (0.60 %)
	

Tabla 5.8: Matrices de confusión y exactitudes para cada experimento, calculadas mediante validación cruzada de 10 iteraciones, para MEEI como conjunto de validación. Los valores están dados como *media % (desvío estándar %)*.

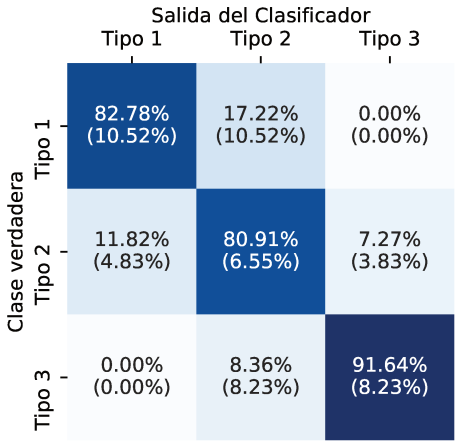
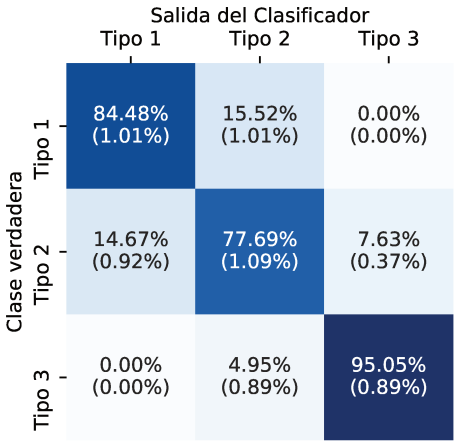
Validación: SVD	
Entrenamiento: SVD	Entrenamiento: MEEI
Exactitud: 83.36 % (4.49 %)	Exactitud: 82.71 % (0.43 %)
	

Tabla 5.9: Matrices de confusión y exactitudes para cada experimento, calculadas mediante validación cruzada de 10 iteraciones, para SVD como conjunto de validación. Los valores están dados como *media % (desvío estándar %)*.

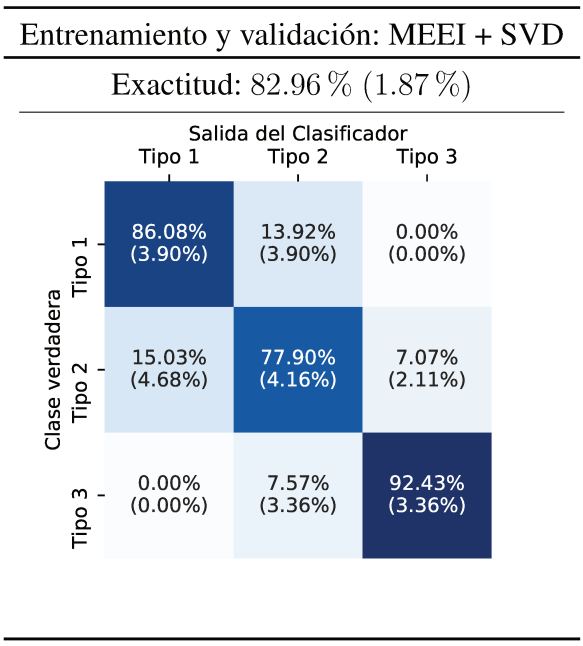


Tabla 5.10: Matriz de confusión y exactitud, calculadas mediante validación cruzada de 10 iteraciones, para el conjunto de señales completo. Los valores están dados como *media* % *(desvío estándar* %*)*.

5.5.4. La probabilidad a posteriori como medida de confianza

Dado que no existe un sistema perfecto para la tipificación automática de señales, antes de emplear un software automático en un contexto clínico debe considerarse la posibilidad de que el tipo asignado por un sistema como el propuesto sea erróneo, así como también acciones a tomarse para prevenir la clasificación incorrecta. Teniendo esto en cuenta, se propuso el uso de las probabilidades a posteriori del clasificador como medidas de la *confianza* en la clase asignada de forma automática.

Las probabilidades a posteriori (p.p.) *por clase* son las probabilidades de que una señal sea de una determinada clase una vez conocidos todos sus parámetros. Si bien las máquinas de vectores de soporte no estiman estas probabilidades para la clasificación, existe la posibilidad de aproximarlas como se explicó en la Sección 4.3.4 [151]. La clase con la p.p. más alta se corresponde con el tipo finalmente asignado.

Un valor umbral podría hacer el uso de las probabilidades a posteriori más sencillo y aplicable en un contexto real. Así, si el valor de la p.p. supera un umbral, el tipo asignado puede ser aceptado con confianza. En caso contrario, si el valor está por debajo del umbral, el especialista o usuario del sistema podría optar por revisar la clasificación, y cambiarla si así lo desea.

Utilizando el modelo entrenado con ambos conjuntos de datos, descripto en la Tabla 5.10, se guardaron las probabilidades a posteriori más altas de cada señal, es decir las correspondientes a la clase asignada, en cada repetición del experimento de validación cruzada. Estas probabilidades, tanto para las señales bien clasificadas por un lado, y para las mal clasificadas por el otro, se consideraron variables aleatorias. De esta forma, se estimaron dos densidades de probabilidad: 1) p.p. de las señales bien clasificadas, 2) p.p. de las señales mal clasificadas. La Figura 5.3 muestra ambas densidades y el umbral calculado $p_{thr} = 0.7990$. Puede observarse que la moda de la densidad de la probabilidad a posteriori de las señales bien clasificadas se ubica más cerca de 1, mientras que aquella para las señales mal clasificadas se ubica aproximadamente en 0.6. El valor de p_{thr} fue calculado de manera tal de minimizar la probabilidad de aceptar la clasificación de una señal mal clasificada con una p.p. por arriba del umbral, y maximizar dicha probabilidad para

señales clasificadas correctamente (ver Apéndice A.1).

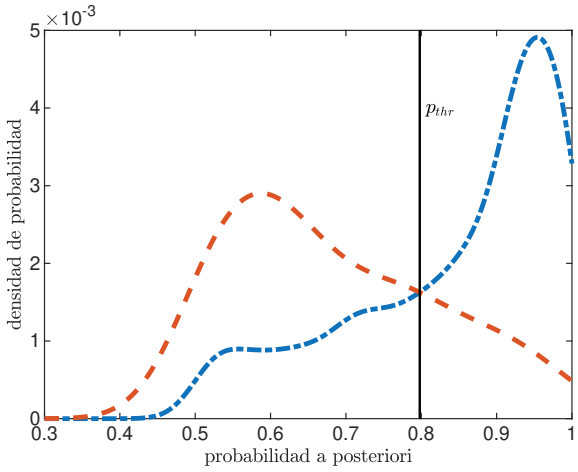


Figura 5.3: Densidades de probabilidad de la probabilidad a posteriori para señales mal clasificadas (línea de trazos) y correctamente clasificadas (línea de trazos con puntos).

5.6. Discusión

Los resultados reportados muestran que el enfoque propuesto para distinguir entre los tres tipos de señales supera el estado del arte, considerando a [61] como referencia, utilizando un clasificador lineal multiclase y características de uso difundido en el área de cuidado de la voz. El uso de máquinas de vectores de soporte lineales puede ser ventajoso para la interpretación de cómo los descriptores son empleados para clasificar las señales, lo que constituye una propiedad útil en el contexto de problemas biomédicos. Asimismo, utilizar parámetros conocidos dentro de la comunidad de especialistas clínicos de la voz puede hacer la propuesta más atractiva para su uso en la práctica diaria, así como también la utilización de las probabilidades a posteriori como una medida de la robustez del tipo asignado automáticamente.

En contraste con trabajos previos [59, 60, 62], las medidas de dinámicas no lineales utilizadas en este trabajo fueron calculadas con un método independiente del usuario [157] y los parámetros de inmersión fueron elegidos para maximizar el desempeño del clasificador utilizando un método automático de selección de características.

Las características nuevas propuestas para este trabajo (VNCP y DSVNCP), estuvieron entre aquellas con mayor ponderación por el método de SCVMC, lo que significa que ambos descriptores son de hecho relevantes para el problema de clasificación. Asimismo, diferencias estadísticamente significativas fueron encontradas entre los valores que estas características toman para los distintos tipos de señales. Estos parámetros fueron propuestos como una forma de describir la regularidad de la forma de onda de una señal, siendo sensibles a los cambios en la morfología de señales aproximadamente periódicas. Un script de PRAAT para calcular VNCP y DSVNCP se encuentra disponible para fonoaudiólogos y profesionales del cuidado de la voz [1].

Los resultados reportados se estimaron sobre un conjunto de señales más grande que todos los trabajos previos [60–62]. Esto es relevante porque permite encontrar una estimación más exacta del desempeño del enfoque propuesto al evaluarlo en condiciones mucho más cercanas a una aplicación en el mundo real. A pesar de que todas las características mostraron diferencias estadísticas significativas para casi todos los tipos de señal y conjuntos de datos, como muchos otros parámetros descriptos con anterioridad [60–62], los resultados revelan que el desempeño de un sistema

[1] https://github.com/jmiramont/PRAAT-scripts

entrenado con un corpus no puede ser directamente extrapolado a otros conjuntos de señales. Por ejemplo, la Tabla 5.8 muestra que el desempeño es menor cuando el sistema es entrenado con un conjunto de datos y luego validado en otro. Adicionalmente, la exactitud general fue sistemáticamente más baja para el conjunto SVD que para MEEI. Esto conduce a la conclusión de que, al menos para el problema de tipificación de voces, se debe ser cuidadosos al hacer suposiciones sobre la habilidad para generalizar de un sistema o parámetro cuando sólo se utiliza un único conjunto de señales reducido, como sucede en numerosos trabajos previos [59–62] . En el futuro, considerando la posibilidad de construir un sistema capaz de utilizarse en la práctica clínica, las nuevas características y algoritmos propuestos para la tipificación automática y objetiva de voces deben evaluarse sobre varios conjuntos, más grandes y diferentes entre sí. Asimismo, un defecto del que adolecen tanto la presente propuesta como todos los trabajos previos con la misma metodología es que, al utilizar sólo las señales para las que la clasificación de las/los especialistas coincide, se deja fuera de consideración la variación interprofesional. Esto implica que el desempeño real del algoritmo podría ser menor al estimado, ya que adopta el sesgo de las especialistas que clasificaron los datos en primer lugar.

Las Tablas 5.8, 5.9 y 5.10 permiten ver que las características utilizadas pueden discriminar correctamente las voces tipo 3 de los otros dos tipos con mayor facilidad, incluso en experimentos inter-corpus. Esto no es sorprendente ya que, en general, las voces tipo 3 se corresponden con señales altamente desorganizadas cuyos parámetros fácilmente se desvían de los valores que tomarían en presencia de algún tipo de periodicidad. Asimismo, ni una sola señal tipo 3 fue clasificada como tipo 1, lo que es deseable, ya que las voces tipo 3 no deberían ser analizadas con medidas de perturbación bajo ninguna circunstancia. En contraste con esta situación, distinguir entre los tipos 1 y 2 resulta una tarea más desafiante, lo que se refleja en el porcentaje de confusión entre ambos tipos. A futuro, se invertirá mayor esfuerzo en estudiar este problema particular.

Con respecto al uso de probabilidades a posteriori como una medida de la confiabilidad en la predicción automática, cabe mencionar que el umbral calculado fue estimado utilizando el modelo con todas las señales por considerarlo el modelo más general posible. El valor calculado se corresponde con un umbral *conservador* (es decir, un valor bastante alto), dado que el clasificador no es perfecto. En el futuro, sistemas con mejor desempeño que utilicen la probabilidad a posteriori como medida de la confiabilidad podrían tener umbrales más bajos, entendiendo que las densidades en la Figura 5.3 se superpondrían menos, y sus modas se encontrarían más alejadas, para un clasificador con un desempeño más cercano al perfecto. Esto reduciría el número de predicciones que deberían revisarse por el especialista, lo que constituiría un paso más hacia una herramienta automática para la tipificación de voces de uso clínico.

5.7. Conclusión

Lo presentado en este capítulo constituye un esfuerzo para la clasificación automática de señales en los tipos descriptos. El enfoque propuesto se basa en herramientas del aprendizaje maquinal para resolver un problema típico de reconocimiento de patrones, empleando descriptores conocidos y un clasificador lineal. La utilización de un número de señales mayor a otros trabajos, y los experimentos intra e inter-corpus proveen resultados más cerca de la aplicación en el ámbito clínico, y marcan un precedente para futuras investigaciones. Los resultados expuestos indican que las características propuestas pueden utilizarse junto a un clasificador lineal en forma efectiva para la discriminación de los tres tipos de señales. Adicionalmente, se propuso el uso de las probabilidades a posteriori como medida de confianza en la predicción del clasificador, con el fin de proporcionar al profesional de la salud vocal un indicador que le ayude a tomar una decisión a la hora de aceptar o rechazar el tipo asignado automáticamente por el sistema. Estos resultados podrían allanar el camino hacia una herramienta automática de clasificación que reduzca la subjetividad de la tipifi-

cación, y que pueda utilizarse en la práctica clínica en el futuro. Los resultados presentados en este capítulo han sido publicados en [161].

Capítulo 6

Estimación robusta de jitter relativo

6.1. Introducción

La producción de la voz puede considerarse como el resultado de una serie de osciladores acoplados de naturaleza biomecánica, neurales y acústicos, cuyo sistema oscilante principal son los pliegues vocales *verdaderos* [5, 7, 42]. Para fonemas vocales, como se describió en capítulos anteriores, la interacción entre el flujo de aire proveniente de los pulmones y los pliegues vocales produce una onda pulsátil conocida como flujo glótico, que es modulado por sucesivos osciladores pasivos y activos [42]. El promedio del tiempo entre pulsos adyacentes del flujo glótico es denominado periodo fundamental promedio $\bar{T}_0$, mientras que su recíproco es definido como la frecuencia fundamental promedio $\bar{F}_0$. Estas definiciones están basadas en la suposición, bastante utilizada, de que los cambios ocurridos en un ciclo, a lo largo del segmento analizado, son despreciables. Y es debido a esto que la señal es, entonces, considerada estacionaria [16]. No obstante, sin importar qué tanto esfuerzo haga un sujeto para producir una alocución perfectamente periódica, pequeñas desviaciones en amplitud, frecuencia y forma de onda, denominadas *perturbaciones*, siempre están presentes. Esto hace necesario la definición de versiones instantáneas del periodo y la frecuencia fundamental: $T_0(t)$ y $F_0(t)$, respectivamente [41, 42].

El jitter vocal ha sido definido como una perturbación aleatoria de la longitud del ciclo glótico [95]. El nivel de jitter puede aumentar considerablemente en el caso de algunas enfermedades laríngeas, de allí la importancia que tiene la cuantificación del jitter para los especialistas en patologías vocales. De entre las varias formas que se han propuesto para estimar el jitter (ver Sección 2.5.2 y [5, 71]), se pondrá el foco en este capítulo en el jitter *relativo*, dado que es una de las medidas más difundidas de la perturbación del periodo [67–69]. Como se expuso en el Capítulo 2, el jitter relativo se define formalmente como:

$$jitter_{\%} = \frac{\frac{1}{M-1}\sum_{j=1}^{M-1}|T_{j+1} - T_j|}{\frac{1}{M}\sum_{j=1}^{M} T_j} \times 100\,\%, \tag{6.1}$$

donde T_j es la duración del j-ésimo periodo (en segundos), M la cantidad de periodos y $\{T_j\}_{j=1}^{M}$ la serie o secuencia de periodos. La Ecuación (6.1) es simplemente el promedio del valor absoluto de las perturbaciones normalizada por la duración promedio de los ciclos $\bar{T}_0$.

Los métodos más comunes para estimar la secuencia de periodos son: detección de picos, cruces por cero y coincidencia de forma de onda (ver Sección 2.5 o [41, 71, 93]). Todos ellos basados en calcular la diferencia de tiempo entre puntos fiduciarios que indican el principio y el fin de cada ciclo de la señal. La hipótesis principal detrás de este procedimiento es que $F_0(t)$ se mantiene constante a lo largo de la duración de cada periodo [5, 41, 68]. Algunas desventajas de estos métodos son:

1. Fallan al estimar valores de jitter altos.

2. Pueden verse afectados por el ruido presente en la señal.

Debido a la primera de estas desventajas se ha planteado el debate acerca de la validez del jitter relativo para valores mayores al $5\,\%$ y en presencia de ruido en la señal [68, 72]. Sin embargo, trabajos más recientes muestran que algunas herramientas para la estimación de jitter pueden ser confiables para valores de hasta $15\,\%$ [67–69, 185].

6.2. Jitter relativo y variación total de la frecuencia fundamental

En esta sección se describirá un nuevo método para la medición de jitter relativo que utiliza la variación total (VT) de una aproximación del periodo instantáneo en lugar de la serie de periodos (equivalente a un muestreo no uniforme de $T_0(t)$). Se explicará cómo $F_0(t)$ y $F_0'(t)$ estimadas a partir de los operadores de *synchrosqueezing* de orden superior (FSSTN, ver Sección 3.5), pueden emplearse para la medición del jitter vocal. También se darán algunos detalles del método, como el orden de *synchrosqueezing* utilizado en esta aplicación y el ancho de la ventana de análisis $g(t) = \frac{1}{\sigma} e^{-\frac{\pi}{\sigma^2} t^2}$, determinado por el valor del parámetro σ. Finalmente, se presentará el algoritmo completo que resume todos los pasos para la estimación del jitter vocal.

6.2.1. Jitter relativo como la variación total de una estimación local de la FI

En primer lugar, se definirá la variación total de una secuencia y de una función real de tiempo continuo. Para una secuencia $\{T_j\}_{j=1}^{M}$, la variación total está definida como

$$\mathrm{VT}\left[T_j\right]_1^M = \sum_{j=1}^{M-1} \left|T_{j+1} - T_j\right|, \tag{6.2}$$

mientras que la variación total para una función real $z(t)$ para $a \leq t \leq b$ se define como:

$$\mathrm{VT}\left[z(t)\right]_a^b = \int_a^b \left|\frac{d}{dt}\left[z(t)\right]\right| dt. \tag{6.3}$$

Teniendo en cuenta la Ecuación (6.2), podemos reescribir la Ecuación (6.1) como

$$\begin{aligned} jitter_{\,\%} &= \frac{\mathrm{VT}\left[T_j\right]_1^M}{(M-1)\bar{T}_0} \times 100\,\% \\ jitter_{\,\%} &= \frac{1}{M\bar{T}_0 - \bar{T}_0} \mathrm{VT}\left[T_j\right]_1^M \times 100\,\%, \\ jitter_{\,\%} &\approx \frac{1}{L - \bar{T}_0} \mathrm{VT}\left[T_j\right]_1^M \times 100\,\%, \end{aligned} \tag{6.4}$$

donde $M\bar{T}_0$ es aproximadamente la duración de la señal L en segundos y $\mathrm{VT}\left[T_j\right]_1^M$ es la variación total de la secuencia de periodos. Para ver cómo esta nueva expresión puede conducir a una nueva forma de estimar el jitter, consideremos la siguiente proposición (una demostración de la misma se ofrece en el Apéndice A.2):

Proposición 6.2.1. *Sea $x(t) = \cos(2\pi\phi(t))$ un chirp lineal real, con $\phi(t) = \alpha t^2 + \beta t$ y $\alpha, \beta \in \mathbb{R}$. Entonces su escala local $s(t) = 1/\phi'(t)$ evaluada en los máximos locales t_ℓ satisface la siguiente expresión:*

$$s(t_{\ell+1}) < t_{\ell+1} - t_\ell < s(t_\ell). \tag{6.5}$$

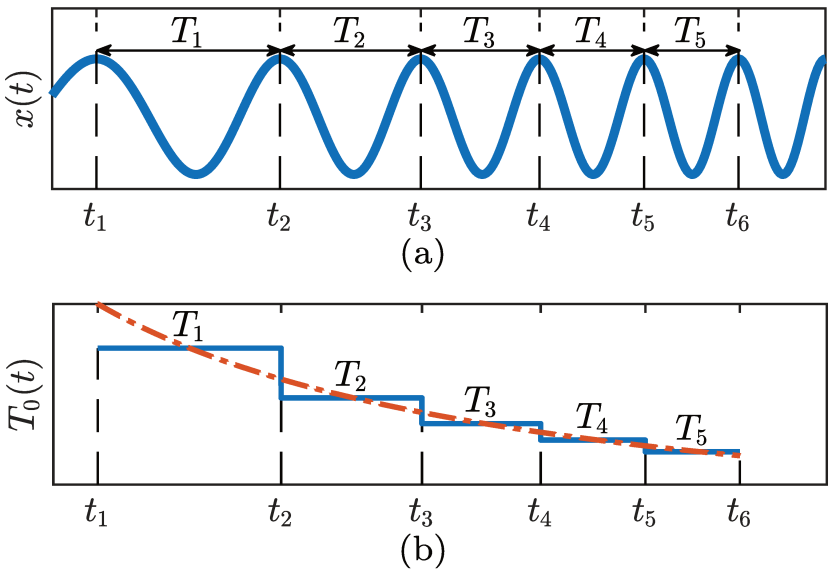


Figura 6.1: (a) Un *chirp* lineal y su segmentación en ciclos, donde T_j, $j = 1, 2, 3, 4, 5$, es la duración del j-ésimo ciclo. (b) Se muestra una aproximación constante a trozos de $T_0(t)$, en la que la altura de cada escalón es igual a su duración. El recíproco de la verdadera $F_0(t)$ se muestra como referencia (con linea de trazos y puntos, en color naranja).

Para un *chirp* lineal con $\alpha > 0$, como el representado en la Figura 6.1a, la escala local $s(t)$ es el recíproco de la FI, $\phi'(t)$, y es una función monótona decreciente. Lo que la Proposición 6.2.1 indica es que, para un *chirp* lineal, el valor de la distancia entre dos máximos sucesivos se encontrará entre los valores que la *verdadera* escala local tome en esos puntos, tal como puede verse en la Figura 6.1b. Esto significa que podemos construir una estimación local de $T_0(t)$ que sea constante a trozos, denotada como $\widetilde{T}_0(t)$, utilizando la diferencia entre los instantes de tiempo en los que la señal alcanza un máximo [125, 170]:

$$\widetilde{T}_0(t) = t_{j+1} - t_j \text{ , para } t_j < t < t_{j+1} \tag{6.6}$$

donde $\{t_j\}_{j=1}^{M}$ es, a la vez, la secuencia de puntos fiduciarios que coinciden con el principio y el final de los ciclos de la señal.

A pesar de que la Proposición 6.2.1 es válida para un *chirp* con fase cuadrática, es posible aproximar *localmente* cualquier *chirp* mediante uno lineal, siempre que $|\phi^{(3)}(t)|$ sea pequeña [170]. En consecuencia, es posible ver que la serie de periodos de una señal aproximadamente periódica constituye una aproximación constante a trozos de $T_0(t)$.

Tomando el valor absoluto de la derivada (generalizada) de $\widetilde{T}_0(t)$, se encuentra una serie de impulsos localizados donde esta función escalonada no es continua [171]. Utilizando esto para calcular la variación total de $\widetilde{T}_0(t)$, según la Ecuación (6.3), se obtiene el siguiente resultado:

$$\mathrm{VT}\left[\widetilde{T}_0(t)\right]_{t_1}^{t_M} = \int_{t_1}^{t_M} \left| \frac{d}{dt}\left[\widetilde{T}_0(t)\right] \right| dt \tag{6.7}$$

$$= \int_{t_1}^{t_M} \sum_{i=1}^{M-1} |T_{j+1} - T_j|\, \delta(t - t_{j+1}) dt \tag{6.8}$$

$$= \sum_{j=1}^{M-1} |T_{j+1} - T_j| \tag{6.9}$$

$$= \mathrm{VT}\left[T_j\right]_1^M . \tag{6.10}$$

Por lo tanto, la variación total de la aproximación constante a trozos de $T_0(t)$ es exactamente igual a la variación total de la secuencia de periodos, para el caso de un *chirp* lineal, permitiendo

el reemplazo de $\mathrm{VT}\,[T_j]_1^M$ por $\mathrm{VT}[\widetilde{T}_0(t)]_0^L$ en la Ecuación (6.4), lo que conduce a una nueva interpretación de la Ecuación (6.1) como la variación total de una *aproximación* de tiempo continuo (ya no discreto, como la secuencia de periodos) del periodo fundamental instantáneo normalizada por un factor $(L - \widetilde{T}_0)^{-1}$.

La hipótesis central detrás del nuevo método aquí propuesto es que la aproximación grosera utilizada en el método tradicional, escalonada y de *orden cero*, es la causa de sus principales desventajas: la dependencia de $\bar{F}_0$ y el fallo en la estimación de grandes cantidades de jitter [41, 68, 72]. En consecuencia, una aproximación local, de orden superior podría mejorar la estimación al poder reflejar cambios más rápidos en la frecuencia fundamental instantánea. Al mismo tiempo, se evitaría la necesidad de determinar los puntos fiduciarios si es posible descartar el uso de la secuencia de periodos para encontrar esta nueva aproximación. Considerando estos elementos, se propone la siguiente definición del *jitter de variación total*:

$$\begin{aligned}
\mathrm{jitter}_{\%}^{VT} &= \frac{1}{L - \bar{T}_0}\,\mathrm{VT}\,[T_0(t)]_0^L \times 100\,\%, \\
&= \frac{1}{L - \bar{T}_0}\int_0^L \left|\frac{d}{dt}\,[T_0(t)]\right|\,dt \times 100\,\% \\
&= \frac{1}{L - \bar{T}_0}\int_0^L \left|\frac{d}{dt}\left[\frac{1}{F_0(t)}\right]\right|\,dt \times 100\,\% \\
&= \frac{1}{L - \bar{T}_0}\int_0^L \left|\frac{F_0'(t)}{[F_0(t)]^2}\right|\,dt \times 100\,\%,
\end{aligned} \tag{6.11}$$

para la que se necesitan estimaciones de $\bar{T}_0$, $F_0(t)$ y $F_0'(t)$.

Estimación de $\bar{T}_0$

$\bar{T}_0$ es determinado mediante el cálculo de la mediana del recíproco de una estimación sencilla de la FI. Con este objetivo, la cresta del primer modo es detectada a partir del espectrograma de la señal. Esto se realiza mediante un algoritmo voraz que maximiza la energía sobre la cresta utilizando un enfoque *hacia adelante y hacia atrás* con iniciaciones aleatorias, al mismo tiempo que mantiene la suavidad de la solución, descrito en la Sección 3.4 y en [127, 128]. Como resultado, se obtiene la cresta $r(t)$, a partir de la cual se calcula $\bar{T}_0$ como:

$$\bar{T}_0 = \mathrm{mediana}\,(1/r(t))\,. \tag{6.12}$$

Es importante notar que el ancho de la ventana de análisis empleada para el cálculo del espectrograma no es particularmente importante en este paso, siempre que permita la detección del primer modo de manera adecuada.

Estimación de $F_0(t)$ y $F_0'(t)$

La estimación de $F_0(t)$ y $F_0'(t)$ se realizará a partir de los operadores de FSSTN, tal como se describió en la Sección 3.6. Sin embargo, algunas pruebas preliminares para estimar $F_0(t)$ y $F_0'(t)$ mostraron que el valor de σ tiene un impacto en las aproximaciones obtenidas a partir de $\tilde{\omega}_x^{[N]}(t, f)$ y $\tilde{q}_x^{[2,N]}(t, f)$ respectivamente. Por un lado, una ventana más angosta favorece la suposición de que la función de fase es localmente aproximable por un polinomio. Por lo tanto, para una señal con una fase arbitraria, es posible obtener una buena aproximación de dicha función siempre que la ventana sea lo suficientemente angosta. Pero, por otro lado, esto significa que los dominios de los modos en el plano tiempo-frecuencia se volverán más anchos, incrementando la interferencia entre modos adyacentes, lo que dificulta la extracción de la cresta y contradice las hipótesis del modelo de señal multicomponente [75].

Para reducir este efecto, se propuso un procedimiento de dos pasos para estimar la FI y su derivada. En primer lugar, el primer modo es extraído utilizando una ventana ancha y luego se sintetiza una versión de un sólo modo de $x(t)$, denotada aquí como $\widetilde{x}(t)$. Esta señal simplificada aún posee toda la información necesaria para estimar el valor de jitter, aunque no toda la influencia de los modos de frecuencias más altas pueden eliminarse dado que la ventana $g(t)$ no es de soporte compacto. En segundo lugar, $F_0(t)$ y $F_0'(t)$ de $\widetilde{x}(t)$ (no de $x(t)$) son estimadas utilizando una ventana más angosta.

Para extraer el primer modo, la TFTC de $x(t)$ es calculada utilizando

$$\sigma = \frac{6\bar{T}_0}{\sqrt{2\pi}} \approx 2.39\bar{T}_0,$$

donde $\bar{T}_0$ es previamente estimado utilizando la Ecuación (6.12). El valor de σ utilizado en este paso permite confinar el primer modo en una franja del plano tiempo-frecuencia cuyo ancho es aproximadamente $1/\bar{T}_0$ y que está centrada en la cresta correspondiente $r(t)$ (ver demostración en el Apéndice A.3). Luego, $\widetilde{x}(t)$ puede sintetizarse como:

$$\widetilde{x}(t) = \frac{1}{g(0)} \int_{\{f:|f-r(t)|<1/(2\bar{T}_0)\}} V_x^g(t,f)\, df. \tag{6.13}$$

utilizando la fórmula de extracción de modo de la Sección 3.4.

Una vez calculada $\widetilde{x}(t)$, $F_0(t)$ y $F_0'(t)$ son estimadas a partir $\tilde{\omega}_{\widetilde{x}}^{[N]}(t,f)$ y $\tilde{q}_{\widetilde{x}}^{[2,N]}(t,f)$, evaluando estos operadores en la cresta correspondiente al primer (y único) modo de $\widetilde{f}(t)$, denotada aquí como $\widetilde{r}(t)$ para evitar confundirla con la cresta equivalente de la señal original. $\widetilde{r}(t)$ es detectada de la misma manera que $r(t)$ para la estimación de $\bar{T}_0$, es decir, calculando el espectrograma de $\widetilde{x}(t)$ y aplicando un algoritmo de detección de crestas como el descripto en la Sección 3.4. Una vez calculados $\tilde{\omega}_{\widetilde{x}}^{[N]}(t,f)$ y $\tilde{q}_{\widetilde{x}}^{[2,N]}(t,f)$ utilizando las Ecuaciones (3.36) y (3.40) [172], la frecuencia instantánea y su derivada son determinadas como:

$$F_0(t) = \mathcal{R}e\left\{\tilde{\omega}_{\widetilde{x}}^{[N]}(t,\widetilde{r}(t))\right\} \tag{6.14}$$

y

$$F_0'(t) = \mathcal{R}e\left\{\tilde{q}_{\widetilde{x}}^{[2,N]}(t,\widetilde{r}(t))\right\}. \tag{6.15}$$

Para este último paso, el ancho de la ventana de análisis, proporcional a σ, y el orden N de los operadores de *synchrosqueezing* fueron determinados de forma empírica, minimizando el error entre el valor real de jitter y el jitter de variación total calculado utilizando la Ecuación (6.11). Este error fue calculado para varios valores de σ y para $N = 2, 3, 4$. El mejor desempeño fue encontrado para $N = 4$ y

$$\sigma = 3.46\frac{\bar{T}_0\sqrt{2\pi}}{6} \approx 1.45\bar{T}_0.$$

Los detalles de este experimento, y cómo se obtuvo el factor 3.46 de la ecuación anterior, se describen más adelante en la Sección 6.5.2. El Algoritmo 3 resume todos los pasos para la estimación del jitter relativo por el método de la variación total del periodo fundamental instantáneo utilizando operadores de FSST4 para la estimación de FI y su derivada.

Algoritmo 3: Estimación de jitter mediante la variación total del periodo fundamental instantáneo.

1 Calcular $V_x^g(t,f)$ utilizando la Ecuación (3.4) (empleando un valor arbitrario de σ).
2 Detectar la cresta $r(t)$ correspondiente al primer modo de $x(t)$ a partir de $|V_x^g(t,f)|^2$.
3 Estimar $\bar{T}_0$ a partir de $r(t)$ mediante: $\bar{T}_0 = \text{mediana}\left(\frac{1}{r(t)}\right)$.
4 Extraer el primer modo, usando $\sigma = \frac{6\bar{T}_0}{\sqrt{2\pi}}$, y sintetizando $\tilde{x}(t)$ mediante la Ecuación (6.13)
5 Calcular $V_{\tilde{x}}^g(t,f)$, $\tilde{\omega}_{\tilde{x}}^{[4]}(t,f)$ y $\tilde{q}_{\tilde{x}}^{[2,4]}(t,f)$, como en las Ecuaciones (3.4), (3.36) y (3.40) (ver también [172]), empleando $\sigma = 3.46\frac{\bar{T}_0\sqrt{2\pi}}{6}$ y $N = 4$.
6 Detectar la cresta $\tilde{r}(t)$ correspondiente al único modo de $\tilde{x}(t)$ a partir de $|V_{\tilde{x}}^g(t,f)|^2$.
7 Estimar $F_0(t) = \mathcal{R}e\left\{\tilde{\omega}_{\tilde{x}}^{[4]}(t,\tilde{r}(t))\right\}$ y $F_0'(t) = \mathcal{R}e\left\{\tilde{q}_{\tilde{x}}^{[2,4]}(t,\tilde{r}(t))\right\}$.
8 Calcular $\text{jitter}_{\%}^{VT}$ mediante la Ecuación (6.11).

6.2.2. Voces sintéticas

Con el propósito de evaluar la capacidad del algoritmo propuesto de estimar el jitter relativo, se emplearon voces artificiales con un valor de jitter conocido. Estas señales fueron sintetizadas por medio del clásico modelo fuente-filtro descripto anteriormente (ver Capítulo 2 y [15]):

$$x[k] = -\sum_{j=1}^{J} h[j]x[k-j] + u[k], \tag{6.16}$$

donde los coeficientes $h[j]$ fueron extraídos mediante predicción lineal [15, 85] a partir de una voz real de un sujeto con voz sana[1]. La frecuencia de muestreo utilizada para la síntesis fue de 50 kHz, la duración de las señales fue fijada en 0.65536 s, para obtener $2^{15} = 32768$ muestras y se utilizó un valor de $J = 53$ (aproximadamente un polo por kHz de f_s según [15]). Para generar una función glótica $u[k]$ adecuada, de manera tal que el jitter de la señal sea conocido, se usaron los modelos de jitter de duración de periodos independiente y correlacionada descritos en Capítulo 2 (ver Sección 2.5.3). Los parámetros empelados para el modelo de duración de periodos correlacionada fueron: $a_0 = 0.4513$, $a_1 = 0.7331$ y $a_2 = -0.4647$, para una frecuencia de microtremor de 6 Hz y un ancho de banda de 4 Hz según [16].

6.3. Gráficos de Bland-Altman

Ante el surgimiento de una nueva técnica o dispositivo para estimar una magnitud, ya sea temperatura, presión arterial, o incluso el jitter vocal, es necesario contar con un método que permita evaluar el nivel de concordancia entre el método nuevo y el método *de referencia* anterior. Una manera efectiva y de uso muy difundido de realizar esta comparación entre un método novedoso y otro método utilizado previamente, es mediante la gráfica de Bland-Altman [173].

El objetivo de esta gráfica es evaluar el comportamiento del error del nuevo método respecto al método de referencia. Lo deseable para un método nuevo es que dicho error sea lo menor posible, idealmente cero. Esto reflejaría la exactitud del nuevo método. No obstante, también es importante estudiar la dispersión del error cometido. Con el fin de estudiar tanto la exactitud como la dispersión del error, Bland y Altman propusieron el uso de una gráfica del error vs. valor de referencia.

La medida del error utilizada en el eje vertical puede ser simplemente la diferencia entre el valor de referencia y la estimación del método nuevo, o una versión relativa, normalizando por el

[1] Archivo "43-a_n.wav" de Saarbruecken Voice Database [158].

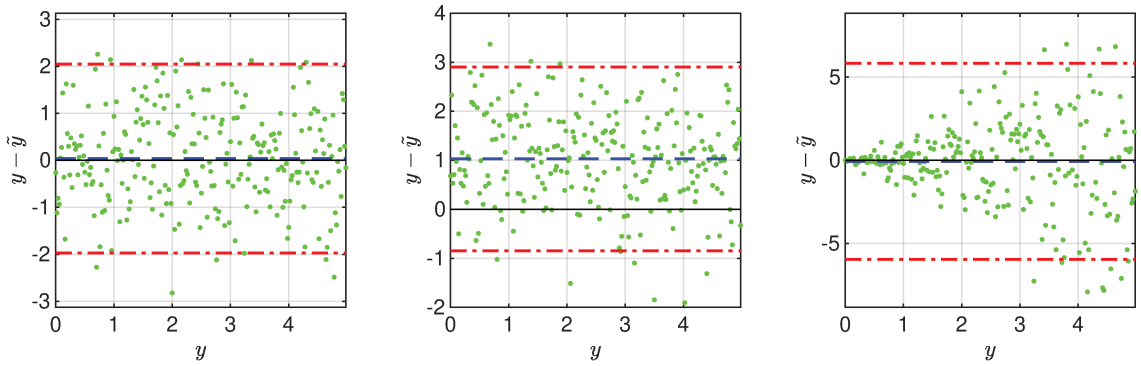


Figura 6.2: Gráficos de Bland-Altman, donde y es el valor de referencia y $\tilde{y}$ es una estimación de y. La línea de trazos azul marca el promedio de las diferencias $y - \tilde{y}$, o sesgo. Las líneas de trazos y puntos rojas delimitan los límites de concordancia. (a) Sesgo nulo, los errores se distribuyen en torno a cero. (b) Sesgo no nulo. El error es igual para todos los valores de y. (c) Sesgo nulo pero con dispersión proporcional al valor de y. En este caso los límites de concordancia no describen adecuadamente el comportamiento del error para cada valor de y (caso heterocedástico).

valor de referencia. También es posible el uso de transformaciones sobre el error para cambiar su distribución [173].

Es importante reconocer una cuestión clave con respecto al valor de referencia. En ocasiones, el verdadero valor de una magnitud no es conocido o no es accesible. En ese caso, el valor de referencia será una estimación realizada con el *mejor* método posible, y es recomendable utilizar el promedio entre el valor de referencia y la estimación obtenida con el método nuevo en el eje horizontal de la gráfica [174]. Si el valor de referencia es efectivamente conocido, como es el caso de las voces con jitter simuladas en este trabajo, entonces puede utilizarse este valor en el eje horizontal, ya que no existe una estimación mejor [174].

La Figura 6.2, muestra distintos gráficos de Bland-Altman. La línea de trazos azul en la figura marca el promedio de las diferencias $y - \tilde{y}$, también llamado sesgo o error sistemático. Las lineas de trazos y puntos (en rojo) delimitan los *límites de concordancia*, que dejan entre ellos el $95\,\%$ de las muestras. En el caso 6.2a, el error sistemático es nulo, mientras que en el caso 6.2b, el error sistemático es aproximadamente 1. Una vez caracterizado el sesgo, puede corregirse restándose de la estimación $\tilde{y}$. En contraste, el caso 6.2c muestra un sesgo nulo, pero los errores no se distribuyen en forma homogénea en torno al promedio de las diferencias. En su lugar, el valor absoluto del error aumenta con el valor de y, por lo que no puede corregirse restando una constante a todas las medidas. Asimismo, el promedio de las diferencias y los límites de concordancia no son útiles para describir el comportamiento del error del método, dado que la varianza del error aumenta con la magnitud estimada. En este caso, también llamado heterocedástico, puede intentar utilizarse el error relativo $\frac{y-\tilde{y}}{y}$ para obtener una distribución homogénea como en los casos anteriores.

Idealmente, un método nuevo será aceptable si tiene un error sistemático cercano a cero y si los límites de concordancia son aceptables bajo algún criterio, generalmente clínico en el caso de variables fisiológicas. Esto significa que los limites de concordancia se encuentren dentro de un margen "aceptable" de error según el uso. Por ejemplo, para algunas magnitudes, un error relativo de $1\,\%$ podría no ser importante, mientras que para otras podría significar la diferencia entre un valor fisiológico y un valor patológico. En consecuencia, la evaluación de la información dada por los límites de concordancia debería valorarse según un criterio de uso [173, 175].

6.4. Metodología

La validación de un método de estimación de jitter utilizando señales de voz reales plantea algunas dificultades. Dado que el verdadero nivel de jitter relativo de una señal real no es conocido, la única forma de caracterizar la exactitud y precisión un nuevo método sobre señales reales consiste en la comparación con otra medida de referencia utilizada como *gold standard*. Una posible medición de referencia es la obtenida por el *software* PRAAT aunque, por lo descrito en la Sección 2.5, es sabido que este método falla para la estimación de valores altos de jitter y depende fuertemente de la frecuencia fundamental de la voz analizada. Por esa razón, una alternativa más adecuada es evaluar nuevos métodos para estimación de jitter, o la comparación entre métodos existentes, utilizando voces sintéticas cuyo valor de jitter es conocido de antemano [67–69].

Teniendo esto en cuenta, se llevaron a cabo una serie de experimentos numéricos empleando señales de voz sintéticas cuyo valor de jitter relativo es conocido de antemano y puede variarse a voluntad. Dichas señales fueron sintetizadas como se explicó anteriormente, mediante el modelo fuente-filtro y los modelos de jitter descriptos en la Sección 2.5.3: el Modelo 1, donde la duración de los periodos es una variable aleatoria con distribución Gaussiana y cada duración es independiente de la del ciclo anterior, y el Modelo 2, en el que la duración de los periodos está correlacionada con la duración de ciclos anteriores. El objetivo de utilizar dos modelos consistió en estudiar la dependencia del desempeño del método con el modelo de perturbación.

Para demostrar que es posible estimar el jitter relativo a partir de la variación total de $T_0(t)$, se realizó una comparación de métodos de estimación de jitter al aplicarse sobre un *chirp* lineal. Esto se debe a que, para esa señal, es posible conocer exactamente su serie de periodos. En consecuencia, puede calcularse sin error el jitter relativo obtenido en forma tradicional. Asimismo, es posible hallar una solución analítica de la integral en la Ecuación (6.11) para calcular el jitter de variación total, así como también aplicar el Algoritmo 3. Por lo tanto, el objetivo de este experimento fue verificar que el jitter de variación total, calculado en forma analítica o numérica, se aproxima al valor del jitter relativo determinado a partir de la serie de periodos.

Seguido de esto, se realizaron experimentos con señales de voz sintéticas. En primer lugar, se determinaron los valores óptimos, para esta aplicación, de σ, proporcional al ancho de la ventana de análisis, y de N, el orden de los operadores de *synchrosqueezing*. Luego, se comparó el desempeño del método propuesto contra PRAAT, al aplicarse sobre señales con dos valores diferentes de $\bar{F}_0$ y dos rangos de jitter relativo. También se estudió la robustez al ruido de la técnica propuesta al aplicarla sobre señales degradadas con ruido blanco real Gaussiano, para diferentes órdenes de los operadores de *synchrosqueezing*. El último experimento realizado sobre voces sintéticas consistió en la comparación entre el método propuesto y PRAAT para señales con $\bar{F}_0$ entre 75 Hz y 225 Hz.

Finalmente, como una evaluación preliminar, se empleó el método presentado anteriormente para estimar el jitter relativo de señales de voz reales. El valor de jitter de referencia para este experimento se obtuvo a partir de señales de electroglotografía adquiridas simultáneamente con las señales de voz, ya que es posible obtener valores de jitter más confiables con PRAAT al utilizarse estas señales más sencillas.

En resumen, se llevaron a cabo los siguientes experimentos:

1. Estimación de jitter de un *chirp* lineal.
2. Estimación de jitter de voces sintéticas:
   - *a*) Experimentos para encontrar los valores de σ y N óptimos para la estimación de jitter.
   - *b*) Comparación de desempeño con PRAAT para señales sin ruido (para $\bar{F}_0$ de 100 o 200 Hz y jitter verdadero entre 0.2 y $15.0\,\%$).
   - *c*) Comparación de desempeño con PRAAT para señales con ruido.
   - *d*) Comparación de desempeño con PRAAT para $\bar{F}_0$ entre 75 y 225 Hz.

3. Evaluación preliminar con señales reales (EGG en lugar de señal de voz como referencia).

6.5. Resultados

A continuación se describirán los resultados de los experimentos realizados. En primer lugar se detallará la comparación, sobre un *chirp* lineal, entre el jitter relativo tradicional y el jitter de variación total propuesto en este trabajo. En segundo lugar se describirá cómo se determinaron empíricamente los parámetros σ y N para la estimación de $F_0(t)$ y su derivada mediante los operadores de *synchrosqueezing* de orden superior. En tercer lugar, se exhibirán los resultados obtenidos al comparar el método propuesto con PRAAT para señales sintetizadas con distintos modelos de jitter, distintas frecuencias fundamentales y con diferentes rangos de jitter relativo verdadero. Por último, se reportará la evaluación preliminar de la técnica descripta en este trabajo sobre señales reales.

En adelante se utilizarán las siguientes medidas de error:

- Error Relativo (ER):

$$ER = \frac{(jitter\ verdadero - jitter\ estimado)}{jitter\ verdadero} \times 100\,\% \tag{6.17}$$

- Promedio de Error Relativo (MER):

$$MER = \frac{1}{J}\sum_{j=1}^{J} ER_j \tag{6.18}$$

  donde ER_j es el error relativo obtenido para la estimación de jitter de la j-ésima señal.

- Promedio del valor absoluto del Error Relativo (MAE):

$$MAE = \frac{1}{J}\sum_{j=1}^{J} |ER_j| \tag{6.19}$$

  donde $|ER_j|$ es el valor absoluto del error relativo obtenido para la estimación de jitter de la j-ésima señal.

6.5.1. Estimación de jitter de un *chirp* lineal

Sea $x(t)$ un *chirp* lineal dado por la expresión:

$$x(t) = \cos(2\pi\phi(t)) \tag{6.20}$$

donde $\phi(t) = \alpha t^2 + \beta t$. Entonces, su frecuencia instantánea está dada por la expresión $\phi'(t) = 2\alpha t + \beta$, y su derivada por $\phi''(t) = 2\alpha$. En este caso, es posible encontrar una solución exacta para la integral de la Ecuación (6.11) en términos de α y β (ver Apéndice A.4), de manera tal que el jitter de variación total de esta señal es:

$$jitter_{\%}^{VT} \approx \frac{1}{L - \bar{T}_0}\,\mathrm{VT}\,[T_0(t)]_0^L \times 100\,\% \tag{6.21}$$

$$jitter_{\%}^{VT} \approx \frac{1}{L - \bar{T}_0}\left[\frac{1}{\beta} - \frac{1}{\beta + 2L\alpha}\right] \times 100\,\%. \tag{6.22}$$

β	α	Jitter	Error utilizando la Ecuación (6.11)	Error utilizando la Ecuación (6.22)
100	1	$0.0197\,\%$	$1.72\,\%$	$< 0.01\,\%$
	2	$0.0390\,\%$	$1.48\,\%$	$0.02\,\%$
	4	$0.0760\,\%$	$1.56\,\%$	$0.06\,\%$
200	1	$0.0050\,\%$	$0.20\,\%$	$< 0.01\,\%$
	2	$0.0099\,\%$	$0.79\,\%$	$< 0.01\,\%$
	4	$0.0195\,\%$	$0.84\,\%$	$< 0.01\,\%$

Tabla 6.1: Valor absoluto del error relativo de la estimación de jitter utilizando el método de la variación total (VT) de la Ecuación (6.11) y la Ecuación (6.22) para un *chirp* lineal.

También es posible calcular el jitter relativo por el método tradicional dado en la Ecuación (6.1), ya que el principio y fin de cada ciclo pueden ser determinados exactamente en este caso como:

$$t_j = \frac{-\beta + \sqrt{\beta^2 + 4j\alpha}}{2\alpha}, \tag{6.23}$$

con $j \in \mathbb{N}_0$ y, en consecuencia, la longitud de cada ciclo puede ser calculada como $T_j = t_{j+1} - t_j$. Por lo tanto, es posible obtener la secuencia de periodos para esta señal sencilla.

Si el enfoque desarrollado en este trabajo es válido, es decir, si es posible estimar el jitter a partir de una aproximación de tiempo continuo del periodo fundamental en lugar de la secuencia de periodos, entonces los valores de jitter medidos mediante la Ecuación (6.22) y mediante el Algoritmo 3 deberían aproximarse al valor dado por el método tradicional expresado en la Ecuación (6.1).

La Tabla 6.1 consigna los valores de los parámetros α, β y el jitter estimado con dichos valores para un *chirp* lineal como el descripto anteriormente. La cuarta y quinta columna muestra el valor absoluto del error relativo estimado por medio de las Ecuaciones (6.11) y (6.22) respectivamente. Se destaca que la estimación utilizando la Ecuación (6.22), que implementa una solución analítica de la integral de la Ecuación (6.11), es bastante exacta, como puede verse de los bajos valores de error de la última columna de la Tabla 6.1, donde el error más alto es de $0.06\,\%$. Asimismo, la estimación del jitter utilizando el Algoritmo 3 y la Ecuación (6.11), resultaron ser también excelentes aproximaciones al valor real de jitter, aunque el error más alto en este caso es de $1.72\,\%$. Esta diferencia puede deberse mayormente al número de muestras que deben eliminarse del comienzo y fin de la señal, con el objetivo de reducir el impacto de los efectos de borde que afectan las estimaciones de la FI y su derivada (ver Sección 3.6).

6.5.2. Estimación de jitter de voces sintéticas

σ y N óptimos para la estimación de jitter relativo

Con el objetivo de encontrar un valor de σ que provea la mejor estimación del jitter relativo de una señal, la relación entre este parámetro y el error en la estimación de jitter fueron estudiados. Para ello, σ fue redefinido como:

$$\sigma_p = p\frac{\bar{T}_0\sqrt{2\pi}}{6}, \tag{6.24}$$

donde $p \in \mathbb{R}^+$ es simplemente el número de periodos fundamentales que caben dentro del ancho efectivo de la ventana. El valor de p fue seleccionado empíricamente, minimizando la mediana del MAE sobre un grupo de 2000 señales sintetizadas utilizando los Modelos 1 (1000 señales) y 2 (1000 señales) al variar p entre 2 y 4 con pasos de 0.02. Las señales de ambos modelos en este grupo

	Error Mínimo	p	Rango intercuantil 2.5 % - 97.5 %
FSST2	2.51 %	2.04	29.73 %
FSST3	2.37 %	2.06	23.85 %
FSST4	2.05 %	3.46	8.27 %

Tabla 6.2: Se reportarn los valores de error mínimos para el correspondiente valor de p, para FSST2, FSST3 y FSST4. También se muestra el ancho del rango intercuantil 2.5 %-97.5 % para su comparación.

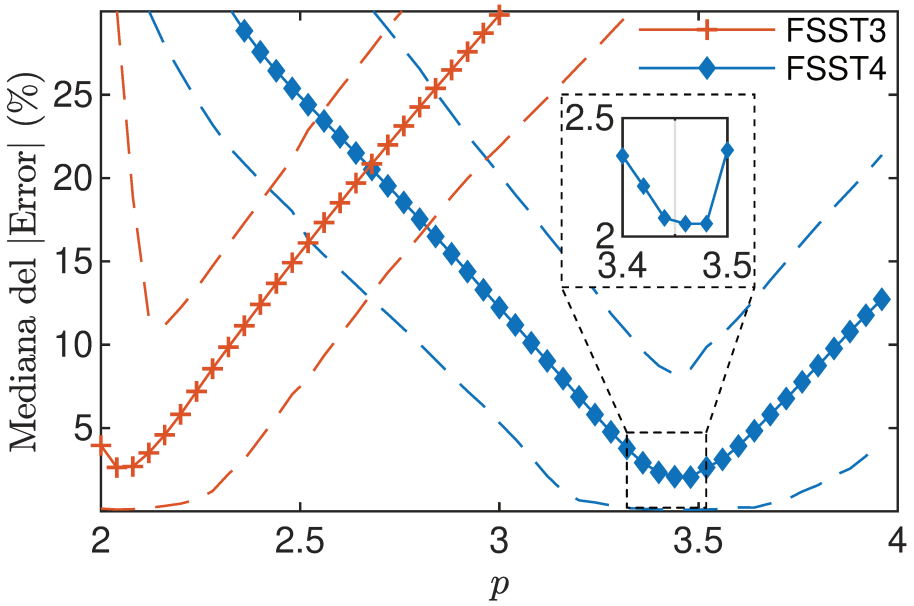


Figura 6.3: Mediana del valor absoluto del error vs. p, donde p es el número de periodos fundamentales ($\bar{T}_0$) que caben en el ancho efectivo de $g(t)$. "+" corresponde a la estimación de jitter con operadores de FSST3, mientras que "$\diamond$" corresponde a FSST4. Las líneas discontinuas muestran los cuantiles 2.5 % y 97.5 %. El mínimo error para FSST4 es 2.05 % y corresponde a $p = 3.46$.

fueron sintetizadas de manera tal que cubrieran el rango de jitter entre 0.2 y 15 uniformemente. Este procedimiento fue repetido para distintos órdenes de *synchrosqueezing*, entre 2 y 4, de manera tal que el orden que minimice el error también sea seleccionado.

La Tabla 6.2 muestra el mínimo error para cada orden de *synchrosqueezing* junto al ancho del rango intercuantil 2.5 %-97.5 %. Puede verse que el error más pequeño se corresponde con FSST4, así como también el rango intercuantil más angosto. El error para FSST2 y FSST3 no difiere significativamente, pero un rango intercuantil más ancho indica que el método es menos preciso para esos órdenes. El comportamiento del error para diferentes valores de p es mostrado en la Figura 6.3 para FSST3 y FSST4. Aquí puede verse que el error mínimo es alcanzado para un valor de $p = 3.46$. Dado que el error mínimo y el rango intercuantil son más grandes para FSST2 y FSST3, el uso de operadores de cuarto orden resulta el más adecuado para esta aplicación.

Comparación con PRAAT (señales sin ruido)

Para ambos modelos de jitter descritos en la Sección 2.5.3, se estudió la estimación de jitter para dos $\bar{F}_0$ diferentes y dos rangos de jitter verdadero. Los valores de $\bar{F}_0$ utilizados fueron 100 y 200 Hz, mientras que los rangos de jitter utilizados fueron $[0.2\,\%, 1.2\,\%]$ y $[1\,\%, 15\,\%]$. Estos valores de jitter fueron seleccionados en base a los niveles de jitter descritos en la literatura correspondientes con voces sanas (entre $0.2\,\%$ y $\sim 1\,\%$) y patológicas (mayores a $1\,\%$) [16]. Esto resulta en cuatro grupos de 250 señales sintetizadas por modelo, donde el verdadero jitter en cada grupo está en el mismo rango ($[0.2\,\%, 1.2\,\%]$ o $[1\,\%, 15\,\%]$) y comparte la misma $\bar{F}_0$ (100 Hz o 200 Hz).

Las Figuras 6.4 y 6.5 muestran las gráficas de Bland-Altman donde cada punto representa a una señal [173]. También se muestra el MER en línea de trazos así como los límites de concordancia (LoA, por sus siglas en inglés) en lineas de trazos y puntos. Los LoA indican aquí la dispersión del error. Desde la Figura 6.4a hasta 6.4h se muestra, para el Modelo 1, el comportamiento de ambos métodos (jitter por variación total y PRAAT) para valores de jitter en el intervalo $[0.2\,\%, 1.2\,\%]$ y ambas frecuencias fundamentales evaluadas. Los mismos resultados, pero para el Modelo 2, pueden apreciarse desde la Figuras 6.5a hasta 6.5h. Puede observarse que ambos métodos tienen un desempeño similar para $\bar{F}_0 = 100$ Hz, siendo PRAAT ligeramente superior, con menor distancia entre los LoA y un error sistemático más cercano a 0. Lo opuesto ocurre para $\bar{F}_0 = 200$ Hz, como puede observarse de las Figuras 6.4c y 6.4d, para el Modelo 1, y 6.5c y 6.5d, para el Modelo 2, donde PRAAT exhibe un sesgo más alto que el método de VT. Debe notarse que, para estos experimentos, se utilizaron operadores de FSST4 para la estimación de la FI y su derivada, siguiendo los pasos resumidos en el Algoritmo 3.

Un resultado interesante puede observarse de la tercera y cuarta fila de las Figuras 6.4 y 6.5, donde se muestra el desempeño de ambos métodos para valores de jitter entre $1\,\%$ y $15\,\%$. Las Figuras 6.4f, 6.4h, 6.5f y 6.5h revelan que PRAAT falla al estimar el jitter cuando el verdadero valor de jitter es mayor a $8\,\%$ (para $\bar{F}_0 = 100$ Hz) y $5\,\%$ (para $\bar{F}_0 = 200$ Hz), sin importar el modelo de jitter utilizado. En contraste, el método aquí presentado provee medidas de jitter con un error sistemático que es cercano 0 para todo el rango de jitter, a la vez que mantiene una distancia más angosta que PRAAT entre los LoA.

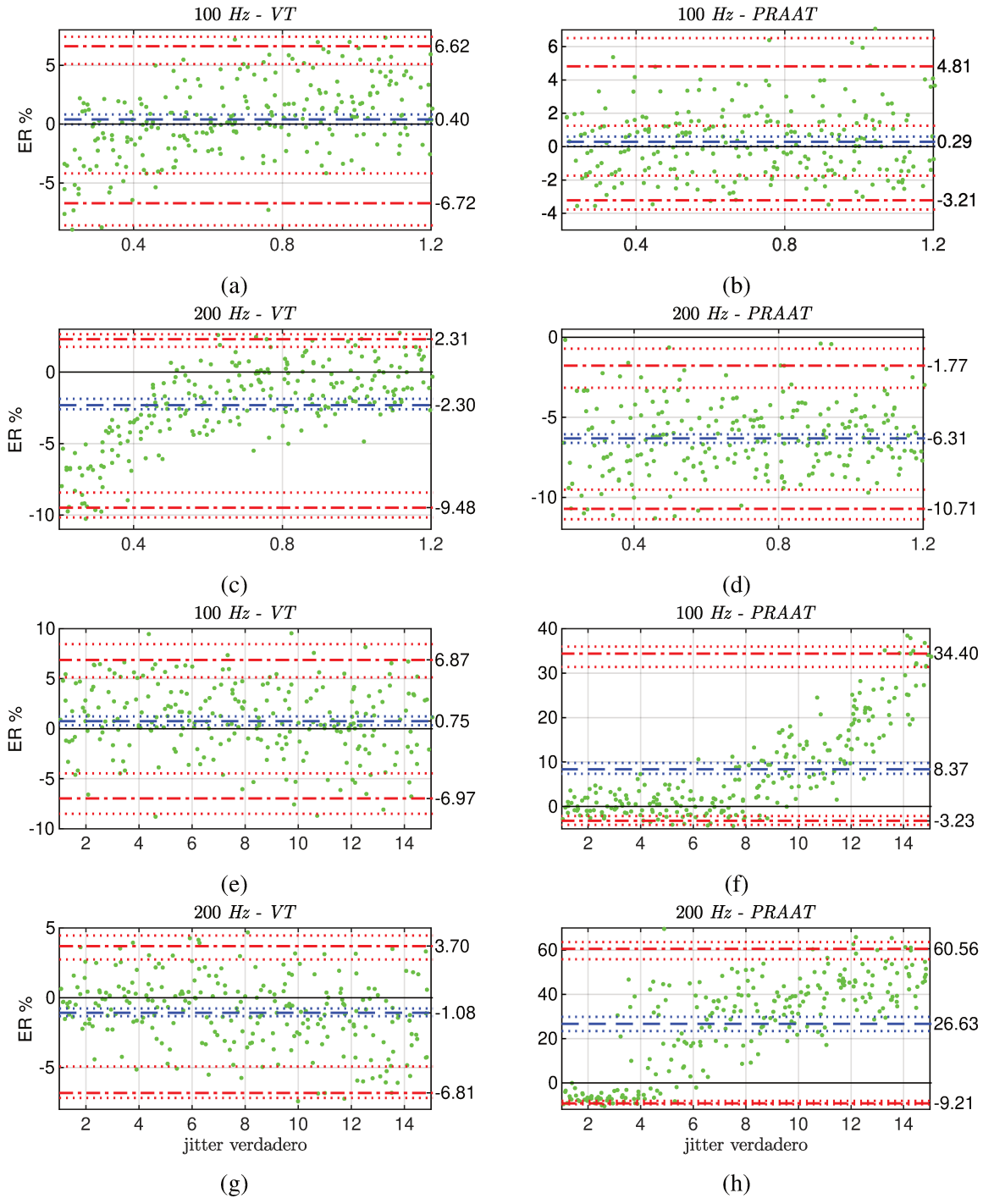


Figura 6.4: Gráficas de Bland-Altman para el modelo de duraciones de ciclos independientes e idénticamente distribuidas (Modelo 1). Cada punto representa el error para una señal. Las líneas de trazos en azul representan el sesgo, mientras que las líneas de trazos y puntos en rojo representan los límites de concordancia superior e inferior. En líneas punteadas se muestran los intervalos de confianza del $95\,\%$ para cada estadístico. La primera y la segunda fila corresponden a valores de jitter en el rango $[0.2\,\%, 1.2\,\%]$, mientras que las filas tercera y cuarta corresponden al rango $[1\,\%, 15\,\%]$. La columna izquierda muestra los resultados para el método de variación total (VT) mientras que la columna derecha muestra los resultados para PRAAT. La frecuencia fundamental promedio se muestra en el título de cada gráfica (100 Hz o 200 Hz).

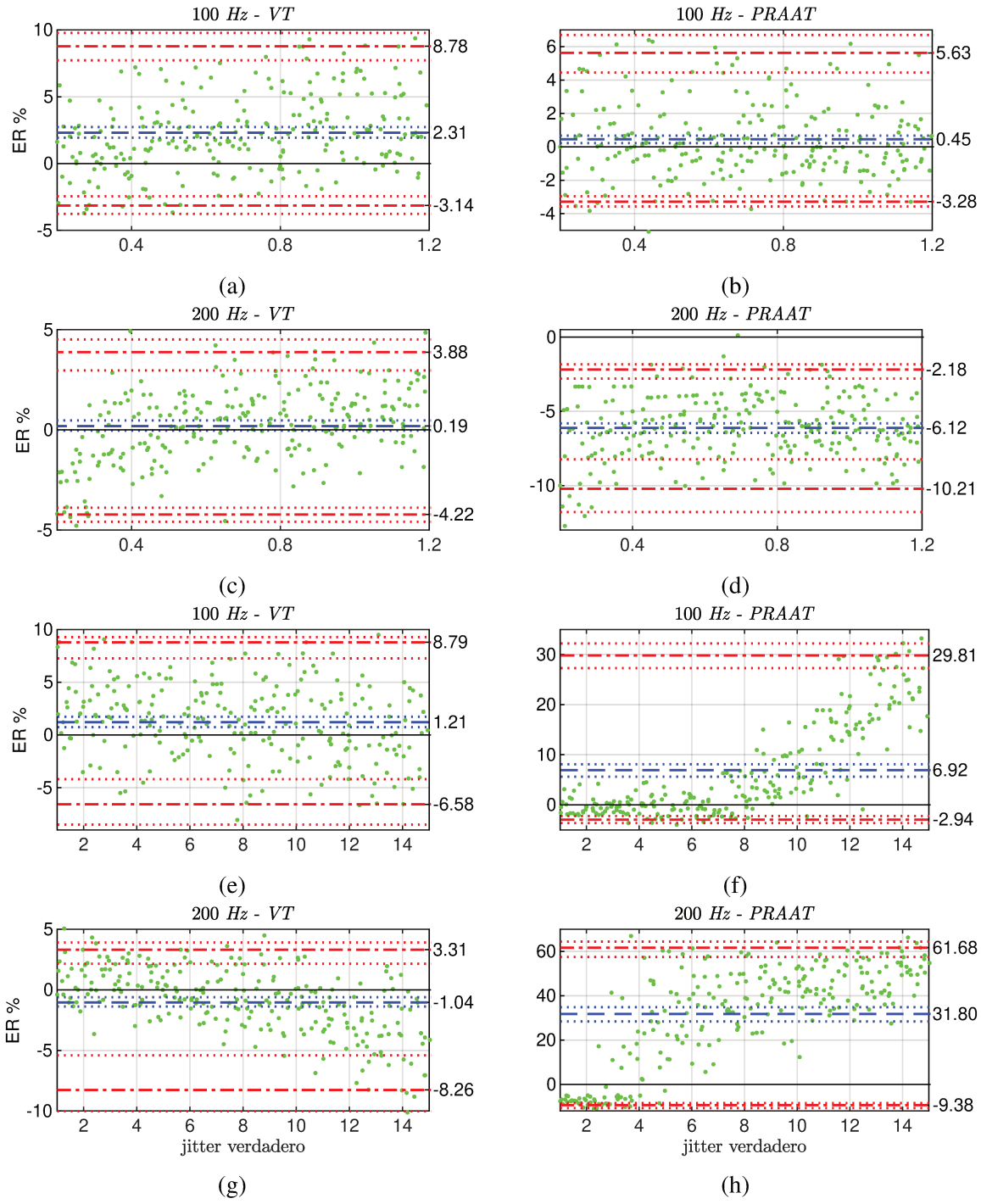


Figura 6.5: Gráficas de Bland-Altman para el modelo de duraciones de ciclos independientes e idénticamente distribuidas (Modelo 2). Cada punto representa el error para una señal. Las líneas de trazos en azul representan el sesgo, mientras que las líneas de trazos y puntos en rojo representan los límites de concordancia superior e inferior. En líneas punteadas se muestran los intervalos de confianza del $95\,\%$ para cada estadístico. La primera y la segunda fila corresponden a valores de jitter en el rango $[0.2\,\%, 1.2\,\%]$, mientras que las filas tercera y cuarta corresponden al rango $[1\,\%, 15\,\%]$. La columna izquierda muestra los resultados para el método de variación total (VT) mientras que la columna derecha muestra los resultados para PRAAT. La frecuencia fundamental promedio se muestra en el título de cada gráfica (100 Hz o 200 Hz).

Modelo 1: Duración de periodos independiente								
	$\bar{F}_0 = 100$Hz				$\bar{F}_0 = 200$Hz			
	Rango de Jitter: $[0.2, 1.2]$		Rango de Jitter: $[1, 15]$		Rango de Jitter: $[0.2, 1.2]$		Rango de Jitter: $[1, 15]$	
	VT	PRAAT	VT	PRAAT	VT	PRAAT	VT	PRAAT
MAE	2.55	1.71	2.96	9.24	2.80	6.31	2.18	29.57
95 % CI	(2.28, 2.76)	(1.56, 1.84)	(2.63, 3.25)	(8.01, 10.42)	(2.46, 3.09)	(6.04, 6.64)	(1.68, 2.44)	(27.31, 31.72)
MER	0.40	0.29	0.75	8.37	-2.30	-6.31	-1.08	26.63
95 % CI	(-0.02, 0.82)	(0, 0.57)	(0.36, 1.18)	(7.12, 9.45)	(-2.59,-1.87)	(-6.55,-6.08)	(-1.34,-0.77)	(24.30, 29.22)
U-LoA	6.62	4.81	6.87	34.40	2.31	1.77	3.70	60.56
95 % CI	(5.09, 7.41)	(1.20, 6.56)	(4.38, 7.99)	(31.76, 36.36)	(1.77, 2.65)	(-2.85,-0.67)	(2.75, 4.46)	(55.28, 64.00)
L-LoA	-6.72	-3.21	-6.97	-3.23	-9.48	-10.71	-6.81	-9.21
95 % CI	(-8.61,-4.18)	(-3.75,-2.23)	(-8.57,-4.28)	(-4.30,-2.03)	(-10.16,-8.42)	(-11.30,-9.35)	(-7.18,-4.46)	(-9.88,-7.67)
Modelo 2: Duración de periodos autocorrelacionada								
	$\bar{F}_0 = 100$Hz				$\bar{F}_0 = 200$Hz			
	Rango de Jitter: $[0.2, 1.2]$		Rango de Jitter: $[1, 15]$		Rango de Jitter: $[0.2, 1.2]$		Rango de Jitter: $[1, 15]$	
	VT	PRAAT	VT	PRAAT	VT	PRAAT	VT	PRAAT
MAE	3.00	1.75	3.29	7.88	1.55	6.12	2.26	34.49
95 % CI	(2.75, 3.26)	(1.58, 1.92)	(2.97, 3.63)	(6.64, 9.12)	(1.39, 1.70)	(5.80, 6.37)	(2.01, 2.45)	(32.03, 36.95)
MER	2.31	0.45	1.21	6.92	-0.19	-6.12	-1.04	31.80
95 % CI	(1.93, 2.73)	(0.19, 0.70)	(0.74, 1.72)	(5.56, 7.92)	(-0.04,-0.48)	(-6.44,-5.80)	(-1.38,-0.61)	(29.22, 34.87)
U-LoA	8.78	5.63	8.79	29.81	3.88	2.18	3.31	61.68
95 % CI	(7.73, 9.78)	(4.39, 6.69)	(7.26, 9.28)	(27.26, 32.33)	(2.96, 4.51)	(-2.88,-1.73)	(2.15, 3.9)	(57.63, 64.52)
L-LoA	-3.14	-3.28	-6.58	-2.94	-4.22	-10.21	-8.26	-9.38
95 % CI	(-3.76,-2.44)	(-3.61,-2.96)	(-8.50,-4.20)	(-3.58,-2.30)	(-4.58,-3.89)	(-11.92,-8.88)	(-10.02,-5.40)	(-10.69,-8.09)

Tabla 6.3: Resumen de los errores límites de concordancia (LoA) para los Modelos 1 y 2. Todos los valores están dados en porcentajes. Se reportan los resultados para ambos métodos, rangos de jitter y frecuencia fundamental promedio. U-LoA límite de concordancia superior, y L-LoA límite de concordancia inferior. Los intervalos de confianza del $95\,\%$ están dados debajo de cada medición.

La Tabla 6.3 resume los resultados mencionados para los Modelos 1 y 2, y adicionalmente muestra el MAE para cada grupo de señales sintéticas. Para el rango de jitter $[0.2\,\%, 1.2\,\%]$ y $\bar{F}_0 = 100$ Hz, ambos métodos tienen un desempeño similar y, como se indicó anteriormente, el método propuesto resulta mejor para el caso de $\bar{F}_0 = 200$ Hz. A su vez, el error para PRAAT aumenta al incrementar la frecuencia fundamental promedio. Por ejemplo, para el Modelo 1, el MAE de PRAAT se incrementa desde $1.71\,\%$ ($\bar{F}_0 = 100$ Hz) a $6.31\,\%$ ($\bar{F}_0 = 200$ Hz) para jitter en el rango $[0.2, 1.2\,\%]$. En contraste, el método aquí propuesto muestra un MAE similar para ambas frecuencias: $2.55\,\%$ (100 Hz) y $2.80\,\%$ (200 Hz). Aun más, el método basado en la variación total muestra un desempeño claramente superior a PRAAT en el rango de jitter verdadero [$1\,\%$ - $15\,\%$], siendo tanto el MER como el MAE más bajos en este caso. Adicionalmente, los límites de concordancia se encuentran más cerca para dicho método para este rango de jitter, por lo que la dispersión del error también es menor.

Comparación con PRAAT (señales con ruido)

Con el objetivo de evaluar la robustez al ruido del método propuesto, fueron sintetizadas señales con $\bar{F}_0 = 100$ Hz para ambos intervalos de jitter verdadero empleando el Modelo 1, y se contaminaron con ruido blanco Gaussiano real para obtener relaciones señal a ruido de 20 a 50 dB, con pasos de 10 dB. El rango de SNR fue tomado basándose en estudios previos [69, 72]. La Figura 6.6 muestra los boxplots del valor absoluto del error para cada SNR. Diferentes órdenes de *synchrosqueezing* fueron utilizados para estimar la FI y su derivada para el método basado en la variación total.

De la primera fila de la Figura 6.6 puede observarse que, para un jitter verdadero en el rango $[0.2\,\%, 1.2\,\%]$, el error para el método basado en VT tiene una mediana más baja para órdenes más altos de *synchrosqueezing* (sin importar el valor de SNR), siendo el error más bajo aquél

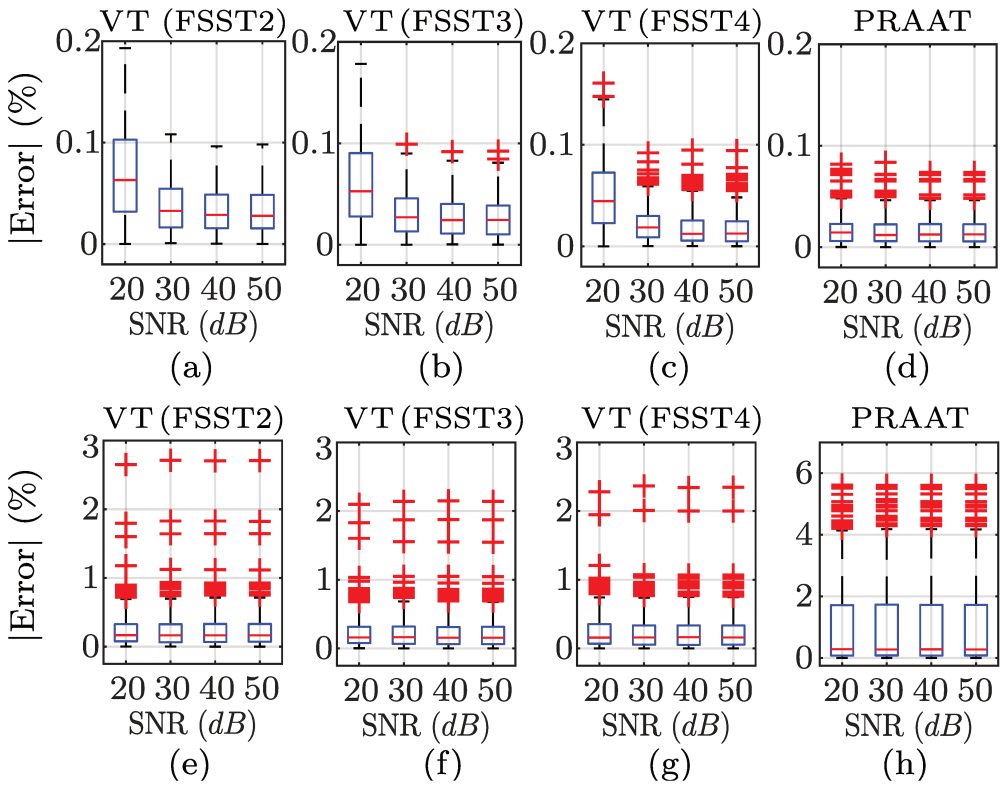


Figura 6.6: Valor absoluto del error vs. SNR. Los gráficos de caja muestran el comportamiento del método de variación total (utilizando FSST2, FSST3 y FSST4) en presencia de ruido (notar el cambio de escala en la subfigura (h)). La primera fila corresponde a señales en el rango $[0.2\,\%, 1.2\,\%]$, mientras que la segunda fila corresponde a señales con jitter en el rango $[1\,\%, 15\,\%]$. El error es calculado aquí como $|Error| = |jitter\ verdadero\text{-}jitter\ estimado|$.

encontrado para FSST4. Asimismo, la dispersión de los errores es a la vez menor para este orden. Considerando los valores de SNR, el error para el método de VT incrementa para una SNR de 20 dB cuando se la compara con valores más altos de SNR, para todos los órdenes de FSST. En contraste, PRAAT mantiene aproximadamente el mismo desempeño para todos los valores de SNR.

La segunda fila de la Figura 6.6 muestra que, el método basado en la VT obtiene un mejor desempeño en términos de la mediana del error y su dispersión que PRAAT (nótese el cambio de escala en el eje vertical), para cualquier nivel de ruido. La diferencia entre los distintos órdenes de FSST aquí es menos evidente en la Figura 6.6, aunque para los operadores de FSST4 la dispersión del error es ligeramente menor.

Comparación con PRAAT para distintas $\bar{F}_0$

Si bien los resultados anteriores se detallaron para dos frecuencias fundamentales promedio, la Figura 6.7 muestra el error obtenido para señales sintéticas con distintas $\bar{F}_0$ entre 75 y 225 Hz. De las Figuras 6.7a y 6.7c puede verse que PRAAT falla a medida que la frecuencia fundamental sube (tal como se describió en la Sección 2.5.2). En contraste, las Figuras 6.7b y 6.7d muestran que el error del método basado en la variación total mantiene una mediana del error por debajo de $5\,\%$ para cualquier frecuencia y para valores de jitter verdadero tanto en el intervalo $[0.2\,\%, 1.2\,\%]$ como en el intervalo $[1\,\%, 15\,\%]$.

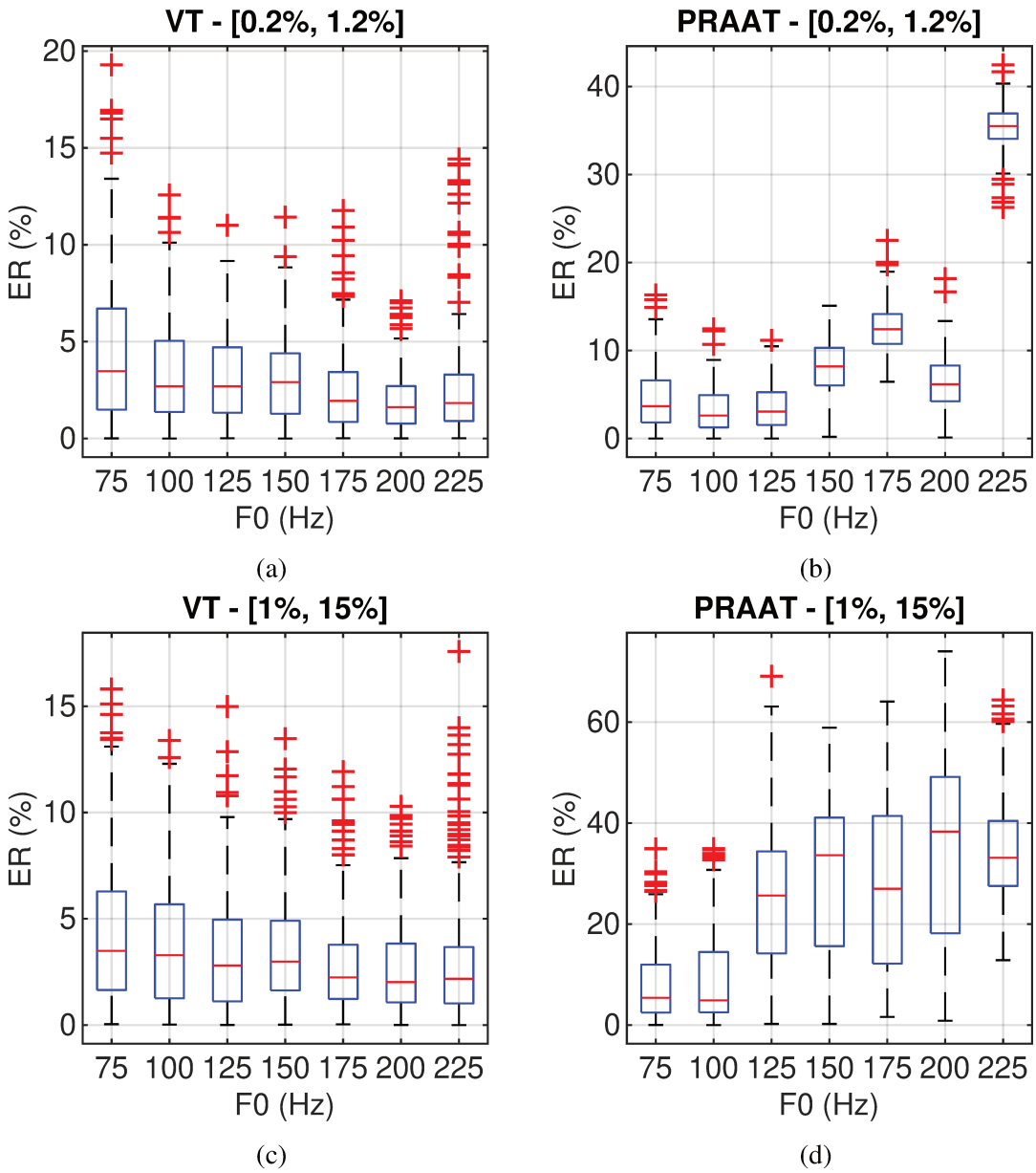
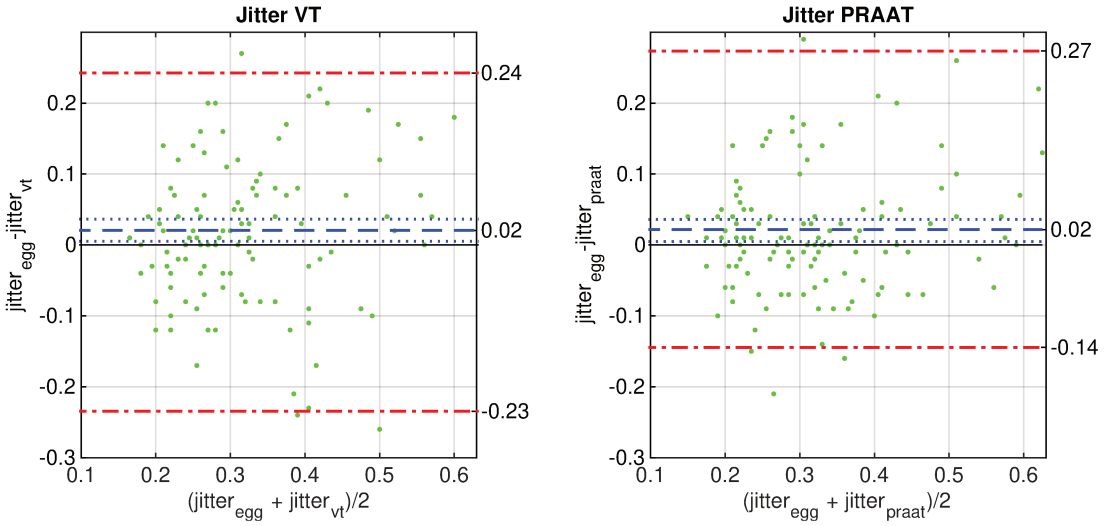


Figura 6.7: Gráficos de caja y bigotes que muestran el desempeño de PRAAT y el método basado en la variación total (VT) para la estimación de jitter para distintas frecuencias fundamentales promedio. La fila superior (gráficas a y b) se corresponde con señales cuyo jitter real se encuentra en el intervalo $[0.2\,\%, 1.2\,\%]$, mientras que para la fila inferior (gráficas c y d) se encuentra en el intervalo $[1\,\%, 15\,\%]$.



(a) (b)

Figura 6.8: Gráficos de Bland-Altman para señales reales. (a) Resultados para el método propuesto. (b) Resultados para PRAAT. Tanto en (a) como en (b) se utiliza como referencia el promedio de los valores hallados por el método correspondiente y una estimación del jitter obtenida a partir de la señal de EGG adquirida simultáneamente.

6.5.3. Resultados preliminares con señales reales

Se realizó un experimento para evaluar la utilidad del método descripto utilizando señales reales provenientes de la base de datos Saarbruecken [158] (ver listado de señales utilizadas en el Apéndice A.5). Con el fin de obtener una estimación de jitter que pueda ser utilizada como referencia con la mayor confianza posible y hallado con PRAAT, se emplearon señales de electroglotografía (EGG) en lugar de las señales de voz. Se utilizaron únicamente señales tipo 1 de acuerdo a la clasificación explicada en la Sección 2.6.1. Luego, se estimó el jitter relativo a partir de las señales de voz adquiridas simultaneamente con la señal de EGG, mediante PRAAT y mediante el método presentado en este trabajo.

En la Figura 6.8 pueden observarse los gráficos de Bland-Altman, tanto para PRAAT como para el método de variación total. En este caso se consideró al error como la diferencia entre el valor de referencia (PRAAT con EGG) y los valores dados por los otros dos métodos (PRAAT y método de la VT con la señal de voz). Es posible ver que tanto PRAAT como el método propuesto tienen un sesgo similar ($0.02\,\%$), aunque los límites de concordancia de PRAAT se alejan menos de la media de las diferencias. Asimismo, el promedio del error absoluto en cada caso fue de $0.0702\,\%$ (intervalo de confianza $[0.0562\,\%; 0.0820\,\%]$) y $0.0859\,\%$ (intervalo de confianza $[0.0699\,\%; 0.0985\,\%]$) para PRAAT y el método de la VT respectivamente. Esta diferencia no es clínicamente relevante ya que, teniendo en cuenta los valores de jitter de referencia de estas señales, no alteraría la conclusión sobre la presencia de patologías (aún con el máximo error por exceso posible, ninguna señal superaría el valor entre $1\,\%$ y $1.2\,\%$ considerado como cota superior para una voz sana). Estos resultados indicarían que la estimación de jitter por variación total tiene un desempeño equivalente a PRAAT incluso para voces reales.

6.6. Discusión

Los resultados obtenidos sugieren que el método propuesto, basado en la VT de una aproximación de mayor orden del periodo instantáneo, puede utilizarse para calcular el jitter de voces sintéticas para niveles de jitter entre $0.2\,\%$ y $15\,\%$ con un desempeño superior al de PRAAT, el *software* libre y gratuito más utilizado en la práctica clínica. La hipótesis detrás de este nuevo método es que un nivel más elevado de jitter implica mayores oscilaciones en la frecuencia fundamental. En consecuencia, para una ventana de análisis dada, una aproximación local polinomial debería ajustarse mejor a la FI que una aproximación constante a trozos utilizada en la fórmula clásica de jitter relativo. Esto resalta un defecto inevitable de la Ecuación (6.1): el uso de una aproximación de *orden cero* para la FI. Con el propósito de mostrar las diferencias entre la aproximación constante a trozos y una aproximación polinomial de orden superior para la estimación de jitter, se utilizó PRAAT como referencia. El comportamiento heterocedástico del error observado en las Figuras 6.4f, 6.4h, 6.5f y 6.5h, que fue descripto con anterioridad en otros estudios [67, 72], evidencia los problemas que conlleva utilizar una mala aproximación del periodo fundamental instantáneo. En contraste con el caso de PRAAT, el error cometido por el método basado en VT parece no sufrir de este problema, lo que permite una caracterización del error que es válida en todo el rango de jitter analizado. Además parece ser menos sesgado que PRAAT para frecuencias fundamentales más altas, como se muestra en la Tabla 6.3, lo que constituye otra ventaja del método aquí propuesto.

Los resultados para las simulaciones utilizando un *chirp* lineal de la Sección 6.5.1 reafirman la intuición sobre la que se basa el método, es decir, que es posible estimar el jitter reemplazando la secuencia de periodos por una mejor estimación del periodo fundamental instantáneo que pueda reflejar mayores cambios en esta magnitud. Estos resultados fueron los mismos sin importar el orden de *synchrosqueezing* utilizado. En contraste, los experimentos con señales reales muestran que utilizar un valor de $N = 4$ redunda en mejores resultados cuando se trata de estimar el valor de jitter que órdenes más bajos, reduciendo tanto el error como su varianza. Esto parece indicar que aproximaciones locales de segundo y tercer orden no son suficientes, en general, para aproximar la fase de voces sintéticas con los modelos empleados a lo largo de este trabajo. Adicionalmente, la Tabla 6.2 y la Figura 6.3 muestran que cuanto más bajo es el orden de *synchrosqueezing*, más angosta es la ventana para el error mínimo. Como fue explicado en la Sección 6.2.1, esto podría ser una consecuencia de que ventanas más cortas favorecen la aproximación polinómica local de la fase, limitada por el compromiso entre resolución frecuencial y temporal.

Los resultados mostrados en la primera fila de la Figura 6.6 muestran que el método basado en VT es afectado por la presencia de ruido, particularmente para SNR bajas del orden de 20 dB y en el rango de jitter verdadero entre $0.2\,\%$ y $1.2\,\%$, donde la robustez del método propuesto es menor a la de PRAAT. Esto puede ser consecuencia del impacto del ruido en los operadores de FSSTN, así como de las perturbaciones causadas en la estimación de la cresta [127]. Sin embargo, la presencia de estos niveles de ruido en un consultorio o instalación clínica debería ser poco frecuente [176, 177].

A pesar de que es un hecho conocido que un orden más alto de FSSTN vuelve a los operadores más sensibles al ruido [178], puede observarse de las Figuras 6.6a hasta la 6.6c que el desempeño del método basado en VT mejora a medida que aumenta el orden desde $N = 2$ hasta $N = 4$ cuando el jitter verdadero se encuentra en el intervalo $[0.2\,\%, 1.2\,\%]$. No obstante, esto no es evidente al observar la segunda fila de la Figura 6.6 donde el verdadero valor de jitter se encuentra entre $1\,\%$ y $15\,\%$. Puede verse que, en ese caso, el desempeño para FSST4 es sólo ligeramente mejor que para órdenes menores. Dado que estos resultados son similares al caso sin ruido, parece ser que la mejora dada por el aumento del orden de los operadores utilizados para estimar la FI y su derivada supera al impacto negativo del ruido para esta aplicación y para los valores de SNR considerados, teniendo la presencia de ruido una influencia más significativa para niveles más altos de jitter.

La extracción del primer modo, es decir, el de menor frecuencia instantánea, para sintetizar

$\widetilde{x}(t)$ alivia la interferencia entre modos adyacentes y es además equivalente a un filtrado pasabajos de la señal. Este procedimiento produce una estimación de jitter más exacta que utilizando la señal original. A la luz de la interpretación del jitter basada en la variación total, de la Ecuación (6.11), una conclusión que puede guiar el procesamiento de la señal de voz para la estimación de jitter es que cualquier estrategia de filtrado debe preservar la información del primer modo de la señal. Siguiendo esta idea, la presencia de una cresta discontinua constituye una limitación del método, dado que la fase no puede ser aproximada localmente por un polinomio en este caso. No obstante, este tipo de señales son raramente analizadas en la práctica, dado que habitualmente se exige cierto grado de periodicidad para calcular medidas como el jitter y el shimmer, por ejemplo que las señales sean de tipo 1, según lo descripto en el Capítulo 2 [16, 42]. Otra situación problemática podría surgir en el caso de señales en las que el primer modo no es el modo *dominante*. Este es el caso para algunas señales reales, incluso para hablantes con fonación normal, cuando el efecto de la modulación del tracto vocal reduce dramáticamente la amplitud de los coeficientes correspondientes al primer modo en el espectrograma, mientras que se preservan los modos cercanos a las frecuencias formantes del tracto [15]. Esto, sin embargo, puede superarse mediante el clásico filtrado inverso [15] o por técnicas más modernas como la TFTC *de-shape* [132].

La necesidad de encontrar una buena aproximación de $\bar{T}_0$ no debe ser vista como una limitación del método aquí propuesto, ya que este requisito es común a todas las medidas *relativas* de jitter [5, 42, 93]. Asimismo, cualquier otro método diferente al empleado en el presente trabajo podría haberse utilizado con el mismo propósito, como la autocorrelación o el cepstrum real [15].

6.7. Conclusión

En este capítulo se describió una nueva aplicación de los operadores de *synchrosqueezing* en el marco de un método novedoso para la estimación del jitter relativo sin necesidad de segmentar en ciclos la señal de voz. Esta nueva técnica fue basada en una generalización de la fórmula de jitter relativo, interpretada como la variación total del periodo fundamental instantáneo $T_0(t)$. Luego, esta nueva formulación puede expresarse en términos de la frecuencia fundamental instantánea y su derivada, obtenidas a partir de los operadores de *synchrosqueezing*. Los resultados obtenidos indican que el método propuesto posee un desempeño superior al *software* PRAAT al compararse ambos métodos sobre señales de voz sintéticas y valores de jitter entre $0.2\,\%$ y $15\,\%$. La robustez al ruido del método también fue evaluada, obteniéndose resultados similares a PRAAT para SNRs por arriba de 20 dB, empleando operadores de cuarto orden (FSST4). Asimismo, el método propuesto es menos dependiente de la frecuencia fundamental de la voz analizada, habiéndose hallado medianas del error menores a $5\,\%$ en voces con frecuencias fundamentales entre 75 Hz y 225 Hz. Finalmente, una evaluación preliminar sobre señales de voz reales parece indicar que el método tiene potencial para su uso en la práctica clínica. Parte de los resultados aquí reportados fueron publicados en [179].

Capítulo 7

Conclusiones y trabajos a futuro

7.1. Conclusiones

El análisis de perturbaciones permite obtener parámetros relevantes de la señal de voz que reflejan la presencia de patologías en el aparato fonador y en otros sistemas asociados (nervioso, respiratorio, digestivo). Las características obtenidas mediante dicho análisis también permiten medir de manera cuantitativa la calidad vocal, una aplicación de importancia para fonoaudiólogos/as y otros profesionales de la salud de la voz. Como se ha visto, las principales hipótesis detrás de las medidas de perturbación pueden resumirse como:

1. Hipótesis de periodicidad: La señal es aproximadamente periódica, es decir, una señal que consiste en la repetición de una forma de onda con cambios pequeños entre ciclos.

2. Hipótesis de estacionariedad local: El parámetro cuya perturbación se desea estudiar se mantiene constante durante cada ciclo.

La primera se debe a que las medidas de perturbación necesitan de una sucesión con los valores del parámetro perturbado para cada ciclo (ya sea frecuencia, amplitud máxima, etc), y en consecuencia, los ciclos deben estar bien definidos. Por otra parte, la hipótesis de estacionariedad local es necesaria para considerar dicha sucesión como una colección representativa de los valores del parámetro para cada ciclo. En consecuencia, las series de valores del parámetro bajo estudio pueden interpretarse como aproximaciones constantes a trozos de la versión instantánea de dicho parámetro. Por ejemplo, la serie de periodos puede verse como una aproximación de orden cero del periodo fundamental instantáneo, como se explicó en la Sección 6.2.

En este documento se han descrito dos desarrollos en torno a dichas hipótesis. La primera propuesta (Capítulo 5) consiste en un sistema automático para la clasificación de voces en tres tipos, que busca evitar el uso de medidas de perturbación sobre señales que no satisfagan la hipótesis de periodicidad. Con ese fin, se propuso un conjunto de características como parámetros objetivos para la clasificación. Luego se llevaron a cabo una serie de experimentos para evaluar el desempeño de una máquina de vectores de soporte, entrenada a partir de dos conjuntos de señales (MEEI [112] y SVD [158]) con un total de más de 1200 señales, para la tipificación automática de voces. Los resultados obtenidos permiten afirmar que el enfoque es efectivo para la clasificación. Adicionalmente, se sugirió un método que permita la aceptación o rechazo, por parte de un profesional de la salud vocal, de la clase asignada automáticamente sobre la base de la probabilidad a posteriori estimada a la salida del clasificador.

La segunda propuesta de esta tesis doctoral (Capítulo 6) pretende relajar la hipótesis de estacionariedad del periodo fundamental empleada en la estimación de jitter relativo, una medida de perturbación ampliamente difundida. Con ese fin, se propuso la reinterpretación del jitter relativo como una cantidad proporcional a la variación total del periodo fundamental instantáneo. De esta

forma, el nuevo método propuesto requiere que la fase de la señal sea aproximable localmente mediante un polinomio de cuarto orden, permitiendo reflejar una mayor variabilidad del periodo fundamental. Los experimentos realizados sobre señales sintéticas con jitter conocido indican que el método propuesto es más robusto que PRAAT, el *software* libre y gratuito más difundido en la práctica, para la estimación de jitter relativo sobre señales con niveles altos de jitter y en presencia de ruido blanco Gaussiano en la señal.

A continuación se enumeran los principales aportes de esta tesis:

1. Se propuso una metodología para la clasificación automática de señales de voz en tres tipos, basada en descriptores ampliamente utilizados en la práctica clínica y un clasificador lineal, que constituye un nuevo estado del arte en el área.

2. Se demostraron resultados con experimentos intra e inter base de datos, permitiendo estudiar la capacidad de generalización del algoritmo propuesto y sentando un precedente para investigaciones futuras: es necesario emplear distintos corpus de voces para validar nuevas medidas para tipificación automática.

3. Se propuso el uso de la probabilidad a posteriori obtenida a la salida de una SVM como medida de confianza en la clasificación. Esto constituye un paso más hacia la obtención de un sistema capaz de asistir al profesional de la salud vocal que pueda utilizarse en la práctica.

4. Se introdujeron nuevas medidas de perturbación de la forma de onda: varianza normalizada de la componente principal (VNCP) y el desvío estándar de una versión de tiempo corto de VNCP (DSVNCP). Ambas fueron parte del conjunto de características más relevantes.

5. Se presentó una nueva forma de medir el jitter relativo, basada en la estimación de la frecuencia fundamental instantánea y su derivada, que no requiere la determinación de puntos fiduciarios.

6. Se propuso una forma novedosa de interpretar el jitter relativo como la variación total del periodo fundamental instantáneo. Esta nueva interpretación permitió la derivación del método propuesto.

7. Se aportó evidencia de que el nuevo método no sufre las principales limitaciones del jitter relativo basado en la estimación de la serie de periodos: dificultad para estimar valores de jitter mayores a 5-$8\,\%$ y aumento del error con la frecuencia fundamental.

8. El método propuesto es más robusto al ruido (para SNR$>$ 20 dB y $F_0 = 100$ Hz) que el método clásico.

9. Se demostró una nueva aplicación de los operadores de *synchrosqueezing* a la señal de la voz.

10. Los experimentos con señales de voz reales y patológicas aquí reportados, demuestran la potencialidad del método para su uso clínico.

Mediante estos aportes se consumaron cada uno de los objetivos específicos establecidos en el Capítulo 1. Con respecto a los objetivos específicos 1 y 2, debe destacarse que las características utilizadas en el sistema de análisis de condición de la voz descripto en el Capítulo 5 fueron seleccionadas por su uso difundido en el estudio clínico de la salud vocal y por ser capaces de traducir criterios subjetivos como la percepción de voces roncas o con ruido aéreo (CPP y HNR, por ejemplo). No se utilizaron características basadas en otras señales provenientes del aparato fonador, como la vibración en la piel de cuello o electromiogramas, quedando su exploración como objetivo de futuros trabajos.

Las nuevas características introducidas en el Capítulo 5, VNCP y DSVNCP pueden considerarse como criterios novedosos para la clasificación de voces en tres tipos, a la vez que constituyen, como se mencionó anteriormente, medidas de perturbación de la forma de onda. De esta forma se cubrió el tercer objetivo específico consignado previamente.

El cuarto y último objetivo específico se satisfizo a partir del desarrollo presentado en el Capítulo 6, en el que se presentó una nueva técnica para estimar la perturbación de la duración de los ciclos, que no necesita la identificación de puntos fiduciarios sobre la señal y es capaz de estimar valores de jitter relativo de hasta un $15\,\%$.

Finalmente, a través de estos aportes se dio cumplimiento al objetivo general consignado en el Capítulo 1, principalmente en la forma de contribuciones que permiten mejorar el análisis de perturbaciones a partir del aprendizaje maquinal y el análisis tiempo-frecuencia.

7.2. Trabajos a futuro

Los trabajos desarrollados a lo largo de esta tesis doctoral no son, ni pretenden ser, definitivos. En su lugar, son contribuciones que dejan abierta la posibilidad de nuevos aportes en las direcciones planteadas. Considerando el desarrollo de algoritmos de clasificación automática en los tres tipos descritos [42], estudios venideros deberían explorar con mayor énfasis la variabilidad interprofesional y su impacto en el desempeño del clasificador. Por ejemplo, mediante una evaluación de la concordancia entre los especialistas y el sistema automático mediante un test de kappa de Cohen/ Fleiss [180]. Asimismo, la evaluación de la consistencia de los observadores, considerando las etiquetas en tiempos sucesivos, puede utilizarse para ponderar la clasificación de los observadores, dándole más importancia a las etiquetas provenientes de aquellos observadores más consistentes en detrimento de aquellos menos congruentes [181]. Dado que la forma de onda de la señal se modifica con los cambios en las estructuras resonantes y articulares, otro aspecto a evaluar en trabajos prospectivos es el uso de otras vocales.

Dejando de lado parte de la perspectiva considerada en este trabajo, también podrían evaluarse otras posibilidades para la extracción de características y su clasificación. Si bien es cierto que el enfoque presentado tiene como ventaja conservar cierta interpretabilidad de la clasificación al utilizar descriptores bien difundidos y un clasificador lineal, sería posible la obtención de resultados superadores empleando otro tipo de características y/ u otros algoritmos de clasificación no lineales. Los coeficientes cepstrales de escala de mel [182] o la transformada *scattering* [114, 115] proveen características espectrales que podrían utilizarse para la clasificación. De hecho, los coeficientes de *scattering* de primer nivel son equivalentes a los coeficientes cepstrales de mel [116]. Esta técnica permite extraer características de uso general con la exploración de unos pocos hiperparámetros necesarios para ajustar los coeficientes al problema de clasificación planteado. Estudios previos muestran un excelente desempeño en señales de audio utilizando coeficientes *scattering* y máquinas de vectores de soporte con *kernel* Gaussiano [116]. Aunque este enfoque perdería gran parte de la interpretabilidad que ofrece la propuesta presentada en el Capitulo 6, tendría como beneficio el uso de un único tipo de características, lo que haría más fácil su extracción en la práctica.

Considerando la posibilidad de un cambio de paradigma en el futuro, la clasificación en tres tipos podría reemplazarse por un criterio más objetivo y basado en un modelo computacional. Un problema crucial de los algoritmos ajustados bajo el paradigma de aprendizaje supervisado es que el sistema adquiere los sesgos del observador que clasificó las señales utilizadas como referencia. Dado que el objetivo último de la clasificación en tres tipos es la discriminación de aquellas señales que cumplen la hipótesis de periodicidad de aquellas que no la cumplen, una forma de estudiar esta situación de forma directa consiste en el uso de modelos de señales periódicas. Modelos de este tipo podrían ser la clásica serie de Fourier, así como también la transformada periódica basada en series de Ramanujan [183, 184]. También existen modelos de señales aproximadamente periódicas

como *wave-shape function* [132]. Seleccionado el modelo, sería posible evaluar la bondad de ajuste de este y utilizarla como criterio de discriminación.

La interpretación del jitter como la variación total de una aproximación del periodo fundamental instantáneo constituye un esfuerzo para expandir la definición de jitter desde un punto de vista matemático, apoyado en el uso de técnicas de modelado de señales más modernas. Así, se formalizó un método para la estimación de jitter relativo a partir de una versión de tiempo continuo del periodo fundamental instantáneo, en lugar de la versión discreta que ofrece la serie de periodos. Esta idea podría generalizarse a otras medidas de jitter (ver Sección 2.5.2), o a otras medidas de perturbación (como el shimmer).

En el futuro se llevarán a cabo comparaciones adicionales entre el método propuesto y otros algoritmos modernos para la estimación de jitter [185], como así también evaluaciones sobre conjuntos de señales de voz reales para avanzar en su validación clínica. Algunos detalles a tener en cuenta serían los casos de diplofonía (ver Sección 2.6.1) o de señales cuyo modo dominante, esto es, el de mayor energía, no es el de frecuencia más baja, ya que en esos casos la detección de la cresta necesaria para la estimación de $F_0(t)$ y su derivada es dificultosa. Una posible solución consiste en la aplicación de una técnica de filtrado inverso, con el objetivo de cancelar el efecto del tracto vocal. También podría explorarse el uso de técnicas más modernas como la TFTC *de-shape* [132], que permite obtener la cresta del primer modo de manera más directa. Una vez superadas las dificultades mencionadas, sería posible evaluar la estimación de jitter mediante la técnica propuesta en el contexto de la separación en los tres tipos de voces. Los resultados de la selección de características (ver Sección 5.5.1) muestran que el jitter relativo calculado con PRAAT no figura entre las características más relevantes para la clasificación. Esto puede deberse a la ya discutida insensibilidad de esa medida para valores de jitter altos. En consecuencia, una medida de jitter capaz de reflejar mayores niveles de jitter podría ser más útil a la hora de realizar la tipificación de señales.

Por otro lado, la estimación de la frecuencia fundamental instantánea a partir de los operadores de FSST4 podría utilizarse para complementar algoritmos basados en la segmentación ciclo a ciclo con el objetivo de lograr métodos híbridos más exactos. Por ejemplo mediante la evaluación de una aproximación de $F_0(t)$ en los puntos fiduciarios calculados con los métodos tradicionales [41]. Asimismo, nuevos avances en el cómputo de los operadores de FSSTN en presencia de ruido podrían traducirse en estimadores de $F_0(t)$ y de jitter aún más robustos.

Apéndice A

A.1. Umbral de la probabilidad a posteriori

Sea X es una variable aleatoria que representa la probabilidad a posteriori (p.p.) obtenida a la salida de una SVM. Suponiendo dos densidades de probabilidad $p_A(X)$ y $p_B(X)$ unimodales, con modas μ_A y μ_B respectivamente, y $\mu_A < \mu_B$.

$p_A(X)$ es la densidad de probabilidad de la p.p. dada una clasificación incorrecta, mientras que $p_B(X)$ es la densidad de la p.p. dada una clasificación correcta (ambas son densidades de probabilidad condicional, teniendo en cuenta que ya se conoce si la señal fue correcta o incorrectamente clasificada). Se buscará un umbral c con el objetivo de minimizar la probabilidad de que una p.p mayor a c se corresponda a una señal mal clasificada, y maximizar la probabilidad de que corresponda a una señal correctamente clasificada.

Esto implica encontrar un valor de c que satisfaga:

$$\arg\min_c P_A(X > c), \tag{A.1}$$

y

$$\arg\max_c P_B(X > c), \tag{A.2}$$

donde $P_A(X > c) = \int_c^1 p_A(q)dq$ y $P_B(X > c) = \int_c^1 p_B(q)dq$.

Dado que minimizar $P_A(X > c)$ implica maximizar $P_A(X < c)$ es posible encontrar un valor de c maximizando la siguiente función:

$$W(c) = P_A(X < c) + P_B(X > c) = P_A(X < c) + 1 - P_B(X < c) \tag{A.3}$$

Derivando $W(c)$ con respecto a c e igualando a cero, encontramos el punto crítico en el que $W(c)$ alcanza un máximo como:

$$p_A(c) = p_B(c) \tag{A.4}$$

es decir, aquel donde ambas densidades se intersecan.

A.2. Demostración de la Proposición 6.2.1

Demostración. Sean t_k los valores t donde el chirp lineal $x(t) = \cos(2\pi\phi(t))$, con $\phi(t) = \alpha t^2 + \beta t$, alcanza un máximo local como se describe en la Ecuación (6.23). Esto se desprende de que los puntos para los que $x(t)$ alcanza un máximo local satisfacen:

$$\phi(t) = n \; ; n \in \mathbb{Z}, \tag{A.5}$$

y en consecuencia:

$$\alpha t^2 + \beta t - n = 0. \tag{A.6}$$

Luego, los valores de t que satisfacen esta última expresión se describen como:

$$t_n = \frac{-\beta \pm \sqrt{\beta^2 + 4\alpha n}}{2\alpha}, \tag{A.7}$$

aplicando la fórmula resolvente para las soluciones de una ecuación cuadrática.

Considerando esto, para probar la Proposición 6.2.1, tomemos la diferencia entre dos máximos locales sucesivos:

$$(t_{k+1} - t_k) = \frac{\sqrt{\beta^2 + 4\alpha(k+1)} - \sqrt{\beta^2 + 4\alpha k}}{2\alpha}. \tag{A.8}$$

Ahora, sea $\phi'(t) = 2\alpha t + \beta$, y $s(t) = \dfrac{1}{\phi'(t)}$. Entonces:

$$s(t_k) = \frac{1}{\sqrt{\beta^2 + 4\alpha k}} \tag{A.9}$$

Demostremos la desigualdad $(t_{k+1} - t_k) < s(t_k)$ en primer lugar

$$\sqrt{\beta^2 + 4\alpha(k+1)}\sqrt{\beta^2 + 4\alpha k} - (\beta^2 + 4\alpha k) < 2\alpha \tag{A.10}$$

$$\left(\beta^2 + 4\alpha k\right)^2 + 4\alpha\left(\beta^2 + 4\alpha k\right) < \left(2\alpha + \beta^2 + 4\alpha k\right)^2 \tag{A.11}$$

$$0 < \alpha^2. \tag{A.12}$$

Ahora probemos que $s(t_{k+1}) < (t_{k+1} - t_k)$:

$$\sqrt{\beta^2 + 4\alpha(k+1)}\sqrt{\beta^2 + 4\alpha k} < \left(\beta^2 + 4\alpha(k+1)\right) - 2\alpha \tag{A.13}$$

$$\left(\beta^2 + 4\alpha k\right)^2 + 4\alpha\left(\beta^2 + 4\alpha k\right) < \left(\beta^2 + 4\alpha k + 2\alpha)\right)^2 \tag{A.14}$$

$$0 < \alpha^2. \tag{A.15}$$

lo que finaliza la demostración. □

A.3. Determinación de σ para la extracción de modo

Considerando a $x(t)$ un tono ligeramente perturbado, su TFTC sera aproximadamente:

$$V_x^g(t, f) = x(t)\hat{g}(t - \phi'(t)), \tag{A.16}$$

donde $\hat{g}(f)$ es

$$\hat{g}(f) = e^{-\sigma^2 \pi f^2}, \tag{A.17}$$

teniendo en cuenta la ventana Gaussiana:

$$g(t) = \frac{1}{\sigma} e^{-\frac{\pi}{\sigma^2} t^2}. \tag{A.18}$$

De esta forma, el ancho en frecuencia del dominio de la señal en el plano tiempo-frecuencia será aproximadamente el de $\hat{g}(f)$. El desvío estándar de $\hat{g}(f)$ estará dado por:

$$\begin{aligned} 2\mathrm{std}_{\hat{g}}{}^2 &= \frac{1}{\pi\sigma^2} \\ \mathrm{std}_{\hat{g}} &= \frac{1}{\sqrt{2\pi}\sigma} \\ &. \end{aligned} \tag{A.19}$$

y considerando su ancho efectivo como $6std_{\hat{g}}$, podemos despejar el valor de σ para que el dominio de $x(t)$ en el plano tiempo-frecuencia quede confinado a una franja de ancho aproximadamente igual a $\bar{F}_0 = \frac{1}{\bar{T}_0}$:

$$\begin{aligned} \frac{6}{\sqrt{2\pi}\sigma} &= \frac{1}{\bar{T}_0} \\ \frac{6\bar{T}_0}{\sqrt{2\pi}} &= \sigma, \end{aligned} \tag{A.20}$$

donde $\bar{T}_0$ es estimado previamente.

A.4. Jitter de variación total para un *chirp* lineal

Sea un chirp lineal $x(t) = \cos(2\pi\phi(t))$, con $\phi(t) = \alpha t^2 + \beta t$, entonces:

$$F_0(t) = \phi'(t) = 2\alpha t + \beta \tag{A.21}$$

y

$$F_0'(t) = \phi''(t) = 2\alpha \tag{A.22}$$

Luego, reemplazando en la Ecuación (6.11) y considerando $\alpha > 0$ podemos calcular el jitter de un chirp lineal como:

$$\begin{aligned} jitter^{TV} &\approx \frac{1}{L - \bar{T}_0} \int_0^L \left| \frac{2\alpha}{(2\alpha t + \beta)^2} \right| \, dt \times 100\,\% \\ &\approx \frac{1}{L - \bar{T}_0} \int_0^L \frac{2\alpha}{(2\alpha t + \beta)^2} \, dt \times 100\,\% \\ &\approx \frac{1}{L - \bar{T}_0} \int_0^L \frac{u'}{u^2} \frac{du}{u\prime} \times 100\,\% \end{aligned} \tag{A.23}$$

con $u = 2\alpha t + \beta$, luego:

$$\begin{aligned} jitter^{TV} &\approx \frac{1}{L - \bar{T}_0} \int_0^L \frac{1}{u^2} \, du \times 100\,\% \\ &\approx \frac{1}{L - \bar{T}_0} \left[\frac{-1}{2\alpha t + \beta} \right]_0^L \times 100\,\% \\ &\approx \frac{1}{L - \bar{T}_0} \left[\frac{1}{\beta} - \frac{1}{2\alpha L + \beta} \right] \times 100\,\%. \end{aligned} \tag{A.24}$$

A.5. Listado de señales reales utilizadas en el Capítulo 6

#	Nombre de la señal	Jitter de referencia (de EGG con PRAAT)	Jitter (por método de VT)	Jitter (por PRAAT)	#	Nombre de la señal	Jitter de referencia (de EGG con PRAAT)	Jitter (por método de VT)	Jitter (por PRAAT)
1	1041-a_n.wav	0.56	0.44	0.42	62	1483-a_n.wav	0.34	0.28	0.38
2	1043-a_n.wav	0.26	0.24	0.21	63	149-a_n.wav	0.29	0.52	0.35
3	1044-a_n.wav	0.39	0.29	0.22	64	1502-a_n.wav	0.53	0.33	0.45
4	1050-a_n.wav	0.64	0.31	0.38	65	154-a_n.wav	0.21	0.3	0.23
5	107-a_n.wav	0.51	0.3	0.3	66	1589-a_n.wav	0.25	0.28	0.32
6	1083-a_n.wav	0.28	0.36	0.3	67	1591-a_n.wav	0.33	0.32	0.32
7	1087-a_n.wav	0.43	0.52	0.5	68	1592-a_n.wav	0.17	0.16	0.13
8	1113-a_n.wav	0.27	0.51	0.34	69	1593-a_n.wav	0.37	0.29	0.25
9	1115-a_n.wav	0.22	0.2	0.17	70	1597-a_n.wav	0.27	0.27	0.27
10	1156-a_n.wav	0.33	0.28	0.3	71	1607-a_n.wav	0.33	0.2	0.18
11	1159-a_n.wav	0.63	0.48	0.56	72	1611-a_n.wav	0.69	0.51	0.56
12	118-a_n.wav	0.22	0.34	0.21	73	1615-a_n.wav	0.31	0.29	0.31
13	1187-a_n.wav	0.24	0.28	0.31	74	1628-a_n.wav	0.30	0.28	0.33
14	1196-a_n.wav	0.21	0.22	0.3	75	1640-a_n.wav	0.26	0.32	0.4
15	1198-a_n.wav	0.45	0.18	0.16	76	1645-a_n.wav	0.28	0.28	0.3
16	1204-a_n.wav	0.32	0.18	0.18	77	1646-a_n.wav	0.59	0.52	0.59
17	1219-a_n.wav	0.18	0.21	0.3	78	1648-a_n.wav	0.19	0.25	0.22
18	1222-a_n.wav	0.40	0.32	0.26	79	1663-a_n.wav	0.43	0.44	0.39
19	1230-a_n.wav	0.28	0.35	0.37	80	1664-a_n.wav	0.35	0.24	0.25
20	1241-a_n.wav	0.34	0.29	0.3	81	1680-a_n.wav	0.41	0.34	0.48
21	1244-a_n.wav	0.37	0.25	0.34	82	1681-a_n.wav	0.27	0.31	0.3
22	1260-a_n.wav	0.44	0.54	0.45	83	1684-a_n.wav	0.20	0.24	0.19
23	1265-a_n.wav	0.25	0.21	0.2	84	1685-a_n.wav	0.28	0.23	0.24
24	1266-a_n.wav	0.33	0.5	0.41	85	1690-a_n.wav	0.26	0.19	0.18
25	1272-a_n.wav	0.53	0.31	0.33	86	1694-a_n.wav	0.21	0.17	0.17
26	1275-a_n.wav	0.28	0.36	0.29	87	1742-a_n.wav	0.41	0.43	0.36
27	1276-a_n.wav	0.17	0.27	0.23	88	1754-a_n.wav	0.28	0.49	0.44
28	1281-a_n.wav	0.35	0.24	0.33	89	1759-a_n.wav	0.22	0.19	0.22
29	1282-a_n.wav	0.58	0.39	0.57	90	1761-a_n.wav	0.23	0.3	0.19
30	1283-a_n.wav	0.25	0.24	0.27	91	1790-a_n.wav	0.38	0.29	0.37
31	1294-a_n.wav	0.16	0.28	0.37	92	1834-a_n.wav	0.18	0.18	0.24
32	1296-a_n.wav	0.21	0.22	0.28	93	1862-a_n.wav	0.25	0.24	0.18
33	1297-a_n.wav	0.41	0.38	0.42	94	2243-a_n.wav	0.30	0.38	0.31
34	1298-a_n.wav	0.39	0.42	0.37	95	2360-a_n.wav	0.26	0.25	0.23
35	1299-a_n.wav	0.56	0.56	0.46	96	2363-a_n.wav	0.25	0.21	0.19
36	1300-a_n.wav	0.33	0.28	0.33	97	2558-a_n.wav	0.33	0.3	0.3
37	1309-a_n.wav	0.37	0.63	0.44	98	2599-a_n.wav	0.53	0.49	0.49
38	1310-a_n.wav	0.53	0.51	0.59	99	349-a_n.wav	0.20	0.28	0.23
39	1317-a_n.wav	0.61	0.44	0.26	100	355-a_n.wav	0.31	0.31	0.33
40	1318-a_n.wav	0.20	0.23	0.2	101	366-a_n.wav	0.37	0.21	0.21
41	1323-a_n.wav	0.23	0.25	0.22	102	494-a_n.wav	0.31	0.63	0.36
42	1325-a_n.wav	0.28	0.24	0.27	103	499-a_n.wav	0.22	0.24	0.23
43	1326-a_n.wav	0.26	0.18	0.17	104	627-a_n.wav	0.36	0.45	0.41
44	1378-a_n.wav	0.16	0.24	0.31	105	634-a_n.wav	0.16	0.2	0.19
45	1379-a_n.wav	0.30	0.23	0.39	106	665-a_n.wav	0.32	0.4	0.29
46	1388-a_n.wav	0.34	0.31	0.34	107	666-a_n.wav	0.17	0.34	0.25
47	1394-a_n.wav	0.37	0.29	0.23	108	668-a_n.wav	0.46	0.29	0.41
48	1395-a_n.wav	0.23	0.18	0.2	109	673-a_n.wav	0.73	0.69	0.51
49	140-a_n.wav	0.43	0.35	0.43	110	693-a_n.wav	0.34	0.18	0.18
50	1425-a_n.wav	0.18	0.21	0.17	111	712-a_n.wav	0.32	0.34	0.33
51	1439-a_n.wav	0.49	0.42	0.46	112	721-a_n.wav	0.38	0.34	0.44
52	144-a_n.wav	0.21	0.33	0.25	113	817-a_n.wav	0.28	0.14	0.14
53	1441-a_n.wav	0.29	0.28	0.35	114	822-a_n.wav	0.26	0.26	0.27
54	1442-a_n.wav	0.14	0.26	0.24	115	830-a_n.wav	0.53	0.49	0.55
55	1444-a_n.wav	0.37	0.17	0.21	116	853-a_n.wav	0.75	0.36	0.43
56	1454-a_n.wav	0.28	0.32	0.37	117	881-a_n.wav	0.38	0.18	0.2
57	1469-a_n.wav	0.29	0.17	0.28	118	889-a_n.wav	0.35	0.46	0.45
58	1471-a_n.wav	0.44	0.29	0.27	119	890-a_n.wav	0.24	0.24	0.23
59	1476-a_n.wav	0.21	0.17	0.19	120	911-a_n.wav	0.37	0.3	0.38
60	1477-a_n.wav	0.59	0.55	0.55	121	937-a_n.wav	0.44	0.29	0.38
61	1479-a_n.wav	0.32	0.44	0.41	122	939-a_n.wav	0.32	0.33	0.32

Apéndice B

Lista de abreviaturas

LPC	Coeficientes de Predicción Lineal
EGG	Electroglotograma
TFTC	Transformada de Fourier de tiempo corto
MR	Método de reasignación
SST	*Synchrosqueezing Transform*
FSST	*Fourier-based Synchrosqueezing Transform*
FSST2	*Second order Fourier-based Synchrosqueezing Transform*
FSSTN	*High order Fourier-based Synchrosqueezing Transform*
SCVMC	Selección de Características por Vecinos Más Cercanos
SVM	*Support Vector Machine*
ROC	*Receiver Operator Characteristic*
KKT	Condiciones de optimalidad de Karush-Kuhn-Tucker
MEEI	*Massachusetts Eye and Ear Infirmary database*
SVD	*Saarbruecken Voice Database*
HNR	*Harmonics-to-Noise Ratio*
CPP	*Cepstral Prominence Peak*
VNCP	Varianza Normalizada de la Componente Principal
DSVNCP	Desvío Estándar de la Varianza Normalizada de la Componente Principal
VT	Variación Total
ER	Error Relativo
MER	Media del Error Relativo
MAE	Media del valor Absoluto del Error

Bibliografía

[1] M. T. Aguilar, "Descartes y el cuerpo máquina," *Pensamiento. Revista de Investigación e Información Filosófica*, vol. 66, no. 249 S. Esp, pp. 755–770, 2010.

[2] S. Silbernagl y A. Despopoulos, *Fisiología: texto y atlas*. Ed. Médica Panamericana, Buenos Aires, 2008.

[3] B. P. Ingalls, *Mathematical modeling in systems biology: an introduction*. MIT press, Cambridge, 2013.

[4] M. Small, *Applied nonlinear time series analysis: applications in physics, physiology and finance*, vol. 52. World Scientific, Singapur, 2005.

[5] R. Baken y R. Orlikoff, *Clinical Measurement of Speech and Voice*. Speech Science, Singular Thomson Learning, Boston, 2000.

[6] P. Lieberman, "Perturbations in Vocal Pitch," *Journal of the Acoustical Society of America*, vol. 33, no. 5, pp. 597–603, 1961.

[7] I. Titze, *Principles of voice production*. National Center for Voice and Speech, 2000.

[8] M. Latarjet y A. Liard, *Anatomía humana*. No. v. 2 in Anatomía humana, Editorial Médica Panamericana, Buenos Aires, 2006.

[9] Lord Akryl, Jmarchn, "Conducting passages of the human respiratory system." `https://en.wikipedia.org/wiki/File:Illu_conducting_passages.svg`, 2010. [Último Acceso 3 de Mayo de 2021].

[10] Olek Remesz, "Larynx external view." `https://commons.wikimedia.org/wiki/File:Larynx_external_en.svg`, 2008. [Último Acceso 3 de Mayo de 2021].

[11] J. West, *Respiratory Physiology: The Essentials*. Lippincott Williams & Wilkins, Filadelfia, 2012.

[12] H. L. Rufiner, *Análisis y modelado digital de la voz: Técnicas recientes y aplicaciones*. Editorial UNL, Santa Fe, 2009.

[13] I. R. Titze y F. Alipour, *The myoelastic aerodynamic theory of phonation*. National Center for Voice and Speech, 2006.

[14] B. Fritzen, B. Hammarberg, J. Gauffin, I. Karlsson, y J. Sundberg, "Breathiness and insufficient vocal fold closure," *Journal of Phonetics*, vol. 14, no. 3-4, pp. 549–553, 1986.

[15] J. Deller, J. Hansen, y J. Proakis, *Discrete-Time Processing of Speech Signals*. An IEEE Press classic reissue, Wiley, Nueva York, 2000.

[16] J. Schoentgen, "Stochastic models of jitter," *The Journal of the Acoustical Society of America*, vol. 109, no. 4, pp. 1631–1650, 2001.

[17] I. R. Titze, "A four-parameter model of the glottis and vocal fold contact area," *Speech communication*, vol. 8, no. 3, pp. 191–201, 1989.

[18] D. A. Rahn III, M. Chou, J. J. Jiang y Y. Zhang, "Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis," *Journal of Voice*, vol. 21, no. 1, pp. 64–71, 2007.

[19] S. Cho y H. Byeon, "Acoustic characteristics of vowel sounds in patients with amyotrophic lateral sclerosis," *Advanced Science and Technology Letters*, vol. 132, pp. 204–207, 2016.

[20] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, y L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[21] M. S. Holi, "Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks," in *Point-of-Care Healthcare Technologies (PHT), 2013 IEEE*, pp. 76–79, IEEE, 2013.

[22] C. Gupte y S. Gadewar, "Diagnosis of Parkinson's disease using acoustic analysis of voice," *Int J Sci Res Netw Secur Communication*, vol. 5, pp. 14–18, 2017.

[23] S. S. Upadhya, A. Cheeran, y J. Nirmal, "Statistical comparison of jitter and shimmer voice features for healthy and Parkinson affected persons," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*, pp. 1–6, IEEE, 2017.

[24] C. E. Stepp, R. A. Lester-Smith, D. Abur, A. Daliri, J. Pieter Noordzij, y A. A. Lupiani, "Evidence for auditory-motor impairment in individuals with hyperfunctional voice disorders," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 6, pp. 1545–1550, 2017.

[25] N. Jafari, F. Izadi, A. Salehi, P. Dabirmoghaddam, F. Yadegari, A. Ebadi, y S. T. Moghadam, "Objective voice analysis of pediatric cochlear implant recipients and comparison with hearing aids users and hearing controls," *Journal of Voice*, vol. 31, no. 4, pp. 505–e11, 2017.

[26] K. Wu, D. Zhang, G. Lu, y Z. Guo, "Influence of sampling rate on voice analysis for assessment of Parkinson's disease," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1416–1423, 2018.

[27] P. Gillivan-Murphy, N. Miller, y P. Carding, "Voice tremor in Parkinson's disease: an acoustic study," *Journal of Voice*, vol. 33, no. 4, pp. 526–535, 2019.

[28] A. Ma, K. K. Lau, y D. Thyagarajan, "Voice changes in Parkinson's disease: What are they telling us?," *Journal of Clinical Neuroscience*, vol. 72, pp. 1–7, 2020.

[29] R. Chiaramonte y M. Bonfiglio, "Acoustic analysis of voice in bulbar amyotrophic lateral sclerosis: a systematic review and meta-analysis of studies," *Logopedics Phoniatrics Vocology*, vol. 45, no. 4, pp. 151–163, 2020.

[30] K. W. Ruckart, M. E. Moya-Mendez, M. Nagatsuka, J. L. Barry, M. S. Siddiqui, y L. L. Madden, "Comprehensive evaluation of voice-specific outcomes in patients with essential tremor before and after deep brain stimulation," *Journal of Voice*, 2020.

[31] M. Vashkevich y Y. Rushkevich, "Classification of als patients based on acoustic analysis of sustained vowel phonations," *Biomedical Signal Processing and Control*, vol. 65, p. 102350, 2021.

[32] A. Tena, F. Claria, F. Solsona, E. Meister, y M. Povedano, "Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: Diagnostic decision support development study," *JMIR Medical Informatics*, vol. 9, no. 3, p. e21331, 2021.

[33] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, y S. B. Gaigg, "Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis," *Autism Research*, vol. 10, no. 3, pp. 384–407, 2017.

[34] M. Asiaee, A. Vahedian-Azimi, S. S. Atashi, A. Keramatfar, y M. Nourbakhsh, "Voice quality evaluation in patients with COVID-19: An acoustic analysis," *Journal of Voice*, 2020.

[35] D. Chitkara y R. Sharma, "Voice based detection of type 2 diabetes mellitus," in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, pp. 83–87, IEEE, 2016.

[36] D. A. M. Ramírez, V. M. V. Jiménez, X. H. López, y P. A. Ysunza, "Acoustic analysis of voice and electroglottography in patients with laryngopharyngeal reflux," *Journal of Voice*, vol. 32, no. 3, pp. 281–284, 2018.

[37] K. Lopez-de Ipina, P. Calvo, M. Faundez-Zanuy, P. Clave, W. Nascimento, U. Martinez de Lizarduy, D. Alvarez, V. Arreola, O. Ortega y J. Mekyska, *et al.*, "Automatic voice analysis for dysphagia detection," *Speech, Language and Hearing*, vol. 21, no. 2, pp. 86–89, 2018.

[38] A. Behrman, C. J. Agresti, E. Blumstein, y N. Lee, "Microphone and Electroglottographic Data from Dysphonic Patients: Type 1, 2 and 3 Signals," vol. 12, no. 2, pp. 249–260, 1998.

[39] Y.-C. Kao, S.-H. Chen, Y.-T. Wang, P.-Y. Chu, C.-T. Tan, y W.-Z. D. Chang, "Efficacy of voice therapy for patients with early unilateral adductor vocal fold paralysis," *Journal of Voice*, vol. 31, no. 5, pp. 567–575, 2017.

[40] M. Hirano, *Clinical examination of voice*, vol. 5. *Springer*, Nueva York, 1981.

[41] R. M. Roark, "Frequency and Voice: Perspectives in the Time Domain," *Journal of Voice*, vol. 20, no. 3, pp. 325–354, 2006.

[42] I. R. Titze, *Workshop on acoustic voice analysis: Summary statement*. National Center for Voice and Speech, 1995.

[43] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, y G. S. Berke, "A comparison of voice analysis systems for perturbation measurement," *The Journal of the Acoustical Society of America*, vol. 93, no. 4, pp. 2337–2337, 1996.

[44] Y. Maryn, P. Corthals, M. De Bodt, P. Van Cauwenberge, y D. Deliyski, "Perturbation measures of voice: a comparative study between multi-dimensional voice program and PRAAT," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 4, pp. 217–226, 2009.

[45] A. Lovato, W. De Colle, L. Giacomelli, A. Piacente, L. Righetto, G. Marioni, y C. de Filippis, "Multi-dimensional voice program (MDVP) vs PRAAT for assessing euphonic subjects: a preliminary study on the gender-discriminating power of acoustic analysis software," *Journal of Voice*, vol. 30, no. 6, pp. 765–e1, 2016.

[46] C. R. Watts, S. N. Awan, y Y. Maryn, "A comparison of cepstral peak prominence measures from two acoustic analysis programs," *Journal of Voice*, vol. 31, no. 3, pp. 387–e1, 2017.

[47] S. Vaz-Freitas, P. M. Pestana, V. Almeida, y A. Ferreira, "Acoustic analysis of voice signal: Comparison of four applications software," *Biomedical Signal Processing and Control*, vol. 40, pp. 318–323, 2018.

[48] K. Richardson, D. Matheron, V. Martel-Sauvageau, y I. Vincent, "A comparative normative study between multidimensional voice program, PRAAT, and TF32," *Perspectives of the ASHA Special Interest Groups*, vol. 4, no. 3, pp. 563–573, 2019.

[49] P. Boersma and V. van Heuven, "Speak and unspeak with PRAAT," *Glot International*, vol. 5, no. 9-10, pp. 341–347, 2001.

[50] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, y R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," 2009.

[51] J. Y. Lee, "Sample selection approach using moving window for acoustic analysis of pathological sustained vowels according to signal typing," *Phonetics and Speech Sciences*, pp. 99–108, 2011.

[52] S. H. Choi, Y. Zhang, J. J. Jiang, D. M. Bless, y N. V. Welham, "Nonlinear dynamic-based analysis of severe dysphonia in patients with vocal fold scar and sulcus vocalis," *Journal of Voice*, vol. 26, no. 5, pp. 566–576, 2012.

[53] C. Fabris, W. De Colle, y G. Sparacino, "Voice disorders assessed by (cross-) Sample Entropy of electroglottogram and microphone signals," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 920–926, 2013.

[54] S. H. Choi y C.-H. Choi, "The Utility of Perturbation, Non-linear dynamic, and Cepstrum measures of dysphonia according to Signal Typing," *Phonetics and Speech Sciences*, vol. 6, no. 3, pp. 63–72, 2014.

[55] C. T. Herbst, "Glottal efficiency of periodic and irregular in vitro red deer voice production," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 724–733, 2014.

[56] D. Stone, P. McCabe, C. E. Palme, R. Heard, C. Eastwood, F. Riffat, y C. Madill, "Voice outcomes after transoral laser microsurgery for early glottic cancer - Considering signal type and smoothed cepstral peak prominence," *Journal of Voice*, vol. 29, no. 3, pp. 370–381, 2015.

[57] S. V. Freitas, P. M. Pestana, V. Almeida, y A. Ferreira, "Integrating voice evaluation: correlation between acoustic and audio-perceptual measures," *Journal of Voice*, vol. 29, no. 3, pp. 390–e1, 2015.

[58] B. Barsties y V. Latoszek, "Treatment Effectiveness of Novafon Local Vibration Voice Therapy for Dysphonia Treatment," *Journal of Voice*, 2018.

[59] Y. Zhang y J. J. Jiang, "Nonlinear dynamic analysis in signal typing of pathological human voices," *Electronics Letters*, vol. 39, no. 7, pp. 1021–1023, 2003.

[60] L. Lin, W. Calawerts, K. Dodd, y J. J. Jiang, "An Objective Parameter for Quantifying the Turbulent Noise Portion of Voice Signals," *Journal of Voice*, vol. 30, no. 6, pp. 664–669, 2015.

[61] J. Y. Lee, "Parameter estimations for signal type classification of Korean disordered voices," *International Journal of Engineering and Technology*, vol. 7, no. 6, pp. 1977–1988, 2016.

[62] W. M. Calawerts, L. Lin, J. Sprott, y J. J. Jiang, "Using rate of divergence as an objective measure to differentiate between voice signal types based on the amount of disorder in the signal," *Journal of Voice*, vol. 31, no. 1, pp. 16–23, 2016.

[63] B. Liu, E. Polce, y J. Jiang, "An objective parameter to classify voice signals based on variation in energy distribution," vol. 33, no. 5, pp. 591–602, *Journal of Voice*, 2018.

[64] B. Liu, E. Polce, y J. Jiang, "Application of local intrinsic dimension for acoustical analysis of voice signal components," *Annals of Otology, Rhinology & Laryngology*, vol. 127, no. 9, pp. 588–597, 2018.

[65] B. Liu, E. Polce, J. C. Sprott, y J. J. Jiang, "Applied chaos level test for validation of signal conditions underlying optimal performance of voice classification methods," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 5, pp. 1130–1139, 2018.

[66] B. Liu, E. Polce, H. Raj, y J. Jiang, "Quantification of voice type components present in human phonation using a modified diffusive chaos technique," *Annals of Otology, Rhinology & Laryngology*, 2019.

[67] C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, y P. Dejonckere, "Validity of jitter measures in non-quasi-periodic voices. Part II: The effect of noise," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 78–89, 2011.

[68] P. Dejonckere, J. Schoentgen, A. Giordano, S. Fraj, L. Bocchi, y C. Manfredi, "Validity of jitter measures in non-quasi-periodic voices. Part I: Perceptual and computer performances in cycle pattern recognition," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 70–77, 2011.

[69] P. H. Dejonckere, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, y C. Manfredi, "To what degree of voice perturbation are jitter measurements valid? A novel approach with synthesized vowels and visuo-perceptual pattern recognition," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 37–42, 2012.

[70] C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, y P. H. Dejonckere, "Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools," *Biomedical Signal Processing and Control*, vol. 7, no. 4, pp. 409–416, 2012.

[71] N. B. Pinto y I. R. Titze, "Unification of perturbation measures in speech signals," *Journal of Acoustic Society of America.*, pp. 1278–1289, 1990.

[72] P. Boersma, "Should jitter be measured by peak picking or by waveform matching?," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 5, pp. 305–308, 2009.

[73] J. Gómez-García, L. Moro-Velázquez, y J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, pp. 181–199, 2019.

[74] J. Gómez-García, L. Moro-Velázquez, y J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.

[75] D.-H. Pham y S. Meignen, "High-order synchrosqueezing transform for multicomponent signals analysis—with an application to gravitational-wave signal," *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3168–3178, 2017.

[76] R. B. Fujiki y S. L. Thibeault, "Examining relationships between grbas ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders," *Journal of Voice*, 2021.

[77] J. A. Gomez García, L. Moro-Velázquez, J. Mendes-Laureano, G. Castellanos-Domínguez, y J. I. Godino-Llorente, "Emulating the Perceptual Capabilities of a Human Evaluator to map the GRB Scale for the Assessment of Voice Disorders," *Engineering Applications of Artificial Intelligence*, vol. 82, pp. 236–251, 2019.

[78] A. E. Stassi, G. A. Alzamendi, G. Schlotthauer, y M. E. Torres, "Vocal fold activity detection from speech related biomedical signals: a preliminary study," in *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*, pp. 520–523, Springer, 2015.

[79] G. A. Alzamendi y G. Schlotthauer, "Modeling and joint estimation of glottal source and vocal tract filter by state-space methods," *Biomedical Signal Processing and Control*, vol. 37, pp. 5–15, 2017.

[80] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, y V. Aharonson, "Sars-Cov-2 detection from voice," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 268–274, 2020.

[81] D. H. Milone, H. L. Rufiner, R. C. Acevedo, L. E. Di Persia, y H. M. Torres, *Introducción a las Señales y los Sistemas Discretos*. EDUNER, Paraná, 2006.

[82] P. Murphy y O. Akande, "Cepstrum-based harmonics-to-noise ratio measurement in voiced speech," *Nonlinear Speech Modeling and Applications*, vol. 3445, pp. 199–218, 2005.

[83] P. J. Murphy, "On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals.," *The Journal of the Acoustical Society of America*, vol. 120, no. 5 Pt 1, pp. 2896–2907, 2006.

[84] G. Fant, *Acoustic theory of speech production*. No. 2, Walter de Gruyter, 1970.

[85] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[86] A. V. Oppenheim, R. W. Schafer, y J. R. Buck, "Tratamiento de señales en tiempo discreto," *Pearson en Español*, Ciudad de México, 2011.

[87] D. Childers, D. Hicks, G. Moore, L. Eskenazi, y A. Lalwani, "Electroglottography and vocal fold physiology," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 2, pp. 245–254, 1990.

[88] I. R. Titze, "Interpretation of the electroglottographic signal," *Journal of Voice*, vol. 4, no. 1, pp. 1–9, 1990.

[89] R. H. Colton y E. G. Conture, "Problems and pitfalls of electroglottography," *Journal of Voice*, vol. 4, no. 1, pp. 10–24, 1990.

[90] E. Yumoto, W. J. Gould, y T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[91] I. R. Titze, Y. Horii, y R. C. Scherer, "Some technical considerations in voice perturbation measurements," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 2, pp. 252–260, 1987.

[92] J. Benesty, M. M. Sondhi, y Y. Huang, *Springer handbook of speech processing*. *Springer*, Nueva York, 2007.

[93] I. R. Titze y H. Liang, "Comparison of F_0 extraction methods for high-precision voice perturbation measurements," *Journal of Speech and Hearing Research*, vol. 36, no. 6, pp. 1120–1133, 1993.

[94] P. Milenkovic, "Least mean square measures of voice perturbation," *Journal of Speech, Language, and Hearing Research*, vol. 30, no. 4, pp. 529–538, 1987.

[95] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 344–353, 1963.

[96] M. H. Hecker y E. J. Kreul, "Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 49, no. 4B, pp. 1275–1282, 1971.

[97] R. Deal y F. Emmanuel, "Pitch effects on vowel roughness and spectral noise," *Journal of phonetics*, vol. 21, pp. 250–264, 1978.

[98] Y. Koike, "Application of some acoustic measures for the evaluation of laryngeal dysfunction," *The Journal of the Acoustical Society of America*, vol. 42, no. 5, pp. 1209–1209, 1967.

[99] R. Yamasaki, A. Montagnoli, E. Z. Murano, E. Gebrim, A. Hachiya, J. V. L. da Silva, M. Behlau, y D. Tsuji, "Perturbation measurements on the degree of naturalness of synthesized vowels," *Journal of Voice*, vol. 31, no. 3, pp. 389–e1, 2017.

[100] J. Schoentgen y R. De Guchteneere, "Searching for nonlinear relations in whitened jitter time series," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 2, pp. 753–756, IEEE, 1996.

[101] W. S. Winholtz y L. O. Ramig, "Vocal tremor analysis with the vocal demodulator," *Journal of speech, language, and hearing research*, vol. 35, no. 3, pp. 562–573, 1992.

[102] J. Schoentgen, "Modulation frequency and modulation level owing to vocal microtremor," *The journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 690–700, 2002.

[103] M. E. Torres, G. Schlotthauer, H. Rufiner, y M. Jackson-Menaldi, "Empirical mode decomposition. spectral properties in normal and pathological voices," in *4th European Conference of the International Federation for Medical and Biological Engineering*, pp. 252–255, Springer, 2009.

[104] G. Schlotthauer, "Análisis de señales con descomposición empírica en modos y aplicaciones a la señal de voz," *Tesis de Doctorado*, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, 2010.

[105] P. Flandrin, *Time-frequency/time-scale analysis*. Academic press, 1998.

[106] P. H. Dejonckere y J. Lebacq, "An analysis of the diplophonia phenomenon," *Speech Communication*, vol. 2, no. 1, pp. 47–56, 1983.

[107] P. Aichinger, M. Hagmüller, I. Roesner, W. Bigenzahn, B. Schneider-Stickler, y J. Schoentgen, "Diplophonia disturbs jitter and shimmer measurement," *Folia Phoniatrica et Logopaedica*, vol. 68, no. 1, pp. 22–28, 2016.

[108] B. Barsties, U. Hoffmann, y Y. Maryn, "The evaluation of voice quality via signal typing in voice using narrowband spectrograms," *Laryngo-rhino-otologie*, vol. 95, no. 2, pp. 105–111, 2016.

[109] Y. Zhang, J. J. Jiang, S. M. Wallace, y L. Zhou, "Comparison of nonlinear dynamic methods and perturbation methods for voice analysis," *The Journal of the Acoustical Society of America*, vol. 118, no. 4, pp. 2551–2560, 2005.

[110] A. J. Sprecher, A. E. Olszewski, J. J. Jiang, y Y. Zhang, "Updating signal typing in voice: addition of type 4 signals," *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3710–3716, 2010.

[111] P. Gassberger y I. Procaccia, "Measuring the strangeness of the strange attractor," *Physica D*, vol. 189, 1983.

[112] "Massachusetts eye and ear infirmary, voice disorders database, version. 1.03 (cd-rom)," *Kay Elemetrics Corporation, Lincoln Park, NJ*, 1994.

[113] C. H. Skokos, G. A. Gottwald, y J. Laskar, *Chaos Detection and Predictability*, vol. 915. *Springer*, Nueva York, 2016.

[114] J. Bruna y S. Mallat, "Classification with scattering operators," in *CVPR 2011*, pp. 1561–1566, IEEE, 2011.

[115] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[116] J. Andén y S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[117] V. Papyan, Y. Romano, J. Sulam, y M. Elad, "Theoretical foundations of deep learning via sparse representations: A multilayer sparse model and its connection to convolutional neural networks," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 72–89, 2018.

[118] P. Flandrin, *Explorations in time-frequency analysis*. Cambridge University Press, 2018.

[119] K. Kodera, C. De Villedary, y R. Gendrin, "A new method for the numerical analysis of nonstationary signals," *Physics of the Earth and Planetary Interiors*, vol. 12, no. 2-3, pp. 142–150, 1976.

[120] F. Auger y P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on signal processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[121] I. Daubechies, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets in medicine and biology*, pp. 527–546, 1996.

[122] T. Oberlin, S. Meignen, y V. Perrier, "The fourier-based synchrosqueezing transform," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 315–319, IEEE, 2014.

[123] T. Oberlin, S. Meignen, y V. Perrier, "Second-order synchrosqueezing transform or invertible reassignment? towards ideal time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1335–1344, 2015.

[124] E. Bedrosian, "A product theorem for hilbert transforms," *Proceedings of the IEEE*, vol. 51, no. 5, pp. 868–869, 1963.

[125] I. Daubechies, J. Lu, y H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Applied and computational harmonic analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[126] H.-T. Wu, *Adaptive analysis of complex data sets*. PhD thesis, Princeton University, 2011.

[127] S. Meignen, D.-H. Pham, y S. McLaughlin, "On demodulation, ridge detection, and synchrosqueezing for multicomponent signals," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 2093–2103, 2017.

[128] M. A. Colominas, S. Meignen, y D. H. PHAM, "Fully adaptive ridge detection based on stft phase information," *IEEE Signal Processing Letters*, 2020.

[129] R. A. Carmona, W. L. Hwang, y B. Torrésani, "Characterization of signals by the ridges of their wavelet transforms," *IEEE transactions on signal processing*, vol. 45, no. 10, pp. 2586–2590, 1997.

[130] G. Thakur, E. Brevdo, N. S. Fučkar, y H.-T. Wu, "The synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications," *Signal Processing*, vol. 93, no. 5, pp. 1079–1094, 2013.

[131] A. Stallone, A. Cicone, y M. Materassi, "New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[132] C.-Y. Lin, L. Su, y H.-T. Wu, "Wave-shape function analysis," *Journal of Fourier Analysis and Applications*, vol. 24, no. 2, pp. 451–505, 2018.

[133] D. Poole, A. Mackworth, y R. Goebel, "Computational intelligence,", Oxford University Press, Nueva York, 1998.

[134] T. M. Mitchell *et al.*, *Machine learning*. McGraw-hill, Nueva York, 1997.

[135] C. M. Bishop, "Pattern recognition and Machine Learning", *Springer-Verlag*, Nueva York, vol. 128, 2006.

[136] R. Duda, P. Hart, y D. Stork, *Pattern Classification*. Wiley, Nueva York, 2012.

[137] M. Last, A. Kandel, y O. Maimon, "Information-theoretic algorithm for feature selection," *Pattern Recognition Letters*, vol. 22, no. 6-7, pp. 799–811, 2001.

[138] S. Nakariyakul y D. P. Casasent, "An improvement on floating search algorithms for feature subset selection," *Pattern Recognition*, vol. 42, no. 9, pp. 1932–1940, 2009.

[139] J. Schenk, M. Kaiser, y G. Rigoll, "Selecting features in on-line handwritten whiteboard note recognition: Sfs or sffs?," in *2009 10th international conference on document analysis and recognition*, pp. 1251–1254, IEEE, 2009.

[140] W. Yang, K. Wang, y W. Zuo, "Neighborhood component feature selection for high-dimensional data.," *Journal of Computers*, vol. 7, no. 1, pp. 161–168, 2012.

[141] J. D. Kelleher, B. Mac Namee, y A. Dárcy, *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press, Cambridge, 2020.

[142] A. Delorme *Encyclopedia of Medical Devices and Instrumentation*, Capitulo "Statistical Methods", American Cancer Society , Atlanta, 2020.

[143] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[144] C. Cortes y V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[145] V. Vapnik, *The Nature of Statistical Learning Theory*. Information Science and Statistics, *Springer*, Nueva York, 1999.

[146] N. Cristianini y J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, Cambridge, 2000.

[147] J. Nocedal y S. J. Wright, *Sequential quadratic programming*. *Springer*, Nueva York, 2006.

[148] M. S. Bazaraa, H. D. Sherali, y C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, Nueva York, 2013.

[149] E. Castillo, A. J. Conejo, P. Pedregal, R. Garcıa, y N. Alguacil, "Formulación y resolución de modelos de programación matemática en ingenierıa y ciencia.," *Escuela Técnica Superior de Ingenieros Industriales, Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos. Universidad de Castilla La Mancha*, 2002.

[150] H. Bourlard y N. Morgan, "A continuous speech recognition system embedding MLP into HMM" in *Advances in neural information processing systems*, pp. 186–193, 1990.

[151] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[152] H.-T. Lin, C.-J. Lin, y R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine learning*, vol. 68, no. 3, pp. 267–276, 2007.

[153] C.-W. Hsu y C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[154] B. Efron y G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983.

[155] B. Schölkopf, A. J. Smola y F. Bach, *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, Cambridge, 2002.

[156] S. Abe, *Support vector machines for pattern classification*, vol. 2. *Springer*, Nueva York, 2005.

[157] J. F. Restrepo y G. Schlotthauer, "Invariant measures based on the u-correlation integral: An application to the study of human voice," *Complexity*, vol. 2018, 2018.

[158] "Saarbruecken voice database." `http://www.stimmdatenbank.coli.uni-saarland.de/index.php4`. [Último Acceso 3 de Mayo de 2021].

[159] M. Jackson-Menaldi, *La voz normal*. Editorial Médica Panamericana, Buenos Aires, 1992.

[160] M. Jackson-Menaldi, *La voz normal y patológica*. Editorial Médica Panamericana, Buenos Aires, 2019.

[161] J. M. Miramont, J. F. Restrepo, J. Codino, C. Jackson-Menaldi, y G. Schlotthauer, "Voice signal typing using a pattern recognition approach," *Journal of Voice*, 2020.

[162] F. F. Severin, B. Bozkurt, y T. Dutoit, "{HNR} extraction in voiced speech, oriented towards voice quality analysis," *Proc. EUSIPCO*, vol. 5, pp. 1–4, 2005.

[163] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, vol. 17.

[164] B. P. Bogert, M. J. Healy, y J. W. Tukey, "The quefrency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the symposium on time series analysis*, vol. 15, pp. 209–243, chapter, 1963.

[165] A. V. Oppenheim y R. Schafer, "Digital signal processing," *Prentice-Hall, Englewood Cliffs, New Jersey*, vol. 6, pp. 125–136, 1975.

[166] J. Hillenbrand, R. A. Cleveland, y R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[167] A. Alpan, J. Schoentgen, Y. Maryn, F. Grenez, y P. Murphy, "Assessment of disordered voice via the first rahmonic," *Speech Communication*, vol. 54, no. 5, pp. 655–663, 2012.

[168] S. Anand, L. M. Kopf, R. Shrivastav, y D. A. Eddins, "Using pitch height and pitch strength to characterize type 1, 2, and 3 voice signals,"*Journal of Voice*, vol. 35, no. 2, pp. 181–193, 2019.

[169] J. F. Restrepo y G. Schlotthauer, "Automatic estimation of attractor invariants," *Nonlinear Dynamics*, vol. 91, no. 3, pp. 1681–1696, 2018.

[170] H.-t. Wu, "Instantaneous frequency and wave shape functions (i)," *Applied and Computational Harmonic Analysis*, vol. 35, no. 2, pp. 181–199, 2013.

[171] G. James y D. Burley, *Matemáticas avanzadas para ingeniería*. Pearson Educación, Londres, 2002.

[172] D. H. Pham y S. Meignen, "High-Order Synchrosqueezing Transform for Multicomponent Signals Analysis-With an Application to Gravitational-Wave Signal," *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3168–3178, 2017.

[173] J. M. Bland y D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[174] J. M. Bland y D. G. Altman, "Comparing methods of measurement: why plotting difference against standard method is misleading," *The lancet*, vol. 346, no. 8982, pp. 1085–1087, 1995.

[175] M. L. McHugh, "Lessons in biostatistics," *Biochemia Medica*, vol. 19, pp. 120–126, 2009.

[176] H. Šrámková, S. Granqvist, C. T. Herbst, y J. G. Švec, "The softest sound levels of the human voice in normal subjects," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, pp. 407–418, 2015.

[177] D. D. Deliyski, H. S. Shaw, M. K. Evans, y R. Vesselinov, "Regression tree approach to studying factors influencing acoustic voice analysis," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 4, pp. 274–288, 2006.

[178] R. Behera, S. Meignen, y T. Oberlin, "Theoretical analysis of the second-order synchrosqueezing transform," *Applied and Computational Harmonic Analysis*, vol. 45, no. 2, pp. 379–404, 2018.

[179] J. M. Miramont, M. A. Colominas, y G. Schlotthauer, "Voice jitter estimation using high-order synchrosqueezing operators," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[180] R. L. Brennan y D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and psychological measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[181] Z. Xie, C. Gadepalli, J. Farideh, B. M. Cheetham, y J. J. Homer, "Machine learning applied to grbas voice quality assessment," *Advances in Science, Technology and Engineering Systems Journal*, vol. 3, no. 6, pp. 329–338, 2018.

[182] S. Davis y P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[183] P. Vaidyanathan, "Ramanujan sums in the context of signal processing—part i: Fundamentals," *IEEE transactions on signal processing*, vol. 62, no. 16, pp. 4145–4157, 2014.

[184] P. Vaidyanathan, "Ramanujan sums in the context of signal processing—part ii: Fir representations and applications," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4158–4172, 2014.

[185] M. S. Morelli, S. Orlandi, y C. Manfredi, "Biovoice: A multipurpose tool for voice analysis," *Biomedical Signal Processing and Control*, vol. 64, p. 102302, 2021.

Doctorado en Ingeniería
mención Inteligencia Computacional, Señales y Sistemas

Título de la obra:
Aportes al análisis de perturbaciones desde el aprendizaje maquinal y el análisis tiempo-frecuencia

Autor: Juan Manuel Miramont

Lugar: Santa Fe, Argentina

Palabras Claves:

análisis de perturbaciones,
máquinas de vectores de soporte,
operadores de synchrosqueezing,
jitter relativo vocal,