UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química

Modelado estocástico de la fonación y señales biomédicas relacionadas:

Métodos en espacio de estados aplicados al análisis estructural, al modelado de la fonación y al filtrado inverso

Gabriel Alejandro Alzamendi

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

2016

Comisión 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

Santa Fe, 21 de junio de 2016.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Modelado estocástico de la fonación y señales biomédicas relacionadas: Métodos en espacio de estados aplicados al análisis estructural, al modelado de la fonación y al filtrado inverso"***, desarrollada por el Bioing. Gabriel Alejandro ALZAMENDI, 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. JOSE LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Dr. Agustín Gravano (*)

Dr. Humberto Torres

Dr. Leandro Di Persia

Dra. Virginia Ballarín

Santa Fe, 21 de junio de 2016

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. Gastón Schlotthauer
Codirector de Tesis

Dra. María Eugenia Torres
Directora de Tesis

(*) Participó a través de video conferencia

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

Modelado estocástico de la fonación y señales biomédicas relacionadas:
Métodos en espacio de estados aplicados al análisis estructural, al modelado de la fonación y al filtrado inverso

por

Gabriel Alejandro Alzamendi

***Resumen*:** La voz y las señales relacionadas poseen información que permite caracterizar la fonación. El surgimiento de métodos potentes y flexibles para extraer esta información de forma precisa ha permitido una mejor descripción de la fonación, lo que ha resultado beneficioso para la medicina y otras ciencias. La presente Tesis se enmarca en esta problemática. En ésta, se proponen nuevos métodos para el modelado de la fonación, y sus señales relacionadas, basados en el modelado estocástico y el procesamiento de señales.

Se comienza describiendo la anatomía y la fisiología de la fonación en los seres humanos. Esto motiva la exploración de tres señales relacionadas a este proceso: la voz, el electroglotograma y las vibraciones en la piel del cuello. Seguidamente, se estudian los fundamentos del modelado de la fonación, se describe la teoría fuente y filtro, y se introducen los conceptos de función glótica, filtro de tracto vocal y filtrado inverso de la voz.

Luego, se investigan las series de períodos y de amplitudes para una vocal sostenida, y se introducen los conceptos de perturbaciones y fluctuaciones de la voz. Esto sirve de motivación para el desarrollo de un método novedoso para la síntesis de vocales sostenidas con perturbaciones controladas por dos parámetros acústicos importantes para la medicina. Para ello, desarrollamos modelos estocásticos para el Jitter y el Shimmer.

Los métodos en espacios de estados para modelos lineales y gaussianos se introducen a continuación. Se los describe de forma concisa y precisa, manteniendo una perspectiva unificada. Se obtiene así el marco conceptual requerido para los demás aportes de la tesis.

Contemplando lo anterior, desarrollamos el modelado estructural en espacio de estados para series de períodos y de amplitudes. El objetivo de esta novedosa estrategia es explicar estas señales suponiéndolas compuestas por elementos simples con una interpretación directa. Además, permite estudiar las fluctuaciones y las perturbaciones de la voz, bajo la hipótesis de que son procesos estocásticos no estacionarios. Combinando los modelos estructurales con los métodos en espacio de estados, implementamos el análisis estructural para señales reales. Se demuestra que el análisis estructural es adecuado para estudiar las series de períodos y de amplitudes correspondientes a voces normales.

Por otra parte, investigamos nuevos métodos en espacio de estados para el modelado de la fonación y el filtrado inverso. Formulamos una ecuación en diferencias estocástica no estacionaria para la función glótica, contemplando perturbaciones o aperiodicidades en esta señal. A partir de ésta, construimos un modelo de la fonación capaz de representar una voz vocal de forma precisa y flexible. Con este modelo, implementamos un método para la estimación conjunta de la función glótica y del tracto vocal, denominado filtrado inverso en espacio de estados. Este método produce estimaciones precisas de estos dos fenómenos. A su vez, permite describir detalladamente las transiciones rápidas en emisiones vocales.

Los aportes originales de esta Tesis de Doctorado consisten en el método para la síntesis de vocales sostenidas con perturbaciones controladas por parámetros acústicos construido en el Capítulo 4; el análisis estructural para series de períodos y de amplitudes, basado en métodos en espacio de estados, desarrollado en el Capítulo 6; y la ecuación en diferencias para la función glótica, el modelo de la fonación, y el filtrado inverso en espacio de estados propuestos en el Capítulo 7.

Stochastic modeling of phonation and the related biomedical signals:
State space methods applied to structural analysis, phonation modeling, and inverse filtering

by

Gabriel Alejandro Alzamendi

***Abstract*:** The voice and related biomedical signals carry information characterizing the phonation. The accurate extraction of this information using powerful and flexible methods allows for a better description of phonation, becoming beneficial in medicine and other fields. This Thesis aims to address this predicament. Here, new methods for the modeling of voice production and related biomedical signals, combining the stochastic modeling and digital signal processing, are proposed and evaluated.

First of all, the anatomy and physiology of human voice production are described. These ideas motivate the exploration of three related biomedical signals: voice, electroglottography and neck skin vibration. Next, the theoretical basis of digital modeling of voice production are studied, source-filter theory is described in detail, and the main concepts of glottal function, vocal tract filter and voice inverse filtering are introduced.

Period and amplitude series extracted from sustained vowels are investigated, and the concepts of voice perturbations and fluctuations are introduced. Using these ideas, we propose a new sustained vowel synthesis method, considering perturbations controlled with two acoustical parameters relevant in voice therapy. In order to produce the perturbations, we generate and evaluate stochastic models for Jitter and Shimmer.

Next, linear Gaussian state space models and the related methods are introduced. This matter is presented in a clear and concise manner, taken into account a unified perspective. Therefore, the theoretical framework required to support the contributions in this thesis is constructed.

Considering the above, we propose the state space structural models, new tools for studying period and amplitude series. The aim of these models is to represent a signal assuming that it is composed of simple stochastic elements with a straightforward interpretation. Voice perturbations and fluctuations are considered, under the hypothesis that these phenomena behave as non-stationary stochastic processes. Combining the structural models and state space methods, we successfully implement a structural analysis method. It is proved that the proposed method is suitable for processing real period and amplitude series extracted from normal voices.

Furthermore, we propose new state space based methods to perform phonation modeling and voice inverse filtering. We first propose a stochastic non-stationary difference equation of glottal function, taking into account the perturbations and aperiodicities in this signal. According to this, we formulate a phonation model suitable for accurate and flexible vocal sounds representation. Using this model, we implement the state space based inverse filtering, a method for the join optimal estimation of the glottal function and the vocal tract filter. The estimations generated using this technique are highly accurate. Moreover, this method allows to describe the fast transitions in vocal voice in detail.

The original contributions of this Thesis are in the synthesis method for sustained vowel considering controlled perturbations with two acoustical parameters in Chapter 4; the structural analysis for period and amplitude series, based on state space methods, in Chapter 6; and the difference equation of glottal function, the phonation model, and state space based inverse filtering proposed in Chapter 7.

Agradecimientos

Son muchas las personas y las instituciones que han colaborado a lo largo del proceso de formación como doctorando y en el desarrollo de esta Tesis de Doctorado. Por este motivo, deseo comenzar este documento agradeciendo públicamente a todos ellos.

Quisiera comenzar agradeciendo a los Doctores María Eugenia Torres y Gastón Schlotthauer quienes dirigieron esta Tesis de Doctorado. Gracias a ellos descubrí qué implica hacer ciencia y ser un científico en la Argentina. Esto ocasionó que me interesara en primer lugar por esta profesión y que luego decidiera dedicarme a ella.

En segundo lugar, me gustaría dar las gracias a los Doctores Virginia Ballarín, Leandro Di Persia, Agustín Gravano y Humberto Torres, quienes amablemente participaron en la evaluación de esta Tesis de Doctorado. Sus comentarios y devoluciones fueron muy útiles y sirvieron para enriquecer considerablemente este trabajo.

Por otro lado, merece ser destacada la labor desempeñada por la Universidad Nacional del Litoral en su conjunto, y por la Facultad de Ingeniería y Ciencias Hídricas (FICH-UNL) en especial. Estas instituciones públicas brindan a muchas personas, entre las que afortunadamente me cuento, la posibilidad de seguir perfeccionándose y de alcanzar su meta de formarse en el máximo nivel académico, como lo es un Doctorado en Ingeniería.

También, quiero destacar que este trabajo no podría haberse logrado sin la subvención del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), a través de la Beca Interna Doctoral oportunamente otorgada, y del Ministerio de Ciencia, Tecnología e Innovación Productiva, que financiara los diferentes Proyectos de Investigación en los que se enmarcó esta Tesis de Doctorado.

Deseo agradecer también toda la ayuda brindada en estos años por parte de la Universidad Nacional de Entre Ríos y de la Facultad de Ingeniería (FI-UNER) de dicha institución. En particular, quisiera destacar los beneficios otorgados en el marco del "Programa de Apoyo para la Finalización de la Formación de Posgrado para docentes de la UNER".

Me gustaría dar las gracias a los amigos que hice en el Laboratorio de Señales y Dinámicas no Lineales de la FI-UNER. En este lugar descubrí un grupo de personas estupendas con las que, a lo largo de estos años, supimos compartir muy buenos momentos y entablamos las más variadas discusiones. Además, quisiera mencionar a los amigos y colegas de la FI-UNER, con los que es muy grato compartir las actividades diarias.

Una mención especial merecen mis padres Griselda Valente y Horacio Alzamendi, y mis hermanas Lorena y Andrea Alzamendi. Ellos han estado a mi lado incondicionalmente desde siempre, brindándome su sabiduría y su fortaleza, ayudándome así a afrontar los diferentes desafíos de la vida. Todo lo que soy y seré es gracias a ellos. También quiero destacar al resto de mi familia. A mis abuelas, a mis tíos y tías, a mis primos y primas, a los que se han ido y los que afortunadamente están. Desde ya, las palabras no alcanzan para reconocer y agradecerles todo lo que han hecho por mi.

Por último, pero no por ello menos importante, quisiera agradecer profundamente a Melisa Rivero por su cariño incondicional, por su alegría contagiosa, por su cálida compañía y por su eterna paciencia. Todos estos años ha estado presente a mi lado en todo momento, en los buenos y especialmente en los malos. Es una fortuna poder disfrutar la vida juntos. Por todo esto, y por mucho más, voy a estar agradecido siempre.

Gabriel A. Alzamendi
Julio de 2016

Índice general

Índice de figuras

Índice de tablas

Acrónimos

Algunos de los acrónimos seleccionadas corresponden a las expresiones en idioma inglés, por ser la forma más difundida en la bibliografía.

FICH-UNL	Facultad de Ingeniería y Ciencias Hídiricas de la Universidad Nacional del Litoral
FI-UNER	Facultad de Ingeniería de la Universidad Nacional de Entre Ríos
LSyDnL	Laboratorio de Señales y Dinámicas no Lineales
EGG	Electroglotograma
VPC	Vibraciones en la piel del cuello
BDDV	Base de datos de desórdenes de la voz
TFF	Teoría *fuente* y *filtro*
LP	Análisis predictivo lineal
LPC	Codificación predictiva lineal
AR	Modelo autorregresivo
TAR	Modelo autorregresivo variante en el tiempo
ARX	Modelo autorregresivo con entrada externa
TARX	Modelo autorregresivo variante en el tiempo y con entrada externa
ARMA	Modelo autorregresivo y de media móvil
TARMA	Modelo autorregresivo y de media móvil variante en el tiempo
ARMAX	Modelo autorregresivo y de media móvil con entrada externa
TARMAX	Modelo autorregresivo y de media móvil variante en el tiempo y con entrada externa
LF	Modelo de la función glótica propuesto por Liljencrants y Fant
IAIF	Filtrado inverso iterativo y adaptativo
SA	Series de amplitudes
SP	Series de períodos
MVDP	Multi-dimensional voice program
PESQ	Evaluación perceptual de la calidad en el habla
MEEG	Modelos en espacio de estados lineales y gaussianos
EM	Método de optimización por esperanza y maximización
GCI	Instantes de cierre glótico
GOI	Instantes de apertura glótica
OP	Fase de apertura
CP	Fase de cierre

Símbolos

/a/, /e/	Fonemas
f_0	Frecuencia fundamental
T_0	Período fundamental
F_1,F_2,F_3,F_4	Formantes
B_1,B_2,B_3,B_4	Anchos de banda
f_s	Frecuencia de muestreo
s	Señal de voz
v_g	Función glótica
U_g	Flujo de aire que atraviesa la glotis
U_l	Flujo de aire en los labios
Re	Parte real
Im	Parte imaginaria
$p(x,\, y)$	Función de probabilidad conjunta de x e y
$p(x,\, y;\, \mathbf{\Theta})$	Función de probabilidad conjunta de x e y que depende del conjunto de parámetros $\mathbf{\Theta}$
$\mathcal{E}\left\{\bullet\right\}$	Operador esperanza o valor esperado
$\mathcal{L}\big(\mathbf{\Theta}\big)$	Función verosimilitud
$\arg\max_x \ f(x)$	Valor de x que maximiza $f(x)$
$\hat{\mathbf{x}}[n\|n-1]$	Estimación *a priori* de $\mathbf{x}[n]$
$\hat{\mathbf{x}}[n\|n]$	Estimación *a posteriori* de $\mathbf{x}[n]$
$\hat{\mathbf{x}}[n\|N]$	Estimación suavizada de $\mathbf{x}[n]$
$\mathcal{T}_1$	Modelo de nivel local o camino aleatorio
$\mathcal{T}_2$	Modelo de tendencia lineal suave o camino aleatorio integrado
$\mathcal{T}_3$	Modelo de nivel y pendiente lineal locales
AR_ρ	Modelo AR de orden ρ
$a_1, a_2, \ldots, a_\rho$	Coeficientes que forman el modelo AR de orden ρ
$\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_\rho$	Estimaciones de los coeficientes que forman el modelo AR de orden ρ
$a_1[n], a_2[n], \ldots, a_\rho[n]$	Coeficientes que forman el modelo TAR de orden ρ

Capítulo 1

Introducción

La fonación es una de las habilidades que se aprenden más tempranamente durante el desarrollo de un individuo [169]. Su estudio ha despertado un gran interés debido a que el habla es la principal forma de comunicación entre las personas [25, 122]. Las investigaciones en esta materia tienen como principal objetivo el dilucidar de qué forma se expresan en la voz los procesos fisiológicos involucrados.

Desde hace varias décadas, han aparecido diversos métodos para imitar artificialmente el proceso de fonación, los que han sido aplicados exitosamente en disciplinas como telecomunicación, informática, educación, medicina, seguridad y entretenimiento, entre otras [42, 128, 144]. Estos evolucionaron notablemente impulsados por los avances en la teoría de modelización y el desarrollo de nuevas técnicas de procesamiento de señales [27, 133]. Las estrategias para el modelado de la fonación y para el análisis de la señal de voz se basaron, originalmente, en las hipótesis de determinismo, estacionariedad o linealidad [31, 38, 130]. Sin embargo, estos postulados no se cumplen en la realidad, por lo que sólo sirven para obtener aproximaciones simplificadas que permitan el uso de los métodos clásicos.

Es importante señalar que los sistemas biomédicos suelen alejarse más aún del comportamiento ideal ante la ocurrencia de alguna patología. En particular, el aparato fonador sufre alteraciones importantes en su fisiología ocasionadas por patologías vocales [82, 119]. Por esta razón, en la medicina y en la fonoaudiología la voz es un medio indispensable para caracterizar el estado del aparato fonador, como así también para la detección de patologías y el seguimiento de un tratamiento [41, 78, 113]. De lo expuesto, se desprende la necesidad de aplicar al estudio de la fonación modelos y métodos de procesamiento de señales no convencionales, basados en hipótesis más realistas.

Además de la voz, existen otras señales biomédicas ligadas a la fonación que resultan de gran interés para la medicina y la ingeniería. Algunas de ellas se obtienen mediante equipamiento específico y otras se calculan a partir de la señal de voz. En el primero de estos conjuntos encontramos, por ejemplo, el electroglotograma (EGG), las vibraciones en la piel del cuello (VPC), la videolaringoscopía y la videoquimografía [15, 16, 161]. Más adelante en este documento analizaremos el EGG y las VPC, por tratarse de señales que se obtienen con métodos mínimamente invasivos y que portan información importante de la actividad glótica.

Del conjunto de señales que se calculan a partir de la voz, centraremos nuestra atención en la serie de períodos (SP) y en la serie de amplitudes (SA) extraídas de emisiones vocales sostenidas [162, 163]. Nuestro interés radica en que se ha ar-

gumentado que las perturbaciones presentes en la SP, denominadas *Jitter*, y en la SA, conocidas como *Shimmer*, aportan información clínica importante para el diagnóstico de patologías vocales [15, 41, 162]. Por otro lado, se ha demostrado que incorporar la información de la SP y la SA en el proceso de síntesis permite generar voces artificiales con una mejor calidad perceptual [29, 128, 134]. En la práctica médica, existen diferentes parámetros acústicos para cuantificar objetivamente el *Jitter* y el *Shimmer*. Sin embargo, el desempeño de estos parámetros no es adecuado en voces que presentan fuertes aperiodicidades [39, 167], lo que evidencia la necesidad de nuevos y mejores métodos para el estudio de estos fenómenos.

Retomando el modelado de la fonación, es importante comentar que las principales estrategias se basan en la teoría *fuente* y *filtro* (TFF), propuesta originalmente por Fant en la década del 60. Ésta considera a la fonación como un proceso compuesto por dos etapas principales [54]. La primera se centra en la generación de la fuente glótica, que representa el flujo de aire modificado por las cuerdas vocales al atravesar la glotis [5, 174]. La segunda etapa contempla la modulación del flujo de aire por la acción de las cavidades resonantes del tracto vocal y de los labios [42, 128].

La fuente glótica es muy importante en el procesamiento de la señal de voz, ya que se ha demostrado que acarrea información del estado de las cuerdas vocales y de la dinámica glótica [5, 27, 56, 174]. Existen en la literatura diferentes modelos determinísticos que, al aplicarlos en los métodos de síntesis, permiten generar voces artificiales que son percibidas como naturales por un oyente [46, 57, 128]. Desafortunadamente, el sensado de la fuente glótica de un individuo es un proceso sumamente complejo que requiere de instrumentación específica [16]. Por ello, desde hace un tiempo han surgido estrategias para la estimación conjunta del tracto vocal y de la fuente glótica a partir de una señal de voz. Esto se conoce como el *filtrado inverso*, o denominado también *descomposición*, digital de la voz. Estos métodos han evolucionado en el tiempo, surgiendo recientemente las primera implementaciones automáticas con resultados satisfactorios [2, 5, 174].

Es importante señalar que la fonación se caracteriza por exhibir un comportamiento aleatorio altamente no estacionario [151, 163]. Como consecuencia, podemos observar un comportamiento similar en la voz y en otras señales biomédicas relacionadas a este proceso [42, 130, 133]. Al mismo tiempo, todas ellas se presentan normalmente alteradas por información indeseada o ruido proveniente de diversas fuentes [101, 120].

En este contexto, los modelos en espacio de estados resultan muy atractivos para el estudio de los rasgos característicos de la fonación, debido a que son adecuados para representar fenómenos no estacionarios y a que en su definición se contemplan las incertezas o errores existentes [76, 86]. A su vez, existen algoritmos muy potentes, denominados métodos en espacio de estados, que permiten el estudio de sistemas ingenieriles reales, y de sus respectivas señales, guiado por esta clase de modelos [30, 51, 89].

Objetivos

Los trabajos realizados en esta tesis se enmarcan en los objetivos del *Doctorado en Ingeniería* dictado por la Facultad de Ingeniería y Ciencias Hídrica de la Universidad Nacional del Litoral (FICH-UNL), y en la misión del Laboratorio de Señales

y Dinámicas no Lineales (LSyDnL) de la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos (FI-UNER). Los *objetivos generales* estipulados en esta tesis de doctorado son:

- Ayudar a la comprensión de los mecanismos fisiológicos y fisiopatológicos involucrados en los sistemas biomédicos.
- Realizar aportes originales en la temática modelado de sistemas biomédicos, prestando especial atención al aparato fonador.
- Colaborar en la formación de recursos humanos altamente capacitados en las temáticas análisis de señales y modelado de sistemas.

En esta tesis de doctorado se contemplaron, a su vez, los siguientes *objetivos específicos*:

- Estudiar los diferentes procesos fisiológicos involucrados en la fonación, tanto en condiciones normales como ante la presencia de patologías.
- Analizar las estrategias tecnológicas existentes para la exploración y el sensado de los procesos fisiológicos involucrados en la fonación.
- Desarrollar estrategias de síntesis de voz que incorporen diferentes características o parámetros de interés para la medicina.
- Proponer nuevas estrategias, inspiradas desde la bioingeniería y basadas en métodos en espacio de estados, para el análisis y el modelado de las series de períodos y de amplitudes.
- Abordar la problemática del filtrado inverso de la voz aplicando modelos estocásticos no estacionarios y variantes en el tiempo, en el marco de los métodos en espacio de estados.

Aportes originales

Los aportes originales alcanzados durante el desarrollo de esta tesis de doctorado pueden organizarse en tres grandes grupos, tomando en consideración los objetivos específicos estipulados en la sección anterior:

- Se desarrolló e implementó un método de síntesis, capaz de generar vocales sostenidas con niveles controlados de *Jitter* y de *Shimmer* y que además evidencian una alta calidad perceptual, es decir, las voces artificiales se perciben muy similares a casos reales. El método contempla, a su vez, otros fenómenos de interés, tales como el ruido de aspiración y el ruido acústico del medio circundante. Se desarrollaron modelos estocásticos con el propósito de simular el *Jitter* y el *Shimmer*.
- Se propuso un método alternativo para el estudio y el modelado de las SP y las SA basado en el análisis estructural de series temporales estocásticas mediante métodos en espacio de estados. El análisis estructural permite explicar una serie temporal a partir de componentes simples con una interpretación física directa. Además, este método permite procesar una señal real tomando como referencia un modelo estocástico no estacionario.

- Se investigó el filtrado inverso de la voz en el marco de los métodos en espacio de estados. En este contexto, se propuso un modelo en espacio de estados variante en tiempo de la fonación, basado en la TFF. Para ello, se construyó un modelo estocástico de la función glótica inspirado en el modelo determinístico de este mismo fenómeno propuesto por Liljencrants y por Fant. Así, el modelo de la fonación desarrollado permite la estimación conjunta de la información del tracto vocal y de la función glótica, aplicando métodos en espacio de estados.

Justificación del encuadre de la tesis en la mención *Inteligencia Computacional, Señales y Sistemas*

En esta tesis de doctorado se propone ampliar el estado del arte respecto al análisis y el modelado estocástico de la fonación. Para ello, se exploraron temáticas vinculadas a nuevas estrategias para el tratamiento de señales, el modelado de sistemas y procesos biomédicos, la estimación de parámetros a partir de optimización y la resolución de problemas inversos, entre otras. Además, en los enfoques abordados se contemplaron las características de no linealidad, no estacionariedad, aleatoriedad y presencia de fuentes de perturbaciones, las cuales son intrínsecas al proceso de fonación. Todas estas temáticas se encuadran en la mención en Inteligencia Computacional, Señales y Sistemas del *Doctorado en Ingeniería* de la FICH-UNL.

Estructura del documento

El documento de esta tesis de doctorado se estructura de la sigue forma.

En el próximo capítulo, se describen la anatomía y la fisiología del aparato fonador de los seres humanos, como así también los sonidos y fonemas más importantes en el habla. Luego, se exploran tres señales biomédicas relacionadas a la fonación: la voz, el electroglotograma y las vibraciones de la piel del cuello. Como cierre, se comentan las bases de datos empleadas en los estudios llevados a cabo.

El Cap. 3 se dedica a describir el estado del arte en el modelado de la fonación. Se presentan aquí dos teorías para explicar este proceso. En primer lugar, se describe la teoría *mioelástica*, *aerodinámica* y *acústica* y, luego, se introduce la TFF. En lo que resta del capítulo, se profundiza respecto a esta última teoría analizando sus principales componentes. Finalmente, se presentan los conceptos fundamentales de función glótica y de filtrado inverso de la voz.

En el Cap. 4 se estudian y se modelan las perturbaciones en la voz. Se comienza la exposición describiendo las SP y las SA de vocales sostenidas. Luego, se exploran las fluctuaciones y las perturbaciones observadas en estas señales, prestando especial atención a los fenómenos *Jitter* y *Shimmer*. En la segunda parte de este capítulo se presenta el primer aporte original de esta tesis de doctorado. Consiste en el desarrollo de un método de síntesis de vocales sostenidas con niveles controlados de *Jitter* y de *Shimmer*. Para ello, se proponen modelos estocásticos para simular estos fenómenos. Seguidamente, se implementa una estrategia de síntesis incorporando los modelos propuestos, que permite generar vocales sostenidas con perturbaciones simuladas.

En el Cap. 5 se describen los métodos en espacio de estados indispensables para el desarrollo de los demás capítulos. En primer instancia, se definen los modelos

en espacio de estados lineales y gaussianos. A continuación, se presentan diferentes métodos específicos para esta familia de modelos, que permiten estimar información oculta de un proceso de forma óptima. Como cierre, se introducen los conceptos involucrados en la estimación de los parámetros de un modelo en espacio de estados.

El Cap. 6 se dedica a describir el segundo aporte original de esta tesis de doctorado. Consiste en el desarrollo de un método para el análisis estructural de las SP y las SA aplicando métodos en espacio de estados. Se comienza la exposición analizando los antecedentes más importantes en esta materia. Se explican, también, las características más importantes identificadas en casos reales de estas señales. En función de lo anterior, se desarrolla un modelo estructural para las SP y las SA. A continuación, se explica como utilizar los métodos en espacio de estados y el modelo propuesto para implementar el análisis estructural de estas dos señales.

El último de los aportes de esta tesis de doctorado se expone en el Cap. 7. Comprende el modelado estocástico de la fonación y el desarrollo de una estrategia para el filtrado inverso aplicando métodos en espacio de estados. En primer lugar, se propone una ecuación en diferencias estocástica y lineal para la función glótica. Considerando esta estructura, se construye un modelo en espacio de estados para explicar el proceso de fonación. Seguidamente, se describe cómo emplear los métodos en espacio de estados para, a partir de una señal de voz, calcular los parámetros desconocidos del modelo de la fonación y, luego, estimar de forma óptima la información de la función glótica y del tracto vocal.

Finalmente, en el Cap. 8 se presentan las principales conclusiones derivadas de los trabajos presentados en esta tesis de doctorado, y se indican algunas de las direcciones futuras para continuar estas investigaciones.

Capítulo 2

Bases fisiológicas de la fonación

2.1. Introducción

La fonación es el proceso fisiológico a partir del cual se produce la voz. Como todo sonido, la voz es una onda acústica capaz de ser transmitida por un medio, como por ejemplo el aire. Así, durante el habla un individuo, el hablante, le transmite a otro, el oyente, un mensaje codificado en su voz. Esta codificación se produce mediante un proceso complejo, que involucra la acción continua, dinámica y coordinada de las estructuras que forman el aparato fonador. Sin embargo, es también posible obtener otras señales relacionadas a la fonación empleando dispositivos desarrollados desde la ingeniería. De acuerdo a su naturaleza, estas señales biomédicas brindarán información específica de diferentes aspectos de la fonación [16, 163].

En este capítulo introduciremos los conceptos fundamentales, que servirán de base para los aportes desarrollados en esta tesis. Desde ya, no pretendemos realizar aquí una exposición exhaustiva, sino brindar las nociones mínimas necesarias que ayuden a la interpretación del resto del documento. Para profundizar en cualquiera de los tópicos, el lector interesado puede recurrir a la extensa bibliografía específica existente, como por ejemplo [42, 82, 128, 151, 163, 169].

2.2. La fonación en los seres humanos

Dedicaremos esta sección a describir el proceso de fonación en los seres humanos. Para ello, introduciremos las estructuras anatómicas más importantes que conforman el aparato fonador, prestando especial atención a las relaciones físicas y funcionales que existen entre ellas. Seguidamente, describiremos la fisiología de la fonación, indicando las transformaciones que se producen a lo largo del aparato fonador y que culminan con la emisión de los sonidos de la voz al medio circundante.

2.2.1. Anatomía del aparato fonador

Se denomina aparato fonador, o sistema de producción del habla, al conjunto de estructuras anatómicas involucradas en la generación de los diferentes sonidos de la voz [163]. Es importante destacar que el aparato fonador y el sistema respiratorio comparten las mismas estructuras anatómicas y, por esta razón, es común encontrar tratados de anatomía o de fisiología que consideran a la fonación como un proceso propio del sistema respiratorio [169]. Sin embargo, en libros específicos de fonética,

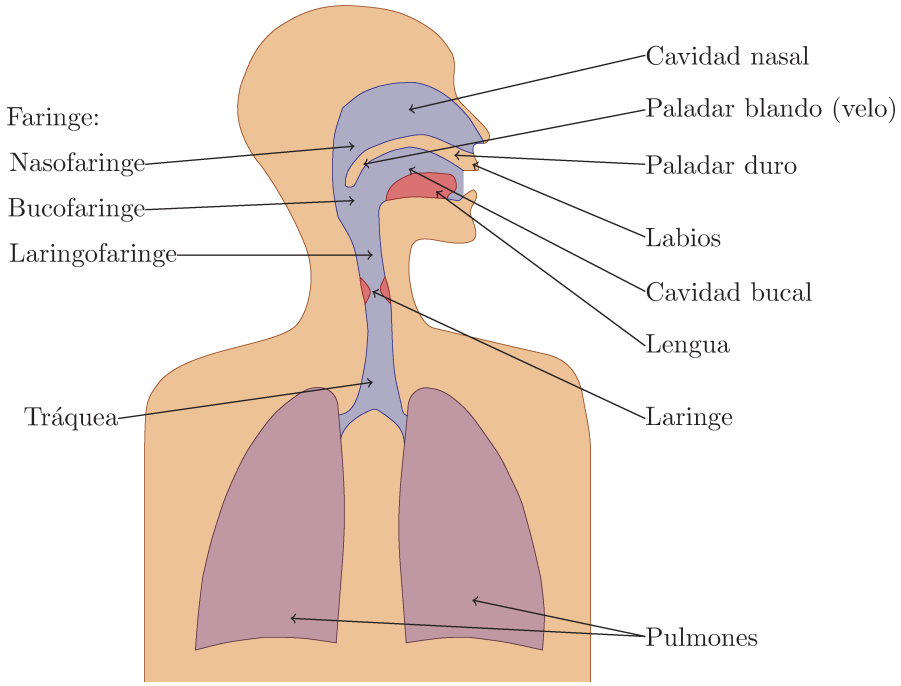


Figura 2.1: Esquema representativo del aparato fonador. Se indican las principales estructuras anatómicas involucradas en el proceso de fonación.

de fonoaudiología, de terapia de la voz y otras ciencias relacionadas, se prefiere la expresión aparato fonador, haciendo énfasis en su relación con el proceso de fonación [25, 64, 151, 163]. Utilizaremos este mismo criterio a lo largo del presente documento.

En la Fig. 2.1 presentamos un esquema del aparato fonador, en el cual se indican las principales estructuras anatómicas que lo conforman. Tomando como referencia a la laringe, agruparemos estos componentes en tres grandes estructuras principales:

- vías subglóticas: formadas por los pulmones, los bronquios y la tráquea;
- la laringe;
- vías supraglóticas: compuestas por la faringe, la cavidad oral, los labios y la cavidad nasal.

Cada uno de estos sistemas posee, a su vez, diferentes estructuras anatómicas y desempeña una función específica. A continuación, describiremos brevemente la anatomía de estos tres sistemas.

Pulmones y tráquea

Los pulmones son órganos pares ubicados en la cavidad torácica, por encima del músculo diafragma y con sus caras anterior, posterior y lateral protegidas por las costillas y los músculos intercostales. Los músculos diafragma e intercostales son los principales responsables de modificar el tamaño de la cavidad torácica durante la ventilación pulmonar, especialmente en la etapa de inspiración.

Dentro de cada pulmón encontramos un árbol bronquial formado por una extensa ramificación de conductos tubulares, desde los bronquios principales derecho e

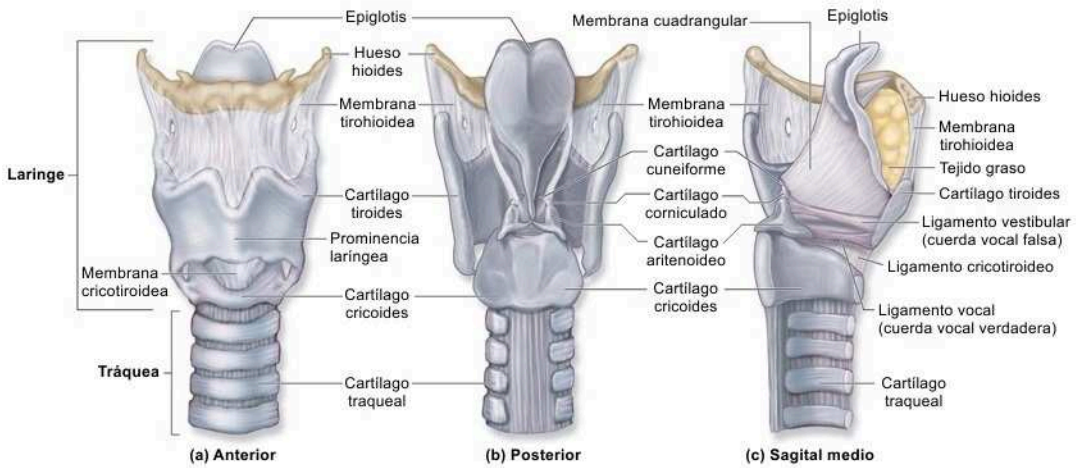


Figura 2.2: Anatomía de la laringe, expuesta a partir de una vista anterior *(a)*, una vista posterior *(b)* y un corte sagital medio *(c)*. Pueden observarse los principales cartílagos, ligamentos y membranas que componen la laringe. Se muestran también las estructuras anatómicas que se comunican directamente con la laringe. Imagen adaptada de `https://rubred.wordpress.com`.

izquierdo, de mayor sección transversal, hasta los conductos más pequeños denominados bronquiolos terminales. Por otro lado, cada uno de los bronquios principales abandona el pulmón por su cara medial y se unifican en el extremo caudal de la tráquea (ver Fig. 2.1). Ubicada por delante del esófago, la tráquea es un conducto tubular que abarca desde la unión de los bronquios principales hasta el cartílago cricoides, en la laringe.

Anatomía de la laringe

La laringe, también llamada órgano vocal, es un sistema complejo formado por cartílagos, músculos y ligamentos. Se ubica en la línea media del cuello, por delante del esófago, y conecta la tráquea con el extremo caudal de la faringe. En la Fig. 2.2 presentamos una vista anterior (izquierda), una vista posterior (centro) y un corte sagital medio (derecha) de la laringe, donde se pueden reconocer las principales estructuras anatómicas que la conforman.

La laringe está compuesta por nueve cartílagos. Tres de ellos son únicos (tiroides, epiglotis y cricoides) y los restantes son pares (aritenoides, cuneiformes y corniculados). Estos cartílagos se unen entre sí gracias a un entramado de ligamentos y membranas, cuya función es controlar la posiciones de estas estructuras durante la fonación. El cartílago tiroides, de mayor tamaño, está compuesto por dos láminas fusionadas generando un ángulo medio anterior, denominado prominencia laríngea. Estas láminas forman la mayor parte de las paredes anterior y laterales de la laringe. A su vez, la prominencia laríngea, también conocida como nuez de Adam, permite localizar fácilmente la laringe por palpación superficial del cuello en hiperextensión [163, 169]. El par de cartílagos aritenoides son muy importantes en la fonación, porque en ellos se fija la parte posterior de cada una de las cuerdas vocales. Profundizaremos sobre esto último, más adelante en esta sección.

Los músculos de la laringe se dividen en extrínsecos, que conectan los cartílagos de la laringe con otras estructuras, y en intrínsecos, que conectan los cartílagos entre

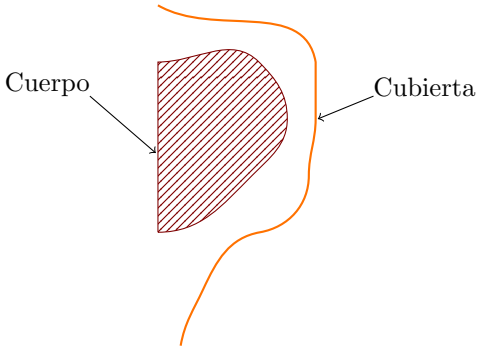


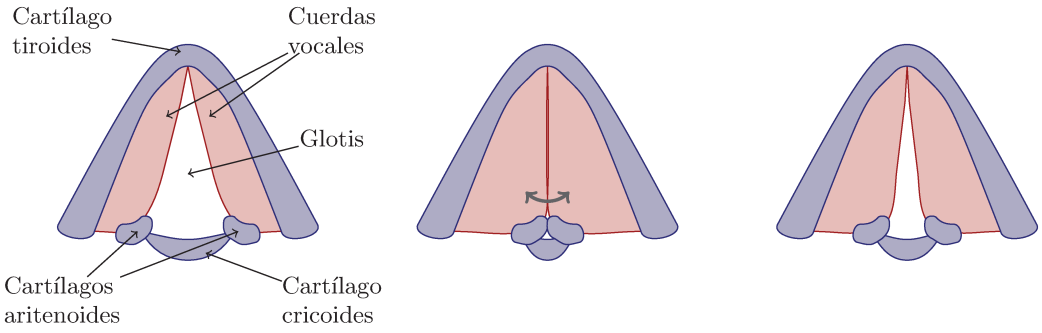
Figura 2.3: Representación de la sección transversal de la cuerda vocal izquierda de acuerdo al esquema *cuerpo* y *cubierta*.

sí [169]. La acción de los diferentes grupos musculares produce cambios en la forma y en la posición de la laringe. En los seres humanos, este fenómeno se evidencia en habla continua, pero no así en emisiones sostenidas y estables [25, 64].

Los órganos de la laringe más importantes para la fonación son los pliegues vocales, o también llamados cuerdas vocales verdaderas. En adelante, nos referiremos a estos simplemente como cuerdas vocales. Se encuentran en la región inferior del cartílago tiroides, donde la cavidad de la laringe es más angosta, y se posicionan paralelos entre sí con orientación anteroposterior. Poseen forma de banda elástica de 1,0-1,5 cm de longitud y un espesor de 2-3 mm, en adultos [151]. Por encima de las cuerdas vocales se encuentran los pliegues vestibulares, o también llamados cuerdas vocales falsas, que no tienen una acción importante en la fonación, sino que participan en el cierre de la cavidad laríngea.

La estructura de las cuerdas vocales puede explicarse fácilmente a partir del esquema *cuerpo* y *cubierta*, propuesta por Titze [155, 165]. Esta construcción es conveniente para explicar y modelar el comportamiento oscilatorio de las cuerdas vocales durante la fonación. En la Fig. 2.3 podemos observar la representación de la cuerda vocal izquierda de acuerdo al esquema *cuerpo* y *cubierta*. De acuerdo a éste, cada cuerda vocal está compuesta por dos capas de tejido con propiedades mecánicas diferentes. Se denomina *cubierta* a la mucosa que reviste las cuerdas vocales y que está compuesta por tejido deformable y no contráctil. Esta capa recubre al *cuerpo*, tejido interno formado por las fibras del músculo tiroaritenoideo (también llamado músculo vocal) y tejido ligamentoso. Retomaremos el estudio del esquema *cuerpo* y *cubierta* en el capítulo 3.

En la Fig. 2.4 representamos, esquemáticamente, la configuración que adoptan las cuerdas vocales en la respiración y en la generación de sonidos sonoros o sordos. La abertura, o hendidura, formada por las caras mediales de las cuerdas vocales recibe el nombre de glotis. Durante la respiración, los cartílagos aritenoides se deslizan lateralmente y los músculos vocales se relajan. Esto ocasiona que las cuerdas vocales se separen, ensanchando la glotis y facilitando el flujo de aire a través de la cavidad laríngea (ver Fig. 2.4.a). Por su parte, en la fonación pueden distinguirse dos situaciones diferentes. En la generación de fonemas sonoros, los cartílagos aritenoides se aproximan medialmente y los músculos vocales se contraen. Como consecuencia, las cuerdas vocales se aproximan entre sí y entran en contacto, cerrando la glotis de forma total o parcial. Por otro lado, para fonemas sordos la configuración es similar



(a) Respiración (b) Sonido sonoro (c) Sonido sordo

Figura 2.4: Configuración de las cuerdas vocales en la respiración *(a)*, y en la generación de sonidos sonoros *(b)*, y sordos *(c)*. Cada esquema representa un corte horizontal de la laringe, a la altura de las cuerdas vocales.

a la adoptada en la respiración, con la diferencia de que las cuerdas vocales están próximas entre sí, sin entrar en contacto, y los músculos vocales se encuentran levemente contraídos (comparar las Figs. 2.4.a y 2.4.c). De lo expuesto hasta aquí, se desprende que tanto la tensión en las cuerdas vocales como el tamaño de la glotis dependerán de la contracción de los músculos vocales y de la configuración de la laringe. En la Sec. 2.2.2 profundizaremos al respecto de los fonemas sonoros y sordos.

Tracto vocal y tracto nasal

Las vías aéreas supraglóticas están formadas, principalmente, por la faringe y las cavidades oral y nasal. La faringe, o garganta, es un conducto en forma de embudo, cuyas principales funciones son permitir el pasaje de aire, en la respiración y en la fonación, y de líquidos y alimentos, en la deglución. Se divide en tres regiones (ver Fig. 2.1):

- Laringofaringe: región inferior ubicada por encima y por detrás de la laringe;
- Bucofaringe: porción central localizada por detrás de la cavidad bucal;
- Nasofaringe región superior ubicada por detrás de la cavidad nasal.

El paladar blando, o velo, se ubica entre la nasofaringe y la bucofaringe, y su posición determina el acople de la cavidad nasal a las vías aéreas supraglóticas [64, 151].

La cavidad bucal, o también denominada comúnmente boca, es la región comprendida entre la bucofaringe y los labios, limitada por las mejillas, el paladar duro, el paladar blando y la lengua. Una práctica muy frecuente, en el estudio de la fonación, consiste en considerar a la bucofaringe y a la cavidad bucal como una única estructura denominada tracto vocal [42, 128, 133]. En adelante, emplearemos este criterio.

Por otro lado, se denomina cavidad nasal a la región interna de la nariz, ubicada por delante de la nasofaringe y por arriba de la cavidad bucal. Se comunica con el exterior a través de las narinas externas, o también llamados orificios nasales. Al igual que en el caso anterior, la nasofaringe y la cavidad nasal suelen agruparse

en una única estructura denominada tracto nasal [42, 128, 133]. Emplearemos esta expresión a lo largo de este documento.

Articuladores

Durante el habla, un individuo modifica continuamente la forma y la longitud del aparato fonador, especialmente en la región del tracto vocal, para producir diferentes sonidos. Esto se logra por el movimiento continuo, controlado y coordinado de las estructuras que lo componen, como por ejemplo el paladar blando, la lengua, los labios y la mandíbula [25, 64, 163]. Estas estructuras móviles reciben el nombre de articuladores. A su vez, se denomina articulación al proceso mediante el cual se modifica la geometría del tracto vocal. La configuración de los articuladores ocasiona que el área de sección transversal, a lo largo del tracto vocal, varíe desde 0, oclusión completa, hasta valores mayores a 20 cm^2 [151].

2.2.2. Fisiología de la fonación

En la fonación, la acción de los músculos intercostales y diafragma ocasiona una reducción en el volumen de la caja torácica. Como consecuencia, ambos pulmones se comprimen, generando así un aumento en la presión del aire almacenado en su interior. Si la presión pulmonar aumenta lo suficiente, el aire es expulsado desde los pulmones hacia las vías aéreas subglóticas. Esto se realiza de forma controlada, con el propósito de mantener estable el proceso a lo largo de una frase o una oración [64, 151]. Se produce así un flujo de aire que atraviesa los bronquios y la tráquea, en dirección a la laringe.

El flujo de aire proveniente de los pulmones es modulado en la laringe, que es la región del aparato fonador donde se genera la mayor resistencia acústica [165]. Además, esta resistencia varía en función de la configuración de la laringe, donde cumplen un papel fundamental las cuerdas vocales [153, 163]. Recordemos que en la fonación existen dos configuraciones principales de las cuerdas vocales, correspondientes a sonidos sonoros o sordos, respectivamente. A continuación, describiremos como influyen estas configuraciones en la modulación del flujo de aire.

La Fig. 2.5 nos ayudará a describir la dinámica oscilatoria de las cuerdas vocales para sonidos sonoros, haciendo uso del esquema *cuerpo* y *cubierta*. Como dijimos anteriormente, en emisiones sonoras las caras mediales de las cuerdas vocales se aproximan entre sí, cerrando la glotis. Esto impide que el aire fluya, por lo que la presión en las vías aéreas subglóticas se equipara a la presión pulmonar. Esta presión actúa sobre las cuerdas vocales, generando fuerzas laterales que favorecen la abducción de las caras mediales, comenzando por la región inferior (Fig. 2.5.a) y alcanzando luego el extremo superior (Fig. 2.5.b). En un instante determinado, se produce la separación total de las cuerdas vocales y la apertura de la glotis, permitiendo así el flujo de aire a través de la laringe (Fig. 2.5.c). Este proceso continúa hasta que las cuerdas vocales alcanzan su máxima separación, instante en que ocurren simultáneamente la máxima apertura de la glotis y el flujo de aire glótico máximo (Fig. 2.5.d).

Las propiedades del tejido interno (músculos vocales y ligamentos) de las cuerdas vocales generan fuerzas elásticas, que actúan en dirección medial y se oponen a la separación de las cuerdas vocales. A su vez, el aumento en el flujo de aire ocasiona una caída en la presión de las vías subglóticas, por lo que se suprimen las fuerzas

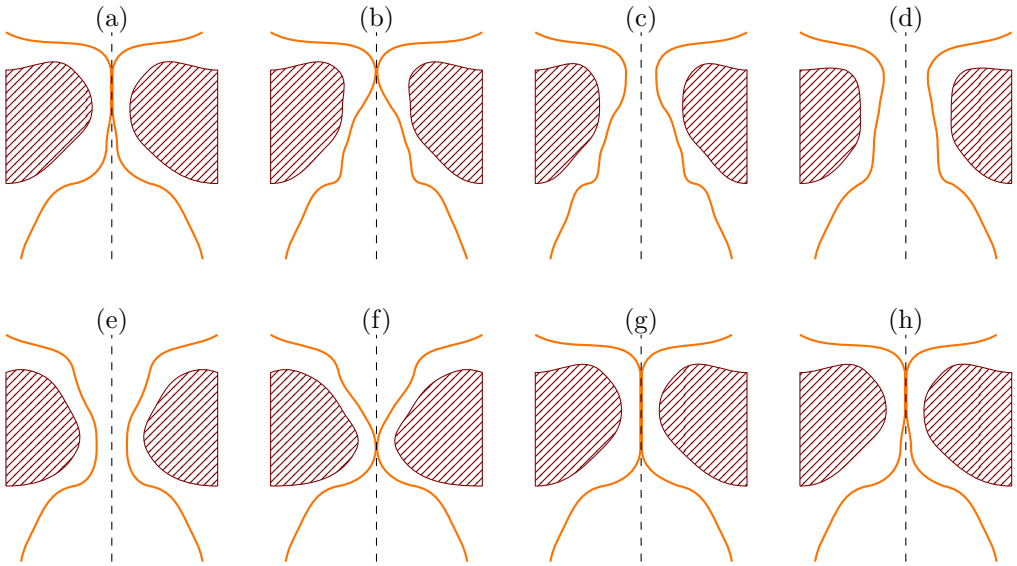


Figura 2.5: Secuencia explicativa de la dinámica glótica. La abducción de las cuerdas vocales comienza por el extremo inferior *(a)*, ésta se transmite al extremos superior *(b)*, hasta alcanzar la separación total *(c)*. La separación continúa hasta que se produce la apertura máxima de la glotis *(d)*. Luego, se genera la aducción de las cuerdas vocales comenzando por el extremo inferior *(e)*, alcanzando luego al extremo superior hasta que se produce la colisión de las cuerdas vocales y el cierre de las glotis *(f)*. Esta configuración se mantiene *(g)*, hasta que se reanuda nuevamente el ciclo *(h)*.

laterales y se interrumpe así la separación de las cuerdas vocales (Fig. 2.5.e). La acción de las fuerzas restaurativas combinado con la disminución de la presión subglótica provoca la aducción de las cuerdas vocales. Luego, se produce la colisión de las cuerdas vocales, comenzando por la región inferior (Fig. 2.5.f) hasta alcanzar el extremo superior (Fig. 2.5.g). En la fase de aducción, el flujo de aire disminuye paulatinamente, hasta que se anula completamente al cerrarse la glotis. Posteriormente, la presión subglótica equipara nuevamente a la presión pulmonar, lo que provoca que inicie un nuevo ciclo (Fig. 2.5.h).

Este comportamiento oscilatorio de las cuerdas vocales se repite de forma regular y es la principal característica distintiva de los fonemas sonoros [25, 64, 151]. En adelante, nos referiremos a este proceso cíclico con la expresión dinámica glótica. Es importante destacar que, además del desplazamiento lateral, se produce una oscilación transversal en las cuerdas vocales. Esto se desprende de que la dinámica del extremo inferior de las cuerdas vocales se produce fuera de fase (se adelanta) con respecto a la región superior. Este fenómeno recibe el nombre de *onda transversal* u *onda de la mucosa* [153, 165], y puede apreciarse claramente en la secuencia mostrada en la Fig. 2.5. Por la acción de las cuerdas vocales, el flujo de aire estable proveniente de los pulmones se convierte en una serie de pulsos o ráfagas de aire regulares con comportamiento cuasiperiódico, es decir, el período es aproximadamente igual para todos los pulsos [151, 153, 163].

El lapso de tiempo en el cual las cuerdas vocales están separadas y existe flujo de aire se denomina *fase de glotis abierta*, mientras que la etapa durante la cual las cuerdas vocales están en contacto y el flujo de aire es nulo se denomina *fase de*

glotis cerrada. La duración de cada ciclo de la dinámica glótica se denomina *período fundamental* de la voz, T_0, mientras que se denomina *frecuencia fundamental* al recíproco de su valor, $f_0 = 1/T_0$. En la literatura, es común el uso de las expresiones *frecuencia fundamental* y *pitch* de forma indistinta, aun cuando en rigor *pitch* se refiere a la altura de un sonido, que es una característica subjetiva [42, 128, 133, 163]. En este documento de tesis, seguiremos este último criterio.

El parámetro f_0 depende de diversos factores, como por ejemplo: el nivel de contracción de los músculos vocales (al aumentar la contracción aumenta f_0), la masa de cada cuerda vocal (al aumentar la masa, disminuye f_0 porque aumenta la inercia), y la presión del aire en los pulmones y en la tráquea (al aumentar la presión de aire aumenta f_0) [25, 64]. En promedio, el rango de f_0 es 60-400 Hz en condiciones normales [42, 133]. Además, habitualmente f_0 es menor en los hombres que en las mujeres, debido a que sus cuerdas vocales son más largas y poseen una mayor masa.

Por otra parte, los sonidos sordos se caracterizan principalmente por la ausencia de oscilaciones en las cuerdas vocales [25, 64]. Como dijimos anteriormente, las cuerdas vocales están próximas entre sí, sin entrar en contacto (ver Fig. 2.4.c). Como consecuencia, se producen turbulencias en el flujo de aire cuando éste atraviesa la glotis. Estas turbulencias ocasionadas por las cuerdas vocales reciben el nombre de *aspiración* [163]. Además, se pueden producir turbulencias de este tipo en cualquier constricción u oclusión formada por los diferentes articuladores a lo largo de las vías aéreas supraglóticas [151].

Los dos flujos de aire glóticos analizados, pulsátil en emisiones sonoras y turbulento en emisiones sordas, se comportan como una fuente sonora. Se genera así una onda acústica en la glotis que se transmite a lo largo del tracto vocal y del tracto nasal, si este último también se encuentra acoplado. Ambas cavidades se comportan como resonadores acústicos, modificando la información espectral de la onda acústica. Este fenómeno se comprende mejor en el dominio de la frecuencia. En primer lugar, consideremos a la fuente sonora como la superposición de tonos puros de diferentes frecuencias. Al mismo tiempo, las vías aéreas supraglóticas presentan un conjunto de frecuencias de resonancia y de antiresonancia, las que dependen de la configuración y de las dimensiones [64, 163]. Como resultado, aquellos tonos que coincidan con las frecuencias de resonancia se amplifican, a la vez que aquellos que coinciden con las frecuencias de antiresonancia se atenúan [42, 151]. Debido a este comportamiento, es común explicar la fisiología del tracto vocal usando como analogía un filtro acústico [101, 163]. Finalmente, la onda acústica es radiada desde los labios al medio circundante, por el cual viaja hasta un oyente.

En habla continua, un sujeto cambia la configuración de sus vías aéreas supraglóticas para generar los diferentes sonidos, gracias al movimiento controlado y sincronizado de los articuladores. Así, esta articulación repercute, entre otros aspectos, en el comportamiento frecuencial del tracto vocal, modificando las frecuencias de resonancia y de antiresonancia.

Mecanismos de excitación

En la fonación, los aspectos que se perciben en los sonidos de la voz dependen, en gran medida, del mecanismo de excitación empleado. En general, podemos distinguir tres mecanismos de excitación principales [42, 133]:

- Un flujo de aire pulsátil, o en forma de ráfaga, con comportamiento cuasiperiódico es generado por las oscilaciones de las cuerdas vocales. Esta excitación es característica de los sonidos sonoros, o también llamados emisiones vocales.
- Constricciones en diferentes regiones del tracto vocal, por ejemplo la glotis, producen turbulencias en el flujo de aire que generan fuentes de ruido acústico de banda ancha. Este mecanismo de excitación es propio de los sonido sordos.
- La oclusión transitoria del tracto vocal produce la interrupción del flujo de aire y el aumento de la presión en ese punto; la posterior apertura de la cavidad permite la rápida liberación de la presión, dando lugar a una fuente sonora con forma de impulso. Esta excitación transitoria produce sonidos plosivos.

Es importante destacar que en la fonación pueden encontrarse otras formas de excitación. Una alternativa, muy frecuente en habla continua, consiste en la ocurrencia simultánea de varias fuentes sonoras al unísono [42, 151].

2.3. Sonidos de la voz y fonemas

A lo largo de la sección anterior estudiamos el proceso de fonación, prestando atención a los principales órganos involucrados y a las transformaciones que se producen. Describiremos aquí los diferentes sonidos empleados para formar las palabras, o significantes, de un lenguaje junto con los atributos que le otorgan sus características perceptuales distintivas.

Se denomina *fonema* a la mínima unidad distintiva, carente de significado, a partir de la cual se construye una palabra [122]. Los fonemas constituyen el conjunto mínimo de unidades necesario para transmitir oralmente todos los significados en un lenguaje, a la vez que permiten clasificar o agrupar los sonidos de la voz. Por si solos no significan nada; pero al combinarlos aplicando las reglas propias de un lenguaje se construyen los significantes en un cierto idioma o lengua [25, 151]. Dos fonemas son distintos si el cambio de uno por otro cambia el significante (por ejemplo, si se modifica un fonema en paso, peso y piso, se obtienen palabras diferentes).

A cada fonema se le asocia un conjunto único de gestos articulatorios, tomando en consideración tanto la naturaleza y ubicación de la fuente de excitación como la configuración y comportamiento del tracto vocal durante la fonación. Se denomina alófono a la realización física de un fonema, es decir, el sonido emitido [122]. Un mismo fonema puede tener asociado un conjunto de alófonos y su ocurrencia dependerá del contexto en que se encuentre el fonema dentro de una palabra [42, 151]. Sin embargo, esta segmentación es sólo teórica ya que, por ser el habla un acto continuo y dinámico, las características de dos alófonos consecutivos se superponen y se mezclan, dificultando su delimitación [25, 64]. Esto repercute, a su vez, en la segmentación de los fonemas [42].

Los sonidos de la voz se representan a partir de alfabetos fonéticos, compuestos por símbolos gráficos para la mayoría de los sonidos de las diversas lenguas del mundo. En la actualidad, se emplea el *Alfabeto Fonético Internacional* (AFI), confeccionado por la Asociación Internacional de Fonética[1]. En la página oficial de este ente (`www.internationalphoneticassociation.org`) se puede obtener la última versión del AFI, acompañada con información complementaria.

[1]En inglés, "International Phonetic Association".

Tabla 2.1: Clasificación de los principales fonemas en el español rioplatense, reproducido de [133].

	Vocales:	/a/, /e/, /i/, /o/, /u/
Consonantes	Fricativas:	/f/, /s/, /θ/, /y/, /x/
	Africativas:	/ʧ/
	Oclusivas:	/p/, /t/, /k/, /b/, /d/, /g/
	Nasales:	/m/, /n/, /ɲ/
	Vibrantes:	/r/, /r̄/
	Laterales:	/l/, /ʎ/

En la Tab. 2.1 presentamos la clasificación de los principales fonemas en el español rioplatense, reproducido de [133]. Esta clasificación se corresponde a su vez con el material desarrollado en [129]. Como se puede apreciar, los sonidos de la voz se dividen fundamentalmente en vocales y consonantes. Es importante destacar que en el español rioplatense existen dos variantes alofónicas para el fonema /e/ ([e] y [ɛ]) y para el fonema /o/ ([o] y [ɔ]), dando lugar así a siete sonidos vocálicos diferentes [64]. Por otro lado, las consonantes se subdividen con el propósito de describirlas de forma detallada. Esto se sustenta tanto en los gestos articulatorios involucrados en su generación como en las características acústicas propias de cada sonido [25, 42, 64, 151]. Desde ya, esta clasificación no pretende ser exhaustiva. Por ejemplo, no se contempla que en el español rioplatense las consonantes sonoras /b/, /d/, /g/ presentan alófonos oclusivos ([b], [d], [g]) y fricativos ([β], [ð], [ɣ]), o que la consonante sonora /y/ presenta alófonos fricativo ([y]) y africado ([ʤ]) [129].

2.3.1. Vocales

En los fonemas vocales, el tracto vocal presenta una configuración relativamente abierta, sin obstrucción oral [64, 129]. Esto se debe a que los articuladores no forman oclusiones o estrechamientos a lo largo del tracto vocal. A su vez, el mecanismo de excitación es siempre de la forma de pulsos cuasiperiódicos, por lo que son emisiones sonoras [163].

Al ser excitado, el tracto vocal actúa como un sistema resonante, caracterizado por cuatro o cinco frecuencias principales de resonancia. Estas frecuencias de resonancia resultan de la forma y la configuración del tracto vocal, y pueden identificarse directamente en el espectro de una vocal, el cual presenta picos o regiones de frecuencias con una mayor amplitud [42, 128]. Las frecuencias de estos picos reciben el nombre de *formantes* [163]. A partir del espectro de una vocal, es posible extraer la información sobre todos los aspectos acústicos relevantes de la configuración del tracto vocal, en un instante determinado. Por otro lado, las propiedades acústicas de las vocales persisten por un tiempo apreciable o cambian muy lentamente, siempre que se mantenga la configuración del tracto vocal. Para profundizar respecto a las propiedades acústicas en fonemas vocales, el lector interesado puede recurrir a [25, 42, 64, 151].

En las vocales, la configuración del tracto vocal queda determinada por la acción de los articuladores, principalmente de la lengua, de la mandíbula y de los labios. En este caso, el paladar cumple un rol secundario. De acuerdo a la teoría articulatoria,

Tabla 2.2: Descripción articulatoria de las cinco vocales neutras en el español rioplatense, reproducido de [64].

Vocal	Localización de la constricción	Grado de constricción	Abertura oral	Acción de los labios
/a/	Faríngea	Amplio	Amplio	Delabializado
/e/	Palatal	Estrecho	Amplio	Delabializado
/i/	Palatal	Estrecho	Reducido	Delabializado
/o/	Velofaríngea	Estrecho	Amplio	Labializado
/u/	Velar	Estrecho	Reducido	Labializado

las vocales pueden describirse a partir de cuatro parámetros [64]:

- **Localización de la constricción:** a lo largo del tracto vocal existen cuatro principales localizaciones de la constricción: en la superficie del paladar duro (vocales *palatales*), en el velo del paladar (vocales *velares*), en la faringe superior (vocales *velofaríngeas*) y en la faringe inferior (vocales *faríngeas*).

- **Grado de constricción:** determina el acoplamiento entre las cavidades anterior y posterior respecto a la constricción. Si la constricción es *amplia*, las dos cavidades resuenan en forma conjuntan; si es *estrecha*, las cavidades influyen poco entre sí.

- **Abertura oral:** las vocales se clasifican según la posición que adopte el maxilar inferior en *abiertas*, abertura bocal más amplia, y en *cerradas*, abertura bocal menor.

- **Acción labial:** en las vocales *labializadas* influyen activamente los labios, a partir de los gestos de redondeado o distendido; mientras que en las vocales *delabializadas* no hay acción labial.

En la Tab. 2.2 presentamos la descripción articulatoria, en función de los parámetros introducidos, de las cinco vocales neutras en el español rioplatense. Esta información se reprodujo de [64].

2.3.2. Consonantes

A diferencia de las vocales, los sonidos consonánticos se producen con una configuración relativamente cerrada del tracto vocal, ocasionada por los mecanismos básicos de cierre y de estrechamiento [25, 64]. Esta configuración repercute directamente en la forma de excitación. En el cierre, el tracto vocal se obstruye transitoriamente y, luego, su rápida liberación genera una ráfaga de aire fugaz. Por otro lado, un estrechamiento a lo largo del tracto vocal ocasiona turbulencias en el flujo de aire, que actúan como fuentes acústicas. En algunos casos, se generan también pulsos glóticos cuasiperiódicos, que actúan como estímulos secundarios en combinación con las excitaciones mencionadas [128, 151].

Tanto el cierre como el estrechamiento se producen en regiones específicas del tracto vocal, gracias a la rápida y geométricamente compleja acción de los articuladores, en especial de la lengua y de los labios [64]. Para ello, los movimientos de

estas estructuras deben desarrollarse de forma coordinada y muy precisa. Como consecuencia, las consonantes presentan propiedades acústicas transitorias fuertemente dependientes del contexto en que se producen [151]. La interacción entre dos sonidos adyacentes se denomina *coarticulación* [42]. Asimismo, en el espectro de una consonante sólo pueden identificarse algunas de las propiedades acústicas que caracterizan la configuración del tracto vocal, en ese instante [128, 133].

A continuación, presentaremos las dos características principales que determinan la naturaleza de una consonante. La primera de éstas, corresponde al *modo de articulación* y permite analizar las consonantes considerando tres propiedades diferentes: [25, 64]:

- **Participación o no de la excitación periódica:** si ésta actúa, la consonante es sonora; si no, es sorda.
- **Transmisión a través de la cavidad nasal u oral:** la consonante es oral si el aire pasa únicamente por la cavidad oral; si pasa también a través de la cavidad nasal, la consonante es nasal.
- **Naturaleza del cierre o estrechamiento:** el mecanismo de cierre o estrechamiento permite la siguiente clasificación (ver Tab. 2.1):
  - **Fricativas:** el flujo de aire es forzado a atravesar una zona estrecha del tracto vocal, localizada en un punto entre la glotis y los labios, generando turbulencias. Se produce así un ruido acústico de banda ancha.
  - **Africativas:** en este tipo de sonido se produce, en primer lugar, un cierre total en algún punto del tracto vocal por un breve lapso de tiempo, seguido luego por una apertura lenta de la cavidad ocasionando un sonido turbulento de duración considerable.
  - **Oclusivas:** se genera un cierre total o parcial en algún punto del tracto vocal, seguido luego por una apertura brusca con liberación de presión que genera un sonido de *explosión*. Se llaman también consonantes *plosivas*.
  - **Nasales:** el velo del paladar ocluye la cavidad oral y direcciona el flujo de aire por la cavidad nasal, generando un murmullo nasal.
  - **Vibrantes:** los articuladores se encuentran relajados y cercanos entre sí; al fluir el aire, se imprimen oscilaciones que los acerca y los aleja alternativamente.
  - **Laterales:** se genera un cierre a lo largo de la línea media del paladar, forzando al aire a fluir por uno o ambos laterales de la lengua.
  - **Semivocales:** el articulador se aproxima a una zona del tracto vocal, pero este estrechamiento es tal que no permite que se produzcan turbulencias.

La característica restante corresponde al *punto de articulación*. De acuerdo a ésta, las consonantes se pueden clasificar como sigue:

- **Bilabiales:** se produce la acción conjunta de ambos labios.
- **Labiodentales:** el labio inferior se aproxima a los incisivos superiores.
- **Dentales:** el ápice de la lengua se coloca detrás de los incisivos superiores.

- **Alveolares:** la lengua se aproxima a la zona alveolar.
- **Palatales:** la lengua se aproxima al paladar duro.
- **Velares:** la parte posterior de la lengua se aproxima al paladar blando.
- **Glotales:** articulación en las cuerdas vocales.

En [25, Pág. 76] se presenta un cuadro con todos los sonidos consonánticos del español, junto con su descripción completa considerando las características introducidas.

2.3.3. Registros vocales

Anteriormente, denominamos sonidos sonoros a aquellos generados a partir de un flujo de aire pulsátil, como principal mecanismo de excitación, producto de la modulación de las cuerdas vocales. Así, el modo de oscilación de las cuerdas vocales influye en las características acústicas percibidas en un sonido sonoro [163]. Los registros vocales permiten comprender los mecanismos de oscilación encontrados normalmente en las cuerdas vocales.

Hasta el momento, no se ha arribado a un consenso, por parte de los especialistas, respecto a la definición y caracterización de los registros vocales. Esto se debe, principalmente, a que los registros vocales se estudian en el marco de la calidad perceptual de la voz, que es un atributo subjetivo, y a que es imposible percibir el punto de inflexión entre los diferentes registros [16, 172]. Otro fenómeno, que influye negativamente, se debe a que en voz hablada y en canto se suelen emplear expresiones distintas para referirse a una misma conducta vocal o, a la inversa, una misma expresión para describir diferentes conductas [163].

De acuerdo a Titze, la expresión registro vocal se emplea para describir aquellos atributos de la calidad vocal, perceptualmente distintos, que se preservan para un rango de frecuencias fundamentales y de intensidades [163]. Para voz hablada, se proponen tres registros vocales fundamentales [16, 172]:

- **Pulsátil o frito:** abarca el rango de fonación correspondiente a bajas frecuencias fundamentales, para el cual la fuente de excitación se percibe pulsátil. En este caso, las cuerdas vocales se presentan relajadas y laxas, a la vez que la presión subglótica es muy baja. La dinámica glótica se caracteriza por una fase abierta larga, seguida por una fase de cierre mucho más corta.
- **Modal o de pecho:** corresponde al rango de la frecuencia fundamental encontrado normalmente en habla y en canto. Comparado con el registro pulsátil, las cuerdas vocales desarrollan un grado de contracción importante y la presión subglótica es mayor. La dinámica glótica se caracteriza por un período muy regular y un ciclo de trabajo mayor o igual al 50 %.
- **Falsette:** abarca las frecuencias fundamentales altas. Las cuerdas vocales presentan una grado de contracción muy alto y una configuración estirada. Únicamente vibra el borde superior de las cuerdas vocales, por lo que no se produce un cierre total de la glotis. La fase abierta abarca la mayor parte o la totalidad de la dinámica glótica.

2.4. Señales biomédicas de la fonación

De lo expuesto hasta aquí, podemos apreciar que, sin lugar a dudas, la gran diversidad de sonidos de la voz es el principal producto de la fonación, el cual es indispensable para la comunicación oral entre las personas. A lo largo del tiempo, han surgido diferentes estrategias para estudiar, por un lado, los sonidos generados y, por otro lado, la fisiología del aparato fonador durante la fonación. Para ello, resulta imprescindible extraer o sensar información que permita describir este proceso. En este capítulo, presentaremos brevemente aquellas señales de la fonación que han sido empleadas en los trabajos realizados por nosotros en el Laboratorio de Señales y Dinámicas no Lineales (LSyDnL).

2.4.1. Señal de voz

Como dijimos anteriormente, los sonidos de la voz consisten en perturbaciones mecánicas, asociadas a cambios locales en la presión o el desplazamiento de las partículas, que se propagan en forma de onda por el aire u otro medio. Estas perturbaciones pueden ser capturadas por un dispositivo transductor, cuya función es convertir la energía mecánica de la onda en otra forma de energía [16, 163]. Aun cuando existen una gran variedad de transductores aptos para esta tarea, los micrófonos son los sensores universalmente empleados para capturar los sonidos de la voz.

Un micrófono es un dispositivo sensible a las ondas sonoras, que genera un correlato eléctrico a partir de una señal de presión acústica. Es decir, transforma la energía acústica en eléctrica. A este correlato eléctrico se lo denomina comúnmente señal de voz [42, 128, 133]. La señal obtenida directamente de un micrófono difícilmente presentará las características temporales y espectrales necesarias. Por ello, normalmente se emplea una etapa de acondicionamiento eléctrico, que involucra la amplificación y el filtrado de esta señal [16]. Así, se modifica el rango de excursión de la amplitud y el contenido frecuencial de la señal de voz, obteniéndose una señal adecuada para ser analizada, mediante un osciloscopio o un altavoz, o para ser digitalizada [16, 163]. En este último caso, se obtiene una versión discreta de la señal de voz apta para ser almacenada o ser procesada por un sistema digital [42, 101, 128].

En la Sec. 2.3, presentamos los fonemas más importantes en el español rioplatense y describimos sus características distintivas más importantes. Estas características se manifiestan en la señal de voz y sus diferentes representaciones alternativas [42, 101, 128, 133]. En la Fig. 2.6, mostramos 50 ms de una vocal /a/ sostenida, fila superior, y dos estimaciones de su espectro de potencia $\hat{P}_{/\mathrm{a}/}$, fila inferior. Para el cálculo de los espectros de potencia aplicamos el método no paramétrico de *Welch*, línea continua, y un método paramétrico basado en el modelado autorregresivo, línea discontinua. Utilizaremos estas tres representaciones para extraer información complementaria de la señal de voz. En el Cap. 3 profundizaremos en el modelado paramétrico de la señal de voz y en el uso de modelos autorregresivos.

A continuación, analizaremos el ejemplo presentado en la Fig. 2.6. En primer lugar, podemos apreciar en la forma de onda temporal la dinámica cuasiperiódica característica de los fonemas sonoros. A su vez, en esta gráfica podemos distinguir el período fundamental T_0 de la señal de voz. En el caso analizado $T_0 = 9{,}31$ ms. Como dijimos anteriormente, este comportamiento cuasiperiódico es el resultado del flujo pulsátil en la glotis, producto de la modulación de las cuerdas vocales. Este fenómeno

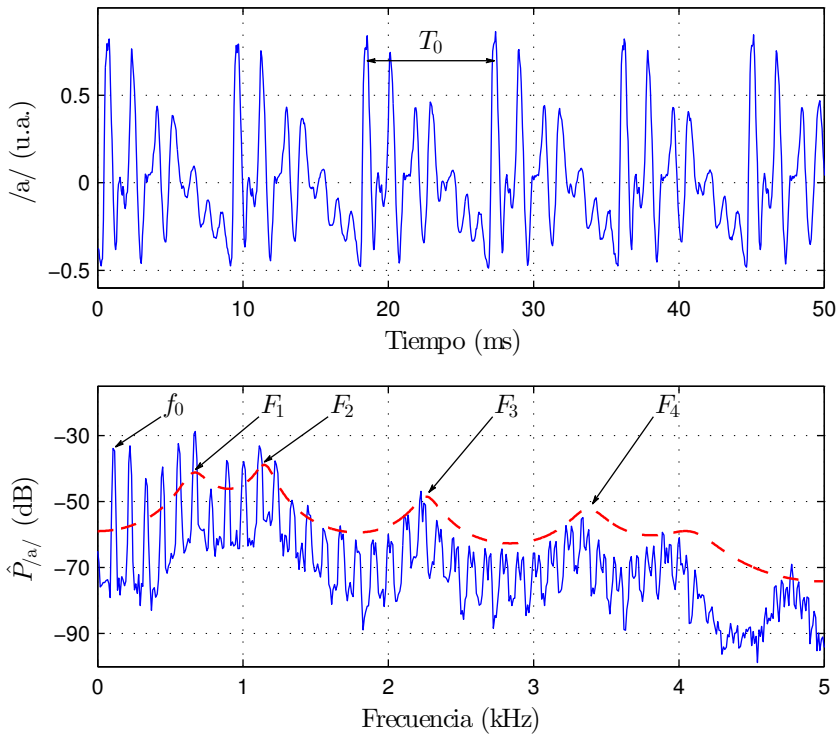


Figura 2.6: Información temporal y espectral para 50 ms de una vocal /a/ neutra, fonada de forma sostenida por un sujeto masculino adulto con voz normal. *Arriba:* forma de onda temporal. Se indica el período fundamental T_0. *Abajo:* espectro de potencia $\hat{P}_{/a/}$ calculado con el método de *Welch* considerando una ventana de *Hamming*, línea continua, y aplicando modelado autoregresivo, línea discontinua. Se señalan la frecuencia fundamental, f_0, y las cuatro primeras formantes: F_1, F_2, F_3, F_4. En este ejemplo: $T_0 = 9{,}31$ ms, $f_0 = 107{,}42$ Hz, $F_1 = 660{,}47$ Hz, $F_2 = 1153{,}15$ Hz, $F_3 = 2264{,}81$ Hz y $F_4 = 3376{,}32$ Hz.

se expresa también en la información frecuencial, pero de un modo diferente. Debido a la dinámica extremadamente regular de la vocal /a/, su espectro de potencia presenta una estructura armónica, es decir, la información frecuencial se concentra en valores de frecuencia específicos. Esto se manifiesta en los picos regulares que se aprecian en el espectro de potencia calculado con el método de *Welch*, línea continua. La ubicación del primer armónico se corresponde con la frecuencia fundamental f_0 de la señal de voz. En el ejemplo analizado, se obtuvo $f_0 = 107{,}42$ Hz como valor promedio para la ventana de señal analizada. Los restantes componentes armónicos ocurren a múltiplos enteros de f_0.

Por otro lado, el espectro de potencia permite estudiar las frecuencias de resonancias, o formantes, resultantes de la configuración del tracto vocal al momento de la fonación. En la estimación con el método de *Welch*, línea continua, podemos observar que existe una modulación de la información espectral, en especial de los armónicos, caracterizada por regiones de la frecuencia con una mayor concentración de la potencia. Podemos apreciar mejor este fenómeno en la estimación calculada con el método paramétrico, línea discontinua. Esta estimación del espectro de potencia se denomina *suavizada*, ya que se representa el comportamiento global del espectro de potencia excluyendo los detalles o estructuras singulares, por ejemplo los armó-

Tabla 2.3: Valores promedios de la frecuencia fundamental f_0 y de las formantes F_1, F_2, F_3 y F_4 de las cinco vocales del español rioplatense, para voces masculinas y femeninas. Todos los valores se indican en Hz. Reproducido de [13].

	Sexo	/a/	/e/	/i/	/o/	/u/
f_0	Fem.	205	205	207	204	204
	Masc.	127	125	130	124	124
F_1	Fem.	330	330	330	546	382
	Masc.	830	430	290	510	335
F_2	Fem.	1553	2500	2765	934	740
	Masc.	1350	2120	2295	860	720
F_3	Fem.	2890	3130	3740	2966	2760
	Masc.	2450	2628	2915	2480	2380
F_4	Fem.	3930	4150	4366	3854	3380
	Masc.	3665	3610	3645	3485	3355

nicos [42, 128]. Así, las formantes corresponden a las ubicaciones de los picos en el espectro suavizado. Se simbolizan F_1, F_2, ... , F_n y se asignan en orden creciente de frecuencia. En la gráfica, podemos distinguir las cuatro primeras formantes para la vocal /a/ estudiada. En este caso, obtuvimos los valores $F_1 = 660{,}47$ Hz, $F_2 = 1153{,}15$ Hz, $F_3 = 2264{,}81$ Hz y $F_4 = 3376{,}32$ Hz.

En la Fig. 2.7 mostramos, en la columna izquierda, las formas de onda temporal para las cinco vocales del español rioplatense emitidas de forma neutra y sostenida y, en la columna derecha, las estimaciones de los correspondientes espectros de potencia $\hat{P}$. Todas estas señales y las de la Fig. 2.6 se obtuvieron de un mismo sujeto masculino adulto con voz normal. En estas gráficas, podemos apreciar nuevamente todas las características temporales y frecuenciales de los sonidos sonoros, las cuales discutimos al analizar el ejemplo anterior. En la Tab. 2.3 presentamos los valores promedios de la frecuencia fundamental f_0 y de las formantes F_1, F_2, F_3 y F_4 para las cinco vocales del español rioplatense [13]. Estos valores se calcularon a partir de una población de 90 participantes (45 masculinos y 45 femeninos), abarcando un rango de edad de 18 a 35 años, los cuales no presentaban patologías vocales ni trastornos del lenguaje al momento del registro.

De la información expuesta, podemos apreciar que los valores de f_0 se comportan de forma notablemente diferente dependiendo del sexo, siendo f_0 mayor en mujeres. Luego, analizando cada sexo por separado, notamos que f_0 varía sutilmente para las diferentes vocales. Además, podemos apreciar que las formantes, en especial los valores de F_1 y F_2, permiten diferenciar entre sí las diferentes vocales e identificar las cualidades propias de cada una [13, 42, 133]. Ejemplificamos esto último en las estimaciones de $\hat{P}$ presentadas en la columna derecha de Fig. 2.7, correspondientes a las diferentes vocales generadas por un mismo individuo. Los valores de las formantes presentados en la Tab. 2.3 resultan, en general, muy similares a los reportados en [64, Tab. 3 de la Pag. 86] para vocales en español, y en [151, Tab. 6.2 de la Pag. 288], para vocales en inglés. Sin embargo, se aprecia una diferencia considerable en los valores de F_1 en las vocales /a/ y /e/ para los sujetos femeninos. De acuerdo a

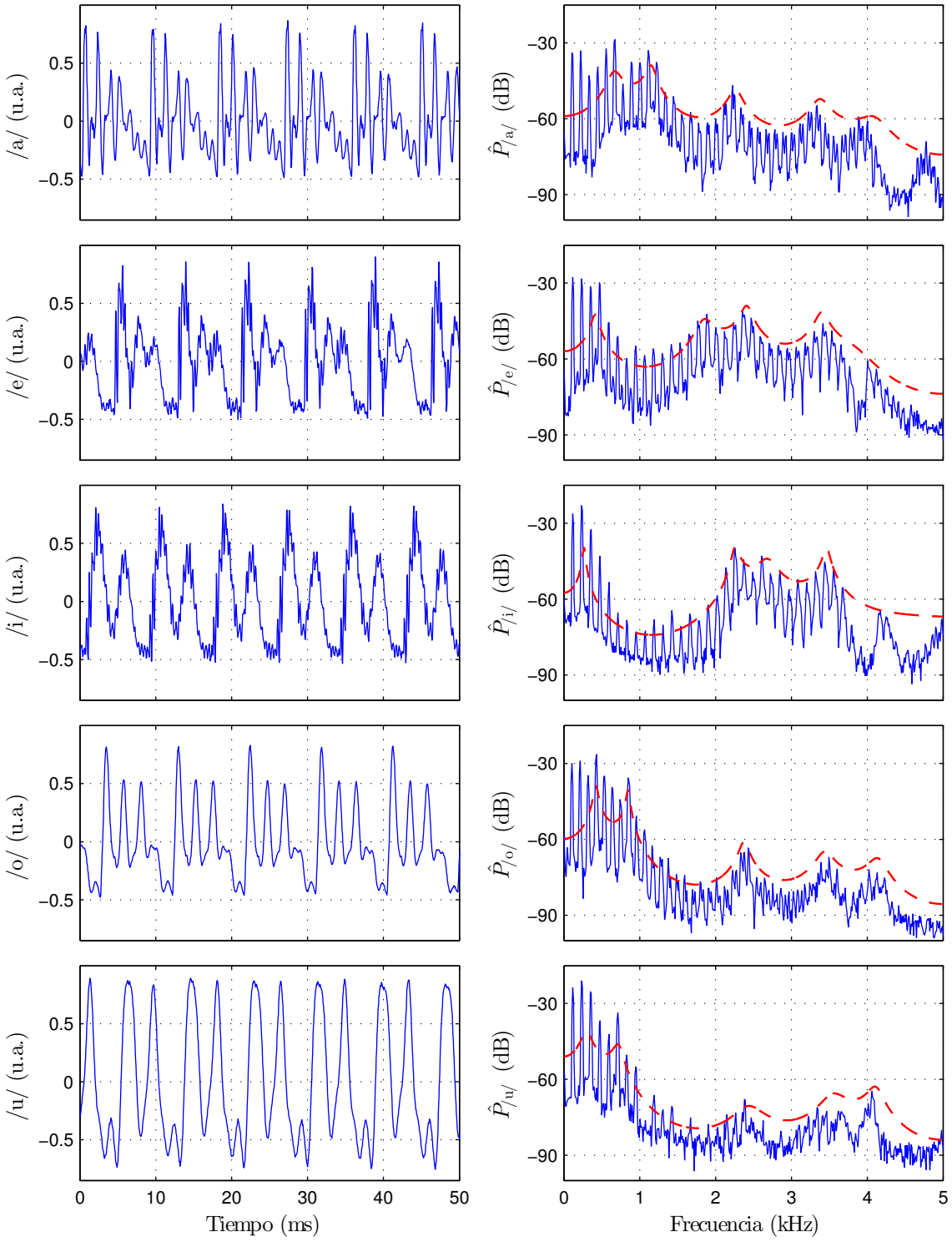


Figura 2.7: Forma de onda temporal, columna izquierda, y espectro de potencia $\hat{P}$, columna derecha, correspondientes a las 5 vocales neutras en el español rioplatense, para un mismo sujeto masculino adulto con voz normal. Cada fila corresponde a una vocal. Las estimaciones de $\hat{P}$ se generaron con el método de Welch considerando una ventana de *Hamming*, línea continua, y utilizando modelado autorregresivo, línea discontinua.

[151], para sujetos femeninos corresponden los valores promedios $F_1 = 850$ Hz en vocales /a/ y $F_1 = 560$ Hz en vocales /e/.

Hasta aquí, hemos dedicado una parte importante de este capítulo a discutir las propiedades de las emisiones vocales sostenidas. Es importante destacar que los aportes originales de esta tesis de doctorado están dirigidos al estudio de este tipo de señales. Por otro lado, en habla continua observamos un comportamiento considerablemente más complejo. Esto es así, principalmente, porque la fonación es un proceso continuo en el cual el aparato fonador modifica su configuración de forma dinámica para generar la señal de voz. Además, la fonación depende fuertemente de la organización y ocurrencia de los fonemas en cada oración. Aun cuando los fonemas ayudan a interpretar el mensaje, la segmentación de una señal de voz en unidades fonéticas no resulta ser una tarea sencilla [42, 101, 132].

En la Fig. 2.8, presentamos la información temporal y frecuencial para la frase corta en español: *"Bueno, muchas gracias"*, fonada por un sujeto masculino adulto con voz normal. En la imagen superior, mostramos la forma de onda temporal de la señal de voz, segmentada en el tiempo de acuerdo a la frase. Además, graficamos la energía por ventana, superpuesta en línea continua gruesa. La curva de energía se escaló para que coincida con el rango de amplitud de la señal de voz, por lo que sólo resultará relevante su comportamiento. En el centro, podemos apreciar las estimaciones por ventanas de f_0 a lo largo del tiempo, calculadas aplicando el método de la autocorrelación [21, 42]. Por último, en la parte inferior presentamos el *espectrograma* de la señal. Es importante señalar que, en este ejemplo, la segmentación fue realizada con fines meramente ilustrativos, por el autor de este documento.

En primer lugar, podemos apreciar en la gráfica superior la naturaleza del habla continua, al extremo de no llegar a distinguir ningún segmento de silencio entre las palabras *"muchas"* y *"gracias"*. Además, observamos que en la señal de voz los fonemas vocales presentan una duración y una energía considerablemente mayores que las consonantes. Para ello, comparar el comportamiento de /a/ y /o/ con respecto a /s/ y /ʧ/ (correspondiente a la grafía *ch*). Estas son dos de las diferencias más importantes entre vocales y consonantes sordas, además de la existencia o no de actividad glótica [42, 101, 128]. Podemos apreciar, también, como fluctúa f_0 a lo largo de la frase analizada. Lógicamente, sólo tiene sentido estimar f_0 en los fonemas sonoros. Este parámetro es uno de los principales responsables de modificar la entonación y el ritmo en el habla. La energía, por su parte, está ligada a la intensidad con que se pronuncia cada elemento de la frase. Todos estos atributos repercuten en la prosodia de un mensaje [133, 151]. En el ejemplo analizado, podemos apreciar que tanto la energía como f_0 decrecen hacia el final de la frase.

En el espectrograma, podemos observar que el comportamiento espectral de la señal de voz es muy variado y complejo. Las vocales presentan un comportamiento armónico, donde las regiones de mayor potencia corresponden a las formantes. Se puede apreciar que en algunas regiones las formantes presentan un comportamiento estable, por ejemplo en las dos instancias de la vocal /a/, mientras que en otras regiones las formantes se modifican dependiendo del contexto o de la entonación, como en el diptongo /ue/ o en la vocal /o/. Por otro lado, en las consonantes la información espectral se distribuye en un rango amplio de frecuencias. En general, se suele considerar que las consonantes sordas presentan una estructura espectral similar al ruido *coloreado* [42, 128, 133]. Este fenómeno puede apreciarse fácilmente en el espectrograma mostrado, para el caso de las consonantes /ʧ/ y /s/.

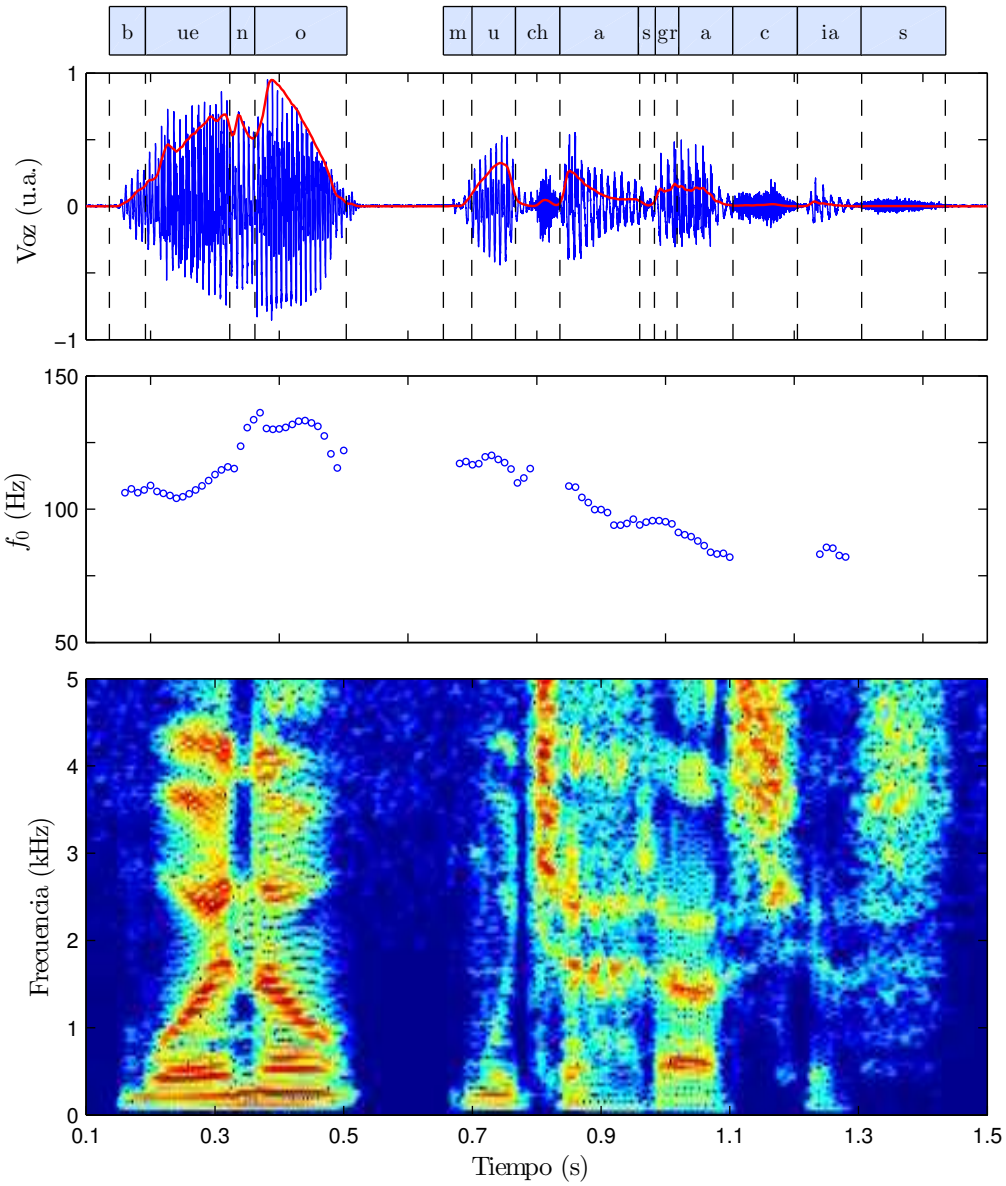


Figura 2.8: Información temporal y espectral de la señal de voz para la frase corta en español *"Bueno, muchas gracias"*. *Arriba:* forma de onda temporal de la señal de voz. Se presenta también la energía por ventana, superpuesta en línea continua gruesa. *Centro:* estimaciones por ventanas de f_0 a lo largo del tiempo. *Abajo:* espectrograma de la señal de voz calculado usando ventanas de *Hamming*, con una superposición del 87,5 %.

Presentamos aquí un ejemplo sencillo de señal de voz para habla continua y analizamos superficialmente algunas de las características más importantes. El objetivo de este análisis fue mostrar que aquellos parámetros que permiten describir satisfactoriamente un fonema generado de forma aislada, en habla continua presentan una dinámica considerablemente más compleja. El lector interesado en el análisis temporal y espectral de la voz para habla continua puede recurrir a la extensa bibliografía específica existente, como por ejemplo: [25, 42, 64, 129, 151].

En esta sección, nos concentramos en el análisis de la voz en el plano temporal y frecuencial, siguiendo la metodología clásica. Sin embargo, se pueden explotar también otras estrategias para su estudio. Por ejemplo, es posible aplicar alguno de los métodos de procesamiento de señales listados a continuación: el análisis de correlación, la transformada *ondita* en sus versiones continua y discreta, las representaciones tiempo-frecuencia no lineales, la descomposición empírica en modos, el análisis basado en diccionarios discretos y las representaciones ralas, entre otras. A su vez, existen otros métodos específicos para la señal de voz, entre los que podemos nombrar: el análisis predictivo lineal por ventanas y su versión sincronizada con la dinámica glótica, el análisis cepstral y la descomposición mediante filtrado inverso, entre otras. En el próximo capítulo, describiremos algunos de estos métodos específicos para la señal de voz. En caso de interés, el lector puede recurrir a la extensa bibliografía específica dedicada a los tópicos aquí listados, por ejemplo: [42, 101, 106, 128, 132, 133, 136].

Ventajas y limitaciones de la señal de voz

Sin lugar a dudas, la principal ventaja de la señal de voz consiste en el gran cúmulo de conocimiento existente, que se ha desarrollado a lo largo del tiempo gracias a los diversos estudios e investigaciones. Es muy sencillo registrar y digitalizar este tipo de señales, debido al fácil acceso al equipamiento requerido, siendo éste prácticamente de uso universal. Por ello, la señal de voz es una de las primeras opciones, como material de prueba, en simulaciones destinadas al estudio de nuevos métodos o algoritmos para el procesamiento de señales [128, 133]. Por otro lado, en la medicina se emplean asiduamente registros de vocales sostenidas o de habla continua para la evaluación y el diagnóstico del aparato fonador [16, 41]. Por ello, la señal de voz es muy importante para la detección y el seguimiento de trastornos de la voz [82, 136]. Sin embargo, presenta tres importantes desventajas.

En primer lugar, la voz es una onda acústica que se transmite por el aire y, por ello, es muy susceptible al ruido generado por diversas fuentes sonoras dispersas en el medio circundante. Este ruido acústico puede enmascarar, deteriorar o incluso modificar la información transmitida por ésta. En general, eliminar estas perturbaciones de la señal de voz es una tarea difícil que ha requerido el desarrollo de diferentes técnicas específicas de filtrado y de realce [101, 132].

En segundo lugar, la forma de onda de esta señal es extremadamente variada y depende de: el proceso de fonación, la naturaleza del fonema, la relación de un fonema con sus vecinos en habla continua y la prosodia, entre otros factores [42, 163]. Si se tiene en cuenta además la contaminación por el ruido acústico, se torna difícil la caracterización y la clasificación de cada uno de los sonidos.

Por último, recordemos que la información de la dinámica glótica, que depende a su vez del comportamiento de las cuerdas vocales, es modificada (filtrada) por el tracto vocal. Además, pueden desarrollarse otras fuentes acústicas a lo largo del

tracto vocal, y actuar en combinación con la fuente ubicada en la glotis. El estudio y la evaluación de esta información a partir de la señal de voz se convierte en una tarea compleja y muy propensa a errores. Ejemplos de esta situación son: el cálculo de f_0, la detección de los eventos glóticos y la estimación del flujo de aire a través de la glotis [16, 128, 136]. Esta dificultad se agrava más aun cuando se toman en cuenta los otros aspectos desfavorables ya enunciados.

2.4.2. Electroglotograma (EGG)

Recientemente, una señal denominada electroglotrograma (EGG), o también laringograma, ha despertado un gran interés por parte de la comunidad médica y científica, por ser adecuada para el estudio del comportamiento de las cuerdas vocales [163]. Esta señal se obtiene a partir de un método denominado *electroglotografía*, propuesto originalmente por Fabre en 1957, y el *electroglotógrafo* es el dispositivo electrónico diseñado para llevarla a cabo. En esta sección, describiremos brevemente la señal de EGG. Para un estudio pormenorizado, el lector interesado puede dirigirse a [15, 16]. En [34] se discuten las principales características y las aplicaciones más importantes del EGG en la práctica clínica, a la vez que se presentan una serie de recomendaciones para su correcta interpretación.

La *electroglotografía* se basa en la impedancia eléctrica que presentan los diferentes tejidos y órganos, y en la forma en que varía esta propiedad ante un cambio de forma o en las relaciones con otras estructuras [15, 16]. En general, los tejidos biológicos presentan una conductividad eléctrica relativamente buena. Por su parte, el aire posee una alta impedancia. Debido a esto, la impedancia eléctrica de la laringe se modifica en cada ciclo glótico, aumentando en el lapso en el cual la glotis está abierta y disminuyendo en la etapa de aducción de las cuerdas vocales [15, 16].

Es posible sensar la variación de la impedancia en la laringe. Para ello, se suministra una corriente alterna de baja amplitud, menor o igual a 10 mA, y de muy alta frecuencia, normalmente en el orden de 2-5 MHz, a través de dos electrodos ubicados sobre el cuello. La posición recomendada de estos electrodos es a la altura de las láminas del cartílago tiroides y a ambos lados de la prominencia laríngea [34]. Luego, la amplitud de la señal portadora es modulada por la dinámica de las cuerdas vocales, las cuales presentan una frecuencia de oscilación mucho menor. Como resultado, la envolvente de la diferencia de potencial, o voltaje, entre los electrodos produce un correlato eléctrico del cambio en la impedancia de la laringe. Finalmente, el EGG se obtiene filtrando la información de muy baja frecuencia, asociada con los movimientos de otras estructuras ajenas a las cuerdas vocales [161]. Esta señal eléctrica puede visualizarse con un osciloscopio o puede digitalizarse para su almacenamiento y posterior procesamiento.

Se recomienda el uso de una fuente de corriente alterna muy estable, de baja amplitud y alta frecuencia [16]. La baja amplitud ayuda a evitar respuestas fisiológicas indeseables en las estructuras excitables de la laringe, por ejemplo en el tejido nervioso o muscular. Emplear una alta frecuencia reduce las sensaciones desagradables o de dolor en las personas. A su vez, se recomienda que las dimensiones de los electrodos sean semejantes a las de la glotis, que la distancia entre los electrodos sea pequeña, que el posicionamiento se mantenga durante el registro y que se logre una adecuada interfaz electrodo-piel. Debido a las diferencias anatómicas existentes, resulta más sencillo obtener registros de EGG en hombres adultos, que en mujeres

adultas y en niños [34]. Sin embargo, el posicionamiento de los electrodos no es una tarea sencilla y, usualmente, se logra luego de varios intentos.

En general, se supone que el EGG representa el área de contacto relativa entre las cuerdas vocales durante la fonación. Se emplea la expresión medida relativa, ya que es imposible asegurar que se produjo un cierre o apertura total de la glotis. En la práctica, un error frecuente consiste en suponer una relación lineal entre el EGG y el área de contacto. Algunos autores han mostrado que la relación es prácticamente lineal en la fase donde se produce la separación gradual de las cuerdas vocales, mientras que en la fase de cierre la relación es más compleja [74]. Asimismo, el EGG puede modificarse por cambios en la forma o en la posición de la laringe durante la fonación, o por cambios en las otras estructuras por las que fluye la corriente, como por ejemplo la epiglotis y los pliegues vestibulares [161].

En la actualidad, existen lineamientos que permiten interpretar el EGG en función de la dinámica glótica, aplicables a registros de señal de alta calidad y correspondientes a sujetos normales [34, 161]. Inclusive, se ha incorporado la información obtenida a partir del estudio de otras señales relacionadas, como por ejemplo la derivada del EGG (dEGG) [77]. Sin embargo, aun no se ha arribado a un consenso generalizado respecto al comportamiento del EGG en condiciones patológicas o en sujetos con una anatomía atípica [16].

La Fig. 2.9 ejemplifica la relación entre tres señales biomédicas de la fonación, correspondientes a una vocal /a/ sostenida fonada por un sujeto masculino adulto. En la fila superior mostramos la señal de voz, mientras que en la segunda fila presentamos la forma de onda temporal del EGG. Estas dos señales se registraron en forma simultánea. En el ejemplo mostrado, valores mayores de EGG representan una mayor conductividad en la laringe, señalando a su vez un mayor contacto entre las cuerdas vocales. Es común encontrar estos valores de forma invertida en la bibliografía, indicando así la resistividad en la laringe [34]. Presentamos también la derivada del EGG (dEGG), superpuesta en línea continua gruesa, cuya amplitud se modificó para que sea comparable. Podemos apreciar que tanto el EGG como la dEGG capturan la dinámica cuasiperiódica característica de los sonidos sonoros. Además, estas dos señales poseen una forma de onda más simple que la señal de voz, en especial la dEGG. Esto facilita considerablemente la estimación de los diferentes parámetros acústicos, como por ejemplo el T_0, la f_0 o el ciclo de trabajo. Para mayor información respecto a la estimación de parámetros a partir de estas señales, tanto en casos normales como patológicos, el lector puede referirse a [15, 16, 77].

En la Fig. 2.10, comparamos el comportamiento de las tres señales biomédicas analizadas anteriormente, en este caso para la frase corta en español: *"Bueno, muchas gracias"*. Señalamos, también, la segmentación de estas señales de acuerdo a la información presente en la frase. En la fila superior, mostramos la señal de voz. Ésta y la segmentación coinciden con aquellas analizadas en el ejemplo de la Fig. 2.8. Por otro lado, en la segunda fila presentamos la señal de EGG. Podemos observar que esta señal desarrolla un comportamiento regular con una amplitud importante para los fonemas sonoros, mientras que para los fonemas sordos la amplitud se reduce drásticamente y se pierde toda regularidad. Esto se puede apreciar fácilmente comparando la morfología del EGG para las vocales /a/ y /o/, por un lado, y su comportamiento para las consonantes sordas /s/ y /ʧ/, por el otro. Este ejemplo nos permite mostrar las virtudes que presenta el EGG para la detección y el estudio de la dinámica glótica.

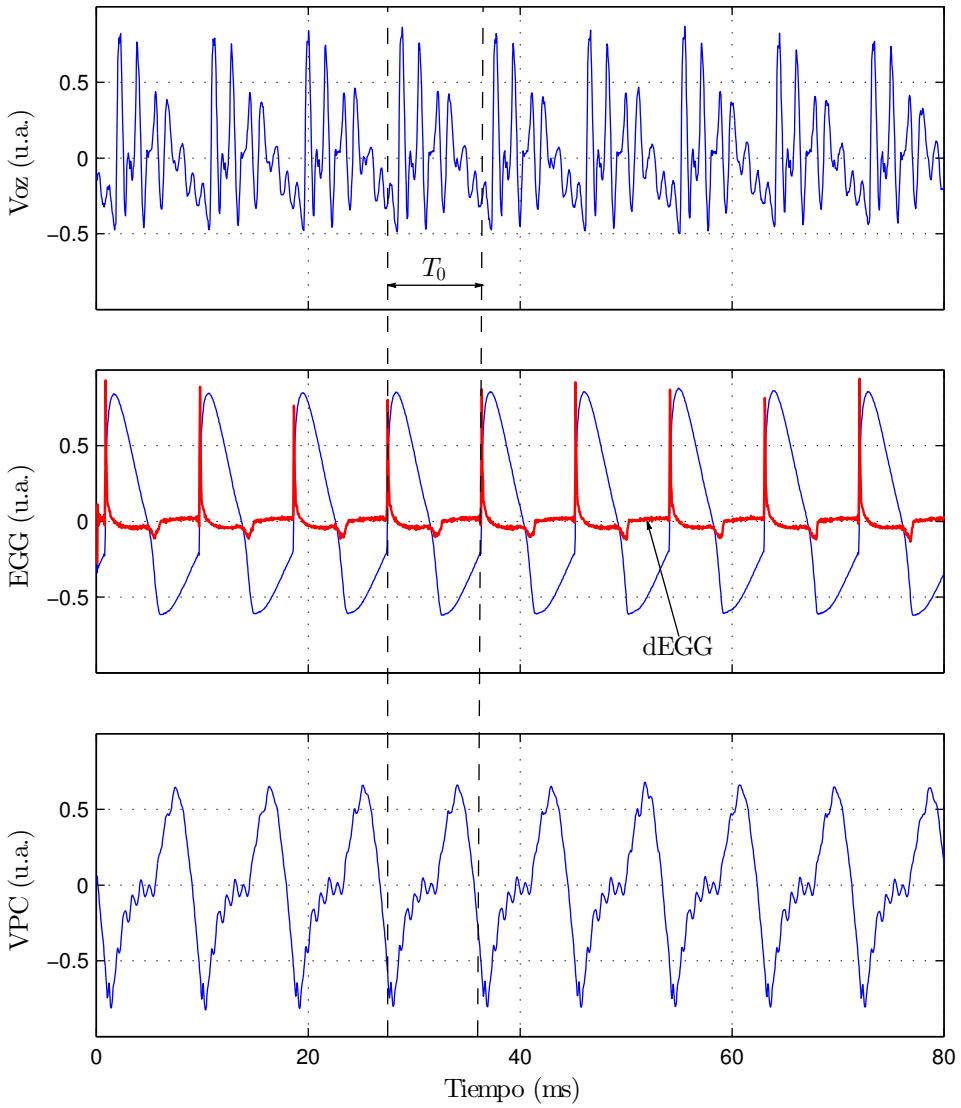


Figura 2.9: Comparación del comportamiento de tres señales biomédicas de la fonación para una vocal sostenida. Las señales corresponden a una vocal /a/ sostenida fonada por un sujeto masculino adulto. *Arriba:* Señal de voz (coincide con la señal mostrada en la fila superior de la Fig. 2.6). *Centro:* Electroglotograma (EGG) y su derivada (dEGG). *Abajo:* Vibraciones en la piel del cuello (VPC). Se indica también el período fundamental T_0. Las tres señales capturan el comportamiento cuasiperiódico de los sonidos sonoros.

Ventajas y limitaciones del EGG

El EGG presenta un conjunto de características importantes tanto para la medicina como para la ingeniería. En primer lugar, esta señal se obtiene libre de la modulación del tracto vocal. Por ello, captura en general información específica del comportamiento de la laringe y, en particular, de las cuerdas vocales [74]. Como consecuencia, el EGG permite inferir diferentes aspectos de la dinámica de las cuerdas vocales, que de otro modo requerirían técnicas o dispositivos más complejos para ser estudiados, como por ejemplo la máscara de Rothenberg o la videoquimografía [16]. Por otro lado, el EGG desarrolla una forma de onda temporal mucho más simple que la señal de voz, lo que resulta sumamente útil para estimar parámetros acústicos representativos de la dinámica glótica [15, 77]. Por último, es importante destacar que la *electroglotografía* es un procedimiento no invasivo, inocuo y que no interfiere con la medición de otras señales, como la voz o el flujo de aire exhalado [16, 161].

A continuación, mencionaremos algunas de las limitaciones del EGG que consideramos más importantes. En primer lugar, la calidad de la señal depende fuertemente de la correcta ubicación de los electrodos: esto es debido a que, en general, el posicionamiento de los electrodos no es sencillo y, por ello, para realizar esta tarea se requieren técnicos o especialistas debidamente entrenados. También influyen en la calidad del EGG las modificaciones en la anatomía del cuello, en especial aquellas resultantes del desplazamiento de la laringe en habla continua. Por otra parte, la señal de EGG es sensible a fuentes de perturbación electromagnética, por ejemplo el ruido de línea o las señales de telecomunicación. Por último, para obtener esta señal se requiere de un *electroglotógrafo*, que es un equipo muy específico y, en general, escaso. Para una mayor información respecto a las ventajas y las limitaciones de esta señal referirse a [15, 34, 161].

2.4.3. Vibraciones en la piel del cuello (VPC)

En esta sección, describiremos las vibraciones en la piel del cuello y, luego, explicaremos cómo a partir de éstas construir una señal biomédica de la fonación. En las últimas décadas, las vibraciones en la piel del cuello han recibido una atención importante por parte de la comunidad clínica y científica. Esto se debe, principalmente, a que se ha demostrado que acarrean información de interés para el estudio la dinámica glótica [14]. Por este motivo, desde hace un tiempo han surgido diferentes monitores ambulatorios, diseñados a partir de la digitalización y el procesamiento de estas vibraciones, los cuales son aptos para la realización de estudios prolongados de la actividad vocal y de la voz ocupacional [16, 112].

Las vibraciones en la piel del cuello tienen su origen en las ondas acústicas generadas en la laringe. Parte de estas ondas se propaga por el tracto vocal hacia los labios, mientras que la parte restante se transmite en dirección caudal por las vías aéreas subglóticas [163]. Estas últimas, las ondas acústicas en retroceso, producen vibraciones en la tráquea, las cuales se transmiten a su vez a través de los tejidos del cuello hacia la superficie de la piel. Recordemos que toda onda acústica puede transmitirse por diferentes medios, además del aire, entre los que se encuentran los tejidos biológicos [151]. Es por ello que estas vibraciones superficiales pueden capturarse empleando dispositivos sensibles a perturbaciones mecánicas o a ondas de presión [121]. Ejemplos de esta clase de sensores son los micrófonos, los acelerómetros, los micrófonos de contacto y los extensómetros. El correlato eléctrico generado

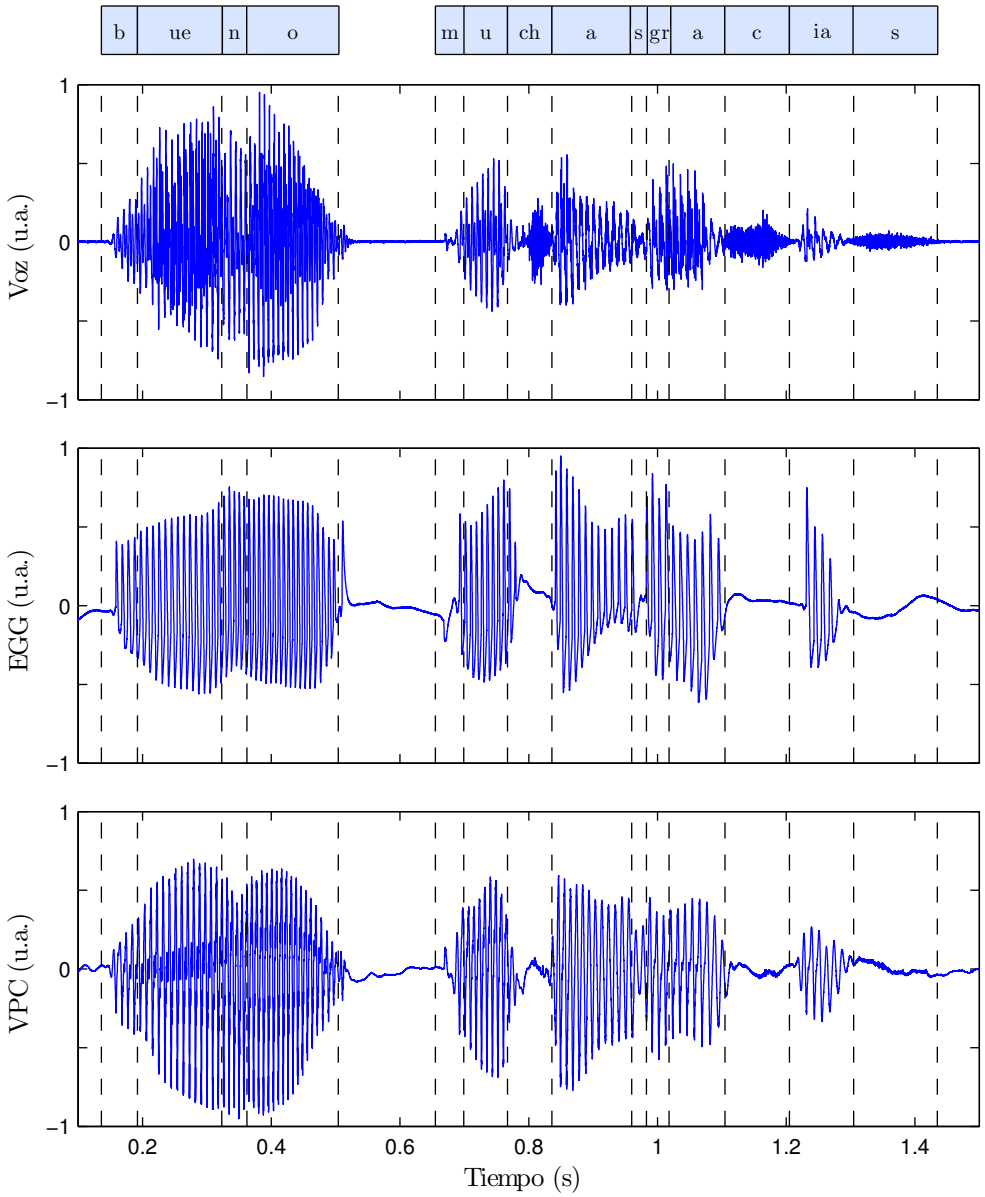


Figura 2.10: Comparación del comportamiento de tres señales biomédicas de la fonación en habla continua. Las señales corresponden a la frase corta en español *"Bueno, muchas gracias."* fonada por un sujeto masculino adulto. *Arriba:* Señal de voz (coincide con la señal mostrada en la fila superior de la Fig. 2.8). *Centro:* Electroglotograma (EGG). *Abajo:* Vibraciones en la piel del cuello (VPC). Las señales se segmentaron de acuerdo a la información presente en la frase. La señal de voz preserva la información de la articulación en el tracto vocal, mientras que las señales EGG y VPC muestran más claramente la información de la dinámica glótica.

por estos sensores se denomina señal de vibraciones en la piel de cuello (VPC), o también llamada señal de aceleraciones en la superficie del cuello [16].

La presión subglótica está formada por la presión resultante del flujo de aire a través de las vías aéreas subglóticas, con dinámica lenta, y por un componente con dinámica rápida asociado a la onda sonora proveniente de la laringe. Por ello, las vibraciones registradas corresponden a la expresión, en la piel del cuello, de la presión compuesta en la región subglótica [121]. La señal generada por estas vibraciones es luego acondicionada con el doble propósito de eliminar la información de muy baja frecuencia y, además, adecuar el rango dinámico de la señal [16]. De este modo, idealmente en la señal de VPC sólo se preserva la información del componente con dinámica rápida [121]. En el dominio frecuencial, la mayor porción de la potencia de esta señal se concentra en el rango de frecuencias de 50-1000 Hz [179]. Por su parte, el rango dinámico depende de la aplicación considerada. Por ejemplo, las VPC pueden graficarse utilizando un osciloscopio para su estudio en tiempo real, o pueden digitalizarse para su almacenamiento y posterior procesamiento.

Dos factores de suma importancia para el registro de la señal de VPC son la forma y el posicionamiento del sensor. Afortunadamente, existen en la actualidad diferentes sensores diseñados con un tamaño reducido y muy livianos. Los especialistas recomiendan ubicar el sensor en la escotadura yugular, localizada en la parte frontal del cuello por debajo de la laringe y apenas por arriba del manubrio del esternón. Las razones de esta recomendación son las siguientes [179]:

- En la escotadura yugular la distancia entre la tráquea y la superficie de la piel del cuello es mínima, por lo que se obtienen señales de máxima amplitud.
- La posición exacta del sensor en esa zona no es un factor crítico.
- Permite una localización discreta, fácil de ocultar y suficientemente confortable para la fijación de un sensor por períodos de tiempo largos.

En la fila inferior de la Fig. 2.9 presentamos un ejemplo de una señal de VPC, correspondiente a una vocal /a/ sostenida emitida por un sujeto masculino adulto. Con fines comparativos, mostramos en la misma figura la señales de voz y de EGG. Todas ellas fueron registradas en forma simultánea. En la gráfica, podemos apreciar que la señal de VPC captura el comportamiento cuasiperiódico característico de los sonidos sonoros. Esto se debe a que el cierre brusco de la glotis es la fuente de excitación más importante en estos sonidos y produce, a la vez, las vibraciones de mayor amplitud tanto en la tráquea como en la superficie del cuello. Es importante destacar que esto ocurre aún ante un cierre incompleto de la glotis [14]. A su vez, en cada ciclo de esta señal se observan oscilaciones rápidas de baja amplitud, resultantes de las frecuencias de resonancia en el tracto subglótico [121].

Finalmente, en la fila inferior de la Fig. 2.10 presentamos la señal de VPC, obtenida para la misma frase corta en español analizada en ejemplos anteriores. Esta gráfica nos permite examinar el comportamiento de esta señal en habla continua. Podemos apreciar rápidamente que existe una marcada diferencia en el comportamiento de la señal de VPC para fonemas sonoros y sordos. En el primer caso, nuevamente la señal presenta una dinámica cuasiperiódica y una importante excursión en su amplitud. Esto se puede observar, fácilmente, en las vocales o en los fonemas nasales. En fonemas sordos, no es posible identificar ninguna forma característica

y, lo que es más importante aún, la amplitud de la señal disminuye considerablemente o se anula. Por estas razones, la señal de VPC ha demostrado ser sumamente útil para la estimación de f_0, de T_0 y de otros parámetros de importancia clínica [14, 16, 112, 179].

Ventajas y limitaciones de las VPC

Enunciaremos a continuación las principales ventajas que presenta la señal de VPC, algunas de las cuales ya fueron expuestas en la sección anterior. Las ventajas más importantes se desprenden de la estrategia de sensado. Como se dijo anteriormente, esta señal se obtiene a partir de las vibraciones mecánicas en la piel del cuello. Por ello, la señal de VPC es robusta ante fuentes de ruido acústico [16]. A su vez, estas vibraciones superficiales son más notorias para los fonemas sonoros, por lo que esta señal es muy útil para el estudio de la dinámica glótica [14].

En general, la señal de VPC se caracteriza por una forma de onda relativamente simple, producto de su información concentrada en las bajas frecuencias [179]. Pudimos apreciar esto en los ejemplos mostrados de las Figs. 2.9 y 2.10. Por último, es importante destacar que la ubicación óptima para el sensor está determinada por una estructura anatómica fácilmente reconocible en sujetos masculinos y femeninos, tanto en adultos como en niños [14]. Todas estas ventajas explican el creciente interés, por parte de los especialistas, en el uso de la señal de VPC para el estudio de la dinámica glótica y para la estimación de parámetros acústicos, como por ejemplo f_0 o T_0.

Sin lugar a dudas, la principal limitación que presenta esta señal es el escaso conocimiento existente respecto a cómo se modifica su forma de onda para los diferentes registros vocales o ante la presencia de una patología vocal [16]. A su vez, se han observado diferencias importantes en la dinámica de la presión subglótica y de su correlato estimado en la superficie del cuello. Por ello, los especialistas sugieren corregir estas diferencias o ser cautelosos al interpretar la señal de VPC [121]. Un factor importante a tener en cuenta durante el sensado es la composición del cuello a la altura de la escotadura yugular. La calidad de la señal será adecuada siempre que la distancia entre la tráquea y la superficie de la piel sea mínima.

2.4.4. Comparación entre señales: voz, EGG y VPC

De las tres señales presentadas, sin ninguna duda la voz es la más estudiada a lo largo del tiempo. Es la única para la que, en la actualidad, existe un consenso por parte de los especialistas de las diferentes disciplinas respecto a sus características y a la información que contiene [64, 151]. En la ingeniería, la gran mayoría de los desarrollos relacionados con la fonación y la comunicación se enfocan en el procesamiento de esta señal. Sin embargo, desde hace algún tiempo las señales de EGG y VPC han cobrado mayor relevancia, particularmente para la exploración de las cuerdas vocales de forma no invasiva y para su estudio en sesiones o en jornadas de larga duración [16, 112]. Discutiremos brevemente en esta sección las principales coincidencias y diferencias entre estas tres señales.

Comparemos, en primer lugar, las técnicas para el registro de estas señales. En el caso de la señal de voz, el procedimiento es universalmente conocido y el equipamiento necesario se consigue fácilmente. Esto es así, al extremo de que hoy en día cualquier persona puede registrar su voz en su propia casa. Sin embargo, en

estudios científicos o en aplicaciones médicas se requieren protocolos confeccionados de modo tal que el registro se realice en condiciones controladas y que de él se obtengan señales de calidad [41, 42, 128]. Para las señales de EGG y VPC la situación es diferente. Para el registro de estas señales se requiere equipamiento específico [16]. En general, se tiene poco acceso a estos instrumentos, incluso por parte de los especialistas. Esto último, es una opinión personal del autor de este documento.

Dos factores importantes relacionados con el registro de estas señales son, por un lado, el posicionamiento de los sensores y, por el otro, la robustez de estas señales a diferentes fuentes de perturbación. En la actualidad, existe una gran variedad de micrófonos, fabricados aplicando diferentes tecnologías y con diseños adaptados a un amplio rango de situaciones. Sin embargo, esta familia de sensores es muy susceptible a perturbaciones sonoras en el medio circundante. Ante la presencia de ruido, es posible aplicar algoritmos diseñados para mejorar la inteligibilidad del mensaje, pero a costa del detrimento en las características perceptuales [101]. A su vez, la técnica de sensado del EGG hace que esta señal sea muy robusta ante perturbaciones sonoras. No obstante, el adecuado posicionamiento de los electrodos es en general un proceso complejo, que depende tanto de la anatomía del cuello del individuo como de la experiencia de quien lleve a cabo esta tarea. Por su parte, la señal de VPC es robusta frente a perturbaciones sonoras y, además, el posicionamiento del sensor es relativamente sencillo. Es importante señalar que estas tres señales son potencialmente susceptibles a contaminación electromagnética, por lo que es necesario un diseño adecuado de la etapa de acondicionamiento [16].

En relación con la información representada en estas señales, podemos hacer una clara distinción entre la voz, por un lado, y las señales de EGG y VPC, por el otro. La voz es el resultado del proceso de fonación. Es una señal compleja generada por la acción combinada de las fuentes de excitación y del tracto vocal. Toda esta información es muy útil en comunicaciones, para el reconocimiento automático en el habla, y en interfaz *hombre-computadora*, entre otras aplicaciones [101, 132]. Sin embargo, su uso para el estudio de la dinámica glótica presenta algunas limitaciones, entre las que se encuentran [42, 128]:

- Posee una forma de onda compleja.
- Es muy susceptible a perturbaciones sonoras.
- La información glótica es enmascarada por la acción del tracto vocal.
- Pueden producirse sonidos de la voz sin actividad glótica.

Por su parte, las señales de EGG y de VPC presentan una forma de onda considerablemente más simple, en comparación con la voz [161, 179]. Además, la forma de onda de estas señales se puede interpretar en relación con la dinámica glótica. Esto último, se puede apreciar fácilmente en los ejemplos de las Figs. 2.9 y 2.10. Por ello, se recomiendan estas señales en algunas tareas como por ejemplo el estudio de la actividad glótica y la estimación de parámetros acústicos [14, 16, 179].

2.5. Bases de datos consideradas en esta tesis

Como cierre de este capítulo, describiremos brevemente los conjuntos de señales de la fonación empleadas para la realización de esta tesis de doctorado. Una parte de

este material fue desarrollado en el LSyDnL, con la finalidad de obtener señales adecuadas para la prueba y puesta a punto de los métodos propuestos y, a su vez, que puedan utilizarse con fines pedagógicos. Por otro lado, contamos también con una base de datos cedida al LSyDnL por una empresa extranjera, la cual utilizamos para realizar las diferentes simulaciones.

2.5.1. Base de datos desarrollada en el LSyDnL

En el LSyDnL de la FI-UNER se han llevado a cabo diferentes trabajos dedicados al estudio de las tres señales de la fonación presentadas anteriormente. Estos dieron como resultado la confección de una pequeña base de datos compuesta por señales de voz, de EGG y de VPC, registradas de forma simultánea. En su desarrollo, se contemplaron una serie de emisiones de interés en la fonoaudiología, abarcando vocales sostenidas neutras, transición entre las vocales con y sin pausa, una frase corta fonéticamente balanceada y la lectura de una fábula. Las dos últimas en español.

En primer lugar, diseñamos un protocolo *ad hoc* para el registro de estas tres señales, contemplando tanto el equipamiento como la infraestructura disponible en el LSyDnL. Luego, procedimos a la etapa de prueba y corrección de este protocolo. Como resultado, el protocolo desarrollado se materializó en un documento técnico [149], redactado de forma tal que pueda ser fácilmente implementado y, de ser necesario, modificado para agregar otro tipo de emisiones o para registrar otras señales.

Posteriormente, procedimos a registrar las señales de nuestro interés aplicando el protocolo desarrollado. En este proceso participaron 43 sujetos adultos, 14 mujeres y 29 hombres, quienes declararon no haber padecido patologías o trastornos vocales, ni haber sufrido ninguna disfonía vocal en los tres meses previos. Los ejemplos presentados en las diferentes figuras de la Sec. 2.4 forman parte del material generado.

Utilizamos esta base de datos para el estudio comparativo de las señales de voz, de EGG y de VPC. En particular, contrastamos el desempeño de estas señales para la detección de la actividad glótica, la estimación de f_0 y el cálculo de diferentes medidas de dosimetría vocal. Encontramos que, en las aplicaciones consideradas, las señales de EGG y de VPC poseen un desempeño comparable, dando lugar a estimaciones similares de los diferentes parámetros. Además, con estas señales se obtuvo un error de estimación considerablemente menor que con la señal de voz. Por otro lado, pudimos comprobar las diferentes características de la señal de VPC, que la vuelven atractiva para monitorear la actividad vocal y para la dosimetría. Los resultados alcanzados concordaron con los trabajos publicados recientemente en esta línea [112, 121, 179]. Es importante destacar que las actividades y los resultados presentados se desarrollaron en el marco del *Proyecto Final* del Bioing. Ariel Esteban Stassi [148], realizado en el LSyDnL bajo la dirección del autor de esta tesis doctoral. Asimismo, parte de los resultados dieron lugar a un trabajo científico presentado en un congreso internacional de la especialidad [150].

2.5.2. Base de datos de desórdenes de la voz (BDDV)

La base de datos de desórdenes de la voz (BDDV) es comercializada en la actualidad por la empresa *Kay Elemetrics Corp* [110]. Sin embargo, fue desarrollada originalmente en el *Massachusetts Eye and Ear Infirmary Voice and Speech Lab*. La BDDV

cuenta con registros de señales de voz, correspondiente a vocales /a/ sostenidas y a la lectura del texto en inglés *"Rainbow Passage"*. Estas señales se obtuvieron de 53 individuos sanos y de 657 pacientes con una amplia variedad de desórdenes de la voz de carácter orgánico, neurológico, traumático y psicogénico [10, 110, 124].

La población de participantes sanos estaba compuesta por 21 hombres y 32 mujeres, con edades de 38,81±8,49 y 34,16±7,87 años, respectivamente [2]. Las señales patológicas contienen muestras de una población de 169 hablantes masculinos, 238 femeninos y 247 de los que no se indican datos de género. Las edades son de 49,80 ± 17,46 y de 46,83 ± 17,41 años para el grupo de hablantes masculinos y femeninos, respectivamente. Todos los registros cuentan con su correspondiente historia clínica, reunida a partir de diferentes estudios específicos y de la opinión de los expertos. Además, se informan los resultados de la evaluación de la voz obtenidos con el programa *Multi-Dimensional Voice Program* (MDVP), comercializado por la misma empresa.

Los registros de los participantes con patologías vocales se grabaron en una cabina insonorizada, empleando una frecuencia de muestreo de 25 kHz y una cuantificación de 16 bits. Las señales de los sujetos normales fueron grabadas por separado en *Kay Elemetrics Corp* bajo condiciones acústicas equivalentes, con una frecuencia de muestreo de 50 kHz y una cuantificación de 16 bits. No se realizaron exámenes en búsqueda de desórdenes de la voz en los sujetos normales, pero ninguno de ellos presentaban antecedentes o acusaban haber desarrollado tales trastornos [124]. Las vocales sostenidas presentan una duración de 3 s y de 1 s para las señales normales y patológicas, respectivamente. A su vez, los registros correspondientes a la lectura del texto tienen una duración máxima de 12 s en ambas poblaciones.

2.6. Comentarios finales

Dedicamos este capítulo a introducir y discutir las estructuras y los procesos involucrados en la fonación. En primer lugar, describimos la anatomía del aparato fonador, considerando los principales órganos que lo componen. Seguidamente, analizamos el rol de cada órgano del aparato fonador en la producción de los sonidos de la voz.

Durante el habla, una persona es capaz de generar una gran variedad de sonidos. A su vez, estos sonidos se producen de forma continua, como resultado de la articulación en el aparato fonador. Por ello, en la Sec. 2.3 presentamos una jerarquización de los sonidos de la voz basada en unidades fonéticas. Esto permite asociar los fonemas con un conjunto de características acústicas presentes en los sonidos de la voz. Discutimos también los registros vocales empleados para caracterizar la fuente de excitación glótica.

Luego, en la Sec. 2.4 describimos tres señales biomédicas relacionadas con la fonación, prestando especial atención a las técnicas de sensado, a la información que en ellas se preserva y a su importancia desde el punto de vista de la ingeniería. Finalmente, en la Sec. 2.5 presentamos las dos bases de datos empleadas en el desarrollo de esta tesis doctoral, indicando sus características principales. Aprovechamos también las dos últimas secciones para comentar brevemente algunas experiencias relacionadas con el estudio de las señales biomédicas de la fonación y las diferentes técnicas de sensado, llevadas a cabo recientemente en el LSyDnL.

[2] La edad se presenta bajo la estructura: *edad promedio* ± *desvío estándar*.

Capítulo 3

Modelado del proceso de fonación

3.1. Introducción

En el capítulo anterior, analizamos las principales estructuras anatómicas que conforman el aparato fonador y describimos los procesos fisiológicos involucrados en la generación de la voz. Estudiamos también los diferentes sonidos de la voz, considerando el mecanismo de excitación y la configuración del tracto vocal involucrados en la fonación. Dedicaremos este capítulo a presentar y analizar las estrategias más importantes utilizadas en la actualidad para el modelado de la fonación. En esta exposición, prestaremos especial atención a aquellos conceptos que serán cruciales para el desarrollo de esta tesis doctoral.

Los modelos son construcciones ideales desarrolladas para el estudio, el análisis, la simulación y la predicción de procesos o sistemas reales. Pueden ser abstractos, de naturaleza física o, inclusive, resultar de la combinación de ambas formas. Constituyen herramientas útiles que permiten explicar un fenómeno bajo estudio, de forma concisa y precisa, considerando sus elementos constitutivos, las relaciones que existen entre ellos y las diferentes transformaciones que se producen. Por ello, el modelado es un proceso sumamente importante en las disciplinas científicas.

Usualmente, en la ingeniería los modelos se construyen combinando las leyes de la física con diferentes estructuras matemáticas. Otra alternativa, consiste en describir un proceso usando como analogía sistemas mecánicos, acústicos, ópticos o eléctricos, entre otros. Estas dos alternativas han sido ampliamente utilizadas para el desarrollo de modelos de la fonación [42, 128, 130, 133, 164, 165]. Es importante señalar que todo modelo es una representación parcial, acotada y sesgada de la realidad. Luego, su utilidad dependerá tanto de la complejidad en su formulación como de la aplicación considerada. Como regla general, es posible modificar un modelo para volverlo más realista, considerando una formulación más compleja, a costa de disminuir su aplicabilidad [50, 114, 133].

A lo largo del tiempo han surgido diferentes alternativas para estudiar y simular la fonación. Los primeros dispositivos capaces de producir sonidos similares a los de la voz humana datan de la segunda mitad del siglo XVIII [49, 173]. Estos trabajos revelaron la importancia del tracto vocal en la fonación, como lugar principal donde se produce la articulación acústica [126]. Sin embargo, fue durante el siglo XIX que se produjo el auge de las *máquinas parlantes*, dispositivos mecánicos cuyo diseño estaba inspirado principalmente en la acústica de los instrumentos de viento [42, 143]. Estos desarrollos sentaron las bases de la teoría acústica de la fonación, la cual establece

que el tracto vocal actúa como un resonador acústico, caracterizado por un conjunto de frecuencias de resonancia, que modifica la información frecuencial de las diferentes fuentes acústicas desarrolladas a lo largo del aparato fonador [54, 143].

Posteriormente, a comienzo del siglo XX aparecieron los primeros sintetizadores de voz fabricados completamente con tecnología electrónica [144]. Algunos de ellos eran capaces de generar palabras o frases cortas inteligibles. A partir de estos trabajos, los científicos comenzaron a prestar mayor atención a las características subjetivas que se perciben en la voz. Los logros alcanzados durante este período sirvieron para el desarrollo de la teoría *fuente* y *filtro* de la fonación [38, 54]. Esta teoría permite un análisis simplificado de este proceso aplicando los conceptos de la teoría de señales y sistemas [42, 130]. Más adelante en este capítulo, describiremos con mayor detalle esta teoría ya que es esencial para las contribuciones propuestas en esta tesis de doctorado.

A fines de la década de 1960, y gracias al avenimiento de las computadoras digitales, comenzaron a implementarse los primeros sintetizadores digitales de voz. Desde entonces, se han propuesto diversas estrategias de modelado, más versátiles y realistas, que permitieron entender con mayor precisión los procesos involucrados en la fonación [128, 163]. La combinación de estos modelos con diferentes técnicas para el procesamiento de señales ha posibilitado grandes avances en el análisis y procesamiento digital del habla y de otras señales de la fonación (ejemplos de este tipo de señales fueron presentadas en la Sec. 2.4). Estos avances han sido muy importantes en la evolución de las diferentes tecnologías de comunicación disponibles actualmente [143]. Para mayor información respecto a la evolución del modelado de la fonación y otros tópicos relacionados el lector puede referirse a [38, 42, 56, 128, 133, 143].

En las siguientes secciones, describiremos brevemente las dos teorías de la fonación más importantes en la actualidad, las cuales son ampliamente utilizadas en la ingeniería. Es importante destacar que si bien ambas aportan un marco conceptual para analizar e imitar la fonación, se basan en hipótesis y principios diferentes. Por ello, pueden considerarse como complementarias entre sí.

3.2. Teoría *mioelástica*, *aerodinámica* y *acústica*

La teoría *mioelástica*, *aerodinámica* y *acústica* de la fonación surgió de los esfuerzos dedicados a interpretar y explicar, de modo minucioso, los diferentes procesos fisiológicos involucrados en la generación de la voz. Describe a la fonación considerando la morfología de las estructuras anatómicas que conforman el aparato fonador, sus interrelaciones y las transformaciones físicas que se producen [153, 165].

Como su nombre lo indica, esta teoría combina una variedad de principios físicos elementales propios de la mecánica, la aerodinámica y la acústica [165]. Éstos permiten modelar las transformaciones que suceden a lo largo del aparato fonador, comenzando por el flujo de aire expulsado desde los pulmones hasta finalizar con la radiación en los labios de la onda sonora que constituye la voz. Principalmente, se estudian en detalle el movimiento oscilatorio de las cuerdas vocales, la modulación del flujo de aire en la glotis, la distribución de la presión en las vías aéreas, la transmisión acústica en el tracto vocal y la emisión de la voz al medio circundante [151, 152, 153, 164].

El mecanismo responsable de las oscilaciones sostenidas de las cuerdas vocales

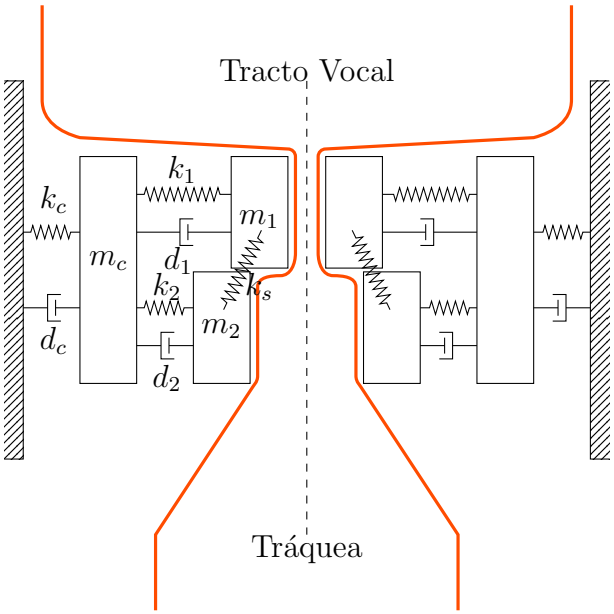


Figura 3.1: Modelo mecánico de las cuerdas vocales considerando el esquema *cuerpo* y *cubierta* propuesto por Titze. Esquema inspirado a partir de [155].

es complejo y depende, en general, de las estructuras anatómicas involucradas, del flujo de aire que atraviesa la glotis y de la distribución de presiones en el tracto vocal. Por este motivo, se recurre a dos simplificaciones importantes. En primer lugar, se supone que las cuerdas vocales son iguales y simétricas entre sí. A su vez, las cuerdas vocales se analizan a partir del esquema *cuerpo* y *cubierta* propuesto por Titze [153, 155, 165], descrito en la Sec. 2.2.1. Estas hipótesis permiten generar modelos mecánicos no lineales de las cuerdas vocales, como por ejemplo el mostrado en la Fig. 3.1. En este esquema, la masa m_c corresponde al *cuerpo* y, a su vez, la cubierta se modela con dos masas más pequeñas, una superior m_1 y otra inferior m_2. Estas masas se acoplan entre sí y con una pared rígida mediante elementos mecánicos elásticos $\{k_c, k_s, k_1, k_2\}$ y amortiguadores $\{d_c, d_1, d_2\}$, representando las propiedades físicas e interrelaciones encontradas en las cuerdas vocales [165].

Las cuerdas vocales se acoplan entre sí y con el flujo de aire a partir de fenómenos físicos y aerodinámicos complejos [152, 155]. La diferencia de presión acústica por arriba y por debajo de la glotis determina el flujo de aire y las fuerzas actuantes sobre las cuerdas vocales. A su vez, las oscilaciones de éstas determinan la apertura en la glotis y, como consecuencia, modulan el flujo de aire hacia el sistema supraglótico [164]. En la Sec. 2.2.2 se describió el comportamiento de las cuerdas vocales.

Aun cuando el modelo mostrado en la Fig. 3.1 es relativamente sencillo, se ha demostrado que captura satisfactoriamente los principales modos de oscilación de las cuerdas vocales [153, 155]. Estos modos corresponden, por un lado, al movimiento lateral de las cuerdas vocales y, por otro lado, a la onda longitudinal en la superficie de la mucosa (ver secuencia explicativa de la Fig. 2.5). A su vez, existen otras representaciones más precisas de las cuerdas vocales, que permiten estudiar en mayor detalle la dinámica de estos órganos [60, 72, 95, 178]. Sin embargo, se demostró que la dinámica de las cuerdas vocales queda determinada por 2 o 3 modos principales de oscilación y que se requieren al menos tres grados de libertad para su representación [18, 153]. Por ello, esquemas como el de la Fig. 3.1 resultan satisfactorios para simular

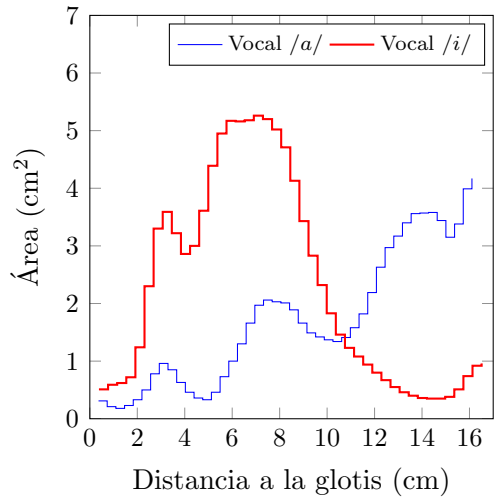


Figura 3.2: Representaciones unidimensionales del tracto vocal de un sujeto masculino, correspondientes a las configuraciones adoptadas para generar las vocales sostenidas /a/ e /i/. Información reproducida de [154].

las características principales de las oscilaciones de las cuerdas vocales.

Por su parte, los conductos que conforman las vías aéreas subglóticas y supraglóticas se representan de forma discreta, a partir de la concatenación de pequeñas unidades cilíndricas con propiedades físicas similares, de longitud fija y de área transversal variable. Estos elementos reciben el nombre de *tubulets* y son guías de onda que permiten la transmisión de una onda acústica [42, 151, 163].

Cuando una onda acústica arriba a una unión entre dos *tubulets* con igual área transversal, ésta se transmite completamente. En caso que las áreas transversales sean diferentes, parte de la onda se transmite y parte se refleja. Esto se repite en cada unión, dando lugar a dos frentes de onda: uno que avanza hacia los labios y otro que retrocede hacia la laringe. Se genera así una distribución irregular de la presión a lo largo del aparato fonador [153, 165]. Este fenómeno, es el principal responsable de la modulación y amplificación de la información espectral de la voz por parte del tracto vocal [42, 130, 163]. En los labios, la onda acústica es emitida al medio circundante. Este fenómeno de radiación se modela mediante una impedancia acústica con un comportamiento similar a un filtro pasa altas [128, 152].

Para implementar la estrategia comentada en el párrafo anterior, es importante contar con una buena caracterización de los conductos subglóticos y del tracto vocal. Esta caracterización se logra con el cálculo de la distribución de las áreas transversales de estos conductos, para diferentes configuraciones del sistema fonador. En la actualidad, esta función se obtiene del procesamiento de imágenes de rayos X o de resonancia magnética nuclear [154, 156, 165]. En la Fig. 3.2 presentamos dos distribuciones de las áreas transversales del tracto vocal de un sujeto masculino. Estas curvas corresponden a las configuraciones adoptadas para generar las vocales /a/, línea fina, e /i/, línea gruesa. La información para esta figura se tomó de [154].

Originalmente, se trabajó bajo las hipótesis de que los *tubulets* presentaban paredes rígidas y que en ellos no se producían pérdidas. Sin embargo, han surgido alternativas superadoras, que permiten considerar pérdidas ocasionadas por: deformaciones en las paredes, viscosidad del aire, radiación acústica en las paredes y

conducción térmica [152, 153, 165]. Es importante destacar que los modelos unidimensionales aquí descriptos son adecuados para representar la información espectral del tracto vocal en el rango de frecuencias acústicas bajas y medias (0-5 kHz). Para modelar la información espectral para frecuencias mayores se recomienda utilizar estrategias más complejas [11, 12, 20, 36].

De lo expuesto en esta sección, podemos inferir que para aplicar la teoría *mioelástica*, *aerodinámica* y *acústica* de la fonación es necesario contar con información muy específica, y muchas veces difícil de obtener, del aparato fonador [154, 164, 178]. Por ejemplo, para modelar la generación de un sonido con esta teoría es necesario conocer las dimensiones de la cuerdas vocales, sus propiedades biomecánicas y la forma tanto del tracto vocal como de las vías aéreas subglóticas, entre otros. Otra desventaja es que los modelos desarrollados a partir de esta teoría presentan una alta complejidad desde el punto de vista computacional [60, 72, 152]. Todo esto, hace que esta representación de la fonación resulte de poca utilidad para el análisis y procesamiento de señales de voz reales. Sin embargo, se ha demostrado que esta estrategia permite generar señales de voz con una muy buena calidad perceptual y, por otro lado, es útil para simular ciertos trastornos vocales [60, 151, 153].

3.3. Teoría *fuente* y *filtro*

La teoría presentada en la sección anterior resulta muy atractiva ya que brinda una alternativa para el estudio de la fonación, considerando directamente las principales estructuras y fenómenos involucrados en este proceso. Sin embargo, hasta donde el autor de esta tesis conoce, ésta no ha sido utilizada plenamente en los sistemas ingenieriles actuales. Esto se debe, principalmente, a las limitaciones expuestas en el párrafo anterior. En esta sección presentaremos una teoría alternativa de la fonación, la cual es aceptada universalmente por la comunidad científica y es estudiada ampliamente en la literatura específica.

Propuesta originalmente por Fant en la década de 1960, la teoría *fuente* y *filtro* (TFF) estudia la fonación empleando conceptos propios del campo de señales y sistemas [27, 54]. Se caracteriza por presentar una formulación relativamente sencilla y por estar respaldada por un sólido marco conceptual. A su vez, ha demostrado ser fundamental en la evolución de las tecnologías involucradas en la comunicación, el procesamiento del habla y en la síntesis de voz. El lector interesado en esta teoría y sus aplicaciones, puede recurrir a la abundante bibliografía específica dedicada al respecto, por ejemplo [42, 101, 128, 130, 133, 144, 163].

Esta teoría considera a la señal de voz como la salida de un sistema formado por componentes específicos con una estructura sencilla. Por esto último, en la bibliografía se la suele denominar *modelo análogo en el terminal*[1]. Esta expresión hace referencia a que aun cuando los componentes de la TFF son, a lo sumo, superficialmente análogos a las estructuras que forman el aparato fonador, ambos sistemas generan señales de voz comparables [42, 128]. En rigor, esta teoría se centra en capturar, dentro de un esquema controlable, las características perceptuales relevantes de la señal de voz, en lugar de modelar en detalle su forma de onda [27, 42].

En la Fig. 3.3 podemos observar una representación esquemática de la fonación de acuerdo con esta teoría. Este esquema es una versión levemente modificada respecto

[1]En inglés, "terminal-analog model".

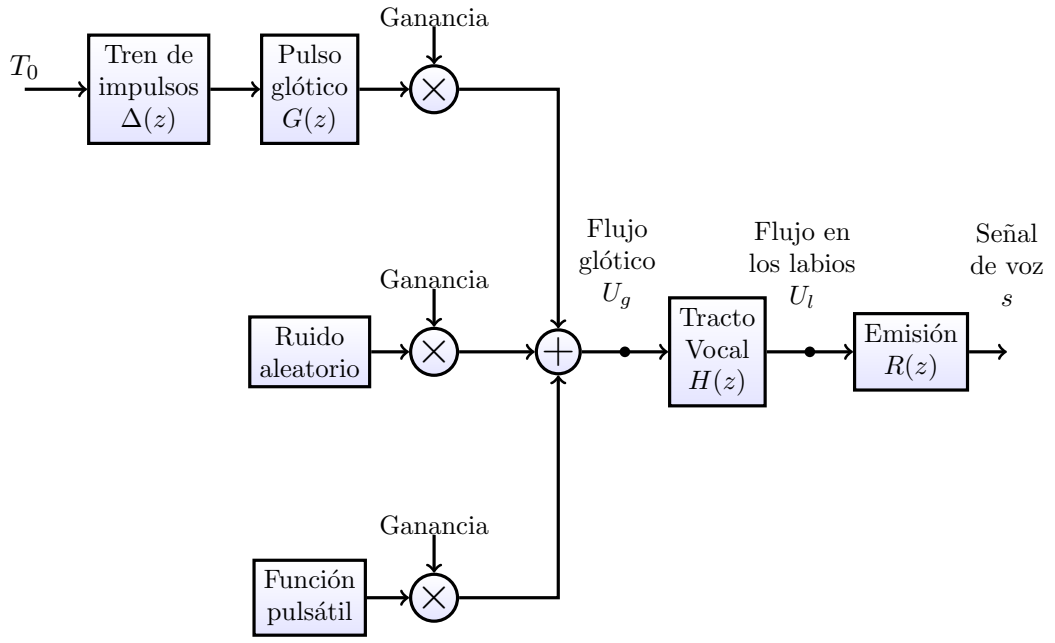


Figura 3.3: Esquema representativo de la fonación de acuerdo a la teoría *fuente* y *filtro* de la fonación.

a las representaciones tradicionales [42, 128, 130]. A continuación, describiremos brevemente su interpretación. En primer lugar, se supone que el flujo de aire que atraviesa la glotis U_g es la fuente del sistema y resulta de la composición de un conjunto de señales con comportamientos notablemente diferentes. En la TFF no se analizan los procesos y las transformaciones que ocurren en la región subglótica, sino que sólo se modela el flujo de aire resultante que atraviesa la glotis.

Luego, la fuente U_g es filtrada por el tracto vocal, representado por la función de transferencia $H(z)$. Este filtro modifica la información frecuencial de U_g, amplificando o atenuando diferentes regiones de su espectro. Se obtiene como resultado una nueva señal U_l que representa el flujo de aire en los labios. Finalmente, el flujo de aire U_l se transforma en una onda acústica en el campo lejano, por el proceso de emisión desde los labios al medio circundante. Se obtiene como resultado la señal de voz s. Generalmente, la emisión se representa con la función de transferencia $R(z)$, que se caracteriza por una baja impedancia y un comportamiento frecuencial similar a un filtro pasa altas [42]. En la práctica, suele definirse a $R(z) = 1 - \gamma_R\, z^{-1}$, con $\gamma_R \lessapprox 1$. En este documento seguiremos este criterio. Es importante destacar que para fonemas nasales se modifica $H(z)$ o se agrega un componente, modelando el comportamiento del tracto nasal [144].

De acuerdo a lo expuesto en la Sec. 2.3, el mecanismo de excitación acústica se modifica dependiendo del fonema considerado. Del mismo modo, es necesario adecuar el comportamiento de U_g según el fonema a modelar [27, 128]. En el caso de las vocales, la principal fuente de excitación consiste en un flujo de aire pulsátil que atraviesa la glotis (ver Sec. 2.3.1). Este mecanismo de excitación se simboliza en la rama superior de la Fig. 3.3. Se supone inicialmente un tren de impulsos cuasiperiódicos, con período fundamental T_0. Se supone también que esta señal tiene transformada $\mathcal{Z}$, representada en adelante por $\Delta(z)$. Esta señal es luego modificada

por el filtro *moldeador del pulso glótico* $G(z)$[2]. Se obtiene así, una fuente U_g formada por un tren de pulsos glóticos. Se define como $G_\Delta(z)$ a la transformada $\mathcal{Z}$ de esta señal, donde $G_\Delta(z) = \Delta(z)\,G(z)$. Más adelante en este capítulo, analizaremos la forma de onda y las propiedades espectrales de los pulsos glóticos.

Para las consonantes, existen dos mecanismos de excitación principales (ver Sec. 2.3.2). Por un lado, se desarrollan turbulencias en el flujo de aire que dan lugar a fuentes acústicas (por ejemplo, en los fonemas fricativos sordos). En la TFF, este fenómeno se modela como ruido estocástico con espectro blanco o coloreado [27, 42]. Por otro lado, el cierre del tracto vocal seguido por una liberación abrupta de la presión da lugar a una fuente acústica semejante a una explosión (por ejemplo, en consonantes oclusivas). En la TFF, esto se suele representar con una función pulsátil o ráfaga transitoria [128, 146, 176]. Estos dos casos se simbolizan en las ramas central e inferior, respectivamente, de la Fig. 3.3. Se supone también que existe la transformada $\mathcal{Z}$ de ambas excitaciones. En adelante, las representaremos como $N(z)$, indistintamente. Es importante destacar que estos tres modelos para las excitaciones acústicas pueden combinarse entre sí para generar una fuente U_g más compleja y realista, aun cuando individualmente presentan una estructura simple. Así, la fuente resultante dependerá de la ponderación, o peso relativo, de cada modelo de excitación. Para mayor información, referirse a [27, 128, 176] y sus referencias correspondientes.

En el marco de la TFF, el filtro del tracto vocal, $H(z)$, suele considerarse lineal e invariante en el tiempo, en ventanas o segmentos de señal de voz de corta duración (20-50 ms). Asimismo, se permite que $H(z)$ varíe para cada ventana analizada. La razón de esto último es que los cambios en el tracto vocal son lentos, en comparación con la dinámica de la señal de voz [42, 54, 128, 163]. Por otro lado, los componentes de la TFF se suponen independientes entre sí (ver Fig. 3.3). Es decir, se desprecia toda interacción entre estos sistemas, bajo la hipótesis de que no repercuten significativamente en el desempeño del modelo [54]. Esta última hipótesis ha sido fuertemente criticada por la comunidad científica, a lo largo del tiempo. Diferentes autores han demostrado que la influencia entre el flujo de aire glótico y el tracto vocal es fuerte y extremadamente compleja, y que ésta no puede representarse adecuadamente con técnicas lineales. [33, 55, 166]. A su vez, la TFF no contempla ningún tipo de realimentación o comportamiento no lineal en sus componentes [42, 128]. Tampoco se contemplan las fuentes acústicas secundarias, que se producen a lo largo del tracto vocal y que contribuyen a la generación de la señal de voz [91, 120].

De lo expuesto hasta aquí, se puede apreciar que la TFF aporta un marco conceptual interesante para el análisis y estudio de la fonación. Sin lugar a dudas, una de sus principales virtudes es su estructura simple y modular (ver Fig. 3.3). Ésta ha permitido importantes avances en el campo de las telecomunicaciones, en particular en la codificación y la transmisión de voz, ya que permite capturar toda la información relevante en un conjunto pequeño de parámetros [68, 144]. Sin embargo, en las implementaciones clásicas de la TFF sólo se preserva adecuadamente la información espectral en el rango de frecuencias acústicas bajas y medias, por debajo de los 5 kHz [42, 144]. Por ello, desde hace un tiempo se está trabajando en implementaciones mejores y más versátiles [132, 133, 144].

[2]En inglés, "glottal pulse shaping filter" [42].

3.3.1. Interpretación temporal y frecuencial

En esta sección, aplicaremos conceptos del campo de señales y sistemas discretos en el marco de la TFF, con el fin profundizar en la interpretación de las características espectrales más importantes del proceso de fonación. Recordemos que, de acuerdo a la TFF, los bloques del esquema de la Fig. 3.3 deben ser lineales, invariantes en el tiempo e independientes entre sí. Sea $S(z)$ la transformada $\mathcal{Z}$ de la señal de voz s, es decir $S(z) = \mathcal{Z}\{s[n]\}$. Entonces, para fonemas vocales la información espectral de la señal de voz queda representada por:

$$S(z) = G_\Delta(z)\, H(z)\, R(z). \tag{3.1}$$

De forma similar, la información espectral para fonemas consonantes se representa como:

$$S(z) = N(z)\, H(z)\, R(z). \tag{3.2}$$

En ambas expresiones $z \in \mathbb{C}$. Note el lector que todas las funciones involucradas fueron presentadas anteriormente. Es posible, también, generar expresiones análogas para la información temporal de la señal de voz. En este caso, la señal de voz s resultará de la convolución en el tiempo de U_g con las funciones de *respuesta al impulso* de $H(z)$ y de $R(z)$ [42, 128].

De acuerdo con la teoría para señales discretas, la transformada $\mathcal{Z}$ y la transformada de Fourier de una señal discreta coinciden para $z = e^{i\omega}$, donde i es la unidad imaginaria, $\omega = 2\pi f$ es una variable real y f representa la variable frecuencia en Hercios. Así, la transformada de Fourier de la señal de voz resulta $S(f) = S(z)|_{z=e^{2\pi f i}}$. Esto es cierto siempre y cuando la región de convergencia de $S(z)$ incluya a la circunferencia unitaria $|z| = 1$ en el plano z [42, 104]. Suponiendo que se cumple esta condición, el espectro de amplitud de la señal de voz para un fonema vocal se calcula de la siguiente forma:

$$|S(f)| = |G_\Delta(f)|\, |H(f)|\, |R(f)|, \tag{3.3}$$

donde $G_\Delta(f)$ es la transformada de Fourier del tren de pulsos glóticos. Para fonemas consonantes, el espectro de amplitud se obtiene reemplazando $G_\Delta(f)$ por la función $N(f)$ en la Ec. 3.3, suponiendo que ésta exista.

Analizando la Ec. 3.3, podemos apreciar que $|S(f)|$ se obtiene de la combinación (producto) de los espectros de amplitudes de cada uno de los sistemas que componen la TFF. Esto último se representa gráficamente en la Fig. 3.4, válido para fonemas vocales. En la fila central, mostramos nuevamente el esquema de cajas de la fonación de acuerdo a la TFF. En la fila superior, podemos observar ejemplos de señales sintetizadas correspondientes al tren de pulsos glóticos U_g, al flujo de aire U_l en los labios y a la señal de voz s resultante. A su vez, en la fila inferior presentamos los espectros de amplitud del tren de pulsos glóticos $|G_\Delta(f)|$, del filtro del tracto vocal $|H(f)|$, del filtro de emisión $|R(f)|$ y de la señal de voz $|S(f)|$, para el rango de frecuencias 0-3,5 kHz. Todas estas señales corresponden a una vocal /a/ sostenida.

En primer lugar, podemos observar el comportamiento cuasiperiódico de U_g y como éste se manifiesta en la estructura armónica de $|G_\Delta(f)|$. Luego, podemos apreciar claramente el efecto de filtrado característico del tracto vocal, que amplifica o atenúa diferentes regiones del espectro de amplitud de $|G_\Delta(f)|$. Apreciamos también las frecuencias formantes, determinadas por las posiciones de los picos de $|H(f)|$. Se

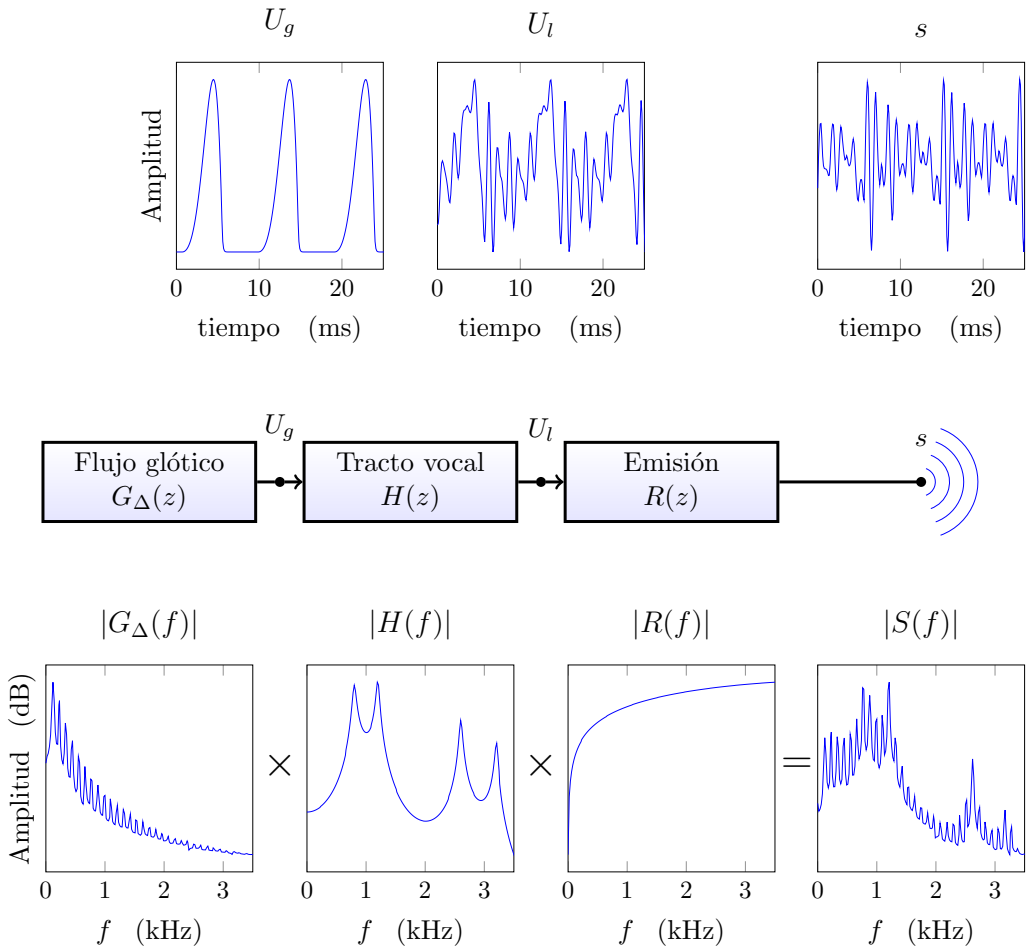


Figura 3.4: Análisis temporal y frecuencial de la fonación en el marco de la TFF, considerando un fonema vocal. *Arriba:* comportamiento temporal de las señales involucradas. *Centro:* esquema de cajas de la fonación. *Abajo:* espectros de amplitud de los sistemas considerados en la TFF.

genera así una señal en los labios U_l con una mayor riqueza espectral que la señal glótica U_g. Por otro lado, $R(z)$ se comporta como un filtro pasa altas y es el responsable de acentuar más aún la información para frecuencias altas, a la vez que elimina las componentes para frecuencias muy bajas. Todo esto da como resultado la señal de voz s de la Fig. 3.4, cuya forma de onda temporal se muestra en el extremo superior derecho y cuyo espectro de amplitud se grafica en el extremo inferior derecho. Podemos comparar estas gráficas con las correspondientes al ejemplo de la Fig. 2.6. Así, inferimos de que manera se modelan en la TFF las principales características de la señal de voz (ver Sec. 2.4.1).

Trabajando de forma similar, se pueden estudiar los fonemas consonantes en el marco de la TFF. Sin embargo, no profundizaremos en esta línea, debido a que los trabajos desarrollados en esta tesis de doctorado se centran principalmente en el estudio de fonemas vocales. El lector interesado puede dirigirse a la bibliografía existente al respecto, por ejemplo: [42, 128, 130, 133]. Dedicaremos las dos próximas secciones a describir diferentes alternativas para modelar, de acuerdo a la TFF, el filtro del tracto vocal y la fuente glótica.

3.4. Filtro del tracto vocal

En esta sección brindaremos mayores precisiones respecto al filtro del tracto vocal, denominado $H(z)$ en la sección anterior. Usualmente, se lo representa a partir de un filtro discreto, lineal e invariante en el tiempo. Sin embargo, desde hace un tiempo han surgido otras alternativas, que permiten una representación más versátil, a la vez que toman en cuenta características propias de la señal de voz o de la fonación.

3.4.1. Modelado lineal e invariante en el tiempo

Comenzaremos describiendo el modelado lineal e invariante en el tiempo de la fonación, por ser la representación más popular y difundida en la literatura. En esta estrategia, se toma como hipótesis principal que la fonación es un proceso estocástico. Luego, la voz resultante de este proceso es, a su vez, una señal estocástica [104]. De acuerdo a esto, la fonación puede analizarse aplicando modelos autorregresivos (AR) de orden ρ, caracterizados por la siguiente función de transferencia [42, 130]:

$$\Theta(z) = \frac{G_0}{\sum_{l=0}^{\rho} a_l \, z^{-l}} = \frac{G_0}{1 + a_1 \, z^{-1} + \ldots + a_\rho \, z^{-\rho}}, \tag{3.4}$$

donde G_0, a_0, a_1, $\ldots$, $a_\rho \in \mathbb{R}$ y $a_0 = 1$. Aquí, G_0 representa un término de ganancia global. Es importante destacar que $\Theta(z)$ no corresponde directamente con la $H(z)$ descripta por la TFF. Más adelante, ahondaremos en este punto.

La función racional $\Theta(z)$ presenta en el denominador un polinomio con coeficientes reales en la variable z^{-1}. Luego, sus polos serán números reales puros o complejos conjugados. Esto último permite escribir a $\Theta(z)$ como sigue:

$$\Theta(z) = \frac{G_0}{\prod_{l=1}^{\rho}(1 - p_l \, z^{-1})} = \frac{G_0 \, z^{\rho}}{\prod_{l=1}^{\rho}(z - p_l)}, \tag{3.5}$$

donde p_l son los polos de $\Theta(z)$. Esto último indica que $\Theta(z)$ presenta, en el plano z, ρ polos para $z \in \{p_1, p_2, \ldots, p_\rho\}$ y un cero de orden ρ para $z = 0$. Por todo esto, los modelos AR reciben también el nombre de *modelos de polos*.

La ubicación de los polos cobra un rol preponderante en la calidad del modelado. Para que una representación de la forma (3.4) sea válida debe dar lugar a un sistema discreto de *fase mínima*, es decir, un sistema discreto causal y estable cuyo sistema inverso sea, a su vez, causal y estable. Esto último se garantiza si todos los polos se hallan en el interior del círculo unitario, es decir, si $|p_l| < 1 \;\; \forall l = \{1, 2, \ldots, \rho\}$ [108, 158]. A su vez, los polos son útiles para obtener estimaciones de las formantes y de sus anchos de banda, a partir de las siguientes expresiones:

$$F_i = \left(\frac{f_s}{2\,\pi}\right) \tan^{-1}\left(\frac{\mathrm{Im}(p_i)}{\mathrm{Re}(p_i)}\right), \quad \text{y} \quad B_i = -\left(\frac{f_s}{\pi}\right) \ln|p_i|, \tag{3.6}$$

donde f_s es la frecuencia de muestreo de la señal de voz, F_i es la i-ésima formante, B_i su correspondiente ancho de banda y p_i es el i-ésimo polo ubicado en el semiplano superior del plano z ($0 < \mathrm{Arg}(p_i) < \pi$) [128].

La gran popularidad de los modelos autorregresivos como representaciones del tracto vocal se debe, principalmente, a las siguiente razones [104, 133]:

1. Este tipo de modelo surge naturalmente del estudio del tracto vocal como una concatenación de *tubulets* de paredes rígidas y con áreas transversales variables.

2. La información espectral más relevante de la señal de voz queda completamente codificada en un conjunto pequeño de parámetros.

3. Existen métodos muy potentes que permiten estimar los parámetros del filtro $\Theta(z)$ usando como única información la señal de voz.

El ítem 3 del párrafo anterior merece ser discutido con mayor detalle. En general, los métodos para la estimación de los parámetros de $\Theta(z)$ se basan en el *análisis predictivo lineal*,[3] que es un caso particular del método de *regresión lineal* propio de la estadística. En este análisis, se considera que la señal de voz en un instante dado puede predecirse a través de una combinación lineal de un conjunto pequeño de valores pasados de la misma señal [104, 133]. Matemáticamente, se trabaja con la siguiente ecuación en diferencias lineal [128]:

$$s[n] = -\sum_{l=1}^{\rho} a_l \, s[n-l] + G_0 \, e[n], \tag{3.7}$$

donde e es una función de excitación desconocida. En estadística, e recibe los nombres de *perturbación*, *error de predicción* o *residuo* y, a su vez, se suele adoptar $e[n] \sim \mathcal{N}(0, \sigma_e^2)$ con $G_0 = 1$. Para el procesamiento del habla en particular, normalmente se supone que la excitación e toma una de las siguientes formas [42, 104]:

1. Tren de impulsos con período instantáneo T_0 para fonemas vocales.

2. Ruido aleatorio no correlacionado con media cero y varianza unitaria para fonemas consonantes.

Consideremos el vector $\mathbf{a} = \begin{pmatrix} a_1 & a_2 & \dots & a_\rho \end{pmatrix}^T$ formado por los coeficientes del modelo. En el marco del *análisis predictivo lineal*, se calcula el vector $\mathbf{a}$ resolviendo la siguiente ecuación lineal:

$$\hat{\mathbf{R}}_s \, \mathbf{a} = -\hat{\mathbf{r}}_s, \tag{3.8}$$

donde $\hat{\mathbf{R}}_s$ y $\hat{\mathbf{r}}_s$ son, respectivamente, estimaciones de la matriz de autocorrelación y del vector de autocorrelaciones de la señal de voz s, para los retardos (o corrimientos) $1, 2, \dots, \rho$ [108, 158]. En la bibliografía, la ecuación anterior recibe el nombre de *ecuación normal*. Existen diferentes alternativas para el calculo de $\hat{\mathbf{R}}_s$ y $\hat{\mathbf{r}}_s$, siendo los más populares el *método de la autocorrelación* y el *método de la covarianza en intervalos de cierre glótico* [5, 46, 104, 174].

Por otro lado, el cálculo del vector $\mathbf{a}$ a partir de la expresión (3.8) se denomina método de *codificación predictiva lineal*[4] (LPC) [42, 133]. La ganancia G_0 se calcula de forma tal que la potencia de la señal de voz coincida con la potencia de la representación obtenida con el modelo (3.7) [128]. Luego, se estiman las formantes y los respectivos anchos de banda aplicando las expresiones (3.6). Sin embargo, se ha demostrado que la técnica de LPC arroja estimaciones sesgadas de las formantes, problema que se acentúa para señales con una frecuencia fundamental alta [104, 105].

[3] En inglés, "linear prediction analysis".
[4] En inglés, "linear prediction coefficients".

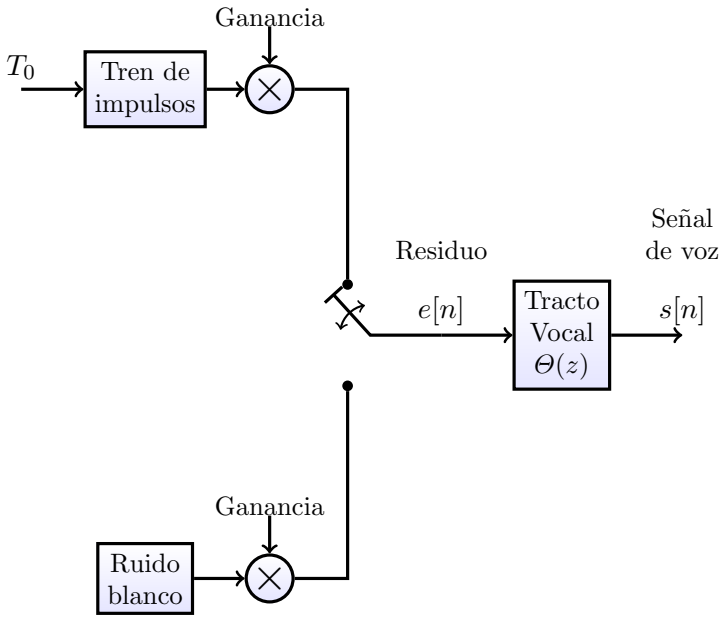


Figura 3.5: Esquema representativo del *análisis predictivo lineal* de la señal de voz.

Por ello, han surgido diferentes alternativas con el objetivo de mejorar la estimación de los parámetros del modelo, basados en medidas espectrales alternativas o en ponderar la información temporal de la señal de voz [6, 52, 102, 103].

Note el lector que las expresiones (3.4) y (3.7) dependen ambas del mismo conjunto de parámetros. Aplicando la transformada $\mathcal{Z}$ en la expresión (3.7) y suponiendo que existe $E(z) = \mathcal{Z}\Big\{e[n]\Big\}$, se puede demostrar que la función de transferencia $\Theta(z)$ corresponde al cociente:

$$\Theta(z) = \frac{S(z)}{E(z)}. \tag{3.9}$$

Esta expresión permite interpretar el modelo $\Theta(z)$ en el marco de la TFF. Tomando en consideración las expresiones (3.7) y (3.9) junto con la formas de excitación e consideradas en el procesamiento del habla, puede inferirse que el modelo propuesto es una aproximación simplificada de la TFF. En la Fig. 3.5 presentamos un diagrama de esta situación. Comparando esta figura con el esquema original para la TFF (ver Fig. 3.3) podemos identificar dos diferencias importantes [27, 42]:

1. La información de la forma de los pulsos glóticos, del tracto vocal y de la emisión en los labios se captura con un único modelo. Es decir, se supone la siguiente igualdad:
$$\Theta(z) = G(z)\, H(z)\, R(z).$$

2. Las formas de excitación contempladas en el modelo $\Theta(z)$ son más simples que las consideradas en la TFF. Sólo se contemplan excitaciones cuasiperiódicas o turbulentas.

De lo expuesto arriba, podemos concluir que el modelo $\Theta(z)$ surge naturalmente de la teoría acústica de la fonación y que el *análisis predictivo lineal* da lugar a

una estimación óptima de sus parámetros, lo que resulta muy atractivo en la práctica. Sin embargo, debemos advertir que este modelo es, en rigor, una aproximación simplificada de los diferentes elementos involucrados en la TFF.

Presentaremos a continuación otra representación lineal e invariante en el tiempo muy utilizada en el estudio de señales de habla. Nos referimos a los modelos autorregresivos y de media móvil (ARMA) de orden (τ, ρ), representados por la función de transferencia [128, 174]:

$$\Theta_{ARMA}(z) = \frac{\sum_{k=0}^{\tau} b_k\, z^{-k}}{\sum_{l=0}^{\rho} a_l\, z^{-l}} = \frac{b_0 + b_1\, z^{-1} + \ldots + b_\tau\, z^{-\tau}}{1 + a_1\, z^{-1} + \ldots + a_\rho\, z^{-\rho}}, \tag{3.10}$$

donde $b_k,\ a_l\ \in \mathbb{R}$ y $a_0 = 1$. En la práctica usualmente se escoge $\tau < \rho$. Los modelos ARMA se definen en el tiempo a partir de la siguiente ecuación en diferencias estocástica [174]:

$$s[n] = -\sum_{l=1}^{\rho} a_l\, s[n-l] + \sum_{k=1}^{\tau} b_k\ e[n-k], \tag{3.11}$$

donde e satisface las mismas características que en la Ec. (3.7).

Sean ζ_k y p_l respectivamente los ceros y los polos de $\Theta_{ARMA}(z)$. Con esta información, podemos reescribir a $\Theta_{ARMA}(z)$ de la siguiente forma:

$$\Theta_{ARMA}(z) = \frac{b_0\ \prod_{k=1}^{\tau}(1 - \zeta_k\, z^{-1})}{\prod_{l=1}^{\rho}(1 - p_l\, z^{-1})} = \frac{b_0\, z^{\rho-\tau}\ \prod_{k=1}^{\tau}(z - \zeta_k)}{\prod_{l=1}^{\rho}(z - p_l)}, \tag{3.12}$$

considerando $\tau < \rho$. Se desprende entonces que $\Theta_{ARMA}(z)$ tiene ρ polos, τ ceros y un cero de orden $\rho - \tau$ en $z = 0$ en el plano z. A su vez, para que $\Theta_{ARMA}(z)$ sea un sistema discreto de *fase mínima*, todos sus polos y sus ceros deben hallarse en el interior del círculo unitario, es decir, $|\zeta_k| < 1\ \ \forall k = \{1, 2, \ldots, \tau\}$ y $|p_l| < 1\ \ \forall l = \{1, 2, \ldots, \rho\}$ [108, 158]. Los modelo ARMA se conocen también como *modelos de polos y ceros* en la literatura.

Los modelos ARMA resultan atractivos porque permiten una representación más versátil, en comparación con los modelos AR. En el estudio del habla en particular, los ceros de $\Theta_{ARMA}(z)$ son muy importantes en la práctica para modelar fonemas consonánticos nasales, fricativos y oclusivos [174]. Sin embargo, la estimación de los parámetros de este modelo no es una tarea sencilla y, en general, conlleva la resolución de un problema de optimización no lineal [108, 158]. En [174] se presentan diferentes estrategias, utilizadas en la actualidad, para la estimación de los parámetros en aplicaciones que involucran señales de voz.

3.4.2. Modelado lineal variante en el tiempo

Los modelos presentados en la sección anterior toman como hipótesis de trabajo que la señal de voz es *aproximadamente* estacionaria en ventanas de corta duración. Si bien esta forma de trabajo ha demostrado ser sumamente útil, la hipótesis de estacionariedad es en general fuerte y no se cumple en la realidad. Para aliviar esta situación, se han propuesto diferentes alternativas para el estudio de la voz a partir de modelos variantes en el tiempo [73]. A su vez, algunos de estos modelos han resultado atractivos para la comunidad científica por su estrecha relación con los métodos en espacio de estados [99, 127].

Presentaremos aquí dos alternativas que surgen de la generalización de los modelos presentados en la sección anterior. La primera de estas alternativas se obtiene de la Ec. (3.7). Se define a un modelo autorregresivo lineal y variante en el tiempo (TAR) de orden ρ a partir de la siguiente ecuación en diferencias [99]:

$$s[n] = -\sum_{l=1}^{\rho} a_l[n]\, s[n-l] + e[n], \tag{3.13}$$

donde $a_l[n]$ con $l = \{1, 2, \ldots, \rho\}$ son en este caso series temporales, e presenta las mismas propiedades que para los modelos AR y $G_0 = 1$. Como podemos apreciar, la principal diferencia es que los coeficientes del modelo pueden cambiar con el tiempo, es decir, son funciones dependientes del tiempo. De forma similar, podemos definir a los modelos autorregresivos y de media móvil variantes en el tiempo (TARMA) de orden (τ, ρ) a partir de la Ec. (3.10).

La capacidad de adaptación de estos modelos permite lograr una representación más precisa, a la vez que su estructura paramétrica captura la información de la dinámica de una señal en un conjunto pequeño de parámetros [73]. Otra ventaja asociada a estos modelos es que, gracias a su adaptación temporal, permiten el estudio de señales en intervalos de tiempos más prolongados [73]. Sin embargo, presentan la importante desventaja que la estimación de la dinámica de los coeficientes del modelo no es una tarea sencilla. Dicha estimación depende a su vez de las diferentes hipótesis sugeridas para explicar la evolución de los coeficientes. Usualmente, se suelen considerar alguna de las siguientes formas [127]:

- No se fija ninguna estructura para los parámetros y se permite que estos cambien libremente.
- Se supone que la dinámica de cada parámetro sigue una regla estocástica, imponiendo alguna restricción de suavidad.
- Se modela la variación de los parámetros como una combinación lineal de una familia de funciones determinísticas.

Para una mayor comprensión respecto a las propiedades principales de los modelos variantes en el tiempo y sus aplicaciones, el lector puede referirse a [69, 127, 131] y sus correspondientes referencias.

3.4.3. Modelos lineales con entradas externas

Todos los modelos presentados hasta aquí coinciden en la fuente de excitación e. En el estudio de señales estocásticas, usualmente se considera una excitación de la forma $e[n] \sim \mathcal{N}(0, \sigma_e^2)$. Ésta es una de las formas consideradas en la fonación y es de utilidad en el caso en que los valores de la señal bajo estudio se mantienen próximos a un nivel de referencia constante o que cambia muy lentamente [26, 108]. Este tipo de comportamiento recibe el nombre de *tendencia*. Es común que esto último no se cumpla para señales no estacionarias, por lo que deben ser estudiadas con modelos capaces de capturar dinámicas más complejas.

Una solución alternativa a la problemática enunciada arriba son los modelos lineales con entrada externa determinística. Nos enfocaremos en esta familia de modelos, ya que ha demostrado ser muy útil para el estudio de sistemas ingenieriles

en los cuales se puede medir, u observar, tanto la señal de salida como la de entrada. Describiremos aquí aquellos modelos que serán importantes para el desarrollo de este documento. Para mayor información respecto a los modelos con entradas externas, el lector puede referirse a [99, 100].

Sea u una señal con dinámica conocida o que puede sensarse directamente. Partiendo de la Ec. (3.7), definimos a un modelo autorregresivo con entrada externa (ARX) de orden ρ a partir de la siguiente ecuación en diferencias lineal:

$$s[n] = -\sum_{l=1}^{\rho} a_l \, s[n-l] + u[n] + e[n], \tag{3.14}$$

donde los a_l se definieron en la Ec. 3.4. De forma similar, definimos a los modelos autorregresivo variantes en el tiempo y con entrada externa (TARX) de orden ρ a partir de la siguiente ecuación en diferencias lineal:

$$s[n] = -\sum_{l=1}^{\rho} a_l[n] \, s[n-l] + u[n] + e[n]. \tag{3.15}$$

donde los $a_l[n]$ se definieron en la Ec. 3.13. En ambos casos, la función u corresponde a la *tendencia* o a la señal de entrada del sistema. Además, e presenta las mismas propiedades que para los modelos AR.

A su vez, trabajando de forma similar a partir de la Ec. (3.10) podemos definir a los modelos autorregresivos y de media móvil de orden (τ, ρ) con entrada externa, los que pueden presentar parámetros constantes (ARMAX) o variantes en el tiempo (TARMAX).

3.5. Fuente glótica

La fuente glótica, también llamada pulso o flujo glótico, es una idealización del flujo de aire que atraviesa la glotis y es modulado por la acción de las cuerdas vocales. Es el principal mecanismo de excitación del tracto vocal en fonemas sonoros y se caracteriza por un comportamiento pulsátil cuasiperiódico (ver Sec. 2.2.2). En el marco de la TFF, este fenómeno pulsátil es usualmente modelado a partir de funciones matemáticas, cuyos parámetros describen características propias del flujo de aire. Estas funciones se utilizan principalmente en fonemas vocales. Sin embargo, pueden aplicarse también en fonemas más complejos, por ejemplo en el caso de los fonemas fricativos o nasales sonoros [151, 163, 176].

Hasta el momento, en la comunidad científica no se ha arribado a un consenso respecto a la dinámica del flujo de aire glótico [5, 46, 174]. Sin embargo, desde hace varias décadas se han propuesto diferentes formas de onda para simular esta señal, desarrolladas gracias a la información extraída del procesamiento de vocales sostenidas reales [42, 57, 128]. Estos modelos han demostrado ser muy útiles para el estudio del esfuerzo vocal, de la variación en la prosodia y de la calidad perceptual en la voz [46, 151].

Afortunadamente, los especialistas han acordado en una serie de características y de propiedades del flujo glótico, las que surgen de la fisiología de la fonación y del procesamiento de la señal de voz. En el campo temporal, se considera que esta señal satisface el siguiente conjunto de propiedades [44]:

- Toma únicamente valores positivos o nulos.
- Es una función suave y aproximadamente periódica (los períodos de los pulsos glóticos difieren levemente entre sí).
- Cada pulso presenta una forma de onda *acampanada*: partiendo del valor nulo, crece hasta llegar a su valor máximo, luego decrece y se anula hacia el final del período.
- Es una señal continua en el tiempo.
- Su derivada debe existir, excepto para los instantes de cierre de la glotis.

A su vez, desde el punto de vista frecuencial el espectro de potencia de esta señal debe comportarse de forma similar a la respuesta en frecuencia de un filtro de segundo orden pasa bajas, con una atenuación de aproximadamente 12 dB por octava para frecuencias superiores a la frecuencia fundamental f_0 [5, 44].

Considerando nuevamente las hipótesis principales de la TFF, presentaremos a continuación una interpretación alternativa de la fuente glótica. De lo expuesto en la Sec. 3.3.1, podemos reescribir la Eq. (3.1) como sigue:

$$S(z) = G_\Delta(z)\, H(z)\, R(z) = R(z)\, G_\Delta(z)\, H(z) = G_\Delta^v(z)\, H(z), \tag{3.16}$$

donde $G_\Delta^v(z) = R(z)\, G_\Delta(z)$. El resultado anterior se obtiene gracias a las hipótesis de linealidad y de la existencia de la transformadas $\mathcal{Z}$ de las señales. Recordemos que la emisión en los labios se modela como un sistema diferenciador. Aplicando la transformada $\mathcal{Z}$ inversa, podemos interpretar a la Eq. (3.16) como si el tracto vocal fuera excitado por una nueva señal, consistente en la derivada con respecto al tiempo del tren de pulsos glóticos. A esta señal se la denomina función, o excitación, glótica v_g. Esta interpretación es muy frecuente en el área del procesamiento del habla, en especial por la notable similitud entre las expresiones (3.9) y (3.16) [42, 128].

Algunos autores han modelado directamente la función glótica, en lugar del flujo glótico. De esta forma, la fonación puede representarse considerando únicamente la función glótica y el filtro del tracto vocal, lo que es una ventaja tanto en el análisis como en la síntesis de señales de voz [5, 46, 57, 174]. Para la síntesis de voz, originalmente se tomaba como función glótica a un tren de impulsos, donde el lapso de tiempo entre dos impulsos sucesivos correspondía al período fundamental T_0. Posteriormente, se demostró que considerar pulsos con morfologías más complejas, en lugar de simples impulsos, permitía mejorar los atributos perceptuales de las voces sintetizadas [56, 101, 151].

En general, v_g preserva las mismas propiedades temporales que el flujo glótico, con la salvedad que su morfología en cada período corresponde a la tasa de cambio temporal del flujo glótico [44, 151]. Así, en cada pulso v_g es positiva en los puntos donde el flujo de aire crece, se anula cuando el flujo llega a su valor máximo, es negativa cuando el flujo decrece y se anula nuevamente cuando el flujo es cero [163]. Además, v_g es derivable salvo en los instantes de cierre de la glotis, donde toma su valor mínimo y su morfología presenta una espiga negativa característica [5, 46]. Por otro lado, el espectro de potencia de v_g se comporta de forma similar a la respuesta en frecuencia de un filtro de segundo orden pasa banda, cuyo máximo absoluto coincide con f_0. Para frecuencias menores (mayores) a f_0 el espectro de potencia presenta una amplificación (atenuación) de 6 (-6) dB por octava [44].

En la bibliografía existe una gran variedad de alternativas para representar el mecanismo de excitación en la glotis, ya sea considerando el flujo glótico U_g o la función glótica v_g. Para una mayor comprensión respectos a estas señales, así como sus diferentes características y propiedades, el lector puede referirse a [42, 44, 46, 57, 128] y su correspondientes referencias. A continuación, centraremos nuestra atención a una representación paramétrica de la función glótica v_g que será de suma importancia, más adelante, en el desarrollo de esta tesis.

3.5.1. Función glótica de Liljencrants y de Fant

El modelo propuesto por Liljencrants y Fant (LF) es, sin duda alguna, una de las representaciones paramétricas de la función glótica v_g más populares. Se propuso originalmente en [57] y, desde entonces, ha sido ampliamente utilizado tanto para la síntesis como el análisis de señales de voz. Esto se debe a que LF se ajusta satisfactoriamente a las formas de onda de la excitación glótica que se observan comúnmente en la práctica, a la vez que satisface todas las propiedades arriba enunciadas [44, 151].

De acuerdo con este modelo, la morfología de cada pulso queda completamente determinada mediante dos conjuntos de parámetros [57]:

- Parámetros de síntesis: $\{\mathrm{E}_0,\, \alpha,\, \omega_g,\, \epsilon\}$,
- Parámetros temporales $\{E_e, N_p,\, N_e,\, N_a,\, N_0\}$,

con $\mathrm{E}_0,\, E_e,\, \alpha,\, \omega_g,\, \epsilon \in \mathbb{R}$ y $N_p, N_e, N_a, N_0 \in \mathbb{N}$. Mostraremos más adelante que estos dos conjuntos de parámetros dependen entre sí, de acuerdo a un conjunto de restricciones [57, 128, 135]. En particular, N_0 es el período fundamental y, luego, la frecuencia fundamental resulta $f_0 = f_s/N_0$, donde f_s es la frecuencia de muestreo.

El modelo LF se define analíticamente, en su forma discreta, de la siguiente manera [57]:

$$v_g^{\mathrm{LF}}[n] = \begin{cases} E_0\, e^{\alpha\, n}\ \mathrm{sen}\,(\omega_g\, n), & \text{si } 0 \leq n \leq N_e, \\ \frac{-E_e}{\epsilon\, N_a}\left(e^{-\epsilon\,(n-N_e)} - e^{-\epsilon\,(N_c-N_e)}\right), & \text{si } N_e < n \leq N_c, \\ 0, & \text{si } N_c < n < N_0, \end{cases} \tag{3.17}$$

sujeto al siguiente conjunto de restricciones:

$$\begin{cases} \sum_{n=0}^{N_0-1} v_g^{\mathrm{LF}}[n] = 0, \\ \omega_g = \frac{\pi}{N_p}, \\ \epsilon\, N_a = 1 - e^{-\epsilon\,(N_c-N_e)}, \\ E_e = -E_0\, e^{\alpha\, N_e}\ \mathrm{sen}\,(\omega_g\, N_e). \end{cases} \tag{3.18}$$

En la expresiones anteriores, se tiene en cuenta además que $\alpha > 1$, $0 < \epsilon < 1$ y $\omega_g, E_0, E_e > 0$.

La Fig. 3.6 nos ayudará a interpretar el modelo LF y a comprender la importancia de sus parámetros. En el renglón superior, presentamos un período de la función glótica v_g^{LF}, mientras que en el renglón inferior podemos observar el correspondiente flujo glótico U_g^{LF}. Rápidamente, podemos afirmar que ambas señales cumplen con las propiedades temporales enunciadas anteriormente [44]. Podemos observar también que los parámetros temporales indican eventos característicos de la dinámica glótica (recordar esquema de Fig. 2.5).

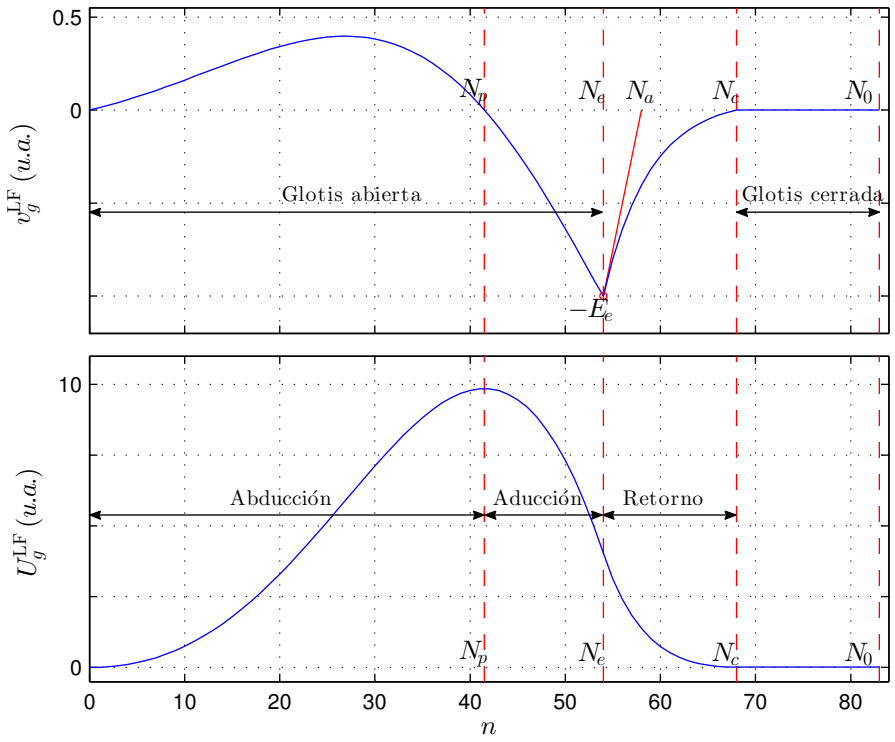


Figura 3.6: Modelo de la función glótica de Liljencrants y de Fant (LF). Arriba: función glótica v_g^{LF}. Abajo: flujo de aire glótico U_g^{LF}. Se representan también los parámetros temporales del modelo LF y las diferentes etapas de la dinámica glótica.

El intervalo de tiempo desde el inicio del ciclo hasta N_p corresponde con la fase de abducción (separación) de las cuerdas vocales, siendo N_p el instante de máxima abducción donde U_g^{LF} toma su máximo valor. N_p puede estimarse a partir del cruce por cero con pendiente negativa de v_g^{LF}, tarea que en la práctica no es sencilla. Seguidamente, observamos la fase de aducción (aproximación) de las cuerdas vocales, que concluye en el instante N_e con la repentina colisión entre las porciones caudales de las cuerdas vocales. Por ello, a N_e se lo denomina *instante de cierre glótico*[5] (GCI). Es en este punto donde v_g^{LF} alcanza su valor mínimo en el período glótico, donde E_e representa la máxima excitación efectiva en el tracto vocal. N_e y E_e pueden estimarse fácilmente calculando el mínimo absoluto de v_g^{LF} en el período [57, 128]. A continuación, se produce el cierre gradual de la glotis y la disminución del flujo de aire hasta anularse en la fase de retorno. Esta fase concluye en el instante N_c, indicando el cierre completo de la glotis. La glotis permanecerá cerrada hasta el comienzo de un nuevo período, determinado por el inicio de una nueva fase de abducción. Este evento se denomina *instante de apertura glótica*[6] (GOI).

De acuerdo a la definición (3.17), el comportamiento de v_g^{LF} se caracteriza por una primera etapa senoidal cuya amplitud crece exponencialmente. A esto le sigue una fase donde la señal toma sólo valores negativos y su magnitud decrece exponencialmente, hasta que finalmente se anula completamente. Con respecto a otras representaciones, el modelo LF presenta una fase de retorno exponencial, la cual permite simular un cierre paulatino o incompleto de la glotis. Además, se ha demos-

[5] En inglés, "glottal closure instant".
[6] En inglés, "glottal open instant".

trado que la fase de retorno influye notablemente en el comportamiento espectral y en las características perceptuales de las señales de voz [46, 57]. N_a es el tiempo de retorno efectivo y se define como la intersección con el eje de las abscisas de la recta tangente a v_g^{LF} en el punto $(N_e, -E_e)$.

De lo expuesto hasta aquí, se desprende que el modelo LF se centra en la información de la dinámica glótica en el dominio del tiempo. Luego, los parámetros de síntesis pueden calcularse de los parámetros temporales aplicando las restricciones (3.18). En la práctica, es usual considerar $N_c = N_0$ indicando que la etapa de glotis cerrada incluye a la fase de retorno [44, 57]. Así, ya no es necesario determinar N_c, cuya estimación no es una tarea sencilla en la práctica. Esto último permite simplificar el modelo LF, volviéndolo más atractivo en algunas aplicaciones.

3.6. Filtrado inverso de la voz

En esta sección introduciremos brevemente el concepto de filtrado inverso de la voz, centrándonos en su aplicación al análisis de señales de voz reales. Desde hace un tiempo, este concepto ha resultado muy atractivo para la comunidad científica, a tal punto que actualmente se ha convertido en una disciplina independiente dentro del área del análisis de la voz [5, 46].

El proceso de filtrado inverso de la voz se desprende directamente de la TFF, a la vez que se relaciona estrechamente con la estimación y el modelado de la fuente glótica [55, 57]. Sea s una señal de voz correspondiente a un fonema vocal. De acuerdo con lo expuesto en la Sec. 3.3, la información de s en el plano z queda completamente determinada por la transformada $\mathcal{Z}$ del flujo glótico $G_\Delta(z)$ y las funciones de transferencia de los filtros de tracto vocal $H(z)$ y de emisión $R(z)$. Suponiendo que ambas funciones de transferencia son conocidas, podemos reescribir la expresión (3.1) de la siguiente forma:

$$G_\Delta(z) = \frac{S(z)}{H(z)\,R(z)}. \tag{3.19}$$

Considerando la transformada $\mathcal{Z}$ inversa y sus propiedades, podemos interpretar la expresión (3.19) en el dominio del tiempo. Rápidamente, podemos estimar el flujo glótico U_g aplicando un método en dos etapas [42, 128, 130]. En primer lugar, se filtra la señal de voz s con el filtro inverso del tracto vocal, caracterizado por la función de transferencia $1/H(z)$. Esto permite eliminar la modulación en la información frecuencial aportada por el tracto vocal. Seguidamente, la señal resultante es integrada, eliminando así la acción derivativa de la emisión en los labios.

Considerando a su vez a la función glótica v_g, es posible simplificar el proceso de filtrado inverso. Recordando la relación $G_\Delta^v(z) = R(z)\,G_\Delta(z)$ presentada en la Sec. 3.5, el filtrado inverso queda determinado por la siguiente expresión:

$$G_\Delta^v(z) = \frac{S(z)}{H(z)}. \tag{3.20}$$

Empleando los mismos criterios que en el caso anterior, esta última expresión nos permite estimar directamente la función glótica v_g aplicando a la señal de voz s el filtro con función de transferencia $1/H(z)$. De esto último, resultan evidentes las razones por las que al método descripto se lo denomina filtrado inverso de la señal de voz.

De lo expuesto hasta aquí, se desprende que la calidad en la estimación de U_g, o de v_g, depende íntimamente de conocer el filtro del tracto vocal $H(z)$. Sin embargo, en la práctica esta información no se conoce *a priori* y es necesario entonces calcularla de forma conveniente. Esto puede estudiarse en el marco de un *problema inverso*, donde se quiere calcular a partir de la señal de voz tanto el filtro del tracto vocal como la información de la fuente glótica, aplicando diferentes modelos matemáticos y métodos de optimización [46, 135]. Desde ya, los resultados dependerán de los modelos considerados y de los métodos aplicados para el calculo de los parámetros correspondientes. Para una mayor comprensión sobre el proceso de filtrado inverso, su aplicación a casos reales y las dificultades que conlleva, el lector puede referirse a [5, 42, 46, 55, 174].

3.6.1. Filtrado inverso iterativo y adaptativo

Centraremos ahora nuestra atención en uno de los método de filtrado inverso de la voz más difundidos y estudiados en la literatura. Este método se denomina *filtrado inverso iterativo y adaptativo*[7] (IAIF) y, como su nombre lo indica, permite estimar la función glótica, el flujo glótico y el filtro del tracto vocal a través de un proceso adaptativo [4]. Se basa en cuantificar iterativamente la contribución de cada uno de los elementos de la TFF por separado, los que combinados dan lugar al espectro de potencia de la señal de voz, y en emplear esta información para obtener mejores estimaciones de los componentes restantes [2].

A continuación describiremos brevemente las etapas principales de IAIF. Partiendo de una ventana de señal de voz, se obtiene una primera estimación de la información frecuencial de la función glótica, $H_{g1}(z)$, usando un modelo AR de primer orden. Luego, se cancela el efecto de la función glótica en la señal de voz aplicando el filtro con función de transferencia $1/H_{g1}(z)$. A partir de la señal resultante, se obtiene una representación preliminar del tracto vocal $H_{tv1}(z)$ considerando un filtro AR de orden p^{IAIF}. En una segunda iteración, se refinan estas estimaciones repitiendo los dos procesos anteriores. Así, la información de la función glótica $H_{g2}(z)$ se representa con un modelo AR de orden g^{IAIF} con $1 < g^{IAIF} < p^{IAIF}$. Finalmente, se calcula un nuevo modelo del filtro del tracto vocal $H_{tv2}(z)$ considerando nuevamente un modelo AR de orden p^{IAIF}. Es importante resaltar que este procedimiento puede realizarse por ventanas de señal de voz o de forma sincronizada con la dinámica glótica [174].

La función glótica v_g puede estimarse aplicando el filtro inverso correspondiente a la última estimación del tracto vocal, es decir, el filtro con función de transferencia $1/H_{tv2}(z)$. Además, integrando esta señal puede estimarse el flujo glótico U_g. En la actualidad, existen versiones de IAIF que, para la estimación de las diferentes funciones de transferencia, emplean el método clásico de LPC o su versión mejorada propuesta en [52]. Para mayor información respecto a este método y sus diferentes versiones, el lector puede referirse a [2, 4, 7, 174].

Para el desarrollo de esta tesis de doctorado, hicimos uso de la implementación de IAIF que forma parte del paquete de funciones *TKK Aparat* [2], disponible libremente en Internet en `http://sourceforge.net/projects/aparat`. Como veremos más adelante, mediante esta herramienta se obtuvieron estimaciones de la función

[7] En inglés, "iterative adaptive inverse filtering".

glótica y del tracto vocal, que sirvieron para comparar y contrastar los resultados arrojados en diferentes simulaciones.

3.7. Comentarios finales

En este capítulo abordamos el estudio y modelado de la fonación, centrándonos fundamentalmente en la producción de la señal de voz. Excluimos de esta exposición el modelado de otras señales relacionadas a la fonación, como por ejemplo el EGG y las VPC (ver Sec. 2.4).

En la primera parte, discutimos las ideas más importantes de las dos principales estrategias, disponibles actualmente en la bibliografía, para estudiar la fonación: la teoría *mioelástica*, *aerodinámica* y *acústica*, y la teoría *fuente* y *filtro*. A lo largo de la exposición, prestamos especial atención a las hipótesis y a los conceptos fundamentales que sustentan cada teoría. Por ello, podemos aseverar que, si bien son diferentes, estas dos teorías se complementan entre sí y garantizan dos interpretaciones alternativas de un mismo fenómeno.

La teoría *fuente* y *filtro* mereció una atención especial. La razón de esto, es que esta construcción es sumamente importante en aplicaciones que involucran a la señal de voz, en particular en las áreas: modelado, comunicación, almacenamiento y procesamiento de esta señal. Además, varios de los aportes presentados en esta tesis se basan en ella. De acuerdo a esta teoría, analizamos diferentes interpretaciones de la fonación, tanto en el dominio del tiempo como en el dominio de la frecuencia.

Seguidamente, analizamos los elementos principales involucrados en la teoría *fuente* y *filtro*. Presentamos algunas de las principales estrategias para representar el filtro del tracto vocal, aplicadas a lo largo del tiempo. Por otro lado, profundizamos en la descripción e interpretación de la fuente de excitación, indicando su relación con el flujo de aire glótico. También, introdujimos el concepto de función glótica, así como el modelo de Liljencrants y de Fant que permite describirla matemáticamente.

Por último, presentamos las ideas fundamentales del proceso de filtrado inverso de la voz, las cuales surgen de la teoría *fuente* y *filtro*. Estas ideas son sumamente importantes para la estimación conjunta de la fuente glótica y del tracto vocal a partir de una señal de voz real. A su vez, describimos una de las implementaciones del filtrado inverso, que es muy difundida y aplicada en la actualidad. Haremos uso de esta implementación más adelante en esta tesis.

Capítulo 4

Análisis y modelado de las perturbaciones en la voz

4.1. Introducción

En los capítulos anteriores, presentamos los fonemas sonoros y explicamos que éstos se caracterizan por un comportamiento cuasiperiódico. Es decir, su forma de onda regular se repite, con pequeñas diferencias, a intervalos de tiempos aproximadamente uniformes. Así, los sonidos de la voz presentan pequeñas perturbaciones, incluso las voces consideradas normales y sanas [162]. En general, este fenómeno se aprecia con mayor facilidad en emisiones vocales sostenidas [41]. Definimos, también, la frecuencia fundamental f_0 que caracteriza la dinámica de las cuerdas vocales durante la fonación. Es importante destacar que, en la práctica, f_0 corresponde a una medida promedio para una ventana de señal de voz [42].

En la medicina, las perturbaciones forman parte del conjunto de características de interés que son estudiadas en el análisis acústico de la voz con fines diagnósticos. Este análisis consiste en la obtención, a partir de una señal de voz, de un conjunto de parámetros que permita caracterizar el estado de las cuerdas vocales de forma objetiva, rápida y no invasiva [16, 41, 162]. Los rasgos que se estudian son las perturbaciones en los períodos y en las amplitudes, la relación entre el contenido armónico y turbulento, la composición biomecánica de las cuerdas vocales y la dinámica glótica, entre otros [16, 46, 70]. Actualmente, existe una gran variedad de programas de computadora (software) que permiten llevar a cabo el análisis acústico de una señal de voz. A modo de ejemplo, podemos nombrar a *Multi-Dimensional Voice Program* (MDVP) de la empresa *Kay Elemetrics* y a *PRAAT* desarrollado por Boersma y Weenink [23]. Lo atractivo de este enfoque es que ha demostrado ser adecuado para monitorear los cambios en la calidad de la voz a lo largo del tiempo [41].

Por otro lado, las perturbaciones en la voz han captado el interés de la comunidad científica, ya que cumplen un rol importante en el grado de *naturalidad* con que se percibe la voz humana. Se ha probado que un oyente reconoce fácilmente una señal de voz artificial perfectamente periódica debido a que, inconscientemente, él está acostumbrado a percibir ciertas irregularidades en los sonidos de la voz [24, 128]. Por otro lado, se han desarrollado sistemas inteligentes para la detección de patologías vocales, diseñados a partir de diferentes parámetros acústicos relacionados a las perturbaciones en la voz [10, 109, 113, 177].

En la primera parte de este capítulo, estudiaremos las series temporales de pe-

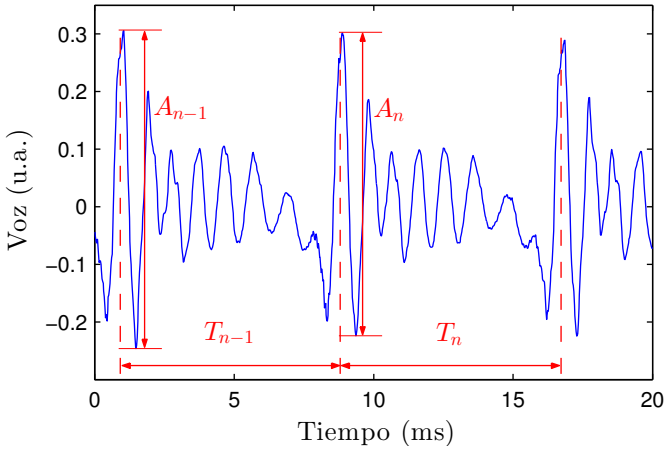


Figura 4.1: Estimación de las series de períodos (SP) y de amplitudes (SA). Se presentan 20 ms de una vocal /a/ sostenida correspondiente a un sujeto masculino sano. Se indican también los instantes de inicio de cada ciclo (líneas discontinuas verticales). La SP y la SA se construyen calculando el período, cota horizontal, y la amplitud, cota vertical, para cada ciclo de la señal de voz.

ríodos y de amplitudes asociadas a la señal de voz. A partir de éstas, introduciremos los conceptos de fluctuación y de perturbación. Luego, presentaremos los parámetros acústicos empleados en la práctica médica para cuantificar las perturbaciones y analizaremos cómo influyen las fluctuaciones en ellos. Dedicaremos el resto de este capítulo a presentar un método de síntesis de vocales sostenidas con perturbaciones controladas en los parámetros acústicos, el cual desarrollamos durante la etapa inicial de esta tesis de doctorado. Describiremos las diferentes partes que lo componen y detallaremos su implementación. Por último, analizaremos las diferentes simulaciones realizadas para evaluar el desempeño del método propuesto.

4.2. Series de períodos y de amplitudes

Como dijimos anteriormente, el análisis acústico de la voz surge como un método indirecto y no invasivo para el estudio de las cuerdas vocales [16, 82, 162]. Su propósito principal consiste en evaluar la dinámica glótica y el estado de las cuerdas vocales usando solamente la información disponible en la señal de voz. En este análisis se emplean, preferentemente, emisiones vocales sostenidas [41]. La regularidad de la dinámica glótica es uno de los diferentes rasgos estudiados. Para ello, es necesario construir dos señales asociadas a la fonación denominadas serie de períodos (SP) y serie de amplitudes (SA). Dedicaremos esta sección a describir estas señales.

Comenzaremos explicando la metodología necesaria para la estimación de la SP y la SA. Como ayuda didáctica para la exposición, utilizaremos el ejemplo mostrado en la Fig. 4.1, donde se presentan 20 ms de una señal de voz correspondiente a una vocal /a/ sostenida de un sujeto masculino sano. Podemos apreciar la regularidad propia de esta clase de fonema. Para la construcción de las señales, es necesario en primer lugar identificar cada ciclo de la señal de voz. La forma más sencilla de llevar a cabo esta tarea consiste en la detección de eventos cíclicos, como por ejemplo los máximos y mínimos más prominentes o los cruces por cero [16, 42, 128]. Sin embargo,

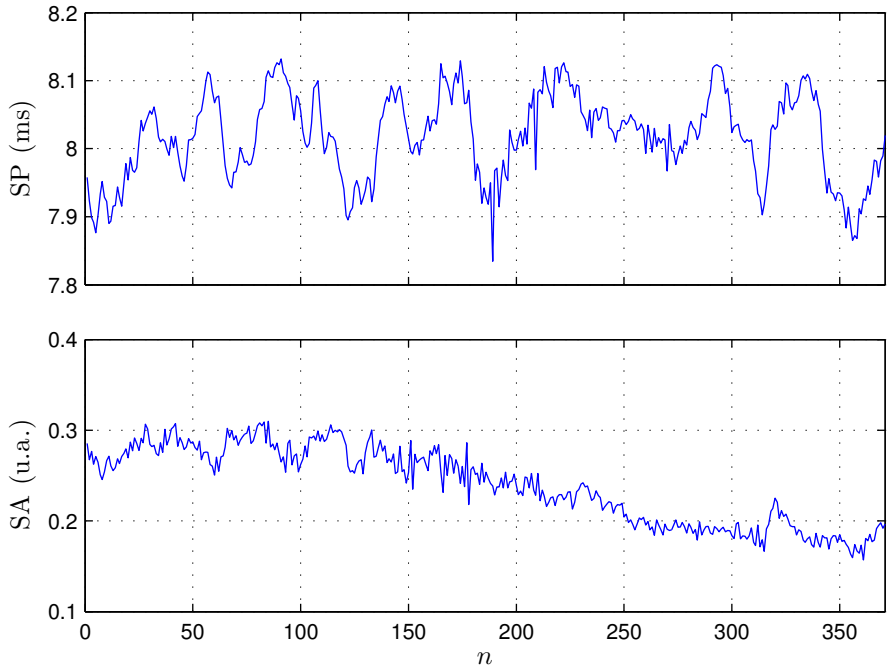


Figura 4.2: Series de períodos (SP) y de amplitudes (SA) extraídas de una vocal /a/ sostenida con una duración de 3 s. Corresponden a un sujeto masculino sano.

la presencia de ruido en la señal de voz dificulta considerablemente esta detección. Por ello, se recomienda emplear varios eventos cíclicos de forma simultánea. En esto consiste el método de *coincidencia en la forma de onda*[1], cuyo objetivo es identificar un conjunto de eventos en la señal de voz para los cuales la forma de onda para dos ciclos sucesivos presentan la máxima similitud [167]. A partir de esta información, se calculan luego los instantes de inicio de cada ciclo. Usualmente, este método se implementa aplicando una medida de autocorrelación [21, 22]. En la Fig. 4.1, las líneas discontinuas verticales indican los instantes de inicio de cada ciclo.

A continuación, se calcula la duración de cada ciclo de la señal de voz como la diferencia entre los inicios de dos ciclos consecutivos. Se construye así la SP:

$$SP = \left\{T_1,\ T_2,\ T_3,\ \ldots\ ,\ T_N\right\} \tag{4.1}$$

donde T_n es el período del n-ésimo ciclo y N es la cantidad de ciclos en la señal de voz. Por último, se mide la amplitud *pico a pico* en todos los ciclos y con esta información se construye la SA:

$$SA = \left\{A_1,\ A_2,\ A_3,\ \ldots\ ,\ A_N\right\} \tag{4.2}$$

donde A_n es la amplitud del n-ésimo ciclo.

En la Fig. 4.2, mostramos la SP y la SA estimadas de una emisión vocal /a/ sostenida de 3 s de duración. Presentamos 20 ms de la parte central de esta señal en la Fig. 4.1. Podemos observar que, como era de esperarse, ni el período ni la amplitud se mantuvieron constantes a lo largo del tiempo, aun cuando se estimaron a partir de una emisión estable. En ambas señales pueden observarse a su vez dos

[1]En inglés, "waveform-matching method".

fenómenos bien diferenciados. Por un lado, existe una dinámica considerablemente lenta y suave, a la cual denominaremos *fluctuación*. A ésta se le superpone una dinámica aleatoria de magnitud pequeña y escala local (algunos ciclos adyacentes), a la cual en adelante llamaremos *perturbaciones*. Podemos concluir entonces que coexisten una dinámica de largo alcance y otra de alcance local [16].

Es importante destacar que la calidad de la SP y la SA dependerá de la precisión en la detección de los ciclos en la señal de voz [19, 167]. Sin embargo, la construcción de estas señales posee un limitación importante. El método explicado no puede aplicarse a señales extremadamente aperiódicas, al menos no en forma confiable [41, 167]. Por otro lado, la SP puede construirse para otras señales biomédicas, como por ejemplo el EGG y las VPC (ver Sec. 2.4). Estas señales resultan atractivas porque, en comparación con la voz, poseen una forma de onda más simple y se obtienen a partir de estrategias de sensado robustas [14, 15]. Desafortunadamente, estas señales no pueden utilizarse para la construcción de la SA, ya que en esos casos la dinámica de la amplitud es diferente a la observada en la señal de voz [16].

4.3. Fluctuaciones y perturbaciones

Es común que existan fluctuaciones y perturbaciones en las señales de voz, lo que puede apreciarse fácilmente en el ejemplo de la Fig. 4.2. Las fluctuaciones se caracterizan por una dinámica suave apreciable a una escala global. Pueden desarrollarse involuntariamente o de forma intencional [162, 163]. Esto último sucede por ejemplo al enfatizar una oración o en el canto. Asimismo, pueden ser el producto de algún trastorno o patología, por ejemplo la tartamudez o las diferentes disfonías [16]. En cambio, las perturbaciones presentan una naturaleza aleatoria y un alcance local [142]. Éstas han demostrado ser útiles para caracterizar la calidad de la voz y, a su vez, para describir el estado de las cuerdas vocales [41, 82, 118]. Sin embargo, desde hace un tiempo existe un fuerte debate en la comunidad científica respecto a si esto es verosímil, y bajo qué condiciones la información obtenida de las perturbaciones es de utilidad [16, 28, 39].

Las fluctuaciones y las perturbaciones son el resultado de los procesos y las transformaciones involucrados en la fonación. Entre sus principales causas, se pueden listar las siguientes [16, 162, 163]:

- *Neurológicas:* La contracción de los músculos vocales es el resultado de la excitación combinada, de forma temporal y espacial, de las motoneuronas que los componen. Sin embargo, esta sincronía no es perfecta. Esto ocasiona pequeñas desviaciones aleatorias en la contracción muscular, que alteran la periodicidad de la dinámica glótica, la abducción de las cuerdas vocales y la intensidad vocal. A su vez, estas desviaciones pueden aumentar considerablemente debido a trastornos neurológicos, como el *mal de Parkinson*.

- *Biomecánicas:* Las imperfecciones en la forma y las propiedades biomecánicas de las cuerdas vocales ocasionan pequeñas alteraciones en la dinámica glótica. Contribuye a esto, además, la falta de simetría en las cuerdas vocales. En casos severos, estas asimetrías dificultan considerablemente la dinámica glótica. Asimismo, las variaciones en la presión pulmonar ocasionan alteraciones adicionales en los períodos y las amplitudes de la voz. Por último, los despla-

zamientos de los articuladores podrían acoplarse mecánicamente modificando el movimiento de las cuerdas vocales.

- *Aerodinámicas:* El flujo de aire glótico puede desarrollar un comportamiento inestable o turbulento en diferentes regiones. Asimismo, su trayectoria es susceptible a modificaciones en la configuración de la glotis y del tracto vocal. Así, este comportamiento aerodinámico errático actúa como una fuente adicional de alteraciones en la señal de voz.

- *Acústicas:* Las variaciones en la distribución de la presión a lo largo del tracto vocal influyen en la dinámica glótica, ya que repercuten en las presiones actuantes en las cuerdas vocales. Esto último se acentúa ante modificaciones rápidas del tracto vocal. Como consecuencia, se alteran de forma diferente los sucesivos ciclos glóticos.

- *Factores de estilo y culturales:* En habla continua se recurre frecuentemente a variaciones voluntarias en el período fundamental o en la intensidad de fonación, por ejemplo al darle énfasis a una oración o para diferenciar entre una pregunta y una respuesta. Por otro lado, existen mutaciones regionales o locales de un mismo idioma, con una marcada impronta social y cultural. Entre otros aspectos, suelen alterarse la cadencia, la entonación, el ritmo y la pronunciación en el habla. Asimismo, en representaciones artísticas es notable el uso de diversas perturbaciones y fluctuaciones para embellecer o transformar la voz de una persona. Ejemplo de esto es el uso de *crescendo*, *diminuendo* y *vibrato* en el canto.

De las causas listadas, se desprende que el estudio y la caracterización de las fluctuaciones en los períodos y las amplitudes no es una tarea sencilla. A su vez, se ha mostrado que es factible el estudio de las diferentes perturbaciones, como consecuencia de su comportamiento aleatorio y su alcance local [16, 82, 125, 162]. Es importante resaltar, sin embargo, que en la actualidad este análisis sólo puede realizarse para señales de voz con una marcada *periodicidad* y que no presenten perturbaciones muy severas [39, 167]. A continuación, introduciremos los parámetros acústicos más importantes, aplicados en la práctica clínica para cuantificar las perturbaciones en los períodos y en las amplitudes.

4.3.1. Perturbaciones en los períodos o *Jitter*

Reciben el nombre de *Jitter* las perturbaciones de corto alcance que presentan los períodos de ciclos sucesivos en la señal de voz [125, 163]. Al igual que ocurre con el término perturbación, la definición de *Jitter* es muy amplia y puede dar lugar a múltiples interpretaciones. Conceptualmente, describe las variaciones ciclo a ciclo de naturaleza aleatoria en los períodos [16]. Al respecto, Schoentgen establece que el *Jitter* es un fenómeno descriptible mediante unos pocos estimadores estadísticos, presente incluso en voces sanas, a partir del cual pueden reconocerse otras formas de aperiodicidades [138]. Este término se emplea también para caracterizar las perturbaciones en la frecuencia instantánea f_i, donde $f_i = 1/T_i$ [19, 167]. Sin embargo, en este documento nos regiremos con la definición introducida en primer lugar.

La exactitud con la que es posible evaluar el *Jitter* depende de la precisión con que se segmentaron los ciclos glóticos. Esta última depende, a su vez, de la frecuencia

de muestreo f_s y del método empleado para estimar el instante de inicio de cada ciclo. El máximo error para la estimación del *Jitter* ($\%\, error_{jitter}$) es [16]:

$$\%\, error_{jitter} = 50\, \frac{f_0}{f_s}. \tag{4.3}$$

Considerando una $f_s = 50$ kHz, para una voz con $f_0 = 100$ Hz, valor típico para hombres adultos, el error máximo es de 0,1 %. Para el mismo valor de f_s y suponiendo ahora una $f_0 = 300$ Hz, valor normal en niños, el error máximo aumenta a 0,3 %. Estos errores son considerablemente altos, en comparación con los niveles de *Jitter* para sujetos con voces normales. Para atenuar este inconveniente puede aumentarse f_s, siempre y cuando el equipamiento empleado lo permita [16, 149]. Otra alternativa consiste en el uso de estrategias de interpolación adecuadas, que permitan mejorar la detección de los ciclos glóticos [167].

En la literatura científica se han propuesto diferentes parámetros acústicos para cuantificar el nivel de *Jitter* presente en una señal de voz. Representan valores promedio de las perturbaciones y, en general, se definen a partir de la función perturbación del período $\Delta T_n = T_n - T_{n-1}$, con $n = 2, 3, \ldots, N$. A su vez, estas medidas de *Jitter* se clasifican en absolutas o relativas a $\hat{T}_0$, donde $\hat{T}_0$ es el período promedio de la SP:

$$\hat{T}_0 = \frac{1}{N} \sum_{i=1}^{N} T_i. \tag{4.4}$$

Medidas de *Jitter* absolutas

Este conjunto de medidas considera la magnitud o la tasa de cambio de signo de la función perturbación [125]. Pertenecen a esta categoría los siguiente parámetros acústicos [16, 59, 157]:

- *Jitter absoluto* (*Jitta*):

$$Jitta = \frac{1}{N-1} \sum_{i=2}^{N} |T_i - T_{i-1}|. \tag{4.5}$$

- *Factor de perturbación*: porcentaje de elementos de la función perturbación del período que cumplen la condición $|\Delta T_i| > 0{,}5$ ms.

- *Factor de perturbación direccional*: porcentaje de cambios de signo en los elementos sucesivos de la función perturbación del período, sin importar la magnitud.

Medidas de *Jitter* relativas a $\hat{T}_0$

El rango de las perturbaciones en los períodos varía con T_0 (o con f_0 respectivamente). Se ha demostrado que grandes diferencias en los períodos de ciclos sucesivos están asociadas a períodos fundamentales mayores [16, 118]. Esto ha dado lugar a diferentes medidas diseñadas para reducir esta dependencia con T_0. A continuación, presentamos las más relevantes [16, 59, 125, 157]:

- *Índice de variabilidad en el período* (PVI): estima la dispersión respecto al valor $\hat{T}_0$, por lo que en rigor no es una medida de *Jitter*. Se calcula de la siguiente forma:

$$PVI = 1000\,\frac{\frac{1}{N}\sum_{i=1}^{N}\left(T_i - \hat{T}_0\right)^2}{\hat{T}_0^2}. \tag{4.6}$$

- *Razón de Jitter porcentual* ($Jitter_{\%}$): es la medida más difundida en la actualidad para cuantificar el nivel de *Jitter* [41, 59]. Se la define como sigue:

$$Jitter_{\%} = 100\,\frac{Jitta}{\hat{T}_0} = 100\,\frac{\frac{1}{N-1}\sum_{i=2}^{N}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i}. \tag{4.7}$$

  La medida análoga basada en las estimaciones de f_n recibe el nombre de *factor de Jitter* en la bibliografía.

- *Perturbación media relativa* (RAP):

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{T_{i-1}+T_i+T_{i+1}}{3} - T_i\right|}{\hat{T}_0}. \tag{4.8}$$

- *Coeficiente de perturbación del período* ($PPQ5$):

$$PPQ5 = \frac{\frac{1}{N-4}\sum_{i=3}^{N-2}\left|\frac{T_{i-2}+T_{i-1}+T_i+T_{i+1}+T_{i+2}}{5} - T_i\right|}{\hat{T}_0}. \tag{4.9}$$

4.3.2. Perturbaciones en las amplitudes o *Shimmer*

Se denomina *Shimmer* a las perturbaciones de naturaleza aleatoria que se desarrollan en las amplitudes de ciclos sucesivos en una señal de voz [41, 163]. Se ha argumentado ampliamente respecto a la importancia del *Shimmer* en la valoración de la ronquera de una voz y en la detección de las voces patológicas [16, 125]. Sin embargo, este fenómeno no ha recibido la misma atención que el *Jitter*, por lo que es considerablemente menor la cantidad de información disponible.

La estimación de las perturbaciones en las amplitudes depende de la calidad con que se calculó la amplitud pico a pico de cada ciclo glótico. En este caso, el factor limitante más importante es la resolución en bits del conversor analógico digital (A/D). Idealmente, el error relativo máximo queda determinado por [16]:

$$\%\,error_{shimmer} = \frac{50}{2^{N_{bits}}}\,\frac{A_{A/D}}{A_{voz}}, \tag{4.10}$$

donde A_{voz} es la amplitud promedio de la señal de voz y los parámetros N_{bits} y $A_{A/D}$ son la cantidad de bits de resolución y el rango de amplitud total del conversor A/D, respectivamente. De la expresión anterior se desprende que para mejorar la precisión en el cálculo de *Shimmer* es conveniente trabajar con un conversor A/D con la mayor resolución posible (10 bits de resolución es el mínimo sugerido) y emplear adecuadamente el rango del conversor [16].

A lo largo del tiempo, se han desarrollado diferentes parámetros para cuantificar el *Shimmer*. Se basan en la función perturbación de la amplitud $\Delta A_n = A_n - A_{n-1}$,

con $n = 2, 3, \ldots, N$. La mayoría de estos parámetros presentan una definición similar a las medidas de *Jitter*, tanto absolutas como relativas. Como valor de referencia para las medidas relativas se considera la amplitud promedio $\hat{A}_0$ de la SA, calculada de la siguiente forma:

$$\hat{A}_0 = \frac{1}{N} \sum_{i=1}^{N} A_i. \tag{4.11}$$

Medidas de *Shimmer* absolutas

Las medidas de *Shimmer* absolutas más relevantes son [16, 157]:

- *Shimmer absoluto* ($Shimma$):

$$Shimma = \frac{1}{N-1} \sum_{i=2}^{N} |A_i - A_{i-1}|. \tag{4.12}$$

- *Shimmer en decibeles* ($Shimmer_{dB}$):

$$Shimmer_{dB} = \frac{1}{N-1} \sum_{i=2}^{N} \left| 20 \log_{10} \left(\frac{A_i}{A_{i-1}} \right) \right|. \tag{4.13}$$

- *Factor de perturbación direccional*: porcentaje de cambios de signo en los elementos sucesivos de la función perturbación de la amplitud, sin importar la magnitud.

Medidas de *Shimmer* relativas a $\hat{A}_0$

Las medidas de *Shimmer* relativas a $\hat{A}_0$ más importantes son [16, 59, 125, 157]:

- *Índice de variabilidad de la amplitud* (AVI): al igual que el PVI, esta medida calcula la dispersión respecto al valor $\hat{A}_0$, por lo que no es una estimación del *Shimmer* en sí misma. Se calcula de la siguiente forma:

$$AVI = 20 \log_{10} \left(1000 \, \frac{\frac{1}{N} \sum_{i=1}^{N} \left(A_i - \hat{A}_0 \right)^2}{\hat{A}_0^2} \right). \tag{4.14}$$

- *Razón de Shimmer porcentual* ($Shimmer_{\%}$): junto con $Jitter_{\%}$ son los parámetros empleados cotidianamente para cuantificar las perturbaciones glóticas en el análisis acústico de la voz [41, 59]. Se define de la siguiente forma:

$$Shimmer_{\%} = 100 \, \frac{Shimma}{\hat{A}_0} = 100 \, \frac{\frac{1}{N-1} \sum_{i=2}^{N} |A_i - A_{i-1}|}{\frac{1}{N} \sum_{i=1}^{N} A_i}. \tag{4.15}$$

- *Coeficiente de perturbación de la amplitud:*

$$APQ3 = \frac{\frac{1}{N-2} \sum_{i=2}^{N-1} \left| \frac{A_{i-1} + A_i + A_{i+1}}{3} - A_i \right|}{\hat{A}_0}. \tag{4.16}$$

  De forma similar, se definen los coeficientes de perturbación de la amplitud $APQ5$ y $APQ11$. Se obtienen calculando la distancia entre la amplitud instantánea y el valor promedio para ventanas móviles de 5 y 11 muestras, respectivamente.

4.3.3. Influencia de las fluctuaciones en la estimación de las perturbaciones

Hasta aquí, estudiamos las dinámicas de la SP y la SA considerando que están formadas por una componente lenta, o fluctuación, y por otra componente instantánea y local que denominamos perturbación. Aun cuando este análisis es conceptualmente adecuado, en la práctica resulta sumamente difícil.

En ocasiones, se generan fluctuaciones por causas específicas, pero con una dinámica similar a la observada en las perturbaciones. Por ejemplo, en voces normales pueden observarse esta clase de fenómenos, llamados comúnmente *microtemblores vocales*, ocasionados por variaciones fisiológicas en la respiración o como producto de la actividad de los músculos laríngeos intrínsecos y extrínsecos [139, 163]. Esta situación se vuelve más notoria en voces patológicas. Por ejemplo, patologías como el *mal de Parkinson* y la disfonía espasmódica pueden dar lugar a espasmos o temblores vocales; en las *diplofonías* se desarrollan simultáneamente dos o más dinámicas oscilatorias [16, 163].

Los parámetros acústicos descriptos en las secciones anteriores arrojan medidas promedios de las perturbaciones en los períodos y en las amplitudes. Sin embargo, las fluctuaciones influyen negativamente en la mayoría de estas medidas, dando lugar a valores considerablemente mayores de las perturbaciones [16, 41, 125]. Los parámetros acústicos RAP y $PPQ5$ en el caso del *Jitter* y $APQ3$, $APQ5$ y $APQ11$ en el caso del *Shimmer* alivian esta problemática, involucrando en su definición la distancia entre cada elemento de la serie y el nivel promedio en una ventana móvil. Esto permite reducir la influencia de fluctuaciones con comportamiento aproximadamente lineal [16, 125].

Por otro lado, para eliminar la problemática anterior se sugiere eliminar la correlación temporal en las series analizadas previo a calcular las medidas de perturbación. Con esta idea, en [141, 142] se propuso eliminar la correlación temporal en un conjunto de SP empleando modelos AR estocásticos. Como resultado, los autores mostraron que esta estrategia produce valores más consistentes en la medida $PPQ5$.

De lo expuesto hasta aquí, podemos concluir que es necesario desarrollar nuevos métodos que permitan obtener mejores estimaciones de las perturbaciones, que además sean robustos ante diferentes dinámicas en las fluctuaciones. A su vez, se aprecia la falta de un criterio o método objetivo para la separación de las fluctuaciones, por un lado, y las perturbaciones, por el otro. Profundizaremos más respecto a estos dos problemas más adelante en este documento de tesis.

4.4. Método para la síntesis de vocales sostenidas con perturbaciones controladas

En la segunda parte de este capítulo, expondremos los primeros aportes realizados en el marco de esta tesis de doctorado. Estos se orientaron al desarrollo y evaluación de un método para la síntesis de vocales sostenidas, que permita simular perturbaciones controladas en los períodos y las amplitudes. Para ello, consideramos dos parámetros acústicos ampliamente utilizados en la fonoaudiología y la logopedia para caracterizar el *Jitter* y el *Shimmer* [41, 59].

Los trabajos realizados tuvieron como meta principal desarrollar un sistema pa-

ra la síntesis de vocales sostenidas que sirva, en un futuro, como material de apoyo para la formación y el entrenamiento de profesionales de la voz y para la evaluación de métodos computacionales para el análisis acústico de la voz. Los requisitos fundamentales de este sistema, tenidos en cuenta en su diseño, son: que genere voces con una alta calidad perceptual y que permita niveles controlados de perturbaciones en sus parámetros acústicos. Esto último, se fundamenta en la evidencia existente de que las perturbaciones permiten caracterizar la voz de una persona, tanto en condiciones normales como en presencia de patologías vocales [16, 41, 118, 119].

La versión inicial de este método fue desarrollada por Schlotthauer y colaboradores en el LSyDnL [136, 137]. Su propósito era sintetizar vocales sostenidas con *Jitter* y el *Shimmer* controlados. Estas señales se utilizaron para contrastar objetivamente el desempeño de diferentes algoritmos para la estimación de f_0. Partiendo de este desarrollo, dirigimos nuestros esfuerzos en mejorarlo concentrándonos principalmente en la calidad perceptual de las señales sintetizadas. De igual modo, adaptamos el diseño para que sea compatible con diferentes estrategias para emular el ruido glótico, o de aspiración, y el ruido acústico en el medio ambiente [79, 101].

El desarrollo del método propuesto se basó en la TFF (ver Sec. 3.3). Sólo se consideraron vocales sostenidas ya que, como dijimos anteriormente, son las principales emisiones estudiadas en el análisis acústico, tanto en el estudio de la voz de un individuo como en el seguimiento de una terapia vocal [16, 41]. Brevemente, el diseño planteado consiste en tres etapas principales: la representación del tracto vocal, la generación de la función glótica y la incorporación de niveles controlados de *Jitter* y *Shimmer*. A continuación, describiremos cada etapa por separado, para luego pasar a describir por completo el proceso de síntesis.

4.4.1. Filtro de tracto vocal y residuo de la voz

El tracto vocal se representó utilizando modelos estocásticos AR. Las características relevantes de estos modelos se describieron en la Sec. 3.4.1. Estos modelos permiten obtener filtros digitales causales que simulan el comportamiento del tracto vocal. Para el cálculo de los parámetros de los modelos, a partir de señales de voz, aplicamos el análisis predictivo lineal, en su versión basada en el método de la autocorrelación [108, 130, 158].

En nuestras simulaciones, empleamos las vocales /a/ sostenidas correspondientes a sujetos sanos, disponibles en la BDDV (ver Sec. 2.5.2), como material para la estimación del filtro del tracto vocal. Trabajamos con cada señal por separado. Originalmente, éstas presentaban una frecuencia de muestreo de 50 kHz. Para la estimación de los parámetros, submuestreamos cada señal a una nueva frecuencia de muestreo de 20 kHz y le aplicamos el filtro de *preénfasis* $P_{enf}(z) = 1 - \beta_{enf}\, z^{-1}$ con $0{,}9 \leq \beta_{enf} \leq 1$. Luego, tomamos ventanas de señal de 25 ms de duración, pesadas por una función de *Hamming*, y para cada una de ellas estimamos los parámetros aplicando el algoritmo de *Levinson* y *Durbin*. Con el fin de modelar adecuadamente la información espectral en el rango de las bajas frecuencias consideramos modelos AR de orden $\rho = 22$, de acuerdo con lo sugerido en [42, 128].

Seguidamente, aplicamos cada uno de los filtros de tracto vocal obtenidos para llevar a cabo el filtrado inverso de la señal de voz original (ver Sec. 3.6), generando de esta forma la señal de residuo, o error de predicción, correspondiente. Se considera que el residuo es una primera estimación de la función glótica. Así, el estudio de esta

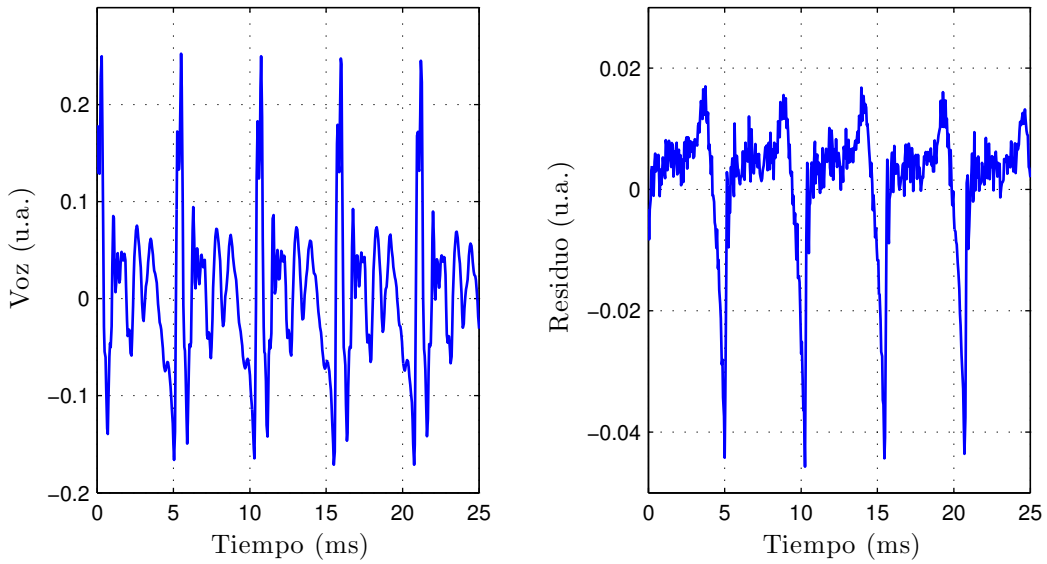


Figura 4.3: Filtrado inverso y estimación del residuo de una señal de voz. *Izquierda:* vocal /a/ sostenida sana de un sujeto de sexo masculino ($f_0 = 189,3$ [Hz], $Jitter_{\%} = 0,269\,\%$ y $Shimmer_{\%} = 1,826\,\%$). *Derecha:* residuo correspondiente calculado mediante filtrado inverso.

señal nos permitió seleccionar un único filtro de tracto vocal más representativo. El criterio empleado consistió en seleccionar aquel filtro capaz de generar un residuo con una dinámica similar al comportamiento teórico de la función glótica (ver Sec. 3.5). Por otro lado, utilizamos la información obtenida de los residuos para mejorar la síntesis de la función glótica. Esto último, se explicará con mayor detalle en la siguiente sección.

En la Fig. 4.3 presentamos, a la izquierda, 25 ms de una vocal /a/ sostenida sana correspondientes a un sujeto de sexo masculino y, a la derecha, el residuo correspondiente calculado mediante el filtrado inverso. Podemos apreciar que los pulsos del residuo cumplen con las principales propiedades de la función glótica, es decir, presentan un comportamiento regular y suave a excepción de las regiones donde ocurren los picos negativos prominentes.

4.4.2. Síntesis de la función glótica

Como dijimos anteriormente, la principal característica de las emisiones vocales sostenidas en sujetos sanos es su destacado comportamiento regular. Es decir, pueden interpretarse como una concatenación de una misma forma de onda característica con pequeñas alteraciones. Esto puede apreciarse en el ejemplo mostrado a la izquierda de la Fig. 4.3. Entre las diferentes modificaciones se encuentran las perturbaciones en los períodos y en las amplitudes.

Inspirados en la TFF, consideramos como hipótesis que las perturbaciones en los períodos y en las amplitudes tiene su origen en la función glótica y que el tracto vocal no las modifica apreciablemente, por lo que se preservan en la señal de voz. Claramente, esto resulta poco realista teniendo en cuenta que las perturbaciones son ocasionadas por mecanismos distribuidos por todo el aparato fonador, como dijimos

en la Sec. 4.3. Sin embargo, decidimos conservar esta hipótesis ya que da lugar a simplificaciones prácticas que permiten modelar mejor el problema.

De acuerdo con la hipótesis propuesta, la regularidad de la función glótica puede modelarse a partir de un tren de impulsos con amplitudes y períodos variables. Analíticamente, se lo define de la siguiente forma:

$$\Delta[n] = \sum_{i=1}^{N} A_i \; \delta \left[n - \sum_{j=1}^{i} T_j \right], \tag{4.17}$$

donde A_n y T_n son la amplitud y el período del n-ésimo pulso, respectivamente, N es la cantidad de ciclos considerados, y δ es la función *delta de Kronecker* discreta. Nuevamente, $f_n = 1/T_n$ es la frecuencia del n-ésimo pulso. La calidad de las voces sintetizadas considerando a $\Delta[n]$ como función glótica no resultó satisfactoria, ya que se percibían como un sonido artificial desagradable y alejado de la voz real que se pretendía emular.

En un intento por mejorar la calidad perceptual de las voces sintetizadas, decidimos construir *plantillas* $g[n]$ para dar forma a los pulsos de $\Delta[n]$. Obtuvimos estas *plantillas* a partir de los residuos estimados aplicando el filtrado inverso. Para ello, en cada residuo procedimos a tomar los pulsos sucesivos y a normalizarlos tanto en amplitud como en duración. Luego, la *plantilla* $g[n]$ corresponde al promedio de estos pulsos. Es así que, finalmente, construimos las fuentes glóticas concatenando copias de la *plantilla* $g[n]$, de forma tal que la amplitud y el período de cada pulso siga el modelo (4.17).

Considerando las fuentes glóticas construidas aplicando este último método, la calidad perceptual de las voces sintetizadas mejoró apreciablemente. Es importante destacar que este proceso presenta las siguientes características importantes:

- Captura la regularidad que caracteriza a las vocales sostenidas.
- La morfología de cada pulso es semejante a la observada en señales reales.
- Permite introducir perturbaciones controladas en la amplitud y en el período de cada pulso.

4.4.3. Perturbaciones controladas

El objetivo propuesto es desarrollar un método para la síntesis de vocales sostenidas con perturbaciones controladas en los períodos y en las amplitudes. Tomando en cuenta la estrategia para la generación de la función glótica introducida en la sección anterior, concluimos que una alternativa interesante consiste en modificar convenientemente los períodos T_n y las amplitudes A_n del modelo (4.17).

Para modificar estos coeficientes se han propuesto diferentes estrategias basadas tanto en leyes determinísticas como estocásticas [29, 116, 134, 142]. Aquí proponemos generarlos a partir de modelos estocásticos para el *Jitter* y el *Shimmer*, que involucren en su definición los parámetros acústicos $Jitter_{\%}$ y $Shimmer_{\%}$, que son utilizados cotidianamente en la práctica médica. A continuación, presentaremos estos modelos.

Shimmer

Como dijimos anteriormente, se denomina *Shimmer* a las perturbaciones en las amplitudes de los pulsos sucesivos en la señal de voz. A su vez, la medida más empleada en la práctica médica para cuantificar este fenómeno es $Shimmer_{\%}$, definido por la Ec. (4.7). Tomando como referencia el modelo (4.17), desarrollaremos a continuación un modelo estocástico para obtener las amplitudes A_n.

En lo que sigue, tomaremos como hipótesis que las amplitudes A_n son variables aleatorias independientes entre sí, idénticamente distribuidas y con comportamiento normal, o gaussiano. Es decir, $A_n \sim \mathcal{N}(\hat{A}_0, \sigma_A^2)$, donde $\hat{A}_0$ es la amplitud media y σ_A^2 su varianza. Estas ideas han sido empleadas anteriormente por otros autores para la síntesis de voz con perturbaciones aleatorias en los períodos y en las amplitudes [92, 117]. De lo expuesto, se desprende que la variable aleatoria $\Delta A_n = A_n - A_{n-1}$ también presenta comportamiento normal, es decir, $\Delta A_n \sim \mathcal{N}(0, 2\,\sigma_A^2)$. Luego, el valor absoluto $|\Delta A_n| = |A_n - A_{n-1}|$ se convierte en una variable aleatoria con distribución heminormal[2], con función densidad de probabilidad:

$$\begin{cases} \frac{1}{\left(4\,\pi\,\sigma_A^2\right)^{1/2}}, & \text{si } |\Delta A_n| = 0, \\ \frac{2}{\left(4\,\pi\,\sigma_A^2\right)^{1/2}}\, e^{\left(\frac{-1}{4\,\sigma_A^2}\,|\Delta A_n|^2\right)}, & \text{si } |\Delta A_n| > 0, \\ 0, & \text{en otro caso.} \end{cases} \tag{4.18}$$

Del análisis anterior y aplicando el operador esperanza a la variable aleatoria $|\Delta A_n|$, se demuestra que:

$$\mathcal{E}\left\{|\Delta A_n|\right\} = \int_0^\infty \frac{2\,|\Delta A_n|}{\left(4\,\pi\,\sigma_A^2\right)^{1/2}}\, e^{\left(\frac{-1}{4\sigma_A^2}\,|\Delta A_n|^2\right)}\, d|\Delta A_n| = \frac{2\sigma_A}{\sqrt{\pi}}. \tag{4.19}$$

Por teoría estadística, en el límite cuando $N \to \infty$ se obtienen las siguientes relaciones $\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i-1}| \to \mathcal{E}\left\{|\Delta A_n|\right\}$ y $\frac{1}{N}\sum_{i=1}^{N} A_i \to \hat{A}_0$. Considerando esto último junto con la Ec. (4.19) y reemplazando lo anterior en la definición (4.15), se demuestra que en el límite cuando $N \to \infty$ se cumple:

$$\sigma_A^2 = \pi \left(\frac{\hat{A}_0\, Shimmer_{\%}}{200}\right)^2. \tag{4.20}$$

Jitter

Por su parte, el parámetro acústico más importante para cuantificar las perturbaciones en los períodos es $Jitter_{\%}$ [41, 59]. Considerando nuevamente el modelo (4.17), presentaremos aquí un modelo estocástico para la generación de los períodos T_n.

Procediendo de igual modo que con las amplitudes, tomaremos como hipótesis que los períodos T_n son variables aleatorias independientes entre sí, idénticamente distribuidas y con comportamiento normal. Es decir, $T_n \sim \mathcal{N}(\hat{T}_0, \sigma_T^2)$, donde $\hat{T}_0$ es el período medio y σ_T^2 su varianza. Luego, trabajando de modo análogo se obtiene la siguiente expresión:

$$\sigma_T^2 = \pi \left(\frac{\hat{T}_0\, Jitter_{\%}}{200}\right)^2. \tag{4.21}$$

[2]En inglés, "half-normal distribution".

Tabla 4.1: Valores medio, máximo y mínimo de las medidas $Shimmer_{\%}$ y $Jitter_{\%}$, correspondientes a las voces sanas y patológicas de la BDDV.

Voces	Parámetro	Media (DE*)	Mínimo	Máximo
Sanas	$Shimmer_{\%}$	2,205 (0,924)	0,963	4,802
	$Jitter_{\%}$	0,615 (0,437)	0,175	2,529
Patológicas	$Shimmer_{\%}$	7,103 (5,027)	1,230	31,296
	$Jitter_{\%}$	2,539 (2,838)	0,212	21,322

* DE: Desvío estándar.

4.4.4. Resumen del método de síntesis

Las expresiones obtenidas en la sección anterior completan el método de síntesis buscado. Éstas permiten la síntesis de vocales sostenidas con perturbaciones en los períodos y las amplitudes y, a su vez, satisfacen las siguientes propiedades:

- Desarrollan un período medio $\hat{T}_0$ y una amplitud media $\hat{A}_0$.
- Las perturbaciones se controlan con los parámetros $Jitter_{\%}$ y $Shimmer_{\%}$.

Lo anterior es cierto siempre que se construya la función glótica de modo tal que su regularidad se rija por el modelo (4.17), al mismo tiempo que los períodos T_n y las amplitudes A_n de sus pulsos se generen aleatoriamente satisfaciendo las funciones densidad de probabilidad $\mathcal{N}(\hat{T}_0, \sigma_T^2)$ y $\mathcal{N}(\hat{A}_0, \sigma_A^2)$, respectivamente. Esto último demuestra la importancia de las expresiones (4.20) y (4.21), ya que permiten calcular las varianzas σ_T^2 y σ_A^2 en función de los parámetros de síntesis $\hat{T}_0$, $\hat{A}_0$, $Jitter_{\%}$ y $Shimmer_{\%}$.

A modo informativo, en la Tab. 4.1 presentamos los valores medio, mínimo y máximo de los parámetros $Shimmer_{\%}$ y $Jitter_{\%}$ para las señales de la BDDV. De acuerdo a la información técnica disponible, para cada señal se calcularon los valores de $Shimmer_{\%}$ y $Jitter_{\%}$, junto con otros parámetros, aplicando el programa de computadora MDVP [110]. Podemos observar que las voces patológicas presentan valores más elevados y una mayor dispersión para ambas medidas, en comparación con las voces normales.

En la Fig. 4.4 presentamos un diagrama de flujo explicativo del método de síntesis aquí desarrollado. Como comentamos anteriormente, los parámetros de síntesis son: amplitud media $\hat{A}_0$ y período medio $\hat{T}_0$ y los valores deseados de $Shimmer_{\%}$ y $Jitter_{\%}$. A su vez, deben especificarse la *plantilla* $g[n]$ y el filtro del tracto vocal. De forma opcional, puede incorporarse a la señal artificial ruido glótico o de medio ambiente de naturaleza diversa [79, 101].

El proceso de síntesis se realiza del siguiente modo. En primer lugar, se considera un tren de impulsos de la forma (4.17), con período $\hat{T}_0$ y amplitud $\hat{A}_0$ constantes, y se incorporan las perturbaciones en los períodos y las amplitudes siguiendo las distribuciones $\mathcal{N}(\hat{T}_0, \sigma_T^2)$ y $\mathcal{N}(\hat{A}_0, \sigma_A^2)$, respectivamente. Luego, se adecua la morfología de cada pulso a partir de la *plantilla* $g[n]$, generándose así la función glótica con perturbaciones controladas. En este punto, se le agrega a la función glótica el ruido de aspiración requerido. A continuación, se procesa la función glótica ruidosa con el filtro del tracto vocal y se obtiene como resultado la señal de voz. Finalmente, se le

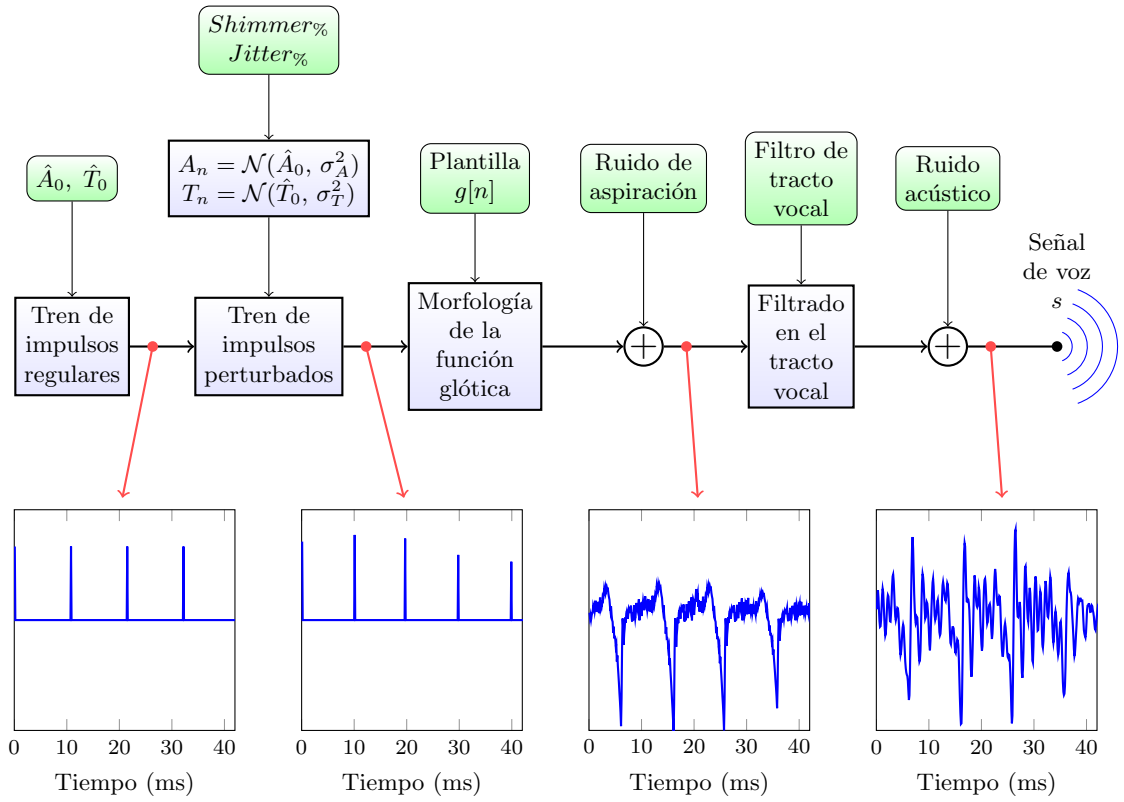


Figura 4.4: Diagrama de flujo explicativo del método de síntesis de vocales sostenidas con perturbaciones controladas en los períodos y en las amplitudes. Se identifican los parámetros de síntesis y las etapas principales del método.

agrega a la señal de voz el ruido acústico especificado. En las simulaciones realizadas consideramos ruido glótico y acústico con comportamiento gaussiano y blanco.

4.4.5. Simulaciones y resultados

En esta sección estudiaremos el desempeño del método de síntesis desarrollado. Para ello, llevamos a cabo un conjunto de simulaciones con el propósito de analizar dos aspectos importantes. En primer lugar, estudiamos el desempeño de los modelos estocásticos para el *Jitter* y el *Shimmer* desarrollados en la Sec. 4.4.3, en relación a su capacidad para generar los niveles de perturbaciones estipulados. Por otro lado, evaluamos la calidad perceptual de las señales sintetizadas con este método. A continuación, describiremos las simulaciones realizadas y analizaremos los resultados más relevantes.

Modelos estocásticos para el *Jitter* y el *Shimmer*

La característica más importante del método propuesto es que permite generar vocales sostenidas con perturbaciones controladas en los períodos y en las amplitudes. Esta capacidad se rige por los modelos estocásticos desarrollados en la Sec. 4.4.3. Presentaremos aquí diferentes simulaciones diseñadas para evaluar el desempeño de estos modelos.

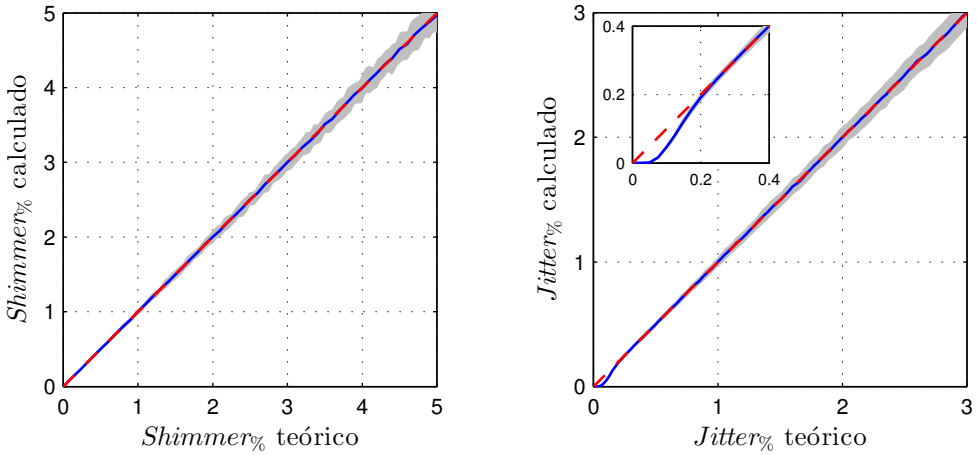


Figura 4.5: Desempeño de los modelos estocásticos de *Shimmer* y de *Jitter*. Valores de $Shimmer_{\%}$, izquierda, y de $Jitter_{\%}$, derecha, arrojados por los modelos estocásticos en función de los parámetros teóricos. En cada figura se presentan: el valor medio en línea continua azul, la incerteza correspondiente en la región llena gris, y el comportamiento ideal en línea discontinua roja.

En una primera instancia, sintetizamos fuentes glóticas con una frecuencia de muestreo $f_s = 50$ kHz y una frecuencia fundamental $f_0 = 189$ Hz. A su vez, consideramos valores de $Jitter_{\%}$ y $Shimmer_{\%}$ en los rangos $0,00 \leq Jitter_{\%} \leq 3,00$ y $0,00 \leq Shimmer_{\%} \leq 5,00$, con un paso de 0,05 en cada caso. Los extremos corresponden a la información clínica para las voces sanas, que acompaña a la BDDV (ver Tab. 4.1). Fijando uno de los parámetros y variando el otro, se construyeron 100 realizaciones de la fuente glóticas a partir de los modelos estocásticos para las amplitudes y los períodos. Luego se calcularon los valores de $Jitter_{\%}$ y $Shimmer_{\%}$ para el conjunto de señales, aplicando las expresiones (4.7) y (4.15), respectivamente.

En la Fig. 4.5 mostramos los valores calculados para el $Shimmer_{\%}$, izquierda, y para el $Jitter_{\%}$, derecha, en función de los valores teóricos, abscisas, correspondientes a cada modelo. En cada gráfico, representamos en línea continua azul el valor medio calculado para cada parámetro y en la región llena gris la incerteza correspondiente, estimada mediante el valor medio más/menos dos veces el desvío estándar. Para ayudar al análisis, indicamos también en línea discontinua roja el desempeño ideal de los modelos.

Los coeficientes de correlación entre los valores medios estimados y los parámetros teóricos son 0,999986 para el $Shimmer_{\%}$ y 0,999939 para el $Jitter_{\%}$. En ambos gráficos, observamos que los valores calculados coinciden con los parámetros teóricos, para casi la totalidad de los rangos analizados. Por otro lado, apreciamos que al aumentar la magnitud de $Shimmer_{\%}$ y de $Jitter_{\%}$, aumenta notablemente la dispersión de los valores generados con los modelos desarrollados. Esto último es una consecuencia de la naturaleza estocástica de estos modelos y puede resultar de suma utilidad para simular las perturbaciones encontradas en las señales de voz reales, en especial en casos patológicos [16, 92, 117].

En el caso particular del $Jitter_{\%}$, gráfica de la derecha en la Fig. 4.5, podemos observar que el desempeño del modelo estocástico se aleja ligeramente del comportamiento ideal para valores menores a 0,2. Esto puede apreciarse mejor en la sección

ampliada para el rango $0{,}00 - 0{,}04$, superpuesta en el extremo superior izquierdo de la gráfica. Esto último se debe, principalmente, al fenómeno de muestreo de la función glótica. Considerando la Ec. (4.7) para valores pequeños de $Jitter_{\%}$, podemos identificar una limitación física para el cálculo de $|T_n - T_{n-1}|$. En señales discretas, éste sólo puede tomar valores múltiplos enteros del período de muestreo $1/f_s$. En consecuencia, la capacidad de reconocer como diferentes dos períodos sucesivos depende exclusivamente de la f_s empleada. Así, para valores de $Jitter_{\%}$ pequeños los períodos toman valores similares, lo que ocasiona que el $Jitter_{\%}$ calculado resulte considerablemente menor al valor teórico correspondiente. A su vez, para $Jitter_{\%}$ cercanos a cero todos los períodos resultan prácticamente iguales, por lo que el valor calculado cae rápidamente a cero.

Para corroborar esta última hipótesis, generamos un conjunto de señales variando la frecuencia de muestreo y evaluamos nuevamente el desempeño del modelo estocástico de *Jitter*. Las f_s consideradas fueron: 35, 50, 75 y 100 kHz. En la gráfica izquierda de la Fig. 4.6 presentamos el $Jitter_{\%}$ calculado, aplicando la Ec. 4.7, en función del valor teórico correspondiente, para las diferentes f_s consideradas. A fines comparativos, mostramos también en línea punteada roja el desempeño ideal del modelo estocástico. En este caso, nos concentramos exclusivamente en el rango $0,00 \leq Jitter_{\%} \leq 0{,}60$. Note el lector que la curva generada para $f_s = 50$ kHz coincide con la que se observa a la derecha de la Fig. 4.5.

Podemos apreciar que al aumentar f_s mejora el desempeño del modelo de *Jitter*. Cabe destacar que el comportamiento cerca del origen para $f_s = 100$ kHz está fuertemente determinado por la grilla de valores teóricos considerados. En este caso, no utilizamos una grilla lo suficientemente densa, lo que ocasionó que el desempeño para $f_s = 100$ kHz resultara muy similar a la situación ideal. Por lo tanto, es esperable que puedan apreciarse mejor los detalles si se aumentan los valores analizados. Un aspecto importante a tener en cuenta es que utilizar una frecuencia de muestreo alta trae aparejado, entre otras cosas, un costo computacional mayor en la síntesis de las vocales y una mayor dificultad en la estimación de los coeficientes del filtro del tracto vocal y de la función glótica a partir del residuo [55, 116].

De la expresión (4.7) podemos inferir que el modelo estocástico de *Jitter* también resulta sensible a los valores del período promedio $\hat{T}_0$. Recordando nuevamente el fenómeno de muestreo de la función glótica al disminuir $\hat{T}_0$, lo que implica aumentar f_0, se reduce la cantidad de muestras para cada pulso glótico. Paralelamente, se disminuye considerablemente la perturbación efectiva en los períodos, para un valor de $Jitter_{\%}$ dado. Esto último, se debe a que en la Ec. (4.21) la varianza σ_T^2 es proporcional a $\hat{T}_0^2$. Todo lo anterior da lugar a valores en los períodos muy similares, lo que ocasiona que el $Jitter_{\%}$ calculado se aleje más aún del comportamiento teórico.

En la gráfica derecha de la Fig. 4.6 mostramos este fenómeno para voces sintetizadas considerando valores de f_0 iguales a 189 y 230 Hz, en comparación con el comportamiento ideal del modelo representado en línea discontinua roja. Estos valores de f_0 pueden interpretarse como correspondientes a sujetos sanos de sexo masculino y femenino, respectivamente. Note el lector que la curva para $f_0 = 189$ Hz coincide con la presentada en la gráfica derecha de la Fig. 4.5. Podemos apreciar que para f_0 mayores, aumenta el rango de valores de $Jitter_{\%}$ para los cuales el modelo se aleja del comportamiento ideal.

En la Tab. 4.1 se observa que el $Jitter_{\%}$ mínimo es 0,175 en la BDDV. Teniendo en cuenta esto, los fenómenos analizados arriba parecen ser un falla del modelo de

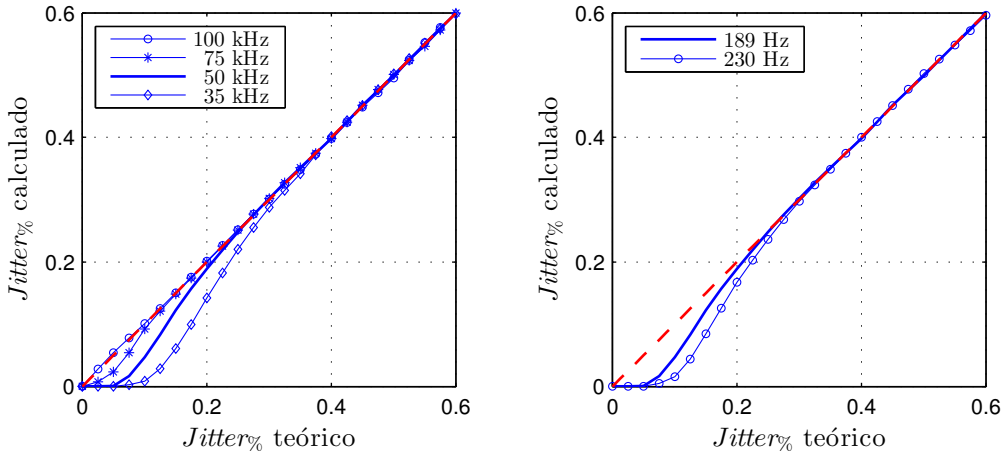


Figura 4.6: Desempeño del modelo estocástico de *Jitter* para diferentes frecuencias de muestreo f_s, a la izquierda, y diferentes frecuencias fundamentales f_0, a la derecha. En cada gráfica, se muestra el $Jitter_{\%}$ estimado en función del valor teórico y se representa el comportamiento ideal en línea discontinua roja. *Izquierda:* los valores de f_s considerados son: 35, 50, 75 y 100 kHz, con $f_0 = 189$ Hz. *Derecha:* los valores de f_0 considerados son: 189 y 230 Hz, con $f_s = 50$ kHz.

Jitter. Sin embargo, al analizar en detalle la BDDV encontramos que el $Jitter_{\%}$ toma valores muy por encima de este mínimo para la mayoría de las señales. Así, lo expuesto hasta aquí nos permite concluir que el método de síntesis propuesto es aplicable a la síntesis de vocales sostenidas con perturbaciones controladas por los parámetros $Shimmer_{\%}$ y $Jitter_{\%}$, con la restricción $Jitter_{\%} > 0{,}2$. Considerando la información expuesta en la Tab. 4.1, podemos afirmar a su vez que este modelo es apto para simular los valores de $Shimmer_{\%}$ y $Jitter_{\%}$ observados tanto en voces sanas como patológicas. Podemos concluir también que para generar vocales sostenidas con un $Jitter_{\%}$ menor a 0,2, será importante seleccionar con una f_s acorde y controlar el desempeño del modelo para la f_0 considerada.

Por otro lado, es importante destacar que en todas las simulaciones realizadas representamos las amplitudes de cada función glótica utilizando números en coma flotante de doble precisión, por lo que el efecto en su cuantificación resultó despreciable. Sin embargo, si se empleara una estrategia de cuantificación de menor precisión, es esperable que el modelo de *Shimmer* se comporte de forma incorrecta para valores pequeños de $Shimmer_{\%}$, de forma similar a lo observado en el modelo de *Jitter*.

Análisis de la calidad perceptual

A los efectos de evaluar la calidad perceptual de las vocales sintetizadas con el método propuesto, empleamos una medida objetiva denominada *Evaluación Perceptual de la Calidad en el Habla* ($PESQ$)[3]. Ésta se define en la norma *ITU P.862: Evaluación de la calidad vocal por percepción*, destinada a establecer criterios objetivos para cuantificar la calidad vocal de extremo a extremo, en redes telefónicas de banda estrecha y aplicando códecs vocales [123]. Esta medida ha sido ampliamente estudiada

[3] En inglés, "Perceptual Evaluation of Speech Quality".

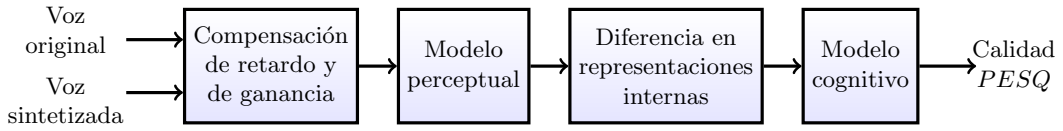


Figura 4.7: Diagrama de flujo explicativo del algoritmo para el cálculo de la medida objetiva *Evaluación Perceptual de la Calidad en el Habla* ($PESQ$).

y se ha demostrado que tiene una correlación muy alta con las mediciones subjetivas de calidad perceptual para señales de voz sintetizadas, en una amplia variedad de situaciones [81, 101].

En la Fig. 4.7 presentamos un diagrama de flujo explicativo del algoritmo para la estimación de $PESQ$. Este algoritmo mide la calidad perceptual de una voz artificial en comparación con otra señal de referencia, que en general es la voz original sin ninguna modificación. Emplea diferentes niveles de análisis en un intento de imitar los procesos involucrados en la percepción humana. La primera etapa consiste en una compensación de ganancia y de retardo, con el propósito de alinear y equiparar las dos señales. Luego, se realiza una transformación a un dominio perceptual y allí se estima la densidad de distorsión a partir de la diferencia entre la voz analizada y la de referencia. Finalmente, se aplican diferentes modelos cognitivos para el cálculo de $PESQ$ [123]. En nuestras simulaciones, aplicamos una versión no lineal de $PESQ$ que arroja valores en el rango $0 - 4{,}5$, con 0 y 4,5 indicando una pobre y una buena calidad perceptual, respectivamente [81]. El algoritmo empleado está disponible libremente en `ecs.utdallas.edu/loizou/speech/software.htm`.

En este estudio, empleamos las señales de la BDDV. Así, generamos un conjunto de 53 vocales sostenidas correspondientes a las señales para sujetos sanos. Para sintetizar cada señal, consideramos una frecuencia de muestreo $f_s = 20$ kHz y seleccionamos los valores de $\hat{T}_0$, $\hat{A}_0$, $Jitter_{\%}$ y $Shimmer_{\%}$ obtenidos de la señal real. Empleamos también el filtro del tracto vocal y la *plantilla* $g[n]$ estimados a partir de la voz original. Seguidamente, las vocales sintetizadas debieron submuestrearse a 16 kHz, ya que la implementación de $PESQ$ aplicada es compatible únicamente con señales muestreadas a 8 y a 16 kHz. Como dijimos anteriormente, el algoritmo evalúa la calidad perceptual de la voz sintetizada en comparación con la señal original. Para ello, extrajimos ventanas de ambas señales y con ellas calculamos el $PESQ$ correspondiente. Escogimos la longitud de las ventanas tomando en consideración que el algoritmo de $PESQ$ se diseñó para la evaluación de la calidad en habla continua y que, en esa situación, la estabilidad de las vocales se garantiza sólo en períodos cortos [42, 81, 123, 130]. A su vez, en registros de vocales sostenidas de larga duración tanto la amplitud como el período fundamental presentan fluctuaciones que no son consideradas por el $PESQ$ y que pueden influir de forma adversa en esta medida [16, 162, 163]. Estas dos situaciones fueron descriptas oportunamente en las Secs. 2.3 y 4.3, respectivamente.

En la Fig. 4.8 presentamos los gráficos de cajas de los valores de $PESQ$ para el conjunto de vocales sostenidas simuladas aplicando el método de síntesis propuesto, considerando ventanas de 0,1 y 1,0 s de duración. A su vez, en la Tab. 4.2 detallamos los estadísticos más relevantes extraídos de los gráficos de caja. De toda esta infor-

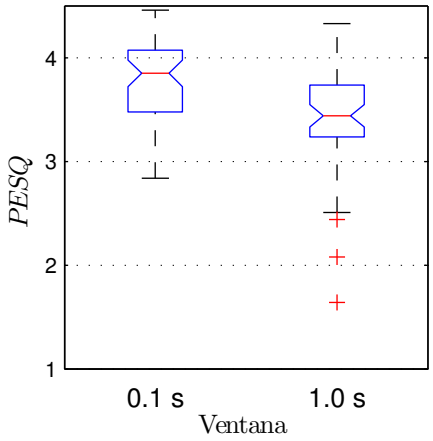


Figura 4.8: Gráfico de cajas de los valores de $PESQ$ calculados para el conjunto de vocales sostenidas simuladas, considerando ventanas de 0,1 s y 1,0 s de duración.

Tabla 4.2: Estadísticos correspondientes a los valores de $PESQ$ calculados en las simulaciones. Valores extraídos del gráfico de cajas de la Fig 4.8, donde Q_1 y Q_3 son el primer y el tercer cuartil, respectivamente.

Ventana	Mediana	Q_1	Q_3	Mínimo	Máximo
0,1 s	3,85	3,48	4,07	2,84	4,46
1,0 s	3,44	3,23	3,74	1,64	4,33

mación, podemos apreciar que el comportamiento de la medida $PESQ$ varía con la longitud de la ventana. Los resultados sugieren que las vocales sintetizadas con el método propuesto presentan una alta calidad perceptual para ventanas de señal de corta duración. Sin embargo, las mismas señales obtienen una calificación menor con la medida $PESQ$ si se consideran ventanas de mayor longitud. Esto puede apreciarse fácilmente comparando las medianas y los mínimos de las cajas.

Corroboramos estadísticamente la idea anterior a partir de la prueba de la *suma de rangos de Wilcoxon* [93]. Establecimos como hipótesis nula H_0 : *la mediana de los valores de* $PESQ$ *para ventanas de* 0,1 s *es menor o a lo sumo igual a la mediana de los valores calculados en ventanas de* 1,0 s y, como hipótesis alternativa, H_1 : *la mediana de los valores de* $PESQ$ *para ventanas de* 0,1 s *es mayor a la mediana de los valores calculados en ventanas de* 1,0 s. Esta prueba arrojó como resultado un *valor p* menor a 0,001. Como conclusión, existe evidencia significativa que respalda la hipótesis alternativa. Es decir, la calidad de las vocales sostenidas es menor para ventanas de 1,0 s de duración.

Evaluamos, también, la calidad de las voces generadas con el método propuesto a través de pruebas subjetivas informales, es decir, pruebas que no siguieron ningún protocolo preestablecido. Participaron 4 voluntarios, todos ellos con experiencia en el área de procesamiento de señales. A cada voluntario, se le hizo escuchar dos vocales sostenidas, una real y otra sintetizada, y se le solicitó que indicara si percibía a estas señales como iguales o diferentes. En el caso de señalar que eran diferentes, se le solicitó también que comparara y evaluara la naturalidad de estas señales, justificando su razonamiento. Esta prueba se repitió considerando ventanas de señal

de corta y de larga duración. En la mayoría de los casos, los voluntarios advirtieron que las señales eran diferentes entre sí. A su vez, estos informaron que percibieron a las voces sintetizadas como naturales en ventanas de cortas duración y no así en ventanas de larga duración. Esto último lo atribuyeron, principalmente, a que percibieron una dinámica poco realista en las voces artificiales o a que notaron diferentes fenómenos transitorios en las voces reales que no advirtieron en las sintetizadas.

Del análisis anterior, podemos apreciar que en las voces reales existen características importantes, desde el punto de vista perceptual, que influyen en la naturalidad y en la calidad asignada a un registro sostenido de larga duración. A su vez, podemos inferir que estas características no son representadas correctamente por el método de síntesis propuesto. Es importante destacar que esto no sería un inconveniente para la síntesis de voz hablada, donde se considera que la estabilidad en vocales se mantiene por períodos cortos [42, 54, 130]. Sin embargo, estas características perceptuales cobran una mayor importancia en aplicaciones en la medicina y en la fonoaudiología, debido a que normalmente se emplean registros de vocales sostenidas para el diagnóstico del aparato fonador y el seguimiento de una terapia [16, 19, 41, 119]. Recordemos que uno de los principales objetivos planteados en esta etapa fue que este método de síntesis permita generar vocales sostenidas con perturbaciones controladas en sus parámetros acústicos, las cuales sirvieran para el entrenamiento y la capacitación de fonoaudiólogos y de terapistas de la voz. Consideramos entonces que debe mejorarse el método desarrollado, incorporando en las señales sintetizadas las características perceptuales encontradas en las vocales reales.

4.5. Comentarios finales

En este capítulo, nos abocamos al estudio y modelado de las perturbaciones en la voz. En particular, nos concentramos en las perturbaciones en los períodos y en las amplitudes. Comenzamos definiendo las series de períodos y de amplitudes, y describiendo el procedimiento necesario para su construcción a partir de la señal de voz. A continuación, analizamos la dinámica de estas señales representando a cada una de ellas como la superposición de dos fenómenos bien diferenciados: las fluctuaciones y las perturbaciones. Motivados por este análisis, presentamos luego diversos parámetros acústicos desarrollados para cuantificar las perturbaciones. Cerramos la sección explicando cómo las fluctuaciones influyen en la caracterización de las perturbaciones.

Dedicamos la segunda parte de este capítulo a presentar el primero de los aportes de esta tesis de doctorado. Se trata del desarrollo de un método para la síntesis de vocales sostenidas que contempla perturbaciones controladas en los períodos y en las amplitudes y, a su vez, permite generar señales con una alta calidad perceptual. Para ello, propusimos modelos estocásticos de *Shimmer* y de *Jitter* ideados a partir de los parámetros acústicos $Shimmer_{\%}$ y $Jitter_{\%}$, respectivamente. Se escogieron estos parámetros por dos importantes razones: son ampliamente empleados en la práctica médica y poseen una definición matemática compatible con los modelos estocásticos considerados. Luego, evaluamos a partir de diferentes simulaciones las principales características del método propuesto y expusimos los resultados alcanzados.

Las simulaciones presentadas en la Sec. 4.4.5 probaron que las voces artificiales, sintetizadas con el método propuesto, mostraron un alta calidad perceptual para ventanas de corta duración. Además, esta calidad se deterioró al considerar venta-

nas de mayor duración. Las pruebas subjetivas llevadas a cabo señalaron que los modelos estocásticos de *Shimmer* y de *Jitter* imprimen una dinámica en las vocales sintetizadas que se aleja considerablemente de la encontrada en los casos reales. Todo esto nos permite afirmar que los modelos estocásticos desarrollados aquí, por su simplicidad, no son capaces de generar perturbaciones y fluctuaciones lo suficientemente realistas. Consideramos, además, que en emisiones de larga duración las fluctuaciones en las amplitudes y en los períodos cobran una mayor preponderancia en la naturalidad con que se percibe una voz. Suponemos entonces que el uso de modelos estocásticos más versátiles permitirá mitigar este inconveniente. Impulsados por esta hipótesis, en el Cap. 6 presentaremos una representación alternativa para las series de períodos T_n y de amplitudes A_n basada en el modelado estructural en espacio de estados.

Otro aspecto a analizar involucra el cálculo de la función glótica y del filtro del tracto vocal. En la Sec. 3.6 introdujimos el concepto de filtrado inverso de la voz. Brevemente, si se conoce el filtro del tracto vocal es posible calcular su filtro inverso correspondiente y, luego, procesar con este último la señal de voz para calcular la función glótica. En la Sec. 4.4.1, aplicamos este método para calcular las *plantillas* $g[n]$. Éstas se emplearon posteriormente para la síntesis de las funciones glóticas con perturbaciones en los períodos y en las amplitudes, de acuerdo a lo expuesto en la Sec. 4.4.2. El éxito del filtrado inverso depende de conocer el filtro del tracto vocal verdadero [52, 55, 104]. Esto no sucede en la práctica, por lo que resulta indispensable obtener buenas estimaciones. Desde hace un tiempo, se ha demostrado que incorporar la información de la función glótica permite mejorar la estimación del filtro del tracto vocal [5, 46, 174]. En el Cap. 7, presentaremos un modelo estocástico del proceso de fonación que, combinado con los métodos en espacio de estados, permite la estimación de la información de la función glótica y del tracto vocal, de forma simultánea.

Como cierre de este capítulo, es importante mencionar que los trabajos dirigidos al desarrollo y la evaluación del método para síntesis de vocales sostenidas con perturbaciones controladas en las amplitudes y los períodos dieron lugar a dos presentaciones en congresos nacionales de la especialidad [1, 2], así como a un artículo en la revista *Latin American Applied Research* [3].

Capítulo 5

Métodos en espacio de estados

5.1. Introducción

Se denominan métodos en espacio de estados a aquellos métodos comprendidos por la teoría del filtrado estocástico guiado por modelos en espacio de estados [30, 86]. Desde los primeros aportes, realizados por Kalman en la década de 1960, estas técnicas han despertado un gran interés en la comunidad científica, no sólo en el campo de la ingeniería sino en diversas áreas como estadística, economía, biología, medicina y física. A su vez, han sido ampliamente utilizados para el estudio de sistemas y procesos en diversas tareas como seguimiento, control, análisis, predicción y detección de fallas, por nombrar sólo algunas.

En el procesamiento de la señal de voz, en particular, los métodos en espacio de estados se han utilizado en aplicaciones como la síntesis de voz [171], la estimación de la frecuencia fundamental [65, 170], el seguimiento de las formantes y anti formantes [111], la descomposición de la señal de voz [67] y el modelado de la función glótica [98], entre otros.

El término filtrado debe interpretarse aquí cuidadosamente. En ingeniería, el filtrado tradicional consiste en enfatizar la información espectral relevante y, al mismo tiempo, atenuar aquellos componentes asociados al ruido o a fenómenos de interferencia [42, 101, 106]. Este proceso se basa principalmente en la distribución de la información en el dominio de la frecuencia. En cambio, el filtrado estocástico permite estimar la información estadística de una señal o un proceso de interés, de forma óptima, a partir de sus características estadísticas propias y de un conjunto de mediciones (observaciones) correspondientes a dicho fenómeno [8, 108]. Este proceso centra su atención en la dinámica temporal y en la naturaleza aleatoria del fenómeno estudiado. La estimación resulta óptima en el sentido que minimiza algún criterio o medida de error.

En los métodos en espacio de estados se considera que existe una dependencia funcional entre la información que se quiere estimar y las observaciones. Esta dependencia se describe matemáticamente mediante modelos en espacio de estados [30, 51, 86]. Los trabajos de esta tesis de doctorado involucran señales y sistemas de tiempo discreto y, por ello, en adelante sólo nos concentraremos en modelos de esta misma naturaleza. En particular, trabajaremos con modelos construidos a partir de sistemas de ecuaciones en diferencias estocásticas [83].

Es importante destacar que los métodos en espacio de estados forman un marco conceptual para el estudio y la representación de sistemas o señales, tanto lineales

como no lineales, con dinámicas aleatorias no estacionarias [30, 51]. Esto último, es una ventaja importante con respecto a otros métodos estocásticos, como el filtro de Wiener, que sólo están diseñados para fenómenos estacionarios [8]. A su vez, se han desarrollado diferentes algoritmos muy potentes, que permiten aplicar estos métodos en situaciones reales.

En la actualidad, existe una gran variedad de material dedicado a los métodos en espacio de estados, y otros tópicos relacionados, que abarca el artículo fundacional de Kalman [86] y otras obras [8, 30, 35, 51, 83]. Dedicaremos este capítulo a presentar los métodos en espacio de estados utilizados en el desarrollo de esta tesis de doctorado.

5.2. Modelos en espacio de estados

En esta sección, introduciremos varios de los modelos en espacio de estados empleados frecuentemente en la práctica. Como dijimos anteriormente, nos concentraremos específicamente en los modelos de tiempo discreto, con variable $n \in \mathbb{Z}$.

En estos modelos, se supone que el vector de estados $\mathbf{x}[n] \in \mathbb{R}^p$ agrupa la información que describe completamente el estado de una señal o un sistema en el instante n, alterada por ruido generado por fuentes internas. Además, se considera que $\mathbf{x}[n]$ no puede determinarse directamente, situación que es usual en sistemas reales. Por ello, se busca estimarlo de forma óptima, a partir de los estímulos (entradas) $\mathbf{u}[n] \in \mathbb{R}^u$ conocidos y de las mediciones (observaciones) ruidosas $\mathbf{z}[n] \in \mathbb{R}^r$ obtenidas del fenómeno. Las diferentes fuentes de ruido contemplan, a su vez, toda dinámica no representada en el modelo, la incerteza propia de los sensores y las perturbaciones involucradas en la medición [8, 83].

De aquí en adelante, $X_N = \big\{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[N]\big\}$ para $n = 1, 2, \ldots, N$ es el conjuntos de vectores de estados hasta el instante N inclusive, donde N es la cantidad de muestras. Luego, $X_n = \big\{\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[n]\big\}$ es el conjunto de vectores de estados hasta el instante n inclusive. De forma análoga, se definen el conjunto de observaciones $Z_N = \big\{\mathbf{z}[1], \mathbf{z}[2], \ldots, \mathbf{z}[N]\big\}$ y el conjunto de entradas $U_N = \big\{\mathbf{u}[1], \mathbf{u}[2], \ldots, \mathbf{u}[N]\big\}$.

5.2.1. Modelos lineales

Desde hace un tiempo, los modelos en espacio de estados lineales han cobrado una gran relevancia. Esto se debe, por un lado, a que su estructura permite representar modelos de importancia en ingeniería y en estadística (por ejemplo los modelos autorregresivos, de media móvil y la combinación de estos) y, por otro lado, los métodos en espacio de estados permiten utilizar estos modelos en aplicaciones reales de forma relativamente simple [35, 51, 76].

Los modelos en espacio de estados lineales se definen a partir del siguiente sistema de ecuaciones en diferencias estocásticas [8, 86]:

$$\mathbf{x}[n+1] = \mathbf{A}[n]\,\mathbf{x}[n] + \mathbf{f}[n]\,\mathbf{u}[n] + \mathbf{B}[n]\,\mathbf{w}[n], \tag{5.1a}$$

$$\mathbf{z}[n] = \mathbf{H}[n]\,\mathbf{x}[n] + \mathbf{v}[n], \tag{5.1b}$$

donde $\mathbf{A}[n] \in \mathbb{R}^{p\times p}$, $\mathbf{B}[n] \in \mathbb{R}^{p\times q}$, $\mathbf{f}[n] \in \mathbb{R}^{p\times u}$ y $\mathbf{H}[n] \in \mathbb{R}^{r\times p}$ son las matrices de transición de estados, de error, de excitación y de observación, respectivamente. A

su vez, los vectores $\mathbf{w}[n] \in \mathbb{R}^r$ y $\mathbf{v}[n] \in \mathbb{R}^q$ representan los errores de estados y de observación, respectivamente.

En la definición anterior, la Ec. (5.1a) gobierna la transición entre los estados internos del sistema, es decir, establece los estados sucesivos partiendo del vector de estados iniciales $\mathbf{x}[0]$. Por ello, se la denomina *ecuación de transición de estados*. Más adelante en este capítulo profundizaremos respecto a $\mathbf{x}[0]$. Por otro lado, la Ec. (5.1b) determina el valor de las observaciones en cada instante, es decir, la observación correspondiente a cada estado. Recibe el nombre de *ecuación de observación*.

Al sistema (5.1), se le agregan las siguientes hipótesis adicionales con respecto a los errores de estados y de observación [86]:

1. Cada uno de los errores tiene media cero:

$$\mathcal{E}\left\{\mathbf{w}[n]\right\} = \mathbf{0} \;\text{ y }\; \mathcal{E}\left\{\mathbf{v}[n]\right\} = \mathbf{0}, \;\; \forall n.$$

2. Los errores son procesos estocásticos *blancos*:

$$\mathcal{E}\left\{\mathbf{w}[n]\,\mathbf{w}[m]^T\right\} = \mathbf{Q}[n]\,\delta[n-m] \;\text{ y }\; \mathcal{E}\left\{\mathbf{v}[n]\,\mathbf{v}[m]^T\right\} = \mathbf{R}[n]\,\delta[n-m], \;\; \forall n, m$$

   donde $\mathbf{Q}[n] \in \mathbb{R}^{q\times q}$ y $\mathbf{R}[n] \in \mathbb{R}^{r\times r}$ son matrices de covarianza simétricas y definidas positivas para todo n, y δ es la función delta de *Kronecker*.

3. Son mutuamente independientes:

$$p(\mathbf{w}[n],\, \mathbf{v}[m]) = p(\mathbf{w}[n])\, p(\mathbf{v}[m]), \;\; \forall n, m.$$

Es importante destacar que los vectores de estados sucesivos forman un proceso de *Markov* de primer orden. Matemáticamente:

$$p\Big(\mathbf{x}[n]\Big|\mathbf{x}[0],\, \mathbf{x}[1],\, \mathbf{x}[2], \ldots,\, \mathbf{x}[n-1]\Big) = p\Big(\mathbf{x}[n]\Big|\mathbf{x}[n-1]\Big). \tag{5.2}$$

Esta propiedad es consecuencia de la causalidad del sistema (5.1) y de que el error de estado es un proceso estocástico *blanco* [8, 51]. Por ello, del conjunto de vectores de estados $\Big\{\mathbf{x}[0],\, \mathbf{x}[1],\, \mathbf{x}[2], \ldots,\, \mathbf{x}[n-1]\Big\}$, sólo $\mathbf{x}[n-1]$ acarrea información relevante para determinar $\mathbf{x}[n]$. Por su parte, las observaciones no cumplen con esta propiedad.

Modelos lineales y gaussianos

Los modelos en espacio de estados lineales y gaussianos (MEEG) son un caso particular de los definidos anteriormente, a los que se agrega una hipótesis adicional respecto a los errores de estados y de observación. Se considera que ambos errores son procesos estocásticos gaussianos. Así, los MEEG se definen matemáticamente de la siguiente forma [8]:

$$\mathbf{x}[n+1] = \mathbf{A}[n]\,\mathbf{x}[n] + \mathbf{f}[n]\,\mathbf{u}[n] + \mathbf{B}[n]\,\mathbf{w}[n], \qquad \mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}[n]), \tag{5.3a}$$

$$\mathbf{z}[n] = \mathbf{H}[n]\,\mathbf{x}[n] + \mathbf{v}[n], \qquad \mathbf{v}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{R}[n]), \tag{5.3b}$$

donde todos los parámetros en la definición fueron presentados anteriormente.

Teniendo en cuenta ahora el modelo (5.3), se desprende que los vectores de estados $\mathbf{x}[n]$ y las observaciones $\mathbf{z}[n]$ se comportan como procesos estocásticos gaussianos. Esto será cierto siempre y cuando $\mathbf{x}[0]$ posea un comportamiento gaussiano. Asimismo, los vectores de estados preservan su dinámica de *Markov* de primer orden.

Los MEEG son sumamente empleados en la práctica, ya que ofrecen dos ventajas importantes. En primer lugar, por ser $\mathbf{x}[n]$ un proceso aleatorio gaussiano, su función densidad de probabilidad queda completamente determinada por el valor medio y la matriz de covarianza [51, 83]. Es decir, estimando estos dos parámetros se obtiene toda la información estadística de los vectores de estados. En segundo lugar, corresponden a unos de los pocos casos para los cuales existen procedimientos numéricos exactos y razonablemente eficientes para estimar la media y la matriz de covarianza a partir de las observaciones [30].

Todos los aportes realizados en la presente tesis de doctorado se basan en el empleo de MEEG. Por esta razón, los métodos y algoritmos descriptos en este capítulo corresponderán a esta clase de modelos.

5.2.2. Modelos no lineales

Antes de pasar a los métodos en espacio de estados, describiremos brevemente los modelos en espacio de estados no lineales. Si bien existe una infinidad de modelos de esta naturaleza, nos centraremos en la familia de modelos en espacio de estados no lineales definidos del siguiente modo [30, 32, 83]:

$$\mathbf{x}[n+1] = a\Big(n,\, \mathbf{x}[n]\Big) + f\Big(n,\, \mathbf{u}[n]\Big) + b\Big(n,\, \mathbf{x}[n]\Big)\, \mathbf{w}[n], \tag{5.4a}$$

$$\mathbf{z}[n] = h\Big(n,\, \mathbf{x}[n]\Big) + \mathbf{v}[n], \tag{5.4b}$$

donde $a(\cdot)$, $f(\cdot)$, $b(\cdot)$ y $h(\cdot)$ son funciones vectoriales, de dimensiones apropiadas, no lineales y diferenciables.

La familia de modelos no lineales considerada puede verse como una extensión de los modelos presentados en la sección anterior. Por este motivo, se han propuesto diferentes métodos en espacio de estados para modelos no lineales, desarrollados a partir de linealizar el sistema (5.4) sobre una trayectoria suave en el espacio de los vectores de estados [30]. Para esto, se trabaja con aproximaciones en serie de Taylor de primer orden. Sin embargo, esta metodología presenta serias limitaciones, lo que ha impulsado el desarrollo de nuevas técnicas para lidiar con ésta y otras familia de modelos no lineales [30, 32].

5.3. Estimación óptima de la información

Introduciremos ahora diferentes métodos en espacio de estados, que permiten estimar la información relevante de un proceso o una señal bajo estudio. En todos ellos, se considera que existe un MEEG, ver Ec. (5.3), que describe matemáticamente el fenómeno bajo estudio. Tomando en cuenta la naturaleza de la información empleada en la estimación, los métodos se pueden clasificar en tres categorías [8, 51]:

- Filtrado: proceso en el que se estima un parámetro correspondiente a un instante determinado, utilizando la información disponibles hasta ese mismo instante inclusive. No debe confundirse con el concepto general de filtrado estocástico.

- Predicción: proceso en el que se considera la información disponible hasta un instante determinado, para estimar un parámetro correspondiente a un tiempo futuro. Comúnmente, se aplica la predicción un paso hacia adelante.
- Suavizado: proceso que permite obtener un parámetro correspondiente a un instante determinado, involucrando información pasada, de ese mismo instante y futura en la estimación.

Los métodos de filtrado y predicción son útiles en aplicaciones en tiempo real. Debido a que el suavizado emplea información futura, es inevitable que exista un retardo temporal en el proceso de estimación. Aun así, el suavizado es preferible en aplicaciones que permitan pequeños retardos o que involucren un procesamiento fuera de línea, ya que se toma en cuenta una mayor cantidad de información en la estimación.

5.3.1. Filtrado de los vectores de estados

Describiremos ahora el filtrado de los vectores de estados, llamado comúnmente filtrado de Kalman. En particular, nos centraremos en su versión *contemporánea* [51]. El propósito de este método consiste en estimar $\mathbf{x}[n]$ de forma óptima, empleando la información disponible hasta el instante n inclusive. Es decir, permite calcular el vector de estados *filtrado* $\hat{\mathbf{x}}[n|n] = \mathcal{E}\left\{\mathbf{x}[n]\middle|Z_n\right\}$ y su respectiva matriz de covarianza $\hat{\mathbf{P}}[n|n] = \mathrm{Var}\left\{\mathbf{x}[n]\middle|Z_n\right\}$, para cada instante $n = 1, 2, \ldots, N$. Este método es óptimo en el sentido que $\hat{\mathbf{x}}[n|n]$ es el estimador lineal no sesgado que minimiza al mismo tiempo la varianza y el error cuadrático medio de la estimación [8].

Actualmente, existen diferentes variantes del filtrado de Kalman. Todas ellas consisten en un proceso iterativo en valores crecientes de la variable n. Su estructura sencilla y su robustez lo convierten en un método sumamente útil en aplicaciones en tiempo real [8, 37]. En este documento consideraremos el algoritmo de filtrado de Kalman para MEEG presentado en [30, 51]. Se supone que el vector de estados iniciales sigue el modelo $\mathbf{x}[0] \sim \mathcal{N}\big(\mathbf{x}_0, \mathbf{P}_0\big)$, por lo que la etapa de inicialización requiere una estimación adecuada de los parámetros $\mathbf{x}_0$ y $\mathbf{P}_0$. A partir de estos parámetros, se aplica iterativamente el filtrado para los instantes $n = 1, 2, \ldots, N$.

En cada iteración, el filtrado se lleva a cabo en tres etapas: predicción, inferencia y corrección [30, 51]. Supongamos que se calcula la información para el instante n. En primer lugar, se predice $\mathbf{x}[n]$ utilizando la información disponible hasta el instante $n-1$. Se obtienen como resultado las estimaciones *a priori* del vector de estados $\hat{\mathbf{x}}[n|n-1] = \mathcal{E}\left\{\mathbf{x}[n]\middle|Z_{n-1}\right\}$ y de su matriz de covarianza $\hat{\mathbf{P}}[n|n-1] = \mathrm{Var}\left\{\mathbf{x}[n]\middle|Z_{n-1}\right\}$. A continuación, se toma $\mathbf{z}[n]$, la observación en el instante n, y se la emplea para inferir el error cometido en la estimación *a priori*. Esto permite corregir la estimación de forma tal de minimizar el error. Se obtienen así las estimaciones *a posteriori*, o estimaciones *filtradas*, del vector de estados $\hat{\mathbf{x}}[n|n]$ y de su matriz de covarianza $\hat{\mathbf{P}}[n|n]$. Este proceso se repite iterativamente para los instantes sucesivos.

En la Tab. 5.1, presentamos las ecuaciones que constituyen el algoritmo de filtrado de los vectores de estados. El vector $\tilde{\mathbf{y}}[n]$ es la inferencia, o error de predicción un paso hacia adelante, y $\tilde{\mathbf{F}}[n]$ es su matriz de covarianza. Estos dos parámetros sirven para analizar la bondad de la representación con el MEEG y de la estimación de la información [75]. La matriz $\tilde{\mathbf{K}}[n]$ se denomina ganancia de Kalman y determina el peso relativo del MEEG y de la observación $\mathbf{z}[n]$ en la etapa de corrección [30, 35].

Tabla 5.1: Ecuaciones que constituyen el algoritmo de filtrado de los vectores de estados, llamado también filtrado de Kalman.

Predicción:

1. $\hat{\mathbf{x}}[n|n-1] = \mathbf{A}[n-1]\,\hat{\mathbf{x}}[n-1|n-1] + \mathbf{f}[n-1]\,\mathbf{u}[n-1]$
2. $\hat{\mathbf{P}}[n|n-1] = \mathbf{A}[n-1]\,\hat{\mathbf{P}}[n-1|n-1]\,\mathbf{A}[n-1]^T + \mathbf{B}[n-1]\,\mathbf{Q}[n-1]\,\mathbf{B}[n-1]^T$

Inferencia:

3. $\tilde{\mathbf{y}}[n] = \mathbf{z}[n] - \mathbf{H}[n]\,\hat{\mathbf{x}}[n|n-1]$
4. $\tilde{\mathbf{F}}[n] = \mathbf{H}[n]\,\hat{\mathbf{P}}[n|n-1]\,\mathbf{H}[n]^T + \mathbf{R}[n]$
5. $\tilde{\mathbf{K}}[n] = \hat{\mathbf{P}}[n|n-1]\,\mathbf{H}[n]^T\,\tilde{\mathbf{F}}[n]^{-1}$

Corrección:

6. $\hat{\mathbf{x}}[n|n] = \hat{\mathbf{x}}[n|n-1] + \tilde{\mathbf{K}}[n]\,\tilde{\mathbf{y}}[n]$
7. $\hat{\mathbf{P}}[n|n] = \left(\mathbf{I}_p - \tilde{\mathbf{K}}[n]\,\mathbf{H}[n]\right)\mathbf{P}[n|n-1]$

Inicialización:

$\hat{\mathbf{x}}[0|0] = \mathbf{x}_0 \quad$ y $\quad \hat{\mathbf{P}}[0|0] = \mathbf{P}_0$

Es importante destacar que el filtrado de Kalman es exacto sólo para los MEEG. Esto es, permite calcular los valores $\mathcal{E}\left\{\mathbf{x}[n]\middle|Z_n\right\}$ y $\mathrm{Var}\left\{\mathbf{x}[n]\middle|Z_n\right\}$ que determinan por completo la función $p(\mathbf{x}[n]|Z_n)$ [37, 51]. Se demuestra que, para los modelos lineales que no satisfacen la hipótesis de gaussianidad de los errores de estados y de observación, este método produce las estimaciones que minimizan el error cuadrático medio en la estimación, aunque ya no minimizan la varianza. Por otro lado, estas estimaciones pueden resultar sesgadas, a la vez que aportan información estadística incompleta [8, 32].

5.3.2. Suavizado de los vectores de estados

El suavizado de los vectores de estados, también llamado suavizado de Kalman, comprende el cálculo de las estimaciones óptimas de los vectores de estados utilizando la información completa disponible en las observaciones. Sea Z_N el conjunto de N observaciones. Este método consiste en calcular el vector de estados *suavizado* $\hat{\mathbf{x}}[n|N] = \mathcal{E}\left\{\mathbf{x}[n]\middle|Z_N\right\}$ y su matriz de covarianza $\hat{\mathbf{P}}[n|N] = \mathrm{Var}\left\{\mathbf{x}[n]\middle|Z_N\right\}$, para cada $n = 1, 2, \ldots, N$ [51]. Se puede apreciar que el suavizado es un método no causal, ya que agrega información futura en el proceso de estimación.

Existen diferentes alternativas para realizar el suavizado de los vectores de estados. Para mayor información respecto a las diferentes versiones referirse a [8, Cap. 7]. Aquí, nos concentraremos en el suavizado para *intervalos fijos*, que se lleva a cabo mediante un proceso en dos etapas [51]. La primera de ellas coincide con el filtrado de Kalman explicado anteriormente, mientras que en la segunda se realiza el suavizado propiamente dicho. Es importante señalar que esta estrategia es aplicable únicamente en situaciones donde se realice un procesamiento fuera de línea.

Tabla 5.2: Ecuaciones que constituyen el algoritmo de suavizado de los vectores de estados, llamado también suavizado de Kalman.

Suavizado:

1. $\mathbf{L}[n] = \mathbf{A}[n]\left(\mathbf{I}_p - \tilde{\mathbf{K}}[n]\,\mathbf{H}[n]\right)$
2. $\mathbf{r}[n-1] = \mathbf{H}[n]^T\,\tilde{\mathbf{F}}[n]^{-1}\,\tilde{\mathbf{y}}[n] + \mathbf{L}[n]^T\,\mathbf{r}[n]$
3. $\mathbf{N}[n-1] = \mathbf{H}[n]^T\,\tilde{\mathbf{F}}[n]^{-1}\,\mathbf{H}[n] + \mathbf{L}[n]^T\,\mathbf{N}[n]\,\mathbf{L}[n]$
4. $\hat{\mathbf{x}}[n|N] = \hat{\mathbf{x}}[n|n-1] + \hat{\mathbf{P}}[n|n-1]\,\mathbf{r}[n-1]$
5. $\hat{\mathbf{P}}[n|N] = \left(\mathbf{I}_p - \hat{\mathbf{P}}[n|n-1]\,\mathbf{N}[n-1]\right)\hat{\mathbf{P}}[n|n-1]$

Inicialización:

$\mathbf{r}[N] = \mathbf{0}$ y $\mathbf{N}[N] = \mathbf{0}$

En los trabajos realizados en esta tesis de doctorado consideramos el algoritmo de suavizado de Kalman para MEEG propuesto en [51]. Este método también es un proceso iterativo, con la diferencia de que se calculan las estimaciones *suavizadas* $\hat{\mathbf{x}}[n|N]$ y $\hat{\mathbf{P}}[n|N]$ *hacia atrás* respecto a la variable n, considerando $n = N, N-1, \ldots, 1$. En la Tab. 5.2, presentamos las ecuaciones que constituyen el suavizado de Kalman. El algoritmo se inicializa considerando $\mathbf{r}[N] = \mathbf{0}$ y $\mathbf{N}[N] = \mathbf{0}$.

En general, el suavizado de Kalman permite obtener estimaciones de los vectores de estados más precisas, con una menor matriz de covarianza, en comparación con el filtrado. Sin embargo, esta mejora trae aparejada un aumento considerable en el costo computacional, debido a los algoritmos *hacia adelante* (Tab. 5.1) y *hacia atrás* (Tab. 5.2) necesarios en el suavizado. Se desprende, entonces, que existe una relación de compromiso entre el aumento en el costo computacional y la mejora en la precisión de las estimaciones.

Suavizado del vector de estados iniciales

El suavizado de los vectores de estados permite, a su vez, generar estimaciones del vector de estados iniciales empleando la información completa disponible en el conjunto de observaciones Z_N. Se obtiene así el vector de estados iniciales *suavizado* $\hat{\mathbf{x}}[0|N] = \mathcal{E}\left\{\mathbf{x}[0]\middle|Z_N\right\}$, junto con su matriz de covarianza $\hat{\mathbf{P}}[0|N] = \mathrm{Var}\left\{\mathbf{x}[0]\middle|Z_N\right\}$ correspondiente. Esto se implementa una vez finalizado el suavizado de Kalman, a partir de las siguiente expresiones:

$$\hat{\mathbf{x}}[0|N] = \mathbf{x}_0 + \mathbf{P}_0\,\mathbf{A}[0]^T\,\mathbf{r}[0], \tag{5.5}$$

$$\hat{\mathbf{P}}[0|N] = \left(\mathbf{I}_p - \mathbf{P}_0\,\mathbf{A}[0]^T\,\mathbf{N}[0]\,\mathbf{A}[0]\right)\mathbf{P}_0, \tag{5.6}$$

donde $\mathbf{x}_0$ y $\mathbf{P}_0$ son los parámetros del modelo estocástico del vector de estados iniciales, presentado en la Sec. 5.3.1.

5.3.3. Otras estimaciones suavizadas

A continuación, describiremos otras estimaciones suavizadas que resultan interesantes tanto por la propia información que acarrean como por su utilidad en la

Tabla 5.3: Ecuaciones que constituyen el algoritmo para el suavizado de las perturbaciones.

Suavizado de las perturbaciones:

1. $\mathbf{d}[n] = \tilde{\mathbf{F}}[n]^{-1}\,\tilde{\mathbf{y}}[n] - \left(\mathbf{A}[n]\,\tilde{\mathbf{K}}[n]\right)^T \mathbf{r}[n]$
2. $\mathbf{D}[n] = \tilde{\mathbf{F}}[n]^{-1} + \left(\mathbf{A}[n]\,\tilde{\mathbf{K}}[n]\right)^T \mathbf{N}[n]\left(\mathbf{A}[n]\,\tilde{\mathbf{K}}[n]\right)$
3. $\hat{\mathbf{v}}[n|N] = \mathbf{R}[n]\,\mathbf{d}[n]$
4. $\hat{\mathbf{P}}_{\mathrm{v}}[n|N] = \left(\mathbf{I}_r - \mathbf{R}[n]\,\mathbf{D}[n]\right)\,\mathbf{R}[n]$
5. $\hat{\mathbf{w}}[n-1|N] = \mathbf{Q}[n-1]\,\mathbf{B}[n-1]^T\,\mathbf{r}[n-1]$
6. $\hat{\mathbf{P}}_{\mathrm{w}}[n-1|N] = \left(\mathbf{I}_q - \mathbf{Q}[n-1]\,\mathbf{B}[n-1]^T\,\mathbf{N}[n-1]\,\mathbf{B}[n-1]\right)\,\mathbf{Q}[n-1]$

Inicialización:

$\mathbf{r}[N] = \mathbf{0}$ y $\mathbf{N}[N] = \mathbf{0}$

verificación de la calidad de la representación y en la estimación de los parámetros de los MEEG [43, 75, 90].

Suavizado de las perturbaciones

Este método hace posible obtener estimaciones de los errores de estados y de observación de un MEEG empleando la información completa disponible en el conjunto de observaciones Z_N. El proceso de suavizado de las perturbaciones consiste en el cálculo de los *errores suavizados* de estados $\hat{\mathbf{w}}[n-1|N] = \mathcal{E}\left\{\mathbf{w}[n-1]\middle|Z_N\right\}$ y de observación $\hat{\mathbf{v}}[n|N] = \mathcal{E}\left\{\mathbf{v}[n]\middle|Z_N\right\}$ junto con sus matrices de covarianza correspondientes $\hat{\mathbf{P}}_{\mathrm{w}}[n-1|N] = \mathrm{Var}\left\{\mathbf{w}[n-1]\middle|Z_n\right\}$ y $\hat{\mathbf{P}}_{\mathrm{v}}[n|N] = \mathrm{Var}\left\{\mathbf{v}[n]\middle|Z_n\right\}$, para los instantes $n = 1, 2, \ldots, N$. Estos parámetros han demostrado ser muy útiles para detectar observaciones atípicas y cambios en la dinámica de un proceso, así como para evaluar el desempeño de la representación basada en MEEG [75, 90].

Al igual que en el caso anterior, el suavizado de las perturbaciones se lleva a cabo en dos etapas. En primer lugar, se aplica el filtrado de Kalman (Tab. 5.1) y, seguidamente, el suavizado de las perturbaciones propiamente dicho. Este último, es un método iterativo *hacia atrás* respecto a la variable n, considerando $n = N, N-1, \ldots, 1$. Podemos apreciar que es un método no causal. En la Tab. 5.3 presentamos las ecuaciones que constituyen el algoritmo para el suavizado de las perturbaciones, de acuerdo a [51]. Los factores $\mathbf{r}$ y $\mathbf{N}$ coinciden con los expuesto en la Tab. 5.2. Es importante destacar que el algoritmo debe inicializarse considerando $\mathbf{r}[N] = \mathbf{0}$ y $\mathbf{N}[N] = \mathbf{0}$, al igual que en el suavizado de los vectores de estados.

Estimación suavizada de la correlación

Es posible generar estimaciones suavizadas de las matrices de correlación y de covarianza entre los diferentes elementos que forman un MEEG, contemplando diferentes retardos temporales. Para mayor información, el lector puede referirse a [51, Sec. 4.5]. Presentaremos aquí algunas de las que resultan de importancia para lo que sigue en este documento.

Tomando en cuenta el suavizado de Kalman, ver Tab. 5.2, se calcula la siguiente estimación suavizada de la matriz de autocorrelación de $\mathbf{x}[n]$:

$$\hat{\mathbf{C}}[n|N] = \mathrm{Cor}\left\{\mathbf{x}[n]\middle|Z_N\right\} = \hat{\mathbf{P}}[n|N] + \hat{\mathbf{x}}[n|N]\,\hat{\mathbf{x}}[n|N]^T. \tag{5.7}$$

Para el caso $n = 0$, la matriz de autocorrelación suavizada del vector de estados iniciales se define de la siguiente forma:

$$\hat{\mathbf{C}}[0|N] = \mathrm{Cor}\left\{\mathbf{x}[0]\middle|Z_N\right\} = \hat{\mathbf{P}}[0|N] + \hat{\mathbf{x}}[0|N]\,\hat{\mathbf{x}}[0|N]^T. \tag{5.8}$$

Nos concentraremos ahora en $\hat{\mathbf{C}}_{n-1,n}[n|N] = \mathrm{Cor}\left\{\mathbf{x}[n-1],\, \mathbf{x}[n]\middle|Z_N\right\}$, es decir, la matriz de correlación cruzada suavizada entre los vector de estados $\mathbf{x}[n-1]$ y $\mathbf{x}[n]$. Para su estimación se aplica la siguiente expresión:

$$\begin{aligned}\hat{\mathbf{C}}_{n-1,n}[n|N] = \hat{\mathbf{P}}[n-1|n-2]\,\mathbf{L}[n-1]^T\,\left(\mathbf{I}_p - \mathbf{N}[n-1]\,\hat{\mathbf{P}}[n|n-1]\right) \\ + \hat{\mathbf{x}}[n-1|N]\,\hat{\mathbf{x}}[n|N]^T \quad \text{para} \quad n = 2,\, 3,\, \ldots,\, N.\end{aligned} \tag{5.9}$$

Se desprende entonces que $\hat{\mathbf{C}}_{n,n-1}[n|N] = \mathrm{Cor}\left\{\mathbf{x}[n],\, \mathbf{x}[n-1]\middle|Z_N\right\} = \hat{\mathbf{C}}_{n-1,n}[n|N]^T$. Para el caso $n - 1 = 0$, se tiene la matriz de correlación cruzada suavizada entre $\mathbf{x}[0]$ y $\mathbf{x}[1]$. La estimación se realiza de la siguiente forma:

$$\hat{\mathbf{C}}_{n-1,n}[1|N] = \mathbf{P}_0\,\mathbf{A}[0]^T\,\left(\mathbf{I}_p - \mathbf{N}[0]\,\hat{\mathbf{P}}[1|0]\right) + \hat{\mathbf{x}}[0|N]\,\hat{\mathbf{x}}[1|N]^T. \tag{5.10}$$

Se demuestra que $\hat{\mathbf{C}}_{n,n-1}[1|N] = \hat{\mathbf{C}}_{n-1,n}[1|N]^T$.

Por otro lado, la matriz de autocorrelación suavizada para el error de observación $\mathbf{v}[n]$ se obtiene a partir de la siguiente expresión:

$$\hat{\mathbf{C}}_{\mathrm{v}}[n|N] = \mathrm{Cor}\left\{\mathbf{v}[n]\middle|Z_N\right\} = \hat{\mathbf{P}}_{\mathrm{v}}[n|N] + \hat{\mathbf{v}}[n|N]\,\hat{\mathbf{v}}[n|N]^T. \tag{5.11}$$

De forma análoga, la matriz de autocorrelación suavizada para el error de estados $\mathbf{w}[n]$ se calcula como sigue:

$$\hat{\mathbf{C}}_{\mathrm{w}}[n|N] = \mathrm{Cor}\left\{\mathbf{w}[n]\middle|Z_N\right\} = \hat{\mathbf{P}}_{\mathrm{w}}[n|N] + \hat{\mathbf{w}}[n|N]\,\hat{\mathbf{w}}[n|N]^T. \tag{5.12}$$

Todas las estimaciones descriptas aquí son importantes en la determinación de la función *verosimilitud* y cumplen un papel central en la estimación óptima de los parámetros de los MEEG [43, 90]. Volveremos respecto al uso de estas estimaciones más adelante en este capítulo y en los siguientes.

5.4. Vector de estados iniciales

Como vimos en la Sec. 5.3.1, el vector de estados iniciales cumple un rol importante en los métodos en espacio de estados. En esta sección, presentaremos dos modelos alternativos para su representación.

5.4.1. Inicialización estocástica

Sin duda alguna, la representación del vector de estados iniciales más utilizada en la práctica es el modelo estocástico [8, 30]:

$$\mathbf{x}[0] \sim \mathcal{N}(\mathbf{x}_0, \mathbf{P}_0), \tag{5.13}$$

donde $\mathbf{x}_0$ y $\mathbf{P}_0$ son conocidos y $\mathbf{P}_0$ es una matriz simétrica definida positiva. Introdujimos este modelo en la Sec. 5.3.1, al describir la inicialización para el filtrado de Kalman. Todos los métodos en espacio de estados presentados hasta aquí se basan en esta representación.

Podemos interpretar al modelo estocástico (5.13) de la siguiente forma. Resulta muy difícil, sino imposible, conocer con exactitud el estado inicial de un sistema. Sin embargo, es posible inferir esta información de manera aproximada. El modelo estocástico simula este último caso, donde $\mathbf{x}_0$ es la información aproximada del estado inicial y $\mathbf{P}_0$ cuantifica la incerteza o el error en esta aproximación.

Si bien este modelo es sencillo y fácil de aplicar, posee la gran desventaja de que requiere conocer *a priori* $\mathbf{x}_0$ y $\mathbf{P}_0$. Usualmente, estos valores se desconocen, de forma parcial o total, por lo que es necesario fijarlos arbitrariamente o estimarlos empleando la información del MEEG y el conjunto de observaciones. Por otro lado, puede suceder que se necesite estudiar un conjunto de señales o sistemas similares a partir de un mismo MEEG. En este caso, será necesario proponer o estimar cada condición inicial, lo que en la práctica se torna engorroso.

5.4.2. Inicialización difusa

Desde hace un tiempo, ha surgido una solución analítica a la problemática descripta en el párrafo anterior. Este método se denomina *inicialización difusa*, y ha cobrado relevancia por su fácil implementación y por resultar computacionalmente eficiente [87, 88]. Se basa en un modelo *difuso* para el vector de estados iniciales, determinado por la siguiente expresión [51]:

$$\mathbf{x}[0] = \mathbf{x}_0 \, + \, \mathbf{T}\,\delta + \mathbf{B}_0\,\mathbf{w}_0, \quad \mathbf{w}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_0), \tag{5.14}$$

donde $\mathbf{x}_0 \in \mathbb{R}^p$ es un parámetro conocido, y $\delta \in \mathbb{R}^s$ y $\mathbf{w}_0 \in \mathbb{R}^{p-s}$ son vectores aleatorios. $\mathbf{T} \in \mathbb{R}^{p\times s}$ y $\mathbf{B}_0 \in \mathbb{R}^{p\times(p-s)}$ son matrices de selección, es decir, sus columnas tomadas conjuntamente, concatenadas y ordenadas, forman la matriz $\mathbf{I}_p$ y satisfacen $\mathbf{B}_0^T\,\mathbf{T} = \mathbf{0}$. Se supone, además, que la matriz de covarianza $\mathbf{Q}_0 \in \mathbb{R}^{(p-s)\times(p-s)}$ es simétrica, definida positiva y conocida.

Podemos apreciar que el modelo difuso es una extensión de la representación estocástica (5.13). En esta idealización se toma como hipótesis principal que la falta de conocimiento en elementos específicos de $\mathbf{x}[0]$ está asociada a una incerteza extremadamente alta en la covarianza correspondiente. Así, la dimensión s indica la cantidad de estos elementos desconocidos, los cuales se consideran agrupados en el vector de *elementos difusos* δ, donde $\delta \sim \mathcal{N}(\mathbf{0}, \kappa\,\mathbf{I}_s)$ con $\kappa \to \infty$. Las matrices de selección $\mathbf{T}$ y $\mathbf{B}_0$ establecen que elementos de $\mathbf{x}[0]$ son estocásticos y difusos, respectivamente, suponiendo que no pueden ocurrir los dos estados al mismo tiempo. Para una descripción detallada de la inicialización difusa referirse a [51, 87, 88].

La versión modificada del filtrado de Kalman, de forma tal que considere el modelo (5.14) y la condición $\kappa \to \infty$, se denomina *filtrado inicial exacto* [88]. Si

bien la condición $\kappa \to \infty$ parece ser un inconveniente, la influencia de κ en las estimaciones desaparece luego de un número de iteraciones d, donde en la práctica $d \ll N$. Usualmente, al trabajar con observaciones unidimensionales ($r = 1$) se cumple $d = p$, donde p es la dimensión del vector de estados [87]. Así, los vectores de estados filtrados se calculan para $n = 1, 2, \ldots, d$ con el filtrado inicial exacto y para $n = d + 1$ en adelante aplicando el filtrado de Kalman (ver Tab. 5.1).

Por otro lado, recibe el nombre de *suavizado inicial exacto* la versión del suavizado de los vectores de estados adecuada a la inicialización difusa y a la condición $\kappa \to \infty$ [87]. De forma análoga al caso anterior, los vectores de estado suavizados se calculan, iterativamente hacia atrás, para $n = N, N-1, \ldots, d+1$ aplicando el suavizado de Kalman (ver Tab. 5.2) y para los d primeros instantes se emplea el suavizado inicial exacto. En esta tesis de doctorado consideramos la inicialización difusa y los algoritmos presentados en [51, 88].

5.5. Estimación de parámetros

La estructura de los MEEG depende de un conjunto de parámetros, ver Ec. (5.3). Podemos encontrar situaciones en las cuales estos parámetros se conocen previamente o se pueden deducir a partir del fenómeno bajo estudio [99]. En otros casos, es necesario estimarlos empleando el conjunto de observaciones Z_N y, si se conocen, los estímulos U_N como única información. Para ello, se han propuesto diferentes estrategias basadas en la inferencia estadística y en métodos de optimización [43, 51, 80]. En esta sección introduciremos las ideas fundamentales involucradas en la estimación de los parámetros de un MEEG.

De aquí en adelante, realizaremos algunas consideraciones, con el propósito de adecuar los desarrollos a las aplicaciones estudiadas en esta tesis de doctorado y, por otro lado, de simplificar las expresiones resultantes. En primer lugar, nuestros trabajos involucran únicamente señales unidimensionales ($r = 1$). Así, $Z_N = \Big\{z[1], z[2], \ldots, z[N]\Big\}$ es el conjunto de observaciones y $v[n]$ es el error de observación. En segundo lugar, consideraremos que los errores de estados y de observación son procesos gaussianos estacionarios, es decir, $v[n] \sim \mathcal{N}(0, \sigma^2)$, con $\sigma^2 > 0$, y $\mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, donde $\mathbf{Q}$ es una matriz simétrica y definida positiva. Por último, emplearemos la igualdad $\mathbf{Q} = \sigma^2 \mathring{\mathbf{Q}}$, donde nuevamente $\mathring{\mathbf{Q}}$ es una matriz simétrica y definida positiva.

5.5.1. Problema de optimización

Consideraremos, de aquí en adelante, que $\boldsymbol{\Theta} \in \mathcal{D}$ representa el conjunto de parámetros desconocidos de un MEEG, donde $\mathcal{D}$ es su dominio de definición. El objetivo aquí es generar una estimación $\hat{\boldsymbol{\Theta}}$ de los parámetros que resulte óptima en algún sentido. Para ello, nos concentraremos en resolver el siguiente problema de optimización:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta} \in \mathcal{D}}{\arg\max} \;\; \mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}, \tag{5.15}$$

donde $\mathcal{L}(\boldsymbol{\Theta})$ es la función *verosimilitud*[1]. En inferencia estadística, se considera que esta función aporta toda la información respecto al proceso estocástico que se puede

[1] En inglés, "likelihood function".

deducir de un conjunto de observaciones, suponiendo que el MEEG es adecuado [26, 30]. Esto último se conoce como el *principio de la verosimilitud*[2]. Luego, resolver el problema de optimización propuesto permite obtener el conjunto de parámetros $\hat{\boldsymbol{\Theta}}$ para el cual se maximiza la probabilidad de que el modelo estocástico considerado haya generado el conjunto de observaciones. Por ello, a $\hat{\boldsymbol{\Theta}}$ se lo conoce como la estimación de *máxima verosimilitud*.

Existe una gran variedad de métodos para resolver el problema (5.15). En nuestros trabajos, hemos utilizado los algoritmos numéricos guiados por gradiente, como por ejemplo el método de *Esperanza* y *Maximización* (EM), el método de *Newton* o los métodos *cuasi Newton* [43, 51, 80]. Por tratarse de un problema de maximización, se trabaja con métodos de gradiente ascendente, denominados también métodos de la dirección de máximo ascenso. Sin embargo, cambiando el signo de la función objetivo $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$ podemos transformar a (5.15) fácilmente en un problema de minimización.

5.5.2. Función objetivo

En el problema de optimización propuesto, la función objetivo es $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$, es decir, el valor esperado del logaritmo natural de la función verosimilitud. A continuación, describiremos como construir esta función.

Consideraremos aquí $p(X_N,\, Z_N;\, \boldsymbol{\Theta})$, es decir, la función densidad de probabilidad conjunta de los vectores de estados y de las observaciones que depende del conjunto de parámetros $\boldsymbol{\Theta}$. El operador valor esperado $\mathcal{E}\left\{\cdot\right\}$ en el problema (5.15) se define respecto a $p(X_N,\, Z_N;\, \boldsymbol{\Theta})$, de forma similar a como se procedió en [43, 80]. Luego, suponiendo conocidos X_N y Z_N, se define la función *verosimilitud*:

$$\mathcal{L}\Big(\boldsymbol{\Theta}\Big) = p(X_N,\, Z_N;\, \boldsymbol{\Theta}). \tag{5.16}$$

Recordemos que los vectores de estados y las observaciones son variables aleatorias, de acuerdo a lo expuesto en la Sec. 5.2.1. Tomando en cuenta las propiedades de una función de probabilidad conjunta de varias variables, obtenemos lo siguiente:

$$\begin{aligned} p\Big(X_N,\, Z_N;\, \boldsymbol{\Theta}\Big) &= p\Big(Z_N\Big|X_N;\, \boldsymbol{\Theta}\Big)\, p\Big(X_N;\, \boldsymbol{\Theta}\Big) \\ &= p\Big(\mathbf{x}[0];\, \boldsymbol{\Theta}\Big) \prod_{n=1}^{N} p\Big(\mathbf{x}[n]\Big|\mathbf{x}[n-1];\, \boldsymbol{\Theta}\Big)\, p\Big(\mathbf{z}[n]\Big|\mathbf{x}[n];\, \boldsymbol{\Theta}\Big). \end{aligned} \tag{5.17}$$

Para construir esta última expresión, consideramos el sistema de ecuaciones que rigen los MEEG, ver Ec. (5.3), y que los vectores de estados desarrollan una dinámica de *Markov* de primer orden, como se comentó anteriormente. Aplicando el logaritmo natural y el operador $\mathcal{E}\left\{\cdot\right\}$ a la Ec. (5.17) se obtiene la función objetivo. Existen dos expresiones alternativas para la función objetivo que resultan de interés para esta tesis de doctorado.

Expresión basada en los vectores de estados y en las observaciones

En esta estructura se pone de manifiesto, de forma explícita, la relación entre $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$ y cada uno de los parámetros del MEEG. De esta forma, permite estudiar como varía la función objetivo al modificar $\boldsymbol{\Theta}$.

[2] En inglés, "likelihood principle".

Aplicando el logaritmo natural en la Ec. (5.17), obtenemos la siguiente expresión:

$$\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\Theta}) = \ln \Bigg[& \frac{1}{(2\,\pi)^{p/2}\,|\mathbf{P}_0|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}[0]-\hat{\mathbf{x}}_0\right)^T \mathbf{P}_0^{-1}\left(\mathbf{x}[0]-\hat{\mathbf{x}}_0\right)\right) \Bigg] \\
& + \sum_{n=1}^{N} \ln \left[\frac{1}{(2\,\pi\,\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\,\sigma^2}\left(z[n]-\mathbf{H}[n]\,\mathbf{x}[n]\right)^2\right)\right] \\
& + \sum_{n=1}^{N} \ln \Bigg[\frac{1}{(2\,\pi\,\sigma^2)^{p/2}\,|\mathring{\mathbf{Q}}|^{1/2}} \exp\Bigg(-\frac{1}{2\,\sigma^2} \\
& \qquad \times \left(\mathbf{x}[n]-\mathbf{A}[n-1]\,\mathbf{x}[n-1]-\mathbf{f}[n-1]\,\mathbf{u}[n-1]\right)^T \\
& \qquad \times \mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T \\
& \qquad \times \left(\mathbf{x}[n]-\mathbf{A}[n-1]\,\mathbf{x}[n-1]-\mathbf{f}[n-1]\,\mathbf{u}[n-1]\right) \Bigg)\Bigg].
\end{aligned} \tag{5.18}$$

Esta última expresión surge de recordar que los errores de estados y de observación son procesos gaussianos estacionarios y de considerar el modelo estocástico de $\mathbf{x}[0]$. En la Ec. (5.18), los valores de las observaciones z y los estímulos $\mathbf{u}$ son conocidos ya que, por hipótesis, pueden medirse directamente del proceso. No ocurre lo mismo con los vectores de estados $\mathbf{x}$. En su lugar, consideraremos las estimaciones suavizadas descriptas en la Sec. 5.3.

Aplicando el operador $\mathcal{E}\{\cdot\}$ y trabajando convenientemente, se obtiene la siguiente expresión para la función objetivo, basada en las observaciones y en los vectores de estados suavizados:

$$\begin{aligned}
\mathcal{E}\{\ln \mathcal{L}(\boldsymbol{\Theta})\} = \tilde{c} & - \frac{1}{2}\ln\left(|\mathbf{P}_0|\right) - \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) \\
& + \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 + \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 \\
& - \frac{N\,(1+p)}{2}\,\ln(\sigma^2) - \frac{N}{2}\,\ln\left(|\mathring{\mathbf{Q}}|\right) \\
& - \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\left[z[n]^2 - 2\,z[n]\,\mathbf{H}[n]\,\hat{\mathbf{x}}[n|N] + \mathbf{H}[n]\,\hat{\mathbf{C}}[n|N]\,\mathbf{H}[n]^T\right] \\
& - \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\Big[\mathrm{Tr}\left(\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\hat{\mathbf{C}}[n|N]\right) \\
& \quad - \mathrm{Tr}\left(\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{A}[n-1]\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
& \quad - \hat{\mathbf{x}}[n|N]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{f}[n-1]\,\mathbf{u}[n-1] \\
& \quad - \mathrm{Tr}\left(\mathbf{A}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) \\
& \quad + \mathrm{Tr}\left(\mathbf{A}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{A}[n-1]\,\hat{\mathbf{C}}[n-1|N]\right) \\
& \quad + \hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{f}[n-1]\,\mathbf{u}[n-1] \\
& \quad - \mathbf{u}[n-1]^T\,\mathbf{f}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\hat{\mathbf{x}}[n|N] \\
& \quad + \mathbf{u}[n-1]^T\,\mathbf{f}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{A}[n-1]\,\hat{\mathbf{x}}[n-1|N] \\
& \quad + \mathbf{u}[n-1]^T\,\mathbf{f}[n-1]^T\,\mathbf{B}[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{B}[n-1]^T\,\mathbf{f}[n-1]\,\mathbf{u}[n-1]\Big],
\end{aligned} \tag{5.19}$$

con $\tilde{c} \in \mathbb{R}$ constante. Sólo nos interesan las variaciones relativas de la función objetivo con respecto a las modificaciones en $\boldsymbol{\Theta}$, por lo que no es necesario conocer el valor de

$\tilde{c}$ en la expresión anterior. Podemos apreciar en la Ec. (5.19) cómo se relacionan los parámetros del MEEG y las diferentes estimaciones suavizadas descriptas en la Sec. 5.3. Esto permite el cálculo de los parámetros aplicando métodos de optimización y propiedades del cálculo vectorial y matricial.

Expresión basada en los errores

Existen aplicaciones donde se conoce por completo la información del MEEG, salvo los parámetros asociados a los errores de estados y de observación. En estos casos, podemos simplificar considerablemente la expresión de la función $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$. Esto resulta conveniente en situaciones, como las consideradas en esta tesis de doctorado, donde sólo se necesita calcular los parámetros σ^2 y $\mathring{\mathbf{Q}}$.

Recordemos una vez más que, de acuerdo a las hipótesis de los MEEG, los errores de estados y de observación son procesos estacionarios gaussianos. A su vez, en la Sec. 5.3.3 presentamos las estimaciones suavizadas de ambos errores, junto con las correspondientes matrices de covarianza y de autocorrelación. Con toda esta información, podemos generar una expresión alternativa de la función $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$.

Considerando nuevamente la Ec. (5.17) junto con el sistema de ecuaciones que caracteriza a un MEEG, podemos obtener la siguiente expresión [51]:

$$\begin{aligned}\ln \mathcal{L}(\boldsymbol{\Theta}) = &\sum_{n=1}^{N} \ln\left[\frac{1}{(2\,\pi\,\sigma^2)^{1/2}}\,\exp\left(-\frac{1}{2\,\sigma^2}\,v[n]^2\right)\right] \\ &+ \sum_{n=1}^{N} \ln\left[\frac{1}{(2\,\pi\,\sigma^2)^{q/2}\,|\mathring{\mathbf{Q}}|^{1/2}}\,\exp\left(-\frac{1}{2\,\sigma^2}\,\mathbf{w}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{w}[n-1]\right)\right].\end{aligned} \tag{5.20}$$

Podemos observar que la expresión resultante depende de los errores de estados $\mathbf{w}$ y de observación v. Sin embargo, no se conocen estos elementos. En su lugar, emplearemos las estimaciones suavizadas de los errores, las cuales se obtienen a partir del algoritmo expuesto en la Sec. 5.3.3.

Aplicando el operador $\mathcal{E}\{\cdot\}$ a la Ec. (5.20), se obtiene la expresión de la función objetivo basada en las estimaciones suavizadas de los errores. De esta forma, arribamos a la siguiente expresión:

$$\begin{aligned}\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\} = \tilde{c} &- \frac{N\,(1+q)}{2}\,\ln(\sigma^2) - \frac{N}{2}\,\ln\left(\left|\mathring{\mathbf{Q}}\right|\right) \\ &- \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\hat{\mathrm{C}}_{\mathrm{v}}[n|N] - \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\right),\end{aligned} \tag{5.21}$$

con $\tilde{c} \in \mathbb{R}$ constante. Al igual que en el caso anterior, estamos interesados únicamente en los cambios relativos en la función objetivo. Por ello, podemos despreciar la constante $\tilde{c}$ en la Ec. (5.21). Por otro lado, podemos apreciar que la función $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$ depende principalmente de los parámetros σ^2 y $\mathring{\mathbf{Q}}^{-1}$.

Función *verosimilitud* difusa

En las expresiones de la función $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$ presentadas anteriormente se consideró el modelo estocástico para $\mathbf{x}[0]$. En el caso de trabajar con el modelo difuso, es necesario modificar adecuadamente la función objetivo. En [51, Sec. 7.2.2] se describe como llevar a cabo esta tarea. De forma análoga a lo que sucede con el filtrado y

el suavizado inicial exacto, es necesario corregir en $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$ las expresiones de los elementos correspondientes a las d primeras iteraciones. Una solución aproximada, más simple, consiste en modificar las expresiones (5.19) y (5.21) para que sólo consideren las estimaciones desde el instante $d+1$ en adelante.

5.5.3. Gradiente y estimación de los parámetros

En la sección anterior, obtuvimos dos expresiones alternativas para la función de costo $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$. Como cierre de este capítulo, ilustraremos cómo calcular el gradiente de esta función con respecto a los diferentes parámetros en $\boldsymbol{\Theta}$ y, luego, cómo emplear esta información para estimar los parámetros óptimos para un determinado MEEG. Recordemos que estamos interesados en obtener parámetros óptimos, de acuerdo al problema (5.15). En este sentido, el gradiente de la función objetivo indica la dirección de máximo crecimiento con respecto a los diferentes parámetros. Nos concentraremos aquí en las expresiones para σ^2, $\mathring{\mathbf{Q}}$, $\mathbf{x}_0$ y $\mathbf{P}_0$. Los restantes parámetros dependen fuertemente del MEEG considerado.

Varianza σ^2

Consideremos en primer lugar la expresión (5.21). Sólo tendremos en cuenta los términos que involucren a σ^2, ya que los restantes se anularán al tomar la derivada. Calculando la derivada parcial de $\mathcal{E}\left\{\ln \mathcal{L}\left(\boldsymbol{\Theta}\right)\right\}$ con respecto a la varianza σ^2, obtenemos la siguiente expresión:

$$
\begin{aligned}
\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\}}{\partial\, \sigma^2} &= \frac{\partial}{\partial\, \sigma^2}\left\{ -\frac{N\,(1+q)}{2}\,\ln(\sigma^2) - \frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\hat{\mathrm{C}}_{\mathrm{v}}[n|N] \right. \\
&\qquad \left. -\frac{1}{2\,\sigma^2}\sum_{n=1}^{N}\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\right)\right\} \\
&= -\frac{N\,(1+q)}{2}\,\frac{1}{\sigma^2} + \frac{1}{2\,(\sigma^2)^2}\sum_{n=1}^{N}\hat{\mathrm{C}}_{\mathrm{v}}[n|N] \\
&\qquad + \frac{1}{2\,(\sigma^2)^2}\sum_{n=1}^{N}\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\right) \\
&= -\frac{N\,(1+q)}{2}\,\frac{1}{\sigma^2} + \frac{1}{2\,(\sigma^2)^2}\sum_{n=1}^{N}\hat{\mathrm{C}}_{\mathrm{v}}[n|N] \\
&\qquad + \frac{1}{2\,(\sigma^2)^2}\,\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\sum_{n=1}^{N}\hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\right).
\end{aligned}
\tag{5.22}
$$

La expresión anterior se obtuvo aplicando las reglas usuales para el cálculo de derivadas de funciones escalares y del operador traza. Recordemos que por hipótesis $\sigma^2 > 0$. Seguidamente, desarrollaremos una regla para la estimación iterativa del valor óptimo de este parámetro. Igualando a cero la Ec. (5.22) y trabajando convenientemente, se desprende que:

$$
\hat{\sigma}^2_{\mathrm{est}} = \frac{1}{N\,(1+q)}\,\left[\bar{\mathrm{C}}^{tot}_{\mathrm{v}} + \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}_{\mathrm{aux}}\,\bar{\mathbf{C}}^{tot}_{\mathrm{w}}\right)\right], \tag{5.23}
$$

donde

$$\bar{\mathrm{C}}_{\mathrm{v}}^{tot} = \sum_{n=1}^{N} \hat{\mathrm{C}}_{\mathrm{v}}[n|N], \tag{5.24}$$

$$\bar{\mathbf{C}}_{\mathrm{w}}^{tot} = \sum_{n=1}^{N} \hat{\mathbf{C}}_{\mathrm{w}}[n-1|N], \tag{5.25}$$

y donde $\mathring{\mathbf{Q}}_{\mathrm{aux}}^{-1}$ es un valor auxiliar conocido de ese mismo parámetro. Por ejemplo, en EM se actualizan iterativamente los parámetros y, en ese caso, $\mathring{\mathbf{Q}}_{\mathrm{aux}}^{-1}$ toma el valor de la estimación de dicho parámetro correspondiente a la iteración anterior.

Matriz de covarianza $\mathring{\mathbf{Q}}$

Trabajando nuevamente con la expresión (5.21), obtendremos la derivada parcial de esta expresión con respecto a $\mathring{\mathbf{Q}}$. Nos concentraremos únicamente en los términos que involucren este parámetro, ya que los restantes se anularán al tomar la derivada. De esta forma, arribamos a la siguiente expresión:

$$\begin{aligned} \frac{\partial \mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\}}{\partial \mathring{\mathbf{Q}}} &= \frac{\partial}{\partial \mathring{\mathbf{Q}}} \left\{ -\frac{N}{2} \ln\left(|\mathring{\mathbf{Q}}|\right) - \frac{1}{2\,\sigma^2} \sum_{n=1}^{N} \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\, \hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\right) \right\} \\ &= -\frac{N}{2} (\mathring{\mathbf{Q}}^{-1})^T + \frac{1}{2\,\sigma^2} \left[\mathring{\mathbf{Q}}^{-1} \sum_{n=1}^{N} \hat{\mathbf{C}}_{\mathrm{w}}[n-1|N]\, \mathring{\mathbf{Q}}^{-1} \right]^T. \end{aligned} \tag{5.26}$$

Para hallar esta última expresión, supusimos que $\mathring{\mathbf{Q}}$ es una matriz no singular y aplicamos las reglas para la derivación de funciones con respecto a variables vectoriales y matriciales [66]. A continuación, desarrollaremos una regla para la estimación iterativa de este parámetro. Igualando a cero la Ec. (5.26), trasponiendo la expresión y, luego, multiplicando a la izquierda y a la derecha por $\mathring{\mathbf{Q}}$, obtenemos:

$$\mathring{\mathbf{Q}}_{\mathrm{est}} = \frac{1}{N\,\hat{\sigma}_{\mathrm{aux}}^2}\, \bar{\mathbf{C}}_{\mathrm{w}}^{tot}, \tag{5.27}$$

donde $\bar{\mathbf{C}}_{\mathrm{w}}^{tot}$ se precisó en la Ec. (5.25). Luego, se tiene que $\hat{\mathbf{Q}}_{\mathrm{est}} = \hat{\sigma}_{\mathrm{est}}^2\, \mathring{\mathbf{Q}}_{\mathrm{est}}$.

Vector de estados iniciales $\mathbf{x}_0$ y su matriz de covarianza $\mathbf{P}_0$

En el desarrollo de la Ec. (5.19), supusimos que el vector de estados iniciales sigue el modelo $\mathbf{x}[0] \sim \mathcal{N}(\mathbf{x}_0, \mathbf{P}_0)$. En esta sección, desarrollaremos las expresiones para calcular iterativamente los parámetros $\mathbf{x}_0$ y $\mathbf{P}_0$.

En primer lugar, nos concentraremos en $\mathbf{x}_0$. Partiendo de la expresión (5.19), tomamos la derivada parcial de esta expresión con respecto a $\mathbf{x}_0$. Sólo prestaremos atención a los términos que involucren este parámetro, ya que se anulan los restantes al derivar. Haciendo esto, arribamos a la siguiente expresión:

$$\begin{aligned} \frac{\partial \mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\}}{\partial \mathbf{x}_0} &= \frac{\partial}{\partial \mathbf{x}_0} \left\{ \frac{1}{2}\, \hat{\mathbf{x}}[0|N]^T\, \mathbf{P}_0^{-1}\, \mathbf{x}_0 + \frac{1}{2}\, \mathbf{x}_0^T\, \mathbf{P}_0^{-1}\, \hat{\mathbf{x}}[0|N] - \frac{1}{2}\, \mathbf{x}_0^T\, \mathbf{P}_0^{-1}\, \mathbf{x}_0 \right\} \\ &= \mathbf{P}_0^{-1}\, \hat{\mathbf{x}}[0|N] - \mathbf{P}_0^{-1}\, \mathbf{x}_0. \end{aligned} \tag{5.28}$$

Para ello, tomamos como hipótesis que $\mathbf{P}_0$ es no singular y aplicamos las reglas para la derivación de funciones con respecto a variables vectoriales y matriciales [66]. Igualando a cero lo anterior y multiplicando a la izquierda por $\mathbf{P}_0$, se obtiene:

$$\hat{\mathbf{x}}_{0\ \mathrm{est}} = \hat{\mathbf{x}}[0|N]. \tag{5.29}$$

Ahora obtendremos una regla similar para calcular $\mathbf{P}_0$. Partiendo nuevamente de la Ec. (5.19), calculamos la derivada parcial de $\mathcal{E}\left\{\ln \mathcal{L}\big(\boldsymbol{\Theta}\big)\right\}$ con respecto a $\mathbf{P}_0$. Sólo se debe prestar atención a los términos de la expresión que contengan a esta variable, ya que al derivar se anularán los restantes. Haciendo esto, obtenemos lo siguiente:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\}}{\partial \mathbf{P}_0} = \frac{\partial}{\partial \mathbf{P}_0}\Big\{ &-\frac{1}{2}\ln\left(|\mathbf{P}_0|\right) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) \\ &+\frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 + \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0\Big\}\end{aligned} \tag{5.30}$$

Es posible simplificar considerablemente la Ec. (5.30). Recordemos que $\hat{\mathbf{C}}[0|N] = \hat{\mathbf{P}}[0|N] + \hat{\mathbf{x}}[0|N]\,\hat{\mathbf{x}}[0|N]^T$. Reemplazando esta expresión y aplicando la regla de la traza, obtenemos lo siguiente:

$$\operatorname{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) = \operatorname{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{P}}[0|N]\right) + \hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N]. \tag{5.31}$$

Por otro lado, en la expresión (5.29) probamos que el vector de estados iniciales óptimo satisface la relación $\hat{\mathbf{x}}_0 = \hat{\mathbf{x}}[0|N]$. Considerando esto último y la relación (5.31), podemos reescribir la Ec. (5.30) de la siguiente forma:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}\big(\boldsymbol{\Theta}\big)\right\}}{\partial \mathbf{P}_0} &= \frac{\partial}{\partial \mathbf{P}_0}\Big\{ -\frac{1}{2}\ln\left(\left|\mathbf{P}_0\right|\right) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{P}}[0|N]\right) \\ &\qquad -\frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] + \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] \\ &\qquad +\frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N]\Big\} \\ &= \frac{\partial}{\partial \mathbf{P}_0}\left\{ -\frac{1}{2}\ln\left(\left|\mathbf{P}_0\right|\right) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{P}}[0|N]\right)\right\} \\ &= -\frac{1}{2}\left(\mathbf{P}_0^{-1}\right)^T + \frac{1}{2}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{P}}[0|N]\,\mathbf{P}_0^{-1}\right)^T.\end{aligned} \tag{5.32}$$

Para arribar a esto último, supusimos que $\mathbf{P}_0$ es una matriz no singular y aplicamos las reglas para la derivación de funciones con respecto a variables vectoriales y matriciales [66]. Igualando a cero la Ec. (5.32), transponiendo la expresión y multiplicando a la izquierda y a la derecha por $\mathbf{P}_0$, se obtiene:

$$\hat{\mathbf{P}}_{0\,\text{est}} = \hat{\mathbf{P}}[0|N]. \tag{5.33}$$

Los resultados alcanzados en (5.29) y en (5.33) cobran importancia, ya que podemos apreciar que los valores del vector de estados iniciales y de su matriz de covarianza que maximizan la función $\mathcal{E}\left\{\ln \mathcal{L}\big(\boldsymbol{\Theta}\big)\right\}$ son, precisamente, las estimaciones suavizadas $\hat{\mathbf{x}}[0|N]$ y $\hat{\mathbf{P}}[0|N]$. Luego, estos elementos presentan una doble interpretación. Por un lado, son estimaciones suavizadas y, por otro lado, son los valores que maximizan la función *verosimilitud* de acuerdo al problema (5.15).

5.6. Comentarios finales

Dedicamos este capítulo a describir los métodos en espacio de estados, enfocándonos en aquellos que serán imprescindibles para los desarrollos presentados más adelante.

Nuestra exposición tuvo como eje principal los modelos en espacio de estados lineales y gaussianos. A partir de ellos, presentamos algunos de los métodos existentes para estimar de forma adecuada la información relevante de un proceso o una señal, que no puede obtenerse por técnicas directas.

A su vez, introdujimos las ideas principales involucradas en la estimación de los parámetros de un modelo a partir de un conjunto de observaciones del fenómeno bajo estudio. Más adelante, aplicaremos estas ideas al estudio de señales biomédicas de la fonación.

Produjimos en este capítulo un compendio sobre los métodos específicos para los modelos en espacio de estados lineales y gaussianos, redactado de forma concisa y precisa. Este material es el resultado del estudio de las bibliografías disponibles en diversas ciencias relacionadas. A lo largo de este desarrollo, perseguimos en todo momento una perspectiva unificada. A su vez, abarcamos algunos tópicos frecuentemente omitidos en ingeniería, como por ejemplo las estimaciones suavizadas de las perturbaciones y de la correlación de los vectores de estados, la estrategia de inicialización difusa y las diferentes formas de la función verosimilitud.

Toda la información desarrollada hasta este punto genera un marco conceptual adecuado para la presentación de los últimos aportes de esta tesis de doctorado, en los dos próximos capítulos.

Capítulo 6

Análisis estructural basado en métodos en espacio de estados aplicado al modelado de series de períodos y de amplitudes

6.1. Introducción

En el Cap. 4, describimos las series de períodos (SP) y de amplitudes (SA), indicando de qué forma calcularlas a partir de una señal de voz correspondiente a un fonema vocal. Señalamos, también, que el comportamiento de estas señales puede explicarse a partir de dos dinámicas diferentes: un comportamiento lento y suave de alcance global llamado fluctuación y, por otro lado, una perturbación aleatoria de acción local. Es importante recordar que estas dinámicas ocurren normalmente, incluso en voces sanas estables [41, 138, 163]. Además, se ha argumentado que constituyen un rasgo distintivo de una persona y que se modifican notablemente ante la presencia de patologías [28, 58, 119]. Dedicaremos este capítulo al desarrollo de un nuevo método para el análisis y el modelado de SP y de SA reales, que contemple explícitamente en su estructura las fluctuaciones y las perturbaciones.

Desde hace varias décadas, los científicos y especialistas de la voz han centrado su interés en el estudio de las fluctuaciones y de las perturbaciones, con el propósito de extraer información útil para el diagnóstico del estado del aparato fonador. Esto ha propiciado la creación de diversos parámetros acústicos para cuantificar las perturbaciones en la SP y la SA (ver Sec. 4.3). No obstante, estas medidas suelen deteriorarse ante la presencia de fluctuaciones acentuadas. Por ello, la caracterización precisa de estas dos dinámicas se convirtió en una tarea sumamente difícil, agravado más aún por la falta de consenso al respecto [16, 39, 163]. Queda claro, entonces, que es necesario desarrollar estrategias más robustas que permitan llevar a cabo esta tarea. En este sentido, presentaremos aquí un método para el análisis estructural de SP y de SA, capaz de separar de forma óptima las fluctuaciones y las perturbaciones.

Como se verá más adelante, el método propuesto se basa en el modelado estocástico de SP y de SA. Recientemente, se ha incrementado el uso de esta clase de modelos en la tecnología moderna. A modo de ejemplo, modelos estocásticos que contemplan fluctuaciones o perturbaciones se han aplicado para mejorar la naturali-

dad de las voces artificiales [62, 134], para simular o reconocer diferentes emociones en interfaces *humano-computadora* [71], con el propósito de verificar la identidad de un individuo en sistemas de seguridad [58] y para simular voces patológicas en condiciones controladas [40, 107, 136]. Desarrollaremos aquí un modelo estructural basado en los modelos en espacio de estados lineales y gaussianos (MEEG), descriptos en la Sec. 5.2.1, capaz de representar las características principales de SP y de SA, identificadas a partir del análisis de señales reales [9, 138]. Mostraremos, además, que este modelo estructural constituye un marco conceptual que favorece el estudio y la caracterización de las fluctuaciones y de las perturbaciones [35, 76].

6.2. Antecedentes

Actualmente, existe en la literatura una gran variedad de estrategias para modelar SP y SA. Sin duda alguna, la forma más simple, y a la vez limitada, consiste en suponer fijos los períodos y las amplitudes sucesivos, ignorando así cualquier aperiodicidad o variación. No obstante, estudiando la dinámica de estas dos señales (ver la Fig. 4.2 a modo de ejemplo) vemos rápidamente que esta estrategia es claramente una simplificación alejada de la realidad.

Tomando en cuenta leyes estocásticas simples, se han desarrollado modelos de SP y de SA aptos para simular las perturbaciones en la voz. Usualmente, el *Jitter* (*Shimmer*) es generado a partir de alguna forma de ruido aleatorio que perturba a los períodos (las amplitudes) sucesivos. Estas estrategias se han aplicado, por ejemplo, para otorgar expresividad a una voz neutra, donde los parámetros de la transformación varían para cada tipo de emoción [29]; y para caracterizar el comportamiento espectral del *Jitter*, del *Shimmer* y del ruido de aspiración [117]. Recordemos, a su vez, que empleamos este tipo de estrategia para el desarrollo del método para la síntesis de vocales sostenidas con perturbaciones controladas descripto en el Cap. 4. Para ello, desarrollamos modelos estocásticos para el *Jitter* y el *Shimmer* controlados por los parámetros acústicos $Jitter_{\%}$ y $Shimmer_{\%}$ [1, 2, 3]. De forma análoga, se han desarrollado modelos estocásticos para representar las fluctuaciones. Por ejemplo, se han aplicado modelos de mezclas de gaussianas a la síntesis de voz expresiva, donde las diferentes emociones se caracterizaron mediante fluctuaciones generadas a partir de procesos multimodales [175].

Las estrategias descriptas en el párrafo anterior suponen que las SP y SA son procesos estocásticos estacionarios, independientes e idénticamente distribuidos. Sin embargo, estas hipótesis no se cumplen en las señales reales, ya que éstas exhiben un comportamiento organizado, que posee una marcada dependencia entre sus muestras (ver Fig. 4.2). Por lo tanto, estas estrategias no son adecuadas para representar los comportamientos complejos observados en estas señales.

Los primeros intentos por explicar la dependencia temporal en SP y SA reales emplearon métodos para el análisis de series temporales mediante modelos AR o ARMA. Estas ideas se aplicaron principalmente al estudio de SP [53, 141, 142]. Esto permitió caracterizar la fuerte autocorrelación evidenciada en SP, tanto para voces normales como patológicas, y se demostró que el orden de estos modelos puede considerarse un rasgo individual de una persona. Otros autores propusieron un método para la síntesis de voz considerando un *banco de Jitter* para caracterizar el nivel y la autocorrelación en la SP, y un modelo estocástico para simular el *Shimmer* en la SA [134]. Su propósito consistió en mejorar la naturalidad al simular ronquera

en voces artificiales. Si bien permiten capturar la autocorrelación, estos métodos requieren de la hipótesis poco realista que las SP y SA son procesos estocásticos estacionarios.

Por otra parte, las ecuaciones en diferencias estocásticas han demostrado ser muy útiles para representar dinámicas estocásticas complejas y no estacionarias [8, 51, 83]. De esta forma, estas estructuras constituyen un marco teórico propicio para el modelado de SP y SA. Aplicando estas ideas, Schoentgen desarrolló un modelo de *Jitter* capaz de representar las oscilaciones aperiódicas de las cuerdas vocales [138]. Este modelo sirvió para discriminar y analizar cómo influyen los factores glóticos y los externos en las perturbaciones de los períodos. Posteriormente, el modelo desarrollado por Schoentgen se aplicó a la síntesis de voz con ronquera, lo que permitió estudiar la estrecha relación entre esta afección y la dinámica del *Jitter* [61, 62]. Asimismo, este modelo de *Jitter* se utilizó para evaluar cómo influyen la experiencia y el entrenamiento de los fonoaudiólogos en la correcta identificación de los períodos en vocales sostenidas con perturbaciones, bajo condiciones controladas de *Jitter* [40] y de ruido aditivo [107].

En la literatura podemos encontrar otras estrategias para el modelado de SP y SA, desarrolladas a partir de modelos biomecánicos o de métodos para el procesamiento de señales [60, 168, 177]. Aoki e Ifukube estudiaron estas dos señales considerando procesos estadísticamente autosimilares e invariantes a la escala [9]. Los autores afirmaron que esta clase de modelos son adecuados para representar las SP y SA, y que aplicar estos modelos a la síntesis de vocales sostenidas ayuda a mejorar la naturalidad con que se perciben. En el LSyDnL, demostramos posteriormente que es conveniente describir estas señales en el marco de los procesos multifractales, aplicando el formalismo multifractal basado en onditas líderes [96, 97]. Estos procesos permitieron analizar los cambios en la regularidad de las señales reales. Además, aplicando el análisis multifractal se generaron parámetros útiles en la tarea de discriminar entre voces normales y patológicas.

6.2.1. Características de SP y de SA reales

Las diferentes estrategias comentadas en la sección anterior han demostrado ser muy útiles para mejorar la calidad en voces artificiales y para el modelado teórico de las perturbaciones en la voz. No obstante, sólo algunas de ellas son adecuadas para el estudio de SP y de SA reales. Nuestro objetivo aquí es desarrollar un método capaz de modelar teóricamente estas series temporales y que, al mismo tiempo, pueda aplicarse para el estudio de casos reales.

A lo largo del tiempo, la SP ha recibido una mayor atención por parte de los especialistas, en comparación con la SA [16, 140]. Es por ello que, a nuestro criterio, se ha avanzado mucho más en la caracterización del comportamiento de la SP. En [138], Schoentgen resumió las principales características observadas en la dinámica temporal de estas señales. A continuación, listaremos aquellas relevantes para el material desarrollado en este capítulo:

- Las perturbaciones en la SP presentan un comportamiento gaussiano.
- El *Jitter* posee una amplitud relativamente pequeña, en el rango del $0{,}1 - 1\,\%$ con respecto al período fundamental.

- Los períodos adyacentes en la SP se encuentran correlacionados y el grado de correlación varía para cada individuo.
- Existen fenómenos denominados *microtemblores* que refuerzan la correlación entre los períodos.
- El *Jitter* es un fenómeno genuinamente aleatorio.
- En promedio, voces con mayor período fundamental presentan un *Jitter* mayor.
- Existe evidencia de que el *Jitter* aumenta ante la presencia de algunas patologías laríngeas.
- Es posible calcular estadísticos significativos para el *Jitter* a partir de emisiones vocales sostenidas.

Asimismo, es poco lo que se conoce actualmente respecto a la dinámica temporal de la SA. En general, se considera que el *Shimmer* es un fenómeno aleatorio y que presenta una amplitud pequeña, en el orden del 1 % de la amplitud promedio de la señal de voz [16, 140].

En [9], Aoki e Ifukube estudiaron un conjunto de SP y de SA extraídas a partir de vocales sostenidas de adultos masculinos sanos. Afirman que es válido considerar a estas señales como procesos estacionarios independientes entre sí y con distribución gaussiana. Demuestran, además, que la SP y la SA desarrollan un comportamiento espectral de la forma $1/f$. De lo expuesto hasta aquí, consideramos que no es correcto aseverar que estas series temporales sean estacionarias. Se ha demostrado que si bien las perturbaciones pueden considerarse como procesos estacionarios, esto no implica que la SP y la SA satisfagan esta propiedad [163]. Además, Aoki e Ifukube no tuvieron en cuenta las fluctuaciones y los microtemblores, que son fenómenos no estacionarios que se observan normalmente en estas señales [16, 139, 140].

En la actualidad, no se ha arribado a un consenso respecto a si existe o no alguna relación entre la SP y la SA. Sin embargo, comúnmente se trabaja bajo el supuesto que estas señales son independientes entre sí. A lo largo de este capítulo consideraremos válida esta hipótesis. Además, en adelante consideraremos que la SA presenta características temporales análogas a las descriptas por Schoentgen para la SP. Si bien es una hipótesis muy fuerte, ésta nos permitirá desarrollar un método único para el análisis de SP y de SA reales.

La mayoría de las estrategias para el modelado de SP y de SA contemplan, a lo sumo, un pequeño subconjunto de las propiedades descriptas en esta sección. En algunos casos, incluso, esta información es absolutamente ignorada, por lo que producen series artificiales con características irreales. Esto demuestra la necesidad de contar con modelos más flexibles para estas señales, capaces de representar adecuadamente sus comportamientos característicos. Además, hasta donde sabemos, no existe ningún modelo que incorpore las propiedades descriptas y que, a su vez, sea aplicable al análisis de SP y de SA reales. Teniendo en cuenta todo esto, en este capítulo desarrollaremos una estrategia basada en métodos en espacio de estados para el análisis y modelado de señales reales. Consideraremos aquí SP y SA correspondientes al conjunto de vocales /a/ sostenidas de sujetos sanos en la BDDV, calculadas aplicando la técnica descripta en la Sec. 4.2. Trabajamos con los métodos en espacio de estados debido a que permiten combinar las propiedades descriptas

y el procesamiento de señales guiado por modelos. A continuación, introduciremos los modelos estructurales para series temporales estocásticas y describiremos cómo aplicarlos al modelado de SP y de SA.

6.3. Modelos estructurales en espacio de estados para SP y SA

El análisis estructural para series temporales consiste en descomponer una señal de interés en elementos con una interpretación simple y directa. Esta metodología proporciona un marco propicio para el procesamiento de señales guiado por modelos, ayudando así a dilucidar la dinámica subyacente en un proceso complejo. Para mayor información, el lector puede referirse a [35, 51, 76, 89]. En este contexto, se denominan modelos estructurales aquellos modelos desarrollados para explicar la dinámica de una serie temporal, empleando reglas estocásticas simples. Dedicaremos esta sección a presentar un modelo estructural novedoso, adecuado para el estudio y modelado de SP y de SA.

En este capítulo, nos concentraremos en el análisis estructural basado en modelos en espacio de estados lineales y gaussianos (MEEG). En el Cap. 5 introdujimos esta familia de modelos y, a su vez, presentamos algunos de los métodos disponibles para el procesamiento de señales a partir de un MEEG. De esta forma, podremos utilizar los métodos en espacio de estados para el estudio de SP y de SA reales. Teniendo en cuenta que ambas señales son unidimensionales, es decir, la dimensión de las observaciones es $r = 1$, en adelante, $Z_N = \big\{z[1], z[2], \ldots, z[N]\big\}$ representa un conjunto de períodos o de amplitudes, según corresponda.

6.3.1. Componentes del análisis estructural

En el desarrollo del modelo estructural para SP y SA consideraremos específicamente tres componentes: tendencia, componente cíclico y perturbaciones [51, 76, 89]. Supondremos, entonces, que $z[n]$ resulta de la combinación de estos componentes de acuerdo con el siguiente modelo estructural:

$$z[n] = \mu[n] + \psi[n] + \varepsilon[n], \qquad \varepsilon[n] \sim \mathcal{N}(0, \sigma^2), \tag{6.1}$$

donde μ, ψ y ε son la tendencia, el componente cíclico y las perturbaciones, respectivamente. Consideraremos, además, que estos componentes desarrollan un comportamiento estocástico a lo largo del tiempo, lo que garantiza una flexibilidad adecuada en el modelo. En particular, de acuerdo a la Ec. (6.1), las perturbaciones ε forman un proceso estocástico gaussiano, estacionario e independiente en el tiempo.

Tomando en cuenta el material de la Sec. 6.2.1, podemos asociar los tres componentes en el modelo (6.1) con las características observadas en las señales reales, siempre y cuando se definan de forma adecuada. En lo que sigue, consideraremos que la tendencia μ representará el comportamiento de las fluctuaciones que se caracteriza por una dinámica suave y lenta identificable a una escala global [163]. En segundo lugar, el componente cíclico ψ simulará los microtemblores o cualquier otro fenómeno que influya en la autocorrelación de las observaciones z [138, 139]. Finalmente, las perturbaciones ε estarán asociadas con el *Jitter* o el *Shimmer*, según corresponda [16, 163]. Veamos ahora como definir matemáticamente la tendencia y el componente cíclico.

Tendencia

En este desarrollo, nos concentraremos en tres diferentes alternativas para representar la tendencia μ en SP y SA reales. La forma más sencilla consiste en el denominado *modelo de nivel local* $\mathcal{T}_1$, conocido también como *camino aleatorio*, representado por la siguiente ecuación en diferencias estocástica:

$$\mu[n+1] = \mu[n] + \nu[n], \qquad \nu[n] \sim \mathcal{N}(0, \sigma_\nu^2). \tag{6.2}$$

Se desprende de esta expresión que la amplitud de la señal en un instante dado es igual a la amplitud en el instante inmediato anterior más una variación aleatoria y de alcance local [35]. Esto explica el nombre del modelo.

Modificando adecuadamente la definición (6.2) podemos generar otras formas para la tendencia. Una alternativa consiste en el *modelo de tendencia lineal suave* $\mathcal{T}_2$, también llamado *camino aleatorio integrado*, cuya expresión es:

$$\begin{aligned} \mu[n+1] &= \mu[n] + \beta[n], \\ \beta[n+1] &= \beta[n] + \zeta[n], \qquad \zeta[n] \sim \mathcal{N}(0, \sigma_\zeta^2), \end{aligned} \tag{6.3}$$

donde β es la pendiente estocástica que rige el cambio en la tendencia [51]. Podemos apreciar que la pendiente β presenta un comportamiento similar a $\mathcal{T}_1$. Más adelante, mostraremos algunas simulaciones con formas de tendencia cercanas a $\mathcal{T}_2$, donde modificamos la pendiente considerando modelos AR.

Por último, podemos generar una representación más flexible de la tendencia mediante la combinación de los dos modelos anteriores. Se denomina *modelo de nivel y pendiente lineal locales* $\mathcal{T}_3$ y se define de la siguiente forma:

$$\begin{aligned} \mu[n+1] &= \mu[n] + \beta[n] + \nu[n], \qquad \nu[n] \sim \mathcal{N}(0, \sigma_\nu^2), \\ \beta[n+1] &= \beta[n] + \zeta[n], \qquad \zeta[n] \sim \mathcal{N}(0, \sigma_\zeta^2). \end{aligned} \tag{6.4}$$

En este caso, se permite que tanto la amplitud como la pendiente varíen de forma estocástica, donde la pendiente nuevamente toma la forma de $\mathcal{T}_1$. Podemos observar que, considerando $\sigma_\zeta^2 \to 0$ y $\beta[0] = 0$, el modelo anterior coincide con $\mathcal{T}_1$. Por otro lado, tomando ahora $\sigma_\nu^2 \to 0$ en la Ec. (6.4) obtenemos $\mathcal{T}_2$.

Es importante destacar que, a pesar de su simplicidad, los tres modelos descriptos son capaces de representar, con diferentes matices, series temporales de larga duración, no estacionarias y fuertemente correlacionadas [35, 51, 89]. Como puede apreciarse de las definiciones, $\mathcal{T}_1$ es el modelo más simple pero a su vez el más limitado, mientras que $\mathcal{T}_3$ es el más flexible. Por su parte, $\mathcal{T}_2$ genera componentes de tendencia considerablemente más suaves que los otros modelos. Esto último suele ser muy útil en aplicaciones en las que se estudian dinámicas lentas fuertemente contaminadas con ruido [76].

Componente cíclico

Para especificar el componente cíclico, escogimos la familia de modelos autorregresivos de orden ρ (AR_ρ). Estos modelos se rigen por la siguiente expresión:

$$\psi[n+1] = -a_1\,\psi[n] - a_2\,\psi[n-1] - \ldots - a_\rho\,\psi[n-\rho+1] + \xi[n], \tag{6.5}$$

donde $\xi[n] \sim \mathcal{N}(0, \sigma_\xi^2)$ y los signos menos se agregaron por conveniencia. Podemos apreciar que la función de transferencia de este modelo es análoga a la presentada en la Ec. (3.4).

Es importante destacar que sólo los modelos AR_ρ estables forman una representación adecuada del componente cíclico [51, 76]. Recordando la Sec. 3.4.1, esto se cumplirá si todos los polos de la función de transferencia de AR_ρ se ubican en el interior del círculo unitario en el plano complejo. Luego, será importante obtener los coeficientes $\left\{a_1, a_2, \ldots, a_\rho\right\}$ garantizando esto último.

6.3.2. Modelo estructural en espacio de estados

Describiremos aquí cómo construir un MEEG adecuado para el análisis estructural de SP y de SA, considerando el modelo estructural (6.1). Para ilustrar la metodología, en esta explicación tomaremos una tendencia de la forma $\mathcal{T}_3$ y un componente cíclico AR_ρ. Los modelos con otra selección de componentes se construyen de forma análoga [51, 89]. Agrupando los componentes de tendencia y cíclico en forma vectorial, definimos el vector de estados como sigue:

$$\begin{aligned}\mathbf{x}[n] &= \begin{pmatrix} x_{(1)}[n] & x_{(2)}[n] & x_{(3)}[n] & x_{(4)}[n] & \ldots & x_{(p)}[n] \end{pmatrix}^T \\ &= \begin{pmatrix} \mu[n] & \beta[n] & \psi[n] & \psi[n-1] & \ldots & \psi[n-\rho+1] \end{pmatrix}^T .\end{aligned} \tag{6.6}$$

Tomando en cuenta las expresiones (6.4) y (6.5), se desprende que la matriz de transición de estados queda determinada por:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & -a_1 & -a_2 & \ldots & -a_{\rho-1} & -a_\rho \\ 0 & 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix}. \tag{6.7}$$

De forma similar, se demuestra que las matrices de error y de observación presentan las siguientes formas:

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{y} \qquad \mathbf{H} = \begin{pmatrix} 1 & 0 & 1 & 0 & \ldots & 0 \end{pmatrix}. \tag{6.8}$$

Por otro lado, agrupando convenientemente los diferentes procesos estocásticos involucrados en el análisis estructural, obtenemos sendas expresiones para los errores de estados y de observación:

$$\mathbf{w}[n] = \begin{pmatrix} \nu[n] & \zeta[n] & \xi[n] \end{pmatrix}^T \qquad \text{y} \qquad v[n] = \varepsilon[n], \tag{6.9}$$

así como las siguientes expresiones para sus varianzas:

$$\mathbf{Q} = \begin{pmatrix} \sigma_\nu^2 & 0 & 0 \\ 0 & \sigma_\zeta^2 & 0 \\ 0 & 0 & \sigma_\xi^2 \end{pmatrix} \qquad \text{y} \qquad \sigma^2, \tag{6.10}$$

respectivamente. Estas expresiones para las varianzas surgen de suponer que los procesos ε, ν, ζ y ξ son independientes con respecto al tiempo y mutuamente independientes.

Revisando las expresiones obtenidas, podemos observar que constituyen un MEEG con parámetros constantes. A su vez, el modelo desarrollado no está sujeto a ninguna entrada o estímulo externo, por lo que $\mathbf{f}[n] = \mathbf{0}$ y $\mathbf{u}[n] = \mathbf{0}$ en la definición (5.3). Del procedimiento descripto, se desprende que las dimensiones del espacio de estados p y del error de estados q dependerán de los componentes escogidos para construir el modelo estructural. Una vez seleccionados el tipo de tendencia y el orden ρ del componente cíclico, los parámetros del MEEG serán: el estado inicial $\mathbf{x}_0$ y su matriz de covarianza $\mathbf{P}_0$, las varianzas $\left\{\sigma^2, \sigma_\nu^2, \sigma_\zeta^2, \sigma_\xi^2\right\}$ y los coeficientes $\left\{a_1, a_2, \ldots, a_\rho\right\}$. Determinados todos estos parámetros, podemos generar series artificiales aplicando las expresiones de la definición (5.3).

En la Fig. 6.2, presentamos ejemplos de series temporales sintetizadas con el método descripto, considerando diferentes formas de tendencia. Las curvas en negro corresponden a la tendencia $\mu[n]$, mientras que las líneas en gris representan las observaciones $z[n]$ formadas por la tendencia y las perturbaciones. Comparar el comportamiento de estos ejemplos con el correspondiente a las señales reales de la Fig. 4.2. Dibujamos además, de forma superpuesta, las estimaciones de la tendencia $\hat{\mu}$ calculadas aplicando el filtrado y el suavizado de Kalman. Podemos observar que las estimaciones resultaron muy similares a las respectivas tendencias. Profundizaremos al respecto más adelante en este capítulo.

6.4. Métodos en espacio de estados

En el capítulo anterior presentamos un abanico de métodos en espacio de estados específicos para los MEEG. A partir de lo desarrollado en las secciones precedentes, se desprende que el modelo estructural propuesto resulta compatible con estos métodos y, como resultado, podremos emplear todas estas técnicas para el análisis estructural de SP y de SA reales. En particular, consideraremos aquí las versiones de los métodos en espacio de estados que contemplan la inicialización difusa. Esta elección se debe a que, como es sabido, los valores de las amplitudes y los períodos varían considerablemente con el sexo o el estado las cuerdas vocales [16, 162, 163] y nuestro objetivo es precisamente que el análisis estructural permita estudiar diferentes series temporales.

6.4.1. Estimación de los componentes

Como dijimos oportunamente, los métodos de filtrado y de suavizado de Kalman permiten, a partir de un conjunto de observaciones, calcular los valores óptimos de los vectores de estados de un MEEG. De acuerdo a la Ec. (6.6), la tendencia y el componente cíclico forman parte del vector de estados del modelo estructural. Por

ello, estas técnicas resultan sumamente útiles para estimar estos componentes en SP y SA. Más adelante en este capítulo analizaremos con mayor detalle ambas estimaciones, comparándolas y señalando sus diferencias. En la Fig. 6.2, podemos observar algunos ejemplos de estimaciones filtradas y suavizadas de la tendencia para diferentes series temporales artificiales. Apreciamos que las estimaciones obtenidas con estos métodos son muy cercanas a la tendencia real. Más adelante en este capítulo describiremos en detalle esta figura.

Por otro lado, para la estimación de las perturbaciones resulta conveniente emplear el método de suavizado específico para ello (ver Sec. 5.3.3). Esto surge porque en el modelo estructural propuesto la información de las perturbaciones queda representada en el error de observación. A su vez, podemos emplear la técnica de suavizado de las perturbaciones para estimar los errores de estados suavizados. Esta información serviría para describir con mayor detalle la dinámica de la tendencia y del componente cíclico. Recordemos que estos dos métodos de suavizado se aplican posteriormente al filtrado de Kalman. Es importante destacar que en este documento no consideramos las estimaciones filtradas de los errores de estados y de observación.

6.4.2. Estimación óptima de los parámetros

Describiremos ahora el procedimiento implementado para calcular el conjunto de parámetros involucrado en la definición del modelo estructural. En adelante, supondremos que estos parámetros se encuentran agrupados en $\boldsymbol{\Theta}$. Ya que consideramos la inicialización difusa, no será necesario entonces calcular la información correspondiente al vector de estados iniciales. Así, en general $\boldsymbol{\Theta} = \left\{\sigma^2,\, \sigma_\nu^2,\, \sigma_\zeta^2,\, \sigma_\xi^2,\, a_1, a_2,\, \ldots\, , a_\rho\right\}$, aunque esto varía de acuerdo a los componentes considerados en el modelo.

El cálculo de $\boldsymbol{\Theta}$ se llevó a cabo mediante la resolución del problema de optimización descripto en la Sec. 5.5. La búsqueda de la solución se realizó a través de un procedimiento iterativo. Para ello, dividimos primeramente el problema de optimización agrupando el cálculo de las varianzas, por un lado, y la estimación de los coeficientes del modelo AR, por el otro. Existen dos importantes razones que impulsaron esta decisión. En primer lugar, contamos con las expresiones desarrolladas en la Sec. 5.5.3 para la estimación de la varianza de los errores de estados y de observación. En segundo lugar, podemos aplicar el *análisis predictivo lineal* (LP) para estimar los coeficientes del modelo AR. Recordemos que el método LP produce estimaciones de los coeficientes $\left\{\hat{a}_1, \hat{a}_2,\, \ldots\, , \hat{a}_\rho\right\}$ que constituyen un modelo AR estable. En la Fig. 6.1 presentamos el diagrama de flujo del procedimiento para la estimación del conjunto de parámetros $\boldsymbol{\Theta}$. A continuación, describiremos en qué consiste este método.

Inicialmente, se seleccionan las observaciones Z_N a estudiar y se construye el MEEG siguiendo el procedimiento descripto en la Sec. 6.3.2. Esta información es indispensable para continuar con el procedimiento. A continuación, se generan valores iniciales para los parámetros del modelo. En esta etapa se emplea el modelo $\mathcal{T}_1$, que es la tendencia más simple. Se producen estimaciones groseras de $\left\{\hat{\sigma}^2,\, \hat{\sigma}_\nu^2\right\}$, a partir de Z_N y de valores iniciales aleatorios, aplicando de forma iterativa las expresiones (5.23) y (5.27). Con estos parámetros, se estima la tendencia para $\mathcal{T}_1$ aplicando el filtrado y el suavizado de Kalman. Luego, se calcula el error $\hat{e}_\mu[n] = z[n] - \hat{\mu}[n|N]$, donde $\hat{\mu}[n|N] = \hat{x}_{(1)}[n|N]$ corresponde a la estimación suavizada de la tendencia. Recordemos que la estructura de $\mathcal{T}_1$ no es adecuada para capturar el componente

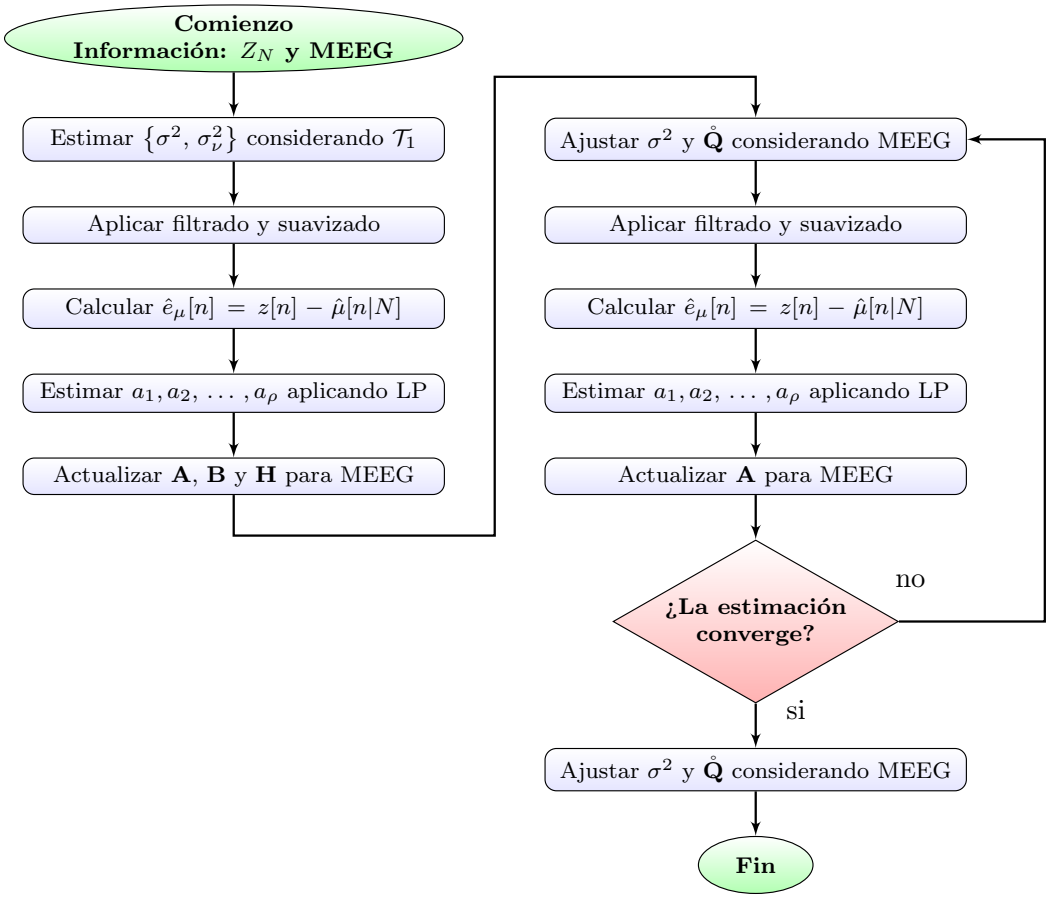


Figura 6.1: Diagrama de flujo del procedimiento para la estimación de los parámetros desconocidos $\boldsymbol{\Theta}$ de los modelos estructurales. El método arroja como resultado los valores óptimos de las varianzas $\left\{\hat{\sigma}^2, \hat{\sigma}_\nu^2, \hat{\sigma}_\zeta^2, \hat{\sigma}_\xi^2\right\}$ y de los coeficientes AR $\left\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_\rho\right\}$.

cíclico, por lo que esta información se preserva en el error $\hat{e}_\mu[n]$. A partir de este error, se generan las estimaciones groseras $\left\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_\rho\right\}$ aplicando el método LP. Seguidamente, se actualiza la estructura del MEEG, donde la matriz **A** se inicializa con los coeficientes estimados y las matrices de covarianza de los errores con $\left\{\hat{\sigma}^2, \hat{\sigma}_\nu^2\right\}$. Esto da paso a la búsqueda iterativa de $\boldsymbol{\Theta}$.

La búsqueda iterativa es similar al procedimiento descripto en el párrafo anterior. Se ajustan las varianzas a partir de Z_N, y se estiman los vectores de estados aplicando los métodos de filtrado y suavizado. Luego, se calcula el error $\hat{e}_\mu[n] = z[n] - \hat{\mu}[n|N]$, el cual preserva la información cíclica. A partir de este error, se estiman los coeficientes $\left\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}\rho\right\}$ aplicando el método LP, y se actualiza la matriz de transición de estados **A**. Se evalúa la convergencia del método con la distancia *euclídea* entre las estimaciones actual y pasada de los coeficientes $\left\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_\rho\right\}$. Si es mayor a un umbral preestablecido, es necesario seguir ajustando las estimaciones y se repite el proceso. En caso contrario, se considera que los coeficientes son lo suficientemente buenos. El umbral se fijó en el rango $(1 \times 10^{-6}, 1 \times 10^{-3})$, dependiendo de la calidad deseada. Por último, se lleva a cabo un ajuste final de las varianzas de los errores de estados y de observación.

Tabla 6.1: Estructura de los MEEG considerados en las simulaciones con series temporales artificiales. Corresponden a diferentes alternativas para la tendencia.

Modelo	I	II	III	IV
H	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}1 & 0\end{pmatrix}$	$\begin{pmatrix}1 & 0\end{pmatrix}$	$\begin{pmatrix}1 & 0 & 0\end{pmatrix}$
A	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$	$\begin{pmatrix}1 & 1\\ 0 & -\varphi_0\end{pmatrix}$	$\begin{pmatrix}1 & 1 & 0\\ 0 & -\varphi_1 & -\varphi_2\\ 0 & 1 & 0\end{pmatrix}$
B	$\begin{pmatrix}1\end{pmatrix}$	$\begin{pmatrix}0\\ 1\end{pmatrix}$	$\begin{pmatrix}0\\ 1\end{pmatrix}$	$\begin{pmatrix}0\\ 1\\ 0\end{pmatrix}$

Es importante destacar que el método descripto se aplica en situaciones donde el componente cíclico participa del modelo estructural. En caso contrario, el problema se simplifica y es suficiente con calcular iterativamente las varianzas empleando las expresiones (5.23) y (5.27). Podemos mencionar además que este procedimiento presenta algunas propiedades interesantes. Como se explicó anteriormente, en cada iteración se estiman por separado las varianzas y los coeficientes del modelo cíclico. Así, el problema de optimización se divide en dos subproblemas más simples y fáciles de resolver, lo que acelera la convergencia del procedimiento. A su vez, el método LP garantiza que las estimaciones $\left\{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_\rho\right\}$ constituyen un proceso estacionario en sentido amplio, lo que es un requisito fundamental. Finalmente, este procedimiento demostró, en estudios preliminares realizados por nosotros, ser robusto con respecto a la inicialización aleatoria de las varianzas.

En lo que resta de este capítulo, describiremos las simulaciones y los experimentos llevados a cabo con el propósito de estudiar el comportamiento del análisis estructural desarrollado y comentaremos los resultados alcanzados.

6.5. Análisis de series artificiales

En una primera etapa, evaluamos las bondades de los métodos en espacio de estados y del análisis estructural en simulaciones con señales artificiales. El propósito de estas simulaciones fue doble. Por un lado, corroborar que el modelo estructural desarrollado en la Sec. 6.3 fuera capaz de generar señales con dinámicas similar a las observadas en SP y SA reales. Por otro lado, estudiar el desempeño, bajo condiciones controladas, de los métodos en espacios de estados contemplados en esta aplicación.

Para las simulaciones, escogimos cuatro MEEG de prueba para la síntesis y posterior análisis de un conjunto de series temporales artificiales. En particular, nos concentramos en diferentes modelos de tendencia, suponiendo que no existe la componente cíclica. En todos los modelos estructurales considerados, las dimensiones de los errores de medición y de transición de estados son $q = 1$ y $r = 1$. Luego, sus varianzas se reducen a σ^2 y σ_*^2, respectivamente. En la Tab. 6.1 se indican los parámetros de cada uno de los MEEG escogidos. Podemos apreciar que los modelos I y II coinciden con $\mathcal{T}_1$ y $\mathcal{T}_2$, respectivamente. Por su parte, los modelos III y IV

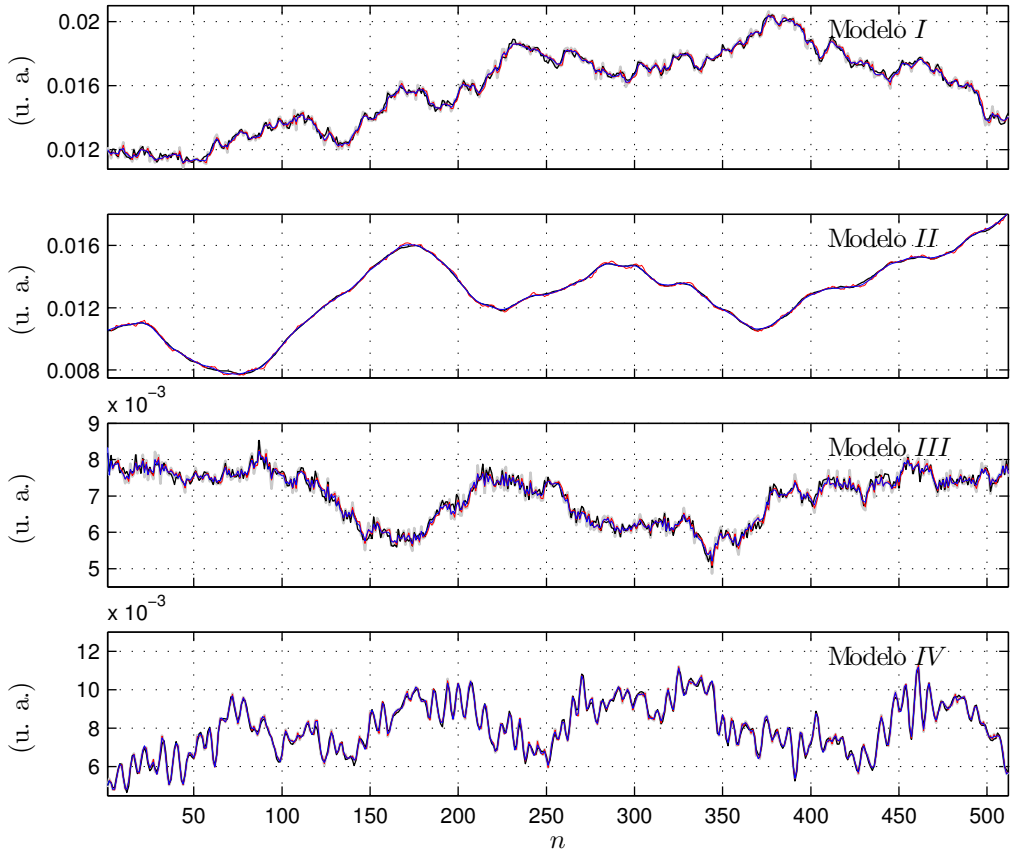


Figura 6.2: Series temporales sintetizadas a partir de los MEEG propuestos. Se presentan las observaciones $z[n]$ en gris, las tendencias $\mu[n]$ en negro, las estimaciones filtradas $\hat{\mu}[n|n]$ en rojo y las estimaciones suavizadas $\hat{\mu}[n|N]$ en azul.

constituyen otras alternativas para la tendencia, las cuales contemplan un cambio de pendiente con dinámica autorregresiva. Para estos casos, definimos $\varphi_0 = 1/2$, $\varphi_1 = -1$ y $\varphi_2 = -13/16$. Estos coeficientes dan lugar a procesos estacionarios en sentido amplio.

En la Fig. 6.2 se reportan ejemplos de series temporales generadas a partir de los cuatro MEEG escogidos. Podemos observar que estos modelos dan lugar a comportamientos muy diversos. Las varianzas se fijaron aleatoriamente de acuerdo a: $\sigma^2 \sim \mathcal{U}_{(1\times e^{-8},\, 5\times e^{-8})}$, $\sigma_*^2 \sim \mathcal{U}_{(1\times e^{-8},\, 5\times e^{-6})}$ para los modelos I, III, IV y $\sigma_*^2 \sim \mathcal{U}_{(1\times e^{-10},\, 5\times e^{-10})}$ para el modelo II. Estos rangos se fijaron de forma tal que las series generadas posean una magnitud comparable a SP reales, en el orden de los milisegundos [138, 163]. Presentamos, en gris, las observaciones $z[n]$ generadas y, en negro, las tendencias (ocultas) $\mu[n]$ correspondientes. Podemos apreciar que cada tendencia es perturbada y enmascarada por el ruido, dando lugar así a las observaciones.

Los MEEG considerados generan componentes de tendencia con dinámicas considerablemente diferentes. Por ejemplo, la tendencia del modelo II desarrolla un comportamiento más suave y con un número menor de transiciones, en comparación con los otros MEEG. A su vez, las tendencias de los modelos III y IV presentan una dinámica local más rica y compleja favorecida por la pendiente autorregresiva, en comparación con el modelo I. Así, los modelos estructurales producen señales

con características similares a las observadas en SP y SA (comparar las Figs. 4.2 y 6.2). Es importante destacar que, pese a su sencillez, los MEEG escogidos permiten simular dinámicas estocásticas no estacionarias.

En las gráficas de la Fig. 6.2 presentamos, de forma superpuesta, las estimaciones de la tendencia calculadas aplicando el filtrado $\hat{\mu}[n|n]$, en rojo, y el suavizado $\hat{\mu}[n|N]$, en azul. Podemos observar que las estimaciones obtenidas resultaron muy similares a las tendencias reales, aún ante la presencia del ruido de observación. Más adelante, describiremos con mayor detalle estas dos estimaciones.

6.5.1. Simulaciones

Considerando los métodos en espacio de estados y los modelos de la Tab. 6.1, procedimos a estudiar dos aspectos importantes: la estimación de los estados iniciales aplicando el filtrado y el suavizado difuso, por un lado, y el cálculo de las varianzas de los MEEG, por el otro. Estas simulaciones nos ayudaron a interpretar y a comprender con mayor detalle estos dos aspectos. A continuación, presentaremos las simulaciones realizadas.

Estimación difusa de los estados iniciales

Como destacamos en Cap. 5, los procesos de filtrado y de suavizado de Kalman dependen de la estrategia de inicialización empleada. Afortunadamente, se ha demostrado que en modelos invariantes en el tiempo una condición inicial errónea produce un fenómeno transitorio que se atenúa rápidamente [30]. Por otro lado, emplear una inicialización difusa nos evita fijar arbitrariamente el vector de estados iniciales [88]. Esto resulta sumamente útil en situaciones donde se requiere analizar un conjunto de señales diferentes. Por ello, en primera instancia estudiamos el comportamiento de la inicialización difusa.

La metodología empleada fue la siguiente. Para cada uno de los modelos de la Tab. 6.1, se generaron diferentes realizaciones $X_N^{\mathrm{k}} = \left\{\mathbf{x}^{\mathrm{k}}[1], \mathbf{x}^{\mathrm{k}}[2], \ldots, \mathbf{x}_N^{\mathrm{k}}\right\}$ y $Z_N^{\mathrm{k}} = \left\{z^{\mathrm{k}}[1], z^{\mathrm{k}}[2], \ldots, z^{\mathrm{k}}[N]\right\}$ a partir de estados iniciales aleatorios, donde $\mathrm{k} = 1, 2, \ldots, 100$ indica el número de la realización. A su vez, los valores de las varianzas σ^2 y σ_*^2 se generaron aleatoriamente respetando las distribuciones introducidas anteriormente. Los parámetros restantes no se modificaron. Por simplicidad, consideramos vectores de estados iniciales de la forma:

$$\mathbf{x}_0^{\mathrm{k}} = \begin{pmatrix} \eta & 0 & \cdots & 0 \end{pmatrix}^T, \tag{6.11}$$

con $\eta \sim \mathcal{U}_{(0{,}4\times10^{-3},\, 1{,}2\times10^{-3})}$. Nuevamente, η se fijó de forma tal de obtener series artificiales de magnitud comparable a SP reales [138, 163].

Luego, a partir de Z_N^{k} y del respectivo MEEG se obtuvieron las estimaciones filtradas $\hat{\mathbf{x}}^{\mathrm{k}}[n|n]$ y suavizadas $\hat{\mathbf{x}}^{\mathrm{k}}[n|N]$ de los vectores de estados. Recordemos que, debido a la inicialización difusa, en el filtrado las d primeras estimaciones no están disponibles [88]. En nuestras simulaciones, el proceso de inicialización arrojó $d = 1$, lo que implica que la inicialización se estabiliza en una iteración. Luego, desde $d+1$ en adelante es posible calcular $\hat{\mathbf{x}}^{\mathrm{k}}[n|n]$. Por su parte, el suavizado no presenta este inconveniente. Con esta información, calculamos el error en las primeras estimaciones

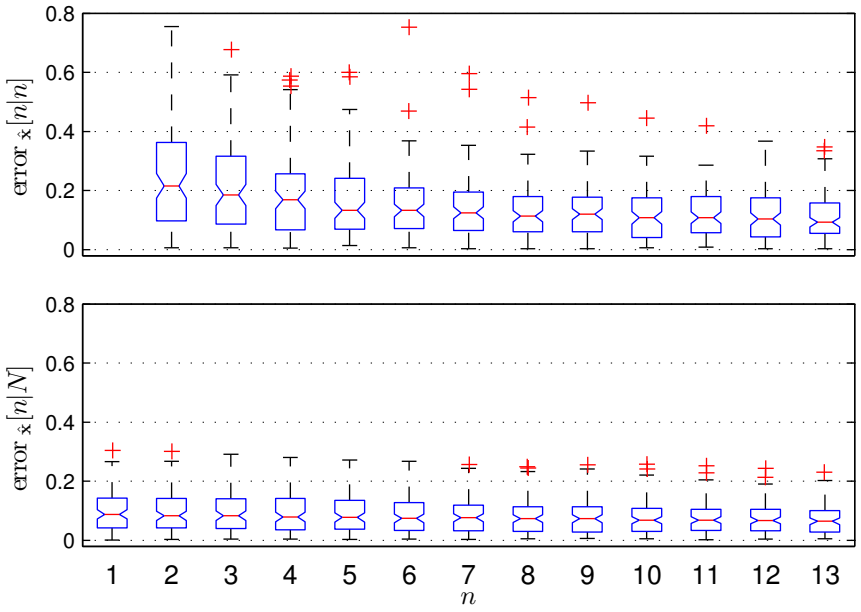


Figura 6.3: Error en la estimación de la información de los estados iniciales aplicando filtrado y suavizado difuso. *Arriba:* Diagrama de cajas de $\text{error}_{\hat{\mathbf{x}}}[n|n]$. *Abajo:* Diagrama de cajas de $\text{error}_{\hat{\mathbf{x}}}[n|N]$. Se utilizó un conjunto de 100 realizaciones.

Tabla 6.2: Valor promedio y desvío estándar de $\text{error}_{\hat{\mathbf{x}}}[n|n]$ para cada MEEG, considerando 100 realizaciones.

n	2	3	4	5	6
I	$0{,}28 \pm 0{,}24$	$0{,}20 \pm 0{,}18$	$0{,}18 \pm 0{,}16$	$0{,}21 \pm 0{,}16$	$0{,}14 \pm 0{,}12$
II	$0{,}29 \pm 0{,}22$	$0{,}20 \pm 0{,}19$	$0{,}17 \pm 0{,}15$	$0{,}21 \pm 0{,}18$	$0{,}15 \pm 0{,}14$
III	$0{,}35 \pm 0{,}33$	$0{,}30 \pm 0{,}27$	$0{,}28 \pm 0{,}27$	$0{,}23 \pm 0{,}19$	$0{,}18 \pm 0{,}16$
IV	$0{,}34 \pm 0{,}29$	$0{,}24 \pm 0{,}23$	$0{,}21 \pm 0{,}20$	$0{,}22 \pm 0{,}17$	$0{,}18 \pm 0{,}12$

de los vectores de estados filtrados y suavizados a partir de las siguientes expresiones:

$$\text{error}^{\text{k}}_{\hat{\mathbf{x}}}[n|n] = 100 \times \frac{\left\|\hat{\mathbf{x}}^{\text{k}}[n|n] - \mathbf{x}^{\text{k}}[n]\right\|}{\left\|\mathbf{x}^{\text{k}}[n]\right\|}, \tag{6.12}$$

$$\text{error}^{\text{k}}_{\hat{\mathbf{x}}}[n|N] = 100 \times \frac{\left\|\hat{\mathbf{x}}^{\text{k}}[n|N] - \mathbf{x}^{\text{k}}[n]\right\|}{\left\|\mathbf{x}^{\text{k}}[n]\right\|}. \tag{6.13}$$

En la Tab. 6.2, informamos el valor promedio de $\text{error}_{\hat{\mathbf{x}}}[n|n]$, con su correspondiente desvío estándar, para cada modelo y para valores de $n = 2, 3, \ldots, 6$. Puede apreciarse que, en promedio, el error de estimación es bajo, inferior al 1 %, y que para valores crecientes de n éste disminuye hasta luego estabilizarse. Algo similar ocurre con su desvío estándar. En la gráfica superior de la Fig. 6.3, se muestra un diagrama de cajas de $\text{error}_{\hat{\mathbf{x}}}[n|n]$ para valores de $n = 2, \ldots, 13$, correspondientes al modelo II. De ésta se pueden extraer conclusiones similares a las que arrojó el análisis anterior. Esto nos permite afirmar que el filtrado difuso proporciona buenas estimaciones de los vectores de estados del sistema para las primeras iteraciones, cualquiera sea la condición inicial real.

Tabla 6.3: Valor promedio y desvío estándar de $\text{error}_{\hat{\mathbf{x}}}[n|N]$ para cada MEEG, considerando 100 realizaciones.

n	1	2	3	4	5	6
I	$0{,}14 \pm 0{,}13$	$0{,}12 \pm 0{,}11$	$0{,}11 \pm 0{,}10$	$0{,}11 \pm 0{,}10$	$0{,}10 \pm 0{,}09$	$0{,}08 \pm 0{,}08$
II	$0{,}10 \pm 0{,}10$	$0{,}09 \pm 0{,}09$	$0{,}09 \pm 0{,}09$	$0{,}08 \pm 0{,}07$	$0{,}07 \pm 0{,}07$	$0{,}07 \pm 0{,}06$
III	$0{,}25 \pm 0{,}27$	$0{,}23 \pm 0{,}23$	$0{,}24 \pm 0{,}23$	$0{,}21 \pm 0{,}14$	$0{,}14 \pm 0{,}13$	$0{,}13 \pm 0{,}13$
IV	$0{,}14 \pm 0{,}16$	$0{,}16 \pm 0{,}16$	$0{,}16 \pm 0{,}16$	$0{,}16 \pm 0{,}08$	$0{,}08 \pm 0{,}07$	$0{,}08 \pm 0{,}08$

Del mismo modo, en la Tab. 6.3 presentamos el promedio de $\text{error}_{\hat{\mathbf{x}}}[n|N]$, con su correspondiente desvío estándar, para valores de $n = 1, 2, \ldots, 7$. En este caso, se aprecia fácilmente que el error de estimación y su desvío estándar se redujeron notablemente e, incluso, su variación con respecto a n también disminuyó. En la gráfica inferior de la Fig. 6.3, podemos apreciar el correspondiente diagrama de cajas de $\text{error}_{\hat{\mathbf{x}}}[n|N]$ correspondientes al modelo II para valores de $n = 1, 2, \ldots, 13$. Su comportamiento respalda los resultados obtenidos. Comparando los diagramas de cajas de la Fig. 6.3, resulta notorio el modo en que el suavizado reduce el error de estimación.

Error en la estimación de los parámetros

En una segunda etapa, nos concentramos en evaluar las reglas para el cálculo de las varianzas σ^2 y σ_*^2, bajo condiciones controladas. Procedimos de la siguiente manera. Para cada uno de los modelos presentados en Tab. 6.1, se generaron diferentes realizaciones $X_N^{\mathrm{k}} = \left\{\mathbf{x}^{\mathrm{k}}[1], \mathbf{x}^{\mathrm{k}}[2], \ldots, \mathbf{x}_N^{\mathrm{k}}\right\}$ y $Z_N^{\mathrm{k}} = \left\{z^{\mathrm{k}}[1], z^{\mathrm{k}}[2], \ldots, z^{\mathrm{k}}[N]\right\}$, donde $\mathrm{k} = 1, 2, \ldots, 100$ indica el número de la realización, a partir de valores aleatorios de σ^2 y σ_*^2. Las funciones de densidad de probabilidad de estos parámetros se describieron anteriormente. Además, los vectores de estados iniciales se generaron a partir de la Ec. (6.11). Los parámetros restantes no se modificaron durante la simulación.

A continuación, estimamos las varianzas $\hat{\sigma}^{2\,\mathrm{k}}$ y $\hat{\sigma}_*^{2\,\mathrm{k}}$ para la k-ésima realización, a partir de Z_N^{k} y del respectivo MEEG, aplicando iterativamente las reglas (5.23) y (5.27) para la actualización de estos parámetros. Finalmente, calculamos el error de estimación para cada una de las realizaciones, relativo al correspondiente valor real. Para ello, empleamos las siguientes expresiones:

$$\text{error}_{\sigma^2}^{\mathrm{k}} = 100 \times \frac{\hat{\sigma}^{2\,\mathrm{k}} - \sigma^{2\,\mathrm{k}}}{\sigma^{2\,\mathrm{k}}}, \tag{6.14}$$

$$\text{error}_{\sigma_*^2}^{\mathrm{k}} = 100 \times \frac{\hat{\sigma}_*^{2\,\mathrm{k}} - \sigma^{2\,\mathrm{k}}}{\sigma^{2\,\mathrm{k}}}. \tag{6.15}$$

En la Fig. 6.4 presentamos los diagramas de cajas de los errores error_{σ^2} y $\text{error}_{\sigma_*^2}$, para cada uno de los modelos considerados. A su vez, en la Tab. 6.4 se reportan los valores del primer cuartil (Q_1), de la mediana y del tercer cuartil (Q_3) extraídos de los diagramas de cajas. Observamos que en promedio el método arrojó estimaciones sesgadas para todos los casos, salvo para la varianza σ^2 del modelo III cuya mediana es aproximadamente 5 % y para la varianza σ_*^2 del modelo I cuya mediana es aproximadamente 7 %. En los modelos II, III y IV el método subestimó la varianza σ_*^2, mostrando el peor desempeño para el modelo IV. Para el caso de la varianza

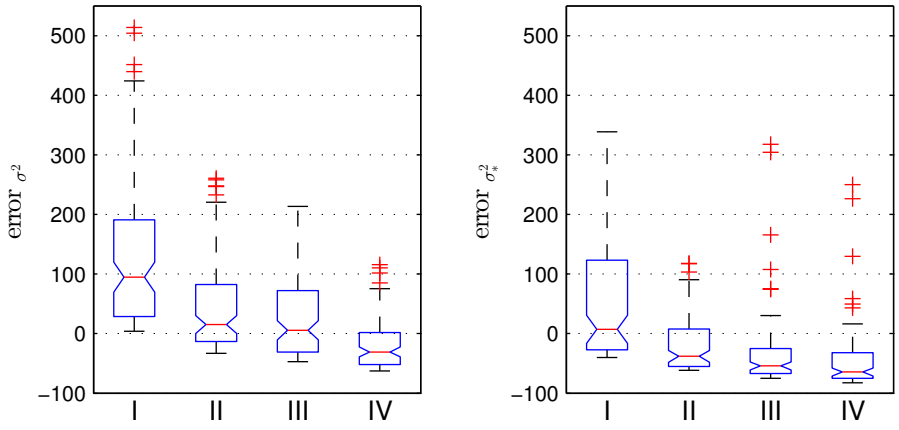


Figura 6.4: Diagramas de cajas de error_{σ^2} y $\text{error}_{\sigma_*^2}$ en la estimación de los parámetros en señales artificiales. Se utilizó un conjunto de 100 realizaciones.

Tabla 6.4: Error en la estimación de los parámetros en señales artificiales. Valores del primer cuartil ($Q1$), de la mediana y del tercer cuartil ($Q3$) de error_{σ^2} y de $\text{error}_{\sigma_*^2}$, para cada modelo considerado.

Modelo		I	II	III	IV
error_{σ^2}	Q_1	28,68	−13,37	−30,90	−52,01
	Mediana	94,72	15,30	5,49	−30,88
	Q_3	190,96	82,21	72,19	1,56
$\text{error}_{\sigma_*^2}$	Q_1	−27,26	−54,99	−67,05	−75,05
	Mediana	7,08	−38,17	−54,32	−64,37
	Q_3	123,40	7,56	−24,89	−31,91

σ^2, el método sobrestimó este parámetro en todos los casos, a excepción del modelo IV que arrojó una mediana negativa. El mejor desempeño para σ^2 se obtuvo con los modelos II y III. Por su parte, el modelo I obtuvo los valores mayores de error_{σ^2}.

Por otro lado, podemos observar también una gran dispersión en los errores obtenidos, lo que implica a su vez una gran dispersión en las estimaciones de las varianzas. Para ello, es suficiente con estudiar la distancia entre Q_1 y Q_3 en la Tab. 6.4. El peor desempeño correspondió a la estimación de σ^2 con el modelo I. Para este caso, se produjeron también los valores atípicos y el máximo más importantes. Comparando los diagramas de cajas, podemos afirmar que las estimaciones de σ^2 muestran una mayor dispersión, y como consecuencia mayor incerteza, en comparación con σ_*^2.

Los resultados anteriores prueban que el desempeño de las reglas (5.23) y (5.27) para la estimación de las varianzas no resultó completamente satisfactorio. En la mayoría de los casos se obtuvieron estimaciones sesgadas y con mucha incerteza. Afortunadamente, los métodos en espacio de estados son robustos a los errores en las estimaciones de los parámetros. Más adelante, presentaremos ejemplos con señales reales donde se muestra que se obtiene un buen desempeño para los métodos en espacio de estados al incluir estas estimaciones.

Tabla 6.5: Composición de los modelos estructurales considerados, de acuerdo a las dimensiones del espacio de estados p y del error de estados q. La dimensión de observación es $r = 1$.

MEEG	Componentes	p	q
$M_{(\mathrm{I})}$	$\mathcal{T}_1$	1	1
$M_{(\mathrm{II})}$	$\mathcal{T}_2$	2	1
$M_{(\mathrm{III})}$	$\mathcal{T}_3$	2	2
$M_{(\mathrm{IV})}$	$\mathcal{T}_3 + \mathrm{AR}_2$	4	3
$M_{(\mathrm{V})}$	$\mathcal{T}_3 + \mathrm{AR}_4$	6	3
$M_{(\mathrm{VI})}$	$\mathcal{T}_3 + \mathrm{AR}_6$	8	3
$M_{(\mathrm{VII})}$	$\mathcal{T}_3 + \mathrm{AR}_8$	10	3

6.6. Análisis estructural de SP y de SA reales

Pasaremos a describir los resultados obtenidos con el método de análisis estructural propuesto al aplicarlo en SP y SA reales, extraídas de la BDDV aplicando el procedimiento descripto en la Sec. 4.2. Nos concentraremos principalmente en la caracterización de los períodos, debido a que han recibido mayor atención por parte de los especialistas y a que hay disponible una mayor cantidad de información al respecto.

6.6.1. Secuencias de períodos

En este estudio tomamos en cuenta las propiedades observadas en SP reales y que listamos oportunamente en la Sec. 6.2.1. Con esto en mente, desarrollamos diferentes modelos estructurales para SP, contemplando las representaciones para la tendencia, el componente cíclico y las perturbaciones. En la Tab. 6.5, exponemos la composición de los modelos estructurales considerados en este estudio, ordenados en orden creciente de complejidad. Podemos observar que los tres primeros modelos sólo incorporan representaciones de la tendencia, mientras que los modelos restantes comprenden además componentes cíclicos de órdenes crecientes. De acuerdo a cómo definimos los modelos estructurales en la Sec. 6.3.2, todos los MEEG contemplan las perturbaciones en el error de observación.

Aplicación del análisis estructural

Explicaremos aquí, de forma general, la información extraída al aplicar el análisis estructural basado en métodos en espacio de estados a una SP real. En la Fig. 6.5, podemos observar una SP normal de un sujeto femenino. Observamos, también, las estimaciones de la tendencia calculadas con los métodos de filtrado $\hat{\mu}[n|n]$ y de suavizado $\hat{\mu}[n|N]$ de Kalman, considerando el modelo más simple $M_{(\mathrm{I})}$.

De esta figura, se desprende que la SP proviene de una vocal /a/ sostenida que presenta un período fundamental considerablemente estable. Podemos afirmar que las estimaciones de la tendencia, calculadas a partir del filtrado y el suavizado, son representaciones adecuadas del comportamiento global de la señal. Esto se puede apreciar fácilmente en la región ampliada para el intervalo (200, 260). Recordemos que esta componente acarrea información propia de un individuo, en particular su

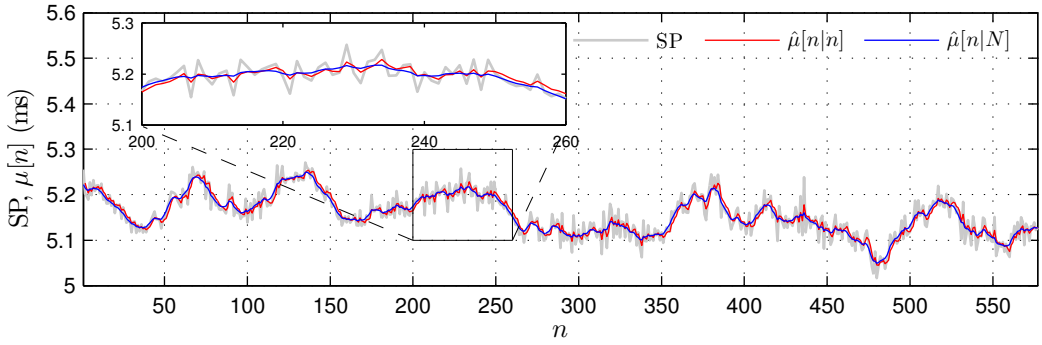


Figura 6.5: Análisis estructural de una SP normal, curva gris, correspondiente a una vocal /a/ sostenida de individuo femenino. Se presentan también las estimaciones de la tendencia obtenidas con el filtrado $\hat{\mu}[n|n]$, en color rojo, y el suavizado $\hat{\mu}[n|N]$, en color azul, de Kalman. Las estimaciones corresponden a $M_{(\mathrm{I})}$.

capacidad y habilidad para la fonación [16, 162, 163]. Como es sabido, $M_{(\mathrm{I})}$ es un modelo extremadamente limitado. Por ello, no permite representar satisfactoriamente las diferentes características de SP reales [138].

Una situación diferente puede apreciarse en la Fig. 6.6. En la parte superior, presentamos una SP normal correspondiente a otro sujeto femenino. Podemos observar que, aun cuando se extrajo de una vocal /a/ sostenida, esta señal presenta fluctuaciones importantes, es decir, oscilaciones de gran amplitud acompañadas con transiciones abruptas. Este comportamiento es característico de personas que no son capaces de mantener un período fundamental suficientemente estable [134]. Asimismo, ocurre como resultado de la acción de los mecanismos de control, tanto voluntarios como involuntarios, que actúan durante la fonación [163].

Aplicamos el análisis estructural para el estudio de esta SP, considerando el modelo $M_{(\mathrm{II})}$. Calculamos las estimaciones de la tendencia $\hat{\mu}[n|n]$ y $\hat{\mu}[n|N]$ empleando los métodos de filtrado y suavizado, respectivamente. Ambas estimaciones se muestran superpuestas en la fila superior de la Fig. 6.6. Analizando estas señales, podemos afirmar que el modelo considerado captura satisfactoriamente la dinámicas de las fluctuaciones. En la segunda fila podemos observar las estimaciones de la pendiente de la tendencia $\hat{\beta}[n|n]$ y $\hat{\beta}[n|N]$, calculadas aplicando el filtrado y el suavizado, respectivamente. Como era de esperarse por la definición (6.3), esta componente se comporta de forma similar a la derivada de la tendencia. Podemos apreciar que estas estimaciones ayudan a caracterizar la dinámica de las fluctuaciones en la SP, a la vez que sirven para anticipar las transiciones abruptas. Esto sugiere que la información de la pendiente podría ser de utilidad para estudiar la estabilidad de una emisión vocal sostenida.

En la fila inferior mostramos la estimación suavizada de las perturbaciones $\hat{\varepsilon}[n|N]$. Podemos observar que esta estimación presenta una dinámica aparentemente aleatoria. Esto insinuaría que los métodos propuestos permiten calcular satisfactoriamente la información de las perturbaciones. Sin embargo, estudiando a $\hat{\varepsilon}[n|N]$ con mayor precisión distinguimos algunas *espigas* que desarrollan un comportamiento aproximadamente regular. Esto nos sugiere que aún persiste cierta estructura en las perturbaciones que podría modelarse con el componente cíclico. Por último, en la Fig. 6.6 podemos apreciar un retardo temporal entre las estimaciones filtradas y

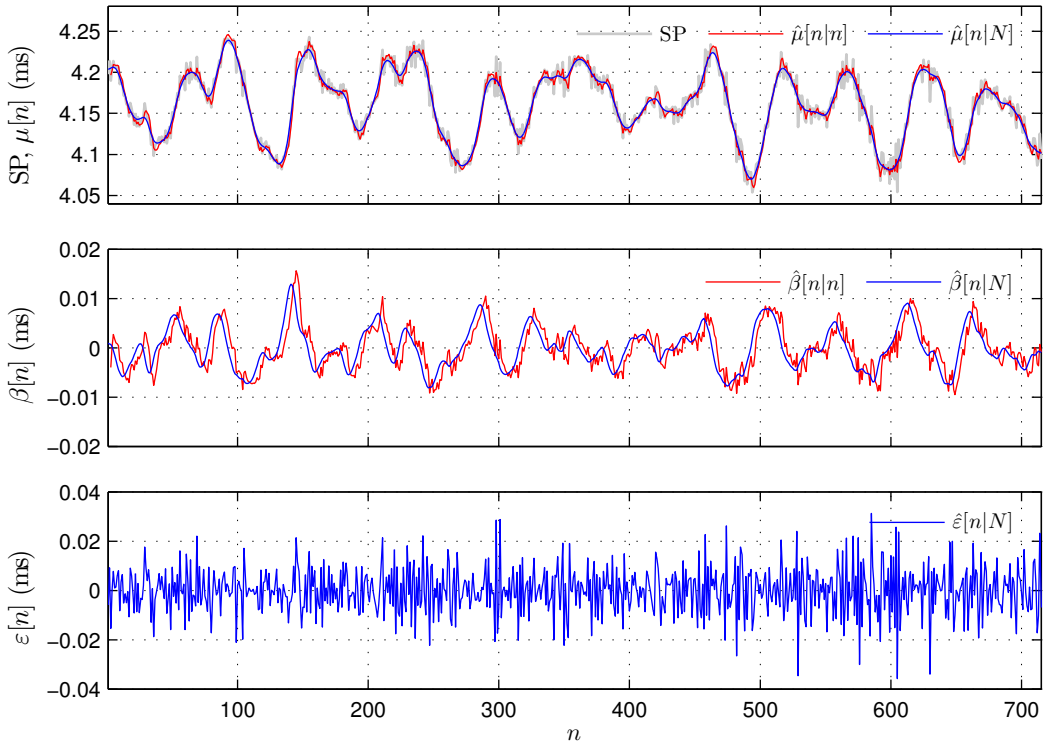


Figura 6.6: Análisis estructural de una SP considerando el modelo $M_{(\mathrm{II})}$. *Arriba:* SP normal extraída de una vocal /a/ sostenida emitida por un sujeto femenino sano. Se presentan superpuestas las estimaciones de la tendencia filtradas $\hat{\mu}[n|n]$ y suavizadas $\hat{\mu}[n|N]$. *Segunda fila:* Estimaciones filtradas $\hat{\beta}[n|n]$ y suavizadas $\hat{\beta}[n|N]$ de la pendiente estocástica de la tendencia. *Abajo:* Estimación suavizada de las perturbaciones $\hat{\varepsilon}[n|N]$.

suavizadas. Más adelante en esta sección, discutiremos respecto a este fenómeno.

Como último ejemplo, en la Fig. 6.7 mostramos las estimaciones calculadas con el análisis estructural, para una SP normal correspondiente a una vocal /a/ sostenida de un voluntario masculino. Para este ejemplo, empleamos el modelo $M_{(\mathrm{V})}$. En la fila superior presentamos la SP. Podemos apreciar que se caracteriza por una dinámica estable con pequeñas oscilaciones. Mostramos también las estimaciones de la tendencia $\hat{\mu}[n|n]$ y $\hat{\mu}[n|N]$, de forma superpuesta. En la segunda fila, presentamos las estimaciones de la pendiente estocástica de la tendencia $\hat{\beta}[n|n]$ y $\hat{\beta}[n|N]$. Analizando estas dos gráficas, podemos aseverar que la estabilidad de la SP bajo estudio se describe completamente a partir de la evaluación conjunta de estas estimaciones. Esto se debe a que, por un lado, las estimaciones de la tendencia describen la dinámica lenta y de largo alcance de la SP mientras que, por otro lado, la pendiente es prácticamente nula indicando que sólo se producen transiciones de pequeña magnitud. Esto último, sugiere además que la tendencia sigue un modelo $\mathcal{T}_1$.

En la tercera fila, observamos las estimaciones filtradas $\hat{\psi}[n|n]$ y suavizadas $\hat{\psi}[n|N]$ del componente cíclico. Ambas estimaciones corresponden a un modelo AR_4. A partir de éstas, podemos afirmar que la SP bajo estudio presenta un componente cíclico con dinámica autorregresiva. Es de esperar que este componente influya, en alguna medida, en la autocorrelación de esta SP. Este tipo de comportamiento ha

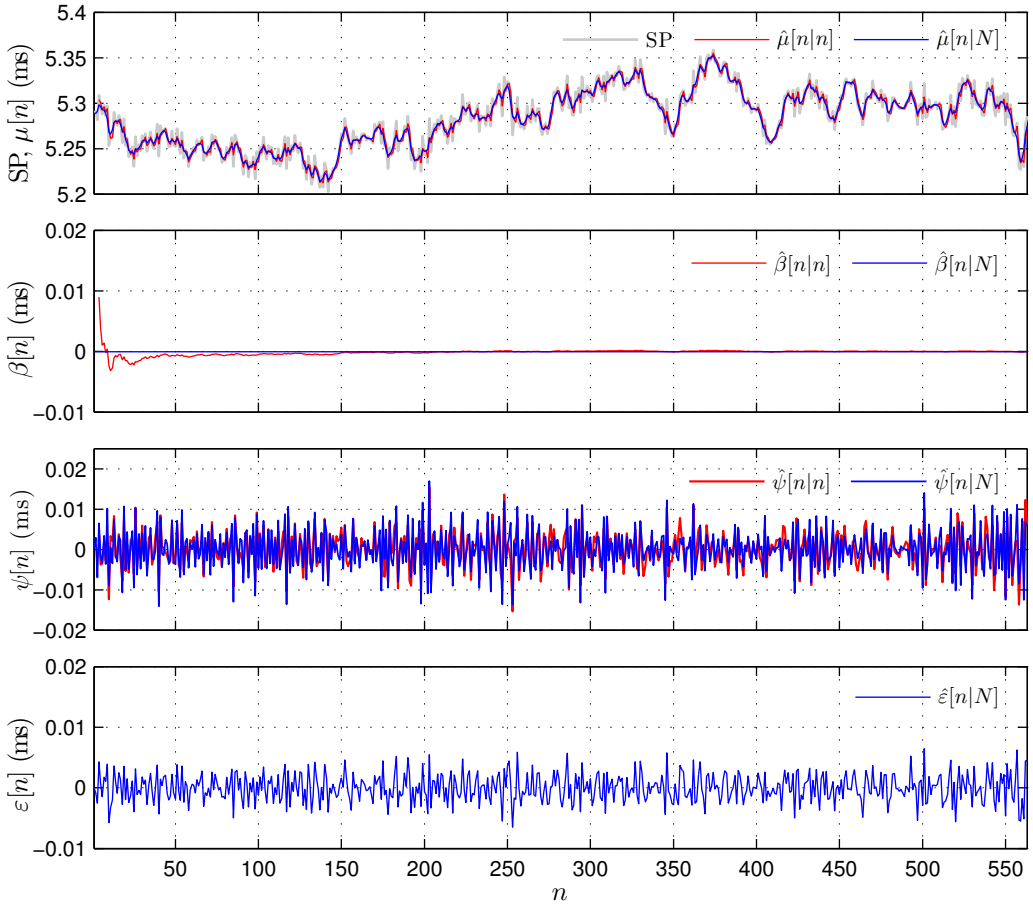


Figura 6.7: Análisis estructural de una SP considerando el modelo $M_{(\mathrm{V})}$. *Arriba:* SP extraída de una vocal /a/ sostenida normal de un sujeto masculino. Se presentan también las estimaciones de la tendencia filtradas $\hat{\mu}[n|n]$ y suavizadas $\hat{\mu}[n|N]$. *Segunda fila:* Estimaciones filtradas $\hat{\beta}[n|n]$ y suavizadas $\hat{\beta}[n|N]$ de la pendiente estocástica de la tendencia. *Tercera fila:* Estimaciones filtradas $\hat{\psi}[n|n]$ y suavizadas $\hat{\psi}[n|N]$ del componente cíclico, considerando un AR_4. *Abajo:* Estimación suavizada de las perturbaciones $\hat{\varepsilon}[n|N]$.

sido observado anteriormente en SP reales [138, 142]. Lo anterior, nos permite afirmar que esta información cíclica puede representarse adecuadamente con el análisis estructural propuesto. Los elementos restantes de los vectores de estados no aportan información nueva para el análisis ya que, de acuerdo a la Ec. (6.6), son versiones retardadas en el tiempo del componente cíclico.

En la fila inferior, presentamos la estimación suavizada de las perturbaciones $\hat{\varepsilon}[n|N]$. Comparando esta estimación con la mostrada en el ejemplo de la Fig. 6.6, podemos apreciar que se caracteriza por un comportamiento genuinamente aleatorio, es decir, no podemos distinguir en ella ninguna estructura o comportamiento regular. Lo anterior nos permite afirmar que $\hat{\varepsilon}[n|N]$ satisface las condiciones para ser considerada una estimación válida del *Jitter* de la SP.

Aplicamos el análisis estructural propuesto al estudio de todas las SP de la BDDV. Los resultados alcanzados mostraron, en general, características análogas a las expuestas en los ejemplos anteriores. Obtuvimos también comportamientos

individuales en algunas señales. Esto demuestra la necesidad de estudiar un conjunto mayor de SP, con el propósito de investigar la utilidad de la información extraída con el análisis estructural. Sin embargo, los resultados mostrados hasta aquí sugieren que el análisis estructural es una herramienta adecuada para modelar y estimar de forma óptima los diferentes componentes de SP reales. A su vez, podemos apreciar que tomar componentes estocásticos simples para construir el modelo estructural facilita la posterior interpretación de las diferentes dinámicas involucradas.

Las diferencias en el desempeño de los métodos de filtrado y suavizado de Kalman pueden apreciarse en las gráficas mostradas en las Figs. 6.6 y 6.7. En primer lugar, en la Fig. 6.6 podemos observar la naturaleza no causal de las estimaciones obtenidas con el suavizado. Debido a que en el filtrado sólo se emplea la información de las observaciones actual y pasadas, podemos apreciar un retardo temporal en las estimaciones calculadas con este método. En cambio, en el suavizado se agrega la información de observaciones futuras y, por ello, no se aprecia ningún retardo en las estimaciones suavizadas. En segundo lugar, el suavizado produce estimaciones más estables y menos fluctuantes que el filtrado, a expensas de un aumento en el costo computacional. Comentamos esto oportunamente en el Cap. 5, pero aquí podemos describirlo fácilmente con ayuda de los ejemplos presentados.

En tercer lugar, comparando las matrices de covarianza $\hat{\mathbf{P}}[n|n]$ y $\hat{\mathbf{P}}[n|N]$ generadas durante el filtrado y el suavizado, respectivamente, observamos que las estimaciones suavizadas se caracterizan por una menor varianza, lo que repercute en una mayor precisión, en comparación con las extraídas con el filtrado. Esto se puede apreciar en la Fig. 6.8, donde se presentan las varianzas de las estimaciones de los componentes μ, β y ψ del ejemplo de la Fig. 6.7. Aquí, $P_\mu[n] = \mathbf{P}_{(1,1)}[n]$ es la varianza de la tendencia μ, $P_\beta[n] = \mathbf{P}_{(2,2)}[n]$ es la varianza de la pendiente β y $P_\psi[n] = \mathbf{P}_{(3,3)}[n]$ es la varianza del componente cíclico ψ. Por último, los estados filtrados desarrollan un comportamiento transitorio inicial, necesario para estabilizar el proceso de estimación. Esto puede observase fácilmente en la segunda fila de la Fig. 6.7. Por el contrario, no se aprecia ninguna dinámica transitoria en los estados suavizados. Por todo esto, para el análisis estructural recomendamos escoger, preferentemente, los vectores de estados suavizados.

Análisis estadístico

El estudio de la sección anterior se basó principalmente en nuestra interpretación (subjetiva) de los resultados obtenidos. Describiremos ahora un análisis estadístico diseñado para estudiar objetivamente el desempeño para las SP de la BDDV del análisis estructural basado en métodos en espacio de estados.

En el Cap. 5, enunciamos las hipótesis fundamentales que rigen los MEEG. A su vez, en la primera parte de este capítulo describimos las hipótesis adicionales necesarias para los modelos estructurales desarrollados. En ambos casos, se supone que los errores de estados y de observación se comportan como procesos gaussianos, con media cero y covarianza constante, y que se caracterizan por ser procesos *blancos* independientes entre sí (ver Sec. 5.2.1). Recordando nuevamente que las observaciones son unidimensionales ($r = 1$), se define el error de predicción un paso hacia

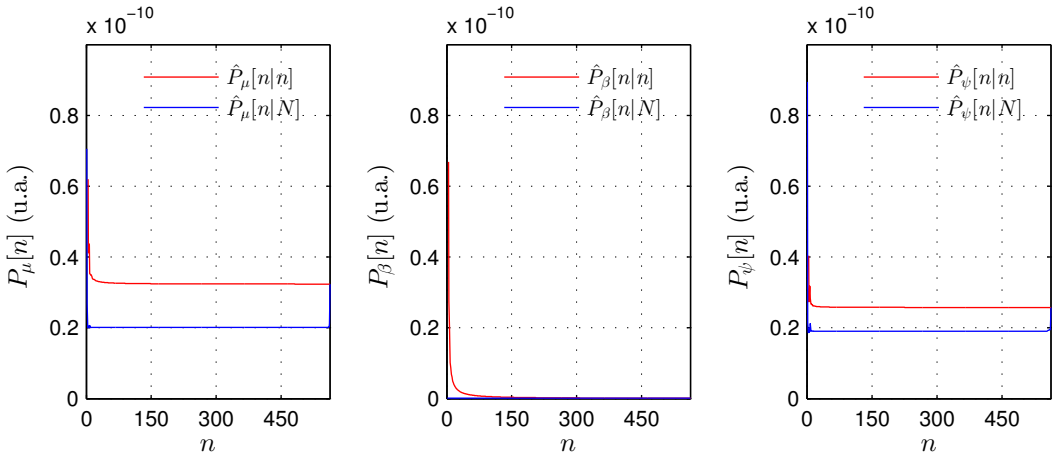


Figura 6.8: Varianzas de las estimaciones filtradas y suavizadas de la tendencia μ, de la pendiente β y del componente cíclico ψ para la SP de la Fig. 6.7. *Izq.:* Varianza de la tendencia $P_\mu[n]$. *Centro:* Varianza de la pendiente de la tendencia $P_\beta[n]$. *Der.:* Varianza del componente cíclico $P_\psi[n]$.

adelante normalizado[1] de la siguiente forma [35, 89]:

$$\tilde{e}[n] = \frac{\tilde{y}[n]}{\tilde{F}[n]^{1/2}}, \tag{6.16}$$

donde $\tilde{y}$ es la inferencia y $\tilde{F}$ su varianza correspondiente. Estos dos elementos se calculan como resultado intermedio en el filtrado de Kalman (ver Tab. 5.1).

Considerando como válidas las hipótesis anteriores, se desprende que el error normalizado $\tilde{e}$ posee un comportamiento estocástico gaussiano, con media cero, varianza unitaria y *blanco*. Por lo tanto, $\tilde{e}$ se convierte en un instrumento útil para evaluar la calidad de la representación de una SP real obtenida con el análisis estructural guiado por un MEEG. Teniendo en cuenta esto, procedimos a estudiar estadísticamente el desempeño del método propuesto para las SP extraídas de la BDDV. Para ello, aplicamos el análisis estructural a cada SP y calculamos el error $\tilde{e}$ de acuerdo con la Ec. (6.16). Seleccionamos el tercio central de cada $\tilde{e}$, correspondiente a 1 s de señal aproximadamente, ya que la mayoría de las señales de voz presentaban su comportamiento más estable en esta porción. Esta es una práctica empleada habitualmente [41]. El objetivo consistió en evaluar la hipótesis nula de que cada error $\tilde{e}$ constituía un proceso aleatorio gaussiano, con media cero y varianza unitaria, homocedástico (varianza constante) y *blanco*. Hasta donde sabemos, no existe una prueba estadística que examine conjuntamente todas estas hipótesis y, por ello, estudiamos cada una de las hipótesis por separado [35, 51].

Empleamos diferentes pruebas estadísticas para estudiar los errores $\tilde{e}$. En la Tab. 6.6, exponemos los resultados arrojados por estas pruebas, para cada MEEG en la Tab. 6.5. En primer lugar, aplicamos el test χ^2 no paramétrico [93]. Tomamos como hipótesis nula que el error $\tilde{e}$ posee distribución gaussiana, en contraposición a la alternativa de que no sea así. En la segunda columna de la Tab. 6.6, informamos el porcentaje de señales en la BDDV que fallaron en rechazar la hipótesis nula, luego

[1] En inglés, "standard one-step ahead forecast error".

Tabla 6.6: Análisis estadístico del desempeño del análisis estructural para SP reales. En cada columna se indica el porcentaje de SP de la BDDV que fallaron en rechazar las hipótesis nulas de gaussianidad (test χ^2), de media cero (test t), de homocedasticidad (test H) y de blancura (test LB) para diferentes MEEG. Para todas las pruebas se escogió un valor de significancia $\alpha = 0{,}05$.

MEEG	χ^2	t	H	LB
$M_{(\mathrm{I})}$	86,79 %	100,00 %	86,79 %	41,51 %
$M_{(\mathrm{II})}$	90,57 %	100,00 %	81,13 %	47,17 %
$M_{(\mathrm{III})}$	88,68 %	100,00 %	83,02 %	50,94 %
$M_{(\mathrm{IV})}$	86,79 %	100,00 %	86,79 %	71,70 %
$M_{(\mathrm{V})}$	90,57 %	100,00 %	88,68 %	79,25 %
$M_{(\mathrm{VI})}$	86,79 %	100,00 %	84,91 %	84,91 %
$M_{(\mathrm{VII})}$	90,57 %	100,00 %	84,91 %	86,79 %

de aplicar el análisis estructural. Estos resultados sugieren que la hipótesis de que el error posee un comportamiento gaussiano es adecuada para la gran mayoría de las señales y para los diferentes MEEG considerados.

En segundo lugar, estudiamos la suposición de que la media del error $\tilde{e}$ es cero. Para ello, aplicamos el tradicional test t, considerando como hipótesis nula que la media es cero y como alternativa que este valor es distinto de cero [115]. En la tercera columna de la Tab. 6.6, informamos el porcentaje de señales en la BDDV que fallaron en rechazar esta hipótesis nula. Podemos deducir que, para cada error $\tilde{e}$ gaussiano, no existe suficiente evidencia que permita refutar la hipótesis nula. Por ello, consideramos que la hipótesis de que la media es cero es apropiada.

En tercer lugar, evaluamos las varianzas de los errores $\tilde{e}$ aplicando el test H, descripto en [35, 76]. Esta prueba establece como hipótesis nula la homocedasticidad (varianza constante) de los errores $\tilde{e}$, y como alternativa la heterocedasticidad (varianza variable en el tiempo). En la cuarta columna de la Tab. 6.6, informamos el porcentaje de señales en la BDDV que fallaron en rechazar esta hipótesis nula, para los diferentes MEEG considerados. Debido a que en la mayoría de los casos no se rechazó la hipótesis nula, los resultados respaldan la hipótesis de homocedasticidad para los errores $\tilde{e}$.

Por último, estudiamos la hipótesis que el error normalizado $\tilde{e}$ es un proceso aleatorio *blanco*. Para ello, empleamos el test LB propuesto por Ljung y Box [26, 51]. Esta prueba considera como hipótesis nula la blancura de $\tilde{e}$, en contraposición a la alternativa de que no sea así. En la quinta columna de la Tab. 6.6, reportamos el porcentaje de señales en la BDDV que fallaron en rechazar la hipótesis nula. Para el caso de los MEEG $\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, M_{(\mathrm{III})}\right\}$ más de la mitad de los errores rechazaron la hipótesis nula. Inferimos que esto se debe a la existencia de alguna forma de correlación temporal que no es modelada adecuadamente por estos modelos y, por ello, se preserva en el error $\tilde{e}$. Sin embargo, la situación fue diferente para los MEEG $\left\{M_{(\mathrm{IV})},\, M_{(\mathrm{V})},\, M_{(\mathrm{VI})}, M_{(\mathrm{VII})}\right\}$. En estos casos, agregar el componente cíclico permitió mejorar la capacidad de representación en los modelos y, por ello, aumentó considerablemente el porcentaje de señales que fallaron en rechazar la hipótesis nula. De lo expuesto hasta aquí, se desprende que es importante considerar el componente

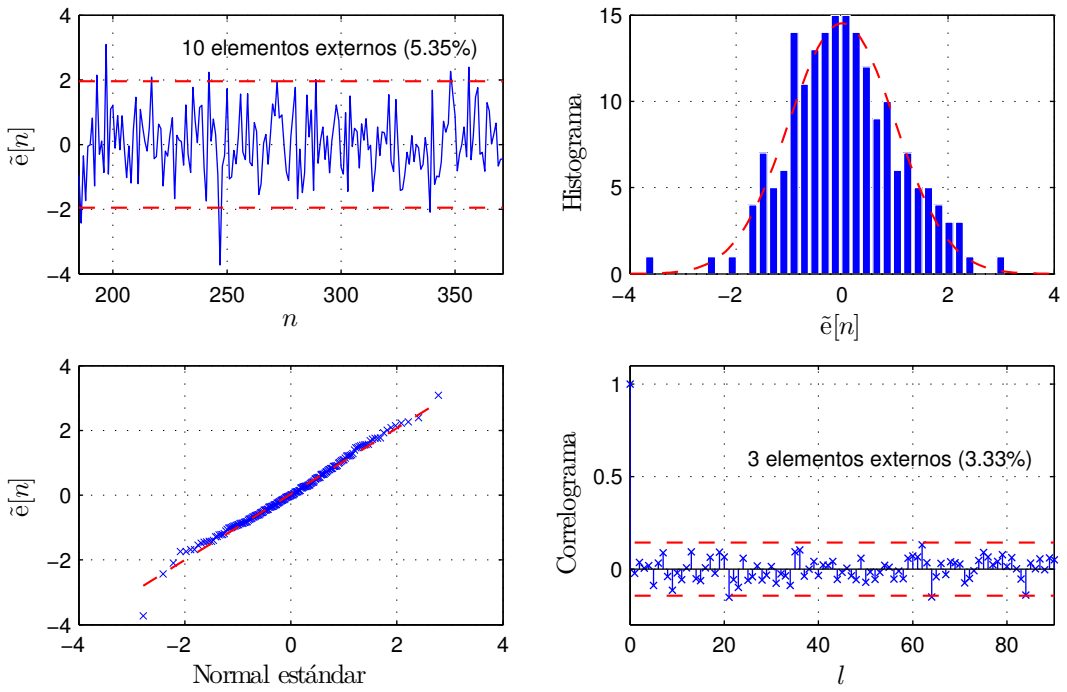


Figura 6.9: Método gráfico para evaluar la bondad del análisis estructural. La información corresponde al tercio central de la SP analizada en la Fig. 6.7. *Arriba izq.:* error normalizado $\tilde{e}$ y el intervalo de confianza del 95 % para un proceso $\mathcal{N}(0,1)$. *Arriba der.:* histograma de $\tilde{e}$ y función densidad de probabilidad para $\mathcal{N}(0,1)$. *Abajo izq.:* gráfico de normalidad de $\tilde{e}$. *Abajo der.:* correlagrama de $\tilde{e}$ y el intervalo de confianza del 95 % para procesos estocásticos blancos.

cíclico en el modelado estructural. Asimismo, podemos aseverar que la hipótesis de blancura es decisiva a la hora de evaluar el desempeño del análisis estructural en SP reales.

Con el propósito de respaldar el estudio estadístico anterior, empleamos también métodos gráficos para evaluar la bondad del análisis estructural. Para ello, construimos las representaciones gráficas para el error $\tilde{e}$ que se muestran en la Fig. 6.9. Esta información corresponde al tercio central de la SP analizada en la Fig. 6.7, empleando el $M_{(\mathrm{V})}$. En la gráfica superior izquierda, presentamos el error normalizado $\tilde{e}$. En línea discontinua, identificamos el intervalo de confianza del 95 % para un proceso con distribución $\mathcal{N}(0,1)$. En este caso sólo el 5,35 % de la señal (10 elementos) quedaron por fuera del intervalo de confianza.

En la región superior derecha observamos el histograma de $\tilde{e}$ y, en línea discontinua, la función de densidad de probabilidad teórica para $\mathcal{N}(0,1)$. Apreciamos que el histograma sigue adecuadamente el comportamiento teórico. En la región inferior izquierda, presentamos un gráfico de normalidad en el que se compara cómo se distribuyen los elementos de $\tilde{e}$ con los correspondientes a la distribución normal estándar $\mathcal{N}(0,1)$. La recta superpuesta en línea discontinua indica la situación favorable en la que los comportamientos coinciden. Estudiando conjuntamente estas tres gráficas, podemos aseverar que el comportamiento del error normalizado $\tilde{e}$ es similar a la distribución normal estándar, con excepción de algunos valores extremos en ambas colas.

Tabla 6.7: Porcentaje de SP en la BDDV correctamente modeladas con el análisis estructural propuesto, para diferentes MEEG. Se consideraron los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$.

MEEG	$\alpha = 0{,}05$	$\alpha = 0{,}01$
$M_{(\mathrm{I})}$	32,08 %	45,28 %
$M_{(\mathrm{II})}$	37,74 %	58,49 %
$M_{(\mathrm{III})}$	41,51 %	54,72 %
$M_{(\mathrm{IV})}$	54,72 %	73,58 %
$M_{(\mathrm{V})}$	66,04 %	81,13 %
$M_{(\mathrm{VI})}$	64,15 %	81,13 %
$M_{(\mathrm{VII})}$	69,81 %	81,13 %

En la región inferior derecha presentamos el correlograma de $\tilde{e}$, para los retardos $l = 0,\ 1,\ 2,\ \ldots\ , 90$. Indicamos, en línea discontinua, el intervalo de confianza del 95 % para un proceso aleatorio gaussiano *blanco*. Aquí, sólo el 3,33 % del correlograma (3 elementos) quedó por fuera del intervalo de confianza. Esto nos permite afirmar que el error normalizado $\tilde{e}$ puede considerarse un proceso estocástico *blanco*. En resumen, el método gráfico confirmó que la SP bajo estudio se representó correctamente con el análisis estructural, empleando el $M_{(\mathrm{V})}$. Este procedimiento resultó muy útil para entender con mayor profundidad los resultados arrojados por el análisis estadístico, mostrado en la Tab. 6.6.

En la Tab. 6.7, informamos el porcentaje de señales en la BDDV para las cuales se falló en rechazar simultáneamente las hipótesis nulas de todas las pruebas estadísticas de la Tab. 6.6, para los diferentes MEEG considerados. En este estudio, escogimos los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$, siendo el primer nivel más restrictivo que el segundo. Podemos apreciar que al incrementar la complejidad del MEEG, volviéndolo de este modo más flexible, pudieron representarse correctamente más señales de la BDDV. Sin embargo, esto trae consigo un aumento en los costos computacionales de los métodos en espacio de estados, en particular en la estimación de los parámetros del MEEG.

A su vez, al incrementar la complejidad de los MEEG, podemos observar una marcada diferencia en el desempeño para los modelos $M_{(\mathrm{III})}$, $M_{(\mathrm{IV})}$ y $M_{(\mathrm{V})}$, la cual no se percibe para los modelos restantes. Es decir, se produce un aumento marcado en el porcentaje de señales representadas correctamente entre los modelos $M_{(\mathrm{III})}$ y $M_{(\mathrm{IV})}$, y entre los modelos $M_{(\mathrm{IV})}$ y $M_{(\mathrm{V})}$. Entre los restantes modelos las diferencias son menores. Esto último nos permite aseverar que en la mayoría de las SP analizadas existe un componente cíclico que no pudo modelarse con los MEEG más simples. Sin embargo, esta dinámica pudo representarse adecuadamente empleando MEEG más flexibles. Esto demuestra la importancia de considerar el componente cíclico en el modelado estructural.

En el estudio anterior, evaluamos el desempeño individual de cada MEEG. Sin embargo, observamos que en ocasiones algunas SP fueron modeladas correctamente con un modelo, pero esta representación se estropeó al modificar el modelo (incluso escogiendo un MEEG más flexible). Tomando en cuenta esto, llevamos a cabo un experimento complementario. En la Tab. 6.8, informamos el porcentaje *acumulado* de SP modeladas correctamente con respecto al aumento en la complejidad en con-

Tabla 6.8: Porcentaje acumulado de SP en la BDDV correctamente modeladas con el análisis estructural, para valores crecientes de complejidad $(p + q + r)$ en conjuntos de MEEG. Se escogieron los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$. En negrita se indica el máximo porcentaje alcanzado de SP correctamente modeladas.

Complejidad	Conjunto de MEEG	$\alpha = 0{,}05$	$\alpha = 0{,}01$
3	$\left\{M_{(\mathrm{I})}\right\}$	32,08 %	45,28 %
4	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})}\right\}$	45,28 %	64,15 %
5	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})}, M_{(\mathrm{III})}\right\}$	47,17 %	64,15 %
8	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{IV})}\right\}$	66,04 %	77,36 %
10	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{V})}\right\}$	73,58 %	84,91 %
12	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{VI})}\right\}$	75,47 %	84,91 %
14	$\left\{\mathbf{M}_{(\mathrm{I})},\, \mathbf{M}_{(\mathrm{II})},\, \ldots,\, \mathbf{M}_{(\mathrm{VII})}\right\}$	**75,47** %	**86,79** %

juntos de MEEG. En este estudio, escogimos los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$. En el contexto de este experimento, la complejidad se definió como la máxima dimensión total $(p+q+r)$ para los conjuntos de modelos descriptos en la segunda columna de la Tab. 6.8. Consideramos que este parámetro representa, de forma indirecta, la flexibilidad máxima de cada conjunto para representar una señal.

Como era de esperarse, el porcentaje de SP en la BDDV correctamente modeladas para cada conjunto de MEEG es mayor que el obtenido para los modelos por separado (comparar las Tabs. 6.7 y 6.8). La razón de esto, es que se fueron acumulando las SP correctamente modeladas, sin importar cuál de los modelos del conjunto resultó adecuado. Además, podemos apreciar nuevamente la marcada diferencia en el comportamiento para los niveles de complejidad 5, 8 y 10, debido a la incorporación del componente cíclico AR en la estructura de los MEEG.

Por otro lado, en la última fila de la Tab. 6.8 indicamos el porcentaje total de las SP de la BDDV correctamente modeladas con el análisis estructural desarrollado. Podemos apreciar que más del 75 %, considerando $\alpha = 0{,}05$, o más del 86 %, considerando $\alpha = 0{,}01$, de las señales fueron representadas correctamente. Estos resultados nos permiten aseverar que el análisis estructural basado en métodos en espacio de estados es una estrategia adecuada para modelar SP reales.

6.6.2. Secuencias de amplitudes

En la Sec. 6.2.1 describimos las principales características observadas en SP reales. Explicamos, a su vez, que es escasa la información disponible respecto al comportamiento de SA reales. Por ello, propusimos trabajar bajo la hipótesis de que estas señales presentan características similares a las identificadas en SP. Esto último sugiere que el análisis estructural propuesto aquí sería también adecuado para el estudio de SA reales. En esta sección, presentaremos los resultados alcanzados al aplicar este método en SA extraídas de la BDDV. En estos estudios consideramos nuevamente los modelos estructurales de la Tab. 6.5.

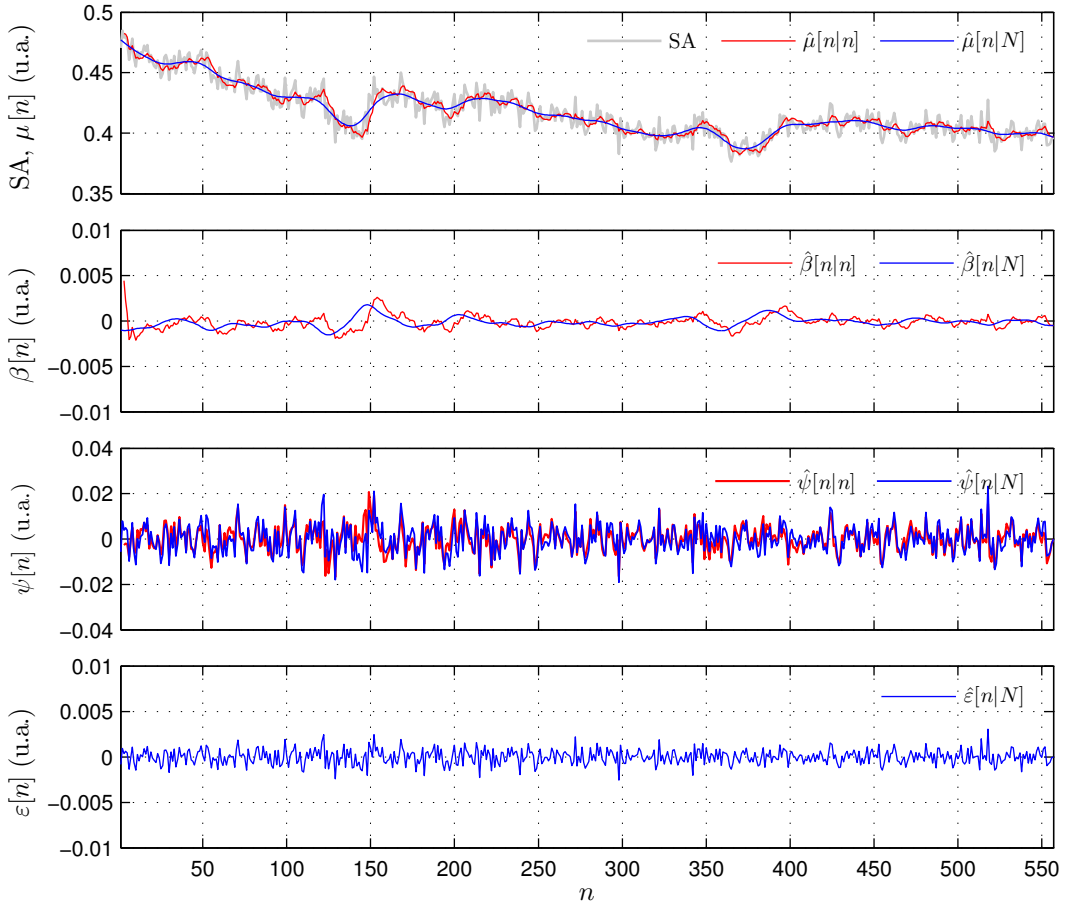


Figura 6.10: Análisis estructural de una SA considerando el modelo $M_{(\mathrm{V})}$. *Arriba:* SA extraída de una vocal /a/ sostenida normal de un sujeto masculino. Se presentan también las estimaciones de la tendencia filtradas $\hat{\mu}[n|n]$ y suavizadas $\hat{\mu}[n|N]$. *Segunda fila:* Estimaciones filtradas $\hat{\beta}[n|n]$ y suavizadas $\hat{\beta}[n|N]$ de la pendiente estocástica de la tendencia. *Tercera fila:* Estimaciones filtradas $\hat{\psi}[n|n]$ y suavizadas $\hat{\psi}[n|N]$ del componente cíclico, considerando un AR_4. *Abajo:* Estimación suavizada de las perturbaciones $\hat{\varepsilon}[n|N]$.

Aplicación del análisis estructural

En la Fig. 6.10, presentamos los componentes estimados para una SA, obtenida de una vocal /a/ sostenida de un sujeto masculino normal, aplicando el análisis estructural y considerando el modelo $M_{(\mathrm{V})}$. Esta SA corresponde al mismo segmento de vocal sostenida de la que se extrajo la SP de la Fig. 6.7. Como era de esperarse, las estimaciones calculadas en este caso muestran dinámicas con características similares a las descriptas anteriormente, en los ejemplos para SP.

Observamos que el comportamiento a escala global de la SA es representado adecuadamente por las estimaciones filtradas $\hat{\mu}[n|n]$ y suavizadas $\hat{\mu}[n|N]$ de la tendencia. Al mismo tiempo, las transiciones acentuadas se manifiestan en las estimaciones de la pendiente estocástica de la tendencia $\hat{\beta}[n|n]$ y $\hat{\beta}[n|N]$ calculadas mediante el filtrado y el suavizado, respectivamente. Podemos apreciar que las magnitudes de estas estimaciones son pequeñas, lo que indica que la SA bajo estudio es estable. Lo mismo concluimos al analizar la SP correspondiente.

Tabla 6.9: Análisis estadístico del desempeño del análisis estructural para SA reales. En cada columna se indica el porcentaje de SA de la BDDV que fallaron en rechazar las hipótesis nulas de gaussianidad (test χ^2), de media cero (test t), de homocedasticidad (test H) y de blancura (test LB) para diferentes MEEG. Para todas las pruebas se escogió un valor de significancia $\alpha = 0{,}05$.

MSSG	χ^2	t	H	LB
$M_{(I)}$	88.68 %	98.11 %	81.13 %	56.60 %
$M_{(II)}$	98.11 %	100.00 %	83.02 %	45.28 %
$M_{(III)}$	96.23 %	100.00 %	83.02 %	66.04 %
$M_{(IV)}$	92.45 %	100.00 %	83.02 %	75.47 %
$M_{(V)}$	94.34 %	100.00 %	84.91 %	75.47 %
$M_{(VI)}$	96.23 %	100.00 %	79.25 %	83.02 %
$M_{(VII)}$	96.23 %	100.00 %	84.91 %	83.02 %

Por otro lado, las estimaciones filtradas $\hat{\psi}[n|n]$ y suavizadas $\hat{\psi}[n|N]$ del componente cíclico representan la dinámica autorregresiva de la SA, considerando un modelo AR_4. Nuevamente, observamos en estas estimaciones cierta regularidad, caracterizada principalmente por el comportamiento de las *espigas*. Por último, la información del componente genuinamente estocástico asociado a las perturbaciones puede apreciarse en la estimación suavizada $\hat{\varepsilon}[n|N]$ del error de observación. Estudiando esta serie temporal, apreciamos que satisface las condiciones necesarias para ser considerada una estimación válida del *Shimmer* en la SA.

Es notable la diferencia de magnitud entre la tendencia y los restantes componentes, cuya excursión es inferior al 1 % del nivel de la tendencia. Esto mismo pudo observarse en los ejemplos para SP analizados anteriormente. Esto último pone en evidencia una de la principales dificultades al querer estimar el *Jitter* y el *Shimmer* [16, 138]. Sin embargo, los resultados mostrados hasta aquí sugieren que el análisis estructural basado en métodos en espacio de estados podría ser de mucha ayuda en esta tarea.

Análisis estadístico

De forma análoga a como procedimos anteriormente con las SP, evaluamos la calidad de la representación alcanzada con el análisis estructural para las SA reales de la BDDV. Para cada SA, aplicamos el análisis estructural y, seguidamente, calculamos el error normalizado $\tilde{\mathrm{e}}[n]$ con la Ec. (6.16). Seleccionamos el tercio central del error $\tilde{\mathrm{e}}[n]$, por las mismas razones ya enunciadas, y con esta información llevamos a cabo el análisis estadístico, empleado oportunamente para SP. En la Tab. 6.9, informamos el porcentaje de SA de la BDDV que fallaron en rechazar las hipótesis nulas de gaussianidad en la segunda columna, de media cero en la tercera columna, de homocedasticidad en la cuarta columna, y de blancura en la quinta columna, para los diferentes MEEG de la Tab. 6.5.

A partir de estos resultados, podemos observar que la gran mayoría de las SA de la BDDV fallaron en rechazar las hipótesis nulas de normalidad, de media cero y de homocedasticidad, independientemente del modelo seleccionado. Esto sugiere que estas hipótesis pueden considerarse válidas para la mayoría de los erro-

Tabla 6.10: Porcentaje de SA en la BDDV correctamente modeladas con el análisis estructural propuesto, para diferentes MEEG. Se consideraron los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$.

MEEG	$\alpha = 0{,}05$	$\alpha = 0{,}01$
$M_{(I)}$	37.74 %	66.04 %
$M_{(II)}$	37.74 %	60.38 %
$M_{(III)}$	52.83 %	71.70 %
$M_{(IV)}$	58.49 %	79.25 %
$M_{(V)}$	62.26 %	83.02 %
$M_{(VI)}$	64.15 %	86.79 %
$M_{(VII)}$	69.81 %	84.91 %

res $\tilde{e}[n]$. Sin embargo, nos enfrentamos a una situación diferente para la hipótesis de blancura. Para los modelos $\left\{M_{(\mathrm{I})}, M_{(\mathrm{II})}\right\}$, cuya estructura contempla las formas más simples de tendencia, aproximadamente la mitad de las señales rechazaron la hipótesis de blancura. Esto implica que las SA poseen una estructura cuya dinámica no puede representarse con estos modelos. Sin embargo, para los modelos $\left\{M_{(\mathrm{III})}, M_{(\mathrm{IV})}, M_{(\mathrm{V})}, M_{(\mathrm{VI})}, M_{(\mathrm{VII})}\right\}$ la situación se invierte. Al emplear una forma de tendencia más flexible y al agregar el componente cíclico más del 60 % de las señales de la BDDV fallaron en rechazar la hipótesis de blancura. A su vez, este porcentaje creció a medida que aumentó el orden del modelo AR del componente cíclico. Esto indica que este componente es muy importante en el análisis estructural de las SA.

A continuación, consideramos de forma conjunta todas las hipótesis con el propósito de evaluar la calidad de la representación para SA alcanzada con el análisis estructural. En la Tab. 6.10, informamos el porcentaje de SA en la BDDV para las cuales se falló en rechazar simultáneamente las hipótesis nulas de todas las pruebas estadísticas de la Tab. 6.9, para los diferentes MEEG considerados. En este estudio, empleamos los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$. Encontramos que para los modelos que incluyen el componente cíclico más de la mitad de las SA analizadas fallaron en rechazar todas las hipótesis. Luego, estos resultados sugieren que estas señales se representaron adecuadamente con nuestro método. Podemos observar, además, que al emplear un MEEG más flexible aumenta el porcentaje de señales correctamente representadas. Esto se cumple para ambos valores de α. Es importante destacar que estos porcentajes resultaron muy similares a los obtenidos para SP, expuestos en la Tab. 6.7.

Al igual que ocurrió con las SP, encontramos casos donde una SA fue correctamente representada aplicando un MEEG en particular, pero al cambiar de modelo se deterioró el ajuste de la señal. Que suceda esto al escoger un modelo más simple es entendible. Sin embargo, observamos esta situación incluso al trabajar con modelos con una estructura más flexible, lo que atrajo nuestra atención. Teniendo en cuenta esto, implementamos nuevamente el experimento complementario de la Sec. 6.6.1.

En la Tab. 6.11, informamos el porcentaje *acumulado* de SA en la BDDV modeladas correctamente con el análisis estructural, en función del aumento en la complejidad $(p+q+r)$ de los conjuntos de MEEG descriptos en la segunda columna. En este estudio, escogimos los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$.

Tabla 6.11: Porcentaje acumulado de SA en la BDDV correctamente modeladas con el análisis estructural propuesto, para valore crecientes de complejidad $(p + q + r)$ de los conjuntos de MEEG. Se escogieron los niveles de significancia estadística $\alpha = 0{,}05$ y $\alpha = 0{,}01$. En negrita se remarca el máximo porcentaje alcanzado de SA correctamente modeladas.

Complejidad	Conjunto de MEEG	$\alpha = 0{,}05$	$\alpha = 0{,}01$
3	$\left\{M_{(\mathrm{I})}\right\}$	37,74 %	66,04 %
4	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})}\right\}$	52,83 %	73,58 %
5	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})}, M_{(\mathrm{III})}\right\}$	56,60 %	75,47 %
8	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{IV})}\right\}$	67,92 %	83,02 %
10	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{V})}\right\}$	71,70 %	86,79 %
12	$\left\{M_{(\mathrm{I})},\, M_{(\mathrm{II})},\, \ldots,\, M_{(\mathrm{VI})}\right\}$	75,47 %	88,68 %
14	$\left\{\mathbf{M}_{(\mathrm{I})},\, \mathbf{M}_{(\mathrm{II})},\, \ldots,\, \mathbf{M}_{(\mathrm{VII})}\right\}$	**77,36** %	**88,68** %

Como era de esperarse, observamos que los porcentajes de señales correctamente representadas resultaron superiores a los obtenidos en el estudio anterior, donde analizamos cada modelo individualmente (comparar las Tabs. 6.10 y 6.11). Explicamos oportunamente cuáles eran, para nosotros, las razones de este comportamiento cuando presentamos el estudio estadístico para SP.

Al momento de redactar este documento, no contamos con una explicación totalmente satisfactoria para el deterioro de las representaciones con el análisis estructural, tanto para SP como para SA, al aumentar la complejidad del MEEG. Sin embargo, consideramos que este problema está ligado a las deficiencias del método para la estimación de los parámetros, ya que cualquier error en las estimaciones puede afectar, posteriormente, la representación de una señal. Además, es esperable que al aumentar la complejidad del MEEG, y con ello su flexibilidad, lo mismo ocurra con el error de estimación. Esta suposición se fundamenta en que el problema de optimización a resolver se torna más complejo y, además, en que crece el dominio de definición $\mathcal{D}$ donde se busca la solución óptima.

El experimento anterior también nos permitió comparar, de forma indirecta, las dinámicas de las SP y SA de la BDDV. Para ello, estudiamos los porcentajes de señales correctamente representadas para los niveles de complejidad {3, 4, 5} en las Tabs. 6.8 y 6.11. Podemos observar que el porcentaje para SA es mayor que el correspondiente para SP, para ambos niveles de significancia estadística. En particular, para los valores de complejidad {4, 5} el porcentaje es superior al 50 %. Todo esto sugiere que las SA en la BDDV desarrollan una dinámica más simple que las SP y, por ello, fue suficiente con emplear modelos de tendencia para representar más de la mitad de las señales. Esta hipótesis amerita ser estudiada en detalle en trabajos futuros.

Finalmente, los resultados de la última fila de la Tab. 6.11 coinciden con el porcentaje total de SA de la BDDV correctamente representadas aplicando el análisis estructural basado en métodos en espacio de estados. Podemos apreciar que más del 77 %, con $\alpha = 0{,}05$, o más del 88 %, con $\alpha = 0{,}01$, de las señales fueron representadas correctamente. Paralelamente, estos porcentajes resultaron muy similares a los obtenidos oportunamente para SP (comparar las Tabs. 6.8 y 6.11). Esto nos permite

concluir que el método propuesto en este capítulo también resulta adecuado para representar SA reales y extraer así información de interés.

6.7. Comentarios finales

Destinamos este capítulo a presentar el segundo de los aportes originales de esta tesis de doctorado. Éste consistió en el desarrollo de una estrategia novedosa, basada en los métodos en espacio de estados, para el análisis estructural de las series de períodos y de amplitudes extraídas de vocales sostenidas.

A lo largo del desarrollo expuesto en este capítulo perseguimos dos objetivos principales. En primer lugar, pretendíamos construir modelos más realistas para la síntesis de series de períodos y de amplitudes, concentrándonos en generar estrategias que superen las limitaciones de las representaciones actuales de *Jitter* y *Shimmer*. Entre ellas, los modelos estocásticos de *Jitter* y *Shimmer* de nuestra autoría, presentados en el Cap. 4. Para ello, discutimos las características más importantes de estas señales, prestando especial atención a cómo se comportan en la realidad. Esto nos impulsó a investigar la manera de incorporar toda esta información en un marco teórico adecuado, aplicando reglas matemáticas y estadísticas. Como resultado, obtuvimos una familia de modelos estructurales desarrollada a partir de modelos en espacio de estados lineales y gaussianos.

En segundo lugar, buscábamos un método que resultara aplicable para el estudio de casos reales, permitiendo así extraer nueva información que ayude a describir objetivamente el comportamiento de estas señales. En este punto, los métodos en espacio de estados demostraron ser sumamente útiles, ya que nos permitieron implementar, de forma satisfactoria, el análisis de estas señales guiado por los modelos estructurales desarrollados. Mostramos que este enfoque facilita la estimación de los diferentes componentes de los modelos estructurales a partir de una señal real y, a su vez, permite construir estrategias para el cálculo de los parámetros desconocidos de estos modelos.

Evaluamos el comportamiento de la estrategia propuesta gracias a diferentes simulaciones y estudios. Las simulaciones nos permitieron mostrar que los modelos estructurales propuestos generan series temporales con un comportamiento complejo, similar al observado en las series de períodos y de amplitudes reales. Por otro lado, los estudios realizados nos permitieron exponer las bondades del análisis estructural para series de períodos y de amplitudes reales. Probamos que el método propuesto estima adecuadamente la información correspondiente a la tendencia, al componente cíclico y a las perturbaciones, de forma conjunta y óptima. Hasta donde sabemos, no existe en la actualidad un método similar que permita esto. Además, introdujimos diferentes pruebas objetivas para evaluar la calidad de la representación alcanzada con el análisis estructural desarrollado. En este contexto, demostramos que el componente cíclico cobra un rol muy importante en el modelo estructural. Los resultados alcanzados nos permiten concluir que el método propuesto resultó adecuado para el estudio de series de períodos y de amplitudes reales.

Las simulaciones sirvieron también para evaluar, bajo condiciones controladas, cómo se comportan los diferentes métodos considerados. Podemos afirmar que los métodos implementados se comportaron en general de forma satisfactoria. Sin embargo, es importante desarrollar nuevas y mejores estrategias para la estimación de los parámetros de los modelos en espacio de estados. Consideramos que esto permiti-

rá mejorar el desempeño de los modelos estructurales, generando como consecuencia mejores estimaciones de los diferentes componentes.

Los trabajos destinados al desarrollo y la evaluación del análisis estructural basado en métodos en espacio de estados, direccionado al estudio de series de períodos, dieron lugar a dos presentaciones en congresos de la especialidad, uno de alcance nacional [5] y otro latinoamericano [6]. Asimismo, este mismo material dio lugar a un artículo publicado en el *Journal of Voice* que, como su nombre lo indica, es una revista dirigida a la investigación clínica y experimental de todos lo aspectos específicos relacionados a la voz y a la fonación [8]. Por otro lado, los resultados correspondientes a las series de amplitudes no han sido publicados hasta el momento.

Capítulo 7

Modelado de la fonación y filtrado inverso de la voz aplicando métodos en espacio de estados

7.1. Introducción

Los métodos en espacio de estados han demostrado ser sumamente útiles en el procesamiento guiado por modelos de series temporales no estacionarias [30, 51, 86]. Dedicamos el Cap. 5 a describir estos métodos, prestando atención a aquellos que consideramos relevantes para el desarrollo de esta tesis de doctorado. Algunas de sus principales características son: (*i*) la formulación de los modelos es directa e intuitiva, (*ii*) existen herramientas analíticas y algorítmicas para la extracción de estimadores (estadísticos significativos) de los procesos involucrados, (*iii*) contemplan las incertezas y los errores cometidos en la modelización, y (*iv*) existen estrategias de optimización para el cálculo de los parámetros desconocidos de un modelo. Debido a que la fonación es un proceso complejo y dinámico, consideramos que los métodos en espacio de estados constituyen un marco adecuado para su estudio [43, 94, 111, 171].

En el Cap. 3, introdujimos dos importantes teorías desarrolladas con el propósito de describir e imitar artificialmente la fonación en los seres humanos. Ambos paradigmas explican las transformaciones involucradas, partiendo de un conjunto específico de hipótesis y aplicando conceptos de diferentes ciencias. De éstas, la teoría *fuente* y *filtro* (TFF) es, sin ninguna duda, la que históricamente ha tenido un mayor protagonismo, tanto en investigación como en las aplicaciones tecnológicas que involucran a la señal de voz [42, 130, 151]. Trabajando en el marco de la TFF, desarrollaremos aquí un modelo que permita analizar y simular la fonación, de forma precisa y flexible. A lo largo de este desarrollo, consideraremos también que la dinámica estocástica de la fonación puede describirse empleando modelos en espacio de estados lineales y gaussianos (MEEG).

En ese mismo capítulo, introdujimos el concepto de filtrado inverso, o también llamado descomposición de la voz, profundizando en su interpretación de acuerdo con la TFF. De forma concisa, consiste en calcular la función glótica procesando convenientemente una señal de voz con la intención de eliminar el efecto del tracto vocal [46, 47, 174]. De este modo, se obtienen como resultado estimaciones de la función glótica y de la información del tracto vocal [2, 7]. Desde luego, este procedimiento sólo es aplicable para fonemas sonoros caracterizados por la modulación

cíclica del flujo de aire al atravesar las cuerdas vocales (ver Sec. 2.2.2). Tomando en cuenta todo esto, en este capítulo presentaremos una estrategia para implementar el filtrado inverso empleando conjuntamente el modelo de la fonación desarrollado y los métodos en espacio de estados. Para ello, introduciremos en primer lugar un modelo alternativo de la función glótica formulado a partir de una ecuación en diferencias estocástica y no estacionaria.

7.2. Antecedentes

Desde que fuera propuesta por Fant, han surgido diferentes estrategias inspiradas en la TFF para implementar el filtrado inverso de la voz. A fines expositivos, podemos agruparlas en dos grandes conjuntos. Por un lado, encontramos los métodos en los que el filtrado inverso se implementa en forma secuencial [5, 46]. En primer lugar, se obtiene una representación del tracto vocal a partir de la señal de voz y, luego, se utiliza esta representación para procesar la señal de voz y calcular así la función glótica. En ocasiones, estas dos etapas se realizan en forma iterativa con el propósito de mejorar la calidad de los resultados [2, 4].

Como explicamos en la Sec. 3.4, el tracto vocal suele representarse empleando modelos (filtros) lineales. De esta forma, el comportamiento espectral del tracto vocal queda determinado paramétricamente con ayuda de un conjunto pequeño de coeficientes [42, 128, 174]. A lo largo del tiempo, se han propuesto diversas alternativas para el cálculo de estos coeficientes. En general, la mayoría se basan en el método de predicción lineal (LP) o en versiones mejoradas de éste. Para mayor información, el lector interesado puede recurrir a [3, 4, 6, 52, 103, 104, 108].

Por otro lado, existen estrategias para la descomposición de la voz en las que se calculan de forma conjunta el filtro del tracto vocal y la función glótica. Cronológicamente, éstas surgieron más recientemente que los métodos explicados arriba. En este capítulo, consideraremos esta forma de trabajo. Habitualmente, el cálculo conlleva la resolución de un problema inverso aplicando algoritmos de optimización o estrategias evolutivas en la búsqueda de la solución [17, 63, 67, 85, 135]. En estos casos, también se emplean filtros lineales para modelar la información del tracto vocal. Sin embargo, existen alternativas desarrolladas considerando tanto coeficientes constantes como variantes en el tiempo.

En general, para describir la función glótica suelen ajustarse modelos determinísticos o se aproxima esta señal aplicando combinaciones lineales de funciones características [17]. Actualmente, existe una gran variedad de plantillas de la función glótica [44, 57]. Sin duda alguna, la más utilizada es la función glótica propuesta por Liljencrants y Fant (LF) [67, 135]. Sin embargo, algunos autores emplean también otras representaciones más simples [84, 85]. Por ejemplo, Fu y Murphy aplicaron la función glótica de Rosenberg para generar valores iniciales de los parámetros del modelo LF, más complejo, para luego refinarlos iterativamente [63].

Los modelos de la función glótica descriptos en el párrafo anterior poseen dos importantes desventajas. En primer lugar, debido a su naturaleza determinística presentan serias limitaciones para capturar adecuadamente las aperiodicidades o comportamientos peculiares que suelen ocurrir a nivel glótico durante la fonación [45, 145]. Esto último evidencia la necesidad de nuevos modelos que permitan una representación más flexible del comportamiento de la función glótica.

La segunda desventaja involucra el cálculo de la función glótica en el filtrado

inverso. Como dijimos anteriormente, existen métodos muy potentes para calcular los parámetros del filtro del tracto vocal. Sin embargo, en general la estimación de la función glótica requiere del ajuste de funciones no lineales [63, 135]. Esto último aumenta considerablemente la complejidad del problema de optimización a resolver. Por ello, es importante generar estrategias más simples para estimar la función glótica, permitiendo así disminuir la complejidad del problema de optimización. En el material desarrollado en este capítulo, propondremos algunas alternativas para hacer frente a estos dos inconvenientes.

7.3. Modelado estocástico de la función glótica y de la fonación

Dedicaremos esta sección al desarrollo de un modelo en espacio de estados de la fonación. El objetivo es generar un modelo con las siguientes características: (*i*) que permita explicar la fonación de acuerdo a la TFF, (*ii*) que sirva para la síntesis de diferentes fonemas sonoros, y (*iii*) que sea aplicable a la descomposición de la voz con ayuda de los métodos en espacio de estados. Para ello, procederemos primeramente a construir una representación estocástica de la función glótica, la cual será un componente importante dentro del modelo de la fonación.

Como dijimos anteriormente, únicamente tendremos en cuenta los fonemas sonoros. De acuerdo con la TFF, estas emisiones son el resultado de la modulación del flujo de aire glótico al atravesar el tracto vocal y los labios (ver Sec. 3.3). En lo que sigue, consideraremos que el efecto derivativo de los labios durante la emisión se incorpora directamente en la fuente, por lo que la función glótica representa la derivada del flujo de aire que atraviesa la glotis.

7.3.1. Modelo estocástico de la función glótica

En la Sec. 3.5, introdujimos el concepto de función glótica, enunciamos sus propiedades principales y explicamos su interpretación en el marco de la TFF. Seguidamente, presentamos el modelo LF de la función glótica. Para facilitar la exposición del material de este capítulo, reproduciremos aquí las Ecs. (3.17) y (3.18) que definen este modelo:

$$v_g^{\mathrm{LF}}[n] = \begin{cases} E_0\, e^{\alpha\, n}\ \mathrm{sen}\,(\omega_g\, n), & \text{si} \quad 0 \leq n < N_e, \\ \frac{-E_e}{\epsilon\, N_a}\left(e^{-\epsilon\,(n-N_e)} - e^{-\epsilon\,(N_c-N_e)}\right), & \text{si } N_e \leq n < N_c, \\ 0, & \text{si } N_c \leq n < N_0, \end{cases} \tag{7.1}$$

sujeto al siguiente conjunto de restricciones:

$$\begin{cases} \sum_{n=0}^{N_0-1} v_g^{\mathrm{LF}}[n] = 0, \\ \omega_g = \frac{\pi}{N_p}, \\ \epsilon\, N_a = 1 - e^{-\epsilon\,(N_c-N_e)}, \\ E_e = -E_0\, e^{\alpha\, N_e}\ \mathrm{sen}\,(\omega_g\, N_e). \end{cases} \tag{7.2}$$

Los parámetros en las expresiones anteriores se describieron oportunamente en la Sec. 3.5.1. A modo de ejemplo, en la Fig. 7.1 mostramos una función glótica generada con este modelo. En la Fig. 3.6 presentamos este mismo ejemplo junto con el

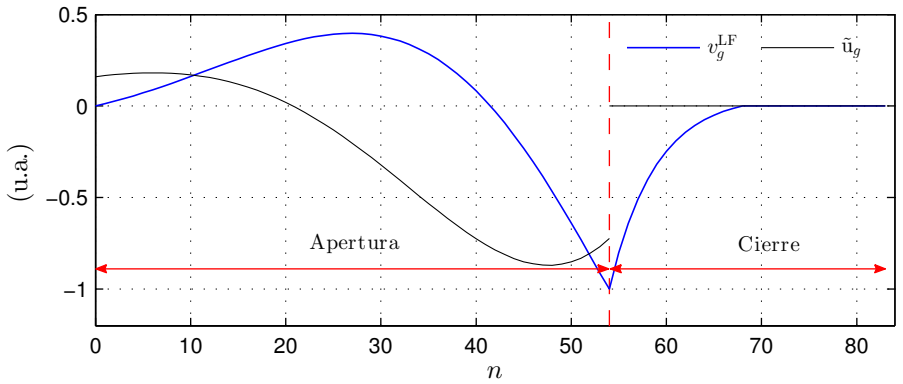


Figura 7.1: Relación entre la función glótica LF y la excitación $\tilde{\mathrm{u}}_g$. Se indican también las fases de apertura (OP) y de cierre (CP).

correspondiente flujo de aire glótico U_g^{LF}. Como puede apreciarse de las definiciones anteriores, el modelo LF se rige por leyes determinísticas variantes en el tiempo.

Recordemos que, para cada ciclo, el inicio de la función glótica se denomina *instante de apertura glótica* (GOI) y que la localización del mínimo de v_g^{LF} determina el *instante de cierre glótico* (GCI). En adelante, denominaremos fase de apertura[1] (OP) al intervalo de tiempo comprendido entre el GOI y su posterior GCI. El primer renglón en la Ec. 7.1 corresponde a la OP. Asimismo, llamaremos fase de cierre[2] (CP) al intervalo de tiempo comprendido entre un GCI y el GOI del siguiente período. De este modo, el segundo y el tercer renglón en la Ec. 7.1 corresponden a la CP.

Inspirados en la Ec. (7.1), presentaremos aquí un modelo estocástico de la función glótica formulado a partir de ecuaciones en diferencias estocásticas lineales. Como vimos en la Sec. 5.2.1, éstas son precisamente las estructuras involucradas en la definición de los MEEG. La primera fila de la Ec. (7.1) se puede escribir como sigue:

$$\begin{aligned} v_g^{\mathrm{LF}}[n] &= E_0\, e^{\alpha\, n}\ \mathrm{sen}\Big(\omega_g\, n\Big) \\ &= E_0\, e^{\alpha}\, e^{\alpha\,(n-1)}\ \mathrm{sen}\Big(\omega_g\,\Big[(n-1)+1\Big]\Big) \\ &= e^{\alpha}\cos\left(\omega_g\right)\Big[E_0\, e^{\alpha\,(n-1)}\ \mathrm{sen}\Big(\omega_g\,(n-1)\Big)\Big] \\ &\quad + e^{\alpha}\ \mathrm{sen}\left(\omega_g\right)\Big[E_0\, e^{\alpha\,(n-1)}\cos\Big(\omega_g\,(n-1)\Big)\Big]. \end{aligned} \tag{7.3}$$

La expresión anterior es válida para $0 \leq n < N_e$.

Se define la función de entrada o excitación auxiliar $\tilde{\mathrm{u}}_g$ de la siguiente forma:

$$\tilde{\mathrm{u}}_g[n] = \begin{cases} E_0\, e^{\alpha\, n}\cos\Big(\omega_g\, n\Big), & \text{si } 0 \leq n < N_e, \\ 0, & \text{si } N_e \leq n < N_0 \end{cases}. \tag{7.4}$$

Nótese que $\tilde{\mathrm{u}}_g$ también depende de los parámetros E_0, α y ω_g del modelo LF, correspondientes a la fase de apertura. Conociendo esta información, $\tilde{\mathrm{u}}_g$ puede generarse a partir de la expresión anterior. Para comprender mejor el comportamiento de esta señal, en la Fig. 7.1 presentamos, en azul, la función glótica v_g^{LF} y, en negro, la excitación $\tilde{\mathrm{u}}_g$ correspondiente.

[1] En inglés, "open phase".
[2] En inglés, "closed phase".

Incorporando la definición de la excitación auxiliar (7.4) en la expresión (7.3), obtenemos la siguiente ecuación en diferencias válida para la OP:

$$v_g^{\mathrm{LF}}[n] = A_g\, v_g^{\mathrm{LF}}[n-1] + B_g\, \tilde{\mathrm{u}}_g[n-1], \quad \text{si} \quad 0 \leq n < N_e, \tag{7.5}$$

donde $A_g = e^{\alpha}\cos(\omega_g)$ y $B_g = e^{\alpha}\,\mathrm{sen}\,(\omega_g)$.

Trabajando de forma similar, podemos escribir la segunda fila de la Ec. (7.1) como sigue:

$$\begin{aligned} v_g^{\mathrm{LF}}[n] &= -\frac{E_e}{\epsilon N_a}\left(e^{-\epsilon(n-N_e)} - e^{-\epsilon(N_c-N_e)}\right) \\ &= -\frac{E_e\, e^{-\epsilon}}{\epsilon\, N_a}\left(e^{-\epsilon(n-1-N_e)} - e^{-\epsilon\,(N_c-1-N_e)}\right) \\ &\approx e^{-\epsilon}\left[-\frac{E_e}{\epsilon\, N_a}\left(e^{-\epsilon(n-1-N_e)} - e^{-\epsilon(N_c-N_e)}\right)\right] \end{aligned} \tag{7.6}$$

La expresión anterior es válida para $N_e \leq n < N_c$. Esto último surge de que, generalmente, para las frecuencias de muestreo y los valores de ϵ usuales se cumple que $N_c - N_e >> 1$ y $e^{-\epsilon\,(N_c-N_e)} \approx e^{-\epsilon\,(N_c-1-N_e)} \approx 0$.

A partir de la expresión anterior obtenemos la siguiente ecuación en diferencias válida para la CP:

$$v_g^{\mathrm{LF}}[n] = C_g\, v_g^{\mathrm{LF}}[n-1], \quad \text{si} \quad N_e \leq n < N_c, \tag{7.7}$$

donde $C_g = e^{-\epsilon}$. Debido a que $\epsilon > 0$, se cumple entonces que $0 < C_g < 1$. Podemos observar que la expresión anterior representa el comportamiento exponencial de la función glótica durante la fase de retorno.

Combinando en una única expresión las ecuaciones en diferencias (7.5) y (7.7), y considerando además a la función glótica como un proceso estocástico, definimos al modelo estocástico (ES) de la función glótica v_g^{ES}:

$$v_g^{\mathrm{ES}}[n+1] = \begin{cases} A_g\, v_g^{\mathrm{ES}}[n] + B_g\, \tilde{\mathrm{u}}_g[n] + \nu[n], & \text{si} \quad 0 \leq n \leq N_e, \\ C_g\, v_g^{\mathrm{ES}}[n] + \nu[n], & \text{si} \;\; N_e < n \leq N_0, \end{cases} \tag{7.8}$$

donde $\nu[n] \sim \mathcal{N}(0, \sigma_\nu^2)$. El modelo propuesto sólo requiere de dos expresiones ya que si $v_g^{\mathrm{ES}}[n] \approx 0$ para $N_c < n < N_0$ y si $\sigma_\nu^2 \to 0$, se obtiene un comportamiento similar al determinado por la tercera fila de la Ec. (7.1). A su vez, para los valores usuales de los parámetros α, ω_g y ϵ se cumplen las siguientes relaciones:

$$1 < A_g^2 + B_g^2, \qquad 0 < B_g/A_g, \qquad \text{y} \qquad 0 < C_g < 1. \tag{7.9}$$

El modelo estocástico (7.8) presenta tres ventajas importantes: (*i*) puede ser estudiado en el marco de los métodos en espacio de estados descriptos en el Cap. 5; (*ii*) la morfología global de la función glótica puede describirse con los parámetros A_g, B_g y C_g; y (*iii*) cualquier error en la formulación del modelo es absorbido por la variable aleatoria ν.

Por último, es importante destacar que los parámetros del modelo LF pueden calcularse a partir de A_g, B_g y C_g, permitiendo obtener a partir de ellos v_g^{LF} y $\tilde{\mathrm{u}}_g$, de la siguiente forma:

$$\begin{aligned} \alpha &= \frac{1}{2}\,\ln\left(A_g^2 + B_g^2\right), \\ \omega_g &= \arctan\left(\frac{B_g}{A_g}\right), \\ \epsilon &= -\ln\left(C_g\right). \end{aligned} \tag{7.10}$$

7.3.2. Modelo en espacio de estados de la fonación

En esta sección, desarrollaremos el modelo en espacio de estados de la fonación propuesto en esta tesis de doctorado. Sea $s[n]$ la señal de voz correspondiente a un fonema sonoro, con $n = 1, 2, \ldots, N$ siendo N la cantidad de elementos. Definimos como $\mathcal{I}_N = \{1, 2, \ldots, N\}$ al conjunto de índices asociados a s. En adelante, tomaremos como hipótesis de trabajo que se conocen los GOI y los GCI para cada ciclo glótico. De esta forma, quedan determinadas las OP y las CP de la señal s. Sean $\mathcal{I}_{op}$ e $\mathcal{I}_{cp}$ los conjuntos de índices correspondientes a las OP y las CP, respectivamente. De acuerdo a las definiciones anteriores, debe cumplirse que $\mathcal{I}_{op} \cup \mathcal{I}_{cp} = \mathcal{I}_N$, $\mathcal{I}_{op} \cap \mathcal{I}_{cp} = \emptyset$ y $N = N_{op} + N_{cp}$, siendo $N_{op} = \#\mathcal{I}_{op}$ y $N_{cp} = \#\mathcal{I}_{cp}$, indicando con $\#$ la cardinalidad del correspondiente conjunto.

Trabajando en el marco de la TFF, consideraremos en lo que sigue que el comportamiento del tracto vocal puede modelarse mediante un filtro autorregresivo variable en el tiempo con entrada externa (TARX). Esta familia de modelos se introdujo en la Sec. 3.4. De este modo, la señal de voz puede modelarse como sigue:

$$s[n] = -\sum_{l=1}^{\rho} a_l[n]\, s[n-l] + G_g\, v_g^{\mathrm{ES}}[n] + \varepsilon[n], \tag{7.11}$$

donde ρ es el orden del modelo TARX, a_l con $l = 1, 2, \ldots, \rho$ son sus coeficientes, G_g es un factor de ganancia y $\varepsilon[n] \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. El signo menos en el primer término de la Ec. (7.11) es sólo por conveniencia algebraica.

En el caso de los fonemas vocales, en particular, durante la emisión las formantes y los anchos de banda se mantienen prácticamente constantes o sufren un cambio leve [42, 128, 133, 163]. Luego, es válido suponer que el modelo TARX del tracto vocal hereda esta característica y que $a_l[n+1] \approx a_l[n]$. En este punto, tomamos como criterio de diseño que el modelo de la fonación buscado debe ser flexible para representar adecuadamente la dinámica del tracto vocal. Por ello, optamos por modelar estocásticamente los coeficientes del modelo TARX [73, 127]:

$$a_l[n+1] = a_l[n] + \xi_l[n], \qquad l = 1, 2, \ldots, \rho, \tag{7.12}$$

donde los ξ_l son procesos gaussianos.

La articulación del tracto vocal es un proceso coordinado. Luego, es esperable entonces que exista cierto grado de dependencia entre los coeficientes a_l del modelo TARX. Para simular este fenómeno, consideraremos que los procesos aleatorios ξ_l desarrollan diferentes grados de correlación entre sí. Todas las hipótesis expuestas hasta aquí serán fundamentales durante el desarrollo y la implementación del modelo en espacio de estados de la fonación.

Para la construcción del modelo de la fonación procedimos de forma similar a [63]. Sin embargo, agregamos además el modelo estocástico la función glótica v_g^{ES}, definido en la Ec. (7.8). Obtuvimos así un modelo en espacio de estados lineal y gaussiano (MEEG) para representar la fonación. Para ello, consideramos un vector de estados $\mathbf{x}[n] \in \mathbb{R}^p$, con $p = \rho + 1$, definido de la siguiente forma:

$$\begin{aligned} \mathbf{x}[n] &= \begin{pmatrix} x_{(1)}[n] & x_{(2)}[n] & \ldots & x_{(p-1)}[n] & x_{(p)}[n] \end{pmatrix}^T \\ &= \begin{pmatrix} a_1[n] & a_2[n] & \ldots & a_\rho[n] & v_g^{\mathrm{ES}}[n] \end{pmatrix}^T. \end{aligned} \tag{7.13}$$

Tomando en cuenta todas las hipótesis de trabajo enunciadas anteriormente, desarrollamos las ecuaciones de transición de los estados y de observación que rigen

el MEEG de la fonación. Al igual que en el desarrollo de la sección anterior, fue necesario trabajar por separado con las OP y las CP. A continuación, describiremos la metodología empleada comenzando con la OP. Para $n \in \mathcal{I}_{op}$, obtuvimos la siguiente regla para la transición de los estados:

$$\mathbf{x}[n+1] = \mathbf{A}_{op}\,\mathbf{x}[n] + \mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n] + \mathbf{w}[n], \quad \text{con} \quad \mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{7.14}$$

y matrices $\mathbf{A}_{op} \in \mathbb{R}^{p\times p}$ y $\mathbf{f}_{op} \in \mathbb{R}^p$ definidas por:

$$\mathbf{A}_{op} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & A_g \end{pmatrix} \quad \text{y} \quad \mathbf{f}_{op} = \begin{pmatrix} \mathbf{0} \\ B_g \end{pmatrix}. \tag{7.15}$$

A su vez, $\mathbf{Q} \in \mathbb{R}^{q\times q}$ es una matriz de covarianza definida positiva, con estructura:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_\xi & \mathbf{0} \\ \mathbf{0} & \sigma_\nu^2 \end{pmatrix}, \tag{7.16}$$

donde $\mathbf{Q}_\xi \in \mathbb{R}^{\rho\times\rho}$ es la matriz de covarianza conjunta para los procesos ξ_l en la Ec. (7.12) y σ_ν^2 es la varianza de ν en la Ec. (7.8). Podemos apreciar que la dimensión del ruido de estado coincide con la dimensión del espacio de estados, es decir, $q = p$. La matriz $\mathbf{Q}_\xi$ simula la correlación entre los coeficientes del modelo TARX, lo que era una de las hipótesis planteadas. Además, de la expresión anterior se desprende que ν es independiente de los procesos estocásticos ξ_l.

Consideremos ahora las expresiones para la CP. Para $n \in \mathcal{I}_{cp}$, encontramos la siguiente regla para la transición de estados:

$$\mathbf{x}[n+1] = \mathbf{A}_{cp}\,\mathbf{x}[n] + \mathbf{f}_{cp}\,\tilde{\mathrm{u}}_g[n] + \mathbf{w}[n], \quad \text{con} \quad \mathbf{w}[n] \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{7.17}$$

con matrices $\mathbf{A}_{cp} \in \mathbb{R}^{p\times p}$ y $\mathbf{f}_{cp} \in \mathbb{R}^p$ definidas por:

$$\mathbf{A}_{cp} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & C_g \end{pmatrix} \quad \text{y} \quad \mathbf{f}_{cp} = \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix}. \tag{7.18}$$

La matriz de covarianza $\mathbf{Q}$ respeta la misma estructura y los requisitos establecidos anteriormente. Recordemos que por definición, Ec. (7.4), $\tilde{\mathrm{u}}_g[n] = 0$ para $n \in \mathcal{I}_{cp}$.

Pasemos ahora a la ecuación de observación. Considerando la expresión (7.11) para la generación de la señal de voz junto con las Ecs. (7.14) y (7.17), obtuvimos la siguiente ecuación de observación válida para $n \in \mathcal{I}_N$:

$$s[n] = \mathbf{H}[n]\,\mathbf{x}[n] + v[n], \quad \text{con} \quad v[n] \sim \mathcal{N}(0, \sigma_v^2), \tag{7.19}$$

donde $\mathbf{H}[n] \in \mathbb{R}^{1\times p}$ se define como sigue:

$$\mathbf{H}[n] = \begin{pmatrix} -s[n-1] & -s[n-2] & \cdots & -s[n-\rho] & G_g \end{pmatrix}. \tag{7.20}$$

Finalmente, las expresiones (7.14), (7.17) y (7.19) constituyen el MEEG de la fonación. Cabe destacar que, debido a su naturaleza estocástica, estas expresiones permiten lidiar con cualquier fenómeno que no haya sido contemplado en el modelo o con las aperiodicidades que ocurren en la fonación.

7.4. Métodos en espacio de estados

En el Cap. 5 describimos diferentes métodos en espacio de estados específicos para los MEEG. Luego, podemos hacer uso de todos estos métodos para el procesamiento de una señal de voz (fonema sonoro) guiado por el MEEG de la fonación desarrollado en la sección anterior. En esta tesis de doctorado, nos concentramos en el uso de estas herramientas para llevar a cabo dos tareas específicas: la estimación de los parámetros del MEEG de la fonación y la descomposición de la señal de voz. Es importante señalar que estos dos procesos están íntimamente relacionados.

Recordando la definición (7.13) para el vector de estados, se desprende que los métodos de filtrado y suavizado de Kalman permiten, por un lado, estimar la función glótica v_g^{ES} y, por el otro, llevar a cabo el seguimiento de los coeficientes del modelo TARX. A partir de esta última información, podemos estudiar el comportamiento espectral del tracto vocal de acuerdo con [127]. También, pueden calcularse las formantes y sus respectivos anchos de banda aplicando las expresiones (3.6). En adelante, denominaremos *filtrado inverso en espacio de estados* a la estrategia para la estimación conjunta de la función glótica y de la información espectral del tracto vocal. Aquí, nos concentraremos principalmente en las estimaciones suavizadas.

En la sección anterior, observamos que el comportamiento del MEEG de la fonación depende de diferentes parámetros. En general, fijar todos estos parámetros no es una tarea sencilla. Esta situación se dificulta más aún si lo que se pretende es procesar un conjunto de señales de voz. Afortunadamente, es posible elaborar estrategias basadas en los métodos en espacio de estados para el cálculo de los parámetros de un modelo específicos para una serie temporal. Dedicaremos la próxima sección a describir la estrategia para calcular iterativamente los parámetros óptimos del MEEG de la fonación a partir de una señal de voz.

7.5. Estimación de los parámetros del modelo

Presentaremos en esta sección una estrategia para generar estimaciones óptimas para los parámetros del MEEG de la fonación. Los parámetros desconocidos son: σ_v^2, $\mathring{\mathbf{Q}}$, A_g, B_g, C_g y G_g. Una vez más, tomamos como hipótesis que $\mathbf{Q} = \sigma_v^2\,\mathring{\mathbf{Q}}$. Consideramos conocido E_0, debido a que sólo se necesita conocer el producto $G_g\,E_0$ y por eso se fija el valor de uno de ellos. En los trabajos realizados seleccionamos $E_0 = 0{,}001$. También, se requiere estimar el estado inicial del MEEG. Para ello, nos inclinaremos por un modelo de la forma $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \mathbf{P}_0)$, donde el vector de estados iniciales $\hat{\mathbf{x}}_0$ y su matriz de covarianza $\mathbf{P}_0$ serán calculados.

Como dijimos antes, si se modifican los parámetros A_g y B_g, la entrada $\tilde{\mathrm{u}}_g$ puede actualizarse a partir de las expresiones (7.4) y (7.10). Afortunadamente, $\tilde{\mathrm{u}}_g$ presenta un comportamiento suave en la OP, el cual se modifica poco al variar ligeramente los parámetros A_g y B_g. En adelante, consideraremos todos los parámetros agrupados en la estructura $\boldsymbol{\Theta}$. Así, $\hat{\boldsymbol{\Theta}}_j$ indica la estimación del conjunto de parámetros en la j-ésima iteración.

7.5.1. Problema de optimización y función objetivo

Abordamos el cálculo de los parámetros del MEEG de la fonación en el marco de un problema de optimización. En particular, nos propusimos resolver el siguiente

problema:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}\in\mathcal{D}}{\arg\max} \;\; \mathcal{E}\left\{\ln \mathcal{L}_{pen}\Big(\boldsymbol{\Theta}\Big)\right\}, \tag{7.21}$$

donde $\mathcal{D}$ es el dominio de definición de los parámetros buscados.

Podemos apreciar que la función objetivo que se busca maximizar es diferente a la considerada en el problema (5.15). En este caso, $\mathcal{E}\left\{\ln \mathcal{L}_{pen}\Big(\boldsymbol{\Theta}\Big)\right\}$ involucra a la función *verosimilitud* $\mathcal{L}\Big(\boldsymbol{\Theta}\Big)$, descripta en la Sec. 5.5.2, con el agregado de un elemento de penalización, también llamado de regularización. Así, trabajamos con la función objetivo:

$$\ln \mathcal{L}_{pen}\Big(\boldsymbol{\Theta}\Big) = \ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big) - \lambda_j\, \Phi, \tag{7.22}$$

donde $\lambda_j > 0$ es el término de penalización asociado a la j-ésima iteración (puede variar con las iteraciones) y Φ es una función de penalización que toma valores no negativos.

Existe una infinidad de candidatos de Φ que cumplen las condiciones anteriores. Su elección depende, entre otros aspectos, del efecto que se busca con la penalización. En nuestras implementaciones, consideramos la siguiente expresión:

$$\Phi = \Phi(A_g,\, B_g,\, C_g) = \frac{1}{2}\left[\Big(A_g - \tilde{A}_g\Big)^2 + \Big(B_g - \tilde{B}_g\Big)^2 + \Big(C_g - \tilde{C}_g\Big)^2\right], \tag{7.23}$$

donde $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$ son valores conocidos que satisfacen dos condiciones fundamentales: (i) producen funciones glóticas v_g^{ES} adecuadas, y (ii) satisfacen las condiciones (7.9). Más adelante, describiremos cómo calcular estos parámetros para una señal de voz en particular.

Como puede apreciarse de su definición, el propósito de la penalización empleada es estabilizar el proceso de optimización con respecto al cálculo de los parámetros A_g, B_g y C_g. Como dijimos en la Sec. 7.2, estimar la función glótica con el modelo LF requiere resolver un problema de optimización no lineal complejo [63, 135]. Por su parte, los modelos lineales desarrollados en la Sec. 7.3 permiten formular un problema de optimización más simple. Sin embargo, en ocasiones esta linealización produce valores de los parámetros A_g, B_g y C_g que no satisfacen las condiciones (7.9), dando lugar así a una función v_g^{ES} inapropiada. Por ello, la función Φ penaliza aquellas estimaciones que difieran considerablemente de los valores adecuados y conocidos: $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$.

Considerando a A_g, B_g y C_g parámetros desconocidos de naturaleza determinística, se cumple entonces que Φ es una función determinística. Así, aplicando el operador valor esperado en la Ec. (7.22), se obtiene lo siguiente:

$$\begin{aligned}\mathcal{E}\left\{\ln \mathcal{L}_{pen}\Big(\boldsymbol{\Theta}\Big)\right\} &= \mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big) - \lambda_j\, \Phi\right\} \\ &= \mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\} - \lambda_j\, \Phi.\end{aligned} \tag{7.24}$$

Recordemos que el MEEG de la fonación se comporta de forma diferente para las OP y las CP. Entonces, es necesario modificar la expresión de la función $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$ de forma tal que contemple este cambio de comportamiento. Considerando la Ec. (5.19), obtenemos la siguiente definición de $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$ válida para el MEEG de

la fonación:

$$
\begin{aligned}
\mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\} = \tilde{c} &- \frac{1}{2}\,\ln\left(|\mathbf{P}_0|\right) - \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) \\
&+ \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 + \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 \\
&- \frac{N\,(1+p)}{2}\,\ln(\sigma_v^2) - \frac{N}{2}\,\ln\left(|\mathring{\mathbf{Q}}|\right) \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_N}\left[s[n]^2 - 2\,s[n]\,\mathbf{H}[n]\,\hat{\mathbf{x}}[n|N] + \mathbf{H}[n]\,\hat{\mathbf{C}}[n|N]\,\mathbf{H}[n]^T\right] \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}[n|N]\right) - \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
&\qquad - \hat{\mathbf{x}}[n|N]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] - \mathrm{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) \\
&\qquad + \mathrm{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}[n-1|N]\right) \\
&\qquad + \hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] - \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{x}}[n|N] \\
&\qquad + \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{x}}[n-1|N] + \tilde{\mathrm{u}}_g[n-1]^2\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\Big] \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{cp}}\Big[\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}[n|N]\right) - \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
&\qquad - \mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) + \mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}[n-1|N]\right)\Big],
\end{aligned}
\tag{7.25}
$$

con $\tilde{c} \in \mathbb{R}$ constante. Para arribar a esta definición de $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$, aplicamos las expresiones que rigen el MEEG de la fonación. Además, tuvimos en cuenta que la excitación $\tilde{\mathrm{u}}_g$ se anula durante las CP. En el Apéndice A describimos con mayor precisión el desarrollo de esta última expresión.

7.5.2. Reglas para el cálculo de los parámetros

A continuación, analizaremos las expresiones necesarias para calcular, de forma iterativa, los parámetros del MEEG de la fonación. Estudiando la Ec. (7.24) y sabiendo que Φ depende únicamente de los parámetros A_g, B_g y C_g, se desprende entonces que resultan aplicables varios de los resultados obtenidos en la Sec. 5.5.3. Este es el caso de las expresiones para el cálculo de σ_v^2 Ec. (5.23), de $\mathring{\mathbf{Q}}$ Ec. (5.27), de $\hat{\mathbf{x}}_0$ Ec. (5.29), y de $\mathbf{P}_0$ Ec. (5.33).

Por otra parte, desarrollamos un conjunto de expresiones útiles para el cálculo de los parámetros A_g, B_g, C_g y G_g del MEEG de la fonación. Con el propósito de evitar que este capítulo sea muy extenso, trasladamos al Apéndice B el material con la deducción de estas reglas. En lo que sigue, presentaremos y analizaremos las expresiones desarrolladas.

Las reglas para actualizar A_g y B_g surgen de resolver un sistema algebraico de dos ecuaciones con dos incógnitas. Así, se obtienen las siguientes expresiones:

$$
\begin{aligned}
\hat{A}_{g\ \mathrm{est}} &= \frac{\gamma_1\,\theta_{22} - \gamma_2\,\theta_{12}}{\mathring{\mathrm{D}}}, \\
\hat{B}_{g\ \mathrm{est}} &= \frac{\gamma_2\,\theta_{11} - \gamma_1\,\theta_{21}}{\mathring{\mathrm{D}}},
\end{aligned}
\tag{7.26}
$$

donde el determinante $\mathring{\mathrm{D}}$ se define como sigue:

$$\mathring{\mathrm{D}} = \theta_{11}\,\theta_{22} - \theta_{12}\,\theta_{21}. \tag{7.27}$$

Los coeficientes involucrados en las expresiones anteriores se calculan de la siguiente forma:

$$\begin{aligned}
\theta_{11} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{n-1})_{(p,p)} + \lambda_j\,\sigma_v^2, \qquad & \theta_{12} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)},\\
\theta_{21} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)}, \qquad & \theta_{22} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,\tilde{\mathrm{u}}^{op}_{n-1} + \lambda_j\,\sigma_v^2,\\
\gamma_1 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op}_{n,n-1})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{op}_{n-1})_{(i,p)} + \lambda_j\,\sigma_v^2\,\tilde{A}_g,\\
\gamma_2 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n})_{(p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(i)} + \lambda_j\,\sigma_v^2\,\tilde{B}_g,
\end{aligned} \tag{7.28}$$

Para generar las reglas para los coeficientes en (7.28), consideramos que $\sigma_v^2 > 0$ y $\mathring{\mathbf{Q}}^{-1}$ son conocidos. Debido a que ambos parámetros se modifican durante el proceso de optimización, se emplean sus últimas estimaciones disponibles para calcular los elementos en (7.28). A su vez, las expresiones $\tilde{\mathbf{C}}^{op}_{n-1}$, $\tilde{\mathbf{C}}^{op}_{n,n-1}$, $\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}$, $\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n}$ y $\tilde{\mathrm{u}}^{op}_{n-1}$ involucradas en las expresiones anteriores se definen en el Apéndice B bajo las referencias (B.7) y (B.17).

El parámetro C_g se actualiza utilizando iterativamente la siguiente regla:

$$\hat{C}_{g\ \mathrm{est}} = \frac{\gamma_3}{\theta_3}, \tag{7.29}$$

donde

$$\begin{aligned}
\theta_3 &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(p,p)} + \lambda_j\,\sigma_v^2,\\
\gamma_3 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{cp}_{n,n-1})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(i,p)} + \lambda_j\,\sigma_v^2\,\tilde{C}_g.
\end{aligned} \tag{7.30}$$

Tomamos como hipótesis que $\sigma_v^2 > 0$ y $\mathring{\mathbf{Q}}^{-1}$ son conocidos, de forma análoga a como procedimos anteriormente. Las expresiones $\tilde{\mathbf{C}}^{cp}_{n-1}$ y $\tilde{\mathbf{C}}^{cp}_{n,n-1}$ involucradas en las expresiones anteriores se definen en el Apéndice B bajo la referencia (B.30).

Por último, el parámetro de ganancia G_g se obtiene de la siguiente forma:

$$\hat{G}_g = \frac{\mu}{\xi}, \tag{7.31}$$

donde

$$\begin{aligned}
\mu &= \sum_{n\in\mathcal{I}_N}\left[s[n]\,(\hat{\mathbf{x}}[n|N])_{(p)} + \sum_{i=1}^{\rho} s[n-i]\,(\hat{\mathbf{C}}[n|N])_{(i,p)}\right]\\
\xi &= \sum_{n\in\mathcal{I}_N}(\hat{\mathbf{C}}[n|N])_{(p,p)}.
\end{aligned} \tag{7.32}$$

Comentarios adicionales respecto a la penalización

Nos valdremos de los resultados expuestos para explicar la conveniencia del uso del término de penalización (regularización) en la función objetivo. De la definición (7.23), se desprende que la penalización sólo afecta el cálculo de A_g, B_g y C_g. Analizaremos aquí la dependencia de las estimaciones de estos parámetros con respecto al coeficiente de penalización λ_j y a la varianza del ruido de medición σ_v^2.

Consideremos en primer lugar la Ec. (7.29) para estimar C_g, junto con las expresiones en (7.30). Observamos que la Ec. (7.29) involucra el cociente entre polinomios de primer orden en la variable $(\lambda_j\,\sigma_v^2)$. El coeficiente que acompaña a $(\lambda_j\,\sigma_v^2)$ en el numerador es $\tilde{C}_g$, mientras que en el denominador el coeficiente es 1. Considerando valores grandes de λ_j o de σ_v^2, convirtiendo a $(\lambda_j\,\sigma_v^2)$ en el término predominante en cada polinomio, se cumple que $\hat{C}_g \to \tilde{C}_g$. A su vez, para $\lambda_j \to 0$ o para $\sigma_v^2 \to 0$, el efecto de la penalización se vuelve despreciable y el parámetro $\hat{C}_g$ queda completamente determinado por la información en los parámetros: $\mathring{\mathbf{Q}}$, $\tilde{\mathbf{C}}_{n,n-1}^{cp}$ y $\tilde{\mathbf{C}}_{n-1}^{cp}$.

Un razonamiento similar puede emplearse para las reglas (7.26) para actualizar A_g y B_g, tomando en cuenta también las expresiones en (7.27) y en (7.28). Estos casos difieren del analizado arriba en que las estimaciones involucran el cociente entre polinomios de segundo orden en la variable $(\lambda_j\,\sigma_v^2)$. En $\hat{A}_{g\,\mathrm{est}}$, el coeficiente que acompaña a la potencia mayor en el polinomio del numerador es $\tilde{A}_g$, mientras que en $\hat{B}_{g\,\mathrm{est}}$ el coeficiente es $\tilde{B}_g$. En ambos casos, el coeficiente que acompaña a la mayor potencia en el polinomio del denominador $\mathring{\mathrm{D}}$ es 1. Se demuestra, entonces, que las estimaciones $\hat{A}_{g\,\mathrm{est}}$ y $\hat{B}_{g\,\mathrm{est}}$ exhiben una dependencia con respecto a λ_j y σ_v^2 análoga a la descripta para $\hat{C}_{g\,\mathrm{est}}$, en el párrafo anterior.

Lo expuesto hasta aquí nos permite extraer dos importantes conclusiones. Por un lado, la función de penalización Φ garantiza que las estimaciones tomen valores similares (cercanos) a $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$. Por ello, si se seleccionan adecuadamente estos parámetros de referencia, de forma tal que satisfagan los requisitos del modelo estocástico de la función glótica, entonces las expresiones (7.26) y (7.29) producirán estimaciones adecuadas de A_g, B_g y C_g.

La segunda conclusión surge de analizar cómo influye la incerteza en las observaciones, caracterizada por la varianza σ_v^2, en el filtrado y el suavizado de Kalman. Cuando las observaciones están fuertemente contaminadas con ruido, cobra mayor relevancia en la estimación del vector de estados la información provista por el modelo [8, 30]. En este escenario, la penalización colabora asegurando que en esta operación se utilicen valores adecuados de A_g, B_g y C_g. Si, por el contrario, las observaciones son confiables, el vector de estados se genera como la suma ponderada entre las observaciones y las estimaciones del modelo [35, 51, 86]. Esta ponderación queda determinada por la ganancia de Kalman $\tilde{\mathbf{K}}$ (ver Tab. 5.1). De forma similar, la penalización ocasiona que los parámetros de la función glótica se obtengan como la suma ponderada entre la información extraída con el modelo y los valores de referencia $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$.

7.5.3. Procedimiento para la estimación de los parámetros

A modo de resumen, describiremos aquí el proceso implementado para calcular los parámetros desconocidos $\boldsymbol{\Theta}$ del MEEG de la fonación. Recordemos que este método involucra la resolución iterativa del problema de optimización (7.21). En la Fig. 7.2 presentamos un diagrama de flujo explicativo.

La información requerida para iniciar este proceso es la siguiente: la señal de voz $s[n]$ con $n \in \mathcal{I}_N$, los GOI y los GCI correspondientes y el orden ρ del modelo TARX del tracto vocal. Se requieren, además, los parámetros de referencia $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$. Una estrategia para obtener estos elementos, que garantiza el cumplimiento de las condiciones descriptas en la Sec. 7.3.1, consiste en estimar la función glótica aplicando alguna técnica de filtrado inverso y, luego, ajustar el modelo LF a esta

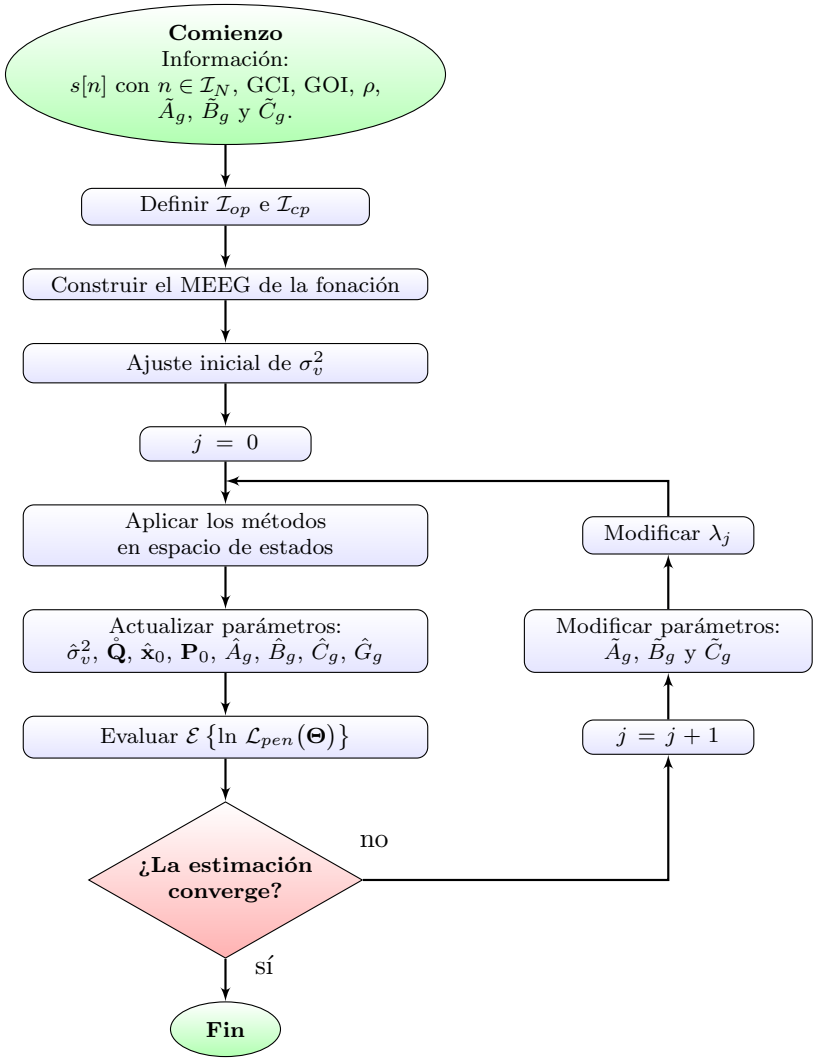


Figura 7.2: Diagrama de flujo del proceso para la estimación de los parámetros desconocidos $\boldsymbol{\Theta}$ para el MEEG de la fonación. Como resultado, se obtienen los valores óptimos de σ_v^2, $\mathring{\mathbf{Q}}$, $\hat{\mathbf{x}}_0$, $\mathbf{P}_0$, A_g, B_g, C_g y G_g.

señal. Haciendo esto, se obtienen los valores de α, ω_g y ϵ que producen el ajuste óptimo. A partir de éstos, se calculan $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$. En nuestros trabajos, para realizar el filtrado inverso utilizamos el método IAIF, ver Sec. 3.6.1. Con toda esta información, se da comienzo a la estimación de los parámetros.

En primer lugar, se definen los conjuntos de índices $\mathcal{I}_{op}$ e $\mathcal{I}_{cp}$ en función de los GOI y los GCI. A continuación, se construye el MEEG de la fonación. Para ello, se inicializan los parámetros como sigue: $\sigma_v^2 = \mathrm{Var}\{s\}$, $\mathring{\mathbf{Q}} = 0{,}001\,\mathbf{I}_p$, $G_g = 1$, $E_0 = 0{,}001$, y $\mathbf{P}_0 = 0{,}001\,\sigma_v^2\,\mathbf{I}_p$. Se asignan también: $A_g = \tilde{A}_g$, $B_g = \tilde{B}_g$ y $C_g = \tilde{C}_g$. Por conveniencia, se trabaja con un instante inicial perteneciente a la CP, preferentemente un GCI. Considerando la definición (7.13), el vector de estados iniciales $\hat{\mathbf{x}}_0$ se construye con la información del filtro de tracto vocal calculada con el método IAIF y tomando $v_g^{\mathrm{ES}}[0] = -E_0$, sabiendo que la función glótica es negativa en la CP. Estudios preliminares realizados por nosotros probaron que esta inicialización da lugar a un comportamiento satisfactorio del MEEG. Seguidamente, se realiza un

primer ajuste de σ_v^2 así su valor se adecua mejor a la señal de voz s, se inicializa el contador $j = 0$ y se continúa con la etapa iterativa de este método.

Se aplican los métodos en espacio de estados con el propósito de estimar, a partir de s, la información intermedia requerida. Seguidamente, se actualizan los parámetros del MEEG de la fonación, de acuerdo a lo explicado en la Sec. 7.5.2. Luego, se estima el valor de la función objetivo $\mathcal{E}\left\{\ln \mathcal{L}_{pen}\left(\boldsymbol{\Theta}\right)\right\}$ y se evalúa la convergencia del método. Para realizar esto último, se calcula la diferencia relativa entre los valores de la función objetivo actual y de la iteración pasada, tomando esta última cantidad como referencia. Si la diferencia es menor a una cota preestablecida consideramos que el procedimiento convergió, arrojando como resultado los valores óptimos de los parámetros σ_v^2, $\mathring{\mathbf{Q}}$, $\hat{\mathbf{x}}_0$, $\mathbf{P}_0$, A_g, B_g, C_g y G_g. En caso contrario, se modifican los valores de referencia $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$, se actualiza el contador j y se da paso a la siguiente iteración. En nuestros trabajos, escogimos la cota en el rango $(1 \times 10^{-3}, 1 \times 10^{-6})$ dependiendo de la calidad deseada.

Es importante señalar que a lo largo del proceso de optimización se modifican los parámetros $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$. La forma de trabajo adoptada fue la siguiente. Durante las primeras iteraciones se mantienen los valores iniciales, generados a partir del método IAIF. Esto otorga robustez a la etapa inicial de la optimización. El número de iteraciones iniciales se tomó en el orden de algunas decenas. Luego, estos parámetros se actualizan en cada iteración, asignándoles el valor de la estimación más reciente disponible. Así, se busca otorgar mayor libertad al cálculo de A_g, B_g y C_g.

Por otro lado, a lo largo de este proceso también se modifica λ_j. Este coeficiente determina la importancia de la penalización en la j-ésima iteración. Un detalle importante al respecto es que la eficacia de la penalización depende, a su vez, de la cantidad de elementos N en la señal de voz. En nuestras implementaciones, adoptamos una solución de compromiso que, por un lado, garantice la estabilidad en la primera etapa del proceso de optimización y, por el otro, favorezca la actualización de A_g, B_g y C_g en las iteraciones posteriores. Para $N \approx 500$, inicializamos λ_j con un valor muy grande $\lambda_j \in (1 \times 10^8, 1 \times 10^{10})$. Luego, se lo decrementa de forma escalonada donde transcurrido un número de iteraciones (≈ 100) se lo modifica como sigue: $\lambda_{j+1} = 0{,}1\,\lambda_j$.

Hasta este punto, nos hemos dedicado a desarrollar el MEEG de la fonación y a describir el procedimiento para el cálculo de sus parámetros. En lo que sigue, estudiaremos las aptitudes de estas herramientas. Nos concentraremos, en particular, en evaluar su utilidad en el marco del filtrado inverso en espacio de estados de una señal de voz.

7.6. Simulaciones con señales artificiales

En esta sección, describiremos el desempeño de los modelos desarrollados al emplearlos, conjuntamente con los métodos en espacio de estados, para la estimación de la función glótica y de la información espectral del tracto vocal en señales de voz sintetizadas. Para ello, llevamos a cabo diferentes simulaciones, con el propósito de estudiar las herramientas propuestas en condiciones controladas. Además, en estos escenarios se conoce con certeza la información buscada.

Recordemos que la construcción del MEEG de la fonación y la etapa inicial del método para la estimación de sus parámetros requieren valores adecuados para $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$ y para los coeficientes del modelo TARX del tracto vocal. Como dijimos

Tabla 7.1: Primeras cuatro formantes, con sus respectivos anchos de banda, para las vocales sostenidas modales /a/ y /e/.

	Formantes (Hz)				**Anchos de Banda** (Hz)			
	F_1	F_2	F_3	F_4	B_1	B_2	B_3	B_4
/a/	800	1200	2600	3200	60	50	105	110
/e/	500	2200	2700	3600	50	40	95	100

antes, calculamos esta información empleando el método IAIF y ajustando el modelo LF. En las simulaciones realizadas, alcanzamos los mejores resultados trabajando con $p^{IAIF} = 8$ y $g^{IAIF} = 2$. Esto se debe a que la estrategia de síntesis empleada se basa íntegramente en la TFF. Para obtener resultados comparables, en el MEEG de la fonación consideramos un modelo TARX de orden $\rho = 8$ para representar el tracto vocal. Es importante destacar, además, que para cada señal sintetizada conocemos con certeza la localización de los GOI y los GCI.

7.6.1. Vocales sostenidas

Sintetizamos diferentes conjuntos de vocales sostenidas con una frecuencia de muestreo $f_s = 10$ kHz. Para ello, empleamos la siguiente estrategia. En primer lugar, se generó un tren de pulsos glóticos v_g^{LF} empleando el modelo LF, y se le adicionó ruido de aspiración siguiendo las recomendaciones en [79, 101]. Luego, se procesó esta señal mediante un filtro AR representativo del tracto vocal, obteniendo como resultado la señal de voz. Por último, se contaminó esta señal agregándole ruido acústico. Los pulsos LF se generaron fijando $E_e = -0{,}001$ y tomando valores aleatorios de los parámetros $\{N_p,\ N_e,\ N_a\}$, de forma similar a [67]. En este punto, sugerimos recordar la descripción del modelo LF expuesta en la Sec. 3.5.1.

Consideramos conjuntos de vocales con diferentes niveles de relación señal de voz a ruido acústico (SNR) y de relación función glótica a ruido de aspiración (GNR), y con diferentes f_0. Generamos conjuntos de vocales modificando sólo una de estas condiciones cada vez, tomando como valores de referencia: SNR = 60 dB, GNR = 60 dB, y $f_0 = 108$ Hz. Los niveles considerados fueron: SNR = $\{0, 5, 10, \ldots, 60\}$ dB, GNR = $\{0, 5, 10, \ldots, 60\}$ dB y $f_0 = \{88, 98, \ldots, 128, 188, 198, \ldots, 228\}$ Hz. Estos conjuntos permitieron evaluar los modelos propuestos simulando diferentes escenarios, contemplando diversos efectos acústicos y fisiológicos. Para cada escenario, se sintetizaron 100 vocales /a/ sostenidas modales, respetando la información espectral de la Tab. 7.1. Empleamos estos conjuntos de vocales para llevar a cabo simulaciones con el objetivo de estudiar diferentes aspectos de los métodos desarrollados. Para cada una de estas señales, se conocía con certeza tanto el comportamiento espectral del tracto vocal como la función glótica v_g^{LF} empleados en la síntesis.

Procesamos las vocales artificiales empleando las estrategias desarrolladas. Para cada señal, se inicializó el MEEG de la fonación empleando el procedimiento descripto anteriormente. Luego, se aplicó el procedimiento de la Sec. 7.5 para calcular los parámetros óptimos del MEEG. De esta forma, se obtuvo el modelo particular para la señal de voz bajo estudio. Con esta información, se trabajó con el método de suavizado de Kalman para estimar la función glótica y los coeficientes del modelo

TARX del tracto vocal. Este último paso constituye el filtrado inverso en espacio de estados, propiamente dicho.

En la Fig. 7.3 mostramos los resultados de la descomposición de una vocal /a/ artificial, aplicando los modelos desarrollados y los métodos en espacios de estados. En la parte superior, podemos observar la forma de onda de la señal de voz sintetizada (SNR=60 dB, GNR=30 dB, f_0=108 Hz). En la segunda fila, presentamos la estimación $\hat{v}_g^{\mathrm{ES}}$ correspondiente al modelo estocástico de la función glótica. A fines comparativos, graficamos también la función glótica v_g^{LF} empleada en la síntesis y su estimación $\hat{v}_g^{\mathrm{IAIF}}$ generada con el método IAIF. Podemos apreciar que tanto $\hat{v}_g^{\mathrm{ES}}$ como $\hat{v}_g^{\mathrm{IAIF}}$ parecen estimaciones satisfactorias de la función glótica. Sin embargo, el uso combinado del modelo estocástico propuesto y de los métodos en espacio de estados da lugar a una estimación más suave de la función glótica. Esto se debe, principalmente, a la estructura estocástica del modelo desarrollado y a que los métodos en espacio de estados permiten estimadores no causales [43, 51].

En la tercera fila, podemos observar el flujo de aire glótico U_g^{LF} correspondiente al modelo LF. Esta señal se calcula integrando numéricamente la función glótica [5, 42, 128]. Presentamos también, de forma superpuesta, las estimaciones de esta señal calculadas, por un lado, con el modelo estocástico $\hat{U}_g^{\mathrm{ES}}$ y, por el otro, con el método IAIF $\hat{U}_g^{\mathrm{IAIF}}$. Se aprecia que estas dos señales son estimaciones satisfactorias de U_g^{LF}. A su vez, observamos en $\hat{U}_g^{\mathrm{IAIF}}$ un leve comportamiento irregular para las CP, que no se percibe en $\hat{U}_g^{\mathrm{ES}}$. En la parte inferior de la Fig. 7.3, presentamos el espectro de potencia en dB del filtro de tracto vocal S^{TV}, junto con las estimaciones calculadas con el MEEG de la fonación $\hat{S}^{MEEG}$ y con el método IAIF $\hat{S}^{IAIF}$. Todas estas curvas corresponden al espectro de potencia promedio para la ventana de señal analizada. Para el MEEG, primero estimamos los espectros de potencia instantáneos siguiendo [127] y luego los promediamos. Podemos apreciar que ambas estimaciones capturan adecuadamente la información espectral del filtro de tracto vocal. Sin embargo, para el ejemplo analizado $\hat{S}^{MEEG}$ muestra un mejor desempeño en la región de las formantes tercera y cuarta.

Estimación de la función glótica

Describiremos aquí las simulaciones llevadas a cabo para evaluar la calidad de la información de la función glótica obtenida empleando, por un lado, la estrategia para la estimación de los parámetros y, por el otro, el filtrado inverso en espacio de estados.

En primer lugar, nos concentramos en el conjunto de parámetros $\{\alpha, \omega_g, \epsilon\}$ debido a que su estimación suele ser difícil en la práctica [2, 57]. Al mismo tiempo, evaluamos las reglas para el cálculo de los valores óptimos de los parámetros A_g, B_g y C_g, ya que con esta información y empleando las expresiones (7.10) pueden obtenerse los parámetros del modelo LF. Como medida de error escogimos la distancia cuadrática media porcentual entre los valores teóricos y sus estimaciones, definida como sigue:

$$e_{\{\alpha,\omega_g,\epsilon\}} = 100\sqrt{\frac{1}{3}\left[\left(\frac{\hat{\alpha}-\alpha}{\alpha}\right)^2 + \left(\frac{\hat{\omega}_g-\omega_g}{\omega_g}\right)^2 + \left(\frac{\hat{\epsilon}-\epsilon}{\epsilon}\right)^2\right]}\ \%, \tag{7.33}$$

donde $\{\hat{\alpha}, \hat{\omega}_g, \hat{\epsilon}\}$ son las estimaciones de los parámetros glóticos. Como dijimos anteriormente, empleamos el método IAIF para inicializar la búsqueda de los parámetros

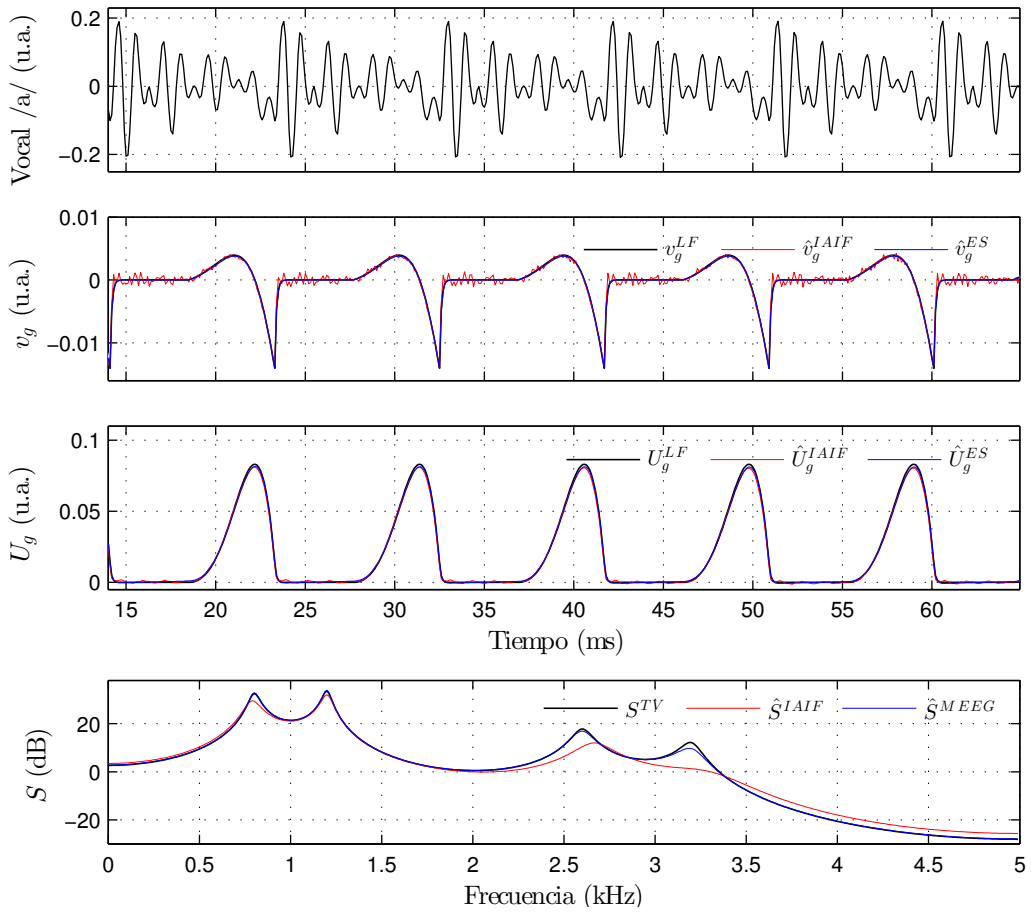


Figura 7.3: Filtrado inverso en espacio de estados de una vocal /a/ sostenida artificial (SNR=60 dB, GNR=30 dB, f_0=108 Hz). *Arriba:* forma de onda de la vocal /a/. *Segunda fila:* función glótica v_g^{LF} empleada en la síntesis, junto con sus estimaciones obtenidas con el modelo estocástico $\hat{v}_g^{\mathrm{ES}}$ y con el método IAIF $\hat{v}_g^{\mathrm{IAIF}}$. *Tercera fila:* flujo de aire glótico U_g^{LF}, y sus estimaciones calculadas con el modelo estocástico $\hat{U}_g^{\mathrm{ES}}$ y con el método IAIF $\hat{U}_g^{\mathrm{IAIF}}$. *Abajo:* espectro de potencia en dB del tracto vocal S^{TV}, junto con sus estimaciones generadas con el MEEG de la fonación $\hat{S}^{MEEG}$ y con el método IAIF $\hat{S}^{IAIF}$.

de la función glótica. Por ello, tomamos a los parámetros obtenidos con este método como valores de referencia para ser usados con fines comparativos.

En la Tab. 7.2 informamos el valor medio y, entre paréntesis, el desvío estándar de $e_{\{\alpha,\omega_g,\epsilon\}}$ para diferentes niveles de SNR, GNR y f_0. Para cada conjunto, se presentan los errores alcanzados aplicando las estrategias desarrolladas en este capítulo y el método IAIF. Recordemos que IAIF es una estrategia desarrollada específicamente para la estimación de la función glótica [7], y que los parámetros correspondientes a este método se calcularon ajustando el modelo LF. Esto explica el desempeño levemente superior de esta técnica. Sin embargo, podemos apreciar que la combinación del modelo estocástico v_g^{ES} y el MEEG de la fonación alcanzó un buen desempeño, similar incluso al obtenido con IAIF, para todos los escenarios evaluados, a excepción de los grupos con baja SNR o con f_0 alta.

En la parte inferior de esta tabla, indicamos aquellos resultados estadísticamen-

Tabla 7.2: Error en la estimación de los parámetros glóticos $\{\alpha, \omega_g, \epsilon\}$ considerando diferentes niveles de SNR, GNR y f_0. En cada caso se presenta el valor medio del error y, entre paréntesis, el desvío estándar correspondiente. Se indican aquellos conjuntos estadísticamente diferentes (* $0{,}05 \geq valor\,p > 0{,}01$; ** $0{,}01 \geq valor\,p > 0{,}001$; *** $0{,}001 \geq valor\,p$).

	SNR (dB)			**GNR** (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
MEEG	8.23 (14.17)	5.09 (1.06)	5.25 (2.12)	5.75 (2.56)	5.31 (2.78)	5.43 (3.29)	11.75 (5.18)	74.79 (11.21)
IAIF	8.23 (14.17)	5.10 (1.06)	4.67 (0.50) ***	5.73 (2.54)	4.66 (1.43)	4.65 (1.31)	10.86 (4.24) **	73.58 (10.83) **

te diferentes de acuerdo con el test de *suma de rangos de Wilcoxon* [93]. Para cada conjunto, consideramos como hipótesis nula que coinciden las medias de los resultados alcanzados con ambos métodos, en oposición a la hipótesis alternativa de que estas medias difieren. Podemos apreciar que en cinco de los escenarios analizados no se pudo rechazar la hipótesis nula. A su vez, resultaron estadísticamente diferentes los errores para niveles altos de SNR y para diferentes f_0. Esto sugiere que los métodos desarrollados producen estimaciones adecuadas de los parámetros $\{\alpha, \omega_g, \epsilon\}$, comparables a las obtenidas con otro método utilizado ampliamente para esta tarea.

En el filtrado inverso, el objetivo principal es estimar la función glótica de forma precisa a partir de una señal de voz. Por ello, en segundo lugar evaluamos la calidad de la estimación de la función glótica calculada aplicando el MEEG de la fonación y los métodos en espacio de estados. Como figura de mérito, escogimos el error cuadrático medio porcentual entre la función glótica real y su estimación, definido como sigue:

$$e_{v_g} = 100 \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{v}_g[n] - v_g[n]}{E_e} \right)^2} \,\%, \tag{7.34}$$

donde $\hat{v}_g$ es la estimación de la función glótica y N es la cantidad de muestras disponibles. En la expresión anterior, el parámetro E_e se empleó como valor de referencia, obteniendo así una medida de error relativa.

En la Tab. 7.3 presentamos el valor medio y, entre paréntesis, el desvío estándar de e_{v_g} para diferentes niveles de SNR, GNR y f_0. Al igual que en el estudio anterior, para cada conjunto se presentan los errores obtenidos con las estrategias desarrolladas y con el método IAIF. Estos resultados sugieren que aplicar los métodos en espacio de estados, tomando en cuenta el MEEG de la fonación, permite obtener estimaciones muy precisas de la función glótica. Además, observamos que la precisión es buena para todos los escenarios analizados, salvo para niveles bajos de SNR (ruido acústico de gran amplitud) donde el error fue mayor al 8 %.

Es notable que las estrategias propuestas muestran un mejor desempeño que el método IAIF. Estudiando la tabla, observamos que las estrategias basadas en métodos en espacio de estados arrojan el menor error de estimación, para todos los escenarios analizados. Nuevamente, indicamos en la parte inferior de la tabla los conjuntos estadísticamente diferentes de acuerdo con el test de *suma de rangos de*

Tabla 7.3: Error en la estimación de la forma de onda de la función glótica considerando diferentes niveles de SNR, GNR y f_0. En cada caso se presenta el valor medio del error y, entre paréntesis, el desvío estándar correspondiente. Se indican aquellos conjuntos estadísticamente diferentes (* $0{,}05 \geq valor\,p > 0{,}01$; ** $0{,}01 \geq valor\,p > 0{,}001$; *** $0{,}001 \geq valor\,p$).

	SNR (dB)			**GNR** (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
MEEG	8.12 (2.70)	3.64 (1.30)	1.40 (1.37)	2.87 (1.76)	1.05 (1.35)	0.97 (1.36)	1.04 (1.15)	4.98 (1.14)
IAIF	29.10 (4.52) ***	13.71 (2.86) ***	5.52 (2.46) ***	10.49 (5.85) ***	2.93 (0.40) ***	2.75 (0.37) ***	6.78 (7.31) ***	23.27 (15.48) ***

Wilcoxon [93]. Así, existe evidencia significativa de que los errores calculados con los dos métodos difieren entre sí. Todo esto nos permite afirmar que el filtrado inverso en espacio de estados genera estimaciones muy precisas de la función glótica.

Estimación de la información espectral del tracto vocal

En tercer lugar, estudiamos el error cometido al estimar la información espectral del tracto vocal con las estrategias desarrolladas. En este caso, consideramos como figura de mérito la distancia cuadrática media en dB entre el espectro de potencia estimado y el correspondiente al filtro de tracto vocal S^{TV} utilizado en la síntesis, para un conjunto de L frecuencias discretas $\{f_1, f_2, \ldots, f_L\}$. Esta medida se define como sigue:

$$e_S = \sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(10\,\log_{10}\hat{S}[f_l] - 10\,\log_{10}S^{TV}[f_l]\right)^2}\ \text{dB}, \tag{7.35}$$

donde $\hat{S}[f_l]$ es valor del espectro de potencia estimado para la frecuencia f_l. Aquí, consideramos el intervalo de frecuencias (0, 5) kHz con $L = 512$.

En la Tab. 7.4 informamos el valor medio y, entre paréntesis, el desvío estándar de la medida e_S para diferentes niveles de SNR, GNR y f_0. A diferencia de los casos anteriores, no sólo se compararon las estrategias propuestas en este capítulo y el método IAIF, sino que presentamos también el error cometido con el método clásico LP. Podemos observar que las estrategias basadas en los métodos en espacio de estados arrojan el menor error de estimación en todos los escenarios, salvo para condiciones muy severas de ruido acústico (SNR $= 0 - 20$ dB) donde el mejor desempeño se obtuvo con LP. A su vez, en casi todos los escenarios analizados el peor desempeño se obtuvo con el método LP. El método IAIF, por su parte, muestra un desempeño intermedio en todos los casos, salvo para valores pequeños de SNR donde obtiene el error mayor.

Comparamos estadísticamente el desempeño de las estrategias desarrolladas con respecto a los métodos IAIF y LP, por separado. En la tabla, indicamos en cada caso los conjuntos de errores calculados con IAIF, o con LP, que son estadísticamente diferentes a los obtenidos con el filtrado inverso en espacio de estados, de acuerdo con

Tabla 7.4: Error en la estimación de la información espectral del tracto vocal para diferentes valores de SNR, GNR y f_0. Se informan los errores calculados con las estrategias propuestas, con IAIF y con LP. En cada caso se presenta el valor medio del error y, entre paréntesis, el desvío estándar correspondiente. Se indican aquellos conjuntos estadísticamente diferentes (* $0{,}05 \geq valor\,p > 0{,}01$; ** $0{,}01 \geq valor\,p > 0{,}001$; *** $0{,}001 \geq valor\,p$).

	SNR (dB)			**GNR** (dB)			f_0 (Hz)	
	0-20	25-40	45-60	0-20	25-40	45-60	88-118	188-218
MEEG	13.49 (1.18)	9.09 (1.45)	2.46 (2.12)	2.46 (1.40)	0.71 (1.07)	0.62 (0.91)	0.81 (1.01)	2.10 (1.71)
IAIF	15.25 (1.47) ***	9.43 (1.68) **	4.09 (1.31) ***	4.07 (1.35) ***	3.00 (0.83) ***	2.86 (0.72) ***	4.49 (1.60) ***	4.14 (1.34) ***
LP	12.93 (1.40) ***	9.92 (0.29) ***	9.66 (0.05) ***	8.59 (1.23) ***	9.64 (0.11) ***	9.65 (0.06) ***	10.01 (0.31) ***	9.91 (0.67) ***

el test de *suma de rangos de Wilcoxon* [93]. En todos los escenarios, el desempeño de las estrategias desarrolladas es diferente al mostrado por los métodos restantes.

Lo expuesto hasta aquí, nos permite afirmar que incorporar el modelo TARX del tracto vocal en el MEEG de la fonación y, paralelamente, aplicar los métodos en espacio de estados produce representaciones más exactas de la información espectral del tracto vocal, superando incluso a los métodos IAIF y LP. Esto último se observó para todos los escenarios analizados, salvo para condiciones muy severas de ruido acústico. Además, esta última simulación prueba que las estrategias propuestas resultan muy robustas ante variaciones de GNR y f_0, pero no así respecto a las variaciones de SNR. Es notable la forma en que se deterioran las estimaciones de la información espectral del tracto vocal al aumentar el ruido acústico, tanto para las estrategias propuestas como para las otras alternativas consideradas.

7.6.2. Transición entre vocales

En el desarrollo de la Sec. 7.3.2 empleamos un modelo TARX, lo que constituye un diferencia importante con respecto a los métodos clásicos [42, 128, 130]. La principal característica de esta familia de modelos consiste en que sus coeficientes varían con el tiempo, es decir, son series temporales. Por ello, es esperable que el MEEG de la fonación sea capaz de representar adecuadamente los fenómenos transitorios que ocurren en la fonación al unirse dos o más fonemas sonoros. En esta sección, describiremos el desempeño del filtrado inverso en espacio de estados en una señal artificial correspondiente a la transición rápida entre dos vocales.

Aquí, nos concentraremos en el estudio del hiato /ea/, simulado como el pasaje de una vocal /e/ estable a una vocal /a/ estable considerando una transición rápida. La información espectral del filtro del tracto vocal para ambas vocales, en su configuración estable, se presenta en la Tab. 7.1. Se informan las cuatro primeras formantes y sus anchos de banda respectivos. Para describir la transición, construi-

mos trayectorias para las formantes y los anchos de banda caracterizadas por una transición marcada, con duración aproximada de 30 ms y con forma sinusoidal. Con esta información, construimos el filtro del tracto vocal variante en el tiempo.

Generamos la función glótica empleando el modelo LF, tomando una frecuencia fundamental constante $f_0 = 108$ Hz. Luego, le adicionamos ruido de aspiración de tal forma que GNR = 40 dB, de acuerdo a lo recomendado en [79]. El propósito de este ruido es mejorar la naturalidad de la señal sintetizada. Obtuvimos la señal de voz filtrando la función glótica con el filtro del tracto vocal construido. Finalmente, agregamos ruido acústico de forma tal que SNR = 50 dB, simulando un entorno de grabación con condiciones acústicas favorables.

Trabajando con la señal de voz, empleamos las expresiones de la Sec. 7.5 para calcular, de forma iterativa, los parámetros óptimos del MEEG de la fonación. Para ello, obtuvimos previamente los valores iniciales de $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$ aplicando el método IAIF para estimar la función glótica y ajustando el modelo LF a la señal obtenida. Contando con el MEEG, procedimos a estimar los vectores de estados empleando el suavizado de Kalman. Recordando la definición (7.13), obtuvimos la función glótica $\hat{v}_g^{\mathrm{ES}}$ y los coeficientes del modelo TARX del tracto vocal

En la Fig. 7.4, presentamos los resultados obtenidos con el filtrado inverso en espacio de estados para el hiato /ea/ sintetizado. En la fila superior, mostramos 80 ms de la señal de voz. En esta imagen, podemos apreciar claramente el cambio en la forma de onda de los ciclos conforme aumenta el tiempo. Este cambio es más pronunciado en la porción central de la señal, correspondiente al intervalo (50, 80) ms. Los métodos tradicionales de procesamiento de la voz suelen emplear ventanas de señal de $20 - 50$ ms de duración. Para el caso analizado, estos métodos producirán como resultado valores promedios de las formantes [104, 130]. Se desprende entonces que las propuestas desarrolladas en este capítulo darán lugar a una estrategia superadora, ya que permitirán un seguimiento puntual de las trayectorias de las formantes. A continuación, probaremos esta afirmación para el ejemplo analizado.

En la segunda fila de la Fig. 7.4 mostramos la estimación de la función glótica $\hat{v}_g^{\mathrm{ES}}$, obtenida con los métodos en espacio de estados. A fines comparativos, presentamos también la función glótica v_g^{LF} empleada en la síntesis, en línea discontinua. Analizando estas dos curvas, se desprende que la estimación obtenida es adecuada, ya que captura la morfología de la función glótica original para todo el intervalo considerado. A su vez, calculamos el error cuadrático medio, relativo al parámetro E_e, entre la estimación y la función glótica original para el intervalo de señal considerado. Obtuvimos un error de 1,48 %, lo que confirma que la estimación resultó satisfactoria.

En la región inferior de la Fig. 7.4, presentamos el espectrograma de la señal de voz estimado de forma paramétrica a partir de los coeficientes del modelo TARX del tracto vocal, ver Ec. (7.11). En líneas discontinuas negras señalamos las trayectoria estimada para cada formante, obtenidas a partir de la expresión (3.6). Se presenta además, en línea continua blanca, el comportamiento teórico impuesto a las formantes, con el propósito de poder comparar los resultados alcanzados. Podemos apreciar que las curvas estimadas coinciden con las trayectorias teóricas, incluso en el intervalo (50, 80) ms donde es más marcada la transición. Calculamos además el error cuadrático medio entre la trayectoria estimada y la teórica para cada formante. Obtuvimos los siguientes errores: 2,12 Hz para F_1, 3,69 Hz para F_2, 1,80 Hz para F_3 y 4,67 Hz para F_4. Todo esto nos permite afirmar que los métodos desarrollados

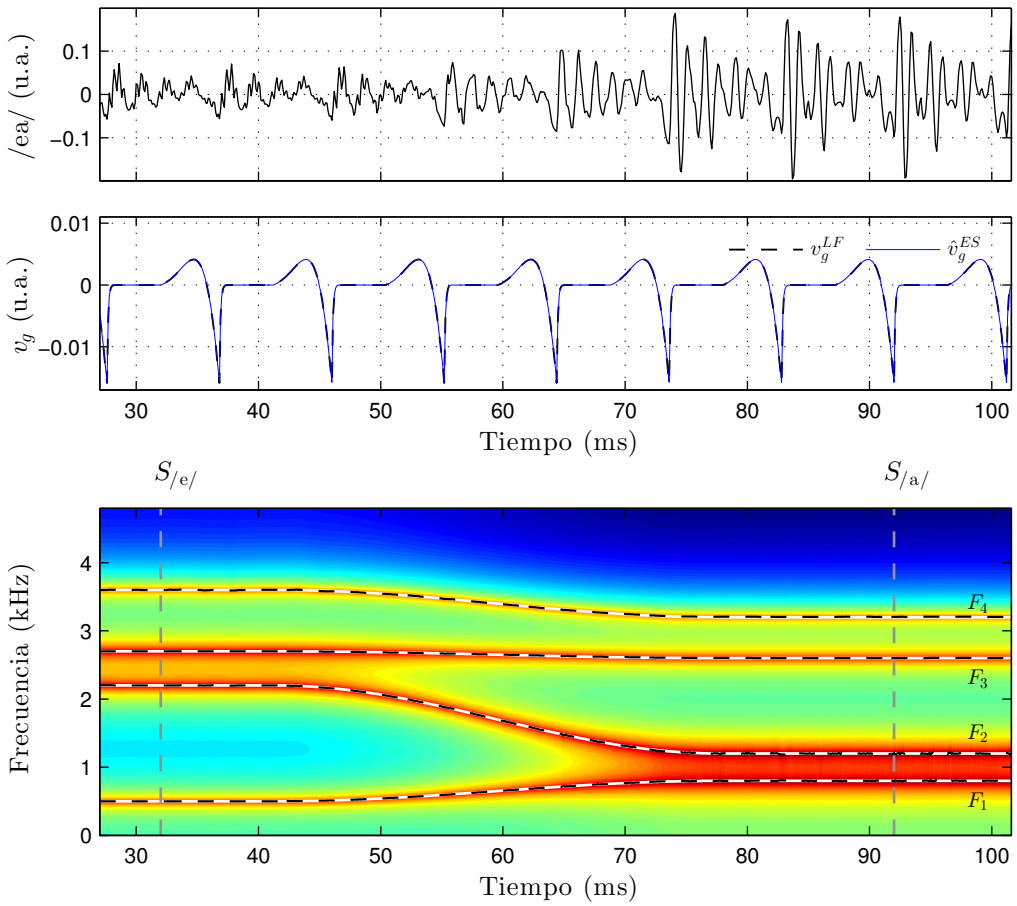


Figura 7.4: Filtrado inverso en espacio de estados aplicado a un hiato /ea/ artificial (SNR = 50 dB, GNR = 40 dB, f_0 = 108 Hz). *Arriba:* forma de onda de la señal de voz. *Segunda fila:* función glótica v_g^{LF} original y su estimación $\hat{v}_g^{\mathrm{ES}}$ aplicando el modelo estocástico. *Abajo:* espectrograma calculado paramétricamente a partir de los coeficientes estimados del modelo TARX del tracto vocal. Se muestran además las trayectorias teóricas y estimadas de las cuatro primeras formantes.

resultaron adecuados para el seguimiento de las trayectorias de las formantes para el intervalo de señal considerado.

Por otro lado, en la Fig. 7.5 mostramos los espectros de potencia estimados $\hat{S}_{/\mathrm{e}/}^{MEEG}$, fila superior, y $\hat{S}_{/\mathrm{a}/}^{MEEG}$, fila inferior, correspondientes a la información indicada por las líneas discontinuas verticales en el espectrograma. Podemos observar que estos espectros pertenecen a las regiones donde se comportan de forma estable las vocales /e/ y /a/, respectivamente. Para cada estimación, presentamos también el correspondiente espectro de potencia teórico construido con la información de la Tab. 7.1. Estas imágenes nos permiten afirmar, nuevamente, que los métodos propuestos son capaces de estimar de forma adecuada la información espectral del tracto vocal para fonemas vocales.

El ejemplo analizado, arroja mayores precisiones respecto a la elección del modelo TARX para construir el MEEG de la fonación. Recordemos que trabajamos bajo la hipótesis de que los coeficientes a_l con $l = 1, 2, \ldots, \rho$ varían de forma estocástica respecto al tiempo, siguiendo la regla (7.12). Durante el desarrollo del MEEG

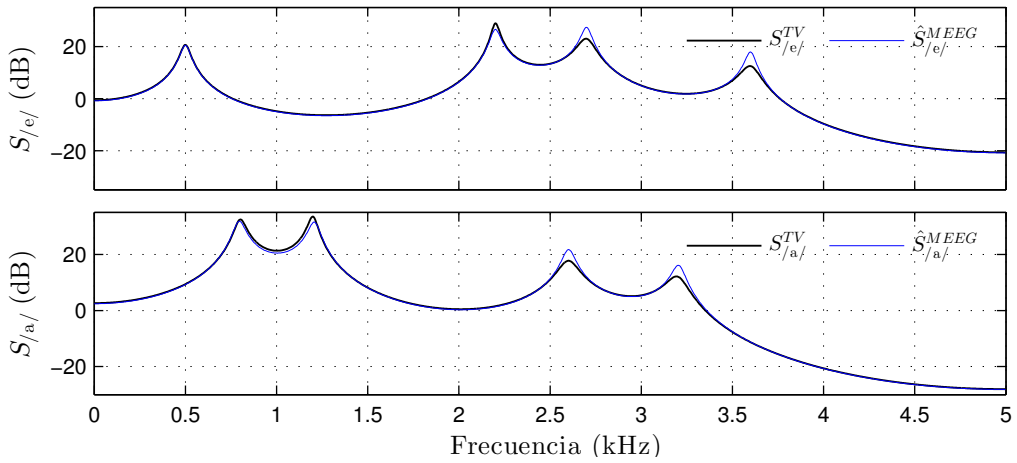


Figura 7.5: Estimaciones de los espectros de potencia correspondientes a las regiones verticales $S_{/e/}$, arriba, y $S_{/a/}$, abajo, indicadas con líneas discontinuas en el espectrograma de la Fig. 7.4. En cada caso, se presentan la estimación $\hat{S}^{MEEG}$ construida de forma paramétrica a partir del modelo TARX del tracto vocal y el espectro de potencia $\hat{S}^{TV}$ empleado en la síntesis.

explicamos, a su vez, que las dinámicas de los coeficientes dependen principalmente de la matriz de covarianza $\mathbf{Q}_\xi$, de acuerdo a la definición (7.16). Como resultado, todo cambio en la matriz $\mathbf{Q}_\xi$ altera la flexibilidad del MEEG de la fonación para representar adecuadamente la información espectral del tracto vocal.

La Fig. 7.6 nos permite ilustrar la afirmación del párrafo anterior. Presentamos los espectrogramas correspondiente al hiato /ea/ artificial (ver la fila superior de la Fig. 7.4) estimados para tres valores diferentes de $\mathbf{Q}_\xi$. En todos ellos, indicamos la trayectoria teórica de cada formante en línea continua blanca, y en líneas discontinuas negras las estimaciones de esas trayectorias calculadas con los coeficientes del modelo TARX del tracto vocal, empleando la expresión (3.6). Centramos la atención en el intervalo (45, 75) ms, donde la transición entre las vocales es más marcada.

El espectrograma de la izquierda se generó a partir del valor óptimo de $\hat{\mathbf{Q}}_\xi$, calculado con la estrategia de la Sec. 7.5 para la estimación de los parámetros del MEEG de la fonación. Coincide con la franja vertical correspondiente al intervalo (45, 75) ms del espectrograma mostrado en la parte inferior de la Fig. 7.4. Anteriormente analizamos este resultado e informamos el error cometido al estimar las trayectorias de las formantes.

En el centro de la Fig. 7.6, mostramos el espectrograma obtenido escalando la matriz de covarianza óptima de la siguiente forma: $0{,}001\,\hat{\mathbf{Q}}_\xi$. Podemos apreciar que modificar la matriz de covarianza de esa forma deteriora la flexibilidad con que el modelo TARX del tracto vocal puede seguir los cambios en las formantes. No obstante, observamos que las trayectorias estimadas acompañan el comportamiento teórico de las formantes. Analizamos objetivamente esto último, estudiando la diferencia entre la trayectoria estimada y la teórica para cada formante. Los errores cuadráticos medios arrojados fueron los siguientes: 14,05 Hz para F_1, 57,68 Hz para F_2, 14,96 Hz para F_3 y 65,94 Hz para F_4. Comparando estos valores con los reportados antes para el caso óptimo, encontramos que el error de estimación resultó mayor para todas las formantes. Como era de esperarse, los errores mayores se obtuvieron

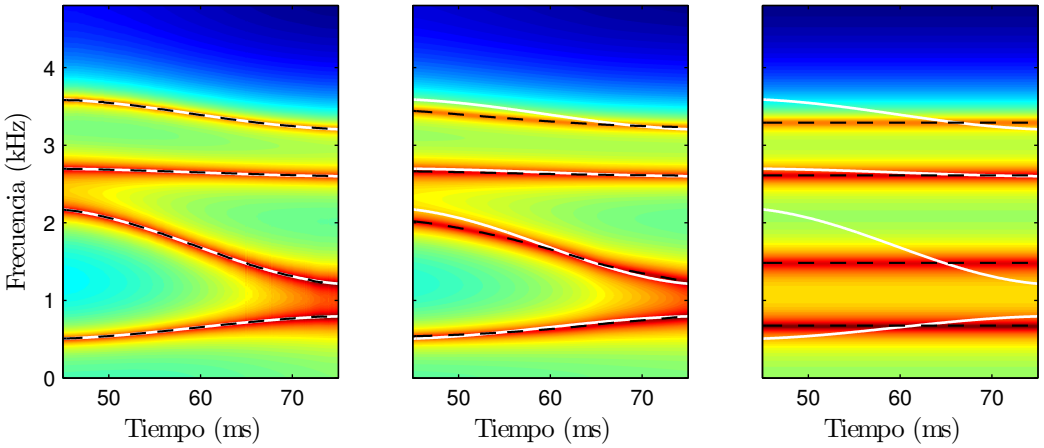


Figura 7.6: Espectrogramas para el hiato /ea/ artificial para diferentes valores de la matriz de covarianza $\mathbf{Q}_\xi$. *Izq.:* valor óptimo $\hat{\mathbf{Q}}_\xi$ de la matriz de covarianza. *Centro:* matriz de covarianza igual a 0,001 $\hat{\mathbf{Q}}_\xi$. *Der.:* condición límite $\hat{\mathbf{Q}}_\xi \to \mathbf{0}$. Las líneas continuas blancas indican las trayectorias teóricas de las formantes. Las líneas discontinuas negras representan las estimaciones de las trayectorias de las formantes.

para las formantes F_2 y F_4, siendo éstas las que sufren una mayor variación entre las vocales analizadas (ver Tab. 7.1).

De forma análoga, presentamos en la imagen de la derecha el espectrograma para el ejemplo bajo estudio, teniendo en cuenta la condición límite $\hat{\mathbf{Q}}_\xi \to \mathbf{0}$. Este escenario permite simular el uso de un modelo ARX para representar el tracto vocal, situación en la cual los coeficientes en la Ec. (7.11) se consideran constantes. De este resultado se desprende que el modelo ARX produce estimaciones promedio o globales de las formantes para la ventana de señal analizada. Por otro lado, los errores de estimación obtenidos en este caso fueron: 139,73 Hz para F_1, 506,01 Hz para F_2, 59,71 Hz para F_3 y 212,50 Hz para F_4. Como era de esperarse, estos valores resultaron considerablemente mayores a los errores obtenidos en los dos escenarios analizados anteriormente. Esto permite afirmar que el modelo TARX garantiza una flexibilidad adecuada para la estimación de la información espectral del tracto vocal.

7.7. Resultados en señales reales

En esta sección, analizaremos el desempeño del filtrado inverso en espacio de estados al aplicarlo en la descomposición de señales reales de voz. Los ejemplos estudiados a lo largo de esta exposición pertenecen a la base de datos desarrollada por nosotros en el LSyDnL, la cual cuenta con señales de voz y de EGG registradas de forma simultánea para diferentes emisiones de interés en la fonoaudiología. Originalmente, estas señales se digitalizaron con una frecuencia de muestreo de 50 kHz. Para llevar a cabo las simulaciones, éstas se submuestrearon posteriormente a 10 kHz.

Como vimos antes, para construir un MEEG de la fonación se necesita información específica de la función glótica. Por un lado, se requieren los sucesivos GOI y los GCI. Para casos reales, estos instantes no se conocen y, por ello, deben calcularse a partir de la señal de voz u otras señales biomédicas. En las simulaciones realizadas procedimos de la siguiente forma. Primeramente, se generan puntos candi-

datos procesando el EGG correspondiente a la señal de voz con el algoritmo SIGMA [160]. Luego, estos valores se adecuan corrigiendo el pequeño retardo, décimas de milisegundos, que presenta la señal de voz con respecto al EGG [46].

Por otro lado, se necesitan valores iniciales adecuados de los parámetros $\tilde{A}_g$, $\tilde{B}_g$ y $\tilde{C}_g$. Para calcularlos, empleamos el método IAIF para estimar la función glótica y ajustamos el modelo LF a la señal obtenida, de forma similar a como procedimos en los ejemplos de la sección anterior. En señales reales, los mejores resultados se obtuvieron para $p^{IAIF} = 10$ y $g^{IAIF} = 4$. A su vez, en los MEEG de la fonación se consideraron modelos TARX de orden $\rho = 10$ para representar el tracto vocal.

7.7.1. Vocales sostenidas

Describiremos, a continuación, cómo se comportan los métodos desarrollados en este capítulo al emplearlos para la descomposición de señales reales de vocales sostenidas. En particular, consideraremos aquí una vocal /a/ sostenida producida por un hombre adulto normal. Para realizar este estudio, seleccionamos una porción de señal de 100 ms de duración.

A partir de la señal de voz, obtuvimos la información de la función glótica empleando el procedimiento previamente explicado. Luego, calculamos los parámetros del MEEG de la fonación, aplicando la estrategia de optimización descripta en la Sec. 7.5. Los parámetros obtenidos sirvieron para construir el MEEG de la fonación particular para la vocal /a/ bajo estudio. Respetando este modelo, estimamos la función glótica y los coeficientes del filtro TARX del tracto vocal haciendo uso del suavizado de Kalman. Esto último, constituye el filtrado inverso en espacio de estados de la señal de voz.

En la Fig. 7.7 presentamos los resultados alcanzados con el procedimiento descripto en el párrafo anterior para la vocal /a/ seleccionada. En la parte superior, podemos observar la forma de onda de esta señal. Se caracteriza por un período fundamental promedio $T_0 = 10{,}90$ ms, lo que implica una frecuencia fundamental $f_0 = 91{,}74$ Hz. Vemos que esta señal exhibe un comportamiento regular y estable para todo el intervalo considerado.

En la segunda fila, mostramos la estimación de la función glótica $\hat{v}_g^{\mathrm{ES}}$ para la señal de voz, en función del modelo estocástico desarrollado al comienzo de este capítulo. Se presenta también la función glótica $\hat{v}_g^{\mathrm{IAIF}}$ calculada con el método IAIF, con el propósito de emplearla a fines comparativos. Nuevamente, podemos apreciar que ambas estimaciones son satisfactorias, mostrando un comportamiento adecuado para los sucesivos ciclos glóticos. Esto se nota especialmente para las muestras cercanas a los GCI. A su vez, observamos que $\hat{v}_g^{\mathrm{ES}}$ exhibe una dinámica más suave, en comparación con $\hat{v}_g^{\mathrm{IAIF}}$. En particular, las fluctuaciones presentes en $\hat{v}_g^{\mathrm{IAIF}}$, se muestran de manera atenuada en $\hat{v}_g^{\mathrm{ES}}$. Esto se aprecia fácilmente en la porción final de la fase de cierre de cada ciclo.

En la tercera fila, presentamos la estimación del flujo de aire glótico $\hat{U}_g^{\mathrm{ES}}$ asociada al modelo estocástico de la función glótica. A fines comparativos, presentamos la estimación de esta señal alcanzada con el método IAIF. Se desprende que ambas estimaciones se comportan de forma similar para cada ciclo, mostrando diferencias en la región posterior a los GCI. Es importante destacar que estas estimaciones y las presentadas en la segunda fila resultaron muy similares a las reportadas por otros autores [1, 3, 4, 67, 147]. Esto sugiere que el filtrado inverso en espacio de estados

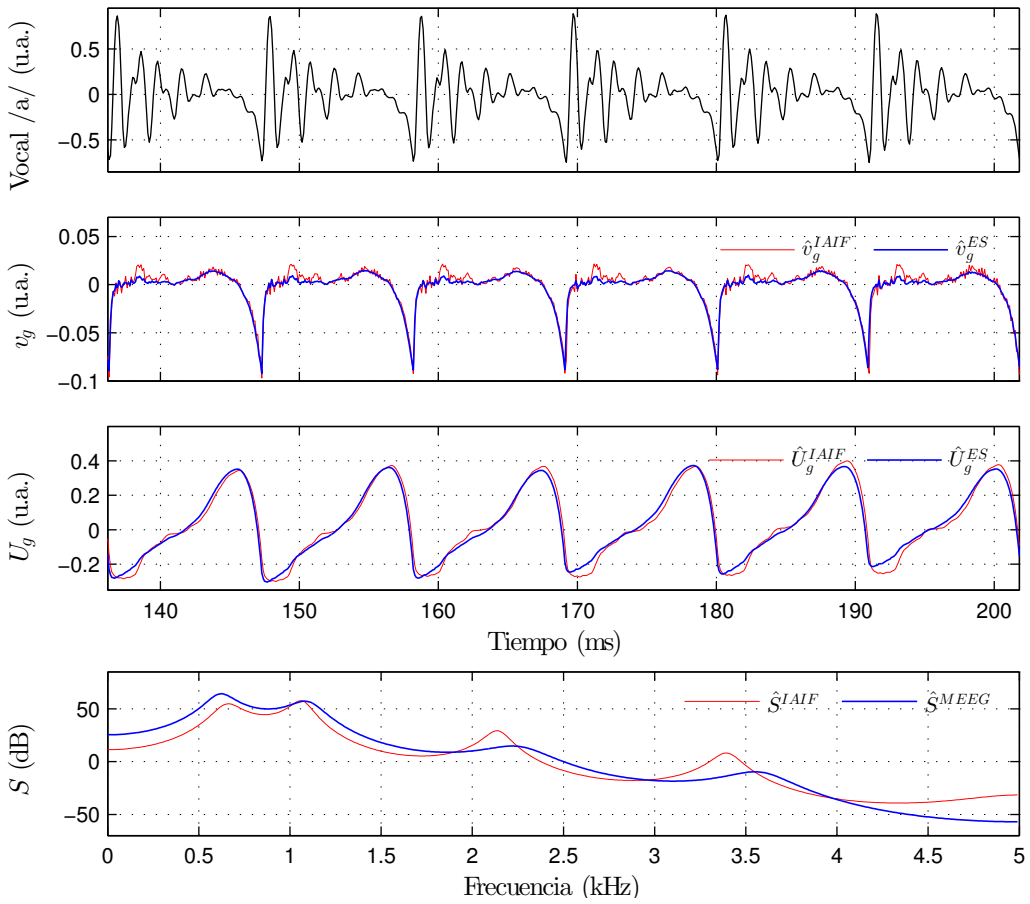


Figura 7.7: Filtrado inverso en espacio de estados de una vocal /a/ sostenida correspondiente a un hombre adulto sano. *Arriba:* forma de onda de la señal. *Segunda fila:* estimaciones de la función glótica generadas con el modelo estocástico $\hat{v}_g^{\mathrm{ES}}$ y con el método IAIF $\hat{v}_g^{\mathrm{IAIF}}$. *Tercera fila:* estimaciones del flujo de aire glótico calculadas con las estrategias propuestas $\hat{U}_g^{\mathrm{ES}}$ y con el método IAIF $\hat{U}_g^{\mathrm{IAIF}}$. *Abajo:* espectros de potencia en dB estimados considerando el MEEG de la fonación $\hat{S}^{MEEG}$ y con el método IAIF $\hat{S}^{IAIF}$.

es una estrategia adecuada para la estimación de la función glótica y el cálculo del flujo de aire a partir de señales reales de vocales sostenidas.

En la parte inferior de la Fig. 7.7, presentamos el espectro de potencia en dB $\hat{S}^{MEEG}$ estimado paramétricamente con el modelo TARX del tracto vocal, considerando las estimaciones de sus coeficientes obtenidas con el MEEG de la fonación. Esta curva corresponde al espectro de potencia promedio para la ventana de señal analizada. Con el objetivo de comparar este resultado, mostramos también el espectro de potencia $\hat{S}^{IAIF}$ calculado paramétricamente con el método IAIF. Podemos apreciar que estas estimaciones difieren entre sí notablemente, tanto en la ubicación de las formantes como en la morfología. Estas diferencias ayudan a explicar las fluctuaciones encontradas en $\hat{v}_g^{\mathrm{IAIF}}$. Cometer un error al calcular la información espectral del tracto vocal ocasiona un comportamiento oscilatorio en la función glótica [4, 55]. Por ello, el comportamiento más suave que presenta $\hat{v}_g^{\mathrm{ES}}$ sugiere que $\hat{S}^{MEEG}$ captura mejor la información espectral del tracto vocal.

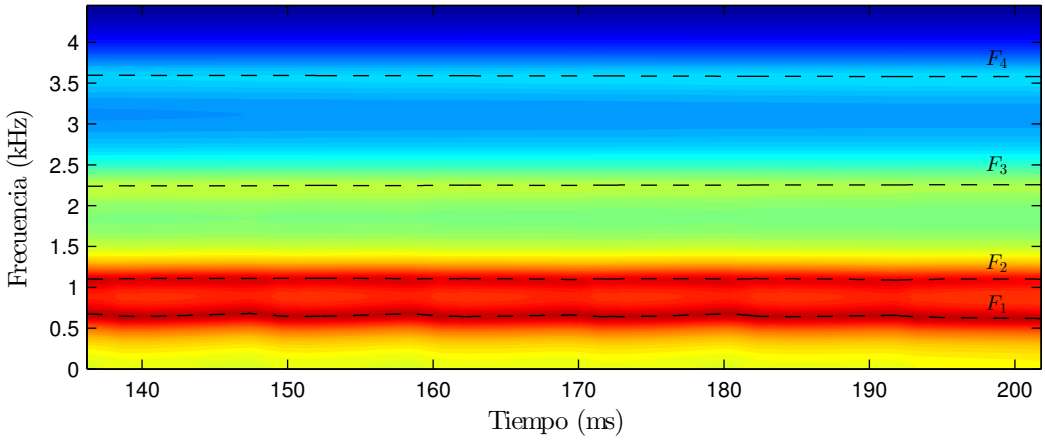


Figura 7.8: Espectrograma correspondiente a la vocal /a/ sostenida de la Fig. 7.7. Las líneas discontinuas negras representan las estimaciones de las trayectorias de las formantes.

En la Fig. 7.8, mostramos el espectrograma para la señal bajo estudio, construido de forma paramétrica a partir de las estimaciones de los coeficientes del modelo TARX del tracto vocal. Presentamos además, de forma superpuesta, las trayectorias de las cuatro primeras formantes, calculadas a partir de estos coeficientes empleando la expresión (3.6). Podemos apreciar que, aun cuando contemplamos un comportamiento no estacionario, las formantes se mantuvieron prácticamente constantes en el intervalo analizado. Esto confirma la validez de la hipótesis de que en vocales sostenidas el tracto vocal puede considerarse estacionario para ventanas de corta duración. Considerando esto, obtuvimos los siguientes valores promedio para cada formante: 650,50 Hz para F_1, 1104,12 Hz para F_2, 2248,56 Hz para F_3 y 3589,44 Hz para F_4. Éstos resultaron muy similares a las formantes reportadas en la Tab. 2.3 para vocales /a/ de sujetos masculinos.

7.7.2. Transición entre vocales

En esta sección, ilustraremos cómo se comporta el filtrado inverso en espacio de estados al aplicarlo en la descomposición de una señal real de voz correspondiente a una transición entre vocales.

De forma análoga a como procedimos en la Sec. 7.6.2, en lo que sigue nos concentraremos en un hiato /ea/. Para este estudio, escogimos una señal de voz registrada para un hombre adulto normal. A partir de esta señal y de su correspondiente EGG calculamos los GOI y los GCI, así como los parámetros iniciales para el modelo estocástico de la función glótica. Esta información nos permitió definir el MEEG de la fonación para esta señal. Luego, calculamos los parámetros óptimos para este modelo. Una vez ajustado el MEEG, empleamos el suavizado de Kalman con el propósito de estimar la función glótica y la información espectral del tracto vocal para la señal analizada.

En la Fig. 7.9 presentamos los resultados alcanzados con el procedimiento descripto arriba. En la fila superior mostramos 75 ms de la señal de voz del hiato /ea/. Podemos observar que la forma de onda de la señal cambia de manera gradual. Además, existe una similitud importante entre las formas de los ciclos al comienzo y al

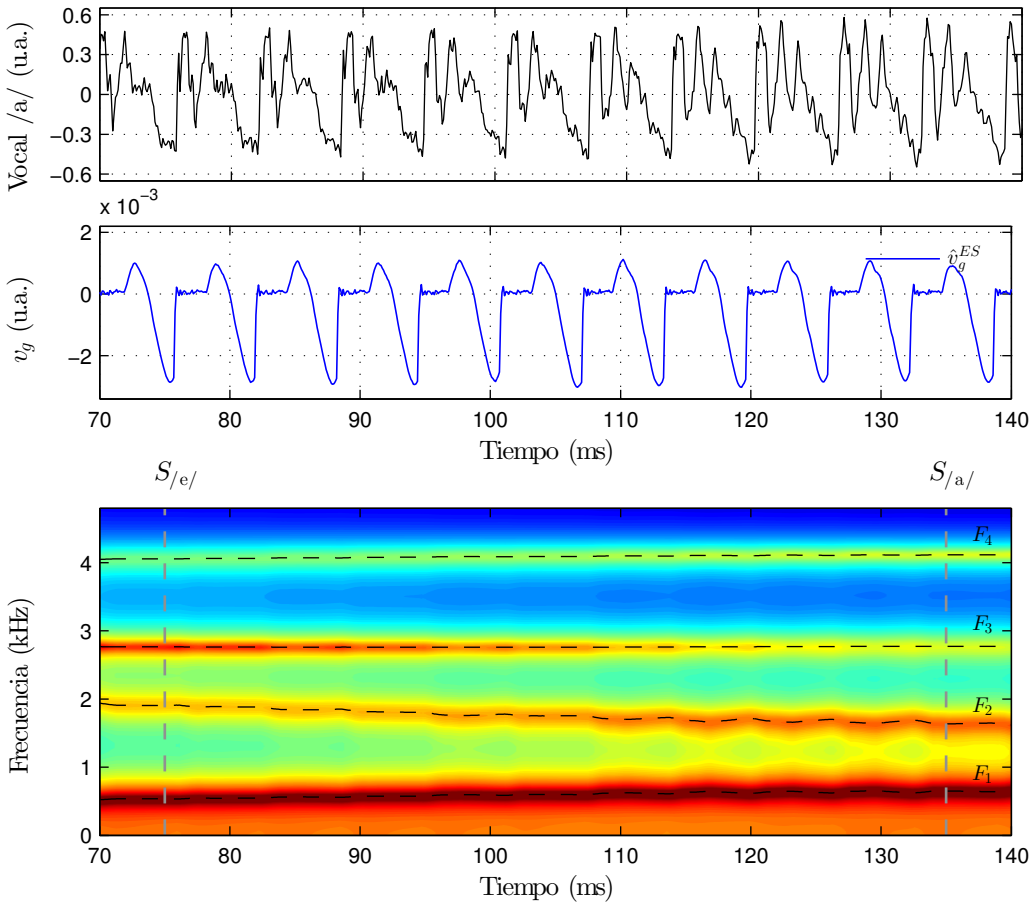


Figura 7.9: Filtrado inverso en espacio de estados aplicado a un hiato /ea/ real correspondiente a un hombre adulto normal. *Arriba:* forma de onda de la señal de voz. *Segunda fila:* estimación de la función glótica $\hat{v}_g^{\mathrm{ES}}$ correspondiente al modelo estocástico. *Abajo:* espectrograma de la señal de voz calculado de forma paramétrica a partir de los coeficientes estimados del modelo TARX del tracto vocal. Se presenta además las trayectorias estimadas de las cuatro primeras formantes.

final de esta señal. En la segunda fila, presentamos la estimación de la función glótica $\hat{v}_g^{\mathrm{ES}}$ de acuerdo al modelo estocástico desarrollado. Esta señal se caracteriza por un comportamiento muy regular, independientemente de los cambios observados en la señal de voz. A su vez, estudiando los ciclos sucesivos podemos apreciar algunas aperiodicidades o comportamientos singulares en la forma de onda. Esto se aprecia mejor en la OP para los elementos posteriores a cada máximo.

En la región inferior de la Fig. 7.9, presentamos el espectrograma para el hiato bajo estudio, estimado de forma paramétrica a partir de los coeficientes del modelo TARX del tracto vocal. En líneas discontinuas negras señalamos las trayectorias estimadas para cada formante, obtenidas a partir de la expresión (3.6). Podemos observar una transición importante en F_2, acompañada por un cambio de menor importancia en F_1. Por su parte, las formantes F_3 y F_4 evidencian un comportamiento aproximadamente constante. Estos resultados explican el cambio gradual observado en la forma de onda de la señal de voz.

En la Fig. 7.10 presentamos los espectros de potencia estimados $\hat{S}_{/\mathrm{e}/}^{MEEG}$, fila

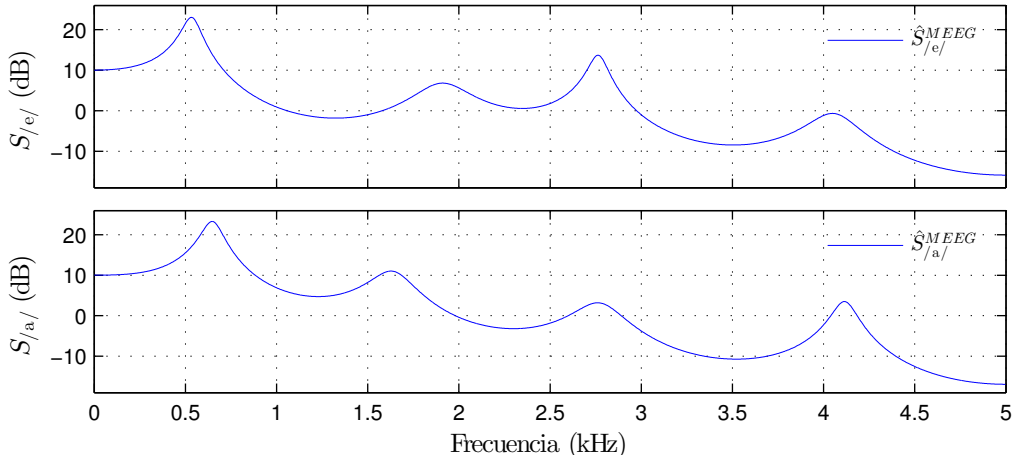


Figura 7.10: Estimaciones de los espectros de potencia correspondientes a las regiones verticales $S_{/e/}$, arriba, y $S_{/a/}$, abajo, indicadas con líneas discontinuas en el espectrograma de la Fig. 7.9. Se presentan las estimaciones $\hat{S}^{MEEG}$ construidas de forma paramétrica a partir de la información obtenida con los métodos propuestos.

Tabla 7.5: Valores de las primeras cuatro formantes y sus anchos de banda correspondientes estimados a partir de los espectros de potencia de la Fig. 7.10.

	Formantes (Hz)				**Ancho de Banda** (Hz)			
	F_1	F_2	F_3	F_4	B_1	B_2	B_3	B_4
/a/	649,0	1638,0	2772,5	4116,0	113,1	225,0	259,0	128,3
/e/	535,9	1907,3	2765,0	4057,0	96,2	294,5	99,2	260,1

superior, y $\hat{S}^{MEEG}_{/a/}$, fila inferior, correspondientes a las regiones indicadas por las líneas discontinuas verticales en el espectrograma estudiado anteriormente. Estas curvas caracterizan las porciones asociadas a las vocales /e/ y /a/ en la señal de voz, respectivamente. Nuevamente, podemos apreciar que estos espectros de potencia difieren principalmente en el comportamiento de la segunda formante. En menor medida, observamos diferencias en las ubicaciones de la primera y cuarta formante, y en la amplitud de la tercera formante.

En la Tab. 7.5, informamos los valores estimados para las cuatro primeras formantes y sus anchos de banda, para los dos espectros de potencias de la Fig. 7.10. La información referida a las formantes corrobora lo expuesto anteriormente. Por otro lado, los valores grandes de los anchos de banda permiten explicar el ancho y la pendiente suave de los picos que caracterizan las formantes. Los estudios expuestos hasta aquí demostraron que en las señales reales el desempeño de las estrategias desarrolladas resultó muy similar al mostrado en los ejemplos artificiales.

Uno de los objetivos de los trabajos expuestos aquí fue generar una representación flexible de la función glótica, que sea capaz de capturar las aperiodicidades o comportamientos particulares que se observan en las voces reales. Esta premisa motivó el desarrollo del modelo estocástico de la función glótica, en la Sec. 7.3.1. El ejemplo analizado en esta sección nos permite brindar mayores precisiones respecto al comportamiento del modelo estocástico obtenido.

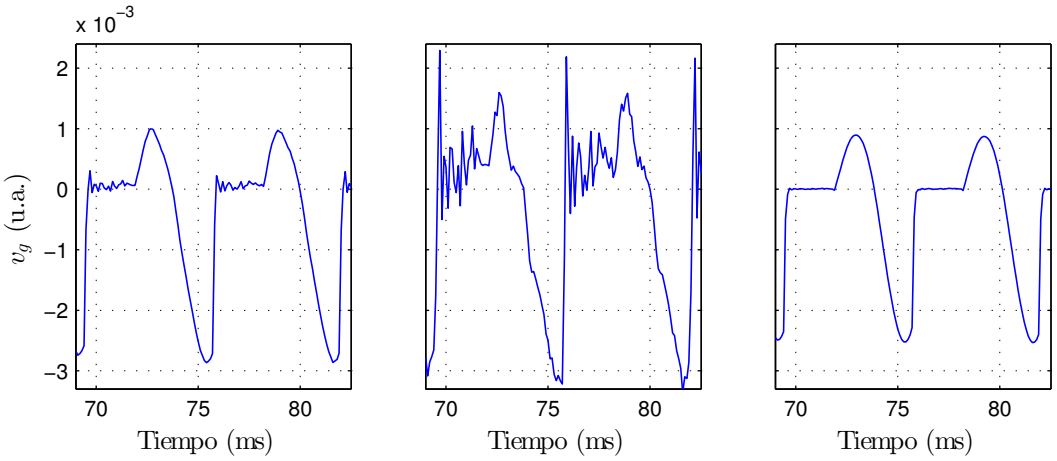


Figura 7.11: Estimaciones de la función glótica para el hiato /ea/ de la Fig. 7.9, generadas a partir de diferentes valores de la varianza σ_ν^2. *Izq.:* varianza óptima $\hat{\sigma}_\nu^2$. *Centro:* varianza igual a $10\,\hat{\sigma}_\nu^2$. *Der.:* varianza igual a $0{,}1\,\hat{\sigma}_\nu^2$.

En la Fig. 7.11, presentamos tres estimaciones de la función glótica calculadas para diferentes valores de la varianza σ_ν^2. Estas tres estimaciones se obtuvieron a partir del hiato /ea/ bajo estudio. En la imagen de la izquierda, presentamos la función glótica obtenida con la varianza óptima $\hat{\sigma}_\nu^2$, que coincide a su vez con la dibujada en la segunda fila de la Fig. 7.9. En el centro, mostramos la estimación calculada empleando una varianza mayor al valor óptimo. Así, aumentar la varianza repercute en la generación de estimaciones que admiten un nivel mayor de aperiodicidades, dando lugar a formas de onda más diversas. Por otro lado, la curva de la derecha corresponde a la función glótica calculada con una varianza menor al valor óptimo. En este caso, disminuir la varianza favorece la obtención de funciones glóticas más suaves, ya que se fuerza a la forma de onda para que resulte similar al modelo LF.

Lo expuesto en el párrafo anterior nos permite afirmar que tanto los parámetros A_g, B_g y C_g como la varianza σ_ν^2 y los valores de los GOI y GCI influyen en la estimación de la función glótica. La forma de onda resultante dependerá, en gran medida, de cómo se combine toda esta información.

7.8. Discusiones

Finalizaremos este capítulo, comentando brevemente algunos detalles importantes, los cuales advertimos durante la realización de las diferentes simulaciones expuestas o que cobraron relevancia a la hora de aplicar los métodos propuestos.

7.8.1. Dificultad en el cálculo de los GOI y GCI

En reiteradas oportunidades señalamos que para implementar el MEEG de la fonación es necesario conocer previamente los GOI y los GCI. Sin embargo, esta tarea es muy difícil en la práctica. En los trabajos realizados encontramos que la calidad de la función glótica estimada con el modelo estocástico desarrollado depende, en gran medida, de la precisión con que se obtuvieron estos instantes. Cualquier error en el cálculo, repercutirá negativamente en la estimación generada.

Afortunadamente, en este último tiempo han surgido diferentes métodos para el cálculo de los GOI y GCI a partir de la señal de voz. Algunos de ellos se apoyan en la información aportada por el EGG [160], mientras que otras estrategias se concentran sólo en la señal de voz [48, 159]. Sin embargo, puede persistir algún error en el cálculo de estos instantes o en el alineamiento con la señal de voz, lo que deteriora el comportamiento de los métodos propuestos. Consideramos que éste es un aspecto que merece ser analizado con mayor profundidad en el futuro.

7.8.2. Estimación de los parámetros del MEEG de la fonación y desempeño de la penalización

Uno de los objetivos considerados durante el desarrollo del MEEG de la fonación era generar una estrategia lineal para describir la función glótica. De este modo, se disminuye la dificultad del problema de optimización involucrado en la estimación de los parámetros. Sin embargo, para garantizar la estabilidad en la búsqueda de la solución fue necesario agregar un término de penalización en la función objetivo. Aun cuando fue una estrategia útil, esta alternativa mostró dos importantes desventajas. Debido a la definición de la función de penalización Φ, ésta no contempla variaciones en la cantidad de elementos N en la señal de voz. Por otro lado, el parámetro de penalización λ_j debe fijarse arbitrariamente, tomando como criterio de selección la convergencia del procedimiento de estimación. Por todo esto, consideramos que la función de penalización debe modificarse teniendo en cuenta estos dos aspectos.

7.8.3. Frecuencia de muestreo de la señal de voz

En todos los ejemplos presentados en este capítulo, consideramos señales con frecuencia de muestreo de 10 kHz. Las razones de esta elección son que los primeros sistemas de comunicación trabajaban con frecuencias en ese orden y que existen numerosos trabajos científicos donde se considera esta frecuencia de muestreo. Sin embargo, en la actualidad se requieren estrategias adecuadas para el procesamiento de señales de voz de mayor resolución, digitalizadas con una frecuencia de muestreo mucho mayor. Teniendo en cuenta esto, creemos que es necesario estudiar el desempeño de los modelos y las estrategias propuestas en este capítulo para el análisis de esta clase de registros de voz.

7.8.4. Evaluación objetiva en señales reales

En la Sec. 7.7 describimos algunos resultados obtenidos con los métodos desarrollados al trabajar con señales reales de voz. La interpretación de estos resultados se llevó a cabo desde un punto de vista principalmente subjetivo. Un inconveniente importante es la falta de consenso en la bibliografía consultada respecto a qué experimentos permiten evaluar y comparar objetivamente el desempeño de esta clase de métodos en señales reales. Este punto ha sido señalado en el pasado por otros autores [2, 5, 46, 174]. Consideramos que es muy importante abordar esta problemática a la brevedad.

7.9. Comentarios finales

En este capítulo, presentamos el último de los aportes originales de esta tesis de doctorado. Propusimos y desarrollamos una estrategia basada en métodos en espacios de estados que permite, por un lado, modelar el proceso de fonación y, por el otro, llevar a cabo el filtrado inverso de una señal de voz.

En primer lugar, formulamos un modelo estocástico de la función glótica, construido a partir de ecuaciones en diferencias estocásticas. Este desarrollo se basó en la función glótica de Liljencrants y de Fant, ampliamente difundida en la comunidad científica, y presenta una estructura en consonancia con los modelos en espacio de estados lineales y gaussianos. A su vez, el modelo desarrollado permite describir esta señal en el marco de los procesos estocásticos no estacionarios. Hasta donde sabemos, no existe un modelo de la función glótica con características similares al propuesto aquí.

Desarrollamos también un modelo en espacio de estados para describir la fonación. Para ello, combinamos la clásica teoría fuente y filtro con el modelo estocástico de la función glótica propuesto y con los modelos auterregresivos variantes en el tiempo. Obtuvimos así una representación de la fonación muy flexible, capaz de adaptarse a las aperiodicidades o fenómenos particulares que se observan frecuentemente en la señal de voz. Además, este modelo demostró ser muy útil para estimar, a partir de una señal de voz, tanto la función glótica como la información espectral del tracto vocal. Probamos todo esto mediante diferentes estudios llevados a cabo con señales artificiales en condiciones controladas y en ejemplos de voces reales.

En el cierre de este capítulo, describimos algunos aspectos importantes de los métodos propuestos, que se manifestaron a lo largo de las etapas involucradas en este trabajo. Consideramos que abordar próximamente los puntos señalados permitirá, por un lado, mejorar el desempeño de los métodos desarrollados y, por el otro, generar nuevas estrategias superadoras.

Los trabajos involucrados en el desarrollo y la evaluación del modelo estocástico de la función glótica y del modelo en espacio de estados de la fonación, en su versión preliminar, dieron lugar a dos presentaciones en congresos de la especialidad, uno de alcance nacional [4] y otro internacional [7]. A su vez, recientemente elaboramos un artículo científico con los desarrollos presentados en este capítulo, el cual fue enviado a la revista científica *Biomedical Signal Processing and Control* para su correspondiente revisión y evaluación [9].

Capítulo 8

Conclusiones finales y trabajos futuros

En esta tesis de doctorado, propusimos y evaluamos nuevas estrategias para el modelado de la fonación y las señales biomédicas asociadas, formuladas a partir de estructuras estocásticas potentes y flexibles. Para concluir este trabajo, a continuación recapitularemos los principales resultados alcanzados. Además, discutiremos las implicancias de estos desarrollos y comentaremos algunas ideas para continuar en el futuro.

Comenzamos describiendo la anatomía y la fisiología de la fonación en los seres humanos y, a su vez, caracterizando el producto principal de este proceso: la señal de voz. Esto nos llevó a explorar el electroglotograma y las vibraciones en la piel del cuello, dos señales complementarias a la voz que brindan información de las estructuras involucradas en la fonación de forma no invasiva. De este modo afianzamos nuestro conocimiento respecto a las señales biomédicas de la fonación y a la información que aporta cada una de ellas, y ahondamos en el correcto uso de los dispositivos tecnológicos requeridos para registrar estas señales. Lo anterior generó un marco conceptual sólido que permitió el desarrollo de una base de datos compuesta por registros de estas tres señales, obtenidos simultáneamente para diferentes emisiones. Hasta donde sabemos, no existe otra base de datos con características similares. En la actualidad, utilizamos este material para llevar a cabo trabajos científicos y actividades didácticas. Esperamos, próximamente, aumentar la cantidad de registros en la base de datos y, en lo posible, incorporar otras señales biomédicas.

Seguidamente, estudiamos los fundamentos del modelado de la fonación. Debido a que éste es un campo muy extenso, centramos nuestra atención en discutir los conceptos esenciales para el desarrollo de este trabajo. Comenzamos analizando las dos principales teorías para explicar la fonación, señalando sus similitudes y sus diferencias. Esto nos permitió concluir que ambas teorías brindan enfoques complementarios de los fenómenos involucrados en la fonación. Profundizamos luego en la descripción de la teoría *fuente* y *filtro*. Para ello, nos concentramos en cada uno de sus componentes, analizando su interpretación y comentando los métodos disponibles en la actualidad para representarlos. Introdujimos, además, los importantes conceptos de función glótica, filtro de tracto vocal y filtrado inverso de la voz, imprescindibles para el desarrollo de esta tesis de doctorado. Todo esto sirvió, por un lado, para acrecentar nuestro conocimiento en este campo y, por el otro, como inspiración para varios de los aportes realizados.

Proseguimos con el estudio de las perturbaciones de la voz. Describimos primeramente las series temporales de períodos y de amplitudes para una vocal sostenida. A partir de estas señales, introdujimos los conceptos de perturbaciones y fluctuaciones. Nuestro interés en estos dos fenómenos radica en que, si bien son estudiados cotidianamente en la medicina y en la fonoaudiología, los parámetros acústicos desarrollados para cuantificarlos presentan severas limitaciones. Al respecto, desde hace un tiempo estamos trabajando en alternativas superadoras. Algunos de los aportes presentados en este documento surgieron como resultado de este proceso. Este es el caso del método desarrollado para la síntesis de vocales sostenidas con perturbaciones controladas. Esta estrategia presenta la novedad de permitir perturbaciones estocásticas en las series de amplitudes y de períodos, controladas a partir de dos parámetros acústicos muy utilizados en la medicina. Paralelamente, demostramos que las voces sintetizadas presentan una calidad perceptual alta para ventanas de corta duración; este atributo se cuantificó con un método objetivo. Sin embargo, en ventanas de mayor duración las fluctuaciones en las amplitudes y en los períodos cobran una mayor preponderancia en la calidad perceptual. Este fenómeno será estudiado en trabajos futuros. Entre otros aspectos, se espera realizar la evaluación de la calidad perceptual de estas señales a partir de experimentos subjetivos de escucha en una población y, también, investigar qué efecto tiene la naturaleza o el comportamiento de las perturbaciones consideradas en la calidad perceptual de las vocales sintetizadas.

A los efectos de proveer un marco conceptual adecuado para los demás aportes de esta tesis de doctorado, se introdujeron previamente los métodos en espacio de estados. Esta revisión no fue exhaustiva, sino que nos concentramos en los métodos específicos para los modelos en espacio de estados lineales y gaussianos, por ser esta familia de modelos fundamental para las estrategias propuestas en esta tesis. Dada la diversidad de disciplinas científicas que aportan fuentes bibliográficas, se redactó el material desde una perspectiva unificada, en forma concisa y precisa. Incluimos también algunos tópicos frecuentemente omitidos en ingeniería, como por ejemplo las estimaciones suavizadas de las perturbaciones y de la correlación de los vectores de estados, o la estrategia de inicialización difusa.

El segundo aporte original de esta tesis de doctorado consistió en una estrategia novedosa para el análisis y modelado estructural basado en métodos en espacio de estados de las series de amplitudes y de períodos. En este sentido, partimos de la hipótesis de que es posible explicar estas señales suponiendo que son el resultado de la combinación de componentes más simples y con una interpretación directa. Al respecto, propusimos una estructura para llevar a cabo este análisis, contemplando los principales fenómenos identificados en casos reales. Esto nos permitió estudiar las fluctuaciones, los fenómenos cíclicos y las perturbaciones de la voz en un marco conceptual adecuado, bajo la hipótesis de que estos fenómenos presentan un comportamiento estocástico no estacionario. A partir de la estructura propuesta, describimos cómo construir un modelo en espacio de estados, haciendo uso de los métodos diseñados específicamente para ellos. De este modo, pudimos combinar estos métodos con el modelo construido para llevar a cabo el análisis estructural de señales reales. Si bien existen diferentes estrategias para el estudio y modelado de las series de períodos y de amplitudes, no conocemos de ninguna alternativa que incorpore todos los fenómenos involucrados y que, al mismo tiempo, explique su dinámica desde una perspectiva estocástica y sea aplicable al análisis de señales reales.

La estrategia aquí propuesta contempla todos estos aspectos.

Las ventajas de los métodos desarrollados fueron discutidas y estudiadas considerando tanto señales artificiales como reales. Mostramos que los modelos estructurales dan lugar a series temporales artificiales con dinámicas muy variadas, que se asemejan a las observadas en los casos reales. A su vez, el análisis estructural permite estimar, de forma óptima, los componentes de una serie real de períodos o de amplitudes. Estudiando estas estimaciones, identificamos algunos elementos que caracterizan la dinámica glótica, como por ejemplo la estabilidad de la fonación, los ajustes de las cuerdas vocales o la existencia de microtemblores vocales. Por último, demostramos que esta estrategia es adecuada para estudiar diferentes señales provenientes de sujetos normales. Todos estos resultados son muy satisfactorios y nos motivan a continuar trabajando en esta temática.

El último de los aportes de este trabajo versó sobre el desarrollo de estrategias para abordar el modelado de la fonación y el filtrado inverso de una señal de voz, inspiradas en los métodos en espacio de estados. A tal fin se propuso primero una representación alternativa de la función glótica, formulada a partir de una ecuación en diferencias estocástica y no estacionaria. Entre otros, esta estructura presenta dos rasgos distintivos respecto a los modelos clásicos. Por un lado, contempla las perturbaciones o aperiodicidades que se observan normalmente en las señales de voz y, por el otro, su formulación es compatible con los modelos en espacio de estados y con los métodos específicos para ellos. Hasta donde sabemos, ninguna otra representación de la función glótica presenta características similares. Combinando la función glótica desarrollada con filtros lineales variantes en el tiempo, propusimos un modelo en espacio de estados de la fonación que permite una representación precisa y muy flexible de una señal de voz compuesta por fonemas sonoros. El principal resultado alcanzado fue el desarrollo de una estrategia novedosa para la estimación, de forma conjunta y óptima, de la función glótica y del filtro del tracto vocal utilizando los métodos en espacio de estados. Denominamos a este proceso *filtrado inverso en espacio de estados*.

Las simulaciones con señales artificiales realizadas demostraron que el filtrado inverso en espacio de estados es capaz de generar estimaciones precisas de la función glótica y de la información espectral del tracto vocal, para los diferentes escenarios considerados. Por otro lado, sabemos que el estudio y modelado preciso de los fenómenos transitorios que ocurren en la voz ha despertado gran interés en la comunidad científica. En este trabajo, mostramos que el método propuesto es capaz de describir, para cada instante y de forma adecuada, las transiciones rápidas entre vocales. Por ello, consideramos que los modelos y las estrategias aquí desarrolladas podrán contribuir a la resolución de esta interesante problemática. A su vez, es importante remarcar la dificultad que conlleva el estudio objetivo de esta clase de métodos en señales reales de voz. En nuestros trabajos, analizamos el desempeño del filtrado inverso en espacio de estados en ejemplos reales basándonos en la información disponible en la bibliografía y en criterios subjetivos. Por ello, estamos trabajando actualmente en la búsqueda y el diseño de experimentos que permitan llevar a cabo de modo objetivo este estudio.

Es de nuestro interés continuar con el análisis de las diferentes estrategias aquí propuestas en escenarios más desafiantes, en particular en el estudio de la fonación en condiciones patológicas. Al momento de la redacción del presente documento, contamos con proyectos de investigación en curso y con trabajos en un estadio inicial.

Los resultados preliminares obtenidos son promisorios y sugieren que, en un futuro próximo, podremos avanzar en esta línea. Considerando la relevancia de los resultados logrados y presentados, así como las publicaciones realizadas a nivel nacional e internacional que los avalan, podemos afirmar que se han cumplido los objetivos propuestos al inicio de este trabajo de formación doctoral.

Apéndice A

Obtención de la función objetivo del problema de optimización asociado al MEEG de la fonación

En este apéndice se describe como se obtuvo la expresión para $\mathcal{E}\left\{\ln \mathcal{L}\Big(\boldsymbol{\Theta}\Big)\right\}$, que forma parte de la función objetivo (7.24) a maximizar a lo largo del proceso de estimación de los parámetros del MEEG de la fonación desarrollado en el Cap. 7.

En primer lugar, es importante recordar que para el MEEG de la fonación las observaciones corresponden a la señal de voz $s[n]$, con $n \in \mathcal{I}_N$. Además, de acuerdo a lo desarrollado en el Cap. 7, la varianza del error de observación queda representada por σ_v^2, la excitación externa es $\tilde{\mathrm{u}}_g$ y la matriz de error satisface la igualdad $B[n] = \mathbf{I}_p$. Teniendo en cuenta estas consideraciones, la expresión (5.19) se puede reescribir como sigue:

$$
\begin{aligned}
\mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\} = \tilde{c} &- \frac{1}{2}\ln\left(|\mathbf{P}_0|\right) - \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) \\
&+ \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 + \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 \\
&- \frac{N\,(1+p)}{2}\,\ln(\sigma_v^2) - \frac{N}{2}\,\ln\left(|\mathring{\mathbf{Q}}|\right) \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_N}\Big[s[n]^2 - 2\,s[n]\,\mathbf{H}[n]\,\hat{\mathbf{x}}[n|N] + \mathbf{H}[n]\,\hat{\mathbf{C}}[n|N]\,\mathbf{H}[n]^T\Big] \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_N}\Big[\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}[n|N]\right) - \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}[n-1]\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
&\quad - \hat{\mathbf{x}}[n|N]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}[n-1]\,\tilde{\mathrm{u}}_g[n-1] - \mathrm{Tr}\left(\mathbf{A}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) \\
&\quad + \mathrm{Tr}\left(\mathbf{A}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}[n-1]\,\hat{\mathbf{C}}[n-1|N]\right) \\
&\quad + \hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}[n-1]\,\tilde{\mathrm{u}}_g[n-1] \\
&\quad - \tilde{\mathrm{u}}_g[n-1]^T\,\mathbf{f}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{x}}[n|N] \\
&\quad + \tilde{\mathrm{u}}_g[n-1]^T\,\mathbf{f}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}[n-1]\,\hat{\mathbf{x}}[n-1|N] \\
&\quad + \tilde{\mathrm{u}}_g[n-1]^2\,\mathbf{f}[n-1]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}[n-1]\Big],
\end{aligned}
\tag{A.1}
$$

con $\tilde{c} \in \mathbb{R}$ constante.

Por otro lado, en la Sec. 7.3.2 se explicó que el MEEG de la fonación desarrolla una dinámica diferente para las OP y las CP, lo que implica que las matrices de

transición de estados $\mathbf{A}[n]$ y de excitación $\mathbf{f}[n]$ se modifican de forma correspondiente, ver Ecs. (7.15) y (7.18). Asimismo, por definición de los conjuntos de índices, se satisfacen las siguientes relaciones: $\mathcal{I}_{op} \cup \mathcal{I}_{cp} = \mathcal{I}_N$ y $\mathcal{I}_{op} \cap \mathcal{I}_{cp} = \emptyset$. Finalmente, se toma en consideración que, por su definición en la Ec. (7.4), la excitación $\tilde{\mathrm{u}}_g[n] = 0$ para todo $n \in \mathcal{I}_{cp}$. Aplicando todo esto, se arriba a la expresión buscada:

$$
\begin{aligned}
\mathcal{E}\left\{\ln \mathcal{L}(\boldsymbol{\Theta})\right\} = \tilde{c} &- \frac{1}{2}\,\ln\left(|\mathbf{P}_0|\right) - \frac{1}{2}\,\mathrm{Tr}\left(\mathbf{P}_0^{-1}\,\hat{\mathbf{C}}[0|N]\right) \\
&+ \frac{1}{2}\,\hat{\mathbf{x}}[0|N]^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 + \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\hat{\mathbf{x}}[0|N] - \frac{1}{2}\,\mathbf{x}_0^T\,\mathbf{P}_0^{-1}\,\mathbf{x}_0 \\
&- \frac{N\,(1+p)}{2}\,\ln(\sigma_v^2) - \frac{N}{2}\,\ln\left(|\mathring{\mathbf{Q}}|\right) \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_N}\left[s[n]^2 - 2\,s[n]\,\mathbf{H}[n]\,\hat{\mathbf{x}}[n|N] + \mathbf{H}[n]\,\hat{\mathbf{C}}[n|N]\,\mathbf{H}[n]^T\right] \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}[n|N]\right) - \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
&\qquad - \hat{\mathbf{x}}[n|N]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] - \mathrm{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) \\
&\qquad + \mathrm{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}[n-1|N]\right) \\
&\qquad + \hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] - \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{x}}[n|N] \\
&\qquad + \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{x}}[n-1|N] + \tilde{\mathrm{u}}_g[n-1]^2\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\Big] \\
&- \frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{cp}}\Big[\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}[n|N]\right) - \mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\
&\qquad - \mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) + \mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}[n-1|N]\right)\Big].
\end{aligned}
\tag{A.2}
$$

Esta última expresión fue presentada oportunamente en el Cap. 7, en la Ec. (7.25).

Apéndice B

Desarrollo de las expresiones para calcular los parámetros óptimos del MEEG de la fonación

Este apéndice contiene los desarrollos de las expresiones para el cálculo de los valores óptimos de los parámetros A_g, B_g, C_g y G_g, pertenecientes al MEEG de la fonación desarrollado en el Cap. 7.

Actualización de A_g y B_g

En primer lugar, desarrollaremos las expresiones para obtener los parámetros A_g y B_g. Estudiando el MEEG de la fonación, se desprende que los parámetros A_g y B_g están involucrados únicamente en las definiciones de $\mathbf{A}_{op}$ y $\mathbf{f}_{op}$ válidas para la OP, ver Ec. (7.15). Manipulando estas dos expresiones, se generan las siguientes expresiones:

$$\mathbf{A}_{op} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & A_g \end{pmatrix} = \begin{pmatrix} \mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix} + A_g \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 \end{pmatrix} = \tilde{\mathbf{A}}_{op_1} + A_g\, \tilde{\mathbf{A}}_{op_2}, \tag{B.1}$$

$$\mathbf{f}_{op} = \begin{pmatrix} \mathbf{0} \\ B_g \end{pmatrix} = B_g \begin{pmatrix} \mathbf{0} \\ 1 \end{pmatrix} = B_g\, \tilde{\mathbf{f}}_{op_1}. \tag{B.2}$$

Nótese que $\tilde{\mathbf{A}}_{op_1}$, $\tilde{\mathbf{A}}_{op_2}$ y $\tilde{\mathbf{f}}_{op_1}$ presentan una estructura rala, es decir, poseen pocos elementos no nulos.

Considerando la expresión (7.24), se genera una expresión para la derivada parcial de $\mathcal{E}\{\ln \mathcal{L}_{pen}\}$ con respecto a A_g. Aplicando propiedades, se obtiene lo siguiente:

$$\frac{\partial \mathcal{E}\{\ln \mathcal{L}_{pen}\}}{\partial A_g} = \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial A_g} - \lambda_j \frac{\partial \Phi}{\partial A_g} = \mathrm{Tr}\left(\frac{\partial \mathbf{A}_{op}{}^T}{\partial A_g} \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{A}_{op}} \right) - \lambda_j \frac{\partial \Phi}{\partial A_g}. \tag{B.3}$$

En la expresión anterior se empleó la regla de la cadena para el cálculo de derivadas de funciones de variable matricial. A partir de la definición (7.23), se demuestra que:

$$\frac{\partial \Phi}{\partial A_g} = \left(A_g - \tilde{A}_g \right). \tag{B.4}$$

Para determinar la traza en la Ec. (B.3), se trabaja con cada factor por separado. Partiendo de la Ec. (7.25), se calcula la derivada parcial de la función $\mathcal{E}\{\ln \mathcal{L}\}$ con

respecto a $\mathbf{A}_{op}$. Sólo importan los términos de la expresión que involucran a $\mathbf{A}_{op}$, ya que al derivar se anularán los restantes.

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}\right\}}{\partial \mathbf{A}_{op}} &= \frac{\partial}{\partial \mathbf{A}_{op}}\Big\{ -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[-\operatorname{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\ &\qquad -\operatorname{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right)+\operatorname{Tr}\left(\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}[n-1|N]\right) \\ &\qquad +\hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] \\ &\qquad +\tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{x}}[n-1|N]\Big]\Big\} \\ &= -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[-2\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]+2\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{C}}[n-1|N] \\ &\qquad +2\,\tilde{\mathrm{u}}_g[n-1]\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\hat{\mathbf{x}}[n-1|N]^T\Big].\end{aligned} \tag{B.5}$$

Para arribar al resultado anterior, se aplicaron las reglas para la derivación matricial de funciones de variable matricial [66]. Considerando las expresiones para $\mathbf{A}_{op}$ y para $\mathbf{f}_{op}$, ver Ecs. (B.1) y (B.2), se obtiene el siguiente resultado:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}\right\}}{\partial \mathbf{A}_{op}} &= -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[-2\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N] \\ &\qquad +2\,\mathring{\mathbf{Q}}^{-1}\left(\tilde{\mathbf{A}}_{op_1}+A_g\,\tilde{\mathbf{A}}_{op_2}\right)\hat{\mathbf{C}}[n-1|N] \\ &\qquad +2\,\tilde{\mathrm{u}}_g[n-1]\,\mathring{\mathbf{Q}}^{-1}\,B_g\,\tilde{\mathbf{f}}_{op_1}\,\hat{\mathbf{x}}[n-1|N]^T\Big] \\ &= \frac{1}{\sigma_v^2}\,\mathring{\mathbf{Q}}^{-1}\left(\tilde{\mathbf{C}}_{n,n-1}^{op}-\tilde{\mathbf{A}}_{op_1}\,\tilde{\mathbf{C}}_{n-1}^{op}-A_g\,\tilde{\mathbf{A}}_{op_2}\,\tilde{\mathbf{C}}_{n-1}^{op}-B_g\,\tilde{\mathbf{f}}_{op_1}\,\tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op}\right),\end{aligned} \tag{B.6}$$

donde

$$\begin{aligned}\tilde{\mathbf{C}}_{n,n-1}^{op}&=\sum_{n\in\mathcal{I}_{op}}\hat{\mathbf{C}}_{n,n-1}[n|N], \qquad \tilde{\mathbf{C}}_{n-1}^{op}=\sum_{n\in\mathcal{I}_{op}}\hat{\mathbf{C}}[n-1|N], \\ \tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op}&=\sum_{n\in\mathcal{I}_{op}}\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n-1|N]^T.\end{aligned} \tag{B.7}$$

Por otro lado, tomando la derivada de la Ec. (B.1) con respecto a A_g se obtiene la siguiente expresión:

$$\frac{\partial\,\mathbf{A}_{op}}{\partial\,A_g}=\frac{\partial}{\partial\,A_g}\left\{\tilde{\mathbf{A}}_{op_1}+A_g\,\tilde{\mathbf{A}}_{op_2}\right\}=\tilde{\mathbf{A}}_{op_2}. \tag{B.8}$$

Tomando en cuenta los últimos resultados alcanzados, se desprende que:

$$\begin{aligned}\mathrm{Tr}\left(\frac{\partial\,\mathbf{A}_{op}}{\partial\,A_g}^T\,\frac{\partial\mathcal{E}\left\{\ln\,\mathcal{L}\right\}}{\partial\,\mathbf{A}_{op}}\right) &= \frac{1}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{op_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}_{n,n-1}^{op}\right) - \frac{1}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{op_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{op_1}\,\tilde{\mathbf{C}}_{n-1}^{op}\right)\\ &\quad - \frac{A_g}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{op_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{op_2}\,\tilde{\mathbf{C}}_{n-1}^{op}\right)\\ &\quad - \frac{B_g}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{op_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{f}}_{op_1}\,\tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op}\right)\\ &= \frac{1}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}_{n,n-1}^{op})_{(p,p)} - \frac{1}{\sigma_v^2}\,\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(i,p)}\\ &\quad - \frac{A_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(p,p)} - \frac{B_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op})_{(p)},\end{aligned} \tag{B.9}$$

donde $(\bullet)_{(k,l)}$ indica el elemento de una matriz ubicado en la fila k y la columna l. Para arribar a este último resultado, se trabajó con diferentes propiedades de las matrices con estructura rala [66].

Reemplazando los resultados intermedios en la expresión (B.3), se arriba a la siguiente expresión para la derivada parcial de $\mathcal{E}\left\{\ln\,\mathcal{L}_{pen}\right\}$ con respecto a A_g:

$$\begin{aligned}\frac{\partial\mathcal{E}\left\{\ln\,\mathcal{L}_{pen}\right\}}{\partial\,A_g} &= \frac{1}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}_{n,n-1}^{op})_{(p,p)} - \frac{1}{\sigma_v^2}\,\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(i,p)}\\ &\quad - \frac{A_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(p,p)}\\ &\quad - \frac{B_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op})_{(p)} - \frac{\lambda_j\,\sigma_v^2\,A_g}{\sigma_v^2} + \frac{\lambda_j\,\sigma_v^2\,\tilde{A}_g}{\sigma_v^2}.\end{aligned} \tag{B.10}$$

Igualando a cero la ecuación anterior y reagrupando convenientemente los términos, se obtiene la siguiente expresión:

$$\theta_{11}\,A_g + \theta_{12}\,B_g = \gamma_1, \tag{B.11}$$

donde

$$\begin{aligned}\theta_{11} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(p,p)} + \lambda_j\,\sigma_v^2, \qquad \theta_{12} = (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}^{op})_{(p)}\\ \gamma_1 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}_{n,n-1}^{op})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}_{n-1}^{op})_{(i,p)} + \lambda_j\,\sigma_v^2\,\tilde{A}_g.\end{aligned} \tag{B.12}$$

En el resultado anterior se consideró que $\sigma_v^2 > 0$ y $\mathring{\mathbf{Q}}^{-1}$ son conocidos.

En una segunda etapa, se trabaja con la derivada parcial de $\mathcal{E}\left\{\ln\,\mathcal{L}_{pen}\right\}$ con respecto a B_g. Para ello, se considera nuevamente la expresión (7.24). Aplicando propiedades se obtiene lo siguiente:

$$\frac{\partial\mathcal{E}\left\{\ln\,\mathcal{L}_{pen}\right\}}{\partial\,B_g} = \frac{\partial\mathcal{E}\left\{\ln\,\mathcal{L}\right\}}{\partial\,B_g} - \lambda_j\,\frac{\partial\Phi}{\partial\,B_g} = \frac{\partial\,\mathbf{f}_{op}}{\partial\,B_g}^T\,\frac{\partial\mathcal{E}\left\{\ln\,\mathcal{L}\right\}}{\partial\,\mathbf{f}_{op}} - \lambda_j\,\frac{\partial\Phi}{\partial\,B_g}. \tag{B.13}$$

Para arribar a esta expresión, se utilizó nuevamente la regla de la cadena para el cálculo de derivadas de funciones de variable matricial. A partir de la definición (7.23), se demuestra que:

$$\frac{\partial\Phi}{\partial\,B_g} = \left(B_g - \tilde{B}_g\right). \tag{B.14}$$

A continuación, se trabaja con los factores desconocidos en la Ec. (B.13). Partiendo de (7.25), se calcula la derivada parcial de $\mathcal{E}\{\ln \mathcal{L}\}$ con respecto a $\mathbf{f}_{op}$. Sólo se debe prestar atención a los términos que contengan este parámetro, ya que al derivar se anularán los restantes.

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{f}_{op}} &= \frac{\partial}{\partial \mathbf{f}_{op}}\Big\{ -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[-\hat{\mathbf{x}}[n|N]^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] - \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{x}}[n|N] \\ &\qquad + \tilde{\mathrm{u}}_g[n-1]^2\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op} + \hat{\mathbf{x}}[n-1|N]^T\,\mathbf{A}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1] \\ &\qquad + \tilde{\mathrm{u}}_g[n-1]\,\mathbf{f}_{op}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\hat{\mathbf{x}}[n-1|N]\Big]\Big\} \\ &= -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[-2\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n|N] + 2\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{f}_{op}\,\tilde{\mathrm{u}}_g[n-1]^2 \\ &\qquad + 2\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{op}\,\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n-1|N]\Big].\end{aligned} \tag{B.15}$$

Para arribar al resultado anterior, se emplearon las reglas para la derivación matricial de funciones de variable matricial [66]. Considerando las expresiones para $\mathbf{A}_{op}$ y para $\mathbf{f}_{op}$, ver Ecs. (B.1) y (B.2), se demuestra que:

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{f}_{op}} &= \frac{1}{\sigma_v^2}\sum_{n\in\mathcal{I}_{op}}\Big[\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n|N] - \mathring{\mathbf{Q}}^{-1}\,B_g\,\tilde{\mathbf{f}}_{op_1}\,\tilde{\mathrm{u}}_g[n-1]^2 \\ &\qquad - \mathring{\mathbf{Q}}^{-1}\left(\tilde{\mathbf{A}}_{op_1} + A_g\,\tilde{\mathbf{A}}_{op_2}\right)\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n-1|N]\Big] \\ &= \frac{\mathring{\mathbf{Q}}^{-1}}{\sigma_v^2}\left(\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n} - B_g\,\tilde{\mathbf{f}}_{op_1}\,\tilde{\mathrm{u}}^{op}_{n-1} - \tilde{\mathbf{A}}_{op_1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}} - A_g\,\tilde{\mathbf{A}}_{op_2}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}\right),\end{aligned} \tag{B.16}$$

donde

$$\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n} = \sum_{n\in\mathcal{I}_{op}}\tilde{\mathrm{u}}_g[n-1]\,\hat{\mathbf{x}}[n|N]^T, \qquad \tilde{\mathrm{u}}^{op}_{n-1} = \sum_{n\in\mathcal{I}_{op}}\tilde{\mathrm{u}}_g[n-1]^2. \tag{B.17}$$

La expresión para $\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}}$ fue presentada en la Ec. (B.7).

Derivando la expresión (B.2) con respecto a B_g, se obtiene la siguiente expresión:

$$\frac{\partial\,\mathbf{f}_{op}}{\partial\,B_g} = \frac{\partial}{\partial\,B_g}\Big\{B_g\,\tilde{\mathbf{f}}_{op_1}\Big\} = \tilde{\mathbf{f}}_{op_1}. \tag{B.18}$$

De las expresiones anteriores, se desprende que:

$$\begin{aligned}\frac{\partial\,\mathbf{f}_{op}}{\partial\,B_g}^T\,\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial\,\mathbf{f}_{op}} &= \frac{1}{\sigma_v^2}\,\tilde{\mathbf{f}}_{op_1}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n} - \frac{1}{\sigma_v^2}\,\tilde{\mathbf{f}}_{op_1}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{op_1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}} \\ &\qquad - \frac{B_g}{\sigma_v^2}\,\tilde{\mathbf{f}}_{op_1}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{f}}_{op_1}\,\tilde{\mathrm{u}}^{op}_{n-1} - \frac{A_g}{\sigma_v^2}\,\tilde{\mathbf{f}}_{op_1}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{op_2}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}} \\ &= \frac{1}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n})_{(p)} - \frac{B_g\,\tilde{\mathrm{u}}^{op}_{n-1}}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)} \\ &\qquad - \frac{1}{\sigma_v^2}\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(i)} - \frac{A_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)},\end{aligned} \tag{B.19}$$

donde $(\bullet)_{(k)}$ indica el k-ésimo elemento de un vector y, como se dijo anteriormente, $(\bullet)_{(k,l)}$ indica el elemento de una matriz ubicado en la fila k y a la columna l. Para arribar a este último resultado, se aplicaron diferentes propiedades de las matrices con estructura rala [66].

Reemplazando el resultado anterior y la expresión (B.14) en la Ec. (B.13), se arriba a la siguiente expresión para la derivada de $\mathcal{E}\{\ln \mathcal{L}_{pen}\}$ con respecto a B_g:

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}_{pen}\}}{\partial B_g} = \frac{1}{\sigma_v^2}(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n})_{(p)} - \frac{B_g}{\sigma_v^2}(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,\tilde{\mathrm{u}}^{op}_{n-1} \\ - \frac{1}{\sigma_v^2}\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(i)} \\ - \frac{A_g}{\sigma_v^2}(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)} - \frac{\lambda_j\,\sigma_v^2\,B_g}{\sigma_v^2} + \frac{\lambda_j\,\sigma_v^2\,\tilde{B}_g}{\sigma_v^2}.\end{aligned} \tag{B.20}$$

Igualando a cero la ecuación anterior y reagrupando convenientemente los términos, se obtiene la siguiente expresión:

$$\theta_{21}\,A_g + \theta_{22}\,B_g = \gamma_2, \tag{B.21}$$

donde

$$\begin{aligned}\theta_{21} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(p)} \qquad & \theta_{22} &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,\tilde{\mathrm{u}}^{op}_{n-1} + \lambda_j\,\sigma_v^2 \\ \gamma_2 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{op\,T}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_n})_{(p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{op}_{\tilde{\mathrm{u}}_{n-1}\,\hat{\mathrm{x}}_{n-1}})_{(i)} + \lambda_j\,\sigma_v^2\,\tilde{B}_g\end{aligned} \tag{B.22}$$

Para arribar a este resultado, nuevamente se supuso que $\sigma_v^2 > 0$ y $\mathring{\mathbf{Q}}^{-1}$ son conocidos.

Considerando conjuntamente las Ecs. (B.11) y (B.21), se genera un sistema de ecuaciones algebraicas de dos ecuaciones con dos incógnitas. Resolviendo este sistema, se obtienen las siguientes reglas para actualizar los parámetros A_g y B_g:

$$\begin{aligned}\hat{A}_g &= \frac{\gamma_1\,\theta_{22} - \gamma_2\,\theta_{12}}{\mathring{\mathrm{D}}}, \\ \hat{B}_g &= \frac{\gamma_2\,\theta_{11} - \gamma_1\,\theta_{21}}{\mathring{\mathrm{D}}},\end{aligned} \tag{B.23}$$

donde el determinante del sistema se obtiene como sigue:

$$\mathring{\mathrm{D}} = \theta_{11}\,\theta_{22} - \theta_{12}\,\theta_{21}. \tag{B.24}$$

Actualización de C_g

A continuación, desarrollaremos la expresión para actualizar C_g. Analizando las definiciones involucradas en el MEEG de la fonación, se desprende que este parámetro se involucra únicamente en la definición de $\mathbf{A}_{cp}$ válida durante la CP, ver Ec. (7.18). Manipulando su definición, se obtiene la siguiente expresión:

$$\mathbf{A}_{cp} = \begin{pmatrix}\mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & C_g\end{pmatrix} = \begin{pmatrix}\mathbf{I}_\rho & \mathbf{0} \\ \mathbf{0} & 0\end{pmatrix} + C_g\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1\end{pmatrix} = \tilde{\mathbf{A}}_{cp_1} + C_g\,\tilde{\mathbf{A}}_{cp_2}. \tag{B.25}$$

Considerando la definición (7.24), se genera una expresión para la derivada parcial de $\mathcal{E}\{\ln \mathcal{L}_{pen}\}$ con respecto a C_g. Aplicando propiedades, se obtiene lo siguiente:

$$\frac{\partial \mathcal{E}\{\ln \mathcal{L}_{pen}\}}{\partial C_g} = \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial C_g} - \lambda_j \frac{\partial \Phi}{\partial C_g} = \mathrm{Tr}\left(\frac{\partial \mathbf{A}_{cp}}{\partial C_g}^T \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{A}_{cp}}\right) - \lambda_j \frac{\partial \Phi}{\partial C_g}. \quad \text{(B.26)}$$

De la definición (7.23) encontramos que:

$$\frac{\partial \Phi}{\partial C_g} = \left(C_g - \tilde{C}_g\right). \quad \text{(B.27)}$$

Se obtendrá ahora una expresión para la traza en la Ec. (B.26). Para ello, se trabaja con cada factor por separado. En primer lugar, se calcula la derivada parcial de $\mathcal{E}\{\ln \mathcal{L}\}$ con respecto a $\mathbf{A}_{cp}$ partiendo de la Ec. (7.25). Sólo deben considerarse los términos que involucren a $\mathbf{A}_{cp}$, ya que al derivar se anulan los restantes.

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{A}_{cp}} &= \frac{\partial}{\partial \mathbf{A}_{cp}}\Big\{ -\frac{1}{2\,\sigma_v^2} \sum_{n\in\mathcal{I}_{cp}} \Big[-\mathrm{Tr}\left(\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}_{n-1,n}[n|N]\right) \\ &\qquad -\mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N]\right) + \mathrm{Tr}\left(\mathbf{A}_{cp}^T\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}[n-1|N]\right)\Big]\Big\} \\ &= -\frac{1}{2\,\sigma_v^2} \sum_{n\in\mathcal{I}_{cp}} \Big[-2\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N] + 2\,\mathring{\mathbf{Q}}^{-1}\,\mathbf{A}_{cp}\,\hat{\mathbf{C}}[n-1|N]\Big].\end{aligned} \quad \text{(B.28)}$$

Para arribar al resultado anterior, se emplearon las reglas para la derivación matricial de funciones de variable matricial [66]. Considerando la expresión para $\mathbf{A}_{cp}$, ver Ec. (B.25), se obtiene el siguiente resultado:

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{A}_{cp}} &= -\frac{1}{2\,\sigma_v^2} \sum_{n\in\mathcal{I}_{cp}} \Big[-2\,\mathring{\mathbf{Q}}^{-1}\,\hat{\mathbf{C}}_{n,n-1}[n|N] + 2\,\mathring{\mathbf{Q}}^{-1}\,(\tilde{\mathbf{A}}_{cp_1} + C_g\,\tilde{\mathbf{A}}_{cp_2})\,\hat{\mathbf{C}}[n-1|N]\Big] \\ &= \frac{1}{\sigma_v^2}\,\mathring{\mathbf{Q}}^{-1}\left(\tilde{\mathbf{C}}^{cp}_{n,n-1} - \tilde{\mathbf{A}}_{cp_1}\,\tilde{\mathbf{C}}^{cp}_{n-1} - C_g\,\tilde{\mathbf{A}}_{cp_2}\,\tilde{\mathbf{C}}^{cp}_{n-1}\right),\end{aligned} \quad \text{(B.29)}$$

donde

$$\tilde{\mathbf{C}}^{cp}_{n,n-1} = \sum_{n\in\mathcal{I}_{cp}} \hat{\mathbf{C}}_{n,n-1}[n|N], \qquad \tilde{\mathbf{C}}^{cp}_{n-1} = \sum_{n\in\mathcal{I}_{cp}} \hat{\mathbf{C}}[n-1|N]. \quad \text{(B.30)}$$

Tomando la derivada de la Ec. (B.25) con respecto a C_g, se obtiene lo siguiente:

$$\frac{\partial \mathbf{A}_{cp}}{\partial C_g} = \frac{\partial}{\partial C_g}\left\{\tilde{\mathbf{A}}_{cp_1} + C_g\,\tilde{\mathbf{A}}_{cp_2}\right\} = \tilde{\mathbf{A}}_{cp_2}. \quad \text{(B.31)}$$

Combinando las expresiones (B.29) y (B.31), se arriba al siguiente resultado:

$$\begin{aligned}\mathrm{Tr}\left(\frac{\partial \mathbf{A}_{cp}}{\partial C_g}^T \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{A}_{cp}}\right) &= \frac{1}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{cp_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{cp}_{n,n-1}\right) - \frac{C_g}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{cp_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{cp_2}\,\tilde{\mathbf{C}}^{cp}_{n-1}\right) \\ &\qquad - \frac{1}{\sigma_v^2}\,\mathrm{Tr}\left(\tilde{\mathbf{A}}_{cp_2}^T\,\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{A}}_{cp_1}\,\tilde{\mathbf{C}}^{cp}_{n-1}\right) \\ &= \frac{1}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{cp}_{n,n-1})_{(p,p)} - \frac{C_g}{\sigma_v^2}\,(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(p,p)} \\ &\qquad - \frac{1}{\sigma_v^2}\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(i,p)},\end{aligned} \quad \text{(B.32)}$$

donde nuevamente $(\bullet)_{(k,l)}$ indica el elemento de una matriz correspondiente a la fila k y a la columna l. Para arribar a este último resultado se hizo uso de diferentes propiedades de las matrices con estructura rala [66].

Reemplazando el resultado anterior en la Ec. (B.26), se obtiene la expresión buscada para la derivada de $\mathcal{E}\{\ln \mathcal{L}_{pen}\}$ con respecto a C_g:

$$\begin{aligned}\frac{\partial \mathcal{E}\{\ln \mathcal{L}_{pen}\}}{\partial C_g} = \frac{1}{\sigma_v^2}(\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{cp}_{n,n-1})_{(p,p)} - \frac{C_g}{\sigma_v^2}(\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(p,p)} \\ - \frac{1}{\sigma_v^2}\sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(i,p)} - \frac{\lambda_j\,\sigma_v^2\,C_g}{\sigma_v^2} + \frac{\lambda_j\,\sigma_v^2\,\tilde{C}_g}{\sigma_v^2}.\end{aligned} \tag{B.33}$$

Igualando a cero la ecuación anterior y reagrupando convenientemente los términos, se genera el siguiente resultado:

$$\hat{C}_g = \frac{\gamma_3}{\theta_3}, \tag{B.34}$$

donde

$$\begin{aligned}\theta_3 &= (\mathring{\mathbf{Q}}^{-1})_{(p,p)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(p,p)} + \lambda_j\,\sigma_v^2 \\ \gamma_3 &= (\mathring{\mathbf{Q}}^{-1}\,\tilde{\mathbf{C}}^{cp}_{n,n-1})_{(p,p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{Q}}^{-1})_{(p,i)}\,(\tilde{\mathbf{C}}^{cp}_{n-1})_{(i,p)} + \lambda_j\,\sigma_v^2\,\tilde{C}_g.\end{aligned} \tag{B.35}$$

Al igual que en los desarrollos anteriores, en la expresión anterior se considera que $\sigma_v^2 > 0$ y $\mathring{\mathbf{Q}}^{-1}$ son conocidos.

Actualización de G_g

Por último, generaremos la expresión para actualizar G_g. De la Sec. 7.3.2, se desprende que este parámetro se involucra únicamente en la definición de $\mathbf{H}[n]$, ver la Ec. (7.20). Partiendo de la definición de $\mathbf{H}[n]$, se arriba la siguiente expresión:

$$\begin{aligned}\mathbf{H}[n] &= \begin{pmatrix}-s[n-1] & -s[n-2] & \cdots & -s[n-\rho] & G_g\end{pmatrix} \\ &= \begin{pmatrix}-s[n-1] & -s[n-2] & \cdots & -s[n-\rho] & 0\end{pmatrix} + G_g\begin{pmatrix}\mathbf{0} & 1\end{pmatrix} \\ &= \mathring{\mathbf{H}}_1[n] + G_g\,\mathring{\mathbf{H}}_2.\end{aligned} \tag{B.36}$$

Considerando la Ec. (7.24), se arriba a la siguiente expresión para la derivada parcial de $\mathcal{E}\{\ln \mathcal{L}_{pen}\}$ con respecto a G_g:

$$\frac{\partial \mathcal{E}\{\ln \mathcal{L}_{pen}\}}{\partial G_g} = \frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial G_g} = \sum_{l\in\mathcal{I}_N}\frac{\partial \mathcal{E}\{\ln \mathcal{L}\}}{\partial \mathbf{H}[l]}\,\frac{\partial\,\mathbf{H}[l]^T}{\partial G_g}. \tag{B.37}$$

Para ello, se utilizó la regla de la cadena para el cálculo de la derivada de funciones de variable vectorial.

A continuación, se estudiarán por separado cada uno de los factores involucrados en la sumatoria de la Ec. (B.37). Partiendo de la expresión (7.25), se calcula la derivada parcial de la función $\mathcal{E}\{\ln \mathcal{L}\}$ con respecto a $\mathbf{H}[l]$ con $l \in \mathcal{I}_N$. Sólo se

presta atención a los términos de la expresión (7.25) que involucran a $\mathbf{H}[l]$, ya que al derivar se anularán los restantes. Haciendo esto, se obtiene el siguiente resultado:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}\right\}}{\partial \mathbf{H}[l]} &= \frac{\partial}{\partial \mathbf{H}[l]}\left\{ -\frac{1}{2\,\sigma_v^2}\sum_{n\in\mathcal{I}_N}\Big[-2\,s[n]\,\mathbf{H}[n]\,\hat{\mathbf{x}}[n|N] + \mathbf{H}[n]\,\hat{\mathbf{C}}[n|N]\,\mathbf{H}[n]^T\Big]\right\} \\ &= -\frac{1}{2\,\sigma_v^2}\Big[-2\,s[l]\,\hat{\mathbf{x}}[l|N]^T + 2\,\mathbf{H}[l]\,\hat{\mathbf{C}}[l|N]\Big].\end{aligned} \tag{B.38}$$

En la expresión anterior se utilizaron las reglas para la derivación matricial de funciones de variable matricial [66]. Considerando la expresión para $\mathbf{H}[n]$, ver Ec. (B.36), se demuestra que:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}\right\}}{\partial \mathbf{H}[l]} &= -\frac{1}{2\,\sigma_v^2}\Big[-2\,s[l]\,\hat{\mathbf{x}}[l|N]^T + 2\left(\mathring{\mathbf{H}}_1[l] + G_g\,\mathring{\mathbf{H}}_2\right)\hat{\mathbf{C}}[l|N]\Big] \\ &= \frac{1}{\sigma_v^2}\Big[s[l]\,\hat{\mathbf{x}}[l|N]^T - \mathring{\mathbf{H}}_1[l]\,\hat{\mathbf{C}}[l|N] - G_g\,\mathring{\mathbf{H}}_2\,\hat{\mathbf{C}}[l|N]\Big].\end{aligned} \tag{B.39}$$

Calculando la derivada de la Ec. (B.36) con respecto a G_g, se obtiene lo siguiente:

$$\frac{\partial\,\mathbf{H}[l]}{\partial\,G_g} = \frac{\partial}{\partial\,G_g}\left\{\mathring{\mathbf{H}}_1[l] + G_g\,\mathring{\mathbf{H}}_2\right\} \;= \mathring{\mathbf{H}}_2. \tag{B.40}$$

Reemplazando estos últimos resultados en la Ec. (B.37), se arriba a la siguiente expresión para la derivada parcial de $\mathcal{E}\left\{\ln \mathcal{L}_{pen}\right\}$ con respecto a G_g:

$$\begin{aligned}\frac{\partial \mathcal{E}\left\{\ln \mathcal{L}_{pen}\right\}}{\partial\,G_g} &= \sum_{l\in\mathcal{I}_N}\frac{1}{\sigma_v^2}\Big[s[l]\,\hat{\mathbf{x}}[l|N]^T - \mathring{\mathbf{H}}_1[l]\,\hat{\mathbf{C}}[l|N] - G_g\,\mathring{\mathbf{H}}_2\,\hat{\mathbf{C}}[l|N]\Big]\mathring{\mathbf{H}}_2^T \\ &= \frac{1}{\sigma_v^2}\sum_{l\in\mathcal{I}_N}\Big[s[l]\,\hat{\mathbf{x}}[l|N]^T\,\mathring{\mathbf{H}}_2^T - \mathring{\mathbf{H}}_1[l]\,\hat{\mathbf{C}}[l|N]\,\mathring{\mathbf{H}}_2^T - G_g\,\mathring{\mathbf{H}}_2\,\hat{\mathbf{C}}[l|N]\,\mathring{\mathbf{H}}_2^T\Big] \\ &= \frac{1}{\sigma_v^2}\sum_{l\in\mathcal{I}_N}\Big[s[l]\,(\hat{\mathbf{x}}[l|N])_{(p)} - \sum_{i=1}^{\rho}(\mathring{\mathbf{H}}_1[l])_{(i)}\,(\hat{\mathbf{C}}[l|N])_{(i,p)} - G_g\,(\hat{\mathbf{C}}[l|N])_{(p,p)}\Big] \\ &= \frac{1}{\sigma_v^2}\sum_{l\in\mathcal{I}_N}\Big[s[l]\,(\hat{\mathbf{x}}[l|N])_{(p)} + \sum_{i=1}^{\rho}s[l-i]\,(\hat{\mathbf{C}}[l|N])_{(i,p)} - G_g\,(\hat{\mathbf{C}}[l|N])_{(p,p)}\Big],\end{aligned} \tag{B.41}$$

donde $(\bullet)_{(k)}$ indica el k-ésimo elemento de un vector y $(\bullet)_{(k,l)}$ es el elemento de una matriz ubicado en la fila k y la columna l. En esta última expresión, se emplearon las definiciones de los vectores de la Ec. (B.36) y su relación con la señal de voz s.

Igualando a cero la expresión anterior y agrupando convenientemente sus términos, se obtiene la regla para actualizar el parámetro de ganancia G_g:

$$\hat{G}_g = \frac{\mu}{\xi}, \tag{B.42}$$

donde

$$\begin{aligned}\mu &= \sum_{n\in\mathcal{I}_N}\Big[s[n]\,(\hat{\mathbf{x}}[n|N])_{(p)} + \sum_{i=1}^{\rho}s[n-i]\,(\hat{\mathbf{C}}[n|N])_{(i,p)}\Big] \\ \xi &= \sum_{n\in\mathcal{I}_N}(\hat{\mathbf{C}}[n|N])_{(p,p)}.\end{aligned} \tag{B.43}$$

Note el lector que se cambió el índice de la sumatoria por conveniencia.

Lista de publicaciones

[1] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, y M. E. Torres. Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos. En *XVIII Congreso Argentino de Bioingeniería y VII Jornadas de Ingeniería Clínica (SABI 2011)*, 2011.

[2] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, y M. E. Torres. Evaluación de un nuevo modelo de síntesis de vocales con perturbaciones en los parámetros acústicos. En *XIV Reunión de Procesamiento de la Información y Control (RPIC 2011)*, pp. 306–311, 2011.

[3] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, y M. E. Torres. Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters. *Latin American Applied Research*, 43(3):225–230, 2013.

[4] G. A. Alzamendi, G. Schlotthauer, y M. E. Torres. Modelado estocástico de la fuente glótica aplicado a la descomposición de fonemas vocálicos mediante métodos en espacio de estados. En *XVI Reunión de Procesamiento de la Información y Control (RPIC 2015)*.

[5] G. A. Alzamendi, G. Schlotthauer, y M. E. Torres. Análisis de secuencias de períodos de la voz mediante modelos en espacio de estados. En *XV Reunión de Procesamiento de la Información y Control (RPIC 2013)*, pp. 421–426, 2013.

[6] G. A. Alzamendi, G. Schlotthauer, y M. E. Torres. A new method for structural analysis of perturbed pitch period series. En *VI Latin American Conference on Biomedical Engineering (CLAIB 2014)*, 2014.

[7] G. A. Alzamendi, G. Schlotthauer, y M. E. Torres. Formulation of a stochastic glottal source model inspired on deterministic liljencrants-fant model. En C. Manfredi, editor, *Models and Analysis of Vocal Emissions for Biomedical Applications: 9th International Workshop*, pp. 15 – 18. Firenze University Press, 2015.

[8] G. A. Alzamendi, G. Schlotthauer, y M. E. Torres. State-space approach to structural representation of perturbed pitch period sequences in voice signals. *Journal of Voice*, 29(6):682 – 692, 2015.

[9] 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*, 2016. **En evaluación**.

Referencias

[1] M. Airaksinen, T. Raitio, B. Story, and P. Alku. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(3):596–607, 2014.

[2] M. Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008.

[3] O. O. Akande and P. J. Murphy. Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, 46(1):15 – 36, 2005.

[4] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2):109 – 118, 1992.

[5] P. Alku. Glottal inverse filtering analysis of human voice production - A review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011.

[6] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, 134(2):1295–1313, 2013.

[7] P. Alku, B. Story, and M. Airas. Estimation of the voice source from speech pressure signals: evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatrica et Logopaedica*, 58(2):102–113, 2006.

[8] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Courier Corporation, 2005.

[9] N. Aoki and T. Ifukube. Analysis and perception of spectral $1/f$ characteristics of amplitude and period fluctuations in normal sustained vowels. *The Journal of the Acoustical Society of America*, 106(1):423–433, 1999.

[10] J. Arias-Londoño, J. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez. Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients. *Biomedical Engineering, IEEE Transactions on*, 58(2):370–379, 2011.

[11] M. Arnela and O. Guasch. Two-dimensional vocal tracts with three-dimensional behavior in the numerical generation of vowels. *The Journal of the Acoustical Society of America*, 135(1):369–379, 2014.

[12] M. Arnela, O. Guasch, and F. Alías. Effects of head geometry simplifications on acoustic radiation of vowel sounds based on time-domain finite-element simulations. *The Journal of the Acoustical Society of America*, 134(4):2946–2954, 2013.

[13] L. Aronson, H. L. Rufiner, H. Furmanski, and P. Estienne. Características acústicas de las vocales del español rioplatense. *Fonoaudiológica*, 46(2):12–20, 2000.

[14] A. Askenfelt, J. Gauffin, and J. Sundberg. A comparison of contact microphone and electroglottograph for the measurement of vocal fundamental frequency. *Journal of Speech, Language, and Hearing Research*, 23(2):258–273, 1980.

[15] R. J. Baken. Electroglottography. *Journal of Voice*, 6(2):98 – 110, 1992.

[16] R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Singular Thomson Learning, San Diego, 2000.

[17] M. A. Berezina, D. Rudoy, and P. J. Wolfe. Autoregressive modeling of voiced speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5042–5045, 2010.

[18] D. A. Berry and I. R. Titze. Normal modes in a continuum model of vocal fold tissues. *The Journal of the Acoustical Society of America*, 100(5), 1996.

[19] S. Bielamowicz, J. Kreiman, B. R. Gerratt, M. S. Dauer, and G. S. Berke. Comparison of voice analysis systems for perturbation measurement. *Journal of Speech, Language, and Hearing Research*, 39(1):126–134, 1996.

[20] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. V. Hirtum, and X. Laval. Effects of higher order propagation modes in vocal tract like geometries. *The Journal of the Acoustical Society of America*, 137(2):832–843, 2015.

[21] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the institute of phonetic sciences*, 17(1193):97–110, 1993.

[22] P. Boersma. Should jitter be measured by peak picking or by waveform matching? *Folia Phoniatrica et Logopaedica*, 61(5):305–308, 2009.

[23] P. Boersma and D. Weenink. Praat: doing phonetics by computer. Computer program Version 5.3.51, 2013. http://www.praat.org/, retrieved 2 June.

[24] H. S. Bonilha and D. D. Deliyski. Period and glottal width irregularities in vocally normal speakers. *Journal of Voice*, 22(6):699–708, 2008.

[25] A. M. Borzone de Manrique. *Manual de fonética acústica*. Librería Hachette, 1980.

[26] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Wiley, third edition edition, June 1994.

[27] A. Breen. Speech synthesis models: a review. *Electronics & Communication Engineering Journal*, 4:19–31, February 1992.

[28] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding. Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of Voice*, 25(1):44 – 53, 2011.

[29] J. P. Cabral and L. C. Oliveira. Emovoice: a system to generate emotions in speech. In *INTERSPEECH 2006*, 2006.

[30] J. Candy. *Model-Based Signal Processing*. John Wiley & Sons, New Jersey, USA, 2005.

[31] R. Carlson. Models of speech synthesis. *Proceedings of the National Academy of Sciences*, 92(22):9932–9937, 1995.

[32] Z. Chen. Bayesian filtering: From Kalman filters to particle filters, and beyond. Technical report, Communications Research Laboratory, McMaster University, Ontario, Canada, 2003.

[33] X. Chi and M. Sonderegger. Subglottal coupling and its influence on vowel formants. *The Journal of the Acoustical Society of America*, 122(3):1735–1745, 2007.

[34] R. H. Colton and E. G. Conture. Problems and pitfalls of electroglottography. *Journal of Voice*, 4(1):10 – 24, 1990.

[35] J. J. Commandeur and S. J. Koopman. *An Introduction to State Space Time Series Analysis*. Oxford University Press, Oxford, 2007.

[36] S. Dabbaghchian, M. Arnela, and O. Engwall. Simplification of vocal tract shapes with different lelvels of detail. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow, The Scottish Consortium for ICPhS 2015 (Ed.), 2015.

[37] F. Daum. Nonlinear filters: beyond the Kalman filter. *Aerospace and Electronic Systems Magazine, IEEE*, 20(8):57–69, 2005.

[38] B. de Boer and W. Tecumseh Fitch. Computer models of vocal tract evolution: An overview and critique. *Adaptive Behavior*, 18(1):36–47, 2010.

[39] P. Dejonckere, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, and 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*, 7(1):37 – 42, 2012. Human Voice and Sounds: From Newborn to Elder.

[40] P. DeJonckere, J. Schoentgen, A. Giordano, S. Fraj, L. Bocchi, and C. Manfredi. Validity of jitter measures in non-quasi-periodic voices. Part I: perceptual and computer performances in cycle pattern recognition. *Logopedics Phoniatrics Vocology*, 36(2):70–77, 2011.

[41] P. H. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, and V. Woisard. A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. *European Archives of Oto-Rhino-Laryngology*, 258(2):77–82, 2001.

[42] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, New York, 1993.

[43] V. Digalakis, J. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 1(4):431–442, 1993.

[44] B. Doval, C. d'Alessandro, and N. Henrich. The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92(6):1026–1046, 2006.

[45] C. Drioli and A. Calanca. Speaker adaptive voice source modeling with applications to speech coding and processing. *Computer Speech & Language*, 28(5):1195–1208, 2014.

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

[47] T. Drugman, B. Bozkurt, and T. Dutoit. A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20 – 34, 2012.

[48] T. Drugman and T. Dutoit. Glottal closure and opening instant detection from speech signals. In *Interspeech*, pages 2891–2894, 2009.

[49] H. Dudley and T. H. Tarnoczy. The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, 22(2), 1950.

[50] S. M. Dunn, A. Constantinides, and P. V. Moghe. *Numerical Methods in Biomedical Engineering*. Academic Pr Inc, 2005.

[51] J. Durbin and S. Koopman. *Time Series Analysis by State Space Methods*. Oxford Univ Pr (Sd), New York, USA, 1 edition, 2001.

[52] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *Signal Processing, IEEE Transactions on*, 39(2):411–423, 1991.

[53] Y. Endo and H. Kasuya. A stochastic model of fundamental period perturbation and its application to perception of pathological voice quality. In *Proc. of Fourth Int. Conference on Spoken Language ICSLP, 1996*, volume 2, pages 772–775, Oct 1996.

[54] G. Fant. *Acoustic theory of speech production*. Number 2. Walter de Gruyter, 1970.

[55] G. Fant. Some problems in voice source analysis. *Speech Communication*, 13(1):7 – 22, 1993.

[56] G. Fant. Phonetics and phonology in the last 50 years. *From Sound to Sense, no. B, MIT*, pages 20–41, 2004.

[57] G. Fant, J. Liljencrants, and Q. G. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985.

[58] M. Farrus and J. Hernando. Using jitter and shimmer in speaker verification. *IET Signal Processing*, 3(4):247–257, 2009.

[59] M. Farrús, J. Hernando, and P. Ejarque. Jitter and shimmer measurements for speaker recognition. In *INTERSPEECH*, pages 778–781, 2007.

[60] R. Fraile, M. Kob, J. I. Godino-Llorente, N. Sáenz-Lechón, V. J. Osma-Ruiz, and J. M. Gutiérrez-Arriola. Physical simulation of laryngeal disorders using a multiple-mass vocal fold model. *Biomedical Signal Processing and Control*, 7(1):65 – 78, 2012. Human Voice and Sounds: From Newborn to Elder.

[61] S. Fraj, F. Grenez, and J. Schoentgen. Synthetic hoarse voices: a perceptual evaluation. In *Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications-MAVEBA*, pages 95–98, 2009.

[62] S. Fraj, J. Schoentgen, and F. Grenez. Development and perceptual assessment of a synthesizer of disordered voices. *The Journal of the Acoustical Society of America*, 132(4):2603–2615, 2012.

[63] Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):492–501, 2006.

[64] M. A. García Jurado and M. Arenas. *La fónetica del español: Análisis e investigación de los sonidos del habla*. Quorum, 2005.

[65] P. Garner, M. Cernak, and P. Motlicek. A simple continuous pitch estimation algorithm. *Signal Processing Letters, IEEE*, 20(1):102–105, 2013.

[66] J. E. Gentle. *Matrix Algebra*. Springer Texts in Statistics. Springer New York, New York, NY, 2007.

[67] P. K. Ghosh and S. S. Narayanan. Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter. *Speech Communication*, 53(1):98 – 109, 2011.

[68] J. D. Gibson. Speech coding methods, standards, and applications. *Circuits and Systems Magazine, IEEE*, 5(4):30–49.

[69] S. J. Godsill, P. J. Wolfe, and W. N. Fong. Statistical model-based approaches to audio restoration and analysis. *Journal of New Music Research*, 30(4):323–338, 2001.

[70] P. Gómez Vilda. *Fonolaringología*, chapter Pasado, presente y futuro del Análisis Acústico de la Calidad de la Voz., pages 217–235. Amplifón Ibérica, 2014.

[71] D. Govind and S. R. Mahadeva Prasanna. Expressive speech synthesis: a review. *International Journal of Speech Technology*, 16(2):237–260, 2013.

[72] H. E. Gunter. A mechanical model of vocal-fold collision with high spatial and temporal resolution. *The Journal of the Acoustical Society of America*, 113(2), 2003.

[73] M. G. Hall, A. V. Oppenheim, and A. S. Willsky. Time-varying parametric modeling of speech. *Signal Processing*, 5(3):267 – 285, 1983.

[74] V. Hampala, M. Garcia, J. G. Švec, R. C. Scherer, and C. T. Herbst. Relationship Between the Electroglottographic Signal and Vocal Fold Contact Area. *Journal of Voice*. Artículo en prensa.

[75] A. C. Harvey and S. J. Koopman. Diagnostic checking of unobserved-components time series models. *Journal of Business & Economic Statistics*, 10(4):377–389, 1992.

[76] A. C. Harvey and N. Shephard. Structural time series models. In G. S. Maddala, C. R. Rao, and H. D. Vinod, editors, *Econometrics*, volume 11 of *Handbook of Statistics*, chapter 10, pages 261–302. Elsevier, 1993.

[77] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, 115(3):1321–1332, 2004.

[78] P. Henriquez, J. Alonso, M. Ferrer, C. Travieso, J. Godino-Llorente, and F. Diaz-de Maria. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1186–1195, 2009.

[79] D. J. Hermes. Synthesis of breathy vowels: Some research methods. *Speech Communication*, 10(5–6):497 – 502, 1991. Speaker Characterization in Speech Technology.

[80] E. E. Holmes. Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models. *ArXiv e-prints*, 2013.

[81] Y. Hu and P. Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238, 2008.

[82] M. C. A. Jackson-Menaldi. *La voz patológica*. Ed. Médica Panamericana, Jan. 2002.

[83] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Courier Corporation, 2007.

[84] P. Jinachitra and J. O. Smith. Joint estimation of glottal source and vocal tract for vocal synthesis using kalman smoothing and em algorithm. In *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, pages 327–330, 2005.

[85] P. Jinachitra and J. O. Smith. Generative model of voice in noise for structured coding applications. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–281–I–284, 2007.

[86] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[87] S. Koopman. Exact initial Kalman filtering and smoothing for nonstationary time series models. *Journal of the American Statistical Association*, 92(440):1630–1638, 1997.

[88] S. Koopman and J. Durbin. Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis*, 24(1):85–98, 2003.

[89] S. J. Koopman and M. Ooms. Forecasting economic time series using unobserved components time series models. In M. P. Clements and D. F. Hendry, editors, *The Oxford Handbook of Economic Forecasting*, chapter 5, pages 129–162. Oxford University Press, 2011.

[90] S. J. Koopman and N. Shephard. Exact score for time series models in state space form. *Biometrika*, 79(4):823–826, 1992.

[91] M. H. Krane. Aeroacoustic production of low-frequency unvoiced speech sounds. *The Journal of the Acoustical Society of America*, 118(1), 2005.

[92] J. Kreiman and B. R. Gerratt. Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4):2201–2211, 2005.

[93] P. H. Kvam and B. Vidakovic. *Nonparametric Statistics with Applications to Science and Engineering*. John Wiley & Sons, New Jersey, USA, 2007.

[94] D. Labarre, E. Grivel, Y. Berthoumieu, E. Todini, and M. Najim. Consistent estimation of autoregressive parameters from noisy observations based on two interacting Kalman filters. *Signal Processing*, 86(10):2863 – 2876, 2006. Special Section: Fractional Calculus Applications in Signals and Systems.

[95] R. Laje, T. Gardner, and G. B. Mindlin. Continuous model for vocal fold oscillations to study the effect of feedback. *Physical Review E*, 64:056201, Oct 2001.

[96] R. F. Leonarduzzi, G. A. Alzamendi, G. Schlotthauer, and M. E. Torres. Análisis multifractal de las secuencias de períodos y amplitudes de la voz: resultados preliminares. In *XIX Congreso Argentino de Bioingeniería y VIII Jornadas de Ingeniería Clínica (SABI 2013)*, 2013.

[97] R. F. Leonarduzzi, G. A. Alzamendi, G. Schlotthauer, and M. E. Torres. Wavelet leader multifractal analysis of period and amplitude sequences from sustained vowels. *Speech Communication*, 72:1 – 12, 2015.

[98] H. Li, R. Scaife, and D. O'Brien. LF model based glottal source parameter estimation by extended Kalman filtering. In *Proc. of the 22nd IET Irish Signals and Systems Conference*, 2011.

[99] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, 2 edition edition, Jan. 1999.

[100] L. Ljung. Prediction error estimation methods. *Circuits, Systems and Signal Processing*, 21(1):11–21, 2002.

[101] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 1 edition edition, June 2007.

[102] C. Ma, Y. Kamp, and L. Willems. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, 12(1):69 – 81, 1993.

[103] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku. Stabilised weighted linear prediction. *Speech Communication*, 51(5):401 – 411, 2009.

[104] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.

[105] J. Makhoul. Spectral linear prediction: Properties and applications. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(3):283–296, Jun 1975.

[106] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3 edition, Jan. 2009.

[107] C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, and P. Dejonckere. Validity of jitter measures in non-quasi-periodic voices. Part II: the effect of noise. *Logopedics Phoniatrics Vocology*, 36(2):78–89, 2011.

[108] D. G. Manolakis, V. K. Ingle, and S. M. Kogon. *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. Artech House Print on Demand, Apr. 2005.

[109] M. Markaki and Y. Stylianou. Voice pathology detection and discrimination based on modulation spectral features. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):1938–1948, 2011.

[110] Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab. Disordered voice database. Model 4337, 2009. http://www.kayelemetrics.com.

[111] D. D. Mehta, D. Rudoy, and P. J. Wolfe. Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *The Journal of the Acoustical Society of America*, 132:1732–1746, 2012.

[112] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, and R. E. Hillman. Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology*, 3(155), 2015.

[113] J. Mekyska, E. Janousova, P. Gomez-Vilda, Z. Smekal, I. Rektorova, I. Eliasova, M. Kostalova, M. Mrackova, J. B. Alonso-Hernandez, M. Faundez-Zanuy, and K. López-de Ipiña. Robust and complex approach of pathological speech signal analysis. *Neurocomputing*, 167:94 – 111, 2015.

[114] Z. Michalewicz and D. B. Fogel. *How to Solve It: Modern Heuristics*. Springer, 2 edition, 2004.

[115] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers*. John Wiley & Sons, 2010.

[116] P. Murphy. Source-filter comparison of measurements of fundamental frequency perturbation and amplitude perturbation for synthesized voice signals. *Journal of Voice*, 22(2):125 – 137, 2008.

[117] P. J. Murphy. Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *The Journal of the Acoustical Society of America*, 107(2):978–988, 2000.

[118] J. Muñoz, E. Mendoza, G. Carballo, M. Fresneda, and A. Cruz. Características acústicas de la voz normal en varones y mujeres mediante el MDVP (Multi-Dimensional Voice Program). *Revista de Logopedia, Foniatría y Audiología*, 21(3):138 – 144, 2001.

[119] J. Muñoz, E. Mendoza, M. Fresneda, G. Carballo, and P. López. Acoustic and perceptual indicators of normal and pathological voice. *Folia Phoniatrica et Logopaedica*, 55(2):102–114, 2003.

[120] S. Narayanan and A. Alwan. Noise source models for fricative consonants. *Speech and Audio Processing, IEEE Transactions on*, 8(3):328–344, 2000.

[121] K. Neumann, V. Gall, H. K. Schutte, and D. G. Miller. A new method to record subglottal pressure waves: Potential applications. *Journal of Voice*, 17(2):140 – 159, 2003.

[122] E. Obediente. *Fonética y fonología*. Universidad de Los Andes, 1998.

[123] I.-T. P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[124] V. Parsa and D. G. Jamieson. Identification of pathological voices using glottal noise measures. *Journal of Speech, Language, and Hearing Research*, 43(2):469–485, 2000.

[125] N. B. Pinto and I. R. Titze. Unification of perturbation measures in speech signals. *The Journal of the Acoustical Society of America*, 87(3):1278–1289, 1990.

[126] B. Pompino-Marschall. Von kempelen et al.-Remarks on the history of articulatory-acoustic modelling. 40:145–159, 2005.

[127] A. Poulimenos and S. Fassois. Parametric time-domain methods for non-stationary random vibration modelling and analysis - A critical survey and comparison. *Mechanical Systems and Signal Processing*, 20(4):763 – 816, 2006.

[128] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, Upper Saddle River, NJ, 1 edition edition, Nov. 2001.

[129] A. Quilis. *Tratado De Fonología Y Fonética Españolas*. Gredos, edición: 2 edition, 2008.

[130] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[131] D. Rudoy, T. Quatieri, and P. Wolfe. Time-varying autoregressions in speech: Detection theory and applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):977–989, May 2011.

[132] H. L. Rufiner. *Análisis y Representación de la Voz Mediante Técnicas No Convencionales*. Disertación doctoral, Universidad Buenos Aires, 2005.

[133] H. L. Rufiner. *Análisis y modelado digital de la voz. Técnicas recientes y aplicaciones*. Ediciones UNL, Santa Fe, 1 edition, 2009.

[134] D. Ruinskiy and Y. Lavner. Stochastic models of pitch jitter and amplitude shimmer for voice modification. In *Proc. IEEEI 2008*, pages 489–493, 2008.

[135] O. Schleusing, T. Kinnunen, B. Story, and J.-M. Vesin. Joint Source-Filter Optimization for Accurate Vocal Tract Estimation Using Differential Evolution. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1560–1572, 2013.

[136] G. Schlotthauer. *Análisis de señales con descomposición empírica en modos y aplicaciones a la señal de voz*. Disertación doctoral, Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, 2010.

[137] G. Schlotthauer, M. E. Torres, and H. L. Rufiner. *Pathological Voice Analysis and Classification Based on Empirical Mode Decomposition*, volume 5967 of *Lecture Notes in Computer Science*, pages 364–381. Springer-Verlag, Berlin Heidelberg, 2010.

[138] J. Schoentgen. Stochastic models of jitter. *The Journal of the Acoustical Society of America*, 109(4):1631–1650, 2001.

[139] J. Schoentgen. Decomposition of Vocal Cycle Length Perturbations into Vocal Jitter and Vocal Microtremor, and Comparison of Their Size in Normophonic Speakers. *Journal of Voice*, 17(2):114–125, 2003.

[140] J. Schoentgen. Spectral models of additive and modulation noise in speech and phonatory excitation signals. *The Journal of the Acoustical Society of America*, 113(1):553–562, 2003.

[141] J. Schoentgen and R. de Guchteneere. Time series analysis of jitter. *Journal of Phonetics*, 23(1–2):189 – 201, 1995.

[142] J. Schoentgen and R. de Guchteneere. Predictable and random components of jitter. *Speech Communication*, 21(4):255 – 272, 1997.

[143] M. R. Schroeder. A brief history of synthetic speech. *Speech Communication*, 13(1):231 – 237, 1993.

[144] M. R. Schroeder. *Computer speech: recognition, compression, synthesis*, volume 35. Springer Science & Business Media, 2013.

[145] Y.-L. Shue and A. Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5134–5137, 2010.

[146] J. Skoglund and W. Kleijn. On time-frequency masking in voiced speech. *Speech and Audio Processing, IEEE Transactions on*, 8(4):361–369, 2000.

[147] R. Sousa, A. Ferreira, and P. Alku. The harmonic and noise information of the glottal pulses in speech. *Biomedical Signal Processing and Control*, 10:137 – 143, 2014.

[148] A. E. Stassi. *Técnicas de sensado y cuantificación para la evaluación objetiva de la actividad vocal*. Proyecto final de bioingeniería, Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, 2015.

[149] A. E. Stassi and G. A. Alzamendi. Protocolo para la adquisición y el almacenamiento de señales biomédicas asociadas a la actividad vocal. Technical report, Laboratorio de Señales y Dinámicas No Lineales, Facultad de Ingeniería de la Universidad Nacional de Entre Ríos, 2014.

[150] A. E. Stassi, G. A. Alzamendi, G. Schlotthauer, and M. E. Torres. Vocal fold activity detection from speech related biomedical signals: a preliminary study. In *VI Latin American Conference on Biomedical Engineering (CLAIB 2014)*, 2014.

[151] K. N. Stevens. *Acoustic Phonetics*. Number 30 in Current studies in linguistic. MIT Press, 2000.

[152] B. H. Story. *Physiologically-Based Speech Simulation Using an Enhanced Wave-Reflection Model of the Vocal Tract*. PhD thesis, University of Iowa, 1995.

[153] B. H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

[154] B. H. Story. Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. *The Journal of the Acoustical Society of America*, 123(1), 2008.

[155] B. H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.

[156] B. H. Story, I. R. Titze, and E. A. Hoffman. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–554, 1996.

[157] J. P. Teixeira, C. Oliveira, and C. Lopes. Vocal acoustic analysis - Jitter, Shimmer and HNR parameters. *Procedia Technology*, 9:1112 – 1122, 2013.

[158] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Pearson Education, 1992.

[159] M. Thomas, J. Gudnason, and P. Naylor. Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):82–91, 2012.

[160] M. Thomas and P. Naylor. The SIGMA algorithm: A glottal activity detector for electroglottographic signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1557–1566, 2009.

[161] I. R. Titze. Interpretation of the electroglottographic signal. *Journal of Voice*, 4(1):1 – 9, 1990.

[162] I. R. Titze. Workshop on acoustic voice analysis: summary statement. Technical report, National Center for Voice and Speech, Denver, USA, 1995.

[163] I. R. Titze. *Principles of Voice Production*. The National Center for Voice and Speech, Iowa, USA, 2 edition, 2000.

[164] I. R. Titze. Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *The Journal of the Acoustical Society of America*, 111(1):367–376, 2002.

[165] I. R. Titze. *The myoelastic Aerodynamic Theory of Phonation*. The National Center for Voice and Speech, Iowa, USA, 1 edition, 2006.

[166] I. R. Titze. Nonlinear source–filter coupling in phonation: Theory. *The Journal of the Acoustical Society of America*, 123(5):2733–2749, 2008.

[167] I. R. Titze and H. Liang. Comparison of f_0 extraction methods for high-precision voice perturbation measurements. *Journal of Speech, Language, and Hearing Research*, 36(6):1120–1133, 1993.

[168] I. R. Titze and B. H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112:1064, 2002.

[169] G. Tortora and B. Derrickson. *Principios de anatomía y fisiología*. Panamericana, México D.F., México, 13 edition, 2014.

[170] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford. Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive kalman filtering. *The Journal of the Acoustical Society of America*, 135(5):2885–2901, 2014.

[171] V. Tsiaras, R. Maia, V. Diakoloukas, Y. Stylianou, and V. Digalakis. Linear dynamical models in speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 300–304, 2014.

[172] M. Uzcanga Lacabe, S. Fernandez Gonzalez, M. Marques Girbau, L. Sarrasqueta-Sáenz, and R. Garcia-Tapia-Urrutia. Voz cantada. *Revista de Medicina de la Universidad de Navarra*, 50(3):49–55, 2006.

[173] W. Von Kempelen. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine.* Wien: JB Degen, 1791.

[174] J. Walker and P. Murphy. A review of glottal waveform analysis. In Y. Stylianou, M. Faundez-Zanuy, and A. Esposito, editors, *Progress in Nonlinear Speech Processing*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer Berlin Heidelberg, 2007.

[175] L. Wang, A. Li, and Q. Fang. A method for decomposing and modeling jitter in expressive speech in chinese. In *Proc. of Speech Prosody*, 2006.

[176] A. Yasmin. *Speech Enhancement Using Voice Source Models.* PhD thesis, University of Waterloo, Waterloo, Ontario, Canada, 1999.

[177] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen. Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis. *Journal of Voice*, 19(4):519–528, 2005.

[178] P. Šidlof, J. G. Švec, J. Horáček, J. Veselý, I. Klepáček, and R. Havlík. Geometry of human vocal folds and glottal channel for mathematical and biomechanical modeling of voice production. *Journal of Biomechanics*, 41(5):985 – 995, 2008.

[179] J. G. Švec, I. R. Titze, and P. S. Popolo. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America*, 117(3):1386–1394, 2005.