UNIVERSIDAD NACIONAL DEL LITORAL

Desarrollo de métodos robustos y fisiológicamente inspirados para el filtrado inverso de la voz

Iván Ariel Zalazar

FICH
FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE MÉTODOS COMPUTACIONALES

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

Tesis de Doctorado **2026**

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

DESARROLLO DE MÉTODOS ROBUSTOS Y FISIOLÓGICAMENTE INSPIRADOS PARA EL FILTRADO INVERSO DE LA VOZ

Iván Ariel Zalazar

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

2026

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

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

DESARROLLO DE MÉTODOS ROBUSTOS Y FISIOLÓGICAMENTE INSPIRADOS PARA EL FILTRADO INVERSO DE LA VOZ

Iván Ariel Zalazar

Lugar de Trabajo:
Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática (IBB), CONICET-UNER
Facultad de Ingeniería, Universidad Nacional de Entre Ríos

Director:
Dr. Gabriel A. Alzamendi — IBB CONICET-UNER

Co-director:
Dr. Gastón Schlotthauer — IBB CONICET-UNER

Jurado evaluador:
Alejandro Weinstein — UTFSM
Humberto Torres — INIGEM-CONICET
Juan Ignacio Godino Llorente — UPM

2026

ACTA DE EVALUACIÓN DE TESIS DE DOCTORADO

En la sede de la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, a los once días del mes de marzo del año dos mil veintiséis, se reúnen en forma online sincrónica los miembros del Jurado designado para la evaluación de la Tesis de Doctorado en Ingeniería, Mención Inteligencia Computacional, Señales y Sistemas, titulada ***"Desarrollo de métodos robustos y fisiológicamente inspirados para el filtrado inverso de la voz."***, desarrollada por el **Ing. Iván Ariel ZALAZAR**, DNI: 38.172.163, bajo la dirección del Dr. Gabriel Alzamendi y la codirección del Dr. Gastón Schlotthauer. Ellos son: Dr. Alejandro Weinstein, el Dr. Humberto Torres, y el Dr. Juan Godino Llorente.----------

La Presentación oral y defensa de la Tesis se efectúan bajo la modalidad online sincrónica según lo establecido por Resolución CS N° 382/21.

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

Que la tesis aborda aspectos muy relevantes en el ámbito del procesado y análisis de la calidad de la voz, utilizando metodologías muy novedosas, que van más allá del estado del arte. Su originalidad ha sido evaluada en distintas publicaciones relacionadas con la corriente principal de la tesis.

El texto de la tesis está excelentemente organizado, detallado y presentado.

La presentación oral fue muy clara y ordenada, destacando lo más importante del trabajo. Ante las preguntas del jurado, respondió con profundidad y solidez, demostrando un amplio conocimiento del tema.

Por ello el Jurado aprueba la Tesis con calificación 10 (Diez) Sobresaliente

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

Dr. Alejandro Weinstein	Dr. Humberto Torres	Dr. Juan Godino Llorente

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS · SANTA FE

Dr. José Luis Macor
Director de Posgrado
FICH - UNL

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

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

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 11 de marzo de 2026

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Desarrollo de métodos robustos y fisiológicamente inspirados para el filtrado inverso de la voz"***, desarrollada por el Ing. Iván Ariel ZALAZAR, 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 la versión digital final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

| --- | --- | --- |
| Dr. Alejandro Weinstein | Dr. Humberto Torres | Dr. Juan Godino Llorente |

Santa Fe, 11 de marzo de 2026

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	Dr. Gabriel Alzamendi Director de Tesis

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS • SANTA FE •

Dr. José Luis Macor
Director de Posgrado
FICH - UNL

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

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

Declaración del Autor

Esta disertación ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería, mención Inteligencia Computacional, Señales y Sistemas, ante la Universidad Nacional del Litoral y ha sido depositada en el Repositorio Institucional de Acceso Abierto (RIAA) de la Facultad de Ingeniería y Ciencias Hídricas para que esté disponible para sus lectores bajo las condiciones estipuladas.

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

Iván Ariel Zalazar

Formato de tesis

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución No255/17 (Expte. No888317-17) por el Comite Académico de la Carrera de Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). La resolución establece que:

"*En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por el director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión.*"

Dedicado a Lucía, a mi familia y a ellas

Agradecimientos

En primer lugar, quiero agradecer a mis directores, Gabriel Alzamendi y Gastón Schlotthauer, por acompañarme a lo largo de estos años en mi formación. Esta tesis es, en gran medida, el resultado de sus enseñanzas, su buena predisposición y su infinita paciencia. Les agradezco a ambos por confiar en mí, por alentarme, por guiarme y por estar siempre presentes.

En segundo lugar, agradezco al Consejo Nacional de Investigaciones Científicas y Técnicas por otorgarme la beca que permitió financiar mi carrera de doctorado. También agradezco a la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral por brindarme la posibilidad de realizar el doctorado de manera gratuita. Por último, agradezco el apoyo brindado por el Programa Integral de Fortalecimiento del Posgrado de la Universidad Nacional de Entre Ríos.

Quiero agradecer también a mis compañeros del Laboratorio de Señales y Dinámicas no Lineales. En particular, a Joaquín Ruiz y a Marcelo Colominas, por sus valiosos aportes al desarrollo de esta tesis. Asimismo, agradezco a Felipe Restrepo y a Joaquín Monti por su acompañamiento.

En lo personal, deseo expresar mi eterno agradecimiento a mi madre Sandra, a mis hermanos Gonzalo, Lucas y Tamara, y a mi sobrina Nicasia, por estar siempre a mi lado, brindarme su cariño y alentarme en cada etapa. También quiero agradecer a Catalina, Juan, Luisina, Sacha y Santiago, quienes han sido una segunda familia para mí.

Finalmente quiero agradecerle a Lucía Alonso por ser mi compañera. Su incondicional cariño y paciencia han hecho que esta etapa de mi vida sea una de las más felices. Gracias por acompañarme en cada momento.

Iván Ariel Zalazar

Resumen

El flujo glótico, la principal fuente acústica en la fonación humana, resulta de las complejas interacciones biomecánicas en la glotis. Por lo tanto, proporciona información sobre la dinámica de las cuerdas vocales. El filtrado inverso de la voz permite la estimación no invasiva del flujo glótico a partir de la señal de voz. Esto se logra eliminando primero la contribución del tracto vocal mediante el ajuste de un filtro digital, obteniéndose así la función glótica, una señal que contiene información sobre el flujo glótico y la radiación de los labios. Luego, al eliminar las modulaciones debidas a la radiación de los labios, se obtiene el flujo glótico. En general, la precisión de la estimación depende de cancelar correctamente las contribuciones de estas estructuras.

Esta tesis introduce nuevos métodos para mejorar los dos pasos fundamentales del filtrado inverso de la voz. En primer lugar, se examinan los problemas asociados con el uso del método de predicción lineal para ajustar el filtro del tracto vocal. A partir de este análisis, se proponen dos estrategias de predicción lineal ponderada que aplican atenuación Gaussiana para reducir los errores en el ajuste del filtro debidos a la influencia adversa de las muestras ubicadas en los instantes de cierre glóticos. Las estrategias propuestas extienden el método de predicción lineal con atenuación Gaussiana, permitiendo un análisis adaptado a la periodicidad de la señal y una ponderación de fase casi cerrada, lo que resulta en un mejor desempeño para aplicaciones de filtrado inverso.

Adicionalmente, se desarrolló un método de predicción lineal basado en el criterio de máxima correntropía, resultando en una estrategia robusta para filtrado inverso. Este método implementa un esquema de ponderación que enfatiza automáticamente las muestras de la señal de voz en la fase cerrada, las cuales contienen información más precisa del tracto vocal, mientras atenúa simultáneamente las muestras alrededor de los instantes de cierre glóticos que generan errores. Esto proporciona una ventaja significativa sobre los métodos que requieren conocer a priori los instantes glóticos.

Finalmente, se propuso un modelo adaptativo no armónico para mejorar la estimación del flujo glótico a partir de la función glótica. En base a esta formulación, se desarrolló una versión regularizada del modelo que permite obtener estimaciones con una fase cerrada plana, lo cual es una característica fisiológicamente relevante de la forma de onda del flujo glótico. Este enfoque reduce las distorsiones de baja frecuencia causadas por errores que surgen durante el filtrado inverso.

En conjunto, los métodos desarrollados en esta tesis constituyen contribuciones significativas al campo del filtrado inverso de la voz y complementan los métodos establecidos. Estas contribuciones mejoran las herramientas disponibles para el análisis de la fonación y sientan las bases para futuras investigaciones en esta temática.

Abstract

The glottal airflow, the principal acoustic source in human phonation, results from complex biomechanical interactions at the glottis. Thus, it provides insight into key aspects of the underlying vocal fold dynamics. Voice inverse filtering enables the non-invasive estimation of glottal airflow from the voice signal. This is achieved by first removing the vocal tract contribution through a properly tuned digital filter; it results in the glottal function, a signal carrying information about glottal airflow and lip radiation. Once the modulations due to lip radiation are removed, the glottal airflow is obtained. In general, estimation accuracy depends critically on effectively cancelling the contributions of these structures.

This thesis introduces novel methods for improving the two fundamental steps in voice inverse filtering. First, the problems associated with using the linear prediction method for tuning the vocal tract filter are examined. Based on this analysis, two weighted linear prediction strategies are proposed that apply Gaussian attenuation to reduce vocal tract misadjustments due to the adverse influence of voice samples around the glottal closure instants. The proposed strategies extend the Gaussian linear prediction method, allowing a pitch-adaptive analysis and a quasi closed phase weighting, resulting in improved performances for inverse filtering applications.

Additionally, a linear prediction method based on the maximum correntropy criterion was developed, resulting in a robust inverse filtering approach. This method implements a weighting scheme that automatically emphasizes voice samples in the closed phase that carry more accurate vocal tract information, while simultaneously attenuating the samples around the glottal closure instants that yield increased estimation errors. This provides a significant advantage over methods that require prior knowledge of glottal instants.

Finally, an adaptive non-harmonic model was proposed to enhance the estimation of glottal airflow from the glottal function. Based on this formulation, a regularized version of the model was developed to produce estimates with a flat closed phase, which is a physiologically relevant characteristic of the glottal airflow waveform. This approach reduces low-frequency distortions caused by errors that arise during inverse filtering.

Collectively, the methods developed in this thesis constitute significant contributions to the field of voice inverse filtering and complement the established methods. These contributions enhance the tools available for phonation analysis and establish a framework for future research in this field.

Índice general

Índice de figuras

Lista de Tablas

Capítulo 1

Introducción

1.1. Evaluación no invasiva de la fonación humana

En la fonación humana, las cuerdas vocales constituyen los principales órganos de la laringe responsables de la producción de sonidos vocales [8]. Analizar su comportamiento, también denominado *función laríngea*, permite obtener información valiosa para caracterizar la calidad de la fonación y el estado de salud de los órganos laríngeos involucrados [9]. Desafortunadamente, el estudio directo de la laringe es un proceso complejo que requiere técnicas invasivas y equipamiento especializado de alto costo, tales como laringoscopios, estroboscopios o cámaras de alta velocidad. Además, este tipo de procedimientos demanda la intervención de profesionales de la salud altamente calificados [10].

Una alternativa a la observación directa consiste en caracterizar de forma indirecta la función laríngea mediante el análisis de la señal de flujo glótico, también conocida como *fuente glótica* [11]. Esta señal representa la corriente de aire proveniente de los pulmones que, al atravesar la laringe, es modulada por la acción de apertura y cierre de las cuerdas vocales [12]. Como resultado, el flujo glótico presenta una forma de onda cuasiperiódica y pulsátil con dos fases bien diferenciadas (ver Fig. 1.1). Por un lado, la fase abierta (1 a 5) corresponde a los instantes en los que las cuerdas vocales permanecen abiertas y el flujo glótico toma valores positivos. Por otro lado, la fase cerrada (6 a 1) se caracteriza por un flujo mínimo debido al cierre de las cuerdas vocales. Los puntos 1 y 5 de la Fig. 1.1 marcan los instantes de apertura y cierre glóticos, conocidos respectivamente como GOI y GCI[1] [13].

La importancia del flujo glótico radica en que porta información relevante acerca de la dinámica de las cuerdas vocales, la cual es propia de cada hablante [13]. Por esta

[1] Siglas correspondientes a las expresiones en inglés: glottal opening instant y glottal closure instant.

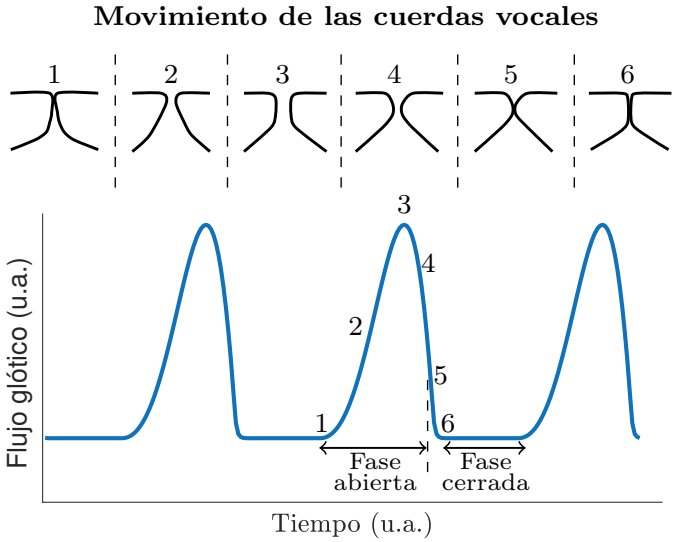

FIGURA 1.1: Relación entre el movimiento de las cuerdas vocales y el flujo glótico en un ciclo glótico ilustrativo. La parte superior de la figura representa esquemáticamente la posición de las cuerdas vocales, en un corte transversal, en instantes determinados del ciclo glótico.

razón, se la utiliza en diversas aplicaciones tales como la identificación de hablantes [14], [15], el reconocimiento de emociones [16], [17] y la detección de trastornos y enfermedades de la voz [18]-[20].

Desafortunadamente, el sensado del flujo glótico sufre de importantes dificultades de instrumentación y aplicación, lo que ha limitado su uso en la práctica clínica [10]. Esta situación ha impulsado el desarrollo de estrategias indirectas para estimar el flujo glótico basadas en el procesamiento digital de señales asociadas a la fonación humana y en la resolución de problemas inversos [12]. Entre las diferentes estrategias propuestas, las más difundidas involucran el análisis de señales de voz, de flujo de aire a nivel de los labios, de aceleración de la piel del cuello y de electroglotografía [13], [21]-[24]. Recientemente han surgido también enfoques basados en aprendizaje profundo [25]. Sin embargo, el entrenamiento de estos modelos requiere grandes bases de datos que incluyan una amplia variedad de señales de flujo glótico. Lamentablemente, las bases de datos públicas con estas características son escasas [26].

Entre las distintas alternativas, el filtrado inverso de la señal de voz constituye la técnica más estudiada y ampliamente aplicada [13]. Esto se debe a que únicamente requiere equipamiento para la adquisición de señales de voz, una tecnología de bajo costo y fácil acceso [12]. Sin embargo, los métodos existentes presentan diversas limitaciones que dificultan obtener de forma correcta el flujo glótico.

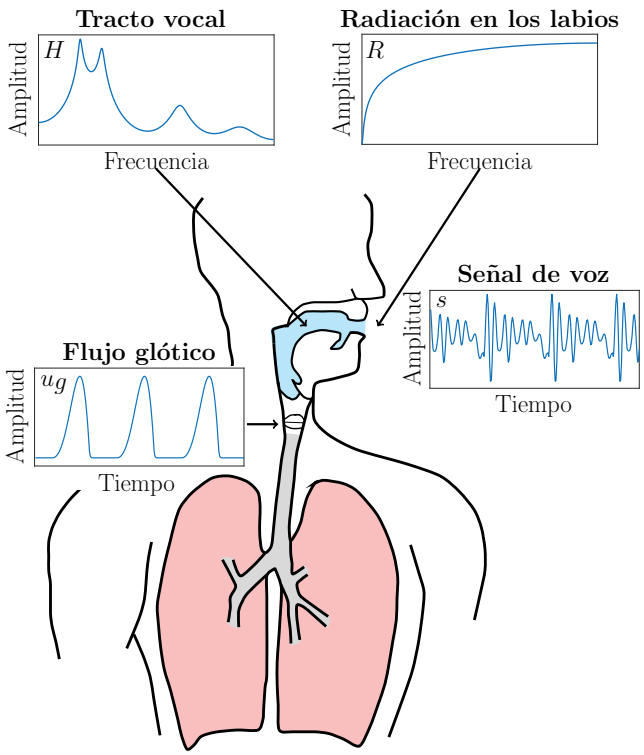

FIGURA 1.2: Esquema ilustrativo de la teoría *fuente-filtro* de la fonación para la producción de la voz sonora.

La presente tesis doctoral tiene como propósito contribuir al filtrado inverso de la voz mediante el desarrollo de métodos que permitan estimar el flujo glótico con mayor precisión, mejorando así la evaluación no invasiva de la función laríngea.

1.2. Filtrado inverso de la voz basado en la teoría *fuente-filtro*

El filtrado inverso se basa en la teoría *fuente-filtro* de la fonación, la cual permite explicar la producción de la voz sonora a partir de la combinación de tres simples elementos: una fuente de excitación glótica (el flujo glótico), un filtro acústico con las resonancias del tracto vocal, responsable de modular la riqueza espectral de los distintos sonidos, y una impedancia de radiación en los labios [27], tal como se ilustra en la Fig 1.2.

La teoría *fuente-filtro* supone un comportamiento lineal del sistema fonador, despreciando las interacciones acústicas no lineales presentes en el proceso de la fonación

[28]. Además, se toma como hipótesis que el tracto vocal se mantiene invariante en ventanas de corta duración (entre 20 y 50 ms) [29]. Por otro lado, esta teoría permite representar adecuadamente la información espectral por debajo de los 5 kHz [30]. Bajo estos supuestos, la producción de la voz puede modelarse matemáticamente como:

$$s[n] = u_g[n] * h_{vt}[n] * r_l[n], \tag{1.1}$$

donde $*$ es el operador convolución, s es la señal de voz, u_g es el flujo glótico y las respuestas al impulso h_{vt} y r_l modelan el aporte realizado por el tracto vocal y el efecto de radiación de los labios, respectivamente. En el dominio de la transformada $\mathcal{Z}\{\cdot\}$, la ecuación (1.1) puede expresarse como:

$$S(z) = U_g(z)H(z)R(z), \tag{1.2}$$

donde $S(z)$, $U_g(z)$, $H(z)$ y $R(z)$ representan las transformadas $\mathcal{Z}$ de cada uno de los elementos introducidos en la expresión anterior.

Basado en el esquema de la Fig. 1.2, el filtrado inverso aborda el problema inverso de la fonación, cuyo objetivo es obtener el flujo glótico a partir de cancelar los aportes del tracto vocal y la radiación de los labios de la señal de voz [12]:

$$U_g(z) = \frac{S(z)}{H(z)R(z)} = S(z)H(z)^{-1}R(z)^{-1}, \tag{1.3}$$

donde $H(z)^{-1}$ y $R(z)^{-1}$ son los modelos inversos del tracto vocal y el efecto de radiación de los labios, respectivamente.

Bajo la hipótesis de linealidad del sistema fonador, los modelos inversos en la Ec. (1.3) pueden aplicarse en la señal de voz de manera conjunta o secuencial, obteniéndose el mismo resultado independientemente del orden de aplicación [12]. A continuación, se describe la implementación secuencial del filtrado inverso de la voz.

1.3. Filtrado inverso de la voz secuencial

En la práctica, la mayoría de los métodos de filtrado inverso de la voz emplean un procedimiento secuencial para estimar el flujo glótico [13], tal como se ilustra en la Fig. 1.3.

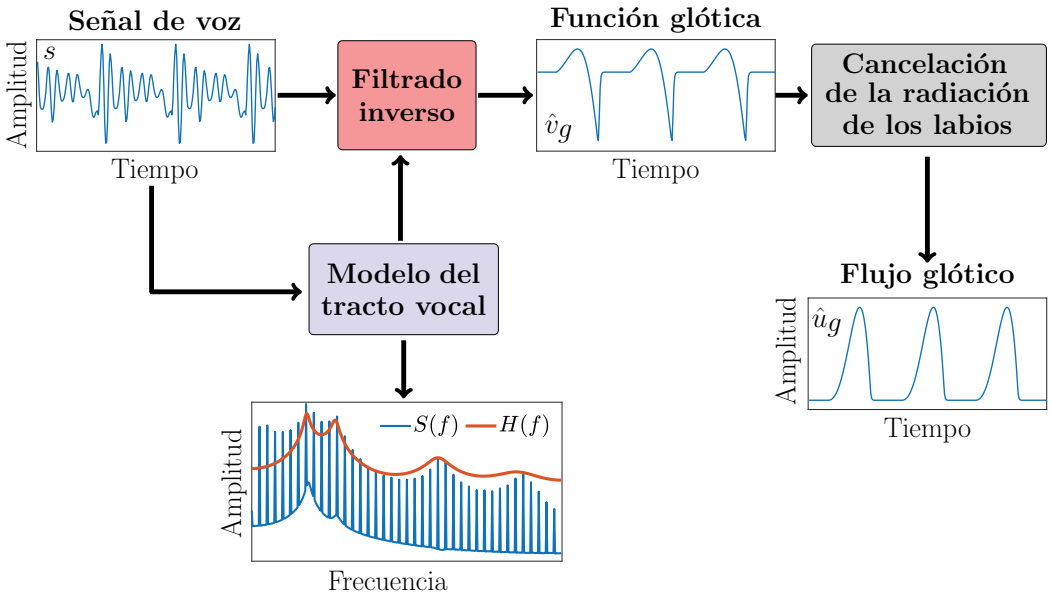

FIGURA 1.3: Esquema secuencial del filtrado inverso de la voz.

La primera etapa consiste en eliminar el aporte del tracto vocal [12]. Para ello, el modelo $H(z)$ se ajusta a partir de la información espectral en la señal de voz con el fin de capturar las principales resonancias (frecuencias *formantes*) introducidas por el tracto vocal [12]. Luego, se aplica el filtro inverso del tracto vocal $H(z)^{-1}$ a la señal de voz, obteniendo así una señal intermedia denominada *función glótica*, v_g, que contiene información del flujo glótico y de la modulación generada por el efecto de radiación de los labios. Matemáticamente, se expresa:

$$V_g(z) = \mathcal{Z}\{v_g(n)\} = S(z)H(z)^{-1} = U_g(z)R(z). \tag{1.4}$$

La segunda etapa del filtrado inverso secuencial se centra en cancelar la modulación introducida por el efecto de radiación de los labios de la función glótica empleando el modelo inverso $R(z)^{-1}$:

$$U_g(z) = V_g(z)R(z)^{-1}, \tag{1.5}$$

lo que permite obtener una estimación del flujo glótico.

El éxito del filtrado inverso de la voz depende de eliminar completamente los aportes del tracto vocal y el efecto de la radiación de los labios. En consecuencia, la precisión en la estimación de u_g requiere de la correcta caracterización y modelización de $H(z)$ y $R(z)$ [13]. A continuación, se describen los modelos clásicos utilizados para cada uno de estos elementos, así como los inconvenientes asociados a su implementación en

el contexto del filtrado inverso.

1.3.1. Modelado del tracto vocal

El aporte acústico del tracto vocal puede modelarse mediante un filtro digital autorregresivo [29]:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{1.6}$$

donde P es el orden del filtro y a_k sus coeficientes, los cuales codifican la información del tracto vocal [31].

El método clásico para estimar los coeficientes de $H(z)$ es la predicción lineal (LP[2]), donde el ajuste computacional se basa en minimizar el error cuadrático de predicción [32]. Esta estrategia funciona adecuadamente cuando el error de predicción posee una distribución similar a la del ruido blanco [32]. Sin embargo, en el contexto de la fonación sonora, esta condición no se cumple. En general, los errores de predicción generados al analizar la señal de voz mediante LP suelen presentar valores atípicos de gran amplitud concentrados alrededor de los GCIs [33]. Estos errores afectan negativamente al método LP debido a la elevada sensibilidad del error cuadrático a muestras de gran amplitud [34], generando filtros que no modelan correctamente la modulación del tracto vocal.

Como consecuencia, aplicar filtrado inverso basado en el método LP produce una cancelación deficiente del aporte del tracto vocal en la señal de voz, dando lugar a distorsiones en la forma de onda de la función glótica, especialmente en la fase cerrada [35], [36].

La figura 1.4 muestra un ejemplo de los problemas generados por un filtrado inverso ineficiente con el método LP para una señal de voz sintética correspondiente a una vocal /u/. La columna izquierda muestra el espectro de magnitud del filtro del tracto vocal estimado con LP, H_{LP}, junto con la ubicación de las cuatro primeras frecuencias formantes (F_i, con $i = 1,2,3,4$). La columna derecha muestra la estimación de la función glótica, $\hat{v}_g$, que resulta de aplicar el filtro inverso H_{LP}^{-1} en la señal de voz.

Como se puede observar, el filtro del tracto vocal obtenido presenta errores significativos en la ubicación de sus picos de resonancia, específicamente en las dos primeras formantes F_1 y F_2. Estos errores en el filtro dan lugar a oscilaciones espurias en la función glótica estimada, las cuales posteriormente afectan la forma de onda final del flujo glótico.

[2]Sigla correspondiente a la expresión en inglés: linear prediction.

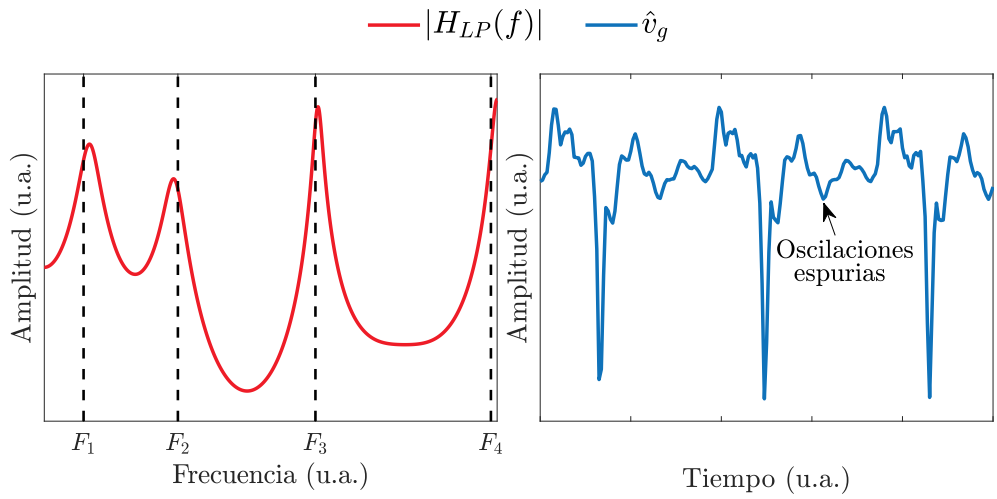

FIGURA 1.4: Errores de filtrado inverso de la voz derivados de una estimación inadecuada del filtro del tracto vocal con el método LP, correspondiente a una vocal /u/. Izquierda: espectro de magnitud del filtro obtenido junto con la ubicación de las frecuencias formantes. Derecha: función glótica estimada mediante filtrado inverso.

1.3.2. Cancelación de la radiación de los labios

La radiación de los labios se encarga de transformar el flujo de aire en ondas de presión sonora [8]. Usualmente, este efecto puede modelarse como la respuesta de un filtro FIR diferenciador de primer orden [29]:

$$R(z) = 1 - \alpha z^{-1} \text{ con } 0 < \alpha \leq 1, \tag{1.7}$$

por lo que, de la Ecs. (1.4) y (1.7), se desprende que la función glótica puede interpretarse como la derivada temporal del flujo glótico [12].
La radiación de los labios puede compensarse aplicando a la función glótica un filtro integrador con pérdidas (LIF[3]) de la forma:

$$R(z)^{-1} = \frac{1}{1 - \alpha z^{-1}}. \tag{1.8}$$

El resultado de este proceso de integración corresponde a una estimación del flujo glótico.

En general, el parámetro α en la Ec. (1.8) se fija manualmente dentro del intervalo $(0, 1]$. Un criterio habitual consiste en seleccionar el valor que produzca la forma de onda del flujo glótico con la fase cerrada más plana (similar a la mostrada en la Fig. 1.1)

[3]Sigla correspondiente a la expresión en inglés: leaky integrator filter.

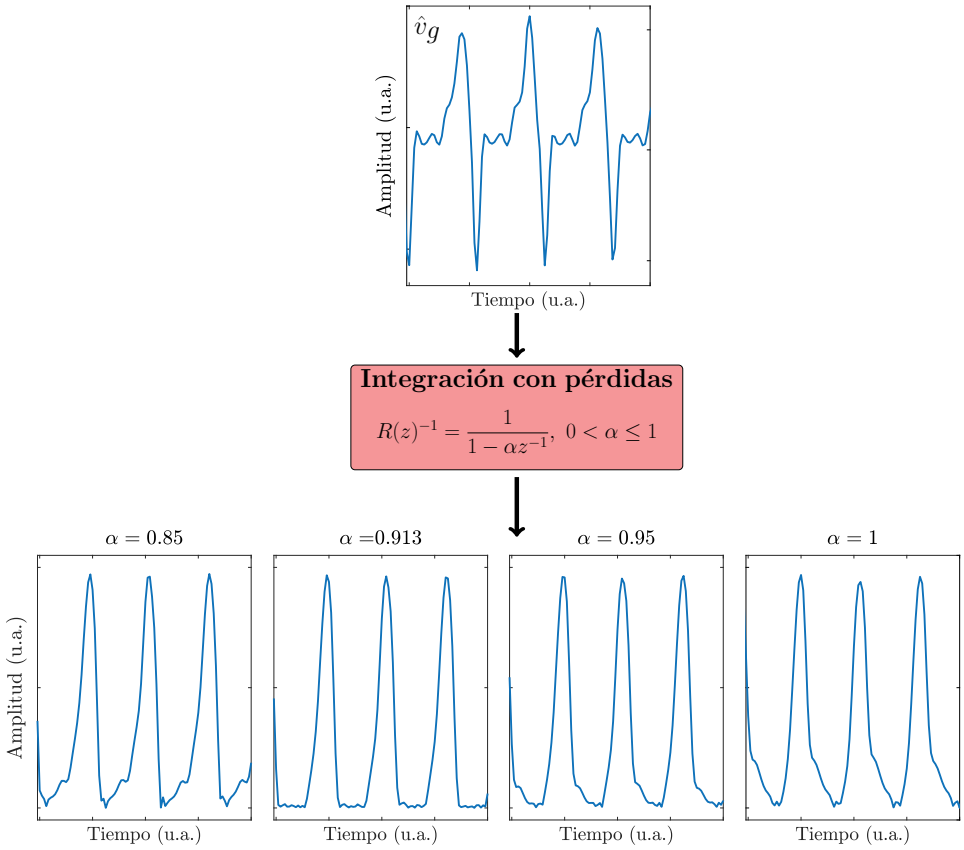

FIGURA 1.5: Efecto de variar el parámetro α en la estimación del flujo glótico aplicando el filtro integrador con pérdidas.

[37]. No obstante, este ajuste resulta laborioso cuando se procesan grandes cantidades de señales. En [37] se propone una selección automática basada en detectar cambios en el área del flujo glótico estimado. Lamentablemente, la selección de α resulta muy sensible al umbral de decisión empleado para detectar dichos cambios.

Si bien el filtro (1.8) es simple de implementar, en la práctica una elección inadecuada de α puede distorsionar la forma de onda del flujo glótico [37]. Estas distorsiones suelen manifestarse como una componente de muy baja frecuencia ascendente o descendente en la fase cerrada del ciclo glótico, tal como se observa en la Fig. 1.5.

1.4. Objetivos

La presente tesis doctoral tiene como propósito contribuir al campo del filtrado inverso de la señal de voz mediante el desarrollo de métodos más precisos, robustos,

simples y fisiológicamente inspirados. Para ello, se siguen los objetivos que se detallan a continuación.

1.4.1. Objetivo general

Desarrollar nuevos métodos para el filtrado inverso de la señal de voz que mejoren la estimación del flujo glótico.

1.4.2. Objetivos específicos

- Comparar diferentes estrategias de filtrado inverso de la voz en términos de su desempeño, robustez, aplicabilidad y sustento fisiológico.
- Aportar al desarrollo de nuevas estrategias de predicción lineal inspiradas en la fisiología de la fonación que permitan estimar con mayor precisión la información del tracto vocal.
- Investigar el uso de funciones de costo alternativas para desarrollar nuevas estrategias robustas de predicción lineal útiles para el filtrado inverso de la voz.
- Estudiar distintos enfoques de integración para mejorar la cancelación del efecto de radiación de los labios en la función glótica.

1.5. Motivación y aportes

Tal como se analizó en la Sección 1.3, la precisión del flujo glótico estimado mediante filtrado inverso depende del ajuste apropiado del filtro del tracto vocal y de la correcta integración de la función glótica. En esta sección se presentan las motivaciones y los aportes principales de esta tesis orientados a mejorar estos aspectos relevantes del filtrado inverso de la voz.

1.5.1. Aportes a la predicción lineal con atenuación Gaussiana

El principal problema de la LP clásica radica en la sensibilidad que posee su función de costo, el error cuadrático, ante los errores de predicción de gran amplitud generados alrededor de los GCIs durante el análisis de la señal de voz [33], [34].

Para superar este inconveniente, se han propuesto variantes del método LP que buscan ponderar selectivamente las muestras de la señal de voz. El objetivo de esta ponderación es enfatizar las muestras de la señal que ayudan a ajustar con precisión el filtro del tracto vocal, a la vez que se atenúan aquellas que generan errores en dicho ajuste, como las ubicadas alrededor de los GCIs [35].

Estas estrategias emplean una función de ponderación w que controla la contribución de cada muestra de la señal de voz durante la minimización del error de predicción [35]. La LP clásica puede interpretarse como un caso particular, en el que la función de ponderación es constante para toda la señal de voz, y así todas las muestras aportan por igual al ajuste del filtro del tracto vocal (ver Fig. 1.6.a).

Otro ejemplo representativo es el método de covarianza de fase cerrada [36], el cual está inspirado en la fisiología de la fonación. Este método ajusta los coeficientes del filtro del tracto vocal empleando sólo muestras de la señal de voz contenidas en la fase cerrada de un ciclo glótico [38]. Para ello, se emplea una función de ponderación como la ilustrada en la Fig. 1.6.b, que prescinde de toda la señal de voz a excepción de un conjunto de muestras ubicadas en la fase cerrada.

El análisis de fase cerrada descrito permite desacoplar en la señal de voz la información del tracto subglótico de la correspondiente al tracto supraglótico, evitando así la influencia perjudicial que tienen las muestras cercanas a los GCIs o pertenecientes a la fase abierta [35]. Sin embargo, una limitación de este método radica en que requiere conocer con precisión la ubicación de los GCIs y GOIs para definir correctamente el intervalo de fase cerrada. En la práctica, la detección exacta de estos instantes glóticos resulta una tarea difícil, especialmente en el caso del instante de apertura [39].

La predicción lineal de fase casi cerrada (QCP[4]) [40] constituye una alternativa más simple, ya que requiere únicamente conocer los GCIs. En este enfoque, los instantes de cierre se utilizan en una función de ponderación paramétrica configurada de modo que atenúa significativamente la información correspondiente a la fase abierta y enfatiza la correspondiente a la fase cerrada, tal como se muestra en la Fig. 1.6.c.

Un estudio en la etapa temprana de formación del autor de esta tesis [1], mostró que el análisis de QCP mejora el cálculo de los coeficientes del filtro del tracto vocal para filtrado inverso, debido a que su función de ponderación permite aprovechar la información de varios ciclos glóticos, en contraste al método de covarianza de fase cerrada. Sin embargo, la correcta configuración de los parámetros de su función de

[4]Sigla correspondiente a la expresión en inglés: quasi closed phase.

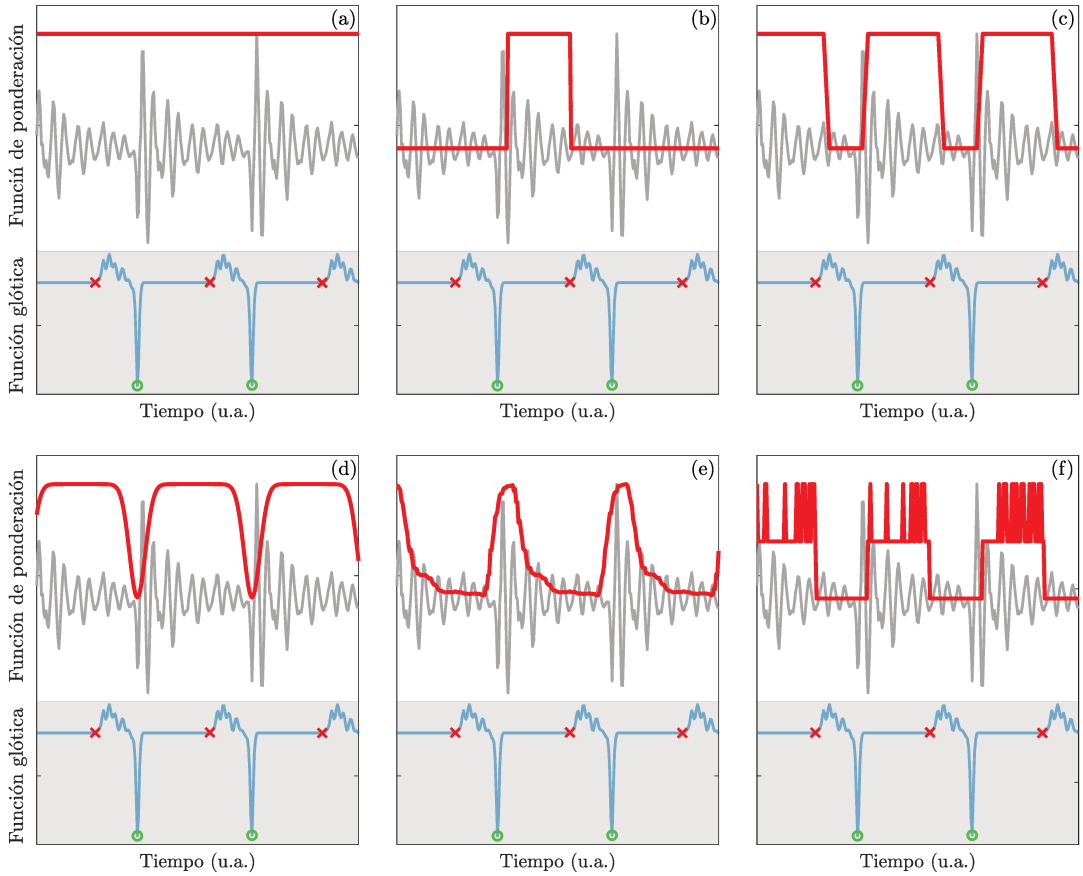

FIGURA 1.6: Estrategias de ponderación aplicadas a la señal de voz, correspondientes a los métodos: a) LP clásico, b) covarianza de fase cerrada, c) QCP, d) GLP, e) SWLP, f) PWLP. En cada figura se muestran la función de ponderación (curva roja), la señal de voz (curva gris) y la función glótica (curva celeste). Se indican además los GCIs y GOIs mediante marcas '○' y '×', respectivamente.

ponderación no es trivial; una parametrización inadecuada puede afectar negativamente el ajuste del filtro, comprometiendo así los resultados de filtrado inverso de la voz.

Una alternativa más sencilla a los métodos anteriormente mencionados es la predicción lineal con atenuación Gaussiana (GLP[5]) propuesta en [41]. En este enfoque se emplea una función de atenuación Gaussiana construida a partir de una serie de ventanas Gaussianas simétricas centradas en cada GCI:

$$w_{\text{sim}}[n] = 1 - \kappa \sum_{l=1}^{L} g_s[n - n_l], \tag{1.9}$$

donde n_l denota la ubicación del l-ésimo GCI (de un total de L), κ controla el nivel de

[5]Sigla correspondiente a la expresión en inglés: gaussian linear prediction.

atenuación alrededor de n_l, y $g_s[n - n_l]$ es una ventana Gaussiana centrada en n_l y y varianza σ_1^2:

$$g_s[n - n_l] = e^{-(n-n_l)^2/2\sigma_1^2}. \tag{1.10}$$

El subíndice "*sim*" en (1.9) indica que la función de ponderación atenúa simétricamente a ambos lados de cada n_l. El alcance de la región de atenuación alrededor de cada GCI es controlado por el parámetro σ_1 de la ventana Gaussiana. La Fig. 1.6.d muestra la atenuación Gaussiana implementada por w_{sim}.

El método GLP presenta la ventaja que su función de ponderación requiere establecer únicamente dos parámetros: κ y σ_1. Si bien este método ha sido utilizado en el contexto del filtrado inverso de la voz [42], demostrando además ser robusto frente a errores en la ubicación de los GCIs, no se ha estudiado en detalle cómo afectan sus parámetros a los resultados del filtrado inverso.

El primer aporte de esta tesis, presentado en el Anexo A, consistió en una revisión del método GLP que dio lugar a una parametrización alternativa para la función de atenuación Gaussiana [2]. Un aspecto fundamental considerado en la parametrización propuesta es la periodicidad de la voz, una característica que varía en función del sexo del hablante, la constitución corporal, la entonación y la calidad de la voz [13]. Asimismo, se desarrolló una nueva versión de la función de atenuación Gaussiana considerando una asimetría en su morfología que permite realizar una ponderación de fase casi cerrada durante el ajuste del filtro del tracto vocal [3]. Esta modificación permitió mejorar los filtros obtenidos con GLP y, en consecuencia, disminuir los errores de filtrado inverso de la voz.

1.5.2. Filtrado inverso de la voz basado en el criterio de máxima correntropía

Algunas variantes de LP realizan su ponderación de forma guiada por los datos, es decir, emplean exclusivamente la información contenida en la señal de voz para definir la función ponderación.

La predicción lineal ponderada [43] y su versión estabilizada (SWLP[6]) [44] son ejemplos de métodos guiados por los datos, los cuales emplean como función de ponderación la energía de la señal de voz calculada en ventanas de corta duración. Esta estrategia permite realzar aquellas muestras de la señal con mejor relación señal-ruido,

[6]Sigla correspondiente a la expresión en inglés: stabilized weighted linear prediction.

siendo útil en escenarios con niveles considerables de ruido. No obstante, en la práctica la función de energía tiende a enfatizar las muestras de la señal de voz con mayor amplitud, como las que ocurren alrededor de los GCIs (ver Fig. 1.6.e), lo que afecta el cálculo de los coeficientes del filtro del tracto vocal [35].

Por otro lado, la predicción lineal con ponderación probabilista (PWLP[7]) introducida en [39] emplea una función de ponderación guiada por los datos la cual tiende a resaltar adaptativamente la información de la señal de voz contenida en la fase cerrada sin necesidad de conocer su ubicación (ver Fig. 1.6.f). Un inconveniente de este método radica en que hace uso de diferentes distribuciones a priori para los parámetros del modelo, cuya interpretación no resulta clara en el contexto de la fisiología de la fonación.

Dada la dificultad que posee la estimación de los instantes de apertura y cierre glóticos, los métodos guiados por los datos resultan una alternativa atractiva, especialmente si permiten realizar un análisis de predicción lineal de fase cerrada de forma automática que sirva para mejorar el ajuste del filtro del tracto vocal [39], [1].

El segundo aporte de esta tesis, presentado en el Anexo B, introduce un nuevo método de filtrado inverso guiado por los datos denominado predicción lineal basada en el criterio de máxima correntropía. Este método emplea la correntropía con núcleo Gaussiano como función de costo, la cual es robusta frente a errores de predicción atípicos y de gran amplitud como los que se dan durante el análisis de la señal de voz [4], [5]. Nuestra propuesta derivó en un algoritmo que ajusta iterativamente los coeficientes del filtro, permitiendo obtener un mejor ajuste de la información del tracto vocal.

1.5.3. Modelado adaptativo no armónico para la estimación del flujo glótico

El cálculo del flujo glótico a partir de su derivada, la función glótica, es fundamental en el filtrado inverso de la voz. Desafortunadamente, este aspecto ha recibido poca atención en la comunidad científica, donde la mayoría de los trabajos se han centrado en mejorar el cálculo de los coeficientes del filtro del tracto vocal, empleando directamente el filtro integrador con pérdidas de la Ec. (1.8) pese a sus inconvenientes.

[7] Sigla correspondiente a la expresión en inglés: probabilistic weighted linear prediction.

En [45] se introdujo una estrategia de filtrado inverso basada en programación cuadrática (QPR[8]), en la cual se propone una formulación alternativa que combina en un único filtro los efectos del tracto vocal y la radiación en los labios. Los coeficientes de este filtro se estiman resolviendo un problema de optimización de programación cuadrática basado en el análisis de fase cerrada. A diferencia del esquema tradicional, este método permite obtener directamente la forma de onda del flujo glótico, sin necesidad de estimar señales intermedias, evitando así los inconvenientes asociados al uso del LIF. Además, el análisis de fase cerrada propuesto garantiza que las estimaciones del flujo glótico presenten una forma de onda plana y libre de distorsiones en dicha fase, lo cual es una característica deseable y fisiológicamente representativa. Sin embargo, estudios recientes indican que el método QPR presenta dificultades para estimar con precisión la forma de onda del flujo glótico fuera de la fase cerrada [28].

El tercer aporte de esta tesis, presentado en el Anexo C, consistió en desarrollar un nuevo método para estimar el flujo glótico a partir de la función glótica mediante un modelo adaptativo no armónico (ANH[9]) [6]. El modelo ANH permite representar señales oscilatorias multicomponentes que varían en el tiempo mediante un número fijo de funciones de forma de onda [46], [47]. Nuestro aporte también incluyó el desarrollo de una versión regularizada de este modelo, capaz de estimar con precisión el flujo glótico a la vez que promueve una forma de onda plana en la fase cerrada [7].

1.6. Organización del documento

El presente documento sigue el formato de tesis por compilación y se encuentra organizado de la siguiente manera:

- En el presente capítulo se describió el filtrado inverso de la señal de voz como alternativa para estimar de forma no invasiva el flujo glótico, junto con los problemas asociados a su implementación y aplicación. También, se plantearon los objetivos y se detallaron las motivaciones que dieron lugar a los aportes de esta tesis.

- En el Capítulo 2 se detallan brevemente la metodología seguida para evaluar los métodos de filtrado inverso de la voz y las dificultades que se afrontan a la hora de validar su desempeño. Asimismo, se describe la base datos considerada en esta

[8] Sigla correspondiente a la expresión en inglés: quadratic programming.
[9] Sigla correspondiente a la expresión en inglés: adaptive non-harmonic.

tesis y las medidas de desempeño empleadas en los experimentos presentados en los capítulos posteriores.

- En el Capítulo 3 se brindan mayores detalles sobre el cálculo de los coeficientes del filtro del tracto vocal mediante predicción lineal y sus inconvenientes en el contexto del análisis de la señal de voz. Luego, se presentan las variantes propuestas para la predicción lineal con atenuación Gaussiana y se discuten los resultados obtenidos.

- En el Capítulo 4 se realiza una revisión sobre el uso de funciones de costo alternativas en predicción lineal y se introduce la correntropía con núcleo Gaussiano junto con sus propiedades. Seguidamente, se presenta el método de predicción lineal basado en el criterio de máxima correntropía propuesto y se discute su desempeño para el filtrado inverso de la voz.

- El Capítulo 5 introduce el modelo adaptativo no armónico, como base para el desarrollo del modelo propuesto para la estimación del flujo glótico a partir de la función glótica. Posteriormente, se desarrolla el modelo regularizado y se analizan los resultados obtenidos.

- Finalmente, el Capítulo 6 presenta el cierra de la tesis, describiendo las conclusiones alcanzadas y las líneas de trabajo futuras.

1.7. Comentarios de fin de capítulo

En este capítulo se presentó la temática general en la que se enmarca esta tesis doctoral, centrada en el filtrado inverso de la señal de voz como estrategia para estimar el flujo glótico y caracterizar la función laríngea de forma no invasiva. Además, se discutieron las bases teóricas en las que se sustenta el filtrado inverso, los modelos empleados para el tracto vocal y la radiación en los labios, así como los desafíos que presenta su correcto ajuste.

Un estudio realizado en la primera etapa de formación del autor de esta tesis permitió analizar distintos métodos y estrategias empleadas para el filtrado inverso de la voz, con el propósito de identificar sus principales características, su relación con la fisiología de la fonación y sus limitaciones. A partir de este estudio se identificaron oportunidades de mejora para el filtrado inverso, las cuales motivaron los objetivos planteados y dieron lugar a los tres aportes principales de esta tesis.

Es importante mencionar que este estudio comparativo dio lugar a una presentación en un congreso nacional [1], constituyendo el primer antecedente del trabajo de investigación del autor en esta temática.

Capítulo 2

Evaluación de los métodos de filtrado inverso de la voz

Un obstáculo importante que se tiene al desarrollar nuevos métodos de filtrado inverso de la voz radica en cómo evaluar y validar su desempeño. Este obstáculo adquiere mayor relevancia especialmente en señales naturales, ya que no disponen de una señal de flujo glótico de referencia; en estos casos, no es posible medir de forma directa y objetiva la calidad de la estimación obtenida con filtrado inverso [48]. Como alternativa, podría compararse la estimación con el flujo glótico obtenido a partir del flujo de aire medido a nivel de los labios utilizando una máscara de Rothenberg [10]. Sin embargo, en la actualidad no existen bases de datos públicas que proporcionen registros simultáneos de la señal de voz y de la máscara de Rothenberg, que permitan llevar a cabo la comparación.

A la falta de señales de flujo glótico de referencia, se suma la dificultad de no contar con un consenso claro dentro de la comunidad científica acerca de cómo evaluar los resultados obtenidos en señales naturales [21]. En algunos estudios, la evaluación se limita a presentar las estimaciones del flujo glótico y verificar visualmente su consistencia respecto a una forma de onda teórica esperada [49]. En otros casos, la presencia de una fase cerrada plana en la estimación del flujo glótico se utiliza como un indicador de desempeño [26], debido a que esta región del ciclo glótico es particularmente sensible a distorsiones generadas por errores en el ajuste del filtro del tracto vocal o en el proceso de integración de la función glótica.

Como alternativa, es posible cuantificar el desempeño de los métodos de filtrado inverso utilizando señales sintetizadas a partir de modelos de la fonación, lo que permite disponer tanto de la señal de voz como del flujo glótico de referencia. De esta forma, el desempeño de los métodos puede evaluarse midiendo, por ejemplo, el error entre la forma de onda del flujo glótico de referencia y su estimación.

Si bien este enfoque es ampliamente utilizado en la literatura [12], [21], debe señalarse que la evaluación puede verse sesgada por el tipo de modelo de la fonación empleado. En particular, si se utilizan modelos lineales basados en la teoría *fuente–filtro*, es esperable que los métodos de filtrado inverso se desempeñen favorablemente, dado que se sustentan en la misma formulación teórica. Una alternativa más desafiante consiste en emplear señales generadas mediante modelos inspirados en la fisiología de la fonación, los cuales consideran con mayor fidelidad los fenómenos físicos subyacentes, incluyendo las interacciones no lineales que ocurren durante el proceso de fonación [50], [51].

A continuación se describen brevemente la base de datos y las métricas utilizadas para evaluar el desempeño de los métodos de filtrado inverso de la voz desarrollados en esta tesis.

2.1. Base de datos Openglot

Openglot es una base de datos libre diseñada específicamente para la evaluación de nuevos métodos de filtrado inverso de la voz [26]. Las señales que la componen se encuentran organizadas en cuatro repositorios (numerados del I al IV). Los tres primeros repositorios contienen señales sintéticas generadas a partir de diferentes modelos de la fonación, los cuales disponen tanto de la señal de voz como del flujo glótico de referencia. El último repositorio proporciona señales naturales de voz de hablantes masculinos y femeninos; es importante destacar que, en estos casos, no se cuenta con las correspondientes señales de flujo glótico.

A continuación se brindan mayores precisiones de cada repositorio.

Repositorio I

El primer repositorio contiene señales generadas a partir de un modelo lineal de la fonación basado en la teoría *fuente–filtro*. En este caso, el flujo glótico se generó utilizando el modelo Liljencrants–Fant de la función glótica [52], mientras que el tracto vocal se modeló mediante un filtro digital autorregresivo de orden 9. La señal de voz se obtuvo filtrando la función glótica con el filtro del tracto vocal. Este repositorio incluye señales correspondientes a cinco vocales (/a/, /e/, /i/, /o/, /u/) y cuatro calidades de fonación, con frecuencias fundamentales comprendidas entre 100 y 360 Hz.

Repositorio II

El segundo repositorio ofrece un conjunto de señales sintetizadas a partir de un modelo no lineal de la fonación basado en la teoría *mioelástica, aerodinámica y acústica*, que simula la interacción entre las cuerdas vocales y el flujo glótico, la propagación acústica a lo largo del tracto vocal y la radiación sonora en los labios [50]. Este modelo reproduce la cinemática de la vibración de las cuerdas vocales y la propagación del sonido a través del tracto vocal configurado para las vocales /a/, /i/ y /u/. Se incluyen simulaciones para voces masculinas y femeninas, variando la frecuencia fundamental (82–220 Hz en hombres y 175–294 Hz en mujeres) y el grado de aducción de las cuerdas vocales.

Repositorio III

El tercer repositorio está compuesto por señales medidas en un sistema físico diseñado para emular la fonación humana. Este sistema emplea tractos vocales plásticos impresos en 3D a partir de imágenes de resonancia magnética de hablantes reales [53]. Estos tractos corresponden a configuraciones para vocales /a/, /e/, /i/ e /u/, y fueron excitados mediante una fuente acústica que emite pulsos sonoros cuya morfología sigue el modelo Liljencrants–Fant, para distintas frecuencias fundamentales en el rango entre los 100 y 500 Hz. Las señales de voz generadas con el sistema se registraron con micrófono en una cámara anecoica.

Repositorio IV

El cuarto repositorio incluye grabaciones de voces naturales de cinco hablantes masculinos y cinco femeninos, quienes produjeron vocales sostenidas con dos calidades de fonación (modal y susurrada). Para cada emisión vocal se registraron simultáneamente tres señales: la señal de voz mediante micrófono, el electroglotograma y el video de alta velocidad de las cuerdas vocales.

Repositorios seleccionados

El repositorio I constituye un punto de partida básico para evaluar los métodos de filtrado inverso de la voz, dado que permite realizar la evaluación bajo condiciones controladas y sin efectos no lineales. Sin embargo, estas condiciones idealizadas lo convierten un caso demasiado simple, por lo que no fue considerado en esta tesis.

Asimismo, el repositorio III también fue descartado debido a que utiliza una fuente acústica en lugar de un flujo de aire para excitar el tracto vocal, introduciendo fenómenos físicos que no ocurren naturalmente durante la fonación humana.

Las señales sintéticas del repositorio II incorporan efectos no lineales inherentes al proceso de fonación, como el acoplamiento aerodinámico, acústico y tisular a nivel glótico, ofreciendo un escenario de evaluación más desafiante para los métodos de filtrado inverso de la voz. Por esta razón, se seleccionaron estas señales para evaluar los métodos desarrollados en esta tesis. También se emplearon las señales del repositorio IV para evaluar los métodos bajo condiciones de habla natural, donde es posible observar fenómenos característicos de la producción de la voz, como variaciones de tono y amplitud, así como la presencia de ruido y artefactos introducidos durante el registro de la señal.

2.2. Medidas de desempeño

Error de forma de onda

Una primera forma de evaluar un método de filtrado inverso consiste en cuantificar el error de forma de onda de las señales estimadas. Este tipo de evaluación sólo es posible en señales sintéticas para las cuales se dispone de la señal glotal de referencia. En este trabajo se consideraron dos métricas distintas, dependiendo de la señal evaluada:

- **Error de estimación de la función glótica**: Para la función glótica, resultante de cancelar la contribución del tracto vocal de la señal de voz, se empleó el error de forma de onda absoluto normalizado:

$$E_{v_g} = \frac{m_e}{\mathrm{RMS}(v_g)}, \tag{2.1}$$

donde m_e denota el valor de la mediana del error de forma de onda absoluto producto de comparar la función glótica de referencia, v_g, y su estimación, $\hat{v}_g$, definido como:

$$e_{v_g}[n] = \left|v_g[n] - \hat{v}_g[n]\right|, \text{ para } n = 1,2,\ldots,N. \tag{2.2}$$

El error en (2.1) está normalizado por el valor RMS[1] de la función glótica de referencia. Las señales en (2.2) se alinean temporalmente para compensar cualquier retardo debido a la propagación de la onda acústica a lo largo del tracto vocal o al filtro inverso aplicado.

Previo a calcular el error de forma de onda, la estimación $\hat{v}_g$ en (2.2) debe normalizarse, dado que no es posible recuperar su amplitud original tras aplicar el filtrado inverso [12]. Una forma de realizar esta normalización consiste en proyectar ortogonalmente $\hat{v}_g$ respecto de la función glótica de referencia, tal como se propone en [42]:

$$\hat{v}_g(n) = \frac{\sum_{i=1}^{N} v_g(i)\tilde{v}_g(i)}{\sum_{i=1}^{N} \tilde{v}_g^2(i)} \tilde{v}_g(n), \text{ para } n = 1,...,N, \tag{2.3}$$

donde $\tilde{v}_g$ es la estimación de la función glótica sin normalizar.

- **Error de estimación del flujo glótico:** Para el flujo glótico, obtenido del procesamiento de la función glótica, se empleó el error cuadrático medio entre las formas de onda de referencia y estimada, u_g y $\hat{u}_g$, respectivamente:

$$E_{u_g} = \frac{1}{N} \sum_{n=1}^{N} (u_g(n) - \hat{u}_g(n))^2, \tag{2.4}$$

donde ambas señales se encuentran normalizadas respecto a su valor máximo.

Es importante justificar el empleo del error absoluto como métrica para evaluar la estimación de la función glótica, en lugar del clásico error cuadrático medio. La razón principal es que la función glótica presenta cambios abruptos en los GCI, tal como se muestra en las Figs. 1.4 y 1.5. Estos cambios generan picos de gran amplitud en el error al comparar la forma de onda teórica con su estimación. Dado que el error cuadrático medio penaliza con mayor peso a los errores localizados de gran amplitud, esta métrica resulta poco adecuada para evaluar la calidad global de la forma de onda estimada. En contraste, el error absoluto ofrece una medida más representativa de la calidad general de la estimación.

[1] Sigla correspondiente a la expresión en inglés: root mean squared.

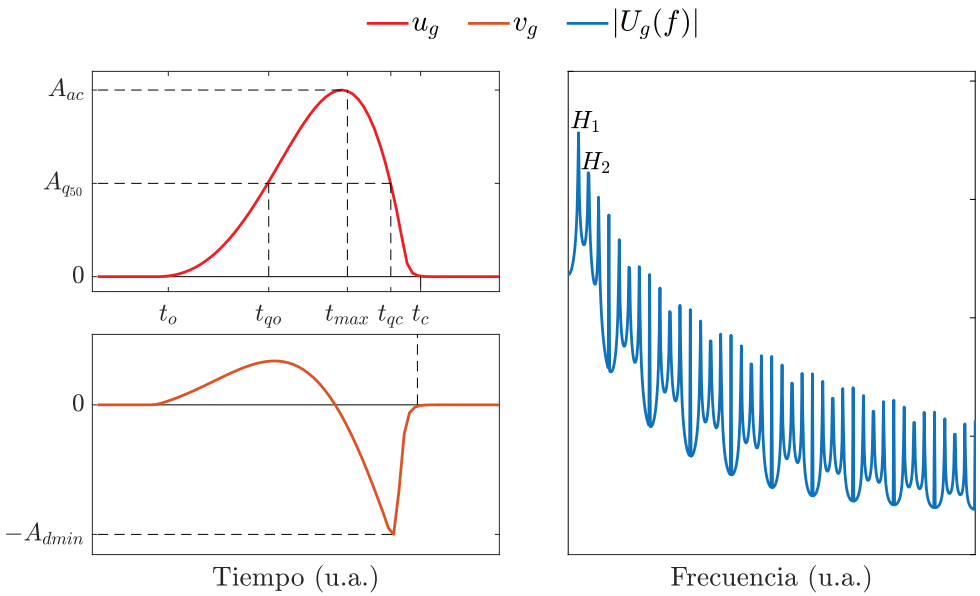

FIGURA 2.1: Instantes de tiempo y amplitudes utilizados en el cálculo de los parámetros aerodinámicos. Panel izquierdo: formas de onda del flujo glótico (superior) y la función glótica (inferior). Panel derecho: espectro de magnitud en escala logarítmica del flujo glótico.

Parámetros aerodinámicos

Otra forma de evaluar las estimaciones es mediante parámetros aerodinámicos que caracterizan el flujo glótico y su derivada. Estos parámetros se han utilizado ampliamente para analizar las relaciones entre las emisiones vocales y la función vocal, el género, la edad, el tono, los trastornos de la voz y la prosodia [21], [54]-[57].

Los parámetros aerodinámicos se obtienen a partir de mediciones sobre las formas de onda glotales en el dominio temporal y sobre el espectro de magnitud del flujo glótico en el dominio frecuencial (ver Fig. 2.1) [58]. A continuación se describen los parámetros considerados en esta tesis:

- **Cociente de amplitud normalizado (NAQ):** mide la relación entre el valor máximo del flujo glótico (A_{ac}) y el valor absoluto del mínimo de la función glótica A_{dmin}, normalizado por el período fundamental T_0:

$$NAQ = \frac{A_{ac}/A_{dmin}}{T_0}. \tag{2.5}$$

- **Cociente de apertura (OQ):** mide la proporción del tiempo correspondiente a la fase abierta dentro de un ciclo glótico:

$$OQ = \frac{t_c - t_o}{T_0}, \tag{2.6}$$

donde t_c es el instante del primer cruce por cero de la función glótica tras alcanzar su mínimo, y t_o es el instante en que el flujo glótico supera el 10 % de su valor máximo.

- **Cociente de cuasi apertura (QOQ):** mide la proporción del ciclo glótico en la cual el flujo glótico supera el 50 % de su valor máximo:

$$QOQ = \frac{t_{qc} - t_{qo}}{T_0}, \tag{2.7}$$

donde t_{qo} y t_{qc} son los instantes en los que el flujo glótico alcanza la mitad de su valor máximo durante la fase de apertura y cierre, respectivamente.

- **Cociente de cierre (ClQ)**: mide la proporción del ciclo correspondiente al cierre glótico, desde que el flujo alcanza su valor máximo hasta el instante t_c:

$$ClQ = \frac{t_c - t_{max}}{T_0}, \tag{2.8}$$

donde t_{max} es el instante en el que el flujo glótico alcanza su máximo.

- **Diferencia de los dos primeros armónicos (H1H2):** se define como la diferencia, medida en decibelios, entre las amplitudes de los dos primeros armónicos del espectro de magnitud del flujo glótico:

$$H1H2 = H_1 - H_2\,(\mathrm{dB}). \tag{2.9}$$

Norma l_1 en fase cerrada

Como se mencionó al inicio del capítulo, la forma de onda del flujo glótico en la fase cerrada constituye un indicador relevante de la calidad del filtrado inverso, ya que se espera un flujo prácticamente nulo debido al cierre casi completo de la glotis. En este contexto, el valor de la norma l_1 en la fase cerrada se utiliza para cuantificar la calidad de las estimaciones [39], [45]. Valores elevados de esta norma indican una distorsión

significativa en la forma de onda estimada del flujo glótico, mientras que valores bajos sugieren una estimación más precisa.

2.3. Comentarios de fin de capítulo

En este capítulo se presentaron los repositorios de señales de la fonación y las métricas empleadas para la evaluación de los métodos desarrollados en esta tesis. En primer lugar, se describió la base de datos Openglot, la cual incluye señales que permiten realizar evaluaciones tanto en condiciones controladas (señales sintéticas) como en escenarios de habla real. Asimismo, se detallaron las distintas medidas de desempeño utilizadas para cuantificar la precisión de las estimaciones obtenidas.

En los capítulos siguientes se presentarán los aportes realizados en esta tesis y se discutirán los resultados obtenidos, todos ellos basados en las señales y métricas descritas en el presente capítulo.

Capítulo 3

Aportes a la predicción lineal con atenuación Gaussiana

Tal como se describió en la Sec. 1.3.1, el efecto acústico del tracto vocal puede modelarse mediante un filtro autorregresivo como en la Ec. (1.6). La forma tradicional de ajustar dicho filtro a partir de la señal de voz es mediante LP. Este método se basa en un modelo lineal de la fonación, en el cual se considera que la señal de voz $s[n]$, para un instante $1 \leq n \leq N$, sigue un proceso autorregresivo [29]:

$$s[n] = \mathbf{a}^T \mathbf{s}[n] + e[n], \tag{3.1}$$

donde $\mathbf{a} = \left[a_1, a_2, \ldots, a_P\right]^T$ es el vector de coeficientes del filtro, $\mathbf{s}[n] = \left[s[n-1], s[n-2], \ldots, s[n-P]\right]^T$ es un vector que contiene las últimas P muestras de la señal de voz, y $e[n]$ es el error de predicción o residuo.

La manera clásica de estimar el vector $\mathbf{a}$ consiste en minimizar el error cuadrático de predicción [32]:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \sum_{n=1}^{N+P} e^2[n], \text{ s.a. } e[n] = s[n] - \mathbf{a}^T \mathbf{s}[n]. \tag{3.2}$$

Uno de los principales inconvenientes del método de predicción lineal radica precisamente en el uso del error cuadrático como función de costo. Tal como se discutió en la Sec. 1.3.1, esta función es muy sensible a valores atípicos de gran amplitud en el error de predicción [33].

En el contexto de la fonación sonora, dichos errores ocurren principalmente alrededor de los GCIs, tal como se muestra en la Fig. 3.1. Allí se puede apreciar que, en torno a cada GCI, el modelo LP de la Ec. (3.1) no logra ajustarse adecuadamente, generando errores de predicción con gran amplitud. Además, en las cercanías a estos instantes la

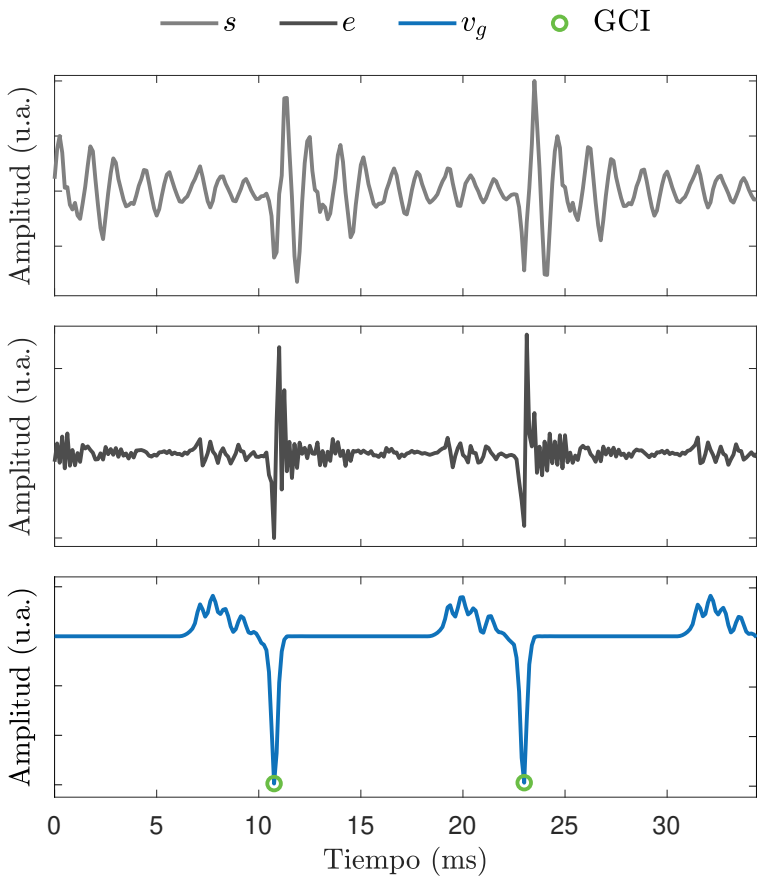

FIGURA 3.1: Ajuste del modelo LP (3.1) en el contexto de la fonación sonora. Superior: señal de voz sintética. Centro: error de predicción. Inferior: función glótica de referencia junto con la ubicación de los GCIs.

señal de voz alcanza su máxima amplitud, mientras que la función glótica presenta su valor mínimo en cada ciclo.

Como consecuencia, durante la minimización del error cuadrático de predicción, estos errores localizados y de gran amplitud tienen un impacto mayor en la función de costo. Puede interpretarse que el ajuste de los coeficientes se concentra más en reducir la amplitud de estos errores localizados, reduciendo la influencia del resto de las muestras de la señal [34], [41]. En consecuencia, los coeficientes $\hat{\mathbf{a}}$ obtenidos mediante (3.2) no codifican de manera adecuada y general la información del tracto vocal [34].

3.1. Predicción lineal con atenuación Gaussiana

Como se discutió en la Sec. 1.5.1, la predicción lineal con atenuación Gaussiana se presenta como una alternativa superadora al enfoque clásico. Este método utiliza

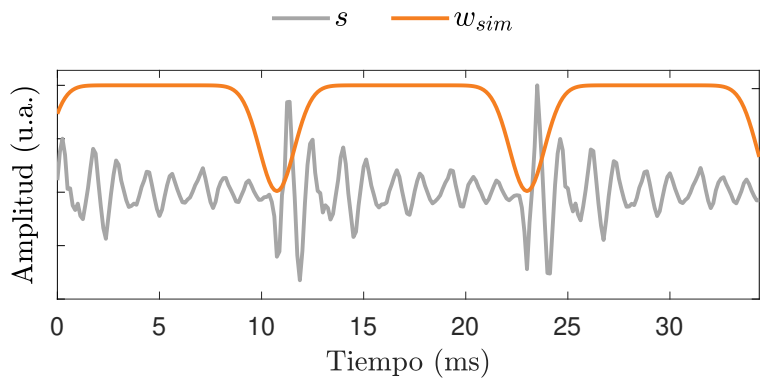

FIGURA 3.2: Ilustración de la función de atenuación Gaussiana del método GLP aplicada a una señal de voz.

la función de atenuación Gaussiana w_{sim} definida en la Ec. (1.9), para controlar el efecto de las muestras de la señal de voz con gran amplitud que se encuentran ubicadas alrededor de los GCIs, tal como se ilustra en la Fig.3.2.

Al emplear la función de ponderación w_{sim}, los coeficientes del filtro del tracto vocal pueden estimarse minimizando una forma ponderada del error cuadrático de predicción [40], [41]:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \sum_{n=1}^{N+P} w_{\text{sim}}[n]e^2[n], \text{ s.a. } e[n] = s[n] - \mathbf{a}^T\mathbf{s}[n], \tag{3.3}$$

cuya solución analítica viene dada por [40], [41], [59]:

$$\hat{\mathbf{a}} = \left(\sum_{n=1}^{N+P} w_{\text{sim}}[n]\mathbf{s}[n]\mathbf{s}^T[n]\right)^{-1}\left(\sum_{n=1}^{N+P} w_{\text{sim}}[n]s[n]\mathbf{s}[n]\right). \tag{3.4}$$

La atenuación Gaussiana en (3.3) mejora el ajuste del filtro al reducir el impacto de las muestras de gran amplitud ubicadas alrededor de los GCIs. Como resultado, los coeficientes $\hat{\mathbf{a}}$ obtenidos mediante (3.4) codifican de manera más precisa la información del tracto vocal presente en la señal de voz [2], [3].

En la práctica, el método GLP suele emplearse para el filtrado inverso de distintas señales de voz utilizando un mismo conjunto de parámetros κ y σ_1 para definir la función de atenuación Gaussiana en la Ec. (1.9) [42]. Sin embargo, esto no es recomendable, ya que una configuración de κ y σ_1 puede resultar adecuada para algunas señales e inadecuada para otras, especialmente cuando las voces presentan grandes diferencias en sus frecuencias fundamentales o en la duración de sus fases cerrada y abierta del

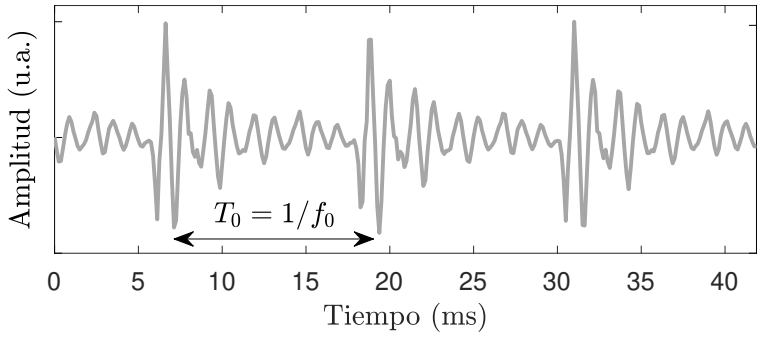

FIGURA 3.3: Periodicidad presente en una señal de voz sonora correspondiente a una vocal /a/.

ciclo glótico.

Por otro lado, el método GLP aplica una función de ponderación que atenúa simétricamente a ambos lados de los GCIs, afectando información relevante de la señal de voz. La evidencia actual indica que las muestras ubicadas en la fase cerrada son cruciales para el correcto ajuste del filtro del tracto vocal [36], [39], [40], [1], mientras que las muestras correspondientes a la fase abierta contribuyen negativamente a dicho ajuste [13], [36], [40].

3.2. Aportes

Esta sección tiene como objetivo resumir los principales aportes del primer artículo del autor de esta tesis [3], incluido en el Anexo A. Dichos aportes se enfocan en el desarrollo de dos estrategias destinadas a mejorar el desempeño del método GLP para el filtrado inverso de la voz.

3.2.1. GLP adaptada a la periodicidad de la voz

Dado que la fonación sonora está determinada por las oscilaciones cuasiperiódicas de las cuerdas vocales, este aspecto debe ser considerado en el método GLP. La periodicidad de la señal de voz está definida por su período fundamental T_0 o, equivalentemente, por su frecuencia fundamental $f_0 = 1/T_0$, tal como se muestra en la Fig. 3.3.

Para señales discretas, como las analizadas en esta tesis, la periodicidad está determinada por el número de muestras por ciclo glótico, N_0. En consecuencia, la distancia entre GCIs consecutivos en la señal de voz es aproximadamente igual a N_0.

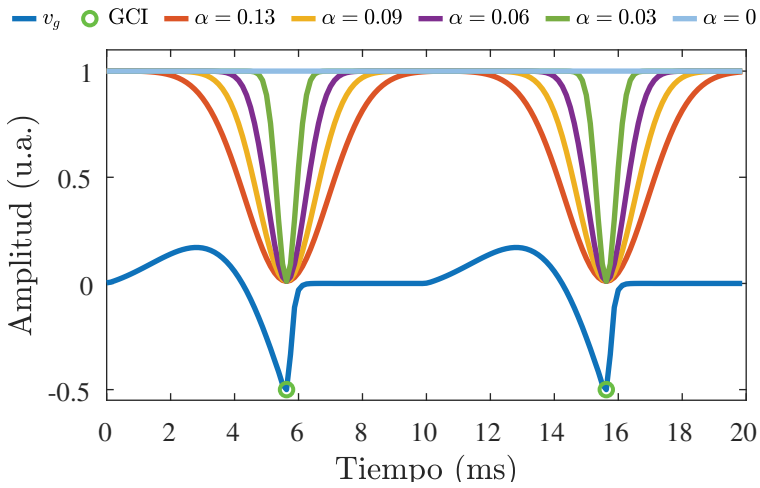

FIGURA 3.4: Efecto de la variación del parámetro α sobre el tamaño de la región de atenuación aplicada por la función de atenuación Gaussiana w_{sim}, en comparación con una función glótica de referencia y la ubicación de los GCIs.

Un primer aporte orientado a mejorar el método GLP consistió en adaptar el parámetro σ_1 de la función de atenuación Gaussiana en base al valor de N_0 de cada señal [2], [3]:

$$\sigma_1 = \alpha N_0, \; 0 \leq \alpha \leq \alpha_{\text{max}}, \tag{3.5}$$

donde el límite superior α_{max} se determina considerando la superposición máxima permitida entre dos ventanas Gaussianas consecutivas.

La Fig. 3.4 muestra ejemplos de funciones de atenuación Gaussianas w_{sim} con distintos valores de σ_1, obtenidos al variar α, en comparación con una función glótica de referencia v_g. Como es de esperarse, valores mayores de α generan regiones de atenuación más amplias alrededor de cada GCI. Nótese además que la parametrización propuesta en la Ec. (3.5) es tal que $\alpha = 0$ equivale a aplicar una ponderación constante, recuperando así el método LP clásico (ver Fig. 1.6.a).

Resultados

Se investigó la influencia de los parámetros κ y α de la función de atenuación Gaussiana en el desempeño del método GLP en el contexto del filtrado inverso, considerando segmentos de 50 ms de señales de voz sintéticas del repositorio II. La Fig. 3.5 muestra el valor promedio del error de forma de onda de la función glótica E_{v_g} para distintos valores de α y niveles de atenuación κ, considerando diferentes órdenes P del filtro del tracto vocal.

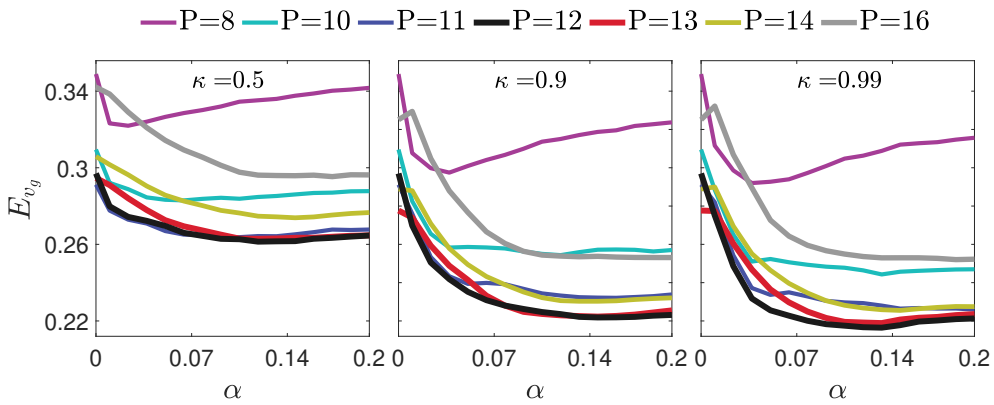

FIGURA 3.5: Efectos de la variación de los parámetros de la función de atenuación Gaussiana en el desempeño del método GLP para filtrado inverso. Se ilustra el error de forma de onda E_{v_g} promedio de la función glótica para las señales del repositorio II al variar α, κ y el orden P del filtro del tracto vocal.

Como se esperaba, atenuar las muestras alrededor de los GCIs mediante w_{sim} con $\alpha > 0$ conduce a una mejora en las estimaciones de filtrado inverso (menor error E_{v_g}) respecto al uso de una ponderación constante ($\alpha = 0$).

Asimismo, se observa que el error disminuye al incrementar el orden del filtro de $P = 8$ a $P = 12$; sin embargo, para $P = 14$ y $P = 16$ el error aumenta significativamente en comparación con $P = 12$. Los resultados muestran también que incrementar κ desde 0.5 hasta 0.99 reduce el error E_{v_g}, mientras que valores superiores a 0.99 no ofrecen mejoras perceptibles, por lo que no fueron incluidos en la figura.

En general, todas las curvas de error presentan un mínimo local dentro de un rango específico de α. En particular, la curva correspondiente a $P = 12$ y $\kappa = 0{,}99$ exhibe el menor error en el intervalo $0{,}1 < \alpha < 0{,}15$. Para $\alpha > 0{,}15$, el error vuelve a aumentar, lo cual se atribuye a que valores elevados de σ_1, obtenidos de la Ec. (3.5) al aumentar α, amplían demasiado la región de atenuación alrededor de los GCIs, atenuando información relevante de la señal de voz en la fase cerrada [39].

3.2.2. GLP con atenuación asimétrica

En el contexto del filtrado inverso de la voz, la función w_{sim} presenta la desventaja de atenuar simétricamente las muestras ubicadas alrededor de los GCIs, sin distinguir entre la información proveniente de las fases abierta y cerrada del ciclo glótico.

Sin embargo, como se discutió al final de la Sec. 3.1, la estimación de los coeficientes LP basada principalmente en la información de la fase cerrada tiende a producir un

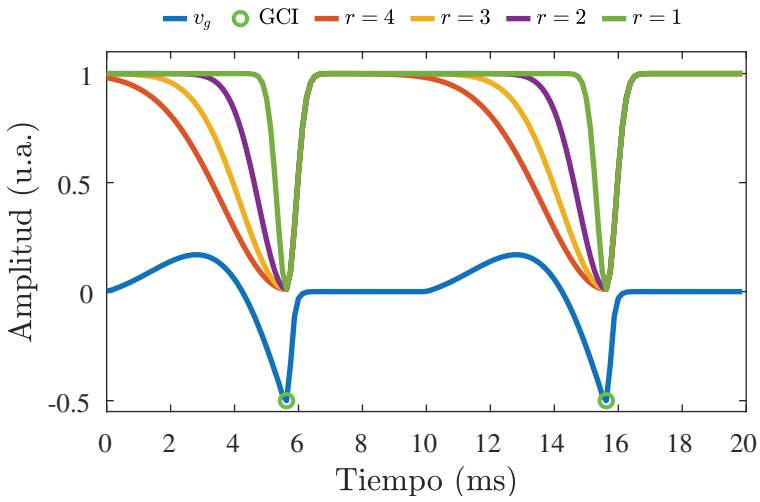

FIGURA 3.6: Ejemplos de funciones de atenuación Gaussiana asimétrica para diferentes valores de r (considerando $\sigma_1 = 0{,}03\,N_0$), en comparación con una función glótica de referencia y la ubicación de los GCI.

ajuste más preciso del filtro del tracto vocal [36], [39], [40], [45]. Por el contrario, las muestras correspondientes a la fase abierta y a los GCIs suelen contribuir negativamente [35].

Para superar esta limitación, se propuso una versión asimétrica de la función de atenuación Gaussiana, diseñada para atenuar de forma diferenciada las fases cerrada y abierta de cada ciclo glótico [3]. Esta función se construyó a partir de una serie de ventanas Gaussianas asimétricas centradas en cada GCI:

$$w_{\text{asim}}[n] = 1 - \kappa \sum_{l=1}^{L} g_a[n - n_l], \tag{3.6}$$

donde $g_a[n - n_l]$ es una versión asimétrica de la ventana Gaussiana, definida como [60]:

$$g_a[n - n_l] = \begin{cases} e^{-(n-n_l)^2/2\sigma_1^2}, & \text{si } n \geq n_l, \\ e^{-(n-n_l)^2/2\sigma_2^2}, & \text{si } n < n_l. \end{cases} \tag{3.7}$$

El parámetro σ_1 determina el alcance de la atenuación aplicada por w_{asim} sobre la fase cerrada, mientras que σ_2 controla la atenuación aplicada en la fase abierta.

Nuestra propuesta establece que el parámetro σ_1 se fija en base a la Ec. (3.5), mientras que σ_2 se define como una proporción de σ_1 [3], es decir:

$$\sigma_2 = r\sigma_1 = r(\alpha N_0),\ 1 \leq r \leq r_{\max}, \tag{3.8}$$

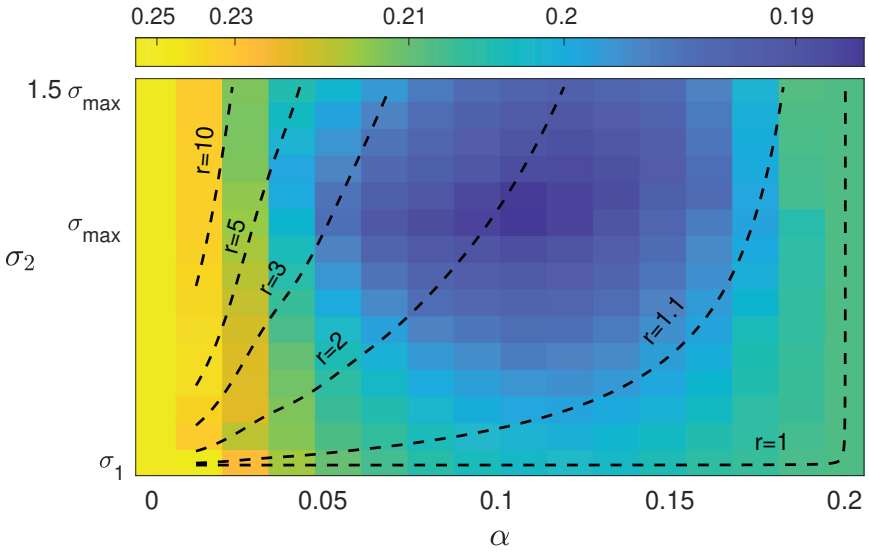

FIGURA 3.7: Mapa del error promedio de forma de onda de la función glótica v_g estimada a partir de las señales del repositorio II para diferentes valores de α y σ_2. Se aplicó un mapeo de color no lineal para facilitar la interpretación visual. Para fines de interpretación se dibujan superpuestas las isolíneas del parámetro de asimetría r.

donde el parámetro r controla el grado de asimetría, y $r_{\max} = \alpha_{\max}/\alpha$. El caso $r = 1$ permite recuperar la atenuación simétrica ($w_{\text{asim}} = w_{\text{sim}}$), mientras que $r = r_{\max}$ genera la máxima región de atenuación permitida sobre la fase abierta sin exceder la superposición máxima entre ventanas Gaussianas, tal como se explicó en la Sec. 3.2.1.

La Fig. 3.6 muestra ejemplos de funciones de atenuación Gaussiana asimétrica para distintos valores de r, en comparación con una función glótica de referencia v_g. Puede observarse que incrementar r amplía progresivamente la región de atenuación correspondiente a la fase abierta (a la izquierda de cada GCI), sin alterar la atenuación aplicada en la fase cerrada.

Resultados

La Fig. 3.7 presenta un mapa del error promedio de forma de onda de la función glótica E_{v_g} para las señales del repositorio II, obtenido mediante el método GLP con la función de atenuación Gaussiana asimétrica. El mapa se construyó variando α en el rango $[0, \alpha_{\max}]$ y σ_2 entre $[\sigma_1, 1{,}5\sigma_{\max}]$, donde $\sigma_{\max} = \alpha_{\max} N_0$. En las simulaciones se consideraron los parámetros fijos $P = 12$, $\kappa = 0{,}99$ y $\alpha_{\max} = 0{,}2$. Además, se incluyen las isolíneas del parámetro r, que describen la razón de asimetría para w_{asim} en base a la Ec. (3.8).

Como muestra la figura, la atenuación asimétrica produce un menor error E_{v_g} respecto de la versión simétrica (caso $\sigma_2 = \sigma_1$, correspondiente a $r = 1$). En particular, el error mínimo se obtiene para $\alpha \approx 0{,}1$ y $\sigma_2 \approx \sigma_{\max}$, lo que coincide con la isolínea $r \approx 2$. Este valor indica una función de atenuación Gaussiana asimétrica que aplica en la fase abierta una región de atenuación aproximadamente dos veces mayor que en la fase cerrada.

Finalmente, puede observarse que para $\sigma_2 > \sigma_{\max}$ el error E_{v_g} aumenta para todos los valores $\alpha > 0$. Este incremento se debe a la superposición excesiva entre ventanas Gaussianas adyacentes, lo que provoca la pérdida de información relevante para ajustar correctamente el filtro del tracto vocal.

3.2.3. Comparación con otros métodos

Las variantes del método GLP con funciones de atenuación Gaussianas simétrica y asimétrica estudiadas en este capítulo, y a las que nos referiremos como SGLP y AGLP[1], se compararon con los métodos LP clásico, LP ponderada y estabilizada (SWLP) [44], y LP de fase casi cerrada (QCP) [40].

Resultados en señales sintéticas

Para las señales sintéticas del repositorio II, la comparación se efectuó utilizando el error de forma de onda de la función glótica E_{v_g} y los errores en la estimación de los parámetros aerodinámicos definidos en la Sec. 2.2.

Los resultados de la Sec. 5.1 del Anexo A muestran que SGLP y AGLP superan a LP y SWLP en señales que incluyen distintos fonemas vocálicos y diversas frecuencias fundamentales. Además, AGLP presenta el menor error E_{v_g} en todas las categorías en comparación con SGLP, lo cual respalda la hipótesis de que la atenuación Gaussiana asimétrica mejora el desempeño del método GLP respecto de su versión simétrica original. No obstante, ambos métodos GLP mostraron un rendimiento ligeramente inferior al del método QCP.

Tendencias similares se observaron en los errores asociados a los parámetros aerodinámicos, donde QCP obtuvo sistemáticamente el menor error para la mayoría de los parámetros, seguido por los métodos basados en atenuación Gaussiana. En señales con baja frecuencia fundamental, AGLP alcanzó un desempeño comparable al de QCP,

[1] Siglas correspondientes a las expresiones en inglés: symmetric/asymmetric Gaussian linear prediction, respectivamente.

excepto en el parámetro NAQ. Por otro lado, para la vocal /i/, se observó que AGLP supera ligeramente a QCP en los parámetros OQ y ClQ.

Resultados en señales naturales

Para las voces naturales del repositorio IV, se emplearon la norma l_1 en fase cerrada y el parámetro NAQ, para realizar la comparación entre los métodos.

Los resultados presentados en la Sec. 5.2 del Anexo A muestran que AGLP logra una mejora marginal respecto de SGLP y SWLP en términos de la norma l_1 en fase cerrada, mientras que QCP presenta los valores más bajos. En cuanto al parámetro NAQ, no se observaron diferencias relevantes entre los métodos.

Finalmente, la inspección visual de las formas de onda glotales estimadas (Fig. 9 del Anexo A) indica que las estimaciones obtenidas mediante QCP, AGLP y SGLP presentan una fase cerrada más plana que aquellas obtenidas con SWLP y LP.

3.3. Comentarios de fin de capítulo

En este capítulo se estudiaron dos variantes del método de predicción lineal con atenuación Gaussiana en el contexto del filtrado inverso de la voz. Estas variantes, inspiradas en la fisiología de la fonación, se sustentan en la evidencia de que atenuar las muestras de la señal de voz alrededor de los instantes de cierre glótico mejora el ajuste del filtro del tracto vocal y los resultados de filtrado inverso. Asimismo, la atenuación asimétrica propuesta permite enfatizar la información contenida en la fase cerrada, lo que mejora aún más la precisión del filtro y, en consecuencia, el filtrado inverso.

Con un desempeño cercano al de QCP, los métodos basados en atenuación Gaussiana representan una alternativa atractiva debido a su simplicidad. Vale la pena destacar que una limitación inherente a estos métodos es la necesidad de contar con la ubicación de los instantes de cierre glótico; sin embargo, existen numerosos algoritmos capaces de obtenerlos a partir de la señal de voz o de señales asociadas a la fonación [13]. Además, la robustez del método GLP frente a errores en la determinación de los GCIs [42] hace que las variantes propuestas sean especialmente valiosas para el análisis señales de fonaciones atípicas o alteradas, en las cuales la detección precisa de los GCIs puede resultar difícil.

Cabe destacar que la versión del método GLP adaptada a la periodicidad de la voz fue presentada en un congreso nacional [2]. Asimismo, la versión asimétrica de la

función de atenuación Gaussiana y sus resultados, dieron lugar a una publicación en la revista *Speech Communication* de la editorial Elsevier [3].

Capítulo 4

Filtrado inverso de la voz basado en el criterio de máxima correntropía

En el contexto del filtrado inverso, la mayoría de las estrategias de predicción lineal ponderadas han sido desarrolladas con el objetivo de mitigar la principal limitación del método LP clásico. Específicamente, la gran sensibilidad del error cuadrático como función de costo ante errores de predicción atípicos de gran amplitud, los cuales son ocasionados por la fuente de excitación acústica durante la producción de la voz.

El estudio de funciones de costo alternativas, más adecuadas para el tipo de error de predicción que se genera en el análisis de la señal de voz, ha recibido escasa atención por parte de la comunidad científica.

Algunos autores han propuesto utilizar la norma l_1 como función de costo en el esquema de predicción lineal. Esto permite aprovechar la mayor robustez de la norma l_1 frente a valores atípicos durante el cálculo de los coeficientes de predicción lineal [31], [34]. Otros enfoques alternativos han empleado la distancia de Itakura-Saito como función de costo para ajustar los coeficientes en el dominio de la frecuencia [61].

Los resultados presentados en [31], [34], [61] muestran la importancia de buscar funciones de costo más adecuadas al contexto de la predicción lineal de la señal de voz. Considerar funciones de costo alternativas puede conducir al desarrollo de nuevas estrategias de estimación de los coeficientes del filtro del tracto vocal que permitan mejorar los resultados de filtrado inverso.

En esta línea, la correntropía, una medida no lineal y robusta de similitud entre variables aleatorias, ha demostrado ser una función de costo adecuada para modelos lineales con errores no Gaussianos y valores atípicos de gran amplitud [62]. Este aspecto le otorga un gran potencial en el contexto de la predicción lineal aplicada a la señal de voz. Si bien la correntropía ha sido empleada, por ejemplo, para la estimación de la frecuencia fundamental de la voz [63], el realce de señales de habla [64] y el

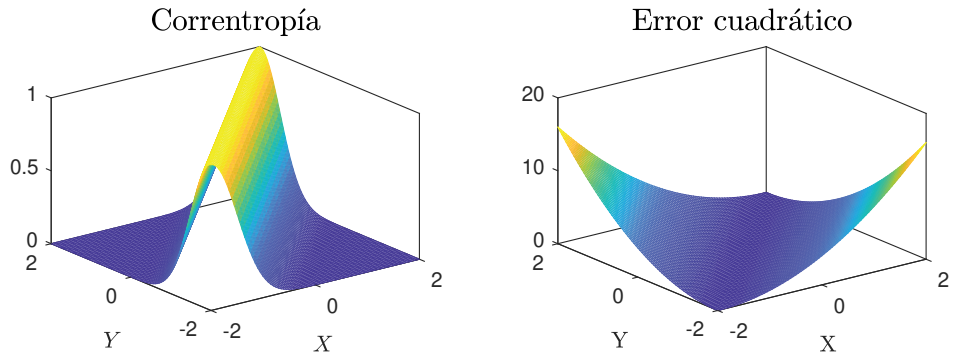

FIGURA 4.1: Superficies de error de las variables X e Y medidas mediante distintas funciones de costo. Izquierda: correntropía con núcleo Gaussiano. Derecha: error cuadrático.

reconocimiento de hablantes [65], no existen antecedentes de su uso en esquemas de predicción lineal y filtrado inverso.

A continuación brindaremos más detalles sobre como se define la correntropía y discutiremos sus principales propiedades.

4.1. Correntropía con núcleo Gaussiano

La correntropía con núcleo Gaussiano entre dos variables aleatorias escalares arbitrarias X e Y se define como [66]:

$$V(X,Y) = \mathcal{E}\{G_\sigma(X-Y)\}, \tag{4.1}$$

donde $\mathcal{E}\{\cdot\}$ es el operador esperanza y G_σ es el núcleo Guassiano con varianza σ^2, definido como:

$$G_\sigma(X-Y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X-Y)^2}{2\sigma^2}\right). \tag{4.2}$$

La correntropía (4.1) mide la similitud entre las variables aleatorias en una vecindad de su espacio conjunto, determinada por el parámetro σ del núcleo Gaussiano [62].

El panel izquierdo de la Fig. 4.1 ilustra la superficie de error correspondiente a evaluar la correntropía entre X e Y. Como puede observarse, la correntropía alcanza su valor máximo cuando ambas variables coinciden, mientras que, ante diferencias grandes, su valor se vuelve muy pequeño. En contraste, el error cuadrático presenta un comportamiento opuesto. Como se aprecia en el panel derecho de la Fig. 4.1, esta función crece considerablemente ante diferencias significativas entre X e Y.

Esta diferencia en el tratamiento de los errores de gran amplitud hace que la correntropía sea una función de costo robusta y más adecuada que el clásico error cuadrático [67].

Además de su robustez, la correntropía posee una serie de propiedades relevantes [66], [68], [69]. A continuación se presentan las más importantes:

- *Propiedad 1:* La correntropía es definida positiva y acotada, es decir, $0 < V(X,Y) \leq 1/(\sqrt{2\pi}\sigma)$. Además, alcanza su valor máximo si $X = Y$.

- *Propiedad 2:* La correntropía involucra todos los momentos estadísticos pares de la diferencia entre X e Y:

$$V(X,Y) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n n!} \mathcal{E}\left\{ \frac{(X-Y)^{2n}}{\sigma^{2n}} \right\}. \tag{4.3}$$

  A diferencia del error cuadrático medio, que coincide con el momento estadístico de segundo orden, la correntropía incluye momentos de orden superior. Nótese que, a medida que σ aumenta, los momentos de mayor orden decaen rápidamente, y el momento de segundo orden se vuelve predominante en la correntropía, aproximándose así al comportamiento del error cuadrático medio.

- *Propiedad 3:* Sea una muestra de pares de datos i.i.d. $\{(x_i,y_i)\}_{i\in N}$ extraída de la función de densidad de probabilidad conjunta $f_{X,Y}(x,y)$. Se define además, el estimador de Parzen, con un parámetro de núcleo σ, de la función de densidad de probabilidad de las muestras de error $e_i = x_i - y_i$ como $\hat{f}_\sigma(e)$. Entonces, el valor de $\hat{f}_\sigma(e)$ evaluado en $e = 0$ coincide con el valor de $\widehat{V}(X,Y)$ [62], donde

$$\widehat{V}(X,Y) = \frac{1}{N} \sum_{i=1}^{N} G_\sigma(x_i - y_i), \tag{4.4}$$

  es una aproximación de la correntropía utilizando un estimador para N muestras del operador esperanza [70].

  Dado que $\hat{f}_\sigma(0)$ estima la probabilidad de que X e Y sean iguales, es decir, $\hat{f}_\sigma(0) \approx p(X = Y)$, entonces maximizar la correntropía equivale a aumentar la probabilidad de que las muestras coincidan [69].

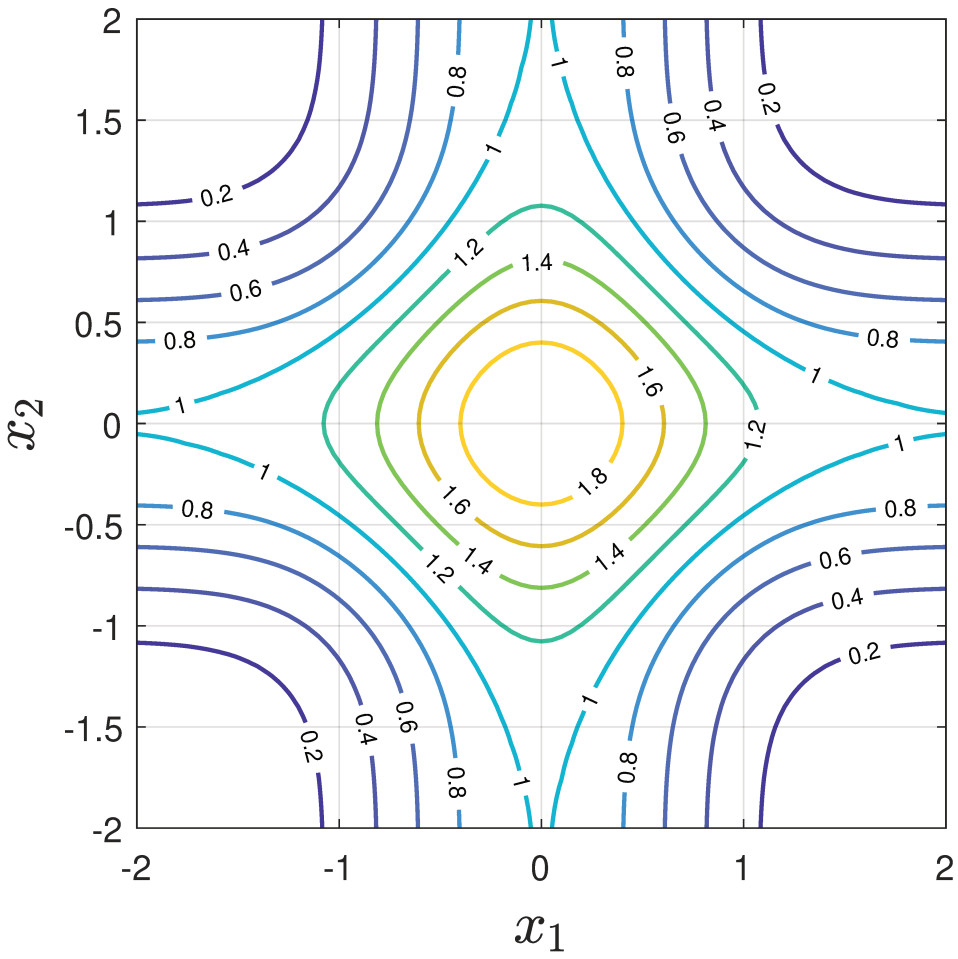

FIGURA 4.2: Curvas de nivel de $V(\mathbf{x},\mathbf{0})$ en el espacio muestral bidimensional. El parámetro del núcleo Gaussiano esta fijado en $\sigma = 0{,}6$.

- *Propiedad 4:* Si se expresa la correntropía entre X e Y como una función de costo:

$$J = V(X,Y), \tag{4.5}$$

entonces J resulta una función cóncava en el rango de $[-\sigma,\sigma]$ (como muestra la figura.4.1). Esta concavidad garantiza la existencia y unicidad de una solución óptima en problemas que buscan maximizar la correntropía [68].

- *Propiedad 5:* La correntropía, como estimador muestral, induce una métrica con características geométricas importantes.

Sea el vector $\mathbf{x} = [x_1, x_2]^T$ y vector origen $\mathbf{0} = [0,0]^T$, ambos definidos en un espacio bidimensional. La correntropía $V(\mathbf{x},\mathbf{0})$ define una métrica en dicho espacio. La Fig. 4.2 muestra curvas de nivel de la métrica definida. Cuando x_1 y

x_2 están cerca del origen, la correntropía se comporta como la norma L_2 (zona euclidiana caracterizada por contornos circulares). A medida que aumenta la diferencia entre los componentes de $\mathbf{x}$ y el origen, la correntropía se aproxima a la norma L_1 (zona de transición caracterizada por contornos en forma de diamante). Finalmente, ante diferencias muy grandes, la correntropía se comporta como la norma L_0, volviéndose robusta (zona de rectificación) [69]. El parámetro σ controla el tamaño de cada zona [62]; aumentar σ amplía la zona euclidiana y reduce la zona de rectificación, mientras que disminuirlo produce el efecto opuesto.

Todas estas propiedades hacen que la correntropía sea una función de costo robusta e idónea para problemas de estimación de parámetros [68], [71].

4.2. Aportes:

A continuación, se presenta el desarrollo de un nuevo método de predicción lineal que utiliza la correntropía como función de costo, el cual propusimos especialmente para el filtrado inverso de la voz. Además, se discuten los principales resultados obtenidos. Hasta donde sabemos, nuestro artículo [5], incluido en el Anexo B, constituye el primer antecedente del uso de la correntropía para esta aplicación.

4.2.1. Predicción lineal basada en el criterio de máxima correntropía

En el modelo LP de la Ec. (3.1), el objetivo es determinar los coeficientes $\mathbf{a}$ que reduzcan el error de predicción $e[n] = s[n] - \mathbf{a}^T\mathbf{s}[n]$. Para ello, se plantea un problema de optimización en el cual se busca minimizar (o maximizar) una función de costo que dependa de $e[n]$.

Empleando como función de costo en el esquema LP a la correntropía con núcleo Gaussiano (propiedad 4) [4], [5], se obtiene:

$$\begin{aligned} J &= \mathcal{E}\left\{G_\sigma(e[n])\right\}, \\ &= \mathcal{E}\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(s[n]-\mathbf{a}^T\mathbf{s}[n]\right)^2}{2\sigma^2}\right)\right\}. \end{aligned} \tag{4.6}$$

El objetivo entonces será determinar los coeficientes que maximicen (4.6) (propiedad 3):

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} \mathcal{E}\left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(s[n] - \mathbf{a}^T \mathbf{s}[n]\right)^2}{2\sigma^2} \right) \right\}. \tag{4.7}$$

A este método para estimar los coeficientes $\hat{\mathbf{a}}$ lo denominamos *predicción lineal basada en el criterio de máxima correntropía* (MCLP[1]).

Cálculo de los coeficientes MCLP

Para obtener una expresión que permita calcular los coeficientes $\hat{\mathbf{a}}$ a partir de la Ec. (4.7), se buscan las condiciones bajo las cuales se satisface $\frac{\partial J}{\partial \mathbf{a}} = 0$, lo que conduce a:

$$\mathcal{E}\left\{ \exp\left(-\frac{\left(s[n] - \mathbf{a}^T \mathbf{s}[n]\right)^2}{2\sigma^2} \right) \left(s[n] - \mathbf{a}^T \mathbf{s}[n]\right) \mathbf{s}[n] \right\} = 0. \tag{4.8}$$

Considerando que el error de predicción es un proceso estocástico estacionario en tiempo discreto, es posible reemplazar el operador esperanza de (4.8) por su estimador para N muestras [70]:

$$\frac{1}{N} \sum_{n=1}^{N} h_e[n] \left(s[n] - \mathbf{a}^T \mathbf{s}[n]\right) \mathbf{s}[n] = 0, \tag{4.9}$$

donde $h_e[n]$ puede interpretarse como una función de ponderación positiva:

$$h_e[n] = \exp\left(-\frac{\left(s[n] - \mathbf{a}^T \mathbf{s}[n]\right)^2}{2\sigma^2} \right). \tag{4.10}$$

Luego, manipulando algebraicamente la ecuación (4.9), se obtiene:

$$\sum_{n=1}^{N} h_e[n] \mathbf{s}[n] \mathbf{s}[n]^T \mathbf{a} = \sum_{n=1}^{N} h_e[n] s[n] \mathbf{s}[n]. \tag{4.11}$$

De esta última expresión se pueden despejar los coeficientes, obteniendo:

$$\begin{aligned} \mathbf{a} &= \left[\sum_{n=1}^{N} h_e[n] \mathbf{s}[n] \mathbf{s}[n]^T \right]^{-1} \left[\sum_{n=1}^{N} h_e[n] s[n] \mathbf{s}[n] \right], \\ &= \mathbf{R}_h^{-1} \mathbf{r}_h, \end{aligned} \tag{4.12}$$

[1] Sigla correspondiente a la expresión en inglés: maximum correntropy linear prediction.

donde $\mathbf{r}_h$ y $\mathbf{R}_h$ representan estimaciones ponderadas del vector de correlación y la matriz de autocorrelación, respectivamente [70].

La expresión (4.12) es similar a la solución de *Wiener–Hopf* [29]; sin embargo, se destacan dos diferencias claves [67]:

- Se aplica la función de ponderación $h_e[n]$, que depende de σ y del error de predicción. Esta función enfatiza los errores pequeños en relación con σ, mientras que atenúa aquellos de gran amplitud (propiedad 5). Por lo tanto, el cálculo de $\mathbf{a}$ mediante (4.12) es inherentemente robusto frente a errores atípicos.

- La ecuación (4.12) no posee una solución en forma cerrada, como la presentada oportunamente en la Ec. (3.4), debido a que la función de ponderación $h_e[n]$ depende de los coeficientes $\mathbf{a}$ según (4.10).

Para abordar el problema del cálculo de los coeficientes, desarrollamos una solución iterativa de punto fijo similar a la propuesta en [67]:

$$\mathbf{a}_{k+1} = [\mathbf{R}_h(\mathbf{a}_k)]^{-1}\,\mathbf{r}_h(\mathbf{a}_k), \quad \text{para } k = 0,1,2,\ldots, \tag{4.13}$$

donde, dada una estimación inicial $\mathbf{a}_0$, se obtiene una sucesión de estimaciones de los coeficientes aplicando (4.13). La convergencia de esta solución iterativa está garantizada debido a que la correntropía es una función cóncava (propiedad 4). Basado en esta solución de punto fijo, se desarrolló el Algoritmo 1; los coeficientes resultantes, luego de varias iteraciones, se emplean para el filtrado inverso de la voz.

Un aspecto destacable del Algoritmo 1 es que realiza una actualización del parámetro σ del núcleo Gaussiano basada en la regla de Silverman [72]:

$$\sigma_S = 1{,}06\,\sigma_e\,N^{-1/5}, \tag{4.14}$$

donde σ_e representa el mínimo valor entre la desviación estándar y el rango intercuartil (escalado por 1.34) del error de predicción obtenido en cada iteración al emplear los coeficientes calculados con (4.13).

Por otro lado, se debe resaltar que la actualización de σ no se realiza en cada iteración del Algoritmo 1, sino que se realiza de manera escalonada. Para ello se emplea el umbral ϵ_1, que establece cuando dos soluciones consecutivas de los coeficientes resultan muy similares.

Algoritmo 1: Cálculo de los coeficientes **a** basado en MCLP.

Inicializar: $\mathbf{a}_0$, σ, ϵ_1, ϵ_2
Preénfasis y normalización de $s[n]$
$k = 0$
do
  Calcular : $h_e[n]$, $\mathbf{R}_h$, $\mathbf{r}_h$
  $\mathbf{a}_{k+1} = \mathbf{R}_h^{-1}\mathbf{r}_h$
  $e_{k+1}[n] = s[n] - \mathbf{a}_{k+1}^T\mathbf{s}[n]$
  if $\left(||\mathbf{a}_{k+1} - \mathbf{a}_k||_2^2 < \epsilon_1\right)$
    Calcular: σ_S
  end
  $k = k + 1$
while $\left(||\mathbf{a}_{k+1} - \mathbf{a}_k||_2^2 > \epsilon_2\right)$
return $(\mathbf{a}_{k+1})$

El algoritmo finaliza cuando una nueva actualización de σ no produce cambios significativos entre dos soluciones consecutivas, lo cual depende del umbral ϵ_2, elegido de modo que sea mucho menor que ϵ_1.

Resultados: ajuste iterativo de MCLP

La Fig. 4.3 ilustra el ajuste iterativo llevado a cabo por el método MCLP, aplicando el Algoritmo 1, para una señal de voz sintética del repositorio II. Las columnas de la Fig. 4.3 representan distintas iteraciones k, y en cada una de ellas se indica el valor del parámetro σ calculado mediante la Ec. (4.14).

La primera fila de la figura muestra la señal de voz s y la función de ponderación h_e obtenida en cada iteración. La segunda fila presenta el error de predicción e. Finalmente, en la tercera fila se comparan la forma de onda teórica de la función glótica v_g con su correspondiente estimación $\hat{v}_g$, obtenida mediante filtrado inverso con el filtro del tracto vocal calculado en cada iteración.

Como se aprecia en la primera fila de la Fig. 4.3, el método MCLP propuesto implementa un esquema de predicción lineal ponderado e iterativo que tiende a enfatizar la información de la señal de voz contenida en la fase cerrada. Esta ponderación se realiza completamente guiada por los datos mediante la función h_e.

En la segunda fila se puede apreciar una disminución localizada en la amplitud del error de predicción a lo largo de las iteraciones, especialmente en los segmentos correspondientes a la fase cerrada, lo cual hace que el error se vuelva más ralo (es decir, con un menor número de elementos distintos de cero [73]).

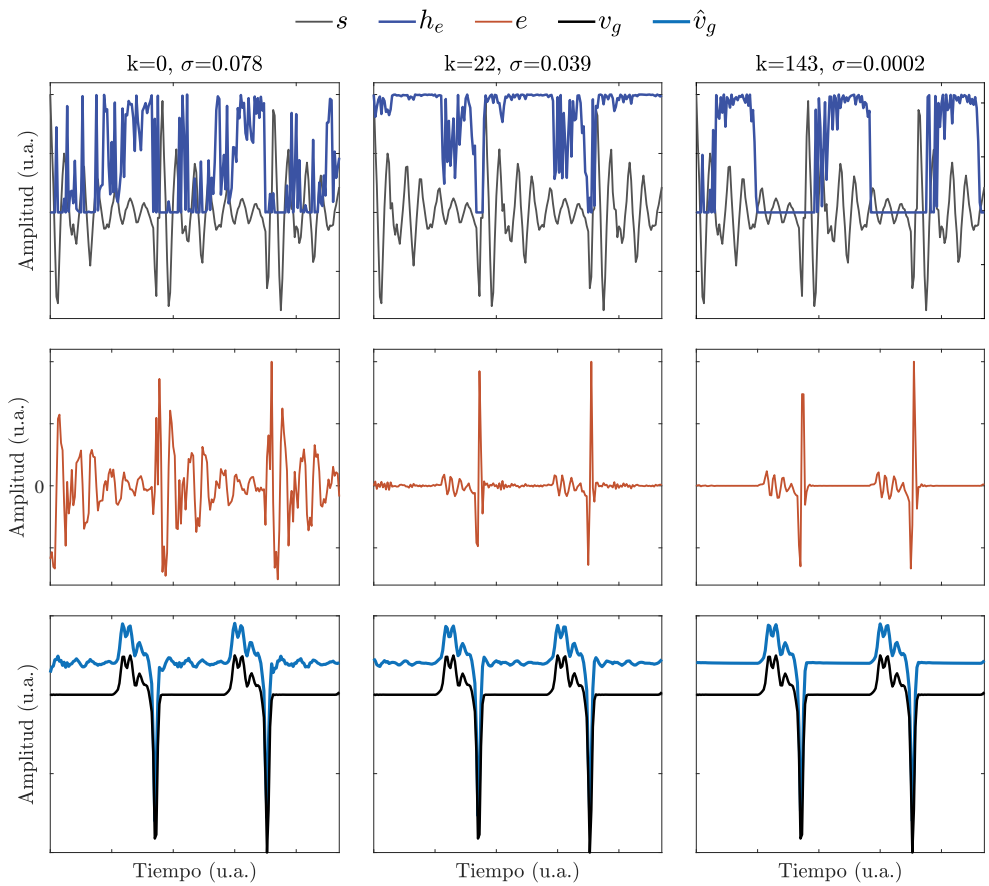

FIGURA 4.3: Ajuste iterativo de MCLP para una señal sintética. Primera fila: señal de voz s y función de ponderación h_e. Segunda fila: error de predicción e. Tercera fila: función glótica teórica v_g y su estimación de filtrado inverso $\hat{v}_g$ (con un desplazamiento vertical aplicado para facilitar la comparación visual). En cada columna se informa el número de iteración k y el valor de σ.

Estos cambios en el error de predicción resultan de la actualización iterativa de los coeficientes **a**, lo cual produce modificaciones en el valor del parámetro σ y en la forma de onda de la función de ponderación h_e que toma valores máximos en las muestras donde $e[n] \approx 0$ y valores mínimos si $|e[n]| > \sigma$.

A medida que el error de predicción se vuelve más ralo durante la fase cerrada, el valor de σ tiende a disminuir y $h_e[n]$ centra la ponderación en las muestras ubicadas alrededor de dicha fase, atenuando simultáneamente aquellas pertenecientes a la fase abierta, donde se cumple que el error de predicción es distinto de cero. Además, la actualización escalonada de σ en el Algoritmo 1 contribuye a reforzar este comportamiento de h_e.

La ponderación progresiva de h_e sobre las muestras de la señal de voz en la fase

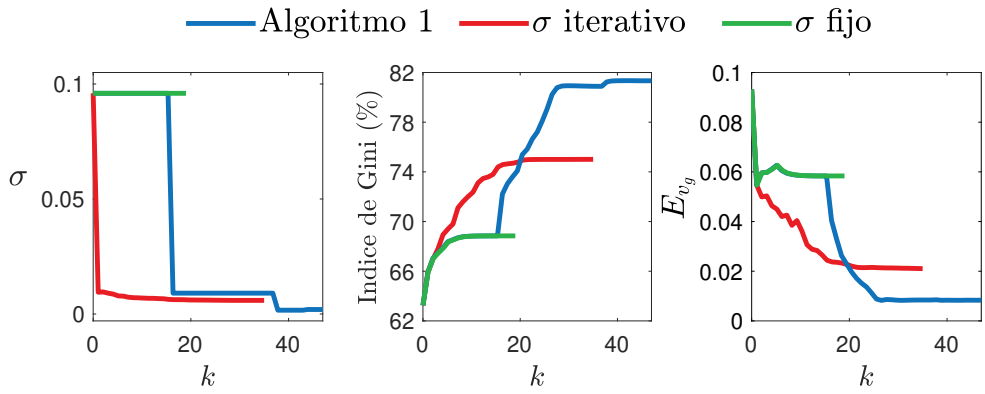

FIGURA 4.4: Análisis del desempeño del método MCLP para tres estrategias de actualización de σ aplicadas a una señal sintética. Izquierda: valores de σ en cada iteración. Centro: índice de Gini del error de predicción. Derecha: error de forma de onda de la función glótica E_{v_g}.

cerrada produce un mejor ajuste de los coeficientes del filtro del tracto vocal en cada iteración. Como consecuencia, se obtienen estimaciones más precisas de la función glótica mediante filtrado inverso, tal como se observa en la tercera fila de la figura.

Resultados: efectos de la actualización del núcleo

Se buscó mostrar que la actualización de σ influye notablemente en el desempeño para el filtrado inverso. Se exploraron tres estrategias para actualizar este parámetro, tomando como base la regla de Silverman (4.14) en todos los casos:

- Inicializar σ y mantenerlo fijo.
- Actualizar σ en cada iteración.
- Actualizar σ de manera escalonada, como se propone en el Algoritmo 1.

La Fig. 4.4 muestra el efecto en el desempeño del método MCLP al aplicar estas tres estrategias a la misma señal sintética utilizada en la Fig. 4.3. En la figura se ilustra, para cada iteración, el valor de σ (panel izquierdo), el índice de Gini del error de predicción (panel central) y el error de forma de onda de la función glótica estimada E_{v_g} (panel derecho). El índice de Gini se consideró debido a que es una medida que cuantifica qué tan ralo es el error de predicción, de manera tal que un valor alto de este índice indica un error más ralo.

Cuantificar la raleza del error de predicción es relevante ya que un error de predicción con numerosos elementos nulos favorece la ponderación realizada por h_e en la señal de voz, especialmente cuando dichos elementos se concentran alrededor de

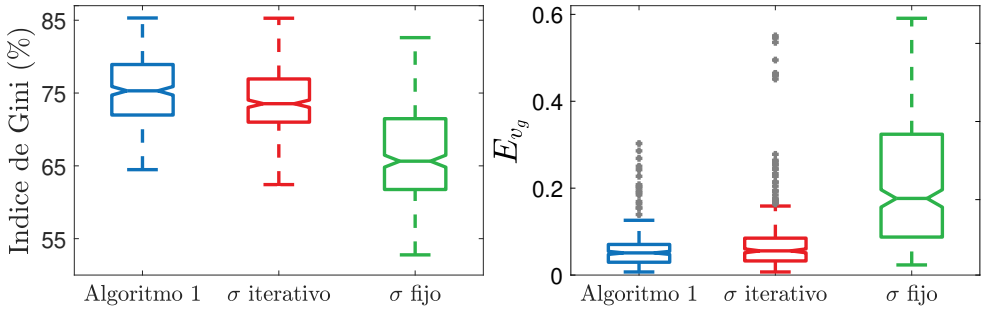

FIGURA 4.5: Diagramas de caja para tres estrategias de actualización de σ, obtenidos a partir de las señales de voz del repositorio II. Izquierda: índice de Gini del error de predicción. Derecha: error de forma de onda de la función glótica E_{v_g}.

la fase cerrada, tal como se observó en la Fig. 4.3. Además, como se ha mencionado en los capítulos anteriores, promover este tipo de ponderación de fase cerrada mejora sustancialmente el ajuste del filtro del tracto vocal y, por consiguiente, los resultados del filtrado inverso.

Los resultados indican que para σ fijo, el cálculo de los coeficientes con MCLP requiere pocas iteraciones (convergencia más rápida); sin embargo, esta estrategia obtiene el índice de Gini más bajo y el mayor error E_{v_g}. En contraste, actualizar σ en cada iteración produce una disminución pronunciada en su valor, acompañada de un aumento en el índice de Gini y una reducción en E_{v_g}. Sin embargo, esta estrategia requiere un número mayor de iteraciones que el caso anterior. No obstante, se observa que los mejores resultados en términos del índice de Gini y del error E_{v_g} se obtienen mediante la actualización escalonada de σ utilizada en el Algoritmo 1.

Para realizar un análisis más general del desempeño, las tres estrategias se aplicaron a todas las señales de voz del repositorio II. En cada caso se empleó la última estimación de los coeficientes para calcular el índice de Gini del error de predicción y el error E_{v_g}. La Fig. 4.5 presenta diagramas de caja del índice de Gini (panel izquierdo) y del error E_{v_g} (panel derecho) para las tres estrategias consideradas. Los resultados muestran que la estrategia con σ fijo produce el índice de Gini más bajo y el mayor error E_{v_g}. En cambio, entre las estrategias que actualizan σ, la actualización escalonada del Algoritmo 1 obtiene el índice de Gini más alto y los errores E_{v_g} más bajos.

4.2.2. Comparación con otros métodos

El método MCLP se comparó con otro enfoque guiado por los datos, como es la predicción lineal con ponderación probabilística (PWLP) [39]. Asimismo, se incluyó

en la comparación a el método QCP, que mostró el mejor desempeño en el estudio presentado en el Cap. 3, junto con una versión mejorada del mismo que incorpora una compensación de la inclinación espectral del filtro del tracto vocal, denominada en adelante como QCP-ST. Para todos estos métodos, el análisis de filtrado inverso se efectuó en segmentos de 50 ms no superpuestos de señales de voz.

Resultados en señales sintéticas

Para las señales sintéticas del repositorio II, la comparación entre los métodos se llevó a cabo considerando el error de forma de onda de la función glótica E_{v_g}, la norma l_1 en la fase cerrada y el error en los parámetros aerodinámicos.

Los resultados de la Sec. 5.1 del Anexo B mostraron que los métodos guiados por los datos, como MCLP y PWLP, presentaron los mejores desempeños en términos de E_{v_g} y de la norma l_1, en contraste con QCP y QCP-ST. Sin embargo, para la mayoría de las categorías evaluadas no se hallaron diferencias estadísticamente significativas entre MCLP y PWLP.

Por otra parte, en cuanto al error en los parámetros aerodinámicos, tanto MCLP como PWLP mostraron un desempeño superior al de los métodos basados en QCP. Se observó además que MCLP presentó un desempeño ligeramente inferior a PWLP en la estimación del parámetro NAQ, especialmente en señales con frecuencia fundamental alta.

Adicionalmente, se estudió la carga computacional del método MCLP. Para ello, se midió el tiempo requerido para calcular los coeficientes del filtro del tracto vocal a partir de cada segmento de señal de voz de 50 ms del repositorio II.

La Tabla 4.1 informa el primer cuartil (Q_1), la mediana (Q_2) y el tercer cuartil (Q_3) de los tiempos de cómputo (medidos en segundos) para los métodos QCP, QCP-ST, PWLP y MCLP. Los resultados muestran que MCLP requiere, en promedio, 60 veces menos tiempo computacional que PWLP. Por otro lado, el método QCP es el método que requiere menor tiempo de cómputo, lo cual era esperable debido a que emplea una solución no iterativa para calcular los coeficientes del filtro del tracto vocal. En cambio, QCP-ST demanda tiempo adicional debido al procesamiento asociado a la compensación de la inclinación espectral del filtro.

TABLA 4.1: Tiempo de cómputo requerido para calcular los coeficientes del filtro del tracto vocal para todas las señales de voz del repositorio II. El tiempo se reporta en segundos para el primer cuartil (Q_1), la mediana (Q_2) y el tercer cuartil (Q_3).

	Q_1	Q_2	Q_3
QCP	2,48x10^{-4}	2,71x10^{-4}	2,88x10^{-4}
QCP-ST	7,02x10^{-4}	7,28x10^{-4}	7,68x10^{-4}
MCLP	1,68x10^{-2}	2,27x10^{-2}	3,13x10^{-2}
PWLP	1,29x10^{0}	1,42x10^{0}	1,51x10^{0}

Resultados en señales naturales

Para las señales naturales del repositorio IV, la comparación entre los métodos se realizó considerando la norma l_1 en la fase cerrada y el valor del parámetro NAQ.

Los resultados presentados en la Sec. 5.2 del Anexo B muestran que el método MCLP propuesto obtuvo los valores mínimos de la norma l_1 en la fase cerrada, con diferencias estadísticamente significativas respecto de PWLP y de las variantes de QCP. Por otro lado, en cuanto al parámetro NAQ, no se encontraron diferencias estadísticamente significativas entre los valores estimados con cada uno de los métodos.

Al analizar las formas de onda de las señales glotales obtenidas para estas señales (véase la Fig. 7 del Anexo B), se observa que MCLP proporciona estimaciones con una fase cerrada más plana que el resto de los métodos, lo cual concuerda con los valores de norma l_1 reportados en la Sec. 5.2 de dicho anexo.

4.3. Comentarios de fin de capítulo

En este capítulo se presentó la predicción lineal basada en el criterio de máxima correntropía (MCLP). La incorporación de la correntropía en el esquema de predicción lineal proporciona una solución robusta frente al tipo de error generado por la fuente de excitación acústica involucrada en la producción de la voz.

En base a las simulaciones, se describió como el método MCLP implementa un análisis iterativo de predicción lineal ponderada, guiado por los datos, que tiende a enfatizar automáticamente la información de la señal de voz contenida en la fase cerrada. Esto le confiere una clara ventaja sobre otros métodos de filtrado inverso que requieren conocer la ubicación precisa de los instantes glotales para realizar una ponderación similar, como por ejemplo los estudiados en el Anexo A.

Los análisis presentados incluyeron un estudio detallado del ajuste iterativo del método y de la influencia que tiene el parámetro σ del núcleo de Gaussiano de correntropía en el desempeño de filtrado inverso. En particular, se evidenció que la actualización escalonada propuesta en el Algoritmo 1 favorece el ajuste de los coeficientes del filtro del tracto vocal, reduciendo así los errores de filtrado inverso de la voz.

Los resultados obtenidos dieron lugar a diversas publicaciones científicas. Las ideas preliminares sobre el uso de la correntropía como función de costo para predicción lineal, junto con resultados iniciales, fueron presentadas en un congreso internacional [4], donde el trabajo recibió la distinción al *mejor artículo del evento*. Posteriormente, el algoritmo propuesto y los resultados presentados en este capítulo fueron publicados en la revista *IEEE Transactions on Audio, Speech and Language Processing* [5].

Capítulo 5

Modelado adaptativo no armónico para la estimación del flujo glótico

Tal y como se discutió en la Sec. 1.3.2, el proceso de estimación del flujo glótico a partir de observaciones de su derivada, la función glótica, ha recibido relativamente poca atención por parte de la comunidad científica. La manera convencional de obtener dicha estimación consiste en aplicar el LIF, definido en la Ec. (1.8), a pesar de las distorsiones que este filtro puede introducir en la forma de onda del flujo glótico.

El objetivo de este capítulo es presentar un nuevo método, basado en el análisis tiempo-frecuencia, para estimar el flujo glótico a partir de la función glótica. Los métodos tiempo-frecuencia han recibido creciente atención en aplicaciones que involucran el estudio de señales caracterizadas por patrones oscilatorios complejos y no armónicos [74].

El modelado adaptativo no armónico constituye una herramienta conveniente para analizar señales oscilatorias multicomponentes y variables en el tiempo. Tal como se propone en [46], el modelado ANH busca representar de manera óptima una señal a partir de estimaciones de la amplitud y fase instantáneas de sus componentes. Además, cada uno de los patrones oscilatorios presentes en la señal se describen mediante funciones de forma de onda que permiten caracterizar con mayor precisión a la señal [47], [75].

5.1. Modelado de señales multicomponentes

5.1.1. Modelo adaptativo armónico

La manera tradicional de describir una señal oscilatoria monocomponente consiste en modelarla como una oscilación tipo coseno modulada en amplitud y frecuencia [75]:

$$x[n] = A[n]\cos(2\pi\phi[n]), \text{ con } A[n] > 0, \phi'[n] > 0, \text{ para } 1 \leq n \leq N, \tag{5.1}$$

donde N es la longitud de la señal, $A[n]$ la amplitud instantánea y $\phi[n]$ la fase instantánea. A partir de $\phi[n]$, es posible definir otra magnitud asociada a la modulación en frecuencia como es la frecuencia instantánea: $f[n] = \phi'[n]$ [76].

Por otro lado, las señales multicomponente pueden modelarse como la superposición de señales monocomponentes [77], [78]:

$$\begin{aligned} x[n] &= \sum_{k=1}^{K} x_k[n], \\ &= \sum_{k=1}^{K} A_k[n]\cos(2\pi\phi_k[n]), \end{aligned} \tag{5.2}$$

donde el $k-$ésimo componente posee su propia amplitud $A_k[n]$ y fase $\phi_k[n]$ instantáneas, las cuales satisfacen que $A_k[n] > 0$ y $\phi'_k[n] > 0$, para $1 \leq n \leq N$.

La Ec. (5.2) constituye el modelo adaptativo armónico [79]. Una limitación importante de este modelo es que cada componente se representa mediante un patrón oscilatorio simple, típico de las funciones trigonométricas [75]. Si bien este modelo ha sido aplicado en numerosos problemas [80]-[83], resulta insuficiente para describir señales biomédicas que presentan patrones oscilatorios de mayor complejidad [84], [85].

5.1.2. Estimación de la amplitud y fase instantáneas

Los modelos (5.1) y (5.2) requiere estimaciones de la amplitud y la fase instantáneas para poder describir cada uno de los componentes de la señal de la forma $x_k[n] = A_k[n]\cos(2\pi\phi_k[n])$.

Estos elementos pueden estimarse a partir de una representación tiempo-frecuencia de la señal [85]. Para ello puede utilizarse, por ejemplo, el espectrograma, definido como el módulo al cuadrado de la transformada de Fourier de tiempo corto de la señal

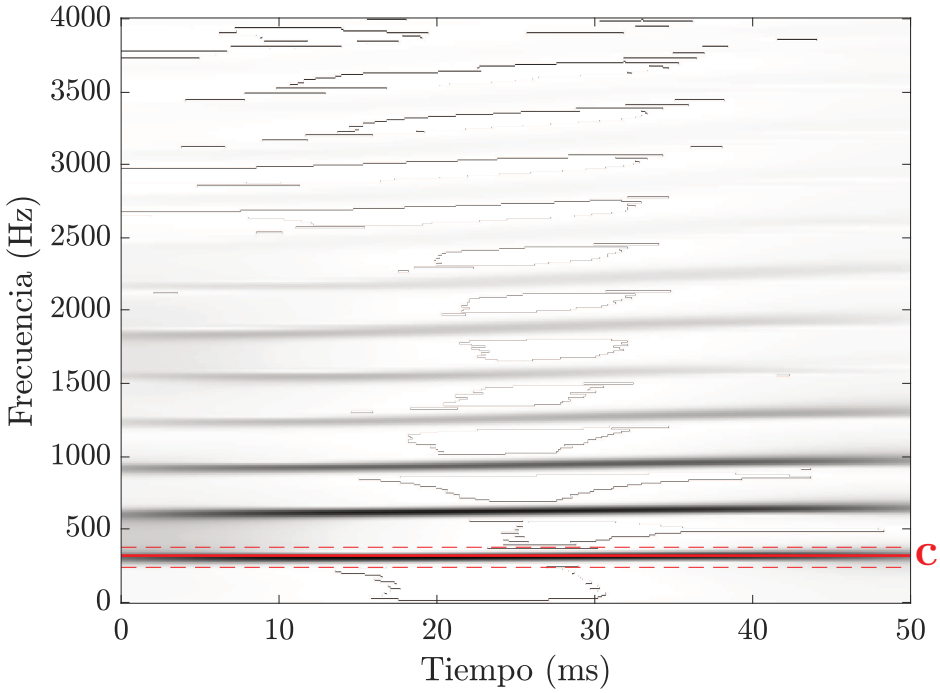

FIGURA 5.1: Espectrograma de una función glótica estimada. Se destaca en color rojo la cresta de mayor energía c.

$x[n]$ para una ventana análisis $h[n]$:

$$\mathbf{F}_x^h[n,m] = \sum_{i=1}^{N} x[i]h[i-n]e^{-j2\pi\frac{m}{M}f_m(i-n)}, \tag{5.3}$$

con $n = 1, \cdots, N$ y $m = 0, \cdots, M-1$, siendo M el numero de intervalos de frecuencia, y f_m la frecuencia de muestreo.

La Fig. 5.1 muestra, a modo de ejemplo, el espectrograma de una función glótica estimada mediante filtrado inverso. Debido a la naturaleza cuasiperiódica de esta señal, su espectrograma presenta una estructura armónica en la que la energía se concentra en un conjunto de crestas, o *ridges*, con trayectorias aproximadamente constantes, las cuales coinciden con los armónicos de la señal.

Para este ejemplo, el armónico fundamental se encuentra en la región de baja frecuencia y corresponde a la cresta de mayor energía, denotada como c y destacada en color rojo en la Fig. 5.1. El resto de los armónicos se encuentran ubicados, aproximadamente, en múltiplos enteros de la frecuencia fundamental.

Es importante resaltar que el ejemplo analizado es representativo de una señal monocomponente. Para una señal multicomponente, su espectrograma presentaría crestas

con trayectorias distintas a las observadas en la Fig. 5.1, debido a las modulaciones en frecuencia propias de cada componente x_k de la señal. Para más ejemplos de este tipo de espectrograma, consulte [86].

Las crestas de mayor energía permiten estimar la amplitud y fase instantáneas de cada componente. Para ello, es necesario emplear un algoritmo de detección de crestas como el descrito en [85]. El objetivo de esta tarea es identificar en el espectrograma un conjunto de crestas c_k que definan las trayectorias asociadas con las frecuencias instantáneas de las componentes de la señal: $c_k[n] \approx \phi_k'[n]$ para $k = 1, 2, \ldots, K$.

Una vez detectadas las crestas c_k, se realiza una reconstrucción vertical utilizando la información de la transformada de Fourier de tiempo corto en una franja de frecuencia de ancho 2Δ, donde Δ es un parámetro elegido por el usuario [86]:

$$y_k[n] = \frac{1}{h[0]} \sum_{|m-c_k[n]|<\Delta} \mathbf{F}_x^h[n,m]. \tag{5.4}$$

Esta reconstrucción alrededor de cada c_k da como resultado una señal $y_k[n]$ asociada al k-ésimo componente a partir de la cual es posible estimar su amplitud y fase instantáneas considerando que [85]:

$$A_k[n] = |y_k[n]|, \qquad \phi_k[n] = \arg\{y_k[n]\}, \tag{5.5}$$

donde $|\cdot|$ y $\arg\{\cdot\}$ denotan el módulo y el argumento de una cantidad compleja, respectivamente.

5.2. Modelo adaptativo no armónico

Con el fin de mejorar el modelo (5.2), H. T. Wu propuso en [46] el modelo adaptativo no armónico para señales multicomponentes:

$$x[n] = \sum_{k=1}^{K} A_k[n] s_k(2\pi\phi_k[n]), \text{ para } 1 \leq n \leq N, \tag{5.6}$$

donde, para cada $k-$ésimo componente, la función coseno es reemplazada por una función forma de onda periódica s_k capaz de describir patrones oscilatorios más generales [85].

La implementación del modelo ANH requiere estimar la amplitud y la fase instantáneas de cada componente, así como las funciones de forma de onda s_k asociadas. Dado que cada s_k es periódica, estas funciones pueden representarse mediante su expansión en serie de Fourier:

$$x[n] = \sum_{k=1}^{K} A_k[n] \sum_{\ell=0}^{r} \left[a_{k,\ell} \cos(2\pi\ell\phi_k[n]) + b_{k,\ell} \sin(2\pi\ell\phi_k[n]) \right], \tag{5.7}$$

donde $a_{k,\ell}$ y $b_{k,\ell}$ son los coeficientes de Fourier de la $k-$ésima función forma de onda, y r es el número máximo de armónicos admisibles según el criterio de Nyquist.

Si se dispone de estimaciones de $A_k[n]$ y $\phi_k[n]$ para cada componente (ver Sec. 5.1.2), el ajuste del modelo ANH se reduce a determinar los coeficientes de Fourier de cada función s_k.

5.2.1. Modelo ANH monocomponente

Sin pérdida de generalidad, a continuación se describe el procedimiento para estimar los coeficientes de Fourier en la Ec. (5.7) para una señal monocomponente ($K = 1$):

$$x[n] = A[n] \sum_{\ell=1}^{r} \left[a_\ell \cos(2\pi\ell\phi[n]) + b_\ell \sin(2\pi\ell\phi[n]) \right]. \tag{5.8}$$

La Ec. (5.8) puede escribirse en forma matricial como:

$$\mathbf{x_r} = \mathbf{C}_r \mathbf{a}, \tag{5.9}$$

donde $\mathbf{x_r} \in \mathbb{R}^N$ representa la aproximación del modelo ANH, $\mathbf{a} = [a_1, \ldots, a_r, b_1, \ldots, b_r]^T \in \mathbb{R}^{2r}$ contiene los coeficientes de Fourier, y $\mathbf{C}_r \in \mathbb{R}^{N\times 2r}$ es un pseudodiccionario de Fourier definido como: $\mathbf{C}_r = [\mathbf{c}_1, \ldots, \mathbf{c}_r, \mathbf{d}_1, \ldots, \mathbf{d}_r]$, con columnas $\mathbf{c}_\ell = A[n] \cos(2\pi\ell\phi[n])$ y $\mathbf{d}_\ell = A[n] \sin(2\pi\ell\phi[n])$, para $\ell = 1,\ldots,r$ y $n = 1,\ldots,N$.

Los coeficientes $\mathbf{a}$ en la Ec. (5.9) pueden obtenerse resolviendo el siguiente problema de mínimos cuadrados:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{x} - \mathbf{C}_r \mathbf{a}\|_2^2, \tag{5.10}$$

donde $\mathbf{x} = [x[1], x[2], \ldots, x[N]]^T \in \mathbb{R}^N$ contiene las muestras de la señal original. Este problema tiene solución analítica [85]:

$$\hat{\mathbf{a}} = (\mathbf{C}_r^T \mathbf{C}_r)^{-1} \mathbf{C}_r^T \mathbf{x}. \tag{5.11}$$

Una vez determinados los coeficientes, la señal reconstruida mediante el modelo ANH se obtiene aplicando la fórmula de síntesis: $\hat{\mathbf{x}}_\mathbf{r} = \mathbf{C}_r\hat{\mathbf{a}}$.

5.3. Aportes

En esta sección se presenta el modelo ANH propuesto para la estimación del flujo glótico a partir de la función glótica, en el contexto del filtrado inverso de la voz. Asimismo, se introduce una versión regularizada del modelo, desarrollada con el fin de mejorar la forma de onda del flujo glótico estimado. Ambos aportes, junto con los resultados discutidos aquí, forman parte de nuestro artículo [7], incluido en el Anexo C.

5.3.1. Modelo ANH para la estimación del flujo glótico

El modelo ANH propuesto para el flujo glótico se formuló considerando dos hipótesis básicas [6], [7]:

- Es monocomponente.
- Posee amplitud constante ($A[n] = 1$ para $1 \leq n \leq N$).

Estas hipótesis se consideran válidas para el flujo glótico cuando se trabaja con ventanas de corta duración, lo cual coincide con el tipo de análisis empleado en el filtrado inverso de señales de voz.

Bajo estos supuestos, y dados unos coeficientes de Fourier $\mathbf{a}$ y un diccionario $\mathbf{C}_r$, el flujo glótico puede explicarse como un modelo ANH monocomponente:

$$\mathbf{u_g} = \mathbf{C}_r\mathbf{a}, \tag{5.12}$$

donde $\mathbf{u_g} = [u_g[1], u_g[2], \ldots, u_g[N]]^T \in \mathbb{R}^N$ contiene las muestras del flujo glótico para cada instante [6].

Desafortunadamente, el problema de optimización (5.10), que permite calcular los coeficientes de Fourier $\mathbf{a}$, no puede resolverse directamente para (5.12), ya que el flujo glótico no es conocido y es precisamente la señal que se desea estimar. No obstante, como ya se ha explicado en los capítulos anteriores, es posible acceder a una estimación de la derivada del flujo glótico, la función glótica, al eliminar la contribución del tracto vocal de la señal de voz, por ejemplo, utilizando los métodos desarrollados en los Caps. 3 y 4.

Por esta razón, en [6] propusimos un modelo auxiliar que permite determinar los coeficientes de Fourier indirectamente a partir de la información contenida en la función glótica. Este modelo se obtiene al considerar la derivada respecto al tiempo de los elementos involucradas en la Ec. (5.12), dando lugar a la ecuación:

$$\mathbf{v_g} = \dot{\mathbf{C}}_r \mathbf{a}, \tag{5.13}$$

donde $\mathbf{v_g} = [v_g[1], v_g[2], \ldots, v_g[N]]^T \in \mathbb{R}^N$ representa el modelo de la función glótica y $\dot{\mathbf{C}}_r = [\dot{\mathbf{c}}_1, \ldots, \dot{\mathbf{c}}_r, \dot{\mathbf{d}}_1, \ldots, \dot{\mathbf{d}}_r]$ es un diccionario cuyas columnas corresponden a las derivadas de las columnas $\mathbf{c}_\ell$ y $\mathbf{d}_\ell$ de $\mathbf{C}_r$. Nótese que los modelos (5.12) y (5.13) comparten los mismos coeficientes $\mathbf{a}$ y el mismo número de armónicos r.

El modelo (5.13) considera además que la función glótica posee la misma fase instantánea $\phi[n]$ que el flujo glótico. Así, la fase puede estimarse a partir de una representación tiempo-frecuencia de la función glótica, tal como se describe en la Sec. 5.1.2. Por otro lado, el número de armónicos r se puede determinar en función de la frecuencia de Nyquist $f_N = f_m/2$ y la frecuencia fundamental f_0 de la señal, en base a la siguiente regla:

$$r = \lfloor f_N / f_0 \rfloor, \tag{5.14}$$

donde $\lfloor \cdot \rfloor$ es un operador que devuelve el entero menor o igual a su argumento.

Del análisis anterior, los coeficientes $\mathbf{a}$ pueden estimarse resolviendo el siguiente problema de mínimos cuadrados:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} ||\tilde{\mathbf{v}}_\mathbf{g} - \dot{\mathbf{C}}_r \mathbf{a}||_2^2, \tag{5.15}$$

donde $\tilde{\mathbf{v}}_\mathbf{g} \in \mathbb{R}^N$ denota la función glótica estimada mediante filtrado inverso. Este problema de optimización posee solución analítica dada por la Ec. (5.11). Una vez estimados los coeficientes $\hat{\mathbf{a}}$, el flujo glótico puede recuperarse mediante la ecuación de síntesis (5.12) [6].

5.3.2. Modelo ANH regularizado

El esquema de optimización de la ecuación (5.15) calcula los coeficientes $\hat{\mathbf{a}}$ que minimizan el error de reconstrucción de la función glótica, sin contemplar ningún aspecto específico en la forma de onda del flujo glótico.

En [45] se demuestra que es conveniente incluir, en el esquema de optimización, un término que minimice la norma l_1 de las muestras del flujo glótico en la fase cerrado. Esto promueve estimaciones con una forma de onda plana y libre de distorsiones en dicha fase. Para ello se emplea una función de ponderación w_{cp} que selecciona las muestras pertenecientes a la fase cerrada, tal como se ilustra en el panel superior de la Fig. 5.2. Esta función de ponderación puede determinarse, por ejemplo, a partir de las ubicaciones de los GCIs y GOIs.

Con el fin de mejorar las estimaciones obtenidas con el modelo (5.12), en [7] desarrollamos una versión regularizada del modelo ANH (RANH[1]). Gracias a la formulación simple del modelo ANH, es posible incorporar un término de regularización en el esquema de optimización, similar al propuesto en [45], con el objetivo de promover una fase cerrada plana en la forma de onda del flujo glótico:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ \underbrace{||\tilde{\mathbf{v}}_{\mathbf{g}} - \dot{\mathbf{C}}_r \mathbf{a}||_2^2}_{\text{Error de reconstrucción}} + \beta \underbrace{||\mathbf{W}\dot{\mathbf{C}}_r \mathbf{a}||_1}_{\text{Regularización}} \right\}, \tag{5.16}$$

donde β controla el peso de la regularización, $\mathbf{W} = \text{diag}(\mathbf{w}_{cp}) \in \mathbb{R}^{N\times N}$, y $\mathbf{w}_{cp} = [w_{cp}[1], w_{cp}[2], \ldots, w_{cp}[N]]^T \in \mathbb{R}^N$ es un vector de ponderación de fase cerrada.

El parámetro β en la Ec. (5.16) establece un equilibrio entre dos aspectos relevantes de las formas de onda glóticas. Por un lado, se garantiza la reconstrucción adecuada de la función glótica $\tilde{\mathbf{v}}_{\mathbf{g}}$; por otro lado, se promueve una forma de onda plana durante la fase cerrada. Si bien esta regularización se aplica sobre la forma de onda reconstruida de la función glótica, también impactará en la forma final del flujo glótico estimado. Nótese que el caso $\beta = 0$ corresponde al esquema sin regularización, es decir, al modelo ANH original.

El panel inferior de la Fig. 5.2 muestra las estimaciones del flujo glótico obtenidas mediante los modelos ANH y RANH (con un desplazamiento vertical aplicado para facilitar la comparación visual). Como puede observarse, ambos modelos producen estimaciones similares durante la fase abierta; sin embargo, el modelo RANH genera una forma de onda más plana en la fase cerrada. Por lo tanto, la regularización de fase cerrada propuesta constituye una incorporación efectiva para mejorar las estimaciones del flujo glótico.

[1] Sigla correspondiente a la expresión en inglés: regularized adaptive non-harmonic.

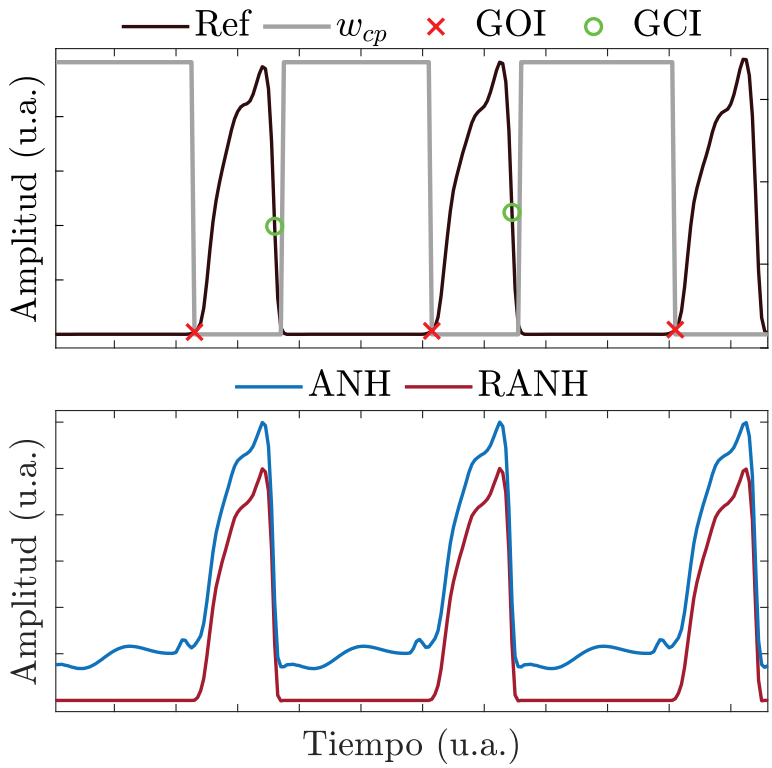

FIGURA 5.2: Estimaciones del flujo glótico obtenidos mediante el modelado no armónico. Superior: ejemplo de una función de ponderación de fase cerrada, w_{cp}, para tres ciclos glóticos de un flujo glótico sintético (indicada como "Ref"). Inferior: estimaciones del flujo glótico obtenidas mediante el modelo ANH y el modelo RANH con $\beta = 1$.

Resultados

Se realizaron simulaciones para evaluar el desempeño de los modelos desarrollados para la estimación del flujo glótico a partir de la función glótica. Para ello se emplearon las señales sintéticas del repositorio II.

Las funciones glóticas se obtuvieron mediante filtrado inverso utilizando un filtro del tracto vocal de orden 12 obtenido con QCP [40], un método bien establecido en la comunidad científica. El análisis se realizó en segmentos no superpuestos de 50 ms de la señal de voz.

El objetivo principal de estas simulaciones fue analizar el impacto del número de armónicos y del peso del término de regularización β en la estimación del flujo glótico. Para ello, se varió el parámetro β en la Ec. (5.16) dentro del rango $[0, 50]$, y se consideraron tres condiciones para el número de armónicos. En la condición base, el valor de r se ajustó individualmente para cada señal en función de su frecuencia fundamental, siguiendo la Ec. (5.14). En la segunda condición, el número de armónicos

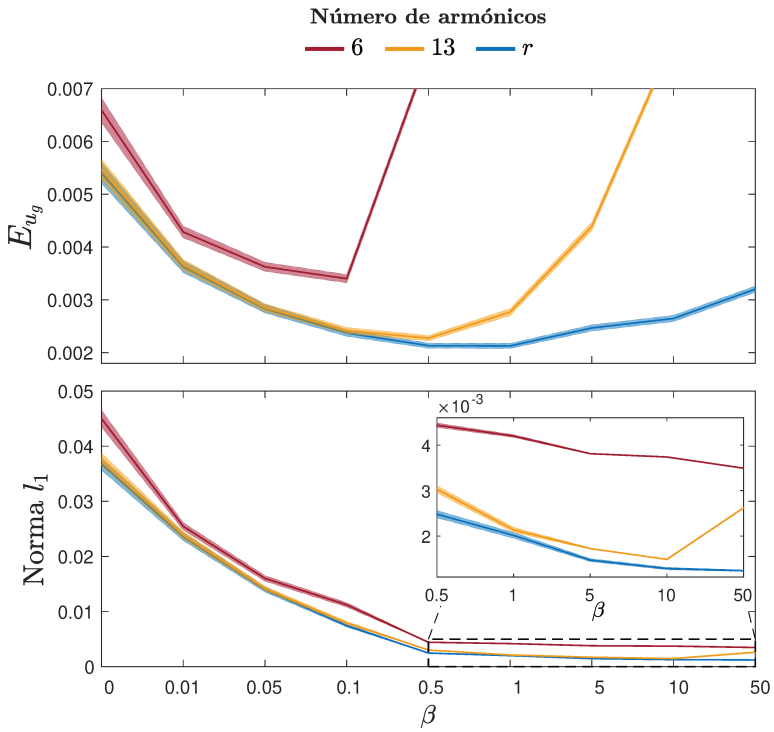

FIGURA 5.3: Error de forma de onda E_{u_g} (superior) y norma l_1 en la fase cerrada (Inferior) en función β para el flujo glótico estimado mediante el modelo RANH. Se analizan tres condiciones para el número de armónicos.

se fijó en 13, correspondiente al menor número de armónicos de la señal con mayor frecuencia fundamental del repositorio (294 Hz). Finalmente, se incluyó una tercera condición con un número fijo de armónicos igual a 6, con el fin de evaluar el desempeño del modelo al considerar un número significativamente limitado de armónicos.

La Fig. 5.3 muestra los valores promedios del error de forma de onda del flujo glótico E_{u_g} y de la norma l_1 en la fase cerrada en función de β, incluyendo intervalos de confianza del 95 %, para las tres condiciones consideradas.

Los resultados en el panel superior indican que la regularización de fase cerrada ($\beta > 0$) reduce significativamente el error de forma de onda E_{u_g} en comparación con el caso sin regularización ($\beta = 0$). A medida que β aumenta, el error disminuye hasta alcanzar un mínimo en el intervalo $0{,}1 < \beta \leq 1$. Luego, para valores $\beta > 1$ el error comienza a incrementarse nuevamente. Este comportamiento en el error E_{u_g} refleja la compensación entre los dos términos de la Ec. (5.16): para valores pequeños de β predomina la reconstrucción de la función glótica, mientras que para valores grandes la regularización de fase cerrada se vuelve dominante, degradando la estimación del flujo glótico, especialmente sobre la fase abierta.

Por otro lado, el panel inferior de la figura muestra que la norma l_1 en fase cerrada disminuye rápidamente para $0 < \beta < 0{,}5$ y se estabiliza para valores mayores, independientemente de la condición del número de armónicos considerada. Esto confirma que la regularización limita eficazmente la amplitud del flujo glótico durante la fase cerrada, cumpliendo el objetivo del modelo RANH. Incrementos adicionales en β no producen mejoras significativas.

Del análisis de la Fig. 5.3 se concluye que ajustar el número de armónicos de acuerdo con la Ec. (5.14) proporciona los mejores resultados tanto en términos de E_{u_g} como de la norma l_1 para $0{,}5 \leq \beta \leq 1$. En contraste, restringir el modelo a un número fijo y reducido de armónicos (6 o 13) disminuye la precisión en la reconstrucción del flujo glótico, aunque el valor de la norma l_1 se ve menos afectado.

5.3.3. Comparación con otros métodos

El modelo RANH con $\beta = 1$ se comparó con el método LIF, considerado el enfoque estándar para obtener el flujo glótico a partir de la función glótica. También se lo contrastó con el método QPR [45], un desarrollo precursor de la estimación del flujo glótico con fase cerrada plana. Adicionalmente, se incluyó el modelo ANH con fines comparativos. Para los modelos ANH y RANH, el número de armónicos se determinó en función de la frecuencia fundamental siguiendo la Ec. (5.14).

Respecto de las métricas de evaluación, para las señales sintéticas del repositorio II se consideraron el error de forma de onda del flujo glótico E_{u_g} y la norma l_1 medida en la fase cerrada. Por otro lado, para las señales naturales del repositorio IV, únicamente se analizó la norma l_1.

Las evaluaciones iniciales con señales sintéticas basadas en el error E_{u_g} (presentadas en la Sec. 5 del Anexo C) mostraron que el modelo ANH alcanza un desempeño comparable al método LIF y superando a QPR. Sin embargo, al evaluar los métodos en señales naturales y sintéticas mediante la norma l_1 en la fase cerrada, QPR exhibió un desempeño superior, logrando los valores más bajos que los obtenidos por el modelo ANH y el método LIF.

Por su parte, el modelo RANH mejora significativamente las estimaciones del flujo glótico para ambos tipos de señales, superando no solo al modelo ANH, sino también a los métodos QPR y LIF en las dos métricas consideradas. Los resultados mostraron que la inclusión del término de regularización de fase cerrada en el modelo RANH contribuye a suprimir distorsiones de baja frecuencia producto de errores en el

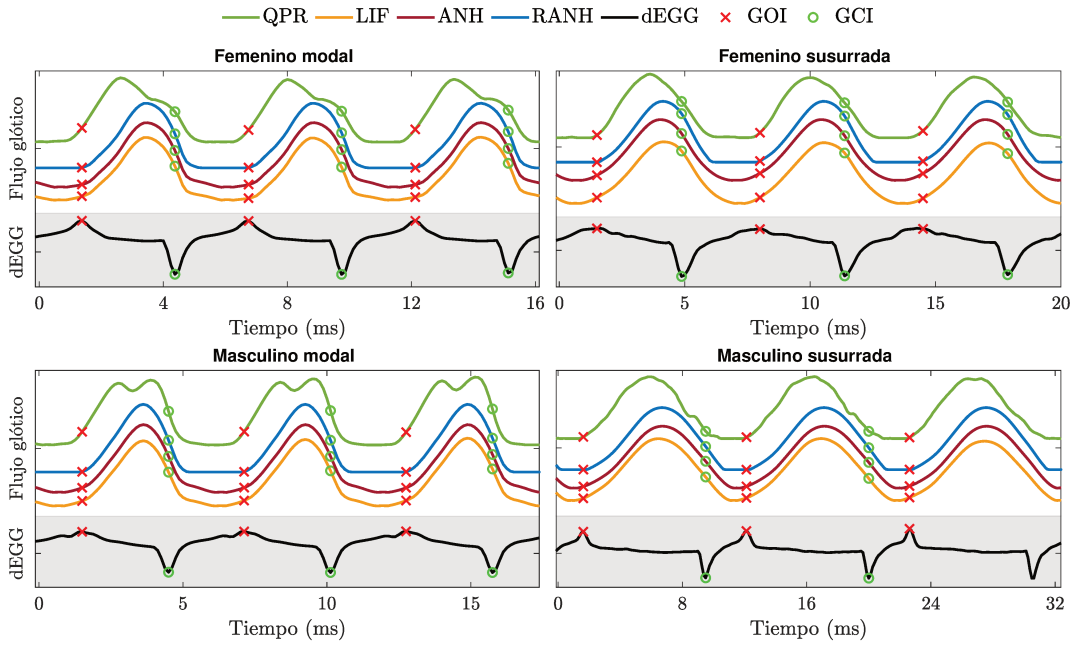

FIGURA 5.4: Flujos glóticos estimados con diferentes métodos a partir de señales de hablantes femeninos (fila superior) y masculinos (fila inferior) con distintas calidades de fonación del repositorio IV: modal (columna izquierda) y susurrada (columna derecha). Cada estimación posee un desplazamiento vertical para facilitar la comparación visual. También se muestra la derivada de la señal del electroglotograma (dEGG), junto con la ubicación de los GCIs y GOIs.

filtrado inverso. Además, la regularización resultó especialmente efectiva para señales caracterizadas por una fase cerrada de larga duración.

Al analizar las estimaciones del flujo glótico mostradas en la Fig. 5.4, obtenidas para señales naturales, se puede apreciarse que el modelo RANH permite obtener reconstrucciones con una fase cerrada más plana en comparación con el modelo ANH y el método LIF. Este comportamiento se refleja también en los valores de la norma l_1 en la fase cerrada reportados en la Sec. 5.2 del Anexo C, donde el modelo RANH y el método QPR alcanzaron los valores más bajos. Por otro lado, en comparación con el método QPR, las estimaciones generadas por el modelo RANH presentan una mejor alineación con los instantes glóticos, especialmente en los GOIs.

5.4. Comentarios de fin de capítulo

En este capítulo se presentó un modelo adaptativo no armónico (ANH) que permite obtener el flujo glótico a partir de la función glótica. Este modelo fue propuesto como una solución a los desafíos prácticos asociados a la estimación del flujo glótico en

el contexto del filtrado inverso de la voz. A partir de dicho modelo se desarrolló una versión regularizada (RANH) que mejora la estimación del flujo glótico al promover una forma de onda plana durante la fase cerrada.

Los resultados obtenidos con señales sintéticas y naturales demuestran que los modelos propuestos constituyen alternativas eficaces para la estimación del flujo glótico, siendo la versión regularizada la que ofrece el mejor desempeño. No obstante, es importante destacar que el modelo RANH requiere conocer la ubicación de los GCIs y GOIs, información indispensable para implementar la regularización en la fase cerrada. Si bien estos instantes pueden estimarse a partir de la señal de voz u otras señales asociadas a la fonación, cualquier error en su determinación puede propagarse y afectar la forma de onda final del flujo glótico. Esto se torna especialmente relevante en casos patológicos o en trastornos de la voz, donde la detección precisa de dichos instantes no siempre es posible. Por lo que en estos escenarios el modelo ANH constituye una alternativa más adecuada para la estimación del flujo glótico.

Otra limitación relevante de los modelos propuestos es que suponen que la forma de onda del flujo glótico presenta una amplitud constante en segmentos de corta duración. Este supuesto restringe su aplicabilidad a señales más extensas, donde pueden aparecer modulaciones de amplitud típicas de la fonación.

El desarrollo del modelo ANH del flujo glótico y su evaluación preliminar dieron lugar a una presentación en un congreso nacional [6]. Asimismo, se elaboró un artículo basado en el modelo regularizado y en los resultados presentados en este capítulo, el cual fue enviado para su revisión a la revista *Biomedical Signal Processing and Control* de la editorial Elsevier [7].

Capítulo 6

Conclusiones

A continuación se dará un cierre a la presente tesis brindando las conclusiones a las que se arribaron al finalizar la etapa de formación doctoral. Estas conclusiones buscan responder si se cumplieron los objetivos generales y específicos que fueron planteados al inicio del proceso, y que se formularon en este documento en la sección 1.4.

En esta tesis se abordó el desarrollo de nuevos métodos de filtrado inverso de la voz destinados a mejorar las estimaciones del flujo glótico, tal como se estableció en el objetivo general. Puntualmente, se realizaron aportes orientados a mejorar el ajuste del filtro del tracto vocal y la obtención del flujo glótico a partir de su derivada, la función glótica.

Se identificó en primer lugar uno de los principales desafíos del filtrado inverso, como es el ajuste adecuado del filtro del tracto vocal. En particular, se encontró que el método de predicción lineal, fundamento de la mayoría de las técnicas empleadas para estimar los coeficientes del filtro, se ve afectado por los errores de predicción de gran amplitud originados por la fuente de excitación acústica involucrada en la producción de la voz, particularmente durante los instantes de cierre glótico.

La causa principal de este problema radica en que el error cuadrático es utilizado como función de costo en predicción lineal, el cual resulta sensible a los errores de gran amplitud que se dan al analizar la señal de voz. Como consecuencia, el filtro del tracto vocal tiende a ajustarse de manera incorrecta, lo que genera distorsiones en las formas de onda obtenidas mediante el filtrado inverso.

Este inconveniente motivó el estudio de diversas estrategias diseñadas para mejorar el ajuste del filtro del tracto vocal en el contexto del filtrado inverso, tal como se planteó en el primer objetivo específico. Se buscó identificar en cada estrategia sus principales características, su relación con la fisiología de la fonación y sus limitaciones, con el fin de detectar oportunidades de mejora.

Abordando el segundo objetivo específico, se contribuyó al desarrollo de nuevas estrategias de predicción lineal con atenuación Gaussiana, orientadas a mitigar los efectos perjudiciales de las muestras de la señal de voz ubicadas alrededor de los instantes de cierre glótico. Nuestros aportes, inspirados en la fisiología de la fonación, consistieron en reformular la estrategia de atenuación Gaussiana para adaptarla a la periodicidad de la señal de voz y, al mismo tiempo, permitir una implementación simple de una ponderación de fase casi cerrada.

Se concluye que atenuar la información de la señal de voz alrededor del instante de cierre glótico y durante una porción significativa de la fase abierta mejora el ajuste del filtro del tracto vocal y, en consecuencia, las estimaciones obtenidas mediante filtrado inverso. Los resultados mostraron que las estrategias propuestas superan a varios métodos del estado del arte; sin embargo, al igual que otras técnicas que consideran la dinámica de la glotis, requieren conocer la ubicación de los instantes glóticos.

Por otro lado, con el fin de superar los problemas que introduce el error cuadrático en el análisis de predicción lineal y en línea con el tercer objetivo específico planteado, en esta tesis se propuso la predicción lineal basada en el criterio de máxima correntropía y se demostró su aplicabilidad para el filtrado inverso de la voz.

Se encontró que la incorporación de la correntropía al esquema de predicción lineal proporciona, de forma simple y efectiva, una solución robusta frente a los efectos perjudiciales de las muestras de la señal de voz ubicadas alrededor de los instantes de cierre glótico. Además, el método propuesto implementa un análisis de predicción lineal ponderado, en el cual una función de ponderación se ajusta iterativamente y de forma guiada por los datos. La función de ponderación resultante enfatiza automáticamente la información de la señal de voz ubicada en la fase cerrada, sin necesidad de determinar a priori los instantes glóticos.

Los resultados indican que nuestro desarrollo basado en correntropía es capaz de superar a métodos de referencia bien establecidos, como la predicción lineal de fase casi cerrada y su variante con compensación espectral.

Finalmente, cumpliendo con el último objetivo específico, se desarrolló un modelo adaptativo no armónico para la estimación del flujo glótico. Este modelo fue diseñado para abordar los desafíos prácticos asociados con la obtención de dicha señal a partir de la función glótica.

Se demostró que es posible ajustar de manera adecuada el modelo adaptativo no armónico del flujo glótico a partir de la información contenida en la función glótica.

Asimismo, se propuso una versión regularizada del modelo que mejora la estimación del flujo glótico al promover una forma de onda plana en la fase cerrada.

Se encontró que la regularización propuesta para la fase cerrada resulta particularmente beneficiosa para señales que presentan una fase cerrada prolongada y para suprimir las distorsiones de baja frecuencia generadas por errores de filtrado inverso. Los resultados indican que el modelo regularizado supera a otros métodos del estado del arte.

Artículos científicos

Los resultados obtenidos durante la realización de esta tesis han sido presentados, preliminarmente, en congresos nacionales e internacionales. Posteriormente, estos resultados fueron publicados (o enviados para su consideración) en revistas científicas indexadas de amplio reconocimiento en el área del procesamiento digital de la voz.

Publicaciones en congresos

- **I. A. Zalazar**, G. A. Alzamendi, and G. Schlotthauer, "Estudio comparativo de técnicas de extracción de fuente glótica basadas en filtrado inverso de la voz," en *XXII Congreso Argentino de Bioingeniería y XI Jornadas de Ingeniería Clínica (SABI)*, 2020.

- **I. A. Zalazar**, G. A. Alzamendi, and G. Schlotthauer, "Gaussian-weighted voice inverse filtering: Effects of varying the attenuation window parameters on the glottal source estimation," en *XIX Workshop on Information Processing and Control (RPIC)*, 2021.

- **I. A. Zalazar**, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Correntropy-based linear prediction for voice inverse filtering," en *18th International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 2022.

- **I. A. Zalazar**, J. V. Ruiz, G. A. Alzamendi, M. A. Colominas, and G. Schlotthauer, "Adaptive Non-Harmonic model for glottal airflow estimation in glottal inverse filtering," en *XXI Workshop on Information Processing and Control (RPIC)*, 2025.

Publicaciones en revistas

- **I. A. Zalazar**, G. A. Alzamendi, and G. Schlotthauer, "Symmetric and asymmetric gaussian weighted linear prediction for voice inverse filtering," *Speech Communication*,vol. 159, p. 103057, 2024, doi: 10.1016/j.specom.2024.103057.
- **I. A. Zalazar**, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Maximum correntropy linear prediction for voice inverse filtering: Theoretical framework and practical implementation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 152-162, 2025, doi: 10.1109/TASLP.2024.3512187.
- **I. A. Zalazar**, G. A. Alzamendi, J. V. Ruiz, M. A. Colominas, and G. Schlotthauer, "Regularized adaptive non-harmonic model for glottal airflow estimation in glottal inverse filtering," *Biomedical Signal Processing and Control*, 2025. En revisión.

Trabajos a futuro

Los métodos desarrollados en esta tesis sientan las bases para continuar avanzando en la mejora de las etapas involucradas en el filtrado inverso de la señal de voz. A continuación, se detallan posibles líneas de trabajo a futuro para cada uno de los aportes realizados.

Respecto al primer aporte, presentado en el Capítulo 3, donde se desarrollaron dos nuevas estrategias de predicción lineal con atenuación Gaussiana, se identifica una oportunidad de mejora mediante la modificación de las ventanas Gaussianas empleadas. En particular, podría incorporarse un parámetro de traslación temporal que permita ajustar la posición de las ventanas con respecto a los instantes de cierre glótico. Nuestra hipótesis es que desplazar dichas ventanas hacia la fase cerrada podría mejorar la estimación de los coeficientes del filtro del tracto vocal. Sin embargo, esta modificación introduce un nuevo parámetro que podría dificultar su aplicabilidad práctica debido a la cantidad de parámetros a configurar. Asimismo, siguiendo las conclusiones de [42], será necesario evaluar con mayor profundidad la robustez de las estrategias propuestas frente a errores en la ubicación de los instantes de cierre glótico.

En cuanto al segundo aporte, relacionado con el método de predicción lineal basado en el criterio de máxima correntropía del Capítulo 4, futuras investigaciones podrían

centrarse en su validación considerando otros conjuntos de datos de la fonación. Además, sería pertinente evaluar su desempeño en escenarios más desafiantes, que incluyan distintos niveles de interacción fuente–filtro, habla continua y fonaciones atípicas. También podrían explorarse estrategias recursivas y adaptativas para la actualización del parámetro del núcleo Gaussiano σ, como las propuestas en [68], [87]. Otra línea de trabajo consistiría en estudiar núcleos alternativos para la correntropía con el objetivo de mejorar aún más el ajuste del filtro del tracto vocal. Una opción interesante es el núcleo Gaussiano generalizado [88], que proporciona una formulación paramétrica flexible que permite modificar la forma del núcleo. Finalmente, el algoritmo desarrollado para estimar los coeficientes de predicción lineal fue diseñado para el filtrado inverso de la voz, por lo que futuras investigaciones podrían evaluar su aplicación en otras señales asociadas a la fonación o en otras señales biomédicas.

Por último, las líneas de trabajo futuro asociadas al tercer aporte del Capítulo 5 incluyen la incorporación de un término para la modulación en amplitud dentro del modelo adaptativo no armónico del flujo glótico, con el fin de capturar de manera más precisa la variabilidad presente en segmentos de voz de larga duración o en habla continua. Evaluaciones adicionales en condiciones ruidosas o reverberantes también serían relevantes para determinar la aplicabilidad del modelo propuesto en escenarios reales. Asimismo, se considera de interés explorar el uso de diccionarios alternativos para el modelo adaptativo no armónico, con el objetivo de reconstruir adecuadamente características fisiológicas del flujo glótico, tales como una fase cerrada plana, sin la necesidad de incluir un término de regularización de fase cerrada.

Anexos

Referido a los artículos incluidos en los Anexos A, B y C:

- **I. A. Zalazar**, G. A. Alzamendi, and G. Schlotthauer, "Symmetric and asymmetric gaussian weighted linear prediction for voice inverse filtering," *Speech Communication*,vol. 159, p. 103057, 2024, doi: 10.1016/j.specom.2024.103057.
- **I. A. Zalazar**, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Maximum correntropy linear prediction for voice inverse filtering: Theoretical framework and practical implementation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 152-162, 2025, doi: 10.1109/TASLP.2024.3512187.
- **I. A. Zalazar**, G. A. Alzamendi, J. V. Ruiz, M. A. Colominas, and G. Schlotthauer, "Regularized adaptive non-harmonic model for glottal airflow estimation in glottal inverse filtering," *Biomedical Signal Processing and Control*, 2025. En revisión.

El tesista declara haber contribuido principalmente en el diseño conceptual, experimental y metodológico, su posterior implementación y la correspondiente evaluación de los métodos descritos y los experimentos realizados para obtener los resultados que allí se presentan. Estas tareas fueron realizadas bajo la guía y supervisión del director Dr. G. A. Alzamendi y el codirector Dr. G. Schlotthauer. En cuanto a la escritura de los artículos, el tesista es el autor principal, guiado por sugerencias y revisiones de los directores y otros coautores que en cada artículo se indican. Los abajo firmantes avalan esta declaración.

Dr. G. A. Alzamendi
Director

Dr. G. Schlotthauer
Codirector

Anexo A

Symmetric and asymmetric gaussian weighted linear prediction for voice inverse filtering

Symmetric and asymmetric Gaussian weighted linear prediction for voice inverse filtering

I. A. Zalazar[a], G. A. Alzamendi[a], G. Schlotthauer[a]

[a]*Instituto de Investigación y Desarrollo en Bioingeniería y Bionformática, CONICET-UNER, Oro Verde, Entre Ríos, Argentina*

Abstract

Weighted linear prediction (WLP) has demonstrated its significance in voice inverse filtering, contributing to enhanced methods for estimating both the vocal tract filter and the glottal source. WLP provides a mechanism to mitigate the effect on the linear prediction model of voice samples that affects the vocal tract filter estimation, particularly those samples around glottal closure instants (GCIs). This article studies the Gaussian weighted linear prediction (GLP) strategy, which employs a Gaussian attenuation window centered at the GCIs to reduce its contribution in the WLP analysis. In this study, the Gaussian attenuation is revisited and a parameterization of the window that adjusts to the typical variability in voice periodicity is introduced. In addition, an asymmetric Gaussian window is proposed to diminish the relevance of voice samples preceding GCIs on the WLP model, thus providing a quasi closed phase inverse filtering method. Characterization of symmetric and asymmetric GLP methods for glottal source estimation is addressed based on synthetic and natural phonation data, resulting in a set of optimal parameters for the Gaussian attenuation windows. The results show that the proposed asymmetric attenuation improves voice inverse filtering with respect to the symmetric GLP method. Comparisons with other state-of-the-art techniques suggest that the proposed GLP approaches are competitive, falling slightly short in performance only when contrasted with the well-known quasi closed inverse filtering analysis. The simplicity of implementing the attenuation windows, coupled with their robust performance, positions the proposed GLP methods as two attractive and straightforward voice inverse filtering techniques for practical application.

Keywords: Voice inverse filtering, Glottal source estimation, Weighted linear prediction, Gaussian attenuation window, Quasi closed phase analysis.

1. Introduction

Linear prediction (LP) theory has become a well-established framework in voice signal processing and modeling [1]. This is partially due to the close relationship between the LP model and the source-filter theory. According to the source-filter theory, human phonation results from the interplay of three simple, physically-relevant components: the acoustic excitation source at the glottal level, the vocal tract as an acoustic filter spectrally modulating the different voice sounds, and the lip radiation producing the free-field acoustic pressure [2, 3]. LP modeling plays a key role in the context of voice inverse filtering, which involves

the numerical estimation of the acoustic excitation, i.e., the glottal volume velocity (or its time-derivative, known as the glottal function), from a voiced speech recording [4, 5]. Tunable digital filters are usually applied to model the contributions of the vocal tract and the lip radiation [6]. Inverse models of these filters are then used to cancel out the effects of the vocal tract and lip radiation in the voice signal, yielding thus the estimated glottal source [7, 8]. Accurate modeling of vocal tract resonances and radiation at the lips is crucial for relevant estimations [9]. Inverse filtering techniques have been applied in the context of speaker identification [10, 11], emotion recognition [12, 13], and the detection of voice disorders and diseases [14, 15].

In the conventional approach to inverse filtering, the lip radiation effect is generally approximated using a digital first-order causal differentiator [3]. In addition, LP model assumes an auto-regressive vocal tract filter with transfer function [16]:

$$V(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{1}$$

where P is the filter order, and a_k are the LP coefficients [3]. The baseline standard for adjusting the coefficients in the model of Eq. (1) from natural voice signals is the classic LP coding based on least-mean-square prediction error [17]. Instead, the weighted linear prediction (WLP) aims to improve the LP scheme, providing further control by including a time-domain weighting function w that weights the prediction errors differently [6]. In WLP, the filter coefficients are computed by minimizing the weighted mean squared prediction error [18]. WLP has proven extremely effective in voice inverse filtering, enhancing model robustness and yielding more relevant vocal tract filters [19].

In WLP, a weighting function drawn upon empirical criteria or prior knowledge regulates the relative importance of the prediction error samples in the LP analysis [20]. For example, the short-time energy (STE) function from voice signal was used as a weighting function in [18]. STE yields large amplitude values in the samples with a higher signal-to-noise ratio. Thus, a weighting function based on STE aided in increasing the robustness of LP analysis to the additive noise. This idea was revisited in [21], where a stabilized WLP (SWLP) analysis was developed to ensure the stability of the vocal tract filter. The application of SWLP as an inverse filtering technique was subsequently evaluated in [22]. Furthermore, it is well known that LP model is less accurate at the glottal closure instants (GCIs), due to the lack of predictability of the impulse-like glottal excitations arising from the abrupt deceleration of the glottal volume velocity [23]. In [16], WLP with an attenuated main excitation (AME) window was introduced to diminish the detrimental effects of the GCIs, thereby improving the formant estimation accuracy, especially for high-pitched voices. Afterward, a modified version of the AME window was applied for inverse filtering in the quasi closed phase (QCP) analysis [24], which places greater emphasis on voice samples in the closed phase than on the open phase to simultaneously reduce the adverse effects of the GCIs and the acoustical coupling between the sub- and supraglottal systems. In [25], a sparse LP model was proposed by applying WLP with an attenuation window based on Gaussian functions centered at the GCIs. This

method seeks to lessen the penalization around the glottal excitations to generate sparser prediction errors. Hereafter, we refer to this method as the Gaussian weighted linear prediction (GLP) analysis. Recently, it was shown that GLP provides a simple and robust alternative for voice inverse filtering [26].

Weighted LP methods inspired by relevant features in the glottal cycle, like QCP and GLP, have shown superior performance in voice inverse filtering [26]. However, they depend on knowing the GCI locations or having sufficiently accurate estimates of the instants, a disadvantage not shared with the classic-LP or SWLP [27]. Compared to other methods, GLP has shown great robustness to estimation errors in the GCIs [26], a valuable characteristic in challenging cases, such as atypical or impaired phonation.

Recently, WLP inverse filtering schemes have been introduced that center iteratively and automatically on the closed phase [27, 28]. The weighting function shape and complexity profoundly impact in the success of inverse filtering. For instance, given the GCIs, the AME window in QCP requires a fine adjustment of four shape parameters to yield suitable results [27], whereas the simpler GLP takes only two shape parameters [25, 26].

GLP is attractive due to the simplicity of the Gaussian attenuation function. Although Gaussian parameters have a straightforward context-dependent interpretation, a common practice in GLP-based inverse-filtering is to use fixed Gaussian standard deviation, neglecting thus the nature and significant variability intrinsic to human phonation [2]. In this article, GLP formulation is revisited and an alternative parameterization for the Gaussian function is proposed. A fundamental aspect that considers this parameterization is the periodicity of voiced phonation, which varies considerably depending on factors such as the speaker's sex, body constitution, intonation, and voice quality [7]. This document is a sequel to a conference paper presented in [29]. Here, the influence of modifying the parameters of the Gaussian attenuation window on the inverse filtering performance is thoroughly studied.

The Gaussian window studied in [29] symmetrically underweights the voice samples around GCIs. However, it is well-known that closed phase analysis improves inverse filtering yielding more accurate estimates [8, 30]. To address this, here an asymmetric Gaussian attenuation is proposed to enable a quasi closed phase GLP, allowing more emphasis on samples during the closed phase. Notably, this asymmetric window maintains the same simple parameterization as the symmetric GLP while yielding improved results in the realm of inverse filtering.

The rest of the paper is as follows. Section 2 revisits the Gaussian attenuation implementation in the GLP scheme, and then the asymmetric GLP is introduced. Section 3 describes the voice inverse filtering data and the experimental setup. Section 4 provides a detailed analysis of parameter selections in voice inverse filtering through symmetric and asymmetric GLP. Section 5 presents the results and discusses their relevance. Finally, we deliver the conclusions of this work in Section 6.

2. Weighted linear prediction

According to the LP model of phonation, the voice signal, $s[n]$, for index time $1 \leq n \leq N$, follows an autoregressive process:

$$s[n] = \mathbf{a}^T \mathbf{s}[n] + e[n], \tag{2}$$

where $\mathbf{a} = [a_1,\, a_2, \cdots, a_P]^T$ are the LP coefficients, with P being the model order, $\mathbf{s}[n] = [s[n-1], s[n-2], \cdots, s[n-P]]^T$ is a vector gathering the P last past voice samples, and $e[n]$ is the prediction error.

In classic LP, vector $\mathbf{a}$ is typically obtained by minimizing the mean squared prediction error in Eq. (2), where each error sample contributes equally to the model adjustment. Instead, in WLP, a weighting function w is included to control the relative weight of each error sample. As a result, the vector $\mathbf{a}$ of WLP is computed by solving the following optimization problem:

$$\hat{\mathbf{a}} = \min_{\mathbf{a}} \sum_{n=1}^{N+P} e^2[n] w[n] \;\; \text{s.t:} \;\; e[n] = s[n] - \mathbf{a}^T \mathbf{s}[n], \tag{3}$$

where the extremes in Eq. (3) are set to use the autocorrelation method [17]. The Eq. (3) has analytical solution [24]:

$$\hat{\mathbf{a}} = \left(\sum_{n=1}^{N+P} w[n] \mathbf{s}[n] \mathbf{s}^T[n] \right)^{-1} \left(\sum_{n=1}^{N+P} w[n] s[n] \mathbf{s}[n] \right). \tag{4}$$

As previously discussed, the weighting function controls the contribution of the prediction error samples for the computation of the LP model coefficients[18]. Prior works in inverse filtering demonstrate that more relevant vocal tract filters are obtained by minimizing the influence of the voice data around the GCIs and by centering the LP analysis on the glottal closed phase [16]. Herein, we address the Gaussian attenuation in GLP-based inverse filtering. In particular, we revisit its formulation to account for the significant variance in voice fundamental frequency and introduce an asymmetric Gaussian attenuation, enabling a quasi closed phase GLP.

2.1. Gaussian attenuation revisited

In GLP, a Gaussian attenuation is applied around the GCIs. As proposed in [25], the weighting function is defined as follows:

$$w_{\text{sym}}[n] = 1 - \kappa \sum_{l=1}^{L} g_s[n - n_l], \tag{5}$$

where n_l denotes the discrete location of the lth-GCI among L instants, κ controls the attenuation level around n_l, and $g_s[n - n_l]$ is a Gaussian function with center in n_l and standard deviation σ_1:

$$g_s[n - n_l] = e^{-(n-n_l)^2/2\sigma_1^2}. \tag{6}$$

Subscript "sym" in w_{sym} denotes symmetric. Fig. 1 illustrates the Gaussian attenuation; the weighting function w_{sym} obtained by Eq. (5) conditioned on the

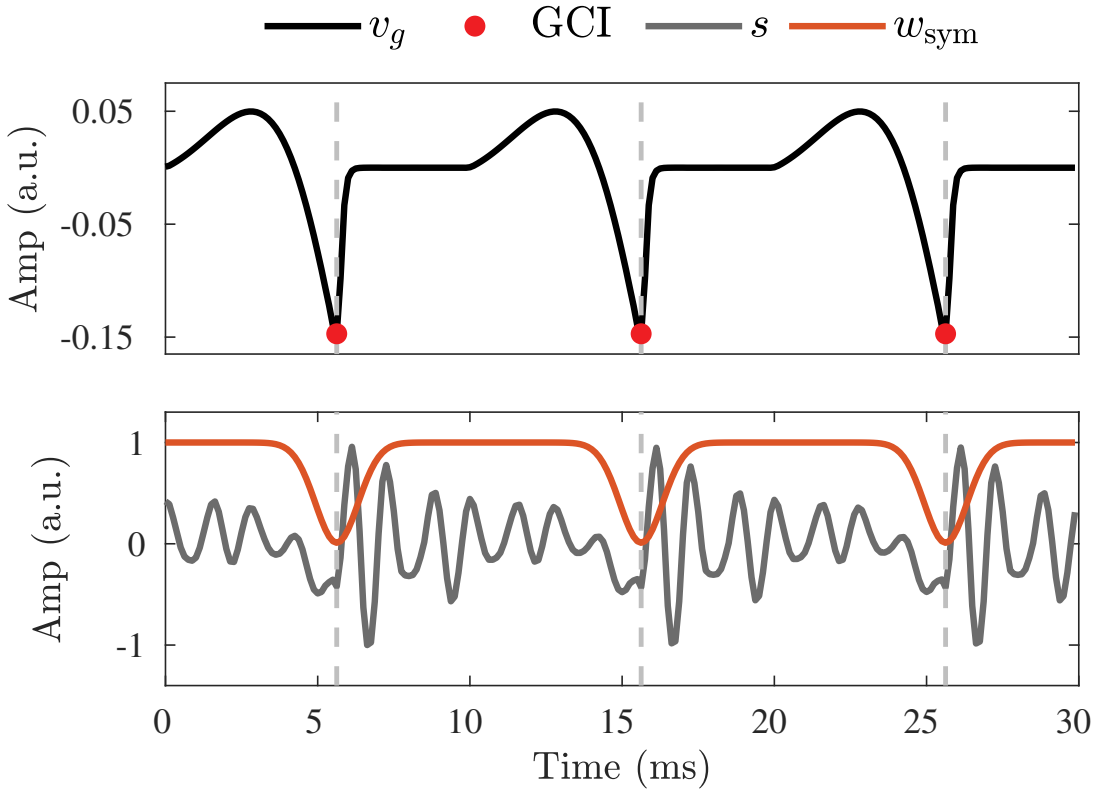

Figure 1: Illustration of the Gaussian attenuation of GCIs in GLP method. Top: Simulated glottal function with GCIs (filled-dot marks). Bottom: Synthetic voice signal (vowel /a/) with the Gaussian attenuation window superimposed.

GCIs is applied to a voiced segment s. This function down-weighs the samples of s with larger amplitudes, which are a manifestation of the main peak-like excitations in the latent glottal function. The attenuation of the voice samples in each GCI is governed by the parameter κ, whose admissible values lie within the range $0 \leq \kappa < 1$. The constraint $\kappa < 1$ is necessary because, for a weighting function with $\kappa = 1$ in Eq. (5), the coefficients $\hat{\mathbf{a}}$ obtained by Eq. (4) could generate an unstable vocal tract filter [21].

Voiced phonation is determined by (almost) periodic vocal fold oscillations. The voice periodicity is assessed by the fundamental period, T_0, the fundamental frequency, $f_0 = 1/T_0$, or in the case of discrete signals, the number of samples per glottal cycle, N_0. The distances between consecutive GCIs in the voice signal are approximately equal to N_0; hence, the window w_{sym} takes the periodicity into account by centering Gaussian functions at every GCI, as shown in Eq. (5). However, the relation of Gaussian width to the fundamental period is usually disregarded, since it is a common practice to use the same σ_1 for inverse filtering across various voice signals [26]. Instead, here we propose to adjust the Gaussian width relative to N_0, as follows:

$$\sigma_1 = \alpha N_0, \quad 0 \leq \alpha \leq \alpha_{\max}, \tag{7}$$

where the upper bound $\alpha_{\max}$ is computed based on the maximum allowed overlap A_0 for consecutive Gaussian functions, as described in the Appendix. The value of A_0 is set to 0.1 to prevent significant overlap of the adjacent Gaussian functions in

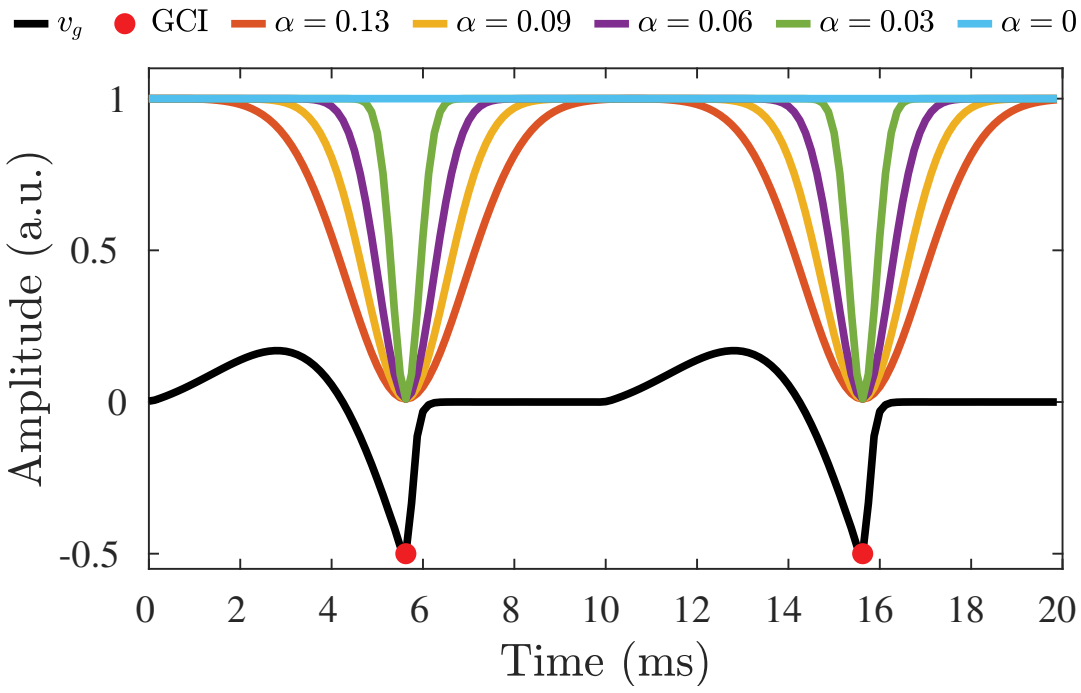

Figure 2: Effect in the Gaussian attenuation by changing the width σ_1 varying α in Eq. (7), compared with a theoretical glottal function with GCIs (dot marks).

Eq. (5). On the other hand, the case $\alpha = 0$ in Eq. (7) is considered equivalent to applying a constant weight as in classic LP. Fig. 2 shows examples of attenuation windows with different widths σ_1 obtained by varying α in comparison with a synthetic glottal function, v_g. It can be seen that larger α values yield increased Gaussian widths and, consequently, broader attenuation regions around the GCIs.

2.2. Asymmetric Gaussian attenuation

The weighting function w_{sym} in Eq. (5) has the limitation of symmetrically attenuating voice samples around the GCIs, without distinguishing between open and closed glottal phases. Current evidence shows that inverse filtering analysis based on the closed phase information yields more reliable and accurate estimates [24, 27, 30, 31]. In contrast, the information on the GCIs and the open phase contributes negatively to the inverse filtering estimation [16]. To address this, we investigate an extension of the Gaussian attenuation window that emphasizes the information in the closed phase while attenuating the open phase. We thus propose the following asymmetric Gaussian attenuation window:

$$w_{\text{asym}}[n] = 1 - \kappa \sum_{l=1}^{L} g_a[n - n_l], \tag{8}$$

where $g_a[n - n_l]$ is an asymmetric Gaussian function [32]:

$$g_a[n - n_l] = \begin{cases} e^{-(n-n_l)^2/2\sigma_1^2}, & \text{if } n \geq n_l, \\ e^{-(n-n_l)^2/2\sigma_2^2}, & \text{if } n < n_l, \end{cases} \tag{9}$$

$\sigma_1 = \alpha N_0$ is the width for the closed phase, and σ_2 controls the Gaussian width for the open phase:

$$\sigma_2 = r\sigma_1 = r(\alpha N_0), \quad 1 \leq r \leq r_{\max}, \tag{10}$$

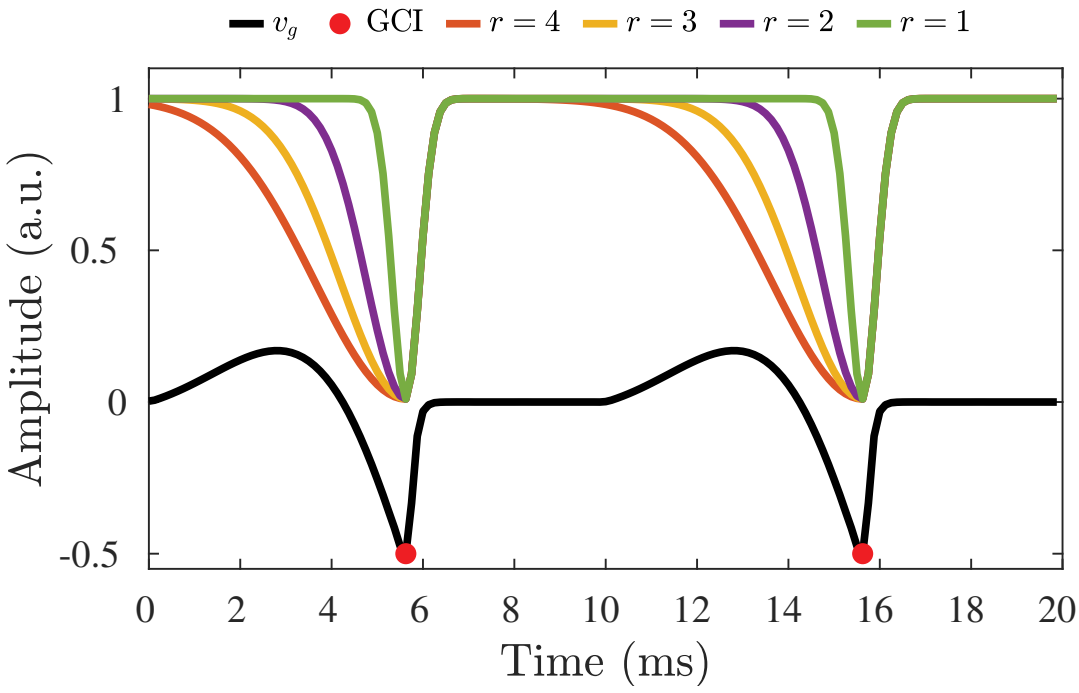

Figure 3: Examples of asymmetric Gaussian attenuation windows with different values of the asymmetry parameter r (considering $\sigma_1 = 0.03N_0$) for a synthetic glottal function with their GCIs.

where r is an asymmetry parameter, and $r_{\max}$ is the maximum asymmetry ratio of σ_2 over σ_1 defined as $r_{\max} = \alpha_{\max}/\alpha$. Then, the range of r depends on the α values.

The parameter r controls the asymmetry of w_{asym}. For example, $r = 1$ corresponds to the symmetric Gaussian function, i.e., $\sigma_2 = \sigma_1$, and $w_{\text{asym}} = w_{\text{sym}}$. On the other hand, if $r = r_{\max}$, σ_2 takes its maximum value $\sigma_{\max} = \alpha_{\max} N_0$ where $\alpha_{\max}$ is as explained above. Fig. 3 shows examples of the asymmetric Gaussian attenuation w_{asym} for different values of the asymmetry parameter r along with a synthetic glottal function v_g. It can be seen that when r increases, the asymmetry proposed in the Gaussian attenuation instruments a WLP analysis that underweights more samples in the open phase (i.e., those that precede the GCIs).

As in the symmetric case, the width σ_2 is bounded by $\sigma_{\max}$. If $\sigma_2 > \sigma_{\max}$, the left-side tail of a Gaussian function could significantly overlap the tail of the predecessor Gaussian function, thus detrimentally affecting the emphasis on the closed phase. Consequently, the asymmetric Gaussian attenuation would not fit its design criteria.

For clarity in the exposition, we hereafter denote with SGLP and AGLP the symmetric and asymmetric Gaussian weighted linear prediction methods introduced in Secs. 2.1 and 2.2.

3. Materials and Methods

3.1. Phonation data for inverse filtering evaluation

In this section, the methods AGLP and SGLP, previously introduced, are investigated for voice inverse filtering, and comparisons with other state-of-the-art techniques are provided. The well-established OPENGLOT database [33] has

been employed due to its versatility in serving as an open environment for the assessment of several inverse filtering methods. OPENGLOT gathers phonation signals for different types of test voice data; in this work, only Repositories II and IV are considered.

Repository II contains signals obtained from a physical model of human phonation that considers vocal fold kinematics, aero-acoustic interactions at the glottal level, and the pressure wave propagation in the vocal tract. The available material includes the voice signal s, the theoretical glottal airflow u_g, with its time-derivative, the glottal function v_g, and the first four formant frequencies, allowing thus the direct assessment of voice inverse filtering methods. The phonation data correspond to three vocal tract configurations for the vowels [/a/, /i/, /u/], four fundamental frequencies, i.e., 82, 110, 156, and 220 Hz for male speakers, and 175, 196, 220, and 294 Hz for female speakers, and three different degrees of vocal fold abduction. The sampling frequency of the signals is $f_s = 44.1$ kHz.

On the other hand, Repository IV contains voice and electroglottogram signals recorded during the natural production of vowel sounds at modal and breathy phonation qualities. The signals correspond to five male and five female speakers that produce a vowel sound with three different pitch levels (low, medium, and high). The sampling frequency of the signals is $f_s = 44.1$ kHz. Glottal airflow and vocal tract information are not available in this repository, thus a direct comparison against a ground truth can not be undertaken.

In accordance with the hypothesis supporting the source-filter theory [2], all signals used in the simulations were resampled at 8 kHz. The inverse filtering analysis was performed for non-overlapping segments of the voice signals, each lasting 50 ms. Pre-processing consisting of pre-emphasis filtering, with transfer function $R(z) = 1 - 0.99z^{-1}$, was applied to cancel the acoustic radiation effects at the lips.

3.2. Experimental setup

The present work aims to study the influence of the parameters κ, α, and r on the outcomes of voice inverse filtering using SGLP and AGLP methods. For this, the following set of parameters are explored: $\kappa \in \{0.1, 0.5, 0.8, 0.9, 0.99, 0.999\}$, and $\alpha \in \{0, 0.01, 0.02, \ldots, \alpha_{\max}\}$, with $\alpha_{\max} = 0.2$ resulting from considering a maximum allowed overlap $A_0 = 0.1$ (see Appendix). For the asymmetry parameter, we consider 15 values in the range $1 \leq r \leq 1.5\, r_{\max}$, where $r_{\max}$ is explained above. Finally, the attenuation windows are normalized using the process described in the Appendix to ensure that their amplitudes vary in the range $[1 - \kappa, 1]$.

Three state-of-the-art inverse filtering methods in the literature are considered as references for comparing the proposed methods: LP [17], SWLP [21], and QCP [24]. In the WLP schemes, the weighting functions are adjusted based on the parameter values informed by the authors [21, 24]. In SWLP, the window length M, for computing the STE function, is set to 12 samples. For QCP, the parameters for the AME window are set as follows: $d = 10^{-5}$, $PQ = 0.05$, $DQ = 0.7$, and $N_{\text{ramp}} = 7$.

Two different strategies were employed to determine the GCIs for QCP, SGLP, and AGLP methods. For the synthetic signals in Repository II, where the theoretical glottal data are available, the GCIs were determined from the location

of the minimum in the glottal function v_g, for each glottal cycle. Before determining the GCIs, the voice signal and the glottal function were time-aligned to avoid any phase lag due to the acoustic wave propagation along the vocal tract. In contrast, for the natural voices, the GCIs and the closed phase detection were computed from the electroglottogram signals using the SIGMA algorithm [34]. The time alignment between the voice and electroglottogram signals was performed to avoid the time delay generated during the recording of the signals. All lag compensations were calculated via cross-correlation measures [3].

Finally, the poles of the estimated vocal tract filter were analyzed. To correctly represent the vocal tract formant frequencies, the poles of the transfer function (see Eq. (1)) in the z-plane must be close to the unit circle [2]. Poles that do not satisfy this condition would provide poor inverse filtering estimates [31]. To address this concern, the computed vocal tract filter was processed to remove the unsuitable poles, previous to applying it for voice inverse filtering [31].

3.3. Performance measure

As mentioned above, direct evaluation of the inverse filter estimates is only possible for the synthesized data from Repository II. As a first measure, we considered the relative waveform error between the theoretical glottal function v_g, and the estimate $\hat{v}_g$, obtained by applying the methods under study. First, the estimate is normalized by an orthogonal projection [26]:

$$\tilde{v}_g(n) = \frac{\sum_{i=1}^{N} v_g(i)\hat{v}_g(i)}{\sum_{i=1}^{N} \hat{v}_g^2(i)}\, \hat{v}_g(n), \qquad n = 1, ..., N, \tag{11}$$

where N denotes the length of the analyzed segment. The signals in Eq. (11) are time-aligned to avoid any phase delay due to the acoustic wave propagation along the vocal tract or the applied inverse filtering method. The lag compensation was computed via cross-correlation measure [3]. Finally, the waveform error for the glottal function relative to the RMS value of the theoretical signal was calculated:

$$E_{v_g} = \frac{1}{N} \sum_{n=1}^{N} \frac{|v_g(n) - \tilde{v}_g(n)|}{\mathrm{RMS}(v_g)}. \tag{12}$$

In addition to the relative waveform error for the glottal function, the inverse filtering evaluation based also on the estimation error for five aerodynamic parameters describing the glottal airflow and its time-derivative: NAQ, OQ, QOQ, ClQ, and H1H2. For example, NAQ quantifies the relative duration of the closed phase in the glottal cycle [35]; any estimation error in this parameter may indicate the presence of spurious oscillations in the closed phase as a result of incorrect inverse filtering [30]. On the other hand, the parameters OQ, QOQ, and ClQ measure the relative duration of the open phase and the closing phase with respect to the fundamental period of the glottal cycle [6]; therefore, errors in these parameters indicate an inaccurate glottal signal recovering over specific parts in the glottal cycles [36]. Finally, H1H2 is a frequency-domain parameter describing the spectral decay of the glottal airflow signal, so any error in H1H2 would evidence alterations in the estimated closed phase [37].

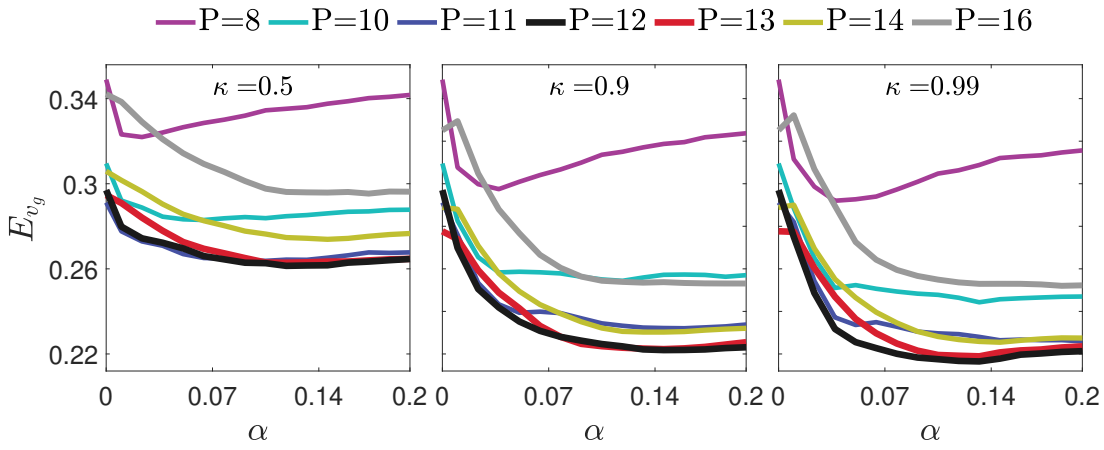

Figure 4: Effects of varying the Gaussian attenuation parameters in SGLP method. Average relative waveform errors of the glottal function in Repository II from varying α, κ, and vocal tract filter order P. Left: $\kappa = 0.5$. Middle: $\kappa = 0.9$. Right: $\kappa = 0.99$.

Estimation errors for the parameters NAQ, OQ, QOQ, and ClQ are reported as average absolute relative error:

$$\text{Error } \Delta = E\left[\frac{|\Delta_{\text{ref}} - \Delta_{\text{est}}|}{\Delta_{\text{ref}}}\right], \tag{13}$$

where Δ denotes any of the parameters NAQ, OQ, QOQ, and ClQ. Additionally, the error for the parameter H1H2 is reported by the average absolute error:

$$\text{Error H1H2} = E\left[|\text{H1H2}_{\text{ref}} - \text{H1H2}_{\text{est}}|\right].$$

For the natural signals in Repository IV, the inverse filtering methods are assessed similarly as in [27, 30]. A first evaluation measures the l_1 norm, i.e., the sum of the absolute values, for the glottal airflow samples in the closed phase. This metric serves as an indicator of the flatness of the closed phase, a desired characteristic in recovered glottal airflow signals [9]; a large value of l_1 norm indicates the presence of spurious oscillations or an incomplete closed phase. In a second evaluation, inverse filtered signals from the different methods are compared with respect to the NAQ parameter, where the results from QCP were considered the benchmark values. As QCP has proved to be one of the best-suited voice inverse filtering methods, we assume that any methods producing similar results as those obtained by QCP could be considered a befitting choice for analyzing natural voice signals.

4. Parameter analysis for GLP methods

In this section, we investigate the influence of the parameters κ and α in the weighting function w_{sym} of the SGLP method for voice inverse filtering. As a complementary study, we analyze the impact of the asymmetry parameter r to ascertain whether the window w_{asym} of the AGLP method enhances voice inverse filtering outcomes when compared to those achieved by the SGLP method. The findings from these investigations provide an optimal parameter range for both the SGLP and AGLP methods.

4.1. Gaussian attenuation

We delve into the analysis of adjusting the parameters governing Gaussian attenuation within the SGLP method. Fig. 4 shows the average relative waveform error of the glottal function, denoted as E_{v_g}, for the synthesized signals from Repository II for seven vocal tract filter orders P. The results correspond to different Gaussian widths from varying α in Eq. (7) and three attenuation levels κ. In Fig. 4, a clear decrease in the error can be observed as the parameter α is increased for all filter orders. As expected, attenuating the samples around the GCIs using the window w_{sym} leads to improved voice inverse filtering estimations when compared to applying a constant weight (case $\alpha = 0$). It is observed that the waveform error decreases as the order P of the vocal tract filter increases from 8 to 12. On the other hand, order $P = 13$ produces similar errors to those from order $P = 12$ except for the range of small α values, where waveform errors for $P = 13$ are slightly high. Finally, for $P = 14$ and $P = 16$, the waveform errors are significantly larger than those from $P = 12$. Similar results are observed for all the considered κ values. The simulations also show that increasing κ from 0.1 (not shown in Fig. 4) to 0.99 results in lower waveform errors. In contrast, for κ values above 0.99, no discernible reduction in E_{v_g} is observed.

Further analysis of Fig. 4 reveals that all error curves exhibit a local minimum at a specific α value. In particular, the curve corresponding to $P = 12$ and $\kappa = 0.99$ displays the lowest error within the range $0.1 < \alpha < 0.15$. In addition, we can observe an upward trend in error for $\alpha > 0.15$ on this curve. The increase in the error is attributed to the larger window widths, which causes a higher proportion of voice signal samples around the GCIs to be attenuated. This excessive attenuation leads to the loss of essential information needed to accurately estimate the vocal tract filter, especially information in the closed phase. As a result, inverse filtering based on this deteriorated information yields inaccurate estimates of the glottal function. Therefore, the best results in our simulations are obtained for $P = 12$, attenuation level $\kappa = 0.99$, and window width in the range $0.1 < \alpha < 0.15$. This underscores the need to attenuate some samples after the GCI, such as the return phase, which is unsuitable for accurate vocal tract filter estimation.

Based on the aforementioned analysis, we propose a simplified implementation of the window w_{sym} assuming that the GCIs are given. Our simulations indicate that suitable parameters for w_{sym} are:

$$\kappa = 0.99, \qquad \alpha = 0.11. \tag{14}$$

Note that this selection yields the lowest relative waveform errors in Fig. 4. Furthermore, the proposed α is approximately half the size of $\alpha_{\max} = 0.2$ (obtained for a maximum overlap $A_0 = 0.1$ in Eq. (A.4)), ensuring the correct implementation of the symmetric window w_{sym}.

4.2. Asymmetric Gaussian attenuation

We will now investigate the effect of asymmetry in the Gaussian attenuation window w_{asym}, defined in the Eq. (8). The purpose of this window is to employ different weighting for the information on the closed and the open phases, thereby enhancing the voice inverse filtering results. Figure 5 presents a map illustrating the average relative waveform errors E_{v_g} from Repository II using the AGLP

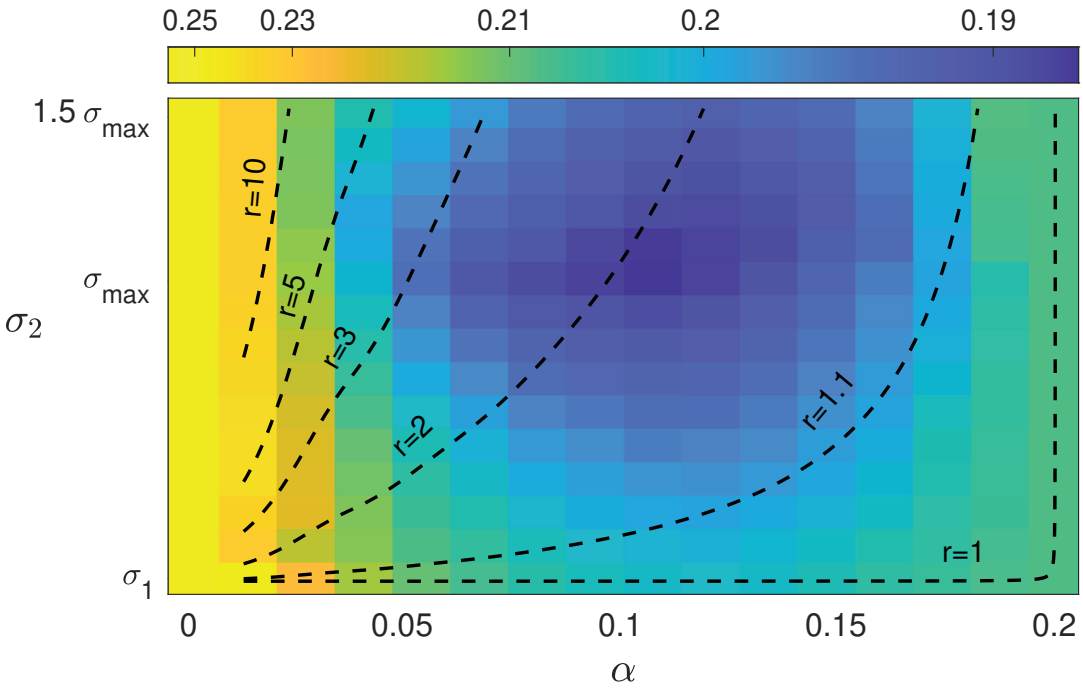

Figure 5: Map of the average relative waveform error for the glottal function v_g obtained from the Repository II for different values of α and σ_2. A non-linear color mapping was applied for better discrimination in the low waveform error region. For interpretation purposes, isolines for the asymmetry parameter r are drawn superimposed.

method with $P = 12$, $\kappa = 0.99$, and $\alpha_{\max} = 0.2$ (obtained from $A_0 = 0.1$). The map was constructed by varying α in the range $[0, \alpha_{\max}]$, and σ_2 in the range $[\sigma_1, 1.5\,\sigma_{\max}]$, as described in Sec. 2.2. In addition, the isolines of the asymmetry parameter r from the definition of the width σ_2 for the open phase (as indicated by Eq. (10)), are superimposed in Fig. 5. These isolines describe the asymmetry ratio in the window w_{asym}. For instance, the isoline $r = 1$ retrieves the symmetric Gaussian attenuation, whereas the isolines $r > 1$ corresponds to a broader attenuation region for the open phase relative to the closed phase (as shown in Fig. 3).

Analysis of Fig. 5 indicates that the error substantially decreases by jointly increasing α and σ_2 quantities, compared to the case of constant weighting ($\alpha = 0$). As expected from the previous section, the symmetric case ($r = 1$) exhibits a reduction in error as α increases. However, the proposed asymmetric attenuation yields a further error reduction. In particular, the minimum error is given for $\alpha \approx 0.1$ and $\sigma_2 \approx \sigma_{\max}$. The latter coincides with the isoline $r \approx 2$ that depicts the asymmetric case with an attenuation region for the open phase twice broader than for the closed phase. Instead, for $\sigma_2 > \sigma_{\max}$, the waveform error increases for the different α values. This performance degradation is due to excessive overlap in w_{asym}, as described in Sec. 2.2. Similar maps were obtained when varying the filter order and the attenuation coefficient; however, the minimum error was observed by setting $P = 12$ and $\kappa = 0.99$. This evidence supports our hypothesis that applying an asymmetric Gaussian attenuation can suitably provide enhanced inverse filtering results.

We propose a simplified implementation of the window w_{asym}, requiring only

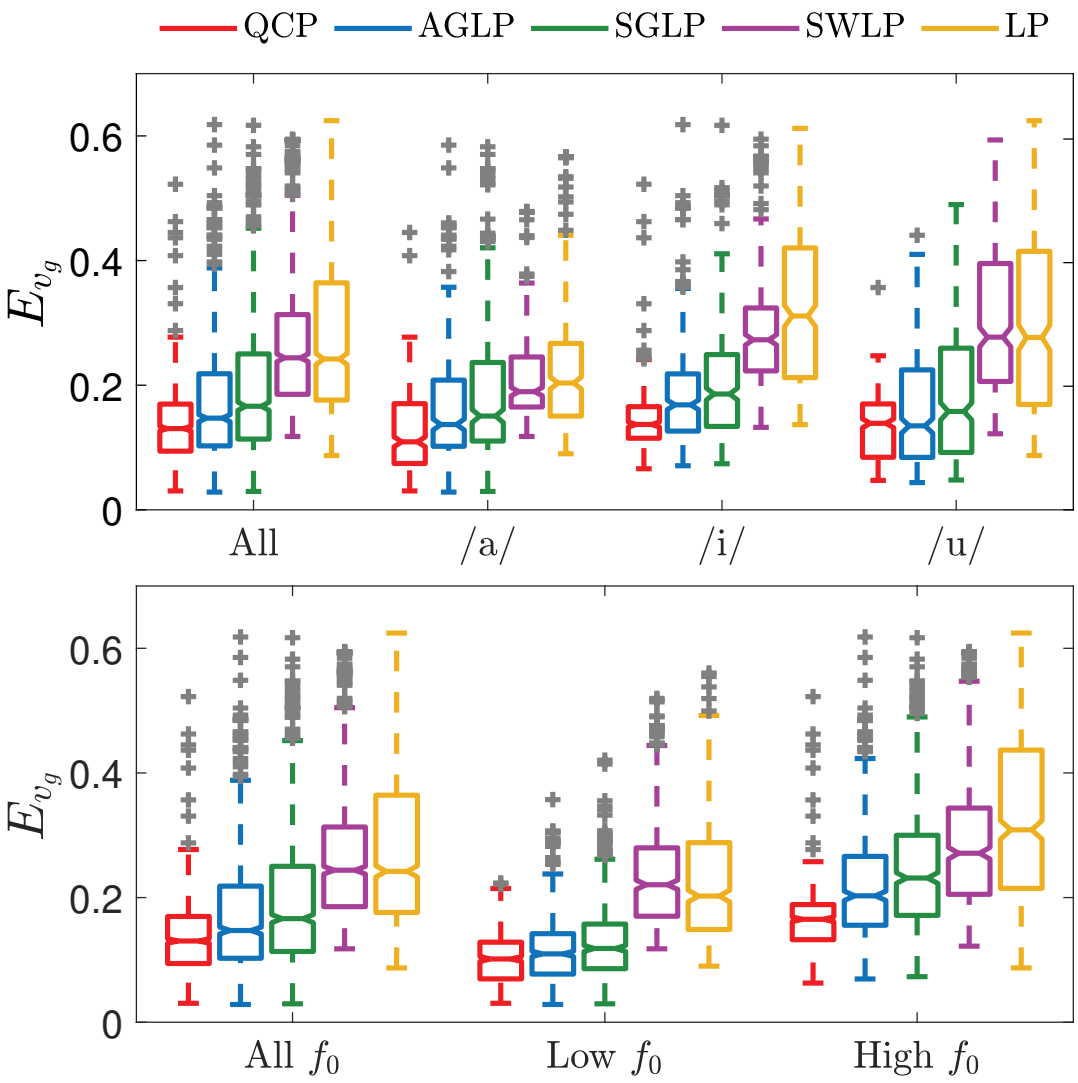

Figure 6: Box plots of the Error v_g for synthesized signals from the Repository II for five inverse filtering methods, shown with respect to (top) vowel sounds and (bottom) f_0 ranges.

the GCI locations, similar to that discussed previously for w_{sym}. Based on the previous results, we suggest the parameters setting:

$$\kappa = 0.99, \qquad \alpha = 0.1, \qquad \sigma_2 = 0.2N_0. \tag{15}$$

The proposed selection yields $r = 2$, in accordance with the minimum waveform error region in Fig. 5. Additionally, the relation $\alpha \leq \alpha_{\max}$ is guaranteed, which is required for the correct implementation of w_{asym}.

5. Results

5.1. Synthetic voice data

The performance for retrieving the theoretical glottal function in Repository II is addressed. Five vocal tract filter orders were considered for processing every voice signal with $P \in \{8, 9, \ldots, 12\}$. The methods SWLP and QCP use the same parameters reported in [21, 24] and in Sec. 3.2, respectively. Instead, for the SGLP and AGLP methods we use the setting proposed in Eqs. (14) and (15).

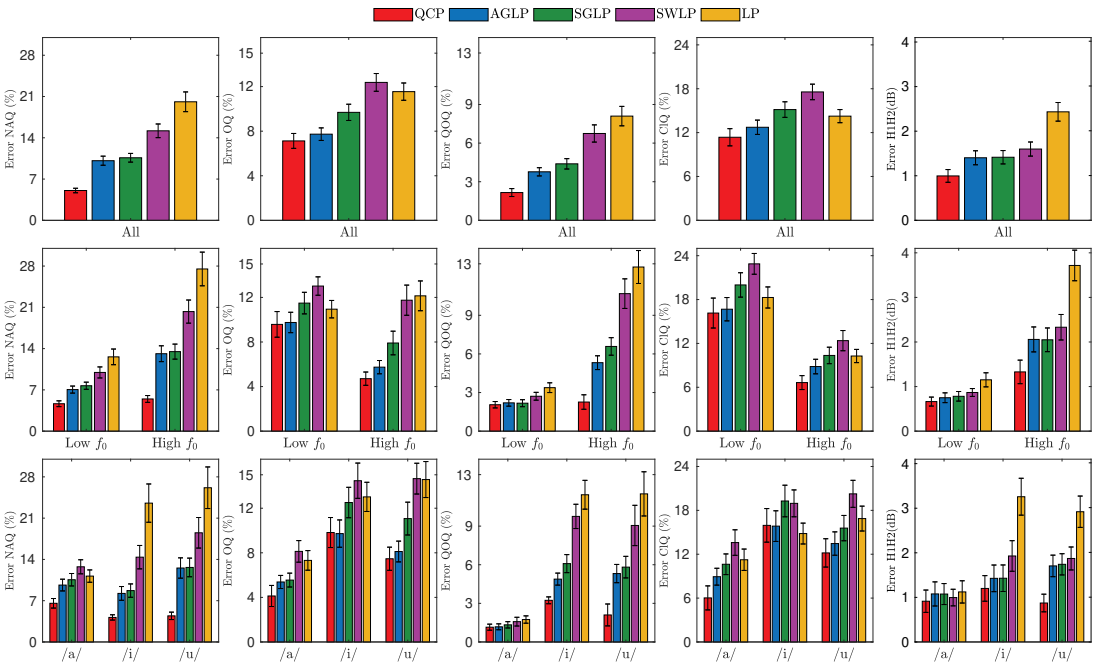

Figure 7: Bar plots of the estimation error for parameters NAQ, OQ, QOQ, ClQ, and H1H2 for the synthesized signals in the Repository II (with the 95% confidence intervals) for five voice inverse filtering methods. The errors are shown according to: all signals (top row), vowels (middle row), and f_0 ranges (bottom row).

Orders P leading to the minimum waveform errors for the glottal function E_{v_g}, are identified, selected, and reported.

Fig. 6 shows box plots of E_{v_g} for different vowels and fundamental frequency ranges (Low: $f_0 < 200$ Hz and High: $f_0 \geq 200$ Hz). As can be seen, classic LP and QCP present, respectively, the worst and best performances among all considered inverse filtering methods. Moreover, the Gaussian attenuation-based methods, SGLP and AGLP, outperform LP and SWLP across all vowel sounds and f_0 ranges. Particularly, the results show that the AGLP method presents lower errors for all categories compared to SGLP. These findings support our assertion that the proposed asymmetric Gaussian attenuation enhances the inverse filtering results, surpassing the performance of the original symmetric alternative.

Concerning the analysis of the aerodynamic parameters, Fig. 7 shows bar plots reporting the estimation errors for parameters NAQ, OQ, QOQ, ClQ, and H1H2 for all signals from Repository II (top row). Additionally, separate bar plots are presented according to the f_0 ranges (middle row) and the vowel sounds (bottom row) to provide a more detailed analysis. Overall, it is evident that QCP consistently exhibits the lowest error across all considered parameters. Also, it can be seen that AGLP and SGLP evidence superior performance relative to LP and SWLP methods. Notably, AGLP outperforms the SGLP method for parameters OQ, QOQ, and ClQ. However, no significant differences are observed for parameters NAQ and H1H2.

The middle row in Fig. 7 describes the estimation error according to low and high f_0 ranges. The results indicate that AGLP performs similarly to QCP in the low f_0 range, except for the NAQ parameter. In contrast, for high f_0 values, QCP exhibits the lowest error across all examined parameters. Furthermore, AGLP outperforms SGLP, SWLP, and LP methods for the OQ, QOQ, and ClQ

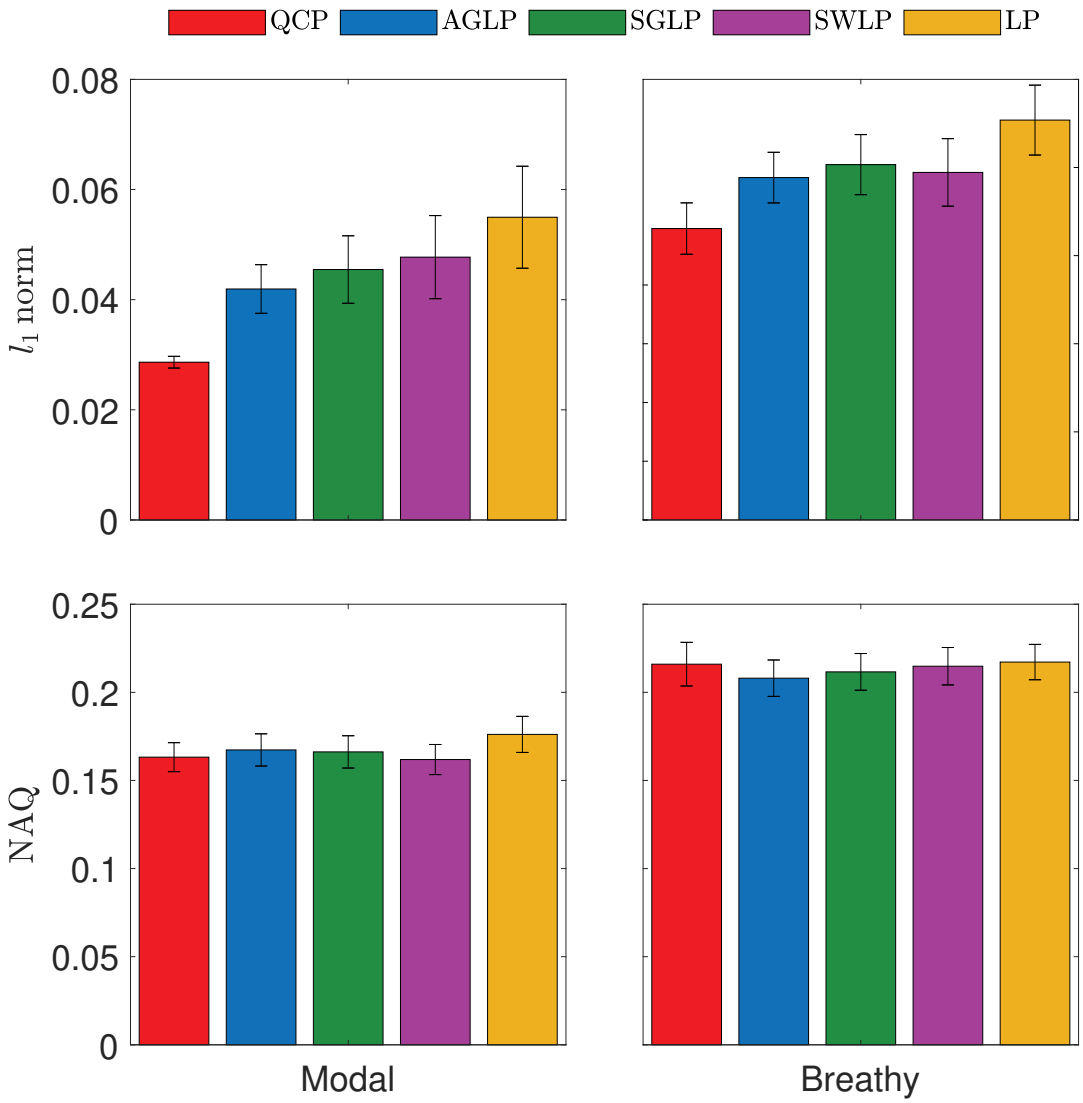

Figure 8: Bar plots of l_1 norm on closed phase and the NAQ estimation error for the natural signals in the Repository IV for five voice inverse filtering methods. The bar plots show the average values (with the 95% confidence intervals) according to modal (left column), and breathy (right column) voice qualities.

parameters, while producing no discernible differences in the estimation error for NAQ and H1H2 parameters compared to SGLP.

Finally, the bottom row in Fig. 7 shows the estimation error according to voice sounds. It is again observed that the QCP method exhibits the best performance across all parameters. However, it can be appreciated that for the vowel /i/, the AGLP method slightly outperforms the QCP method for the OQ and ClQ parameters. Moreover, AGLP and SGLP produce smaller prediction errors than LP and SWLP in most cases.

5.2. Inverse filtering of real voices

The results for natural voice signals from Repository IV are described in this section. Inverse filtering method parameters were adjusted as previously described. For each voice signal, the best results were selected according to the minimum l_1 norm criteria. The first row in Fig. 8 shows bar plots for the average l_1 norm on closed phase for signals having Modal (left column) and Breathy

(right column) voice qualities, for five voice inverse filtering methods. For modal voices, QCP and LP yield respectively the minimum and maximum l_1 norm, with statistically significant differences; the remaining methods show intermediate performances, with AGLP showing a marginal, no significant improvement compared to SGLP and SWLP. Although a similar result is observed for the breathy voices, the differences between methods are much less meaningful. The second row in Fig. 8 depicts bar plots for the average NAQ values obtained from the glottal airflow estimated by five voice inverse filtering methods. There are no observable, significant differences in the NAQ estimates between inverse filtering methods.

Fig. 9 shows examples of the estimated glottal function and glottal airflow obtained from two natural voice signals from Repository IV by applying four inverse filtering methods. All curves are shifted vertically for better visualization. We can observe that the estimates obtained by QCP, AGLP, and SGLP, exhibit decreased spurious oscillations and a flatter closed phase. These features are relevant indicators of suitably estimating the glottal signals through voice inverse filtering [24]. In Fig. 9 we can further see that the estimates obtained by LP display fluctuations and perturbations of great magnitude and a short closed phase for every glottal cycle. These results suggest that the QCP, AGLP, and SGLP methods are able to improve the estimates of the glottal airflow and its time-derivative in voice inverse filtering.

6. Conclusions

This paper studied a voice inverse filtering method based on weighted linear prediction and Gaussian attenuation. This method uses a Gaussian attenuation window centered on the voice GCIs to mitigate the adverse impact of error samples around the maximum glottal excitation instants in the estimates of the vocal tract filter. As a result, this method yields more accurate estimates of the glottal airflow and its time-derivative, the glottal function, obtained by voice inverse filtering.

The parameterization of Gaussian attenuation was revisited to account for the natural variability of voice periodicity. A proportional selection of Gaussian width relative to the fundamental period was proposed, thereby enabling the window to adapt across various voice signals. On the other hand, an asymmetric Gaussian attenuation was proposed to implement a quasi closed phase linear prediction analysis performed by other baseline inverse filtering methods. Finally, based on our simulation results, a parameter set is suggested that facilitates the application of the Gaussian weighted linear prediction methods.

Experimental evaluations illustrated the existence of an optimal window width range, which yields the best outcomes for inverse filtering when compared to classic linear prediction that lacks any signal weighting. Extensive simulations were conducted to identify the optimal parameter set for the symmetrical and asymmetrical Gaussian attenuations. It is worth noting that the proposed methods have a limitation: the definition of both Gaussian attenuation windows depends on knowing the GCI locations. Fortunately, different methods are available for computing the GCIs from the voice signal or other complementary signals of phonation, ensuring the applicability of SGLP and AGLP methods. Furthermore, the robustness of the Gaussian attenuation to inexact estimation of the GCIs [26], makes Gaussian attenuation methods appealing for inverse filtering

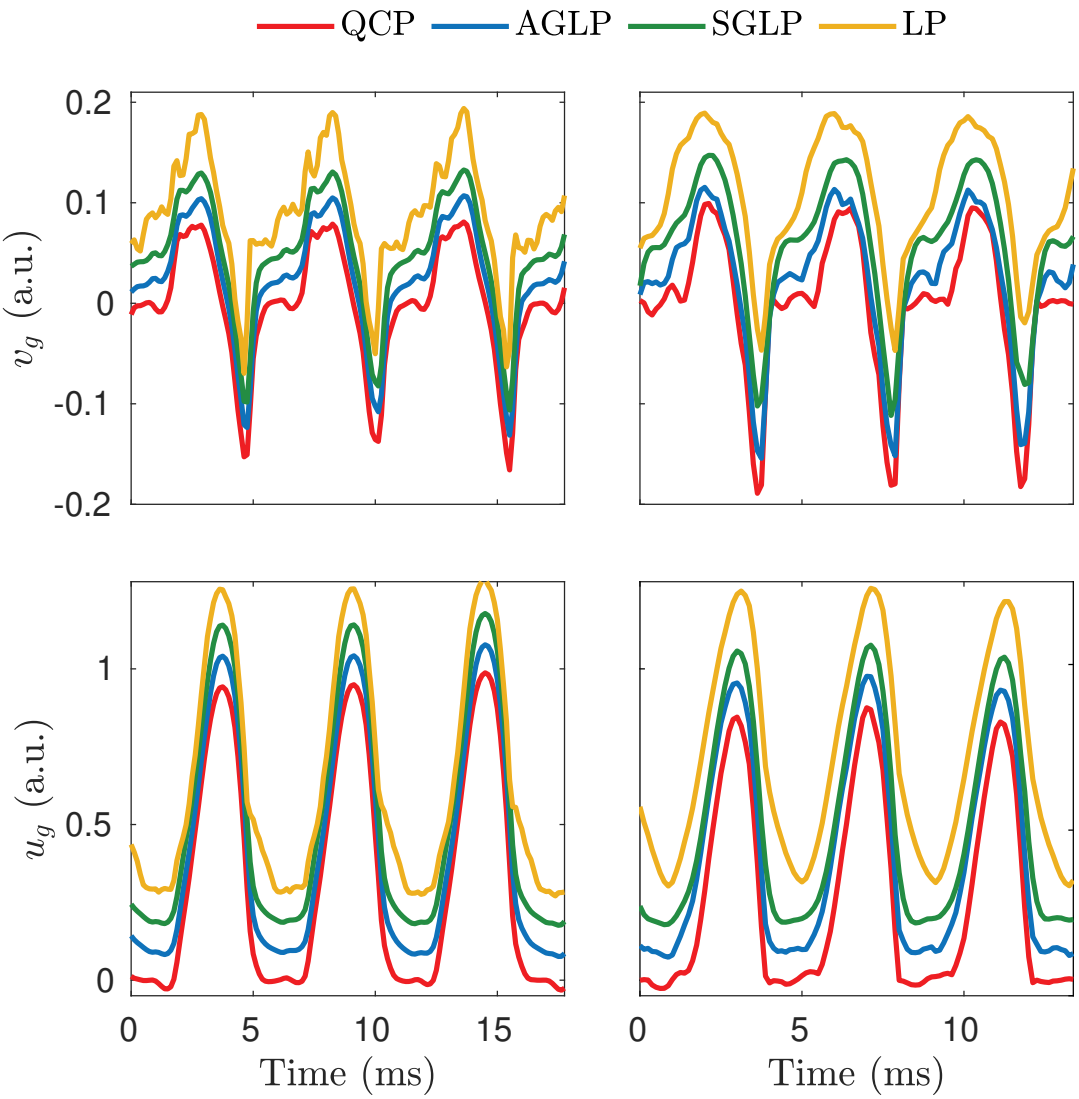

Figure 9: Inverse filtering analysis for natural voice signals with different phonation qualities from the Repository IV. Estimations of the glottal function v_g (top row), and the glottal airflow u_g (bottom row) for four inverse filtering methods are shown. Left column: Modal voice. Right column: Breathy voice.

involving atypical or impaired phonation cases, where the estimation of the GCIs can be difficult.

Upon comparing the symmetric and asymmetric Gaussian weighted linear prediction methods to reference inverse filtering methods, our proposals outperformed some of the reference methods for synthetic and natural voice signals. Nevertheless, the Gaussian strategies produced inferior performance compared to the quasi closed phase linear prediction, i.e., the benchmark method. Despite this lower performance, Gaussian attenuation-based methods can be an interesting alternative for voice inverse filtering practice due to the simplicity of the construction of the attenuation windows used in these strategies.

A future improvement for the Gaussian attenuation windows proposed in this study is the incorporation of a displacement parameter. This parameter would enable windows to shift from their original position, currently centered on the GCIs. We hypothesize that by moving the windows to the closed phase, it is

possible to improve the Gaussian weighted linear prediction results. Following [26], a possibility is to evaluate the impact of shifting the Gaussian attenuation windows and simulated errors in the GCIs on the inverse filtering performance in the OPENGLOT database. However, the proposed improvement introduces an additional parameter in the definition of the Gaussian attenuation windows, which could limit its practical use due to the number of parameters that need to be set.

Appendix A. Maximum window width

The periodicity of the voice signal is the main factor delimiting the allowed maximum width, $\sigma_{\max}$, of the Gaussian functions in Eq.(6). Therefore, the $\sigma_{\max}$ value can be derived from the fundamental period in samples, N_0. Let's consider the sum of two Gaussian functions, g_{s_1} and g_{s_2}, centered at two consecutive GCIs, N_1 and N_2, respectively:

$$g[n] = g_{s_1}[n] + g_{s_2}[n] = e^{-\frac{(n-N_1)^2}{2\sigma_1^2}} + e^{-\frac{(n-N_2)^2}{2\sigma_1^2}}. \tag{A.1}$$

By assuming a periodic voice signal, the functions g_{s_1} and g_{s_2} in Eq. (A.1) are separated by $N_0 = N_2 - N_1$. Let define the width $\sigma_1 = \alpha N_0$ as in Eq. (7), with $0 \leq \alpha \leq \alpha_{\max}$, where $\alpha_{\max}$ represents the value of α for which σ_1 takes its maximum value, i.e., $\sigma_{\max} = \alpha_{\max} N_0$.

When increasing the α value from zero to $\alpha_{\max}$, the width σ_1 also increases, and the Gaussian functions' tails overlap, as shown in Fig. A.1. Due to this overlap, $g[n]$ takes a non-zero value at the midpoint, $\frac{N_2+N_1}{2}$, given by:

$$g\left[\tfrac{N_2+N_1}{2}\right] = e^{-\frac{(N_2-N_1)^2}{2\sigma_1^2}} + e^{-\frac{(N_1-N_2)^2}{2\sigma_1^2}} = A_0. \tag{A.2}$$

We will call A_0 the overlap level in the function $g[n]$ at the midpoint for a given value of α. By considering the definition of σ_1 and N_0, Eq. (A.2) can be written as:

$$2e^{-\frac{N_0^2}{8(\alpha_{\max} N_0)^2}} = A_0, \tag{A.3}$$

Then, we can obtain the $\alpha_{\max}$ value as follows:

$$\alpha_{\max} = \left(\sqrt{8\left[\ln(2) - \ln(A_0)\right]}\right)^{-1}. \tag{A.4}$$

Hence, given a specific overlap A_0, we can compute $\alpha_{\max}$ by Eq. (A.4). For example, setting $A_0 = 0.1$, we obtain $\alpha_{\max} \approx 0.2$, and the range of α will be: $0 \leq \alpha \leq 0.2$.

The previous discussion assumes that the overlap A_0 results from the Gaussian functions centered at two consecutive GCIs, neglecting the contributions of any extra Gaussian function located more than N_0 samples apart from the midpoint.

Appendix B. Normalization process for the attenuation windows

To ensure that the amplitude of the Gaussian attenuation windows falls within the range $[1-\kappa, 1]$, a normalization process is applied for the sum of Gaussian functions $g[n]$. In particular, this normalization is necessary in cases where, due to

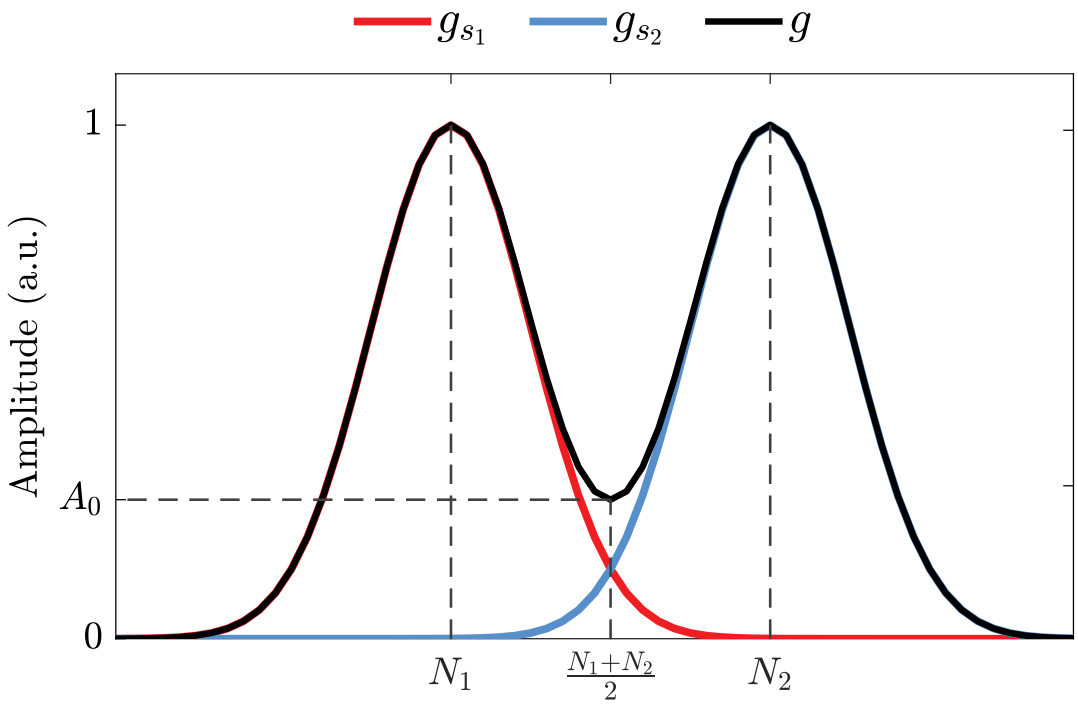

Figure A.1: Example of a function $g[n]$ formed by the sum of two Gaussian functions with the same width σ_1.

the overlapping of the Gaussian tails, $g[n]$ takes a non-zero value at the midpoints of each pair of GCIs (see Fig. A.1).

The normalization process employs the following procedure: First, compute the average value of $g[n]$ at the midpoints of each pair of L-GCIs:

$$\beta = \frac{1}{L-1} \sum_{l=1}^{L-1} g\left[\frac{N_l+N_{l+1}}{2}\right], \tag{B.1}$$

where N_l denotes the discrete location of the l-th GCI. Then, the average value is subtracted from the sum of Gaussian functions, i.e., $g[n] = g[n] - \beta$. Later, for all n such that $g[n] < 0$, set $g[n] = 0$. Finally, $g[n]$ is normalized with respect to its maximum value. As a result, this process guarantees that the amplitude of $g[n]$ oscillates between 0 and 1.

Acknowledgments

This work was financed by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) through project PIP-CONICET 633, the Ministerio de Ciencia, Tecnología e innovación (MINCyT) through projects PICT-ANPCYT 2020 Serie A-01865 and PICT-2021-I-INVI-00122, and the Universidad Nacional de Entre Ríos (UNER) through PID-UNER projects 6224 and 6228.

References

[1] A. C. Singer and M. Feder, "Universal linear prediction by model order weighting," *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2685–2699, 1999.

[2] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. Number 2.

[3] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals*. Institute of Electrical and Electronics Engineers, 2000.

[4] D. O'Shaughnessy, "Review of analysis methods for speech applications," *Speech Communication*, 2023.

[5] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[6] 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*, vol. 36, no. 5, pp. 623–650, 2011.

[7] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Extraction and utilization of excitation information of speech: A review," *Proceedings of the IEEE*, 2021.

[8] A. Palaparthi and I. R. Titze, "Analysis of glottal inverse filtering in the presence of source-filter interaction," *Speech communication*, vol. 123, pp. 98–108, 2020.

[9] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[10] Y. Banaras, A. Javed, and F. Hassan, "Automatic speaker verification and replay attack detection system using novel glottal flow cepstrum coefficients," in *2021 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2021, pp. 149–153.

[11] K. Bharath and M. R. Kumar, "New replay attack detection using iterative adaptive inverse filtering and high frequency band," *Expert Systems with Applications*, vol. 195, p. 116597, 2022.

[12] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, pp. 93–120, 2018.

[13] X. Yao, W. Bai, Y. Ren, X. Liu, and Z. Hui, "Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion," *Neurocomputing*, vol. 410, pp. 328–341, 2020.

[14] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang, and Z. Tao, "Investigation and evaluation of glottal flow waveform for voice pathology detection," *IEEE Access*, vol. 9, pp. 30–44, 2020.

[15] A. B. Aicha and K. Ezzine, "Cancer larynx detection using glottal flow parameters and statistical tools," in *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*. IEEE, 2016, pp. 65–70.

[16] 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*, vol. 134, no. 2, pp. 1295–1313, 2013.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[19] D. Gowda, S. R. Kadiri, B. Story, and P. Alku, "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1901–1914, 2020.

[20] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida, and O. Kakusho, "Analysis of speech signals of short pitch period by a sample-selective linear prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 9, pp. 1233–1240, 1987.

[21] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[22] G. P. Kafentzis, Y. Stylianou, and P. Alku, "Glottal inverse filtering using stabilised weighted linear prediction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5408–5411.

[23] T. Drugman, "Maximum phase modeling for sparse linear prediction of speech," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 185–189, 2014.

[24] 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*, vol. 22, no. 3, pp. 596–607, 2013.

[25] V. Khanagha and K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–9, 2013.

[26] Y.-R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu, and T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1718–1730, 2017.

[27] A. Rao and P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 114–124, 2018.

[28] I. A. Zalazar, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Correntropy-based linear prediction for voice inverse filtering," in *18th International Symposium on Medical Information Processing and Analysis*, vol. 12567. SPIE, 2023, pp. 356–365.

[29] I. A. Zalazar, G. A. Alzamendi, and G. Schlotthauer, "Gaussian-weighted voice inverse filtering: Effects of varying the attenuation window parameters on the glottal source estimation," in *2021 XIX Workshop on Information Processing and Control (RPIC)*. IEEE, 2021, pp. 1–6.

[30] M. Airaksinen, T. Bäckström, and P. Alku, "Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 929–939, 2016.

[31] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *the Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[32] T. Kato, S. Omachi, and H. Aso, "Asymmetric gaussian and its application to pattern recognition," in *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada, August 6–9, 2002 Proceedings*. Springer, 2002, pp. 405–413.

[33] P. Alku, T. Murtola, J. Malinen, J. Kuortti, B. Story, M. Airaksinen, M. Salmi, E. Vilkman, and A. Geneid, "Openglot–an open environment for the evaluation of glottal inverse filtering," *Speech Communication*, vol. 107, pp. 38–47, 2019.

[34] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[35] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *the Journal of the Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[36] T. Hacki, "Classification of glottal dysfunctions on the basis of electroglottography," *Folia Phoniatrica*, vol. 41, no. 1, pp. 43–48, 1989.

[37] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, vol. 2, no. 3, p. 40, 1995.

Anexo B

Maximum Correntropy Linear Prediction for Voice Inverse Filtering: Theoretical Framework and Practical Implementation

Maximum Correntropy Linear Prediction for Voice Inverse Filtering: Theoretical Framework and Practical Implementation

I. A. Zalazar[a], G. A. Alzamendi[a], M. Zañartu[b], G. Schlotthauer[a]

[a] *Institute for Research and Development on Bioengineering and Bioinformatics, CONICET-UNER, Oro Verde, Entre Ríos, Argentina*
[b] *Advanced Center for Electrical and Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso, Chile*

Abstract

Voice inverse filtering methods aim at noninvasively estimating the glottal source information from the voice signal. These inverse filtering strategies typically rely on parametric models and variants of linear prediction for tuning the vocal tract filter. Weighted linear prediction schemes have proved to be the best performing for inverse filtering applications. However, the linear prediction and its variants are sensitive to the impulse-like acoustic excitations triggered by the abrupt glottal closure during voiced phonation. The present study examines the maximum correntropy criterion-based linear prediction (MCLP) for voice inverse filtering. Correntropy is a nonlinear, localized similarity measure inherently insensitive to peak-like outliers. Here, a theoretical framework is established for studying the properties of correntropy relevant for voice inverse filtering and for developing an algorithm to estimate vocal tract filter coefficients. The proposed algorithm results in a robust weighted linear prediction, where a correntropy weighting function is adjusted iteratively by a data-driven optimization scheme. The effects of correntropy kernel parameters on the performance of the MCLP method are analyzed. Characterization of the MCLP method for voice inverse filtering is addressed based on synthetic and natural sustained vowel signals. Simulations show that MCLP naturally overweights samples in the glottal closed phase, where the phonation model is more accurate. MCLP does not require prior information about the glottal instants, nor applying a predefined weighting function. Results show that MCLP performs similarly or better than other well-established inverse filtering methods based on weighted linear prediction.

Keywords: Correntropy, Weighted linear prediction, Voice inverse filtering, Glottal source estimation, Closed phase analysis.

1. Introduction

According to the source-filter theory, human phonation results from the interplay of three simple decoupled components: the glottal airflow source as the acoustic excitation at the glottal level, the vocal tract as an acoustic filter spectrally modulating the different voice sounds, and the lip radiation producing the free-field acoustic pressure [1, 2]. Voice inverse filtering, in turn, addresses the inverse problem of phonation, aiming at noninvasively estimating the acoustic

excitation underlying the voiced phonation through digitally processing the voice signal [3]. First, tunable filters are adjusted from the voice signal to match the main vocal tract resonance frequencies and the lip radiation effect [4, 5]. Then, inverse versions of these filters are applied to cancel out the contributions of the supraglottic structures in the voice signal, which results in an estimate of the glottal airflow [3, 6]. Therefore, airflow estimates from voice inverse filtering depend on suitably modeling the vocal tract resonances and radiation at the lips [7].

Inverse filtering typically uses a first-order causal differentiator filter to approximate the lip radiation effect [4, 2]. Vocal tract contribution, in turn, is modeled as an autoregressive filter with transfer function [8]:

$$V(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}}, \tag{1}$$

where P is the filter order, and a_k are the filter coefficients [2]. The baseline standard for optimally computing the vocal tract filter from a voice signal is the linear prediction (LP) method [9]. In the traditional LP scheme, the filter coefficients in Eq. (1) are computed by minimizing the mean squared prediction error [2]; theoretically, this is optimal only for zero-mean, white Gaussian prediction errors [10]. However, in the voiced phonation context, LP scheme is suboptimal because successive glottal closure instants (GCIs) trigger high-energy acoustic bursts in the glottal excitation, which in turn give rise to prediction errors with exceptionally high local amplitudes [8]. Additionally, it is well known that the mean squared error function is highly susceptible to outlier data [11]. Therefore, the LP method is prone to provide vocal tract filters featuring biased formant frequencies and inaccurate bandwidths [8].

Variants of classic LP have been proposed and applied to improve the vocal tract filter estimation. Weighted LP strategies are specifically addressed here since they have demonstrated the best performance in inverse filtering applications [12, 13, 14]. These schemes take advantage of a weighting function conceived upon empirical criteria or prior knowledge to regulate the relative importance of the prediction error samples in the least squares problem [15]. For example, closed phase covariance (CPC) [16] and quasi closed phase linear prediction (QCP) [17] are two strategies inspired in the time-derivative of the glottal airflow, the so-called glottal function. These methods use a weighting function to emphasize the voice samples in the closed phase (relative to those in the open phase), thereby mitigating the detrimental effects in the vocal tract estimation of the GCIs and acoustic coupling between the sub- and supraglottal systems. However, a drawback of these methods is their inherent reliance on the location of the glottal opening and closing instants for the weighting functions implementation. Consequently, any error in the instants' location affects QCP and CPC performance. Vocal tract filter estimates obtained by QCP may exhibit a residual spectral tilt that affects the estimated glottal signals [18]. To address this issue, a QCP-based method with spectral tilt compensation (QCP-ST) has been recently proposed to enhance the glottal signals estimated using QCP by transferring the estimated tilt from the vocal tract filter [13]. Other methods obtain the weighting functions

in a data-driven manner, i.e., built upon information extracted from the voice signal under consideration. For example, the weighted linear prediction (WLP) [19] and its stabilized version (SWLP) [20] make use of the short-time energy computed from the voice signal as a weighting function. The energy function yields large amplitude values in the samples with a higher signal-to-noise ratio. Thus, this approach increases the robustness of LP analysis to additive noise. However, the resulting weighting function tends to center around the GCIs, yielding a suboptimal estimation of the vocal tract filter [8]. More recently, the probabilistic weighted linear prediction (PWLP) scheme [12] has been developed, which employs a data-driven adaptive function that automatically centers the linear prediction analysis on the closed phase. PWLP employs a probabilistic interpretation of the weighted LP scheme based on the likelihood of voice signal and different model parameters priors, and Gibbs sampling is applied to generate estimates of the LP coefficients from the posterior distribution. This method has the disadvantage of relying on different prior probability distributions for the model parameters, which makes the physiological interpretation challenging.

All weighted LP strategies seek to overcome the inherent problems of the mean squared prediction error function. Instead, LP schemes based on alternative cost functions with well-documented properties have not received equivalent attention. For example, as the mean absolute error function is less sensitive to large-amplitude errors, it becomes an appealing cost function for developing a sparse linear prediction model for speech processing [10]. In turn, the discrete all-pole modeling applies the Itakura-Saito distortion measure to better fit the vocal tract filter in Eq. (1) to a small set of spectral points [21]. Here, we study the correntropy with Gaussian kernel as a cost function to extend the LP scheme.

Correntropy with Gaussian kernel is a robust, nonlinear similarity measure that proved to be a suitable cost function for linear models because of its robustness to non-Gaussian impulse-like prediction errors [22]. Recently, in a conference paper [23], we took advantage of this alternative cost function to develop the maximum correntropy-based LP (MCLP), a data-driven weighted LP scheme that iteratively adjusts both the weighting function and the vocal tract filter coefficients. We showed that MCLP is capable of overweighting the closed phase samples without requiring prior knowledge of glottal instants. Correntropy has been used for processing voice signals (e.g., for fundamental frequency estimation [24], speech enhancement [25], and speech recognition [26]); however, only in [23] has been applied in the context of voice inverse filtering, to the best of our knowledge. The present work extends the ideas in [23] in the following aspects. First, the attention is paid to the theoretical properties of the correntropy measure relevant to the LP model and for voice inverse filtering. This sets up a framework to study the MCLP method in detail, particularly the effects of Gaussian kernel parameters on the local weighting imposed on LP prediction errors. Second, a computational algorithm is proposed for adjusting the vocal tract filter coefficients based on the step-like update of the correntropy kernel parameter. Third, performance of MCLP for voice inverse filtering is thoroughly evaluated in synthetic and natural phonation data using different evaluation metrics. Analysis of the estimated vocal tract filters and glottal waveforms for synthetic signals indicates that MCLP outperforms QCP and QCP-ST, and achieves a comparable performance to PWLP at a significantly reduced computational time. Results for

natural signals show that MCLP yields glottal waveforms characterized by a flatter closed phase with fewer spurious oscillations associated with inverse filtering errors.

The organization of the paper is as follows. Section 2 analyzes the correntropy measure with Gaussian kernel, and addresses the MCLP method. Section 3 describes the voice inverse filtering data and the experimental setup. Section 4 provides a detailed analysis of the proposed method for voice inverse filtering. Section 5 presents the results and discusses their relevance. Finally, we deliver the conclusions of this work in Section 6.

2. Maximum correntropy criterion for linear prediction

2.1. Correntropy with Gaussian kernel

Correntropy with Gaussian kernel is a nonlinear similarity measure between two arbitrary scalar random variables X and Y, given by [27]:

$$V(X,Y) = \mathcal{E}\left\{G_\sigma(X-Y)\right\}, \tag{2}$$

where $\mathcal{E}\{\cdot\}$ denotes the expectation operator, and $G_\sigma(\cdot)$ is the Gaussian kernel with kernel size σ defined as:

$$G_\sigma(X-Y) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X-Y)^2}{2\sigma^2}\right). \tag{3}$$

Kernel G_σ is positive-definite and symmetric about the origin. An alternative expression of the Eq. (2) can be obtained using the finite-sample estimator of the expectation operator [22] :

$$\hat{V}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(x_i - y_i). \tag{4}$$

where N is the number of samples available.

Correntropy assesses the probability of how similar two random variables are with respect to a neighborhood of the joint space determined by the kernel size σ [22]. This proves to be very useful in reducing the detrimental effects of outliers from gross differences between X and Y [28]. Correntropy is endowed with a strong theoretical framework; here, we present the most relevant properties in the context of linear prediction [27, 29]:

Property 1: Correntropy is positive-definite and bounded, i.e., $0 < V(X,Y) \leq 1/(\sqrt{2\pi}\sigma)$. Additionally, it reaches its maximum value if $X = Y$. This property guarantees the existence of an optimal solution for the LP model.

Property 2: Correntropy involves all the even moments of the difference between X and Y:

$$V(X,Y) = \frac{1}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n n!}\mathcal{E}\left\{\frac{(X-Y)^{2n}}{\sigma^{2n}}\right\}. \tag{5}$$

Compared to the mean squared error (second-order statistical moment) used in LP and its variants, correntropy includes statistical information on second and higher-order moments. Notice that, as σ increases, the high-order moments decay

faster and the second-order moment dominates, thus approaching to the mean squared error.

Property 3: Assume an i.i.d. data sample $\{(x_i, y_i)\}_{i\in N}$ drawn from the joint probability density function $f_{X,Y}(x, y)$. The Parzen estimator with kernel size σ of the probability density function of the error samples $e_i = x_i - y_i$ is defined as $\hat{f}_\sigma(e)$. Then, the correntropy measure $\hat{V}(X, Y)$ is the value of $\hat{f}_\sigma(e)$ evaluated at the point $e = 0$, i.e, $\hat{V}(X, Y) = \hat{f}_\sigma(0)$ [22]. Since $f_\sigma(0) = p(X = Y)$ [30], maximizing the correntropy in the context of LP leads to an increase in the probability that the predicted and desired signals are equal.

Property 4: Express correntropy of X and Y as a cost function:

$$J = V(X, Y), \tag{6}$$

J is concave in the range of $[-\sigma, \sigma]$. The concavity guarantees the existence and uniqueness of the optimal solution in problems built around maximizing the correntropy cost function [29].

Property 5: Correntropy, as a sample estimator, induces a metric in the sample space with important geometric characteristics. As explained in [22], the induced metric changes progressively from L_2 norm-like in the Euclidean zone (where samples X and Y are close) to L_1 norm-like in the transition zone away from the Euclidean zone to eventually approaching L_0 norm and becoming insensitive to distance in the rectification zone (where two points are further apart). This explains the inherent robustness of the correntropy measure against outlier samples. Furthermore, the kernel bandwidth σ controls the size of each zone [22]. Then, increasing σ produces a large Euclidean zone and reduces the rectification zone, whereas decreasing σ has the opposite effect.

2.2. Linear prediction based on maximum correntropy criterion

Correntropy has proved to be a suitable cost function for information theoretic learning, establishing the foundations for the maximum correntropy criterion in estimation-related problems [31, 29]. In the same vein, we proposed in [23] the maximum correntropy criterion-based LP (MCLP) method. In this section, we describe in detail the derivation of MCLP based on the correntropy properties discussed previously.

According to the LP model of phonation, the voice signal, $s[n]$, for index time $1 \leq n \leq N$, follows an autoregressive process [2]:

$$s[n] = \mathbf{a}^T \mathbf{s}[n] + e[n], \tag{7}$$

where $\mathbf{a} = \big[a_1, a_2, \ldots, a_P\big]^T$ are the LP coefficients, with P being the model order, $\mathbf{s}[n] = \big[s[n-1], s[n-2], \ldots, s[n-P]\big]^T$ is a vector gathering the P last past voice samples, and $e[n]$ is the prediction error. In classic LP, vector $\mathbf{a}$ is typically obtained by minimizing the mean squared prediction error in Eq. (7) [9]. Instead, as proposed in [23], we consider correntropy as a cost function of the prediction error, e, in the computation of the coefficients $\mathbf{a}$ of Eq. (7). Based on properties 1, 3, and 4 mentioned above, the optimal coefficients $\mathbf{a}$ can be determined by maximizing the correntropy with Gaussian kernel of the difference between the voice signal and the LP model in Eq. (7):

$$
\begin{aligned}
J &= \mathcal{E}\left\{G_\sigma(e[n])\right\} \\
&= \mathcal{E}\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(s[n]-\mathbf{a}^T\mathbf{s}[n]\right)^2}{2\sigma^2}\right)\right\}.
\end{aligned} \tag{8}
$$

This method for computing the LP coefficients is called maximum correntropy criterion based-LP (MCLP). Maximum of Eq. (8) is obtained by setting $\frac{\partial J}{\partial \mathbf{a}} = 0$, which yields:

$$
\mathcal{E}\left\{\exp\left(-\frac{\left(s[n]-\mathbf{a}^T\mathbf{s}[n]\right)^2}{2\sigma^2}\right)\left(s[n]-\mathbf{a}^T\mathbf{s}[n]\right)\mathbf{s}[n]\right\} = 0. \tag{9}
$$

Considering property 3 and assuming that the prediction error is a discrete-time stationary stochastic process, then we can use in Eq. (9) the sample estimator of expectation operator [32], obtaining:

$$
\frac{1}{N}\sum_{n=1}^{N} h_e[n]\big(s[n]-\mathbf{a}^T\mathbf{s}[n]\big)\mathbf{s}[n] = 0, \tag{10}
$$

where, according to the weighted LP scheme, $h_e[n]$ takes the form of a positive weighting function:

$$
h_e[n] = \exp\left(-\frac{\left(s[n]-\mathbf{a}^T\mathbf{s}[n]\right)^2}{2\sigma^2}\right). \tag{11}
$$

After some manipulation in Eq. (10), coefficient vector $\mathbf{a}$ becomes:

$$
\begin{aligned}
\mathbf{a} &= \left[\sum_{n=1}^{N} h_e[n]\mathbf{s}[n]\mathbf{s}[n]^T\right]^{-1}\left[\sum_{n=1}^{N} h_e[n]s[n]\mathbf{s}[n]\right] \\
&= \mathbf{R}_h^{-1}\mathbf{r}_h,
\end{aligned} \tag{12}
$$

where $\mathbf{r}_h$ and $\mathbf{R}_h$ represent the weighted estimates of the correlation vector and the correlation matrix, respectively. As can be seen, Eq. (12) is in the form of the *Wiener-Hopf* solution; two key differences are worth noting, though [28]:

- A time-domain weighting function, $h_e[n]$, is applied from the Gaussian kernel with size σ. By $h_e[n]$, small errors with respect to σ are emphasized, whereas large amplitude errors, i.e., significantly higher than σ, are underweighted (as explained in property 5). Thus, the computation of coefficient vector $\mathbf{a}$ from Eq. (8) is robust against impulse-like prediction errors.

- Equation (12) is not a closed-form solution. Computation of the $\mathbf{a}$ vector involves the weighting function $h_e[n]$, which in turn depends on coefficients $\mathbf{a}$ according to Eq. (11).

Following the work by Singh and Principe [28], an iterative method is used based on assuming that the optimal solution $\mathbf{a}^*$ is a fixed point of Eq. (12). Thus, given an initial coefficient vector, $\mathbf{a}_0$, the solution is obtained through the following fixed-point iteration [28]:

$$
\mathbf{a}_{k+1} = \left[\mathbf{R}_h(\mathbf{a}_k)\right]^{-1}\mathbf{r}_h(\mathbf{a}_k), \tag{13}
$$

Algorithm 1 Computation of LP coefficient vector $\mathbf{a}$ based on MCLP.

Initialization: $\mathbf{a}_0$, σ, ϵ_1, ϵ_2
Pre-emphasis and normalization of $s[n]$
$k = 0$
do
  Compute: $h_e[n]$, $\mathbf{R}_h$, $\mathbf{r}_h$
  $\mathbf{a}_{k+1} = \mathbf{R}_h^{-1}\mathbf{r}_h$
  $e_{k+1}[n] = s[n] - \mathbf{a}_{k+1}^T\mathbf{s}[n]$
  if $(||\mathbf{a}_{k+1} - \mathbf{a}_k||_2^2 < \epsilon_1)$ **then**
    Compute: σ_S
  end
  $k = k + 1$
while $(||\mathbf{a}_{k+1} - \mathbf{a}_k||_2^2 > \epsilon_2)$;
return $(\mathbf{a}_{k+1})$

where $\mathbf{a}_k$ denotes the solution for iteration $k = 0, 1, 2, \ldots$. A condition guaranteeing the convergence of this fixed-point solution was provided by property 5.

Algorithm 1 describes the computation of fixed-point solution in Eq. (13). First, the coefficients $\mathbf{a}_0$ are initialized by classic LP. On the other hand, the kernel size in Eq. (11) is a free parameter that must be chosen by the user using concepts of density estimation [22]. In our case, σ is determined from the prediction error using Silverman's rule [33]:

$$\sigma_S = 1.06\sigma_e N^{-1/5}, \tag{14}$$

where σ_e stands for the minimum between standard deviation and interquartile range scaled by 1.34 extracted from prediction error computed by Eq. (7) using the coefficients $\mathbf{a}_0$. Thresholds, ϵ_1 and ϵ_2, are defined so that ϵ_1 is much greater than ϵ_2. Finally, a pre-emphasis filter is applied to the voice signal, s, and the resulting signal is then normalized to its maximum absolute value. In the second stage of Algorithm 1, the LP coefficients $\mathbf{a}_k$ are computed iteratively according to Eq. (13). If the difference between two consecutive solutions is less than the threshold ϵ_1, the kernel size σ is then updated using Silverman's rule of Eq. (14) from the last estimate of the prediction error. Steplike update of σ improves the vocal tract filter coefficients estimation, compared to updating σ at each iteration or keeping it constant. Finally, the computation stops when the difference between two consecutive solutions is less than ϵ_2.

3. Materials and Methods

3.1. Phonation data for inverse filtering evaluation

The OPENGLOT database [34], a comprehensive data set on glottal excitation during phonation, was considered to benchmark the inverse filtering methods. In this work, only repositories II and IV were considered.

Repository II contains synthetic signals obtained from a physical model of phonation that includes vocal fold kinematics, aero-acoustic interactions at the glottal level, pressure wave propagation through the vocal tract, and voice sound radiation at the lips. This repository of synthetic or theoretical signals includes

the voice signal (radiated pressure) s, the glottal airflow (volume velocity) u_g, and its time-derivative (volume acceleration) v_g. The latter will be subsequently referred to as the "glottal function". Therefore, a direct evaluation of voice inverse filtering methods is possible for these signals. The phonation data correspond to three vowels: [/a/, /i/, /u/], with four fundamental frequency cases: 82, 110, 156, and 220 Hz for male speakers, and 175, 196, 220, and 294 Hz for female speakers. The sampling frequency of the signals is $f_s = 44.1$ kHz.

In contrast, repository IV contains voice and electroglottogram signals from the natural production of vowel sounds at modal and breathy phonation qualities. Additionally, this repository contains high-speed video of the vocal folds. However, inverse filtering methods only consider voice signals, so this information is not taken into account in the present work. The signals correspond to five male and five female speakers that produce a vowel sound with three different pitch levels (low, medium, and high). The sampling frequency of the signals is $f_s = 44.1$ kHz. Glottal airflow and vocal tract information is not provided in this repository. So, a direct comparison with a ground truth signal is impossible.

3.2. Methodology and reference methods

Simulations were conducted to assess the performance of the MCLP method for voice inverse filtering based on different descriptors for the estimated glottal airflow and its time derivative. Three voice inverse filtering methods were considered for comparison: PWLP [12], QCP [17], and QCP-ST [13]. QCP-based approaches have become benchmarks for voice inverse filtering methods due to its good performance; as recent studies [14, 13] have shown that other methods in the literature, such as LP, WLP, and SWLP, exhibit low performance compared to QCP methods, they are thus excluded from the present study. On the other hand, PWLP is a suitable inverse filtering method requiring no prior detection of glottal instants, since the weighting function is obtained in a fully data-driven manner similar to MCLP.

Following the hypothesis supporting the source-filter theory [1], all signals used in the simulations were resampled at 8 kHz. Inverse filtering analysis was performed for every 50 ms non-overlapping segment of the voice signal. A pre-emphasis filtering, with transfer function $R(z) = 1 - 0.99z^{-1}$, was applied to enhance the higher frequencies on the voice signal before computing the vocal tract filter with each voice inverse filtering method.

Vocal tract filters computed through MCLP, QCP, QCP-ST, and PWLP were processed to remove the poles with positive real parts, and magnitude less than 0.8 before applying it for voice inverse filtering. To correctly represent the vocal tract formant frequencies, the poles of the transfer function (see Eq. (1)) in the z-plane must be close to the unit circle [1]. Poles that do not satisfy this condition would provide poor inverse filtering estimates [16].

3.3. Performance measures

Direct evaluation of voice inverse filter estimates is only possible for synthetic data from repository II. As a direct performance metric, a normalized waveform error of the glottal function is defined as:

$$E_{v_g} = \frac{m_e}{\mathrm{RMS}(v_g)}, \tag{15}$$

where m_e denotes the median value of the absolute waveform error between the theoretical glottal function, v_g, and its voice inverse filtering estimate, $\hat{v}_g$, given by:

$$e_{v_g}[n] = \left| v_g[n] - \hat{v}_g[n] \right|, \text{ for } n = 1, 2, \ldots, N. \tag{16}$$

Normalization by the RMS value of the theoretical glottal function is applied. The signals in Eqs. (16) are time-aligned to compensate for any phase delay due to the acoustic wave propagation along the vocal tract, or the applied inverse filtering method. Furthermore, the voice inverse filtering estimates $\hat{v}_g$ are normalized by the orthogonal projection proposed in [35].

Another metric used to evaluate the performance is the l_1 norm, i.e., the sum of the absolute values of the estimated glottal airflow samples on the closed phase. This measure quantifies the waveform flatness in the closed phase, which is a looked-for feature in the estimated glottal airflow, $\hat{u}_g$ [7]; a large value of l_1 norm indicates the presence of spurious oscillations or an incomplete closed phase, resulting from inadequate inverse filtering [36].

In addition to the normalized waveform error E_{v_g} and the l_1 norm, the inverse filtering evaluation of the MCLP method was based on five aerodynamic parameters computed from the estimated glottal airflow and its time-derivative: Normalized Amplitude Quotient (NAQ), Quasi-Open Quotient (QOQ), Closing Quotient (ClQ), the difference between first two harmonic (H1H2), and the Spectral tilt (ST)[1] [4, 5, 13]. Estimation errors for NAQ, QOQ, and ClQ are reported as average absolute relative errors. For example, for the NAQ parameter:

$$\text{Error NAQ} = \mathcal{E}\left\{ \frac{\left| \text{NAQ}_{\text{ref}} - \text{NAQ}_{\text{est}} \right|}{\text{NAQ}_{\text{ref}}} \right\}, \tag{17}$$

where NAQ_{ref} and NAQ_{est} are the NAQ values from the theoretical glottal airflow and the inverse filtered estimate, respectively. Errors for the parameters QOQ, and ClQ are computed similarly to Eq. (17). On the other hand, the error for the parameters H1H2 and ST are reported through the average absolute errors:

$$\text{Error H1H2} = \mathcal{E}\left\{ \left| \text{H1H2}_{\text{ref}} - \text{H1H2}_{\text{est}} \right| \right\}, \tag{18}$$

$$\text{Error ST} = \mathcal{E}\left\{ \left| \text{ST}_{\text{ref}} - \text{ST}_{\text{est}} \right| \right\}. \tag{19}$$

For the natural signals in repository IV, the MCLP method is assessed similarly as in [36, 12]. A first evaluation measures the l_1 norm on the closed phase of the estimated glottal airflow. Another measure to evaluate the MCLP method for the natural data is the NAQ parameter, where the results from QCP were considered the benchmark values. As QCP has proved to be one of the best-suited voice inverse filtering methods [37], we assume that any methods producing similar results as those obtained by QCP could be considered a befitting choice for analyzing natural voice signals.

3.4. Experimental setup

Model order P and the threshold ϵ_1 in Algorithm 1 were explored in the simulations involving the repository II. The following sets were considered: $P \in$

[1] Computed following the code in [13].

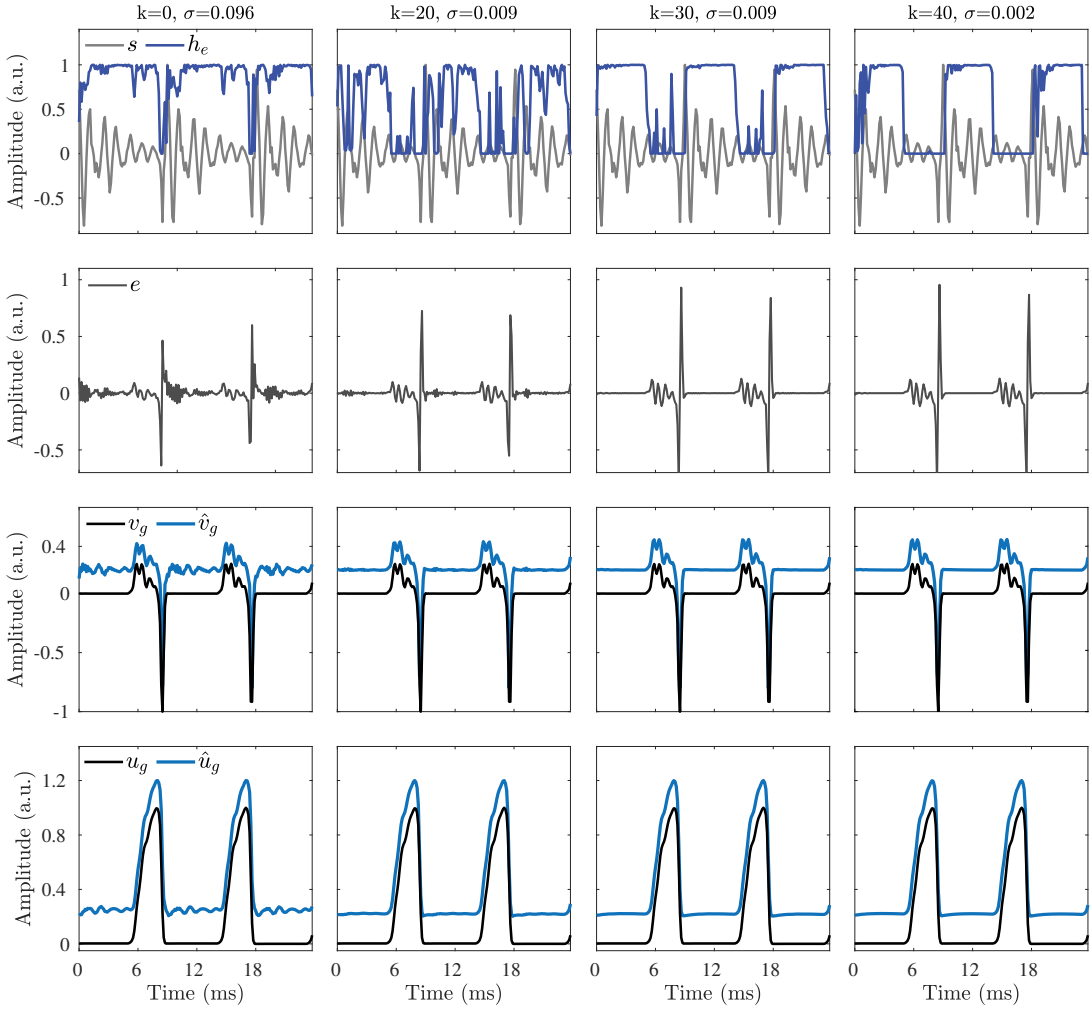

Figure 1: Iterative adjustment of MCLP in a synthetic signal example for different iterations k of the Algorithm 1. First row: voice signal s and resulting weighting function h_e. Second row: prediction error e. Third row: estimated glottal function $\hat{v}_g$ together with the theoretical waveform v_g. Fourth row: theoretical glottal airflow u_g and its resulting estimation $\hat{u}_g$. For better visualization, the estimated $\hat{u}_g$ and $\hat{v}_g$ are shifted vertically in the third and fourth rows, respectively.

$(8, 9, \ldots, 12)$, and $\epsilon_1 \in (0.000001, 0.00001, \ldots, 0.1, 1)$. To fulfill the constraint $\epsilon_1 >> \epsilon_2$ we set $\epsilon_2 = 0.1\epsilon_1$. From P, ϵ_1, and ϵ_2, the vocal tract filter coefficients were computed using Algorithm 1, and the estimated glottal function $\hat{v}_g$ was obtained through inverse filtering. Finally, the parameter combination yielding the lowest average normalized waveform error, E_{v_g}, for each signal was selected.

For QCP and PWLP methods, its parameters are set up based on the values suggested in [17, 12]. Spectral tilt compensation for QCP-ST was performed as described in [13]. Additionally, QCP methods requires prior knowledge about the GCI locations; here, two different strategies are employed. For the synthetic signals in repository II, where the theoretical glottal data are available, the GCIs were determined from the location of the minimum value in the glottal function, v_g, for each glottal cycle. Before determining the GCIs, the voice signal and the glottal function were time-aligned to avoid any phase lag. On the other hand, for the natural voices, the GCIs and the closed phase were estimated from the electroglottogram signal using the SIGMA algorithm [38]. Time alignment

between the voice and electroglottogram signals was performed to avoid any delay generated during the recording of the signals. For all cases, lag compensation was computed via cross-correlation measure [2].

4. MCLP analysis

This section discusses some interesting features of the MCLP method based on the simulations for the synthetic signals from repository II.

4.1. Iterative adjustment of MCLP

In MCLP, the linear prediction coefficients are iteratively computed based on the prediction error, e, and the resulting weighting function h_e (see Eq. (12)). Fig. 1 illustrates the iterative adjustment of MCLP variables for a synthesized signal from the repository II. Columns in Fig. 1 represent different iterations k of Algorithm 1 for a kernel size σ computed by Silverman's rule in Eq. (14). The first row of Fig. 1 displays the voice signal, s, and the weighting function, h_e. The second row shows the prediction error. Finally, the third and fourth rows of Fig. 1 show the theoretical waveform of the glottal airflow, u_g, and its time-derivative, v_g, together with the estimates, $\hat{u}_g$ and $\hat{v}_g$, obtained by using the vocal tract filter coefficients computed in each iteration k. Note that all estimated signals are shifted vertically for better visualization.

The proposed MCLP method becomes an iterative weighted LP scheme that performs a data-driven closed-phase voice inverse filtering analysis over multiple glottal cycles. As shown in Fig. 1, the prediction error becomes sparser during the iterations, i.e., evident local error diminutions are observed, especially in the closed phase segments, while releasing the spike-like components characteristic of the maximum glottal excitations. The changes in the prediction error are the results of updating the kernel size σ, the weighting function h_e, and the LP coefficients $\mathbf{a}$. Algorithm 1 applies a steplike update of kernel size based on prediction errors, where the increases in the sparse level of the prediction error produce a diminution in the σ value computed by Eq. (14). In addition, h_e overweights the glottal closed phase, i.e., open-phase samples with high amplitude prediction error progressively receive significantly lesser weights than the closed-phase ones (see Fig. 1). So, the iterative adjustment of the vocal tract filter becomes increasingly meaningful, as it is based mostly on information from the closed phase, where the effect of the maximum glottal excitation is minimal [17]. The iterative adjustment of the vocal tract filter leads to more accurate estimates of the glottal airflow and its time-derivative obtained by voice inverse filtering. The third and fourth rows in Fig. 1 illustrate the progressive improvement in the estimates $\hat{v}_g$ and $\hat{u}_g$ across the successive iterations, compared to the theoretical waveforms.

4.2. Correntropy kernel size update

The update of the kernel size σ is a worth analyzing factor affecting the performance of Algorithm 1. Fig. 2 describes three different strategies for updating σ from the voice signal of Fig. 1. The analyzed strategies are: initialize σ using Silverman's rule, perform no update and keep it fixed (in green color), update σ in each iteration by Silverman's rule (in red color), and steplike update σ as proposed in Algorithm 1 (in blue color). Fig. 2 shows the σ value (left panel), the

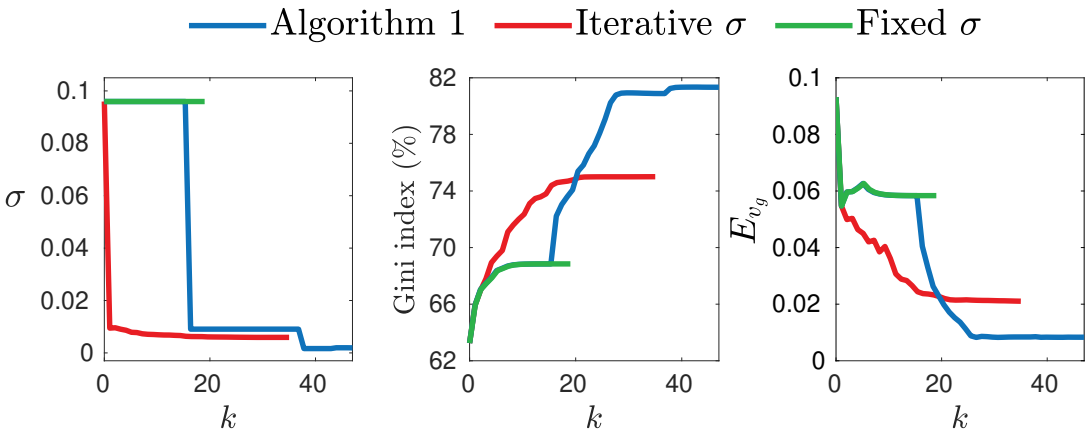

Figure 2: Analysis of different σ update strategies in a synthetic signal from repository II. Left: σ values for each iteration. Middle: Gini index of the prediction error. Right: normalized waveform error of the glottal function.

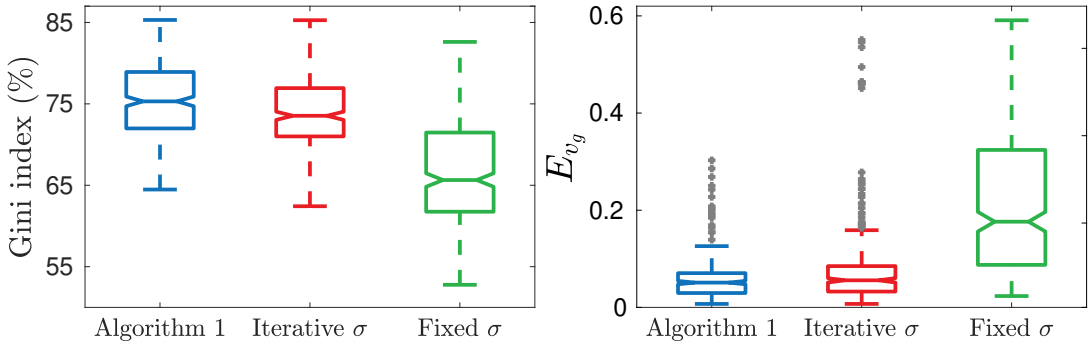

Figure 3: Boxplots for the three alternatives for σ update obtained from the voice signals in the repository II. Left: Gini index for the prediction error. Right: normalized waveform errors E_{v_g}.

Gini index of the prediction error (middle panel), and the normalized waveform error of the glottal function E_{v_g} (right panel) for successive iterations k. The Gini index allows quantifying the sparsity level of the prediction error [39, 11]. As explained above, a correct estimation of the coefficients $\mathbf{a}$ yields sparse prediction errors and, consequently, a high Gini index. On the contrary, a low Gini index indicates less sparse prediction errors resulting from an imprecise adjustment of the LP model.

As shown in Fig. 2, for fixed σ, the computation of the MCLP model converges faster (fewer iterations are required); however, it results in the lowest Gini index, indicating non-sparse prediction errors, and in the highest waveform error E_{v_g}, characteristic of a deficient voice inverse filtering. In the case of updating the kernel size in each iteration, a significant decrease in σ takes place, and a better fitting of the MCLP model results, as suggested by a higher Gini index and a lower waveform error, at the expense of requiring more iterations until convergence. However, the best results (showing the topmost Gini index and minimum waveform error) are obtained by the steplike update of σ, as in Algorithm 1. This strategy seeks to iteratively tune the LP model leveraged by a smooth-changing weighting function that gradually adjusts to the glottal closed phase (see Fig. 1), thus allowing a closed phase weighted linear prediction analysis that improves

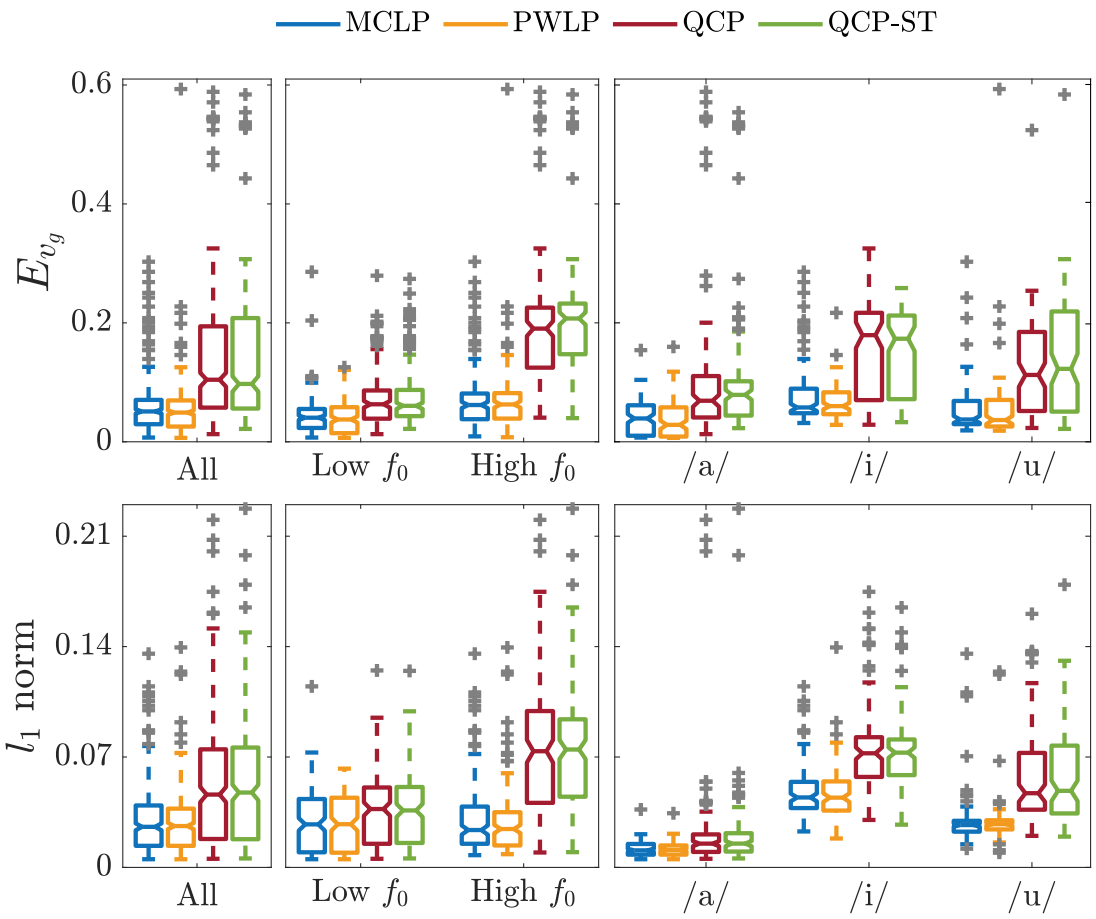

Figure 4: Boxplots of the normalized waveform error E_{v_g} (top) and l_1 norm on closed phase of the estimated glottal airflow (bottom) for synthetic signals from repository II. Left: all signals. Middle: high and low f_0 ranges. Right: vowel sounds.

the voice inverse filtering estimations.

To further study the update of the kernel size, MCLP schemes considering the three update strategies described above were applied for processing all voice signals from repository II, and the performance of the resulting LP models was assessed. Fig. 3 depicts boxplots for the Gini index of the prediction error (left panel) and the normalized waveform error E_{v_g} (right panel). As expected, a deficient model adjustment (e.g., smallest Gini index and highest waveform error overall) results from keeping kernel size fixed. However, when updating σ the steplike strategy of Algorithm 1 yields a significant improvement (e.g., elevated Gini index and reduced estimation error) for LP model tuning. Statistical analysis based on the Friedman test with the Bonferroni correction shows that the Gini indices are different at a significance level greater than 95%. Similarly, normalized waveform errors are also significantly different for all three update strategies.

Based on our simulations for the synthetic signals from repository II, the best parameter combination for Algorithm 1 is $P = 12$, $\epsilon_1 = 0.0001$, and $\epsilon_2 = 0.00001$.

5. Results

This section studies the performance of the MCLP method for inverse filtering the voice signals in repositories II and IV. MCLP is compared against QCP, QCP-

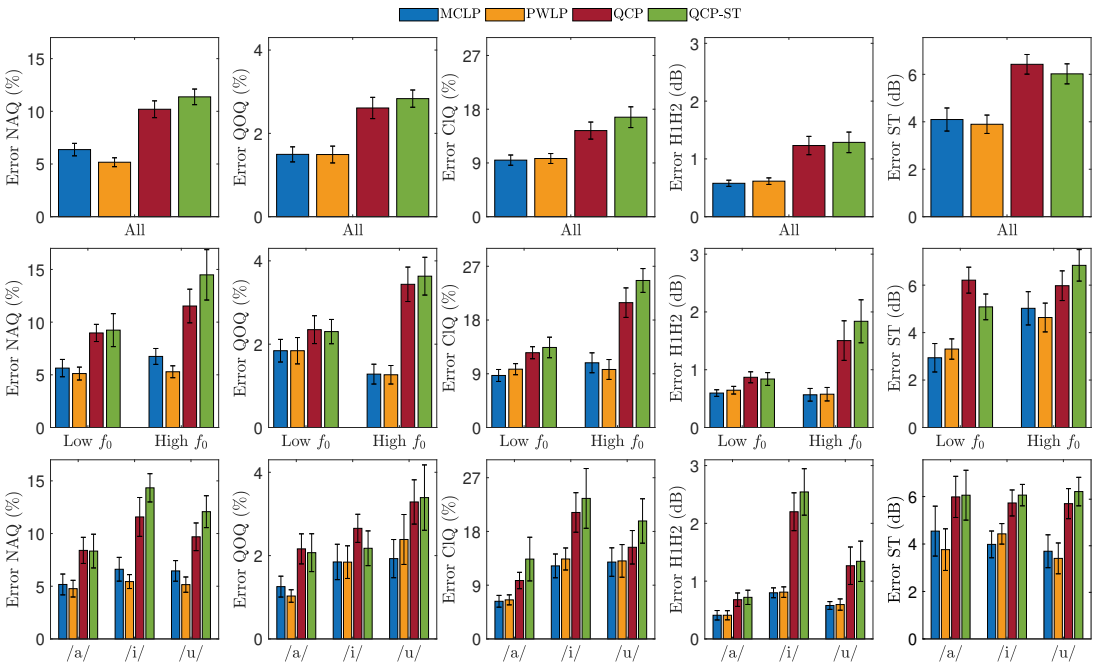

Figure 5: Bar plots of the estimation error for parameters NAQ, QOQ, ClQ, H1H2, and ST for the synthetic signals from repository II (with the 95% confidence intervals) for different voice inverse filtering methods. The errors are shown according to: all signals (top row), f_0 ranges (middle row), and vowel sounds (bottom row).

ST, and PWLP. All voice signals were processed considering $P = 12$ for the vocal tract filters. For the MCLP method, the thresholds for Algorithm 1 were set to $\epsilon_1 = 0.0001$ and $\epsilon_2 = 0.00001$.

5.1. Synthetic voice data

The performance for estimating the glottal function in repository II is addressed. The top row in Fig. 4 shows boxplots for the normalized waveform error E_{v_g} for different vowels and fundamental frequency ranges (Low: $f_0 < 200$ Hz and High: $f_0 \geq 200$ Hz). As can be seen, the data-driven methods, PWLP and MCLP, presented the lowest waveform errors, whereas the QCP-based approaches showed the highest errors for all categories considered. The error values showed statistically significant differences greater than 95% between the MCLP and PWLP methods compared to the QCP-based approaches, while no significant differences were found between QCP and QCP-ST error values. However, for vowels /i/ and /u/, and for the low f_0 range, no statistically significant differences were obtained between the MCLP and PWLP errors. Likewise, the bottom row in Fig. 4 shows boxplots for the l_1 norm on closed phase of the estimated glottal airflow for the same categories. As can be seen, MCLP and PWLP exhibited low l_1 norm values for all categories. According to Friedman test with the Bonferroni correction, statistically significant differences were obtained between l_1 norms from QCP-based approaches and those from MCLP and PWLP for all categories.

Fig. 5 shows bar plots reporting the estimation errors for parameters NAQ, QOQ, ClQ, H1H2, and ST for all signals from repository II (top row). Additionally, separate bar plots are presented according to the f_0 ranges (middle row) and the vowel sounds (bottom row) to provide a more detailed description. As can be

Table 1: Computational time dispersion for adjusting vocal tract filter coefficients from voice signals in repository II. Time is reported in seconds for the lower quartile (Q_1), median (Q_2), and upper quartile (Q_3).

	Q_1	Q_2	Q_3
QCP	$2.48\text{x}10^{-4}$	$2.71\text{x}10^{-4}$	$2.88\text{x}10^{-4}$
QCP-ST	$7.02\text{x}10^{-4}$	$7.28\text{x}10^{-4}$	$7.68\text{x}10^{-4}$
MCLP	$1.68\text{x}10^{-2}$	$2.27\text{x}10^{-2}$	$3.13\text{x}10^{-2}$
PWLP	$1.29\text{x}10^{0}$	$1.42\text{x}10^{0}$	$1.51\text{x}10^{0}$

seen in the top row, MCLP and PWLP methods show superior performance for all parameters in contrast to QCP-based approaches. In particular, no statistically significant differences were obtained between MCLP and PWLP errors for the time-domain parameters, QOQ and ClQ, and the frequency-domain parameters, H1H2 and ST. Instead, PWLP yields lower NAQ errors than MCLP, QCP, and QCP-ST. Results for f_0 ranges indicate that MCLP shows similar performance to PWLP. However, MCLP showed inferior performance for the NAQ parameter in high f_0 cases. These results coincide with those reported in [12], where PWLP shows a lower NAQ error than QCP for physical model-based synthetic signals with high f_0. Finally, similar results were observed for vowel analysis, where the MCLP method obtained comparable results to PWLP, except for the NAQ parameter estimation.

The computational burden of each inverse filtering method was also analyzed based on the time taken to compute the vocal tract filter coefficients from voice signal segments from repository II. Table 1 reports the lower quartile (Q_1), median (Q_2), and upper quartile (Q_3) in seconds describing the dispersion of the computational times for QCP, QCP-ST, PWLP, and MCLP methods[2]. Simulations show that MCLP requires, on average, a 60 times less computational time than PWLP. This differences are due to PWLP being a computationally expensive method compared to MCLP. Table 1 also shows that the QCP method requires the lowest computational time, as expected, since it applies a simple, closed-form rule for computing vocal tract filter coefficients. Instead, QCP-ST demands more computations than QCP because of the processing of the spectral tilt in the estimated vocal tract filter.

5.2. Natural voice data

The results for natural voice signals from repository IV are described in this section. The first column in Fig. 6 shows bar plots for the average l_1 norm on the closed phase of the estimated glottal airflow for four voice inverse filtering methods. As can be seen, MCLP showed the minimum l_1 norm values, where statistically significant differences with l_1 norms from PWLP and QCP-based methods were found. On the other hand, the second column in Fig. 6 depicts bar plots for the average NAQ values obtained from the estimated glottal airflow waveform. No statistically significant differences were found in the estimated

[2]Computational times for all considered methods were measured in the same desktop computer with a Core i7-12700, and 32 GB of RAM.

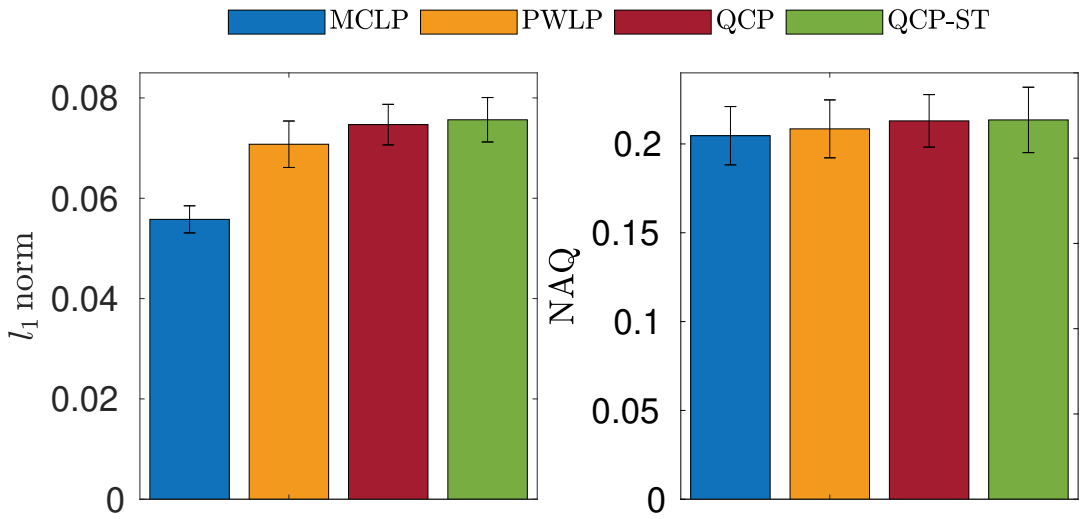

Figure 6: Bar plots of the l_1 norm on the closed phase and NAQ values in the estimated glottal airflows from natural signals in repository IV by four voice inverse filtering methods. The bar plots show the average values with 95% confidence intervals.

NAQ values for the four inverse filtering methods. These findings are similar to those reported in [12] for natural voices, where PWLP demonstrated superior performance over QCP for the l_1 norm while achieving similar NAQ values.

Fig. 7 shows examples of estimated waveforms for the glottal function (top rows) and the glottal airflow (bottom rows) obtained from a modal and a breathy natural voice of repository IV using four inverse filtering methods. All curves are shifted vertically for better visualization. We can observe that the estimates obtained by all methods present small spurious oscillations in the closed phase, which indicates suitably inverse filtering [17]. The results in Fig. 7 indicate that MCLP can provide glottal airflow and glottal function estimates comparable to well-established state-of-the-art inverse filtering methods. However, MCLP estimates exhibit a flatter closed phase, which aligns with the lower l_1 norm values show in Fig. 6. This feature is highly desirable in estimated glottal waveforms [5, 36, 34].

6. Discussion and future works

Results for signals from OPENGLOT indicate that MCLP and PWLP perform similarly, with both surpassing the QCP-based methods. However, our findings on synthetic signals indicate that PWLP outperforms MCLP regarding the NAQ parameter. The performance differences between MCLP and PWLP can be attributed in part to the strategies employed for computing the vocal tract filter coefficients. PWLP better explores the solution space by means of numerous estimations of the filter coefficients randomly generated using the Gibbs sampling method, increasing thus the likelihood of identifying an optimal set of coefficients at the expense of a significant increase in computational complexity. On the other hand, MCLP computes the vocal tract filter sequentially following the proposed algorithm, making the final estimation sensitive to the thresholds, ϵ_1 and ϵ_2, and the initial coefficients, $\mathbf{a}_0$. For initializing the vocal tract filter, classic LP method is applied in this work, which has been shown to perform poorly

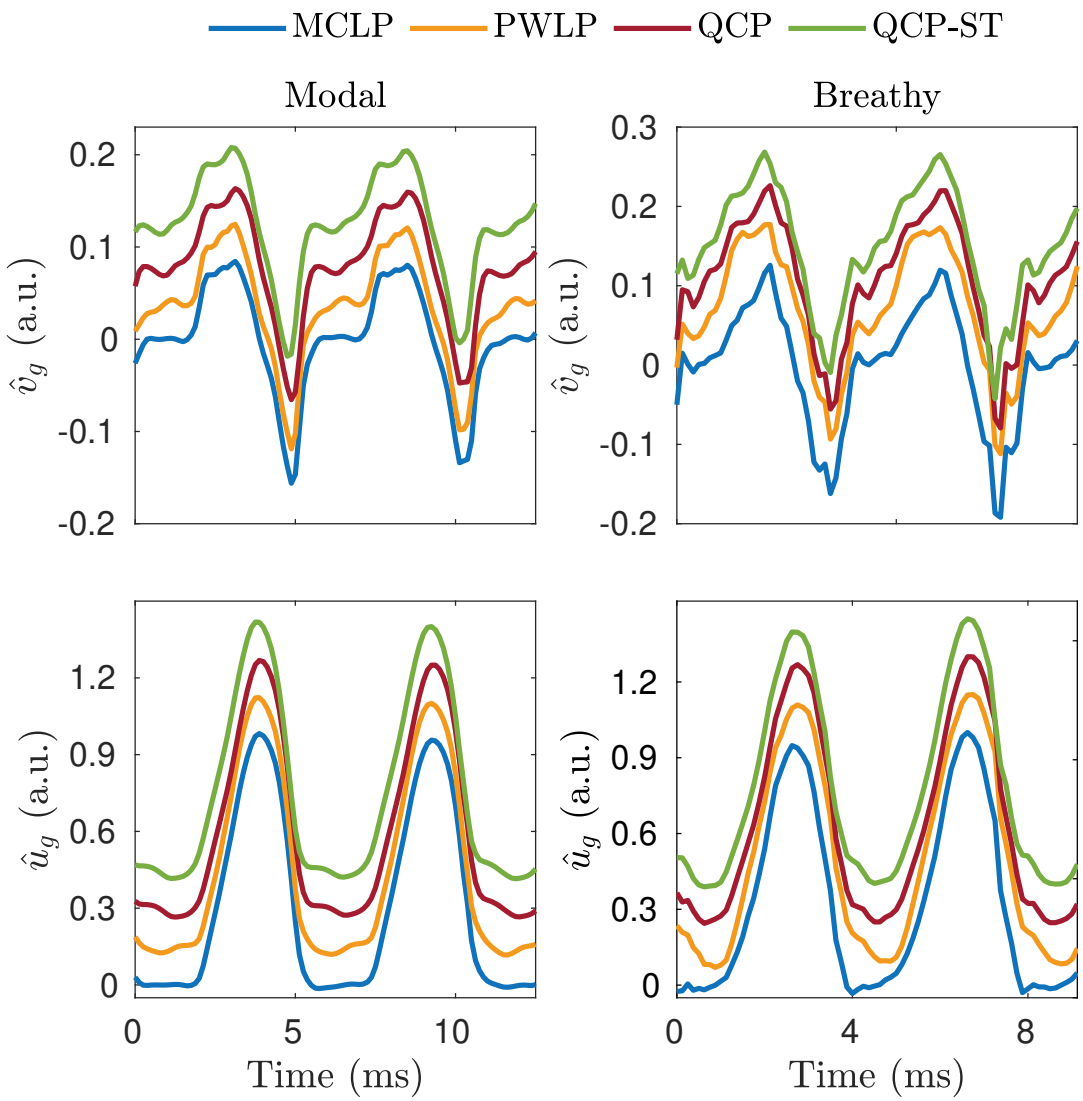

Figure 7: Inverse filtering analysis for natural voice signals with different phonation qualities from repository IV. Estimations of the glottal function $\hat{v}_g$ (top row), and the glottal airflow $\hat{u}_g$ (bottom row) for four inverse filtering methods are shown. All estimated signals are shifted vertically for better visualization. Left column: Modal voice. Right column: Breathy voice.

for voice inverse filtering, specifically for high fundamental frequency phonations. Moreover, the proposed algorithm does not guarantee obtaining the best global estimation of the vocal tract filter coefficients. Another possible reason could be the suggested PWLP hyper-parameters. Here, the parameters for the MCLP method are derived by minimizing the normalized waveform error, E_{v_g}. In contrast, the recommended hyper-parameters for PWLP detailed in [12] (the same considered in our simulations) were determined by minimizing the NAQ error; thus, they would provide the conditions for better performances in terms of NAQ error.

The results by MCLP are promising. However, in future work, we aim to validate the performance of the proposed method on other phonation datasets including increasingly challenging scenarios involving different levels of source-filter interaction, continuous speech, and atypical phonations. [6, 35, 40].

Lastly, it is important to emphasize that the present effort was limited to the

inverse filtering of voice signals; future avenues may study the use of MCLP for inverse filtering schemes based on alternative phonation signals, or for other signal processing applications. The correntropy cost function could also be applied in other inverse filtering methods. For example, an unconventional method was proposed in [41] for addressing the automatic glottal inverse filtering from the voice signal spectrogram, combining weighted linear prediction, time-frequency representations, and non-negative matrix factorization. Time-frequency analysis techniques, such as the spectrogram, provide a better description of the non-stationary dynamics in a signal (for more details, see [42, 43]). Moreover, the non-negative matrix factorization yields a part-based low-rank decomposition of the spectrogram in two components corresponding to the glottal acoustic excitation and the encompassing samples nearby, respectively. Maximum correntropy criterion could thus be applied to enhance the weighted linear prediction (as proposed here) and the non-negative matrix factorization [44, 45].

7. Conclusions

The present study investigated weighted linear prediction based on the maximum correntropy criterion, referred to as MCLP in this study, as a novel method for voice signal analysis. The method involves maximizing the correntropy, a nonlinear similarity measure suitable for optimization problems and digital signal processing. In this work, we discussed the theoretical features of the correntropy with the Gaussian kernel that are relevant in the context of voice inverse filtering and glottal source estimation. Incorporating the maximum correntropy criterion in the linear prediction inverse filtering scheme provides a robust solution against the detrimental effects of impulse-like acoustic excitations during voiced phonation on vocal tract model adjustment.

Utilizing the maximum correntropy criterion, a solution for the MCLP (Maximum Correntropy Linear Prediction) coefficients can be obtained through an iterative fixed-point approach. In this work, we proposed an algorithm for this task based on the effect of the kernel size in the correntropy measure. The kernel size is a critical factor in MCLP model adjustment. We explored three strategies for managing the kernel size and found that using a steplike update of kernel size through Silverman's rule boosts the prediction error sparsity, especially on the closed phase, and inverse filtering performance. We also showed that the MCLP method implements an iterative weighted linear prediction analysis based on a data-driven weighting function emphasizing the closed phase region.

The performance of MCLP for voice inverse filtering was assessed based on the waveform error, the l_1 norm on the closed phase, and the estimation of aerodynamic parameters extracted from the estimated glottal waveforms. The results in a benchmark synthetic dataset showed that MCLP performs similarly or better than other inverse filtering methods well-established in the literature, the probabilistic weighted linear prediction, the quasi closed phase analysis and the quasi closed phase analysis with spectral tilt compensation, respectively. MCLP also showed a proper performance in simulations involving the inverse filtering of natural voices, resulting in smooth glottal airflow estimates with a flatter waveform on the closed phases.

MCLP presents a series of advantages that make it attractive for voice inverse filtering: 1) it is robust to the impulse-like acoustic excitations that impacts

negatively on the estimated vocal tract filter; 2) it employs a weighting function that automatically adjusts in a data-driven manner; 3) it does not require prior information about the glottal instants; 4) the computation of vocal tract filter coefficients requires a low to moderate computational burden. However, MCLP has the disadvantage of not guaranteeing the estimation of the optimal vocal tract coefficients, since the proposed algorithm is sensitive to the initial conditions.

Acknowledgments

This work was financed by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) through project PIP-CONICET 633, the Ministerio de Ciencia, Tecnología e innovación (MINCyT) through projects PICT-ANPCYT 2020 Serie A-01865 and PICT-2021-I-INVI-00122, and the Universidad Nacional de Entre Ríos (UNER) through PID-UNER projects 6224 and 6228. Research in this work was also supported by the NIDCD of the NIH under Award No. P50DC015446, and ANID BASAL AFB240002. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] G. Fant, *Acoustic theory of speech production*. Mounton, The Hague, The Netherlands, 1970, no. Number 2.

[2] J. R. Deller, J. G. Proakis, and J. H. Hansen, "Discrete-time processing of speech signals." Institute of Electrical and Electronics Engineers, 2000.

[3] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Extraction and utilization of excitation information of speech: A review," *Proceedings of the IEEE*, 2021.

[4] 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*, vol. 36, no. 5, pp. 623–650, 2011.

[5] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[6] A. Palaparthi and I. R. Titze, "Analysis of glottal inverse filtering in the presence of source-filter interaction," *Speech communication*, vol. 123, pp. 98–108, 2020.

[7] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] 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*, vol. 134, no. 2, pp. 1295–1313, 2013.

[9] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[10] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[11] T. Drugman, "Maximum phase modeling for sparse linear prediction of speech," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 185–189, 2014.

[12] A. Rao and P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 114–124, 2018.

[13] M. Freixes, L. Joglar-Ongay, J. C. Socoró, and F. Alías-Pujol, "Evaluation of glottal inverse filtering techniques on openglot synthetic male and female vowels," *Applied Sciences*, vol. 13, no. 15, p. 8775, 2023.

[14] I. A. Zalazar, G. A. Alzamendi, and G. Schlotthauer, "Symmetric and asymmetric gaussian weighted linear prediction for voice inverse filtering," *Speech Communication*, vol. 159, p. 103057, 2024.

[15] Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida, and O. Kakusho, "Analysis of speech signals of short pitch period by a sample-selective linear prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 9, pp. 1233–1240, 1987.

[16] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *the Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[17] 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*, vol. 22, no. 3, pp. 596–607, 2013.

[18] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17 230–17 246, 2019.

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

[20] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[21] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on signal processing*, vol. 39, no. 2, pp. 411–423, 1991.

[22] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-gaussian signal processing," *IEEE Transactions on signal processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[23] I. A. Zalazar, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Correntropy-based linear prediction for voice inverse filtering," in *18th International Symposium on Medical Information Processing and Analysis*, vol. 12567. SPIE, 2023, pp. 356–365.

[24] J.-W. Xu and J. C. Principe, "A pitch detector based on a generalized correlation function," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 8, pp. 1420–1432, 2008.

[25] X. Cui, Z. Chen, F. Yin, and X. Xu, "Correntropy-based multi-objective multichannel speech enhancement," *Circuits, Systems, and Signal Processing*, vol. 41, no. 9, pp. 4998–5025, 2022.

[26] R. Singh and J. C. Principe, "Correntropy based hierarchical linear dynamical system for speech recognition," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[27] I. Santamaría, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187–2197, 2006.

[28] A. Singh and J. C. Principe, "A closed form recursive solution for maximum correntropy training," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 2070–2073.

[29] S. Zhao, B. Chen, and J. C. Principe, "Kernel adaptive filtering with maximum correntropy criterion," in *The 2011 International Joint Conference on Neural Networks*. IEEE, 2011, pp. 2012–2017.

[30] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[31] A. Singh and J. C. Principe, "Using correntropy as a cost function in linear adaptive filters," in *2009 International Joint Conference on Neural Networks*. IEEE, 2009, pp. 2950–2955.

[32] D. G. Manolakis, V. K. Ingle, and S. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. McGraw-Hill, 2000.

[33] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[34] P. Alku, T. Murtola, J. Malinen, J. Kuortti, B. Story, M. Airaksinen, M. Salmi, E. Vilkman, and A. Geneid, "Openglot–an open environment for the evaluation of glottal inverse filtering," *Speech Communication*, vol. 107, pp. 38–47, 2019.

[35] Y.-R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu, and T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1718–1730, 2017.

[36] M. Airaksinen, T. Bäckström, and P. Alku, "Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 929–939, 2016.

[37] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.

[38] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[39] N. Hurley and S. Rickard, "Comparing measures of sparsity," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, 2009.

[40] I. Langheinrich, S. Stone, X. Zhang, and P. Birkholz, "Glottal inverse filtering based on articulatory synthesis and deep learning." in *INTERSPEECH*, 2022, pp. 1327–1331.

[41] M. Airaksinen, L. Juvela, T. Bäckström, and P. Alku, "Automatic glottal inverse filtering with non-negative matrix factorization." in *Interspeech*, 2016, pp. 1039–1043.

[42] S. Mallat, "A wavelet tour of signal processing," 1999.

[43] R. B. Pachori, *Time-frequency analysis techniques and their applications*. CRC Press, 2023.

[44] J. J.-Y. Wang, X. Wang, and X. Gao, "Non-negative matrix factorization by maximizing correntropy for cancer clustering," *BMC bioinformatics*, vol. 14, pp. 1–11, 2013.

[45] S. Peng, W. Ser, Z. Lin, and B. Chen, "Robust sparse nonnegative matrix factorization based on maximum correntropy criterion," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2018, pp. 1–5.

Anexo C

Regularized adaptive non-harmonic model for glottal airflow estimation in glottal inverse filtering

Regularized adaptive non-harmonic model for glottal airflow estimation in glottal inverse filtering

I. A. Zalazar[a], G. A. Alzamendi[a], J. V. Ruiz[a,b], M. A. Colominas[a], G. Schlotthauer[a]

[a]*Institute for Research and Development on Bioengineering and Bioinformatics, CONICET-UNER, Oro Verde, Entre Ríos, Argentina*
[b]*Department of Psychiatry, New York University School of Medicine, New York, USA*

Abstract

Glottal inverse filtering enables the estimation of glottal airflow from voice recordings. It involves eliminating the vocal tract effects from the voice signal, thereby obtaining the glottal function. Then, the glottal airflow is estimated by processing the glottal function to compensate for any waveform modulation introduced by the lip radiation process. The accuracy of the estimations critically depends on correctly removing the contributions of the vocal tract and lip radiation. Most methods primarily focus on canceling out the vocal tract filter contribution, often overlooking the impact of lip radiation. Typically, a standard leaky integrator filter is applied to eliminate the lip radiation effects; however, the coarse tuning of the filter response can cause low-frequency distortions in the estimated airflow waveform, particularly noticeable during the glottal closed phase. The present study proposes a regularized Adaptive Non-Harmonic (RANH) model of the glottal airflow. The RANH model is designed for indirect model adjustment from the glottal function using a least-squares optimization scheme with closed phase regularization, which jointly promotes an overall representation of the glottal airflow waveform and a waveform flatness during the closed phase. An automatic approach for the RANH model application is derived that, unlike the leaky integrator filter, requires no user-controlled parameters. The results for both synthetic and natural voice signals indicate that the RANH model outperforms state-of-the-art methods regarding estimation error and waveform flatness in closed phase regions.

Keywords: Glottal inverse filtering, Glottal airflow estimation, Adaptive non-harmonic model, Wave-shape function, Lip radiation cancellation, Closed phase analysis.

1. Introduction

Glottal inverse filtering methods enable the estimation of glottal information from the voice signal. Supported by the *source–filter* theory for voiced phonation, most inverse filtering methods recover the glottal airflow through a sequential process [1]. First, two digital filters, denoted henceforth as $H(z)$ and $R(z)$, are adjusted to represent the acoustic effects of the vocal tract and the lip radiation, respectively. Next, the vocal tract contribution is canceled out from the voice signal by applying the inverse filter $H(z)^{-1}$. The resulting signal, known as

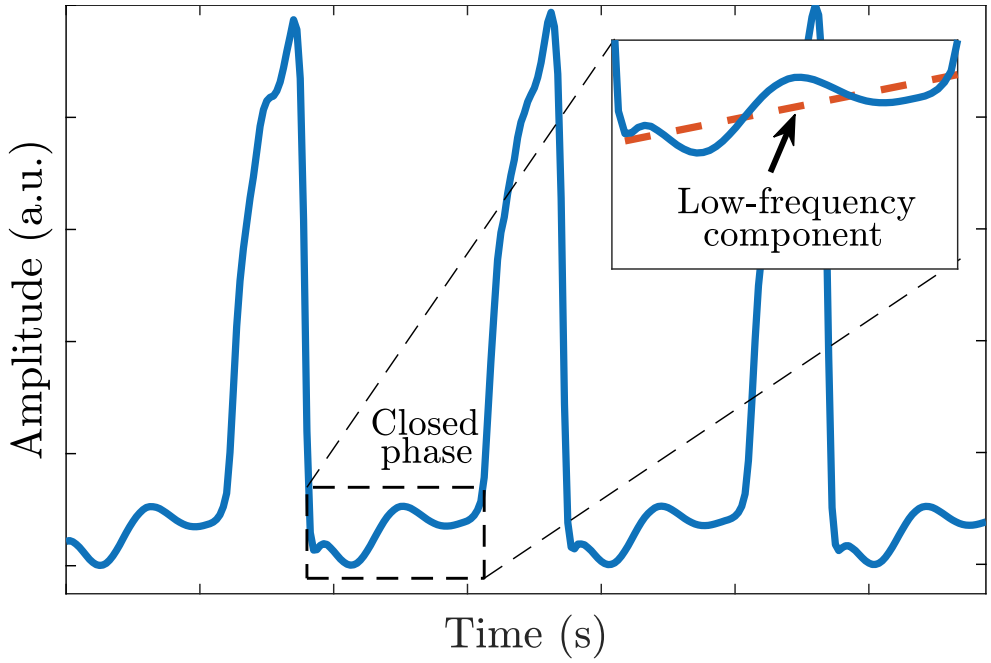

Figure 1: Example of closed phase distortions in an estimated glottal airflow signal. The inset panel illustrates the spurious oscillations caused by an inexact vocal tract inverse filter and a low-frequency component resulting from an inadequate leaky integrator filter.

the glottal function, describes how the airflow at the lips transforms into the radiated acoustic pressure [2]. Thus, an estimate of the glottal airflow is obtained by eliminating the lip radiation effects from the glottal function using the inverse filter $R(z)^{-1}$ [3].

Conventionally, the contribution of the vocal tract is modeled as an autoregressive filter [2]:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} d_k z^{-k}}, \tag{1}$$

where P is the filter order and d_k are the filter coefficients [4]. Ensuring a reliable vocal tract filter is crucial for effective glottal inverse filtering, as inaccuracies in $H(z)$ can severely distort the glottal airflow waveform. Distortions due to improperly fitting of the vocal tract filter typically result in undesirable oscillations during the glottal closed phase, as shown in Fig. 1. Several computational methods have been developed for improving the computation of the vocal tract filter coefficients [5]. The baseline is the linear prediction method [6]; however, variants of the linear prediction method have been developed to enhance the estimation [7, 8, 9, 10, 11, 12]. Among these it is worth mentioning the quasi closed phase linear prediction analysis [13], a well-accepted method for voice inverse filtering applications. It applies a weighted scheme that specifically minimizes the prediction error in the glottal closed phase.

The lip radiation effect can typically be modeled as a first-order differentiator filter $R(z) = 1 - \alpha z^{-1}$, with $0 < \alpha \leq 1$ a user-tunable parameter. Due to the differentiating effect of lip radiation, the glottal function is proportional to the

time-derivative of the glottal airflow. Therefore, the leaky integrator filter (LIF):

$$R(z)^{-1} = \frac{1}{1 - \alpha z^{-1}}, \tag{2}$$

is applied in most glottal inverse filtering schemes to cancel out the lip radiation effect from the glottal function [3]. LIF is simple and computationally efficient. Its performance, however, is sensitive to the α value [3]. An incorrect α can lead to a distorted glottal airflow waveform, particularly in the closed phase regions [14]. The distortions typically manifest as ascending/descending low-frequency components, instead of a smooth, flat waveform consistent with a closed glottis configuration (see Fig. 1). The origin of these distortions arises partly from the combined effect of vocal tract filter and LIF in the range of the very low frequencies. Accurately capturing the spectral information below 200 Hz is challenging when adjusting the vocal tract model (1) [3]. Thus, the application of the inverse vocal tract filter yields low-frequency distortions in the glottal function, which are further amplified by the LIF method, resulting in an unexpected trend in the estimated glottal airflow waveform. A technique to eliminate undesirable low-frequency components is high-pass filtering of the glottal airflow using a linear phase filter with a cutoff frequency of 70 Hz [15]. However, this filter can unintentionally distort the glottal airflow waveform.

Improving the estimation of the glottal airflow typically requires the manual tuning of α in the LIF through trial and error, which can be a tedious and laborious task. A common criterion for α is to select the value that yields the flattest closed phase response in the estimated glottal airflow waveform [14]. In [3], an automatic strategy to find α is introduced. The selection criterion consists in detecting the significant changes in the area of the estimated glottal airflow for different α values. However, this strategy is sensitive to the decision threshold used to detect changes in the area.

An alternative to automatically computing glottal airflow is the quadratic programming glottal inverse filtering (QPR) approach introduced in [14]. QPR is based on a comprehensive formulation that integrates the vocal tract and the lip radiation into a single filter; the filter coefficients are computed using a quadratic programming scheme following a quasi closed phase analysis, aiming at minimizing the prediction error in the excitation-free samples in the closed phase. QPR stands out from traditional glottal inverse filtering methods because it directly obtains the glottal airflow waveform from the voice signal, ruling out the need for other intermediate signals (e.g., the glottal function). Additionally, it avoids the manual tuning required for the LIF. However, recent evidence indicates that QPR struggles to accurately estimate the glottal airflow waveform outside the closed phase [16].

The goal of the current investigation is to propose a new method for computing the glottal airflow waveform based on time-frequency analysis. Time-frequency methods have received significant attention in applications involving the study of natural signals characterized by complex, non-harmonic oscillatory patterns [17]. In particular, the adaptive non-harmonic (ANH) model is an advanced tool for studying time-varying, multicomponent oscillatory signals. As proposed in [18], the ANH model aims to optimally represent the signal waveform using estimates of instantaneous amplitude and phase, and a fixed number of waveshape functions [19].

The present article introduces a regularized ANH (RANH) model for computing the glottal airflow waveform from the inverse-filtered glottal function, under the assumptions that both signals share identical phases and maintain approximately constant amplitudes over short segments. A regularized least-squares optimization scheme is proposed to indirectly adjust the RANH model of the glottal airflow based on observations of its time-derivative. This scheme is designed to ensure an accurate representation of the glottal airflow waveform overall, while promoting a flatter waveform during the closed phase. To the best of our knowledge, this is the first study to propose a regularization-based optimization scheme for the ANH model. Our approach has the advantage of automatically addressing the glottal airflow estimation, without requiring user-tuned parameters. The results on synthetic and natural phonation data indicate that the proposed RANH model outperforms the LIF and the QPR methods.

The organization of the paper is as follows. Section 2 introduces the RANH model for the glottal signals. Section 3 describes the phonation data and the experimental setup. Section 4 provides a detailed analysis of the proposed method for estimating the glottal airflow. In Section 5, the results are compared against state-of-the-art methods. Finally, we deliver the discussions and conclusions of this work in Sections 6 and 7.

2. Regularized adaptive non-harmonic model for glottal airflow computation

2.1. Adaptive non-harmonic model

The ANH model enables to describe a multicomponent signal, $x(n)$, as a sum of oscillatory components [18]:

$$x(n) = \sum_{k=1}^{K} A_k(n) s_k(2\pi\phi_k(n)), \text{ with } 1 \leq n \leq N, \tag{3}$$

where N is the length of the signal, K is the number of components, and s_k is the wave-shape function of the $k-$th component. $A_k(n)$ and $\phi_k(n)$ represent the instantaneous amplitude and phase of the $k-$th component, respectively, satisfying $A_k(n) > 0$, $\phi'_k(n) > 0$ [20]. Representing each s_k by its Fourier expansion, Eq. (3) can be written as follows:

$$x(n) = \sum_{k=1}^{K} A_k(n) \sum_{\ell=0}^{L} \big[a_{k,\ell} \cos\left(2\pi\ell\phi_k(n)\right) + b_{k,\ell} \sin\left(2\pi\ell\phi_k(n)\right)\big], \tag{4}$$

where $a_{k,\ell}$ and $b_{k,\ell}$ are the coefficients of the trigonometric Fourier expansion of the $k-$th wave-shape function, and L denotes the maximum number of admissible harmonics (according to the Nyquist criterion). For monocomponent zero-mean signals, we can simplify the ANH model by considering a single wave-shape function ($K = 1$), thus [20]:

$$x(n) = A(n) \sum_{\ell=1}^{L} a_\ell \cos\left(2\pi\ell\phi(n)\right) + b_\ell \sin\left(2\pi\ell\phi(n)\right). \tag{5}$$

In Eq. (5), the number of harmonics in the ANH model is usually limited by assuming that the magnitudes of the Fourier coefficients are negligible for all ℓ greater than a fixed number of harmonics r:

$$x_r(n) = A(n) \sum_{\ell=1}^{r} a_\ell \cos\left(2\pi\ell\phi(n)\right) + b_\ell \sin\left(2\pi\ell\phi(n)\right), \tag{6}$$

where x_r denotes the r-harmonics approximation of x (would be equal to x if $r = L$). The Eq. (6) can be expressed in matrix notation as:

$$\mathbf{x}_r = \mathbf{C}_r \mathbf{a}, \tag{7}$$

where $\mathbf{x}_r \in \mathbb{R}^{N\times 1}$ is the r-harmonics approximation vector, $\mathbf{a} = [a_1, \ldots, a_r, b_1, \ldots, b_r]^T \in \mathbb{R}^{2r\times 1}$ gathers the Fourier expansion coefficients of s, and $\mathbf{C}_r \in \mathbb{R}^{N\times 2r}$ is a pseudo-Fourier dictionary of the form: $\mathbf{C}_r = [\mathbf{c}_1, \ldots, \mathbf{c}_r, \mathbf{d}_1, \ldots, \mathbf{d}_r]$, with columns $\mathbf{c}_\ell = A(n)\cos\left(2\pi\ell\phi(n)\right)$ and $\mathbf{d}_\ell = A(n)\sin\left(2\pi\ell\phi(n)\right)$, for $\ell = 1, \ldots, r$ and $n = 1, \ldots, N$.

If estimates of the instantaneous amplitude $A(n)$ and phase $\phi(n)$ are available (typically from the signal's time-frequency representation, as in [20]), then x can be approximated by suitably estimating the coefficients $\mathbf{a}$ in the model (7). The coefficients can be estimated by solving the following least-squares problem:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} ||\mathbf{x} - \mathbf{C}_r\mathbf{a}||_2^2, \tag{8}$$

where $\mathbf{x} = [x(1), x(2), \ldots, x(N)]^T \in \mathbb{R}^{N\times 1}$ is a vector that contains all samples of the signal. The Eq. (8) has a closed-form solution:

$$\hat{\mathbf{a}} = (\mathbf{C}_r^T\mathbf{C}_r)^{-1}\mathbf{C}_r^T\mathbf{x}. \tag{9}$$

Thus, the reconstructed signal is obtained using the synthesis formula $\hat{\mathbf{x}}_r = \mathbf{C}_r\hat{\mathbf{a}}$.

2.2. ANH model of the glottal airflow

The ANH model of the glottal airflow is described here, based on two fundamental assumptions. The glottal airflow is assumed to be a monocomponent signal that can be modeled with a single wave-shape function with r harmonics. It is also assumed that the glottal airflow amplitude remains relatively constant in short-duration segments, thus $A(n) = 1$ in Eq. (6). Let $\mathbf{u_g} = [u_g(1), u_g(2), \ldots, u_g(N)]^T \in \mathbb{R}^{N\times 1}$ be the vector gathering the $N-$samples glottal airflow. According to the ANH model in Eq. (7), $\mathbf{u_g}$ can be modeled as follows:

$$\mathbf{u_g} = \mathbf{C}_r\mathbf{a}, \tag{10}$$

given proper Fourier coefficients $\mathbf{a}$ and $\mathbf{C}_r$.

As explained in Section 1, the glottal airflow is seldom available in the glottal inverse filtering context; instead, it is usually estimated from the glottal function, its time-derivative, by using the LIF. Here, we propose an alternative approach to estimate the glottal airflow based on the model (10), and on indirectly determining the coefficients $\mathbf{a}$ from the inverse filtered glottal function. By taking the time-derivative of Eq. (10), we can deduct a related ANH model for the glottal function:

$$\mathbf{v_g} = \dot{\mathbf{C}}_r\mathbf{a}, \tag{11}$$

where $\mathbf{v_g} = [v_g(1),\, v_g(2),\, \ldots,\, v_g(N)]^T \in \mathbb{R}^{N\times 1}$ is the $N-$samples glottal function vector, and $\dot{\mathbf{C}}_r = [\dot{\mathbf{c}}_1,\, \ldots,\, \dot{\mathbf{c}}_r,\, \dot{\mathbf{d}}_1,\, \ldots,\, \dot{\mathbf{d}}_r]$ is the matrix whose columns are the time-derivatives of the columns $\mathbf{c}_\ell$ and $\mathbf{d}_\ell$ of $\mathbf{C}_r$. As a consequence, the coefficients $\mathbf{a}$ can be alternatively obtained by solving the following least-squares problem:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} ||\tilde{\mathbf{v}}_\mathbf{g} - \dot{\mathbf{C}}_r\mathbf{a}||_2^2, \tag{12}$$

where $\tilde{\mathbf{v}}_\mathbf{g}$ denotes the inverse filtered glottal function data. The optimization problem has a closed-form solution as in Eq. (9), and the glottal airflow can be obtained from Eq. (10) by using the computed coefficients $\hat{\mathbf{a}}$.

2.3. Closed phase regularized ANH model

Note that the optimization scheme in Eq. (12) calculates the coefficients $\hat{\mathbf{a}}$ that minimize the waveform reconstruction error of the glottal function. However, this criterion does not contemplate any specific aspects of the glottal airflow waveform. In [14], it is demonstrated that minimizing the l_1-norm of the glottal airflow samples is effective to promote that the estimated waveforms remain flat and free from distortion in the closed phase. A weighting function, w_{cp}, is used to select the samples in the closed phase, as illustrated in the top panel of Fig. 2. This function is determined based on the locations of the glottal opening and closing instants (GOIs and GCIs).

Based on the ideas proposed in [14], the estimation of the Fourier coefficients can be improved by incorporating a closed phase regularization term into the least-squares optimization problem:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left\{ \underbrace{||\tilde{\mathbf{v}}_\mathbf{g} - \dot{\mathbf{C}}_r\mathbf{a}||_2^2}_{\text{Waveform error}} + \beta \underbrace{||\mathbf{W}\dot{\mathbf{C}}_r\mathbf{a}||_1}_{\text{Regularization}} \right\}, \tag{13}$$

where β controls the weight of the regularization term, $\mathbf{W} = \text{diag}(\mathbf{w}_{cp}) \in \mathbb{R}^{N\times N}$ and $\mathbf{w}_{cp} = [w_{cp}(1),\, w_{cp}(2),\, \ldots,\, w_{cp}(N)]^T \in \mathbb{R}^{N\times 1}$ is a N-samples closed phase weighting vector. Hereafter, this approach is referred to as regularized ANH (RANH) model.

The parameter β in Eq. (13) establishes a trade-off between two aspects of the glottal waveform. The first term ensures the waveform reconstruction of the inverse filtered glottal function $\tilde{\mathbf{v}}_\mathbf{g}$, while the regularization term aims to minimize the amplitudes of the estimated glottal signal during the closed phase. Note that the case $\beta = 0$ corresponds to the scheme without the regularization term, i.e., the original ANH model. The bottom panel of Fig. 2 displays two glottal airflow estimates computed by the ANH and the RANH models respectively, with a vertical offset applied for a better visual comparison. As can be seen, both ANH and RANH produce similar waveform reconstruction for the samples during the open phase; however, in the closed phase, RANH produces a flatter estimated waveform than the non-regularized scheme.

2.4. Number of harmonic components

The number of harmonics, r, is a key parameter in the RANH model, as it directly affects the estimation of the coefficients $\hat{\mathbf{a}}$ and the reconstruction of the glottal signals. In general, increasing the number of harmonics improves the fit

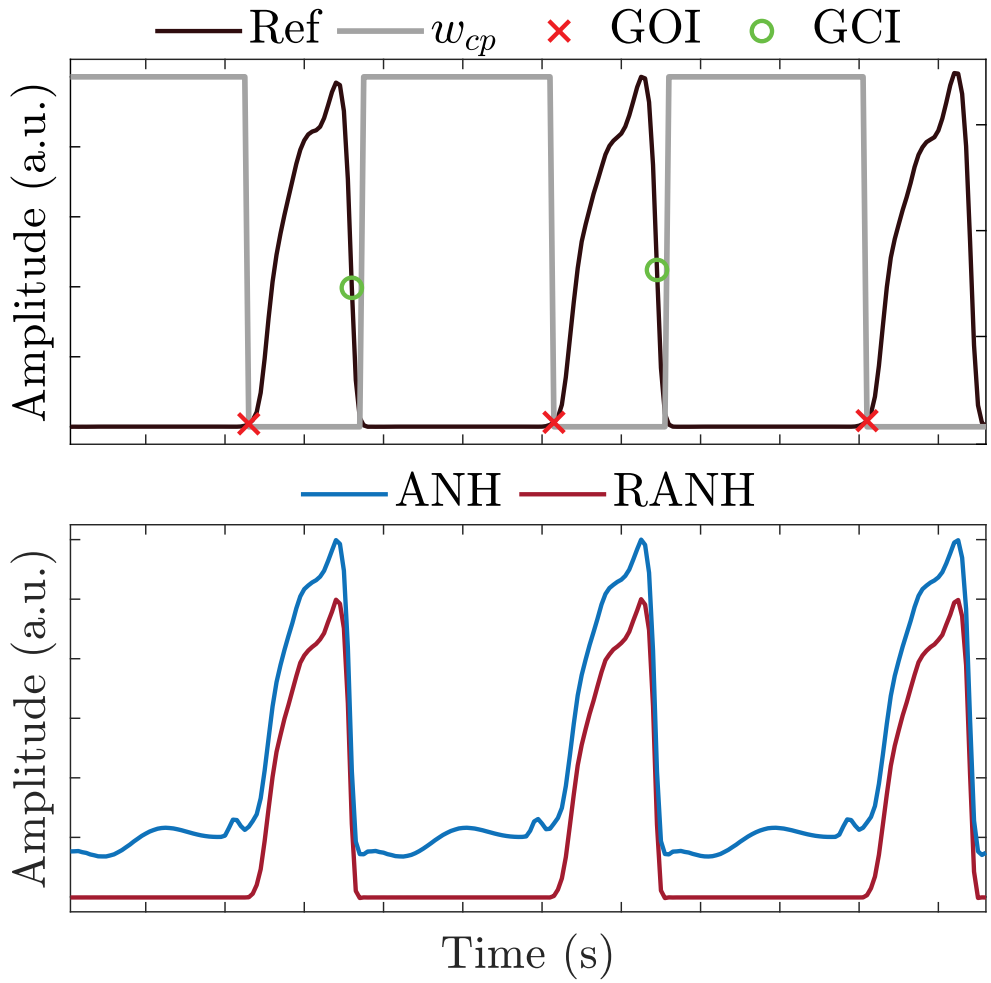

Figure 2: Top: Example of a closed phase weighting function, w_{cp}, for three glottal cycles of a simulated glottal airflow waveform (denoted as "Ref"). Bottom: Estimations of the glottal airflow obtained using the ANH model and the proposed RANH model with $\beta = 1$.

of the model reconstruction. Minimizing the mean squared error (MSE) between the signal and its reconstruction is the classical criterion used to determine the optimal r. In general, for a noise-free signal, the MSE decreases monotonically with the number of harmonics, and the value of r is chosen such that the Fourier coefficients for $\ell > r$ vanish [20]. However, in the presence of noise, a large r may lead to overfitting, as the model captures both the signal waveform and the noise. Conversely, choosing too few harmonics can result in poor reconstructions of the signals. Different criteria have been studied to determine r in the ANH model by minimizing a penalized MSE criterion [20].

In this work, we assume that the inverse filtered glottal function signal is noise-free. Since glottal signals are quasi-periodic with fundamental frequency, f_0, their time-frequency representation contains harmonics approximately located at integer multiples of f_0. The number of harmonics is inherently limited by the Nyquist frequency, f_N. Therefore, we propose defining r as a function of f_N and f_0, as follows:

$$r = \lfloor f_N / f_0 \rfloor, \tag{14}$$

where $\lfloor \cdot \rfloor$ is the floor operator (i.e., $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x). Equation (14) provides a practical and automatic criterion to adjust

the harmonic number for each signal.

3. Material and methods

3.1. Phonation data for glottal airflow evaluation

Simulations were carried out to evaluate the performance of the proposed method for estimating the glottal airflow waveform. The *OPENGLOT* database [21], a comprehensive data set on glottal excitation during phonation, was considered for our simulations. In this work, only repositories II and IV were considered. Repository II contains signals obtained from a physical model of human phonation. The available speech material includes the voice signal, and the theoretical glottal airflow. The phonation data correspond to three vowel sounds: [/a/, /i/, /u/], produced at four fundamental frequencies: 82, 110, 156, and 220 Hz for male speakers, and 175, 196, 220, and 294 Hz for female speakers, and three different degrees of vocal fold adduction: Small, Medium, and Large. On the other hand, the repository IV contains voice and electroglottogram signals from natural production of vowel sounds at Modal and Breathy vocal qualities. All signals used in the simulations were resampled at 8 kHz in accordance to the hypothesis supporting the *source-filter* theory [1]. Any lag between signals was compensated by computing the cross-correlation [4].

3.2. Experimental setup

The methodology for estimating the glottal airflow from the voice signal using the RANH model is explained here. First, the estimated glottal function is obtained from the voice signal by applying voice inverse filtering through quasi closed phase linear prediction method, considering a 12th-order vocal tract filter. The weighting function parameters are set up as suggested in [13]. Inverse filtering analysis is performed on non-overlapping 50-ms segments of the voice signal. A pre-emphasis filtering, with transfer function $R(z) = 1 - 0.99z^{-1}$, is applied to enhance the high-frequency spectral content of the voice signal before computing the vocal tract filter.

For the RANH model, the phase $\phi(n)$ is obtained from the time-frequency representation of the estimated glottal function using steps 1 to 5 of the algorithm used in [20], without applying the de-shape method. The number of harmonics, r, is computed according to Eq. (14), based on the fundamental frequency estimated from the glottal function.

The determination of the weighting functions for the quasi closed phase linear prediction method and the RANH model requires prior knowledge about the locations of the GOIs and GCIs. Two different strategies are employed: (i) for the synthetic signals in repository II, the glottal instants were determined from the theoretical glottal function; and (ii) for natural voices in repository IV, the instants were estimated from the electroglottogram signal as in [22].

The coefficients $\hat{\mathbf{a}}$ of the RANH model are estimated by minimizing Eq. (13) for a given β (considering r and the weighting function $\mathbf{w}_{cp}$ as explained above) using MATLAB's CVX convex programming toolbox [23, 24]. Finally, glottal airflow is computed using Eq. (10).

3.3. Benchmark schemes

The LIF in Eq. (2) was used as the standard method for computing the glottal airflow from the inverse filtered glottal function. The α value in Eq. (2) was automatically determined using the algorithm proposed in [3]. QPR method was also included for comparison. QPR requires three optimization constants γ_1, γ_2, and γ_3. In our simulations, these were set to 40 (200 for natural voice signals), 5000, and 500, respectively, as suggested in [14].

3.4. Performance measures

For assessing the estimation quality, the glottal airflow obtained with each method is amplitude-normalized, and a constant offset is introduced to ensure that the minimum signal value is close to zero [15].

Direct evaluation of glottal airflow estimates is possible only for synthetic data from repository II. We used the MSE between the theoretical and the estimated waveforms of the glottal airflow, u_g and $\hat{u}_g$, respectively:

$$E_{u_g} = \frac{1}{N}\sum_{n=1}^{N}(u_g(n) - \hat{u}_g(n))^2. \tag{15}$$

Before error computing, the signals in Eq. (15) are time-aligned to compensate for any phase delay due to the acoustic wave propagation along the vocal tract, or the applied inverse filtering method.

As discussed in Section 1, glottal airflow is susceptible to disturbances caused by insufficient cancellation of the vocal tract effect or issues with numerical integration, especially during the closed phase [3]. Glottal airflow estimation was further evaluated using the l_1-norm (i.e., the sum of the absolute values) over the closed phase. The l_1-norm is used to quantify the waveform flatness of the estimated glottal airflow in the closed phase [14]. So, a large l_1-norm indicates a significant distortion of the glottal waveform.

As glottal airflow information is not available for natural phonation, a comparison against a ground truth cannot be undertaken for the signals from repository IV. For this reason, analysis of natural voice signals is carried out by measuring l_1-norm in the closed phase of the estimated glottal airflow. Visual inspection was also applied to assess the temporal correspondence between the glottal airflow estimates and the electroglottogram signals. The goal is to verify whether the GOIs and GCIs obtained from the electroglottogram signal suitably align with the characteristic changes in the glottal airflow waveform: an amplitude increase at the end of the closed phase and a sharp decrease after reaching its maximum value, before the next closed phase, for each glottal cycle.

4. Parameter analysis of the RANH model for glottal airflow estimation

This section analyzes the effect of the number of harmonics and the closed phase regularization on the performance of the proposed RANH model to compute glottal airflow using the inverse filtered glottal function estimations from repository II.

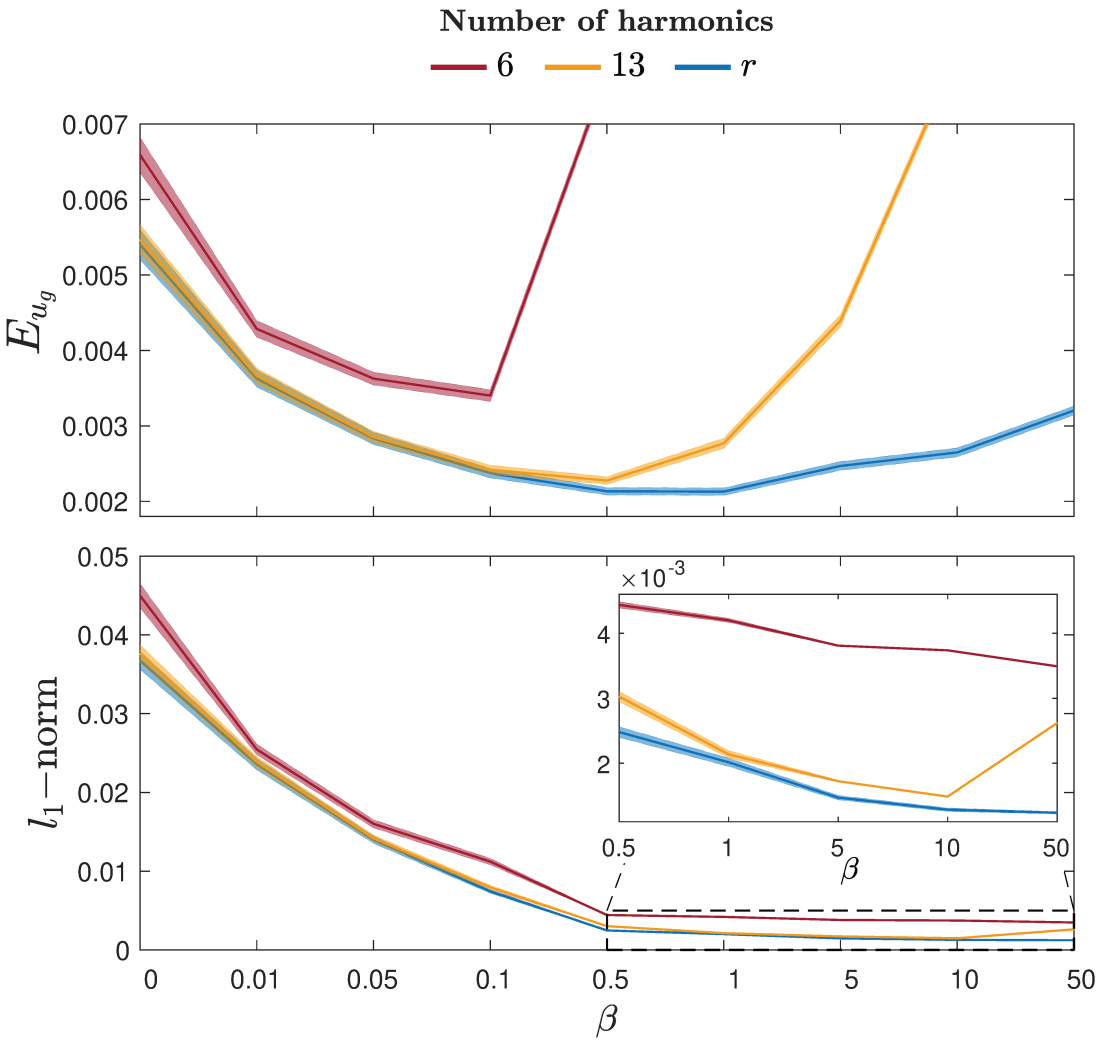

Figure 3: Waveform error E_{u_g} (top) and l_1-norm in the closed phase (bottom) as a function of weight β for the estimated glottal airflow through the RANH model from repository II, for three conditions on the number of harmonics.

To investigate the impact of the number of harmonics in the RANH model of the glottal airflow, three experimental conditions were considered. The baseline condition involves adjusting r individually for each voice signal based on its fundamental frequency as described in Eq. (14). In the second condition, the number of harmonics was fixed at 13, corresponding to the number of harmonics for the voice signal with the highest fundamental frequency (294 Hz) in repository II. Finally, a third condition using a fixed number of harmonics equal to 6 was included to assess the performance of the model when considering a very limited number of harmonics.

For all of the experimental conditions, the weight of the regularization in Eq. (13) was varied to assess the effect of the closed phase regularization in the estimation of the glottal airflow waveform. Figure 3 illustrates the waveform error E_{u_g} and the l_1-norm in the closed phase (with the 95% confidence intervals) for the signals from repository II as functions of the weight β, for the three experimental conditions considered for the number of harmonics.

The results in top panel in Fig. 3 show that introducing the closed phase regularization ($\beta > 0$) significantly reduces the waveform error E_{u_g} compared to

the non-regularized case ($\beta = 0$). As β increases from zero, the error E_{u_g} initially decreases, reaching a minimum value in the range $0.1 < \beta \leq 1$. Higher values of β cause the error to rise again. This reflects the trade-off between the two terms in the optimization scheme in Eq. (13): at low β values, the waveform reconstruction error term of the glottal function dominates, while at high β values, the closed phase regularization term becomes overly dominant, leading to reduced accuracy in the estimation of the glottal airflow waveform in the open phase.

On the other hand, the l_1-norm (bottom panel in Fig. 3) decreases steeply for $\beta < 0.5$ and stabilizes for higher values for all cases. These results confirm that the regularization effectively constrains the glottal airflow amplitude during the closed phase, which was the goal of the proposed RANH model. Further increases in the regularization weight do not substantially reduce the l_1-norm.

Figure 3 shows that setting the number of harmonics according to Eq. (14) provides an effective trade-off in the optimization scheme of Eq. 13 between the waveform reconstruction error and the closed phase regularization term. This follows from the minimum waveform error observed in the range $0.5 \leq \beta \leq 1$, and the lowest l_1-norm during the closed phase, compared to other conditions.

Constraining the RANH model to a fixed lower number of harmonics (specifically 6 and 13 for our simulations) degrades the waveform reconstruction accuracy as β increases, although the l_1-norm is less affected. Note that the 6-harmonics condition consistently yields the highest waveform error E_{u_g} across all β. This results from the limited harmonic representation, which restricts the RANH model to suitably satisfy both terms in the optimization scheme of Eq. (13).

In light of the previous discussion, we select the number of harmonics in the RANH model according to Eq. (14) and set the regularization weight $\beta = 1$ for the remainder of this article. For comparative purposes, we also include the non-regularized ANH model (corresponding to the RANH model with $\beta = 0$) in the analysis.

5. Results

This section evaluates the performance of the proposed RANH model for estimating the glottal airflow waveform from inverse filtered glottal functions in repositories II and IV, compared with the non-regularized ANH model, QPR, and standard LIF. Statistical analyses were performed using the Friedman test with Bonferroni correction.

5.1. Synthetic signals

The top panel of Fig. 4 shows boxplots for the waveform error for the estimated glottal airflow, E_{u_g}, for different fundamental frequency ranges (Low: $f_0 < 200$ Hz, and High: $f_0 \geq 200$ Hz), vowels, and degrees of vocal fold adduction.

As can be seen, the RANH model presented the lowest waveform errors for most categories considered. Additionally, LIF and the non-regularized ANH model exhibited similar performances for most categories, whereas QPR showed the highest errors. Statistical analysis shows statistically significant differences (at an overall significance level of 95%) between the RANH model and the rest of the methods for most categories. However, no significant differences were found between RANH, LIF and the non-regularized ANH model, for signals with large vocal fold adduction.

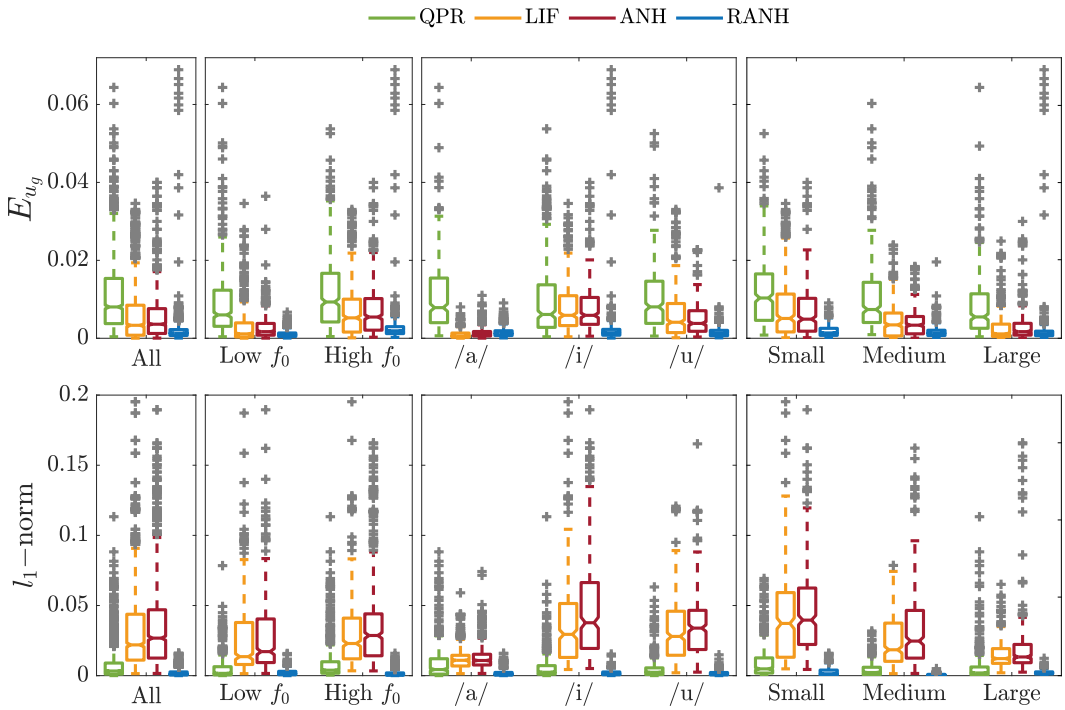

Figure 4: Waveform errors E_{u_g} (top) and l_1-norm in the closed phase (bottom) obtained by using different glottal airflow estimation methods in synthetic signals from repository II. The results are reported for all signals, and grouped by f_0 ranges, vowel sounds, and degrees of vocal fold adduction.

Analysis based on f_0 revealed that, among the estimation methods considered, the RANH model achieved the lowest waveform errors. Furthermore, all methods exhibited a general increase in the error E_{u_g} when applied to voice signals with high f_0. A similar trend is found in the vowel-based analysis: the RANH model yielded the lowest errors for the vowels /i/ and /u/; however, for the vowel /a/, it produced higher errors compared to LIF. Regarding the different degrees of vocal fold adduction, the RANH model outperformed all methods for small and medium vocal fold adductions. However, for the large adduction case, its performance was comparable to that of both the LIF and the non-regularized ANH model.

Bottom panel in Fig. 4 showed that, in terms of the l_1-norm in the closed phase of the estimated glottal airflow, the RANH model yielded lowest values across all categories, closely followed by the QPR method; statistically significant differences (at an overall significance level of 95%) between RANH and the other methods were obtained for all categories. These results agree with the fact that both QPR and the proposed RANH were developed with the focus on estimating a glottal airflow waveform with a minimum l_1-norm in the closed phase. It should be noted that, for the vowel /a/, the LIF method and the non-regularized ANH model produced l_1-norm values close to those achieved by the RANH model. In contrast, for vowels /i/ and /u/, the differences in l_1-norm values between the methods were more noticeable. These behaviors are consistent with the waveform error analysis, which showed comparable error levels among the RANH, ANH, and LIF for vowel /a/, but greater differences for /i/ and /u/.

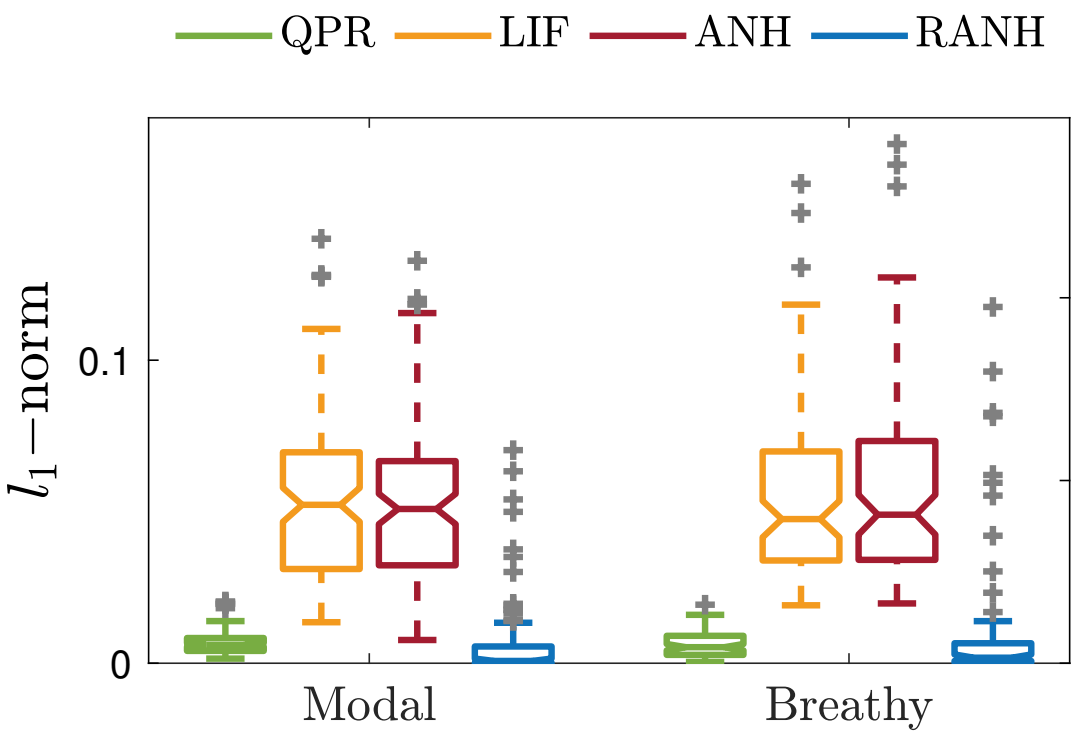

Figure 5: Boxplots for the l_1-norm in the closed phase of the estimated glottal airflows from natural signals in repository IV using different methods.

5.2. Natural signals

The results for natural voice signals from repository IV are presented in this section. Figure 5 shows boxplots of the l_1-norm in the closed phase of the estimated glottal airflow, obtained with different methods for modal and breathy phonation qualities. As shown, the RANH model and QPR method produced the lowest l_1-norm values for both phonation qualities in contrast to the non-regularized ANH model and LIF.

Statistically significant differences were obtained between the RANH model and QPR for modal voice quality, with the RANH model yielding lower l_1-norm values; however, no significant differences were found for breathy quality. In addition, no statistically significant differences were found between the l_1-norm values obtained with the non-regularized ANH model and the LIF method for either voice quality.

Figure 6 shows examples of estimated glottal airflow waveforms obtained from modal (left column) and breathy (right column) natural voice signals produced by female (top row) and male (bottom row) speakers. Additionally, the time-derivative of the electroglottogram signal is shown along with the location of the GCIs and GOIs. All curves are shifted vertically for better visualization.

In Fig. 6, it can be seen that the RANH model and QPR method produce glottal airflow estimations with the flattest closed phase in contrast to the non-regularized ANH model and LIF. These results agree with the results shown in Fig. 5. It is worth noting that the closed phase in the waveforms calculated by the RANH model aligns better with the glottal instants, specially the GOIs, in contrast to the QPR method. LIF method and both ANH models yield similar glottal airflow waveforms for the samples outside the closed phase, that is, over the

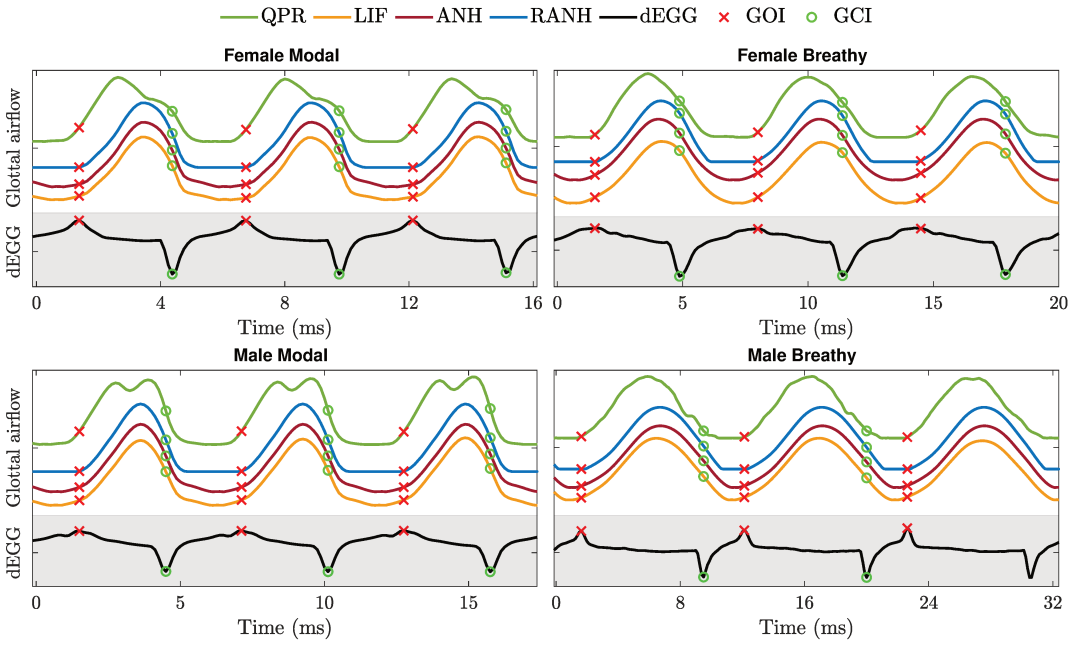

Figure 6: Examples of glottal airflow waveforms estimated with different methods from female (top) and male (bottom) voice signals with different phonation qualities from repository IV: Modal (left) and Breathy (right). dEGG: Time-derivative of the electroglottogram signal, with the location of glottal opening (GOI) and closing (GCI) instants are shown as reference.

open phase. In contrast, the QPR method produces estimations with distorted waveforms, particularly for modal voices. Similar distortions produced by QPR were also reported in [16].

6. Discussion

The proposed ANH model estimates the glottal airflow waveform using the least-squares optimization scheme in Eq. (12) that minimizes the reconstruction error of the glottal function, which is considered as the time-derivative of the glottal airflow in the context of the *source-filter* theory of phonation. Initial assessments, based on the error of the estimated glottal airflow waveform, E_{u_g}, for synthetic signals, show that the ANH model achieves performance comparable to the standard LIF method, outperforming the QPR method (see top panel in Fig. 4). However, when assessing natural and synthetic signals using the l_1-norm in the closed phase, QPR exhibited superior results, achieving lower values than both ANH and LIF (see bottom panel in Fig. 4, and Fig. 5).

To enhance the performance, a regularized ANH (RANH) model is here proposed. The formulation of the ANH model enables the straightforward inclusion of a regularization term that promotes a flatter closed phase in the optimization scheme used for estimating the glottal airflow waveform (see Eq. (13)). The proposed RANH model significantly improves the glottal estimations, outperforming not only the non-regularized ANH model but also the QPR and LIF methods in terms of both waveform error and l_1-norm metrics.

The observed performance differences for synthetic signals are explained in part by the issues discussed in the Sec. 1. Low-frequency inaccuracies in the estimated vocal tract filter introduce distortions during the inverse filtering process, that particularly affect the glottal airflow waveform in the closed phase.

This effect is especially noticeable for /i/ and /u/ vowels, whose typical low first formants make the estimated vocal tract filter more susceptible to low-frequency errors. Applying an inverse vocal tract filter affected by such errors introduces low-frequency distortions in the glottal function waveform. These distortions are then amplified by the LIF method and, inevitably, modeled by the non-regularized ANH model as part of the glottal airflow waveform.

Therefore, the inclusion of the closed phase regularization term in the RANH model helps to suppress these low-frequency distortions in the closed phase, thereby enabling a more accurate estimation of the glottal airflow waveform. This behavior explains the lower E_{u_g} and l_1-norm values achieved by the RANH model compared to the non-regularized ANH model and the LIF method, as shown in Fig. 4. Conversely, for the vowel /a/, whose first formant resides in the higher-frequency region, low-frequency errors are scarce or negligible in the vocal tract filter. Consequently, the LIF method and the ANH models achieve comparable low values of E_{u_g} and l_1-norm.

Further analysis of degrees of vocal fold adduction reveals that the closed phase regularization becomes particularly beneficial for voices characterized by a longer closed phase. For synthetic signals with small and medium vocal fold adductions, the RANH model produces glottal airflow estimates with the lower waveform errors and l_1-norm values than the other considered methods. Conversely, for large vocal fold adductions, where the closed phase is shorter, the influence of the regularization term becomes less relevant, and the results obtained with the proposed ANH models and the LIF method are comparable.

Note that RANH and QPR, two methods based on closed phase regularization, exhibit different performances. As shown in the top panel of Fig. 4, the QPR method yields the highest E_{u_g} values. This lower performance can be attributed to the glottal waveform distortions observed in the QPR estimations during the open phase, as shown in Fig. 6. Instead, the proposed optimization scheme in the RANH model better balances the requirements for a flat closed phase together with the fitting of the glottal airflow waveform overall.

The results in synthetic and natural signals suggest that the ANH and RANH models provide a flexible and effective framework for glottal airflow estimation from an inverse filtered glottal function. Particularly, the RANH model seems to be superior, systematically achieving the most accurate glottal airflow estimations. The non-regularized ANH model, nevertheless, could still be a valuable alternative when a less constrained waveform estimation is required.

7. Conclusion

The present study investigates the Adaptive Non-Harmonic Model of the glottal airflow for inverse filtering applications. A novel regularized version of the ANH model is introduced to optimally estimate the glottal airflow from the inverse filtered glottal function, which is assumed to be its natural time-derivative, while promoting a flatter waveform during the closed phase, as a physiologically-based quality attribute for glottal inverse filtering. To this end, a regularized least-squares optimization scheme was formulated, in which the model coefficients are determined under the assumption that the glottal airflow and the glottal function share the same phase and exhibit approximately constant amplitudes over short segments. The proposed method was designed to address the practical

challenges of estimating glottal airflow using the well-established leaky integrator filter.

Effects of the adjustment of the proposed model parameters on glottal airflow estimations were investigated. Extensive simulations were conducted to identify the optimal number of harmonics and the appropriate weighting for the closed phase regularization term. The performance of the proposed method was assessed based on the waveform estimation error and the l_1-norm in the closed phase.

The results of the experiments demonstrate that the proposed model outperforms other well-established methods for glottal airflow estimation. In the case of synthetic signals, it achieved superior performance—in terms of waveform error and l_1-norm—for vowel sounds characterized by a low-frequency first formant and a longer closed phase. Furthermore, for natural speech signals, it provided smoother glottal airflow estimates, exhibiting a flatter waveform over the closed phase.

It is worth noting that the proposed method is limited by depending on prior knowledge of the glottal instants location for the closed phase regularization. These instants can be estimated from the electroglottogram signal. However, in the absence of this signal, estimating them directly from the voice signal becomes challenging. Although some data-driven methods have shown potential for automatically identifying the closed phase from the voice [10, 12], their suitability for this specific application has not yet been established. Another limitation of the proposed model is the assumption that the glottal airflow waveform maintains constant amplitude over short segments, which restricts its applicability to longer voice signals. For natural signals, it is difficult to find glottal airflow waveforms that do not exhibit any form of amplitude modulation.

Future work will focus on incorporating time-varying amplitude components to better capture the variability in long voice segments or dynamic speech. Additional evaluations under noisy or reverberant conditions could help to assess the applicability of the proposed model in real-world scenarios. Additionally, alternative dictionaries for the Adaptive Non-Harmonic Model could be explored to more effectively represent physiologically relevant aspects of glottal airflow, such as flat closed phase and smoother waveform, thus avoiding the need for the closed phase regularization.

Acknowledgments

This work was financed by the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) through project PIP-CONICET 633, the Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación through project PICT-ANPCYT 2020 Serie A-01865 and PICT-2021-I-INVI-00122, and the Universidad Nacional de Entre Ríos (UNER) through PID-UNER projects 6280 and 6281.

References

[1] G. Fant, *Acoustic theory of speech production*. Mounton, The Hague, The Netherlands, 1970, no. Number 2.

[2] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Extraction and utilization of excitation information of speech: A review," *Proceedings of the IEEE*, 2021.

[3] M. Airaksinen, T. Bäckström, and P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[4] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-time processing of speech signals*. Institute of Electrical and Electronics Engineers, 2000.

[5] S. Gmyrek, R. Hossa, and R. Makowski, "Amplitude spectrum correction to improve speech signal classification quality," *International Journal of Electronics and Telecommunication*, vol. 70, no. 3, pp. 569–574, 2024.

[6] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[8] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[9] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *the Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[10] A. Rao and P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 114–124, 2018.

[11] I. A. Zalazar, G. A. Alzamendi, and G. Schlotthauer, "Symmetric and asymmetric gaussian weighted linear prediction for voice inverse filtering," *Speech Communication*, vol. 159, p. 103057, 2024.

[12] I. A. Zalazar, G. A. Alzamendi, M. Zañartu, and G. Schlotthauer, "Maximum correntropy linear prediction for voice inverse filtering: Theoretical framework and practical implementation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[13] 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*, vol. 22, no. 3, pp. 596–607, 2013.

[14] M. Airaksinen, T. Bäckström, and P. Alku, "Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 929–939, 2016.

[15] T. Koc, "Post-processing method for removing low-frequency bias in glottal inverse filtering," *Electronics Letters*, vol. 51, no. 1, pp. 110–112, 2015.

[16] A. Palaparthi and I. R. Titze, "Analysis of glottal inverse filtering in the presence of source-filter interaction," *Speech communication*, vol. 123, pp. 98–108, 2020.

[17] Y.-W. Su, G.-R. Liu, Y.-C. Sheu, and H.-T. Wu, "Ridge detection for nonstationary multicomponent signals with time-varying wave-shape functions and its applications," *IEEE Transactions on Signal Processing*, 2024.

[18] H.-T. Wu, "Instantaneous frequency and wave shape functions (i)," *Applied and Computational Harmonic Analysis*, vol. 35, no. 2, pp. 181–199, 2013.

[19] C.-Y. Lin, L. Su, and H.-T. Wu, "Wave-shape function analysis: When cepstrum meets time–frequency analysis," *Journal of Fourier Analysis and Applications*, vol. 24, pp. 451–505, 2018.

[20] J. Ruiz and M. A. Colominas, "Wave-shape function model order estimation by trigonometric regression," *Signal Processing*, vol. 197, p. 108543, 2022.

[21] P. Alku, T. Murtola, J. Malinen, J. Kuortti, B. Story, M. Airaksinen, M. Salmi, E. Vilkman, and A. Geneid, "OPENGLOT–an open environment for the evaluation of glottal inverse filtering," *Speech Communication*, vol. 107, pp. 38–47, 2019.

[22] M. R. Thomas and P. A. Naylor, "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1557–1566, 2009.

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," https://cvxr.com/cvx, Mar. 2014.

[24] M. C. Grant and S. P. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent advances in learning and control*. Springer, 2008, pp. 95–110.

Acrónimos

En el presente documento utilizaremos una serie de acrónimos en inglés, por ser la forma más difundida en la bibliografía para referirse a métodos, modelos, o instantes particulares del ciclo glótico.

GOI	Instante de apertura glótica
GCI	Instante de cierre glótico
LP	Predicción lineal
GLP	LP con atenuación Gaussiana
SGLP	LP con atenuación Gaussiana simétrica
AGLP	LP con atenuación Gaussiana asimétrica
SWLP	LP ponderada y estabilizada
QCP	LP de fase casi cerrada
QCP-ST	QCP con compensación de inclinación espectral
PWLP	LP con ponderación probabilística
QPR	Filtrado inverso basado en programación cuadrática
MCLP	LP basada en el criterio de máxima correntropía
ANH	Modelo adaptativo no armónico
RANH	Modelo adaptativo no armónico regularizado
LIF	Filtro integrador con pérdidas

Lista de publicaciones

[1] I. A. Zalazar, G. A. Alzamendi y G. Schlotthauer, "Estudio comparativo de técnicas de extracción de fuente glótica basadas en filtrado inverso de la voz," en *XXII Congreso Argentino de Bioingeniería y XI Jornadas de Ingeniería Clínica (SABI)*, 2020.

[2] I. A. Zalazar, G. A. Alzamendi y G. Schlotthauer, "Gaussian-weighted voice inverse filtering: Effects of varying the attenuation window parameters on the glottal source estimation," en *XIX Workshop on Information Processing and Control (RPIC)*, 2021, págs. 1-6.

[3] I. A. Zalazar, G. A. Alzamendi y G. Schlotthauer, "Symmetric and asymmetric Gaussian weighted linear prediction for voice inverse filtering," *Speech Communication*, vol. 159, pág. 103 057, 2024.

[4] I. A. Zalazar, G. A. Alzamendi, M. Zañartu y G. Schlotthauer, "Correntropy-based linear prediction for voice inverse filtering," en *XVIII International Symposium on Medical Information Processing and Analysis (SIPAIM)*, 2022.

[5] I. A. Zalazar, G. A. Alzamendi, M. Zañartu y G. Schlotthauer, "Maximum Correntropy Linear Prediction for Voice Inverse Filtering: Theoretical Framework and Practical Implementation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[6] I. A. Zalazar, J. V. Ruiz, G. A. Alzamendi, M. A. Colominas y G. Schlotthauer, "Adaptive Non-Harmonic model for glottal airflow estimation in glottal inverse filtering," en *XXI Workshop on Information Processing and Control (RPIC)*, 2025.

[7] I. A. Zalazar, G. A. Alzamendi, J. V. Ruiz, M. A. Colominas y G. Schlotthauer, "Regularized adaptive non-harmonic model for glottal airflow estimation in glottal inverse filtering," *Biomedical Signal Processing and Control*, 2025, En revisión.

Bibliografía

[8] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[9] R. T. Sataloff, *Voice science*. Plural Publishing, 2017.

[10] R. J. Baken y R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[11] T. Drugman, P. Alku, A. Alwan y B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, n.º 5, págs. 1117-1138, 2014.

[12] 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*, vol. 36, n.º 5, págs. 623-650, 2011.

[13] S. R. Kadiri, P. Alku y B. Yegnanarayana, "Extraction and Utilization of Excitation Information of Speech: A Review," *Proceedings of the IEEE*, 2021.

[14] Y. Banaras, A. Javed y F. Hassan, "Automatic speaker verification and replay attack detection system using novel glottal flow cepstrum coefficients," en *2021 International Conference on Frontiers of Information Technology (FIT)*, IEEE, 2021, págs. 149-153.

[15] K. Bharath y M. R. Kumar, "New replay attack detection using iterative adaptive inverse filtering and high frequency band," *Expert Systems with Applications*, vol. 195, pág. 116 597, 2022.

[16] M. Swain, A. Routray y P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, págs. 93-120, 2018.

[17] X. Yao, W. Bai, Y. Ren, X. Liu y Z. Hui, "Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion," *Neurocomputing*, vol. 410, págs. 328-341, 2020.

[18] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang y Z. Tao, "Investigation and evaluation of glottal flow waveform for voice pathology detection," *IEEE Access*, vol. 9, págs. 30-44, 2020.

[19] A. B. Aicha y K. Ezzine, "Cancer larynx detection using glottal flow parameters and statistical tools," en *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, IEEE, 2016, págs. 65-70.

[20] R. Khumukcham y K. Nongmeikapam, "Investigation into a suitable voice pathology detection and classification using glottal inverse filtering," *Emerging Trends and Future Directions in Artificial Intelligence, Machine Learning, and Internet of Things Innovations*, pág. 8, 2025.

[21] S. R. Kadiri y P. Alku, "Glottal features for classification of phonation type from speech and neck surface accelerometer signals," *Computer Speech & Language*, vol. 70, pág. 101 232, 2021.

[22] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *The Journal of the Acoustical Society of America*, vol. 53, n.º 6, págs. 1632-1645, 1973.

[23] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman y G. R. Wodicka, "Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, n.º 9, págs. 1929-1939, 2013.

[24] J. P. Cortés, V. M. Espinoza, M. Ghassemi et al., "Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration," *PLoS One*, vol. 13, n.º 12, e0209017, 2018.

[25] I. Langheinrich, S. Stone, X. Zhang y P. Birkholz, "Glottal inverse filtering based on articulatory synthesis and deep learning," *Proc. Interspeech 2022*, págs. 1327-1331, 2022.

[26] P. Alku, T. Murtola, J. Malinen et al., "OPENGLOT–An open environment for the evaluation of glottal inverse filtering," *Speech Communication*, vol. 107, págs. 38-47, 2019.

[27] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970.

[28] A. Palaparthi e I. R. Titze, "Analysis of glottal inverse filtering in the presence of source-filter interaction," *Speech communication*, vol. 123, págs. 98-108, 2020.

[29] J. R. Deller, J. G. Proakis y J. H. Hansen, "Discrete-time processing of speech signals," Institute of Electrical y Electronics Engineers, 2000.

[30] M. R. Schroeder, *Computer speech: recognition, compression, synthesis*. Springer Science & Business Media, 2013, vol. 35.

[31] D. Gowda, S. R. Kadiri, B. Story y P. Alku, "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, págs. 1901-1914, 2020.

[32] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, n.º 4, págs. 561-580, 1975.

[33] T. Drugman, "Maximum phase modeling for sparse linear prediction of speech," *IEEE Signal Processing Letters*, vol. 21, n.º 2, págs. 185-189, 2014.

[34] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen y M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, n.º 5, págs. 1644-1657, 2012.

[35] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen y B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America*, vol. 134, n.º 2, págs. 1295-1313, 2013.

[36] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström y B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *the Journal of the Acoustical Society of America*, vol. 125, n.º 5, págs. 3289-3305, 2009.

[37] M. Airaksinen, T. Bäckström y P. Alku, "Automatic estimation of the lip radiation effect in glottal inverse filtering," en *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[38] D. Wong, J. Markel y A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, n.º 4, págs. 350-355, 2003.

[39] A. Rao y P. K. Ghosh, "Glottal inverse filtering using probabilistic weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, n.º 1, págs. 114-124, 2018.

[40] M. Airaksinen, T. Raitio, B. Story y P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, n.º 3, págs. 596-607, 2013.

[41] V. Khanagha y K. Daoudi, "An efficient solution to sparse linear prediction analysis of speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, n.º 1, págs. 1-9, 2013.

[42] Y.-R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu y T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, n.º 8, págs. 1718-1730, 2017.

[43] C. Ma, Y. Kamp y L. F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, n.º 1, págs. 69-81, 1993.

[44] C. Magi, J. Pohjalainen, T. Bäckström y P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, n.º 5, págs. 401-411, 2009.

[45] M. Airaksinen, T. Bäckström y P. Alku, "Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, n.º 5, págs. 929-939, 2016.

[46] H.-T. Wu, "Instantaneous frequency and wave shape functions (I)," *Applied and Computational Harmonic Analysis*, vol. 35, n.º 2, págs. 181-199, 2013.

[47] C.-Y. Lin, L. Su y H.-T. Wu, "Wave-shape function analysis: When cepstrum meets time–frequency analysis," *Journal of Fourier Analysis and Applications*, vol. 24, págs. 451-505, 2018.

[48] M. Airaksinen, L. Juvela, T. Bäckström y P. Alku, "Automatic Glottal Inverse Filtering with Non-Negative Matrix Factorization.," en *Interspeech*, 2016, págs. 1039-1043.

[49] T. Drugman, B. Bozkurt y T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, n.º 1, págs. 20-34, 2012.

[50] I. R. Titze y F. Alipour, *The myoelastic aerodynamic theory of phonation*. National Center for Voice y Speech, 2006.

[51] P. Alku, B. Story y 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*, vol. 58, n.º 2, págs. 102-113, 2006.

[52] G. Fant, J. Liljencrants, Q.-g. Lin et al., "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, n.º 1985, págs. 1-13, 1985.

[53] A. Hannukainen, J. Kuortti, J. Malinen y A. Ojalammi, "An acoustic glottal source for vocal tract physical models," *Measurement Science and Technology*, vol. 28, n.º 11, pág. 115 902, 2017.

[54] M. Włodarczak, B. Ludusan, J. Sundberg y M. Heldner, "Classification of voice quality using neck-surface acceleration: Comparison with glottal flow and radiated sound," *Journal of Voice*, vol. 39, n.º 1, págs. 10-24, 2025.

[55] D. G. Childers y C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *the Journal of the Acoustical Society of America*, vol. 90, n.º 5, págs. 2394-2410, 1991.

[56] P. Proutskova, C. Rhodes, T. Crawford y G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *Journal of New Music Research*, vol. 42, n.º 2, págs. 171-186, 2013.

[57] M. Airas y P. Alku, "Comparison of multiple voice source parameters in different phonation types.," en *INTERSPEECH*, 2007, págs. 1410-1413.

[58] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, n.º 1, págs. 49-64, 2008.

[59] D. Gowda, M. Airaksinen y P. Alku, "Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation," *The Journal of the Acoustical Society of America*, vol. 142, n.º 3, págs. 1542-1553, 2017.

[60] T. Kato, S. Omachi y H. Aso, "Asymmetric Gaussian and its application to pattern recognition," en *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops SSPR 2002 and SPR 2002 Windsor, Ontario, Canada, August 6–9, 2002 Proceedings*, Springer, 2002, págs. 405-413.

[61] A. El-Jaroudi y J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on signal processing*, vol. 39, n.º 2, págs. 411-423, 1991.

[62] W. Liu, P. P. Pokharel y J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Transactions on signal processing*, vol. 55, n.º 11, págs. 5286-5298, 2007.

[63] J. W. Xu y J. C. Principe, "A pitch detector based on a generalized correlation function," *IEEE transactions on audio, speech, and language processing*, vol. 16, n.º 8, págs. 1420-1432, 2008.

[64] X. Cui, Z. Chen, F. Yin y X. Xu, "Correntropy-Based Multi-objective Multichannel Speech Enhancement," *Circuits, Systems, and Signal Processing*, vol. 41, n.º 9, págs. 4998-5025, 2022.

[65] R. Singh y J. C. Principe, "Correntropy based hierarchical linear dynamical system for speech recognition," en *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018, págs. 1-7.

[66] I. Santamaría, P. P. Pokharel y J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization," *IEEE Transactions on Signal Processing*, vol. 54, n.º 6, págs. 2187-2197, 2006.

[67] A. Singh y J. C. Principe, "A closed form recursive solution for maximum correntropy training," en *2010 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2010, págs. 2070-2073.

[68] S. Zhao, B. Chen y J. C. Principe, "Kernel adaptive filtering with maximum correntropy criterion," en *The 2011 International Joint Conference on Neural Networks*, IEEE, 2011, págs. 2012-2017.

[69] J. C. Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[70] D. G. Manolakis, V. K. Ingle y S. Kogan, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. McGraw-Hill, 2000.

[71] A. Singh y J. C. Principe, "Using correntropy as a cost function in linear adaptive filters," en *2009 International Joint Conference on Neural Networks*, IEEE, 2009, págs. 2950-2955.

[72] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[73] L. Stanković, E. Sejdić, S. Stanković, M. Daković e I. Orović, "A tutorial on sparse signal reconstruction and its applications in signal processing," *Circuits, Systems, and Signal Processing*, vol. 38, n.º 3, págs. 1206-1263, 2019.

[74] Y.-W. Su, G.-R. Liu, Y.-C. Sheu y H.-T. Wu, "Ridge detection for nonstationary multicomponent signals with time-varying wave-shape functions and its applications," *IEEE Transactions on Signal Processing*, 2024.

[75] J. V. Ruiz, "Procesamiento de señales basado en funciones de forma de onda: contribuciones algorítmicas y aplicaciones biomédicas," Tesis Doctoral, Universidad Nacional del Litoral, Santa Fe, Argentina, 2025. dirección: `https://hdl.handle.net/11185/8529`.

[76] B. Picinbono, "On instantaneous amplitude and phase of signals," *IEEE Transactions on signal processing*, vol. 45, n.º 3, págs. 552-560, 2002.

[77] I. Daubechies, J. Lu y H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Applied and computational harmonic analysis*, vol. 30, n.º 2, págs. 243-261, 2011.

[78] L. Li, H. Cai y Q. Jiang, "Adaptive synchrosqueezing transform with a time-varying parameter for non-stationary signal separation," *Applied and Computational Harmonic Analysis*, vol. 49, n.º 3, págs. 1075-1106, 2020.

[79] J. Ruiz, G. Schlotthauer, L. Vignolo y M. A. Colominas, "Fully adaptive time-varying wave-shape model: Applications in biomedical signal processing," *Signal Processing*, vol. 214, pág. 109 258, 2024.

[80] J. T. Astola, K. O. Egiazarian, G. I. Khlopov et al., "Application of bispectrum estimation for time-frequency analysis of ground surveillance Doppler radar echo signals," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, n.º 9, págs. 1949-1957, 2008.

[81] N. Wang, E. Ambikairajah, B. G. Celler y N. H. Lovell, "Feature extraction using an AM-FM model for gait pattern classification," en *2008 IEEE Biomedical Circuits and Systems Conference*, IEEE, 2008, págs. 25-28.

[82] N. A. Khan y S. Ali, "A new feature for the classification of non-stationary signals based on the direction of signal energy in the time–frequency domain," *Computers in biology and medicine*, vol. 100, págs. 10-16, 2018.

[83] F. Li, R. Li, L. Tian, L. Chen y J. Liu, "Data-driven time-frequency analysis method based on variational mode decomposition and its application to gear fault diagnosis in variable working conditions," *Mechanical Systems and Signal Processing*, vol. 116, págs. 462-479, 2019.

[84] H.-T. Wu, H.-K. Wu, C.-L. Wang et al., "Modeling the pulse signal by wave-shape function and analyzing by synchrosqueezing transform," *PloS one*, vol. 11, n.º 6, e0157135, 2016.

[85] J. Ruiz y M. A. Colominas, "Wave-shape function model order estimation by trigonometric regression," *Signal Processing*, vol. 197, pág. 108 543, 2022.

[86] F. Auger, P. Flandrin, Y.-T. Lin et al., "Time-frequency reassignment and synchrosqueezing: An overview," *IEEE Signal Processing Magazine*, vol. 30, n.º 6, págs. 32-41, 2013.

[87] Z. Wu, J. Shi, X. Zhang, W. Ma, B. Chen et al., "Kernel recursive maximum correntropy," *Signal Processing*, vol. 117, págs. 11-16, 2015.

[88] B. Chen, L. Xing, H. Zhao, N. Zheng, J. C. Prı et al., "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, n.º 13, págs. 3376-3387, 2016.

Doctorado en Ingeniería
mención inteligencia computacional, señales y sistemas

Título de la obra:

Desarrollo de métodos robustos y fisiológicamente inspirados para el filtrado inverso de la voz

Autor: Iván Ariel Zalazar
Director: Gabriel Alejandro Alzamendi
Codirector: Gastón Schlotthauer

Palabras Claves:
Filtrado inverso de la voz,
Análisis de la señal de voz,
Flujo glótico,
Función glótica,
Predicción lineal ponderada,
Correntropía,
Modelado adaptativo no armónico,
Funciones de forma de onda.