Universidad de Las Palmas de Gran Canaria
Departamento de Señales y Comunicaciones

Programa de Doctorado

Cibernética y Telecomunicación

Tesis Doctoral

# Aportaciones sobre Modelado Paramétrico de Procesos Aplicados al Reconocimiento Automático del Habla

AUTOR: Pedro José Quintana Morales

DIRECTORES: Dr. Juan Luis Navarro Mesa

Dr. Antonio G. Ravelo García

Las Palmas de Gran Canaria, Noviembre 2015

*"La preocupación por conseguir un fin*
*nos intranquiliza a todos los hombres,*
*aun a los más desaprensivos,*
*aun a los más indolentes, . . . "*

Las inquietudes de Shanti Andía
Pío Baroja

# Agradecimientos

Quiero dedicar este trabajo a todos los que de alguna manera me han ayudado, me han apoyado, me han animado y también a los que me han empujado. En principio a mi Director de Tesis, Dr. Juan Luis Navarro Mesa, por aportarme otra visión del trabajo y de los resultados y siempre mirar a lo contribuido. También a mi otro Director Toni, por su nuevas ideas, su disposición y ayuda en todo momento, especialmente al final.

Por otro lado, los que estaban ahí detrás, ahí al lado, abajo, arriba, empujando, Victor, Francis, Pablo, Iván, Eugenio y Eduardo, ah y Rafa y Jose. También a todos los que no nombro y con los que siempre me he echado unas risas cuando me los cruzo por el pasillo.

En la otra dimensión, Rosa, Gabriel y Violeta. Al otro lado Marielo, Manolo y Cheque. Y en el más allá, a Mamá y a Papá, que pusieron lo tangible y lo intangible, que vieron el inicio y hoy les muestro el final.

¿Que cuando la leo? ... ya

# Resumen

El objetivo de este trabajo de investigación es el análisis de diferentes herramientas basadas en modelos lineales paramétricos en sistemas de reconocimiento de voz. La amplia literatura y la madurez alcanzada en el uso de estos modelos en diferentes áreas del procesado son una garantía, tanto por su potencialidad como por su flexibilidad y a la vez es un desafío, por buscar nuevos aspectos que puedan servir para mejorar la eficacia con nuevas herramientas o con enfoques diferentes. Nuestra idea es trabajar en un sistema de reconocimiento del habla con voces con patologías y dentro de él, actuar sobre dos subbloques aparentemente dispares dentro del procesador acústico, como son los elementos de modelado paramétrico y modelado acústico, de forma integral.

Dentro del modelado paramétrico, investigamos en la forma de extraer las características de la voz para obtener una mayor información y a la vez más detallada, que pudiera ofrecer un conocimiento más profundo del sistema de producción de voz. Para ello se emplean los modelos de polos comunes sobre tramas contiguas asociadas a las diferentes fases de las cuerdas vocales, abierta y cerrada, buscando la caracterización de los aspectos intrínsecos del proceso de producción y pensando en la robustez de las estimaciones. Por otro lado se introducen aspectos psicoacústicos en la parametrización, para que le confiera un mayor potencial desde el punto de vista del reconocimiento. Y todo lo anterior se configurará de forma que pueda formularse de una manera integral.

El otro elemento sobre el que investigamos es el del modelado acústico, en el cual estamos interesados en encontrar una caracterización estadística para unos parámetros acústicos de los que se desconocen sus propiedades estadísticas. Nuevamente los modelos paramétricos lineales nos brindan la posibilidad de su uso, dentro de la teoría espectral, para poder establecer un marco de predicción de la caracterización acústica por medio de la función de densidad de probabilidad. Esta tiene potentes propiedades para la aproximación de funciones, por lo que se esperan resultados competitivos.

Finalmente y para comprobar las posibilidades que se abren con el uso de las herramientas desarrolladas, éstas se utilizan en sistema de reconocimiento del habla para voces con patologías, en diferentes experimentos controlados.

# Índice general

# Índice de figuras

# Lista de Tablas

# Capítulo 1
# INTRODUCCIÓN

## 1.1. Modelado y Reconocimiento del habla

El modelado es una herramienta útil para el estudio de los procesos reales de cualquier naturaleza. A partir del conocimiento previo que se tiene del proceso y de sus observaciones se puede construir un modelo que permita emular su funcionamiento para someterlo a diversas condiciones que ayuden a profundizar en su conocimiento. Este modelo no es único, pues dependerá del punto de vista que se considere y no sólo de la relación directa que pudiera tener con el proceso físico. Además la elección del modelo tendrá que ver con los recursos, herramientas y tecnologías disponibles para él. Esta diversidad de modelos no hace sino aumentar el enriquecimiento del conocimiento en torno a los procesos.

El proceso de reconocimiento del habla es uno de esos procesos que tiene una importancia indudable por varios motivos. Uno de ellos es por su pertenencia al campo de la comunicación humana, concretamente al área de la comunicación hombre-máquina que siempre quiso emular al primero. Otro es debido a la cercanía con que lo percibimos en el desarrollo tecnológico en que estamos inmersos, con tantas aplicaciones actuales

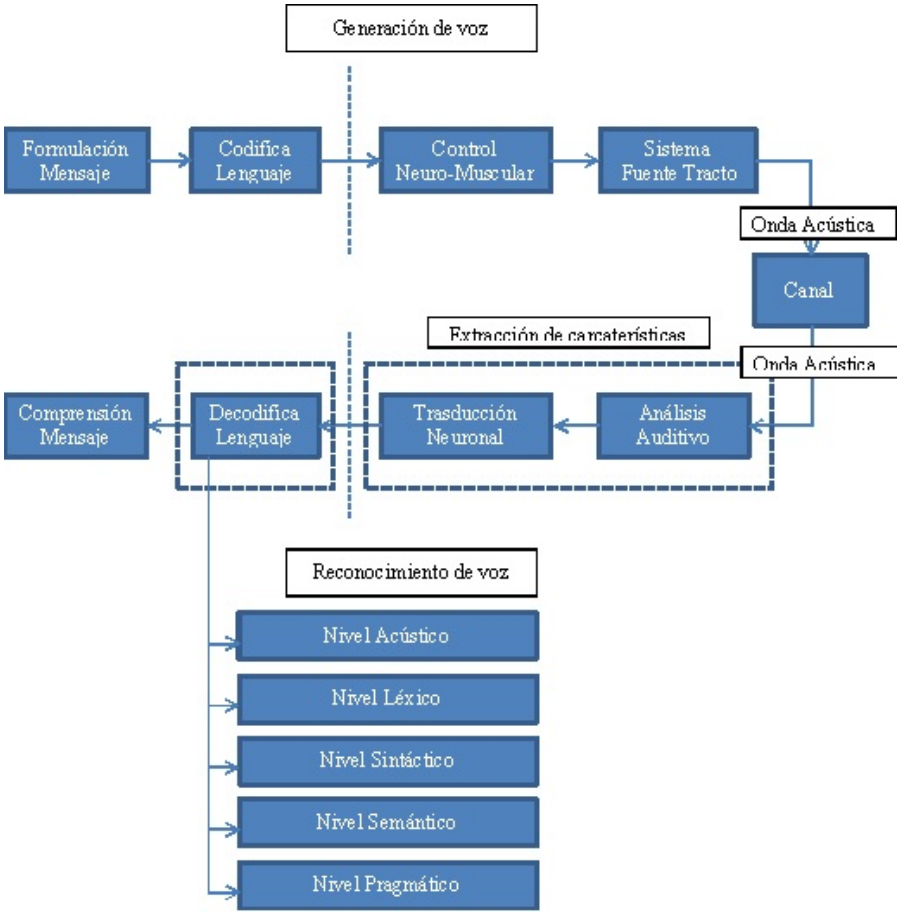



Figura 1.1: Proceso de Comunicación Humana

como posibilidades futuristas tiene. Es por ello que se lleva trabajando en esta área hace más de 60 años, desde el primer modelo de los laboratorios Bell publicado en 1952 y todavía se sigue investigando.

El proceso de comunicación humana se puede representar de una manera simplificada, como se ve en la figura 1.11. En el diagrama se pueden apreciar los dos grandes procesos en que se divide el sistema: el de generación y el de reconocimiento del habla. Nos interesa representar los dos procesos para resaltar la relación entre ambos. Queda claro que el reconocimiento del habla se apoya en el proceso de generación, en tanto en cuanto se analiza la señal acústica producida por aquel. El proceso de reconocimiento se compone de un proceso de análisis de señal, para extraer sus características, seguido de otro más general de decodificación del lenguaje, que englobarían todas las tareas encaminadas a entender el mensaje. Este último se puede analizar desde diferentes puntos de vista, dependiendo del conocimiento que se considere: el nivel acústico, el nivel léxico, el nivel sintáctico, el nivel semántico y el nivel pragmático.

Dentro del sistema de decodificación del lenguaje el nivel acústico ha jugado un pa-

pel motor esencial en el desarrollo de los sistemas de reconocimiento automático del habla (RAH). Tan es así que este nivel ha marcado en gran medida los hitos principales del desarrollo. Además cada uno de ellos define sus propias tareas. Primero fueron los sistemas basados en la teoría acústico-fonética, que segmentaban la señal y la etiquetaban en función de sus propiedades acústicas, [7]. Después se desarrollaron los modelos que empleaban una medida de distancia para comparar las características de la señal, trama a trama, con la de los diferentes patrones, desde un punto de vista determinista. El más relevante fue el Alineamiento Dinámico del Tiempo (DTW por sus siglas en inglés) [6]. Más tarde el concepto de igualación de patrones se extendió al área de los procesos estocásticos, dando lugar al desarrollo de los modelos estadísticos, que son los dominantes actualmente con el uso de los Modelos Ocultos de Markov (HMM por sus siglas en inglés) [8]. El modelado estadístico construye un modelo probabilístico, también llamado modelo acústico, para las características de las tramas de la señal. Mucho se ha investigado también con modelos de redes neuronales [2], que tratan de simular de alguna manera el comportamiento conexionista del cerebro humano. Otras direcciones tomadas han sido en el sentido de los modelos estadísticos, con estructuras más complejos y con modelado acústico basados en criterios discriminativos. Y en todo momento combinacines o estructuras híbridas entre ellos.

Por otro lado la extracción de características, como bloque fundamental, también evolucionaron conforme lo hacían los anteriores. En los primeros sistemas se emplearon las propiedades de la onda acústica, aquellas que surgían del proceso de producción de voz. Primero se usaron el conjunto de ellas que servían para etiquetar cada sonido: resonancias, sonoridad, etc. Después se utilizaron modelos generales de estimación espectral, inicialmente basados en bancos de filtros y con posterioridad en la transformada de fourier. Al mismo tiempo se desarrollaron los modelos fuente-filtro, que se apoyaban en la teoría acústica de fant [3] para caracterizar el proceso de producción de voz y que se definían por medio de los parámetros del filtro. Diferentes parámetros que se podían deducir se emplearon, como los coeficiente de predicción lineal (LPC por sus siglas en inglés) , los de reflexión, los coeficientes log-area (LAR por sus siglas en inglés) o los pares de líneas espectrales (LSP por sus siglas en inglés). Con ello se distinguían éstos, los modelos paramétricos, de los primeros, los no paramétricos. Por otra parte, el procesado homomórfico [10] aplicado sobre el modelo de Fant daba lugar al modelado cepstral, paramétrico si lo hacía desde el modelo de fuente o no paramétrico si lo hacía desde el análisis espectral. En todos estos modelos era evidente la falta de consideración del proceso perceptual, que se fue intercalando en ellos poco a poco. Así se incorporaron diferentes propiedades perceptuales, como el comportamiento no lineal en frecuencia del oído o la dependencia potencial con la intensidad. De estas aportaciones surgieron el modelado de bancos de filtros no uniforme, la transformada de fourier modificada en escala frecuencial perceptual, el modelado paramétrico modificado en frecuencia (LPCW por sus siglas en inglés), el modelado cepstral en escala perceptual (MFCC por sus siglas en inglés) , o el modelado paramétrico perceptual (PLP por sus siglas en inglés) . Otros modelos considerados han sido los basados en reden neuronales o los modelos discrimitativos . Los modelos dominantes en estos años son aquellos que incluyen infor-

mación perceptual, el modelado cepstral en escala perceptual y el modelado paramétrico perceptual. Pueden verse [4, 5, 11] para una revisión más detallada.

## 1.2. Motivación

En términos generales se podría decir que los sistemas de RAH actuales, dominados por el uso de HMM y parámetros MFCC, trabajan adecuadamente, con tasas de error menores al 5 % en condiciones de uso controladas [14, 15]. Esto quiere decir que el conjunto de voces para entrenar los modelos sean suficientemente representativos del conjunto de voces que lo usarán, o que el entorno de trabajo sea similar. En caso contrario la eficiencia del sistema sufre drásticamente y esta es una de las razones del porque no se ha extendido definitivamente a todas las áreas potenciales de aplicación. Entre esas áreas nos interesa resaltar el de reconocimiento del habla con voz patológica donde hay un déficit de representatividad de las voces y también de investigaciones, en parte debido a la falta anterior, del número de bases de datos para la investigación (afortunadamente existe, entre otras, la Kay Elemetric, [13]).

¿Cuales son las causas del mal comportamiento?

Del lado del modelado acústico y de forma general es que, evidentemente, el no tener datos suficientes de todas las situaciones posibles, éstas no se pueden caracterizar adecuadamente. Los HMM se definen de forma genérica con mezclas de distribuciones de densidad. Sin conocer el comportamiento estadístico de las características, la elección más frecuente ha sido la de mezclas de gaussianas, por basarse en el teorema del límite central, con lo que su comportamiento se resiente cuando no hay un número elevado de datos, ya sea por variabilidad o ruido.

Del otro lado, el del modelado de características, la parametrización, la causa será una de dos: o que el análisis no es capaz de extraer las características del proceso verdadero de producción de voz, por la variabilidad de las voces de entrenamiento, o que las características no pueden extraerse adecuadamente en el análisis de la onda acústica ni del proceso perceptual. En el caso de los MFCC, la pobre relación con el proceso de producción de voz le hace más vulnerable. Con un modelo perceptual más potente, los PLP son a veces preferibles a los anteriores. Aún así ambos trabajan con tramas uniformes y asíncronas con el pitch, obteniendo características promedio, vulnerables a la variabilidad. En un intento de mejorar el modelado de esa variabilidad del habla, creemos que los modelos paramétricos lineales pueden seguir ofreciendo oportunidades. El modelado paramétrico se aprovecha del conocimiento a priori del proceso, o de alguna suposición que se haga del mismo, para desarrollar modelos más eficientes y mejores según el principio de parsimonia. Los modelos lineales, además de fiarse de un número reducido de parámetros para representar los procesos, tiene propiedades interesantes, como pueden ser: su relación lineal en dominio temporal (espacial o de características), el numeroso y potente conjunto de herramientas de estimación de los parámetros del modelo, su capacidad para representar características espectrales basadas en resonancia y antiresonancia, incluso características generales, o su capacidad para aproximar curvas.

Para conseguir una parametrización que se acerque un poco más al verdadero proceso de producción de voz, y por tanto se defienda mejor frente a la variabilidad, una alternativa a las parametrizaciones anteriores sería trabajar de forma síncrona con el pitch y particularizando en las fases abierta y cerrada de las cuerdas vocales. En cada una de las fases el modelo fuente-filtro es diferente, por lo que parece lógico pensar en que el modelo se aproxima mejor al proceso real. Es lo que denominamos a lo largo de la tesis como "análisis localizado". Se propone emplear modelos paramétricos lineales para cada fase, por ser apropiados para representar resonancias y antiresonancias y utilizar métodos de estimación robustos basados en la idea del modelado acústico de salas. Adicionalmente, y como consecuencia de los buenos resultados que aportan las características perceptuales en los modelos de características, se plantea la modificación perceptual de los parámetros. Como mejora se sugiere optimizar la implementación.

Por otra parte, diferentes trabajos han planteado alternativas al modelado acústico mediante HMM que utilizan mezclas de distribuciones gaussianas, con distribuciones laplacianas generalizadas [9]. Ello revela que el teorema del límite central no debe ser la causa para dejar de emplear otras distribuciones. Nosotros proponemos usar una distribución basada en modelos paramétricos lineales, a través del modelado de la densidad espectral de potencia y su semejanza con la densidad de probabilidad, [1], por su propiedad de aproximación a cualquier curva arbitraria [12]. La idea se plantea tanto desde el punto de vista de una distribución única como de una mezcla de distribuciones, para que sea un poco más robusta. Con ello se pretende mejorar el modelado frente a la variabilidad por su capacidad de aproximación.

## 1.3. Objetivos

El objetivo principal de esta tesis es el de contribuir al modelado de los diferentes bloques que componen un sistema de reconocimiento automático del habla desde un punto de vista común, el de los modelos lineales paramétricos, generalizados en el de tipo autorregresivo de media móvil, (ARMA por sus siglas en inglés), enriqueciendo el conocimiento del sistema para enfrentarse mejor a la variabilidad (y al ruido). Concretamente, vamos a trabajar con diferentes modelos paramétricos en los bloques de extracción de características y de generación de patrones estadísticos con HMM's y que podemos sintetizar en los siguientes puntos:

Para el modelado de características proponemos una formulación compacta para un análisis localizado, preciso y robusto, frente al tamaño de las tramas de análisis, basado en el modelado lineal paramétrico ARMA. Dicho planteamiento será lineal, por fase, en voz sonora, y sobre un conjunto de tramas consecutivas que mantengan algunas características en común. A su vez se trazará su integración con el análisis de voz sorda. Las tareas serán:

- Estudio del modelado paramétrico de voz por fase glótica.
- Desarrollo de estimadores robustos y fiables para el modelado paramétrico ARMA de voz por fase glótica con información de tramas adyacentes con características

comunes.

- Desarrollo de una formulación matricial compacta para el análisis de voz por fases glóticas que incluya información de tramas adyacentes con características comunes.
  - Con polos comunes y cero particulares.
  - Con polos y ceros comunes.
- Estudio de la incorporación de aspectos psicoacústicos al modelado paramétrico ARMA basados en la modificación no lineal de la escala de frecuencia.
- Reformulación de la estimación del modelado paramétrico ARMA de voz por fase glótica aplicándole modificación no lineal en frecuencia, como principio psicoacústico.
- Aplicación de las técnicas anteriores a voz sorda.
- Validación de resultados con experimentos de modelado y seguimiento de formantes.

A nivel del modelado probabilístico de las señales, o de sus parámetros, proponemos utilizar un modelo lineal paramétrico de tipo ARMA, en una visión general, como una función densidad de probabilidades, a través de la convergencia de las funciones de probabilidad con modelos lineales, utilizando su semejanza con la densidad espectral de potencia y extenderla al caso de mezclas de funciones de densidad de probabilidad. Las tareas serán:

- Análisis de la obtención de la función densidad de probabilidad de un proceso basado en el cálculo de una función densidad espectral de potencia con modelo paramétrico AR y ARMA y su extensión a mezclas de densidades.
- Desarrollo e implementación de algoritmos de cálculo de la función de densidad de probabilidad con modelo paramétrico AR y ARMA y su extensión a mezclas de densidades.
- Validación con experimentos sobre variables aleatorias.

A nivel de la aplicación de reconocimiento proponemos introducir por un lado la formulación obtenida de la estimación de parámetros ARMA en el bloque de caracterización y por otro lado investigar en el modelado ARMA utilizado en el bloque de parametrización para la descripción de las observaciones en los HMM con el objetivo añadido de desarrollar una formulación adecuada para la actualización basada en EM. Las tareas serán:

- Estudio de la parametrización ARMA de voz, con y sin información psicoacústica en sistemas de reconocimiento de voz basados en HMM.

- Análisis del cálculo de la función densidad de probabilidad de observación en los HMM a través del cálculo de la función densidad espectral de potencia paramétrica ARMA.
- Desarrollo de algoritmos de entrenamiento de HMM basados en el cálculo de la probabilidad de observación por medio de la densidad espectral de potencia paramétrica ARMA.
- Validación de resultados con experimentos de reconocimiento .

## 1.4. Principales contribuciones

Las principales contribuciones de este trabajo de investigación han sido las siguientes:

1. El establecimiento de una formulación para diferentes métodos de estimación de parámetros en el modelado ARMA de la señal de voz por fase glótica, teniendo en cuenta las características comunes y particulares del modelo en tramas adyacentes. Primeramente, con ello se dota de un marco general de análisis sobre intervalos estacionarios, en los que trabajar con un periodo, o una parte de él, es solo un caso particular. Cuando se utiliza más de un periodo, disminuye la inestabilidad en el análisis para la extracción de características sobre tramas de tamaño menor a un periodo, sin llegar a promediar. Además permite un seguimiento más eficiente de las características de la señal, tanto las comunes como las particulares asociadas al proceso físico de producción de voz. Este análisis periodio a periodo es lo que hemos denomianado 'localizado' visto esto en el sentido de obtener mayor precisión temporal que si trabajamos por tramas.

2. La incorporación de información psicoacústica al modelado ARMA por fase glótica, dotando a estos modelos de una mayor capacidad para representar la señal de voz con información simultánea del proceso de producción más profundo, el de las fases glóticas, con el proceso de percepción. Como consecuencia el modelado proporcionará una mejor resolución perceptual a la vez que una mejor resolución temporal.

3. La extensión al modelo ARMA del uso de las funciones de densidad espectral de potencia de modelos paramétricos lineales como funciones de densidad de probabilidad, ya desarrollados en la literatura previa sobre modelos AR. Con esta extensión se alcanza una mayor flexibilidad en la definición de la función de probabilidad derivada de dichos modelos.

4. La utilización de las funciones de densidad de probabilidad definidas desde el modelado paramétrico lineal AR para la construcción de funciones de mezclas de probabilidades, obteniendo un modelado probabilístico potente y flexible.

5. Las mejoras alcanzadas en aplicaciones de reconocimiento automático del habla sobre voz patológica por el empleo del conjunto de parámetros obtenidos con el

modelado ARMA por fase glótica de la señal de voz, por su detallada información. En nuestras aportaciones damos las matemáticas que permiten obtener los parámetros que definen las funciones de densidad de probabilidad.

6. El funcionamiento competitivo en aplicaciones de reconocimiento automático del habla por el empleo del modelado probabilístico aproximaciones ARMA y con mezclas de funciones de densidad de probabilidad basadas en modelado paramétrico lineal AR, por su capacidad de aproximaciónen ambos casos.

## 1.5. Estructura de la Memoria

Esta memoria se ha estructurado en siete capítulos en los que se ha tratado de organizar la información de una forma eficiente.

En el primer capítulo se establece el marco de actuación, la motivación, los objetivos, las principales contribuciones y la estructura de la memoria.

En el segundo capítulo se revisa la teoría básica de los modelos paramétricos lineales, sobre los que versarán las aportaciones de este trabajo, señalando sus características, sus formas funcionales, sus métodos de estimación, su funcionamiento y en que aplicaciones se utilizan.

En el tercer capítulo se introduce el trabajo por fase glótica de la señal de voz y se desarrollan diferentes modelos sobre intervalos estacionarios y se deducen los correspondientes métodos de estimación de parámetros.

En el capítulo cuarto se trabaja en la modificación de los modelos para añadirles las características psicoacústicas y se vuelven a deducir los métodos de estimación de los parámetros.

En el capítulo cinco se desarrolla la teoría del modelado probabilístico de variables aleatorias con funciones de densidad espectral de potencia de modelos paramétricos lineales. Primero se extiende sobre modelado ARMA y después sobre mezclas de modelos AR.

En el capítulo seis se exponen las características del sistema en varias aplicaciones de reconocimiento automático del habla y se aplican los modelos desarrollados en los capítulos anteriores en el bloque de extracción de características y de modelado probabilístico y se analiza y evalúa su comportamiento.

En el capítulo siete se revisa el trabajo realizado, los objetivos alcanzados y las líneas que han quedado abiertas.

La bibliografía se incluirá por capítulos.

## Bibliografía

[1] M. B. Priestley, *Spectral analysis and time series*. Academic Press, 1981.

[2] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Computation*, vol. 1, no. 1, pp. 1–38, 1989.

[3] G. Fant, *Acoustic Theory of Speech Production*. Mouton De Gruyter, revised ed., Jan. 1971.

[4] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1993.

[5] J. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, pp. 1215 –1247, Sept. 1993.

[6] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43 – 49, Feb. 1978.

[7] J. Wiren and H. L. Stubbs, "Electronic binary selection system for phoneme classification," *The Journal of the Acoustical Society of America*, vol. 28, no. 6, pp. 1082–1091, 1956.

[8] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, pp. 532 – 556, Apr. 1976.

[9] N. Atsushi and A. I, "Acoustic modeling for speech recognition based on a generalized laplacian mixture distribution.," *IEICE Transactions on Information and Systems, Pt.2 (Japanese Edition)*, vol. J83-D-2, no. 11, pp. 2118–2127, 2000.

[10] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, 1968.

[11] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*, vol. 14, pp. 99–145, June 2011.

[12] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 2, pp. 211–216, 2002.

[13] M. Eye, E. I. Voice, and S. Lab, "Voice disorders database model 4337," 1994.

[14] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv:1001.2267*, Jan. 2010. International Journal of Computer Science and Information Security, IJCSIS, Vol. 6, No. 3, pp. 181-205, December 2009, USA.

[15] Vimala.C and V.Radha, "A review on speech recognition challenges and approaches," 2012.

Capítulo 2

# MODELADO PARAMÉTRICO LINEAL

## 2.1. Introducción

En este capítulo se va a revisar la teoría en la que se basa el modelado paramétrico lineal. Primero se definirá el modelo lineal como paramétrico, frente al modelado no paramétrico y se particularizará en sus diferentes tipos, repasándose sus propiedades. A partir del modelo general se expondrán la forma de implementarlo. A continuación se obtendrán las funciones que caracterizan al modelo en el dominio transformado, la función de transferencia del sistema y la densidad espectral de potencia y se mostrarán sus características. Después se examinarán los métodos de estimación para dichos modelos, sus parámetros y sus propiedades. Finalmente se reflexionará sobre la función de densidad espectral de potencia y su capacidad de representar otras funciones de densidad generales y, en particular, la función densidad de probabilidad.

## 2.2. Modelo paramétrico de sistemas lineales

Los procesos que representan sus observaciones en un instante determinado mediante una combinación lineal de las observaciones previas, junto con una combinación lineal de las muestras de otros procesos de los que dependen, caracterizan una clase de sistemas lineales muy usado en la práctica. Estos modelos se definen en el dominio discreto y se representan con ecuaciones lineales en diferencia de coeficientes constantes, como muestra la ecuación 2.1.

$$\sum_{k=0}^{p} a_k y(n-k) = \sum_{k=0}^{q} b_k x(n-k) \tag{2.1}$$

en el que $y\,(n)$ y $x\,(n)$ son los datos que intervienen en el proceso y los coeficientes $\{a_k\}$ y $\{b_k\}$ son los parámetros que caracterizan el sistema.

El hecho de suponer dicho comportamiento, ya por que sea verdad o por que permita una caracterización aproximada, simplificada o útil en términos de resultados, clasifica el modelo lineal como paramétrico. Parece obvio que los modelos paramétricos consigan representaciones más eficaces, en tanto en cuanto las suposiciones sean acertadas, en relación al modelado no paramétrico, que no utiliza ningún conocimiento a priori. Este comportamiento parece ser más acusado cuanto menor sea el número de datos disponibles de los que derivar el modelo [4]. Por contra, se resienten en entornos que puedan modificar las características del proceso.

El estudio de los modelos paramétricos lineales se ha desarrollado desde diferentes puntos de vista, entre los que destacamos el de los sistemas lineales y el de las series temporales. En el primer caso el estudio se realiza normalmente desde la óptica determinista y conlleva el concepto de sistema como filtro. De esta forma el modelo representa al sistema que actúa modificando la composición en frecuencia de la señal de entrada. En el caso de las series temporales, los datos se consideran de cualquier naturaleza, aunque principalmente desde procesos aleatorios y no siempre como una función temporal sino que se puede extender a cualquier otro dominio, como el espacial. Con ello se alcanza a conseguir un modelo probabilístico de los datos. En cualquier caso, el modelado paramétrico lineal tiene la misma formulación.

El modelo paramétrico de un sistema lineal en el dominio temporal discreto se puede definir, por tanto, a partir de la ecuación 2.1 considerando la señal $y\,(n)$ como la salida de un sistema caracterizado por los parámetros $\{a_k\}$, $0 \leq k \leq p$ y $\{b_k\}$, $0 \leq k \leq q$, a la señal de entrada $x\,(n)$, de acuerdo a la ecuación 2.2.

$$y(n) = \sum_{k=0}^{q} b_k x(n-k) - \sum_{k=1}^{p} a_k y(n-k) \tag{2.2}$$

La señal de salida en el instante actual depende de la salida en los instantes previos y de la entrada en los momentos actual y previos. La dependencia con las muestras previas

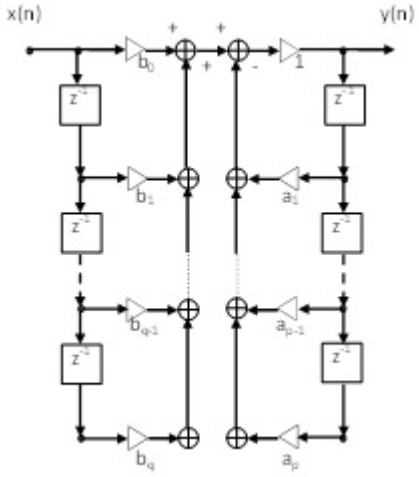



Figura 2.1: Estructura IIR tipo 1

de salida son de alguna manera una predicción, o regresión. En cambio, la dependencia con los datos de entrada es un promedio. Por ello a este modelo se le conoce como autorregresivo de medias móviles, o ARMA ("autoregressive moving average").

La implementación del modelo se puede hacer con estructuras realimentadas para la predicción y no realimentadas para el promediado, siendo la más sencilla la forma directa 1 [5], como se muestra en la figura 12.1.

La definición del sistema en el dominio de la frecuencia la podemos conseguir aplicando la transformada Z al modelo de la ecuación 2.1 y tomando la relación entre la entrada y la salida. Con ello se obtiene $H(z)$ en la ecuación 2.3, que representa la función de transferencia del modelo ARMA en la variable compleja $z = \sigma e^{jw}$, donde $\omega$ simboliza la frecuencia angular.

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{i=0}^{q} b_i z^{-i}}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.3}$$

Podemos ver como la función de transferencia del modelo paramétrico lineal es una función racional, caracterizada por los coeficientes de los polinomios numerador y denominador, que son los mismos que los de la ecuación en diferencia 2.1. Por ello el polinomio del numerador, $B(z) = \sum_{i=0}^{q} b_i z^{-i}$, se considera como la parte de que lleva a cabo el promediado móvil del sistema, o parte MA, mientras que el polinomio denominador, $A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$, se encarga del comportamiento autorregresivo del mismo, o parte AR. Debido al carácter racional de la función, ésta también se puede representar a través de los polos y los ceros del sistema, como se muestra en la ecuación

2.4, por los que al modelo ARMA también se le conoce como modelo polo-cero.

$$H(z) = K\frac{\prod_{i=1}^{q}(1 - c_i z^{-1})}{\prod_{k=1}^{p}(1 - p_k z^{-1})} \tag{2.4}$$

El sistema será estable y causal cuando los polos del modelo, $\{p_k\}$, estén dentro del círculo unidad. Si también estuviera dentro los ceros, $\{c_i\}$, sería invertible. Además, si el modelo es real, los coeficientes del modelo son reales y por tanto los polos, $\{p_k\}$, y los ceros, $\{c_k\}$, deberán aparecer en pares complejos conjugados o ser reales.

La evaluación de la función de transferencia en el círculo unidad, $z = e^{jw}$, nos proporciona la respuesta en frecuencia del sistema, $H(e^{jw})$. Particularizando en la ecuación 2.3 obtendremos dicha función, que es compleja, en la siguiente expresión

$$H(e^{jw}) = \frac{Y(e^{jw})}{X(e^{jw})} = \frac{\sum_{i=0}^{q} b_i e^{-jwi}}{1 + \sum_{k=1}^{p} a_k e^{-jwk}} \tag{2.5}$$

El comportamiento en frecuencia se puede deducir directamente de los polos y los ceros de la función. Siendo $p_k = \rho_k e^{j\theta_k}$ y $c_i = \nu_i e^{j\phi_i}$ y sustituyéndolos en la ecuación 2.4, la función de respuesta en frecuencia queda de la siguiente forma

$$H(e^{jw}) = K\frac{\prod_{i=1}^{q}(1 - \nu_i e^{j\phi_i} e^{-jw})}{\prod_{k=1}^{p}(1 - \rho_k e^{j\theta_k} e^{-jw})} \tag{2.6}$$

Analizando la ecuación 2.6 se puede observar como en las frecuencias vecinas a la de los polos, $\theta_k$, se producirá un pico en la magnitud de la respuesta en frecuencia, cuya altura depende de la cercanía de dicho polo a la circunferencia unidad, $\rho_k$, mientras que en la vecindad de los ceros, $\phi_i$, se producirán valles con profundidades que tendrán que ver con la cercanía de los ceros a la circunferencia unidad, $\nu_i$.

Desde el punto de vista de la teoría del filtrado, el modelo puede verse factorizado en términos reales, de orden 1 para raíces reales y orden 2 para raíces complejas conjugadas. En el caso de orden 1 los modelos representan filtros paso bajo, mientras que en el caso de orden 2 los modelos caracterizan filtros paso banda, centrados en las frecuencias de los polos, que definirían las frecuencias de resonancia. En ambos casos los anchos de banda son proporcionales a los módulos de las raíces, de la forma $BW_k \approx \frac{-ln(\rho_k)}{\pi}$. Desde el ángulo de las series temporales, en cambio, se estarían definiendo modos del modelo probabilístico.

## 2.3. Análisis espectral paramétrico

Cuando los procesos son aleatorios la forma de analizarlos es por medio de los estadísticos de 2º orden: la función de autocorrelación, ACF ("autocorrelation function") y la densidad espectral de potencia, PSD ("power spectral density"). Ambas funciones están

relacionadas por la transformada de Fourier. Los procesos han de ser estacionarios y ergódicos para que el análisis sea posible y existan las transformadas. Ello obliga a que el modelo 2.3 sea estable y causal, o sea, que tenga sus polos en el interior del círculo unidad.

Dado un proceso aleatorio discreto y estacionario en sentido amplio, definido por sistema lineal caracterizado por la ecuación en diferencias 2.2 y cuya función de respuesta en frecuencia viene expresada por 2.5, la densidad espectral de potencia queda determinada de la siguiente forma

$$S_{yy}(e^{jw}) = S_{xx}(e^{jw}) \left|H(e^{jw})\right|^2 = S_{xx}(e^{jw}) \frac{\left|\sum_{k=0}^{q} b_k e^{-jwk}\right|^2}{\left|1+\sum_{k=1}^{p} a_k e^{-jwk}\right|^2} \tag{2.7}$$

Al ser un modelo discreto, el dominio de la frecuencia se determina de manera única en el intervalo $\omega\epsilon[-\pi\ \pi]$ rad/seg.

Por otro lado está la función de autocorrelación, que como se refirió anteriormente se relaciona con la PSD a través del par de transformadas de Fourier siguiente, como establece el teorema de Wold [2] para el caso discreto

$$R_{yy}(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{yy}(e^{jw}) e^{jwk} dw \tag{2.8}$$

$$S_{yy}(e^{jw}) = \sum_{k=-\infty}^{\infty} R_{yy}(k) e^{-jwk} \tag{2.9}$$

Sin embargo, la función de autocorrelación también se puede especificar con el operador esperanza mediante la ecuación siguiente

$$R_{yy}(m) = E\{y^*(n) \cdot y(n+m)\} \tag{2.10}$$

Multiplicando y aplicando el operador esperanza se obtiene la relación entre los parámetros del proceso y la función de autocorrelación del modelo de la manera señalada a continuación.

$$R_{yy}(m) = \sum_{k=0}^{q} b_k R_{yx}(m-k) - \sum_{k=1}^{p} a_k R_{yy}(m-k) \tag{2.11}$$

donde $R_{yx}(m) = E\{y^*(n) \cdot x(n+m)\}$ define la correlación cruzada entre el proceso de entrada y el de salida.

Esta forma de relacionar los parámetros del modelo con las funciones de autocorre-

lación y de la correlación cruzada, es una herramienta útil para caracterizar y analizar los modelos lineales.

## 2.4. Tipos de modelos lineales

El modelo lineal definido por 2.2 se le conoce como **modelo ARMA(p,q)**. El proceso de entrada que lo alimenta puede ser de cualquier naturaleza. Una opción habitual es considerarlo como ruido blanco, con espectro plano. Con ello es la función de respuesta en frecuencia del modelo, expresada por 2.5, quien caracteriza completamente la PSD del proceso, como se ve a continuación.

$$S_{yy}(e^{jw}) = \sigma_x^2 \frac{\left|\sum_{k=0}^{q} b_k e^{-jwk}\right|^2}{\left|1+\sum_{k=1}^{p} a_k e^{-jwk}\right|^2} \tag{2.12}$$

donde $\sigma_x^2$ es la varianza del proceso de entrada $x\,(n)$. El comportamiento en frecuencia caracteriza funciones con picos y valles espectrales.

La relación entre las correlaciones y los parámetros se deduce de 2.11. Teniendo en cuenta que $R_{yx}(m) = \sigma_x^2 h^*(-m)$, se obtiene

$$R_{yy}(m) = \begin{cases} -\sum_{k=1}^{p} a_k R_{yy}(m-k) + \sigma_x^2 \sum_{k=0}^{q-m} h^*(k) b_{k+m} & m = 0, 1, \ldots, q \\ -\sum_{k=1}^{p} a_k R_{yy}(m-k) & m \geq q+1 \end{cases} \tag{2.13}$$

lo cual expresa una relación no lineal entre la ACF y los parámetros del modelo, al igual que entre estos últimos.

Cuando el sistema solo depende de las muestras de la entrada, 2.14, el modelo es conocido como **modelo de Medias Móviles, MA(q)** ("moving average"), ya que se comporta como un promediado de la entrada. En este caso $a_k = 0$, para $k \geq 1$.

$$y(n) = \sum_{k=0}^{q} b_k x(n-k) \tag{2.14}$$

Su función de transferencia será igual a

$$H(z) = \sum_{i=0}^{q} b_i z^{-i} \tag{2.15}$$

por lo que también se le llama modelo todo cero. La estabilidad queda por tanto garantizada.

La PSD quedará de la siguiente forma

$$S_{yy}(e^{jw}) = \sigma_x^2 \left| \sum_{k=0}^{q} b_k e^{-jwk} \right|^2 \tag{2.16}$$

El comportamiento en frecuencia caracteriza adecuadamente funciones con valles espectrales.

Particularizando en 2.13, la autocorrelación se relacionará con los parámetros del proceso MA de la manera siguiente:

$$R_{yy}(m) = \begin{cases} \sigma_x^2 \sum_{k=0}^{q-m} b_k^* b_{k+m} & m = 0, 1, \ldots, q \\ 0 & m \geq q+1 \end{cases} \tag{2.17}$$

Se vuelve a ver una relación no lineal entre los parámetros del modelo.

Cuando el sistema solo depende de las muestras previas de la salida, además de la entrada actual, el proceso se llama **modelo autorregresivo, AR(p)** ("autoregressive") y se representa en 2.18. Al ser ruido blanco el proceso de entrada, el modelo se puede ver como un sistema de predicción lineal. En este caso $b_k = 0$, para $k \geq 1$.

$$y(n) = x(n) - \sum_{k=1}^{p} a_k y(n-k) \tag{2.18}$$

La función de transferencia no tiene ceros, por lo que también se le conoce como modelo todo-polo.

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.19}$$

La PSD será igual a

$$S_{yy}(e^{jw}) = \frac{\sigma_x^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jwk}\right|^2} \tag{2.20}$$

El comportamiento en frecuencia caracteriza bien funciones con picos espectrales.

Las ecuaciones que relacionan la autocorrelación con los parámetros del proceso se obtienen de 2.11 con la condición de incorrelación entre la entrada y la salida, como se muestra a continuación

$$R_{yy}(m) = \begin{cases} -\sum_{k=1}^{p} a_k R_{yy}(m-k) & m \geq 1 \\ -\sum_{k=1}^{p} a_k R_{yy}(-k) & m = 0 \end{cases} \tag{2.21}$$

Estas son las ecuaciones de Yule-Walker, que en forma matricial y para $1 \leq m \leq p$, se puede poner como

$$\boldsymbol{r}_{yy} = \boldsymbol{R}_{yy}\boldsymbol{a} \tag{2.22}$$

donde

$$\boldsymbol{a} = [a_1 \; a_2 \; \cdots \; a_p]^T$$

$$\boldsymbol{r}_{yy} = [R_{yy}(1) \; R_{yy}(2) \; \cdots \; R_{yy}(p)]^T$$

$$\boldsymbol{R}_{yy} = \begin{bmatrix} R_{yy}(0) & R_{yy}(-1) & \cdots & R_{yy}(-p+1) \\ R_{yy}(1) & R_{yy}(0) & \cdots & R_{yy}(-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{yy}(p-1) & R_{yy}(p-2) & \cdots & R_{yy}(0) \end{bmatrix}$$

La matriz de correlaciones de este proceso, $\boldsymbol{R}_{yy}$, es toeplitz y además simétrica y semi definida positiva, debido a las propiedades de la autocorrelación. Estas ecuaciones siguen mostrando una relación no lineal entre los parámetros y la ACF, aunque si se supone conocida la función de autocorrelación del proceso, 2.22 representará un sistema de ecuaciones lineal en los parámetros, que caracteriza un proceso AR aleatorio estacionario.

Los tipos de modelos aún se pueden relacionar por medio de los teoremas de wold y de kolmogorov [4]. Con el primero se puede concluir que un modelo AR o ARMA de orden finito se puede aproximar con un modelo MA de orden ínfinito. Con el segundo se establece que un modelo AR de orden infinito puede representar un modelo MA o ARMA de orden finito. Con esos teoremas la equivocación en la suposición del modelo se puede mitigar aumentando el orden del modelo elegido hasta un valor suficientemente alto.

Otro aspecto a considerar es la capacidad que tiene los modelos lineales de aproximar el comportamiento de cualquier proceso con modelos de orden suficientemente grande, como han apuntado diferentes autores, como Partzen, [1] o [6].

## 2.5. Métodos de estimación

La estimación de los modelos lineales se ha llevado a cabo principalmente desde el punto de vista del principio de mínimos cuadrados. Estos métodos se aplican a sistemas cuyos parámetros se relacionan linealmente con los datos de entrada y salida y tratan

de minimizar los promedios del error cuadrático de las muestras de salida [7]. Aunque no se garantiza la estimación óptima, el tratamiento matemático es muy atractivo. Enfrente se hayan los métodos que si garantizan estimadores eficientes, que son los de máxima verosimilitud. Estos estimadores son de uso general, pero que necesitan conocer la distribución de probabilidad del proceso, de cuya complejidad deriva la de su tratamiento matemático. También se considera los estimadores basados en el método de los momentos, procedimientos empíricos y útiles que tampoco garantizan su eficiencia. En cualquier caso los métodos dependen del modelo a estimar.

En general se considera que el proceso de entrada es aleatorio y blanco, aunque el resultado es idéntico cuando la entrada es un impulso, ya que ambos tienen la misma ACF y por tanto el mismo espectro [7]. Por otro lado se establece que las observaciones forman un conjunto finito de muestras de tamaño N, que es lo que normalmente sucede en situaciones prácticas..

### 2.5.1. Estimación de modelos AR

Dado un modelo AR(p), definido por 2.18, **el método de los momentos** hace uso de la relación entre los parámetros del modelo y los estadísticos de 2º orden del proceso. Sustituyendo la muestras de la ACF, $R_{yy}(m)$ por una estimación de las mismas, $\hat{R}_{yy}(m)$, en 2.21 se obtiene la estimación del modelo resolviendo en $\{a_k\}$, para obtener $\{\hat{a}_k\}$. Se utilizan dos tipos de estimadores para la ACF[2], el estimador insesgado

$$\hat{R}_{yy}^{(is)}(m) = \frac{1}{N-|m|} \sum_{n=0}^{N-1-|m|} y(n)^* y(n+|m|) \tag{2.23}$$

y el estimador sesgado

$$\hat{R}_{yy}^{(se)}(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} y(n)^* y(n+|m|) \tag{2.24}$$

La diferencia entre ambos está en el escalado. Aunque dentro de 2.22 esta diferencia no tiene efecto, si que la tiene a la hora de definir sus propiedades. El estimador sesgado, además de tener una menor varianza en los extremos, es una función semi definida positiva, cosa que no se garantiza en el estimador insesgado, y que la define como una función de autocorrelación de un proceso discreto estacionario en sentido amplio.

Sin embargo la forma más común de enfocar la estimación del modelo AR(p), dado por 2.18 es desde el punto de vista de la predicción lineal [7], en que la salida en cada instante se puede ver compuesta por la contribución de una predicción lineal, $\hat{y}(n) = \sum_{k=1}^{p} a_k y(n-k)$ más una componente de ruido, o de error de la predicción, $x(n)$. Desde este contexto lineal la teoría de **mínimos cuadrados** se emplea apropiadamente para llevar a cabo el proceso de estimación minimizando el error de predicción, que se

define como:

$$e(n) = y(n) - \hat{y}(n) = y(n) - \sum_{k=1}^{p} a_k y(n-k) \tag{2.25}$$

Teniendo en cuenta que el conjunto de observaciones es finito, el error de predicción se podrá expresar entonces completamente en forma matricial de la siguiente manera

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Y}\boldsymbol{a} \tag{2.26}$$

donde

$$\boldsymbol{e} = \begin{bmatrix} e(0) & e(1) & \cdots & e(p) & \cdots & e(N-1) & e(N) & \cdots & e(N-1+p) \end{bmatrix}^T \tag{2.27}$$

$$\boldsymbol{y} = \begin{bmatrix} y(0) & y(1) & \cdots & y(p) & \cdots & y(N-1) & 0 & \cdots & 0 \end{bmatrix}^T \tag{2.28}$$

$$\boldsymbol{Y} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ y(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y(p-1) & y(p-2) & \cdots & y(0) \\ \vdots & \vdots & \ddots & \vdots \\ y(N-2) & y(N-3) & \cdots & y(N-1-p) \\ y(N-1) & y(N-2) & \cdots & y(N-p) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & y(N-1) \end{bmatrix} \tag{2.29}$$

$$\boldsymbol{a} = \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix}^T \tag{2.30}$$

En el caso de considerar el error en todo el dominio temporal la solución de mínimos cuadrados la proporciona el **método de la Autocorrelación**, el cual minimiza

$$E = \sum_{n=-\infty}^{\infty} [e(n)]^2 \tag{2.31}$$

Suponiendo cero las muestras no conocidas, el error 2.31 en forma matricial se expresa

como

$$E = \sum_{n=0}^{N-1+p} [e(n)]^2 = (\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{a})^H (\boldsymbol{y} - \boldsymbol{Y}\boldsymbol{a}) \tag{2.32}$$

Derivando 2.32 respecto al vector de parámetros e igualando al vector cero, se obtiene el modelo estimado, que queda determinado por

$$\boldsymbol{Y}^H \boldsymbol{y} = \left(\boldsymbol{Y}^H \boldsymbol{Y}\right) \hat{\boldsymbol{a}} \tag{2.33}$$

Los elementos de la matriz $\boldsymbol{Y}^H\boldsymbol{Y}$ y del vector $\boldsymbol{Y}^H\boldsymbol{y}$, después de escalarlos con el factor $1/N$, son idénticos a los que vienen dados por el estimador sesgado de la autocorrelación y que son iguales a

$$\hat{R}_{yy}^{(se)}(j-i) = \left[\boldsymbol{Y}^H\boldsymbol{Y}\right]_{ij} = \left[\boldsymbol{Y}^H\boldsymbol{y}\right]_{ij} = \frac{1}{N} \sum_{n=0}^{N-1-|j-i|} y(n+|i|)^* y(n+|j|) \tag{2.34}$$

Con estos elementos el modelo 2.33 es igual al representado por las ecuaciones de Yuler Walker, con el estimador sesgado de la autocorrelación que permite caracterizar la matriz de correlaciones como verdadera, a saber Toeplitz, simétrica y semidiefinida positiva. La estimación de los parámetros viene dado por

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{Y}^H\boldsymbol{Y}\right)^{-1} \boldsymbol{Y}^H \boldsymbol{y} \tag{2.35}$$

Cuando el error a minimizar se contabiliza solamente sobre las muestras del proceso que se obtienen a partir de las muestras observadas, sin hacer suposiciones, la solución la proporciona el **método de la Covarianza**, que minimiza

$$E = \sum_{n=p}^{N-1} [e(n)]^2 = (\boldsymbol{y}_f - \boldsymbol{Y}_f\boldsymbol{a})^H (\boldsymbol{y}_f - \boldsymbol{Y}_f\boldsymbol{a}) \tag{2.36}$$

en el que la matriz $\boldsymbol{Y}_f$ y el vector $\boldsymbol{y}_f$ representan una forma reducida de 2.29 y de 2.28 respectivamente, que contienen únicamente las filas desde la $p$ hasta la $N-1$. El modelo entonces viene dado por

$$\boldsymbol{Y}_f^H \boldsymbol{y}_f = \left(\boldsymbol{Y}_f^H \boldsymbol{Y}_f\right) \hat{\boldsymbol{a}} \tag{2.37}$$

y la la solución de mínimos cuadrados será

$$\hat{\boldsymbol{a}} = \left(\boldsymbol{Y}_f^H \boldsymbol{Y}_f\right)^{-1} \boldsymbol{Y}_f^H \boldsymbol{y}_f \tag{2.38}$$

similar a 2.35. Ahora los elementos de la matriz $\boldsymbol{Y}_f^H \boldsymbol{Y}_f$ y del vector $\boldsymbol{Y}_f^H \boldsymbol{y}_f$, después de escalarlos con el factor $1/N-p$, son similares a los que vienen dados por un estimador insesgado de la covarianza y que se expresan como

$$\hat{C}_{yy}^{(is)}(i,j) = \left[\boldsymbol{Y}_f^H \boldsymbol{Y}_f\right]_{ij} = \left[\boldsymbol{Y}_f^H \boldsymbol{y}_f\right]_{ij} = \frac{1}{N-p} \sum_{n=p}^{N-1} y(n-|i|)^* y(n-|j|) \tag{2.39}$$

Con estos elementos el modelo estimado tiene una matriz de correlaciones simétrica pero no Toeplitz, ni necesariamente semi definida positiva, que garantice que representa a un proceso estacionario en sentido amplio. Sin embargo, para conjuntos pequeños de datos observados el método de la Covarianza produce estimaciones con mejor resolución que las del método de la Autocorrelación, que de alguna manera enventana los datos.

Otros estimadores toman en cuenta no solo el error de estimación basado en estimaciones hacia adelante, o forward, que utiliza datos anteriores, como el considerado hasta ahora y expresado por

$$e^f(n) = y(n) - \sum_{k=1}^{p} a_k y(n-k) \tag{2.40}$$

sino que emplea también el error que hace uso de muestras posteriores, al estilo de una predicción hacia atrás, llamada predicción backward y dada por

$$e^b(n) = y(n) - \sum_{k=1}^{p} a_k^* y(n+k) \tag{2.41}$$

El estimador en estos casos minimizan el error de predicción conjunto, forward y backward, sobre el intervalo del proceso realmente definido por las observaciones

$$E = \sum_{n=p}^{N-1} \left[e^f(n)\right]^2 + \sum_{n=0}^{N-1-p} \left[e^b(n)\right]^2 \tag{2.42}$$

El **método de la Covarianza modificada** minimiza el error de predicción definido por 2.42 y viene determinado por un modelo análogo a 2.37 en el que los elementos de

la matriz de correlaciones vienen dados por

$$\hat{C}_{yy}^{(is)}(i,j) = \frac{1}{N-p}\left(\sum_{n=p}^{N-1} y(n-|i|)^* y(n-|j|) + \sum_{n=0}^{N-1-p} y(n+|i|)y(n+|j|)^*\right) \quad (2.43)$$

Este método no garantiza que la matriz de correlaciones sea definida positiva pero en la práctica obtiene estimaciones de procesos estables, con alta resolución y trabaja mejor que los anteriores con conjunto de datos pequeños al utilizar más información.

El **método de Burg** minimiza también el error 2.42, pero de forma recursiva sobre modelos AR(i) de distintos ordenes, empezando por $i = 1$ y suponiendo conocido el modelo AR(i-1). Los modelos AR(i-1) se estiman asegurando su estabilidad. El método estima los parámetros de reflexión, $\{k_i\}$, que están definidos por $\left\{k_i = a_{ii} = a_i^{AR(i)}\right\}$ y que lo hace de la siguiente forma

$$\hat{k}_i = \frac{-\sum_{n=i}^{N-1} e_{i-1}^f(n) e_{i-1}^b(n-1)^*}{\sum_{n=i}^{N-1}\left(\left|e_{i-1}^f(n)\right|^2 + \left|e_{i-1}^b(n-1)\right|^2\right)} \quad (2.44)$$

En la expresión anterior, $e_{i-1}^f(n)$ y $e_{i-1}^b(n)$ hacen alusión a los errores forward y backard respectivamente, dados por 2.40 y 2.41 cada uno de ellos, pero referidos al modelo AR(i-1).

Este estimador produce estimaciones de procesos estacionarios, con baja varianza y alta resolución.

Estimaciones de **máxima verosimilitud** conducen a sistemas de ecuaciones no lineales complejas de tratar. El problema deriva de la dificultad de manejar la distribución de probabilidad conjunta entre datos y parámetros en los $p$ momentos iniciales del modelo AR(p), en el que las muestras del proceso aleatorio que se quiere caracterizar no quedan definidas completamente por las observaciones disponibles. Aproximaciones basadas en la probabilidad condicional, condicionadas al conocimiento de las $p$ primeras observaciones y suponiendo el proceso de entrada a ser ruido blanco Gaussiano, obtienen resultados idénticos a los de mínimos cuadrados. Cuando el número de observaciones es elevado el efecto de tomar solo la probabilidad condicional puede ser despreciado. Por ello los métodos revisados anteriormente, que fueron soluciones de mínimos cuadrados, pueden considerarse como aproximaciones de máxima verosimilitud.

### 2.5.2. Estimación de modelos MA

Para los modelos MA(q), definidos por 2.14, la falta de relación lineal de sus parámetros en su relación con la ACF, dada en 2.17, o del proceso de entrada como combinación lineal finita de las observaciones ha hecho que los métodos de estimación de los momentos o de máxima verosimilitud requieran soluciones complejas, normalmente no lineales.

Frente a ellas surgen las aproximaciones basadas en las relaciones entre los modelos lineales. La que ha tenido más éxito ha sido el **método de Durbin**, que convierte el modelo MA(q), dado por 2.14, en un modelo AR(p) equivalente, definido por 2.18, de orden suficientemente grande para que se suponga aplicado el teorema de kolmogorov y obtiene la estimación de los parámetros del modelo equivalente $\{\hat{a}_k\}$ por lo métodos de estimación de modelos AR. Entonces aplica la teoría de máxima verosimilitud sobre los parámetros estimados del modelo autorregresivo equivalente para conseguir la estimación del modelo MA(q), dado por $\left\{\hat{b}_k\right\}$.

Esta estimador es una aproximación robusta de máxima verosimilitud que se comporta muy satisfactoriamente con una elección apropiada del orden del modelo intermedio AR(p) y además garantiza invertibilidad [8].

### 2.5.3. Estimación de modelos ARMA

Los modelos ARMA(p,q), definidos por 2.2 y vistos como procesos que contienen una parte AR y otra MA, tendrá las dificultades asociadas a ambos en cuanto a la construcción de estimadores. Los estimadores de máxima verosimilitud conllevan una alta no linealidad. De nuevo las aproximaciones subóptimas basadas en las relaciones entre los modelos lineales son las más adecuadas para tratar con estimadores manejables. Lo más útil de cualquier manera resulta ser estimar separadamente las partes MA y AR.

El **método de las ecuaciones de Yule Walker extendidas** hace uso de la relación entre los parámetros del modelo y la ACF, dado por 2.13. Una estimación de los parámetros de la parte AR se pueden obtener para $m \geq q+1$, sustituyendo las muestras de la ACF, $R_{yy}(m)$, con una estimación de las mismas. Con los parámetros del modelo AR se filtra la salida de forma inversa y se obtiene un proceso MA, del cual se pueden estimar los parámetros del modelo MA con los métodos de estimación vistos en el apartado anterior.

Otra forma de estimar el modelo ARMA(p,q) es convertirlo en un **AR(r) de orden suficientemente grande**, para que sea equivalente según el teorema de Wold. Igualando los modelos, los parámetros quedan relacionados, pudiendo obtener primero la estimación de la parte MA, $\left\{\hat{b}_k\right\}$ y luego de la parte AR, $\{\hat{a}_k\}$. Igualmente se puede convertir en un modelo equivalente **MA(r) de orden suficientemente alto**, siguiendo el teorema de Kolmogorov y obtener los parámetros del modelo ARMA(p,q) en sentido inverso al caso anterior.

Una técnica más es la de **mínimos cuadrados por identificación entrada-salida**, la cual estima, o supone, el proceso de entrada y con ello su correlación cruzada con la salida. El problema se convierte entonces en uno similar a un proceso AR.

## 2.6. Orden del modelo

Un aspecto intrínseco en la estimación del modelo es la determinación del orden de los mismos. Una elección inadecuada puede dejar de representar determinadas características frecuenciales, si el orden es menor del verdadero, o puede provocar la aparición

de espúreos, si el orden es mayor. Los métodos van a depender de la varianza del residuo de proceso, o error de predicción, que indica la eficacia de la predicción asociada a los modelos lineales, del orden del modelo, que penaliza órdenes altos que puedan introducir espúreos y del número de datos.

Diferentes métodos se han propuesto, entre los que destacan: El Error final de predicción, específico de los modelos AR; el Criterio de información de Akaike, de uso general que representa una distancia de Kullback-Leibler con las funciones de densidad de probabilidad; el Criterio de información Bayesiano, que utiliza la estimación de máxima verosimilitud de la varianza del residuo además de la varianza de las observaciones y llegan a órdenes menores que el anterior; el Criterio de la transferencia autorregresiva, que obtiene el orden que aproxima mejor el filtro de error al optimo de tamaño infinito; y el Criterio de mínima longitud de descripción.

## 2.7. Función de densidad

La PSD representa una función de densidad, que se deriva de su relación con la transformada de Fourier de la autocorrelación, dada por 2.8, que en el origen es la potencia del proceso y es igual a la integral en todo el dominio. Además es positiva, por ser de naturaleza cuadrática. Con estas dos propiedades, Priestley [2] ya apuntó su semejanza con la función densidad de probabilidad, cuyo acrónimo en inglés es PDF.

Por un lado, una PDF arbitraria, $f(w)$, tiene las siguientes propiedades:

1.

$$f(w) \geq 0, \qquad \forall w \tag{2.45}$$

2.

$$\int_{-\infty}^{\infty} f(w)dw = 1 \tag{2.46}$$

Por el otro lado, la PSD cumple con lo siguiente:

1.

$$S_{yy}(e^{jw}) \geq 0 \tag{2.47}$$

$$R_{yy}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{yy}(e^{jw})dw \tag{2.48}$$

que arreglándolo adecuadamente queda como

2.

$$\int_{-\pi}^{\pi} \frac{S_{yy}(e^{jw})}{2\pi R_{yy}(0)} dw = 1 \tag{2.49}$$

Si el proceso es real

3.

$$S_{yy}(e^{jw}) = S_{yy}(e^{-jw}) \tag{2.50}$$

Por tanto, se puede decir que

$$p_+(w) = \begin{cases} \frac{S_{yy}(e^{jw})}{\pi R_{yy}(0)} & 0 \leq w \leq \pi \\ 0 & 0 > w > \pi \end{cases} \tag{2.51}$$

cuando el proceso es real y

$$p(w) = \begin{cases} \frac{S_{yy}(e^{jw})}{2\pi R_{yy}(0)} & -\pi \leq w \leq \pi \\ 0 & -\pi > w > \pi \end{cases} \tag{2.52}$$

cuando el proceso no es real, representan una PDF.

En cualquier caso, la igualdad entre ambas permite compartir propiedades y relaciones. Así, la ACF, siendo la transformada de Fourier de la PDS, como se ve en 2.8, será equivalente a la transformada de Fourier de la PDF, que es la función característica [3]. Y lo mismo ocurre con su propiedad de función definida semi positiva, que heredará dicha función característica.

## Bibliografía


[1] A. Poritz, "Linear predictive hidden markov models and the speech signal," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, pp. 1291 – 1294, May 1982.

[2] M. B. Priestley, *Spectral analysis and time series.* Academic Press, 1981.

[3] A. Papoulis, *Probability, random variables, and stochastic processes.* McGraw-Hill, 1984.

[4] S. M. Kay, *Modern spectral estimation: theory and application.* Prentice Hall, 1988.

[5] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing.* Prentice Hall, 1989.

[6] P. Broersen, "Facts and fiction in spectral analysis," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, pp. 766 –772, Aug. 2000.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561 – 580, Apr. 1975.

[8] P. M. T. Broersen, "Autoregressive model orders for durbin's MA and ARMA estimators," *IEEE Transactions on Signal Processing*, vol. 48, no. 8, pp. 2454–2457, 2000.

Capítulo 3

# ANÁLISIS LOCALIZADO DE LA SEÑAL DE VOZ

## 3.1. Introducción

El modelado de la señal de voz requiere de un análisis de características que extraiga los atributos más importantes de la señal. Para ello, aparte del uso que se hace de los métodos no paramétricos basados en los bancos de filtros o en la transformada de Fourier, se utilizan los modelos paramétricos de producción de voz, basados en el modelo de Fant [8]. El objetivo es representar lo más fielmente posible el proceso de generación de la voz, con la intención de obtener características invariantes, robustas y perceptualmente discriminantes.

En este capítulo se llevará a cabo el análisis de la señal de voz con el que buscamos una precisión temporal lo más alta posible, esto es, localizado en el tiempo. Para alcanzar esta precisión con voz sonora hacemos un análisis localizado a nivel de periodo de vibración de las cuerdas vocales pues, como veremos, es posible establecer con suficiente precisión el intervalo de duración de los periodos. Estos intervalos se definen en

función de los instantes de cierre glótico. El punto de vista como sistema está, pues, en el funcionamiento de las cuerdas vocales, que nos dará la oportunidad de encarar un estudio más detallado de las características de la señal. Particularmente, estableceremos un modelo para cada una de las fases de las cuerdas vocales y desarrollaremos diferentes métodos de estimación para alcanzar un análisis robusto y fiable.

Para el caso de voz sorda un análisis localizado es más difícil realizar pues, a diferencia de la voz sonora, no disponemos de unos instantes especiales que permitan delimitar intervalos de análisis. Se impone aquí establecer unos criterios de análisis que permitan obtener una frecuencia de estimación de parámetros equiparable a la obtenida con la voz sonora.

## 3.2. El proceso de producción y la sonoridad

El proceso de producción de voz es llevado a cabo por el aparato fonador, que se compone de diferentes órganos, como son los pulmones, la traquea, la laringe, que contiene a la cuerdas vocales, la faringe, el tracto bucal, los labios, el velo del paladar, el tracto nasal y la nariz. La generación del sonido comienza con la expulsión de aire, a gran velocidad, desde los pulmones, pasando por la traque y la laringe, hasta llegar a las cuerdas vocales. Si éstas se tensan y se juntan suficientemente, puede ocurrir que empiecen a vibrar, abriéndose y cerrándose la abertura que hay entre ellas, llamada glotis. Los sonidos producidos de esta manera dan lugar a señales periódicas, caracterizadas por la frecuencia de vibración de las cuerdas vocales, o pitch y se denominan sonoros. En contraposición, los llamados sonidos sordos se generan manteniendo la glotis completamente abierta. Vibren o no las cuerdas vocales, la columna de aire atraviesa la faringe y se dirige hacia la cavidad bucal, para ser expulsada por los labios y produciendo sonidos orales, o hacia la cavidad nasal, generando sonidos nasales que salen por la nariz. Los sonidos se pueden clasificar de distintas maneras. Por ejemplo, dependiendo del modo de articulación en la cavidad bucal se tiene los tipos de sonidos vocálicos, oclusivos, fricativos, africados, laterales o vibrantes. En cualquier caso, los sonidos son modificados espectralmente por las características del sistema que atraviesa y que varían según la clase de sonido que sea. Esto es debido a las propiedades de las cavidades implicadas que hay por encima y por debajo de la glotis. La sucesión de sonidos es lo que constituye la voz, o el habla si tiene algún sentido.

Las características principales del proceso de producción de voz tienen que ver entonces con las características de los elementos que lo componen, como son, la energía del flujo de aire, si las cuerdas vocales están vibrando o no, la frecuencia de vibración y las resonancias, antiresonancias o amortiguamientos, que están presentes en el sistema que se forma en el proceso de producción.

## 3.3. Modelado de la señal de voz desde el proceso de producción

Los primeros trabajos en el campo de la caracterización del proceso de producción de voz, que encontramos en [8, 15], mostraron poca interacción entre la fuente del sonido, considerada como el volumen de aire expelido por los pulmones, modificado o no por la glotis, y el sistema vocal, que comprende el conducto por donde transita la fuente y que abarca el tracto bucal y el tracto nasal según sea el caso. Esto dio lugar a un modelo, conocido como modelo de Fant, que separa linealmente la contribución de fuente de la del sistema, facilitando su tratamiento. La fuente puede ser representada de forma simplificada como un generador de ruido, más o menos plano, para sonidos sordos, o como un tren de pulsos cuasi-periódicos, para sonidos sonoros. Además de esto la glotis introduce una pendiente de -12 dB/octava cuando vibran las cuerdas vocales, que es equivalente a un sistema con 2 polos, produciendo una señal glótica de aspecto más o menos triangular. Por otro lado está el sistema vocal que se caracteriza por diversas contribuciones. Una de las contribuciones es la del tracto bucal, que va desde la glotis a los labios y cambia de forma con el modo de articulación, se describe por sus frecuencias de resonancia, o formantes y se puede representar por un sistema todo-polo. Otra de las contribuciones es la del tracto nasal, que comprende desde el velo del paladar a la nariz, incluyendo además frecuencias de antiresonancias que se describen con ceros, dando lugar a sistemas polo-cero. También se pueden introducir ceros en el sistema cuando el punto de excitación no está en la glotis, como en algunos sonidos sordos. Finalmente también contribuye el efecto de la radiación del sonido hacia afuera del aparato fonador, que pre-enfatiza la señal con el efecto de un sistema con un cero.

Los modelos de producción de voz se desarrollaron inicialmente en el área de la codificación y la síntesis de voz, proponiéndose modelos AR para caracterizar todo el sistema vocal [18, 19, 20]. Con los modelos todo-polo se intentan aproximar los ceros que pudiera tener el sistema vocal con múltiples polos, siguiendo el teorema de Kolmogorov, para tratar con métodos lineales de estimación [16], como los revisados en el capítulo anterior, obteniéndose resultados satisfactorios. Sin embargo los ceros de los sonidos nasales y de algunos sonidos sordos no son fáciles de aproximar con polos adicionales, habiendo el peligro adicional de que aparezca realzado algunas otras frecuencias no deseadas. La necesidad de una representación más precisa de esos sonidos hizo que se plantearan modelos ARMA para trabajar con modelos polo-cero [21, 22]. La determinación de evitar sistemas no lineales que surgen en el planteamiento de soluciones a los sistemas polo-cero, inclina al uso de métodos de estimación subóptimos [16], basados en los modelos todo-polos y que se revisaron en el capítulo anterior.

Estos modelos representan adecuadamente el proceso de producción de voz, esperándose de ellos características de invarianza y cierta robustez frente al ruido aditivo. Lo invariante que sea tiene algo que ver con su mayor o menor aproximación a algún aspecto del proceso físico que varíe poco respecto al locutor o su circunstancia y ello está relacionado más con las características del tracto y menos con las de la fuente. Mientras, la robustez frente al ruido aditivo se ve beneficiada por su interpretabilidad

física [9] así como su linealidad del modelado [11].

De cualquier forma en las aplicaciones de reconocimiento se buscan que los modelos aporten parámetros, que además de que puedan representar el proceso físico, sean no solo invariantes y robustos, sino discriminativos con las herramientas de reconocimiento utilizadas. En este sentido se han propuesto otros tipos de modelos que se basan en representar el proceso de producción de voz desde un punto de vista no paramétrico. Estos modelos utilizan el procesado homomórfico [23] para separar la contribución de fuente y del sistema vocal en el dominio espectral, en un contexto no lineal, y volver al dominio temporal mediante transformadas, obteniéndose los cepstrum. Esta última transformada, alejando a los parámetros de un significado físico respecto del proceso de producción, les aporta una buena decorrelación entre ellos que les confiere características discriminativas, como podemos ver en los parámetros obtenidos con esta idea, como lo FFT-Cepstrum o los Mel Frequency Cepstrum Coefficient [11, 9].

Otra opción del modelado con cepstrum es obtenerlos a partir de los coeficientes del modelado paramétrico, como los LPC-Cepstrum [11]. Con ello se pueden obtener coeficientes con un mayor grado de decorrelación, que le aporta características discriminativas, desde el tratamiento lineal del modelado paramétrico.

### 3.3.1. Estimación Clásica de los Parámetros del Modelo

El proceso de producción de voz es un proceso estocástico no estacionario. No obstante desde una óptica estacionaria de cada uno de los diferentes sonidos y de la condición de ergodicidad se puede llevar a cabo su estudio. Este punto de vista plantea la definición del intervalo de análisis para la extracción de las características. La manera natural sería mediante la segmentación de tramos localmente estacionarios de la señal de voz, que pudiera incluir un preénfasis, para compensar la caída glotal y un enventanado, que minimizara el error introducido por el proceso de la segmentación [11, 24].

La forma usual de llevarlo a cabo en las aplicaciones de reconocimiento ha sido mediante un proceso asíncrono en términos de eventos relevantes (p.e., ciclos de apertura y cierre de las cuerdas vocales). Tomamos intervalos fijos de entre 20 a 30 ms con solapamientos entre el 30 % y el 60 %. Es una manera automática y fácil, en la que el análisis se efectúa adecuadamente en las mayoría de las vocales, que incluyen varios periodos de pitch y que además son los fonemas más importantes, los que llevan la mayor parte de la energía. Por contra, el análisis no se acomoda a las consonantes, de menor tamaño, donde es dudoso que coincidan exactamente dentro de las tramas de análisis [9, 24]. En todo caso aparecerán conjuntamente con otros sonidos. El mayor o menor tamaño del intervalo de análisis tendrá que ver entonces con la precisión de la estimación, por el posible solapamiento de características de sonidos diferentes, que producirá variaciones en los formantes y sus anchos de banda [9].

El orden de los modelos empleados tanto AR como ARMA tendrían que contemplar por lo menos un par de polos complejos conjugados por cada kHz [3], más 1 o 2 pares más en caso de entornos ruidosos. Los ceros, en el caso de los modelos ARMA, deberían incluir 1 o 2 pares complejos conjugados, para contemplar los sonidos nasales y 1 par más para representar la radiación. Si se utilizan modelos AR, sería conveniente considerar 1

o 2 pares más para sustituir a cada cero de los nombrados anteriormente.

La estimación asíncrona de modelos paramétricos en aplicaciones de reconocimiento tiene diversas ventajas. Matemáticamente es sencilla y precisa en muchos momentos, llevando el trabajo linealmente. Alcanzan una representación adecuada del proceso de producción de voz y son capaces de separar razonablemente la contribución de la fuente de la del tracto vocal. El tracto puede caracterizarse por un número de parámetros finito con una dimensión relativamente baja. Al representar adecuadamente el proceso físico, la estimación se puede conseguir aún con un conjunto de datos reducido [25]. Los modelos estimados llegan a sistemas estables. Y adicionalmente se disponen de una gran variedad de métodos óptimos de estimación.

Realmente, al usar estimaciones asíncronas con intervalos de análisis fijos se consiguen estimaciones en promedio, en las que se puede colar información de fuente, ya sea de la periodicidad, de la características de la glotis o de la cavidad subglótica, que pudiera modificar la estimación de los formantes hacia los armónicos del pitch [12, 26]. Consideremos el modelado desde una perspectiva más detallada de cómo se genera la fuente en el proceso de producción de voz, teniendo en cuenta sus características instantáneas respecto de las fases glóticas, para ver como sortear estos inconvenientes.

## 3.4. Las características instantáneas de la voz sonora

Si nos fijamos más detalladamente en la explicación del apartado 1.2 sobre el proceso de producción de voz, nos podemos dar cuenta de que si la caracterización de un sonido depende de las propiedades de las cavidades que conforma el sistema que se forma en el proceso de generación y estas cavidades varían durante dicho proceso, entonces lo harán la características del sonido. Sería lógico pensar en analizar y modelar la señal en cada uno de esos instantes para representarla adecuadamente, sin que la estimación lleve a resultados imprecisos.

Nos estamos refiriendo a las características instantáneas de la voz durante la generación de sonidos sonoros, cuando la glotis se cierra y se abre periódicamente, creando una fuente periódica. En cada una de esas fases, el sistema de producción, desde los pulmones hasta los labios o la nariz, varía para un mismo sonido. Cuando la glotis está cerrada, el sistema se compone únicamente del tracto bucal y/o del tracto nasal, caracterizándose por las resonancias y antiresonancias de éstos. En ese caso, además, la excitación sería nula por estar cerrado el sistema en dicho punto. Por el contrario, en los momentos en que la glotis permanece abierta, el sistema se compone además de la glotis y de las cavidades subglóticas, que añaden resonancias, antiresonancias y amortiguamiento propios, [15, 2, 26]. La fuente en esta otra situación vendría dada por el aire expulsado y sus propiedades.

### 3.4.1. Modelado de la fase abierta y cerrada

Los trabajos sobre filtrado inverso, para la estimación de la fuente glotal desde la señal de voz, fueron los primeros en poner el foco, principalmente, en el modelado de

las características instantáneas del sonid, ver [27, 2, 28] o [29] para una revisión. Dicha tarea, usada en distintas áreas como codificación, síntesis o reconocimiento de locutores, necesita de una aproximación precisa de todo el sistema en cada instante para llegar a una representación de la excitación glótica que no dependa de aquel. Desde los primeros trabajos se vio que el modelado debía diferenciar los instantes en los que la fase estuviera abierta de aquellos en los que estuviese cerrada y modelarlos adecuadamente, atendiendo a sus distintas características.

El estudio centrado en la fase cerrada se ve desde aquellos trabajos como el más relevante, ya que es cuando el sistema se cierra en la glotis, quedando la señal de voz caracterizada únicamente por las propiedades de las cavidades supraglóticas, o sea el tracto vocal, e independizándola de la contribución de la excitación, que es nula. Los modelos empleados mayoritariamente en dichos trabajos para la fase cerrada son los modelos todo-polos 2.19, con los que se pretende caracterizar no solo los formantes del tracto vocal sino los ceros del tracto nasal o del efecto de la radiación, y lo llegan a hacer satisfactoriamente en muchas situaciones.

Hay sin embargo otras formas de tratar con esos ceros. Por ejemplo, el efecto de la radiación, un modelado con un cero, se aborda en [28] o [30] trasladando su acción a la excitación. Suponiendo un comportamiento estacionario de la radiación en todo momento, e igual en todos los sonidos, el intercambio consigue hacer desaparecer su influencia del modelo del sistema y endosárselo al modelo de la excitación, que a partir de entonces se considerará como la derivada de la excitación original.

Los ceros del tracto nasal, por otro lado, se toman en cuenta utilizando modelos polo-cero, como en [12] o [7]. En el primero trabajo se insinúa que en fase cerrada el modelo se convierte en todo-polo, por la ausencia de excitación, mientras que en el segundo se anima a obtener los ceros correspondientes.

El modelado de la fase abierta se complica algo más, al incorporarse las características de las cavidades subglóticas y su acoplamiento, que introducen polos y ceros adicionales además de producir alguna dispersión en los formantes del tracto vocal, modelados en la fase cerrada, o el ensanchamiento de sus anchos de banda [2, 7]. Una explicación desde el punto de vista de la teoría quantal se puede ver en [31, 32]. En cualquier caso el modelo empleado es de tipo polo-cero, para poder caracterizar todos los dichos efectos, con excitación distinta de cero.

Con este tipo de modelado por fase, no solo se intenta desacoplar la influencia que pudiera tener la excitación sonora y sus armónicos sobre las características del sistema, trabajando sobre muestras en intervalos menores a un periodo de pitch. También se pretende separar e identificar las contribuciones, tanto del sistema, representando al tracto vocal y/o nasal, como de la excitación, asociando a ella la glotis y las cavidades subglóticas, en cada uno de los sonidos.

Distintos aspectos se tienen que tomar en consideración a la hora de trabajar con el modelado de fase, como son la definición de los modelos de cada fase, en función de sus ventajas e inconvenientes o la identificación de las fases, a través de la detección de los instantes de cierre glótico y la determinación del tamaño de las tramas de análisis. Este último matiz, que es tan importante como difícil de establecer, tiene que servir para

representar adecuadamente cada fase y ser útil para obtener estimaciones robustas.

### 3.4.2. Detección de los instantes de cierre glótico

De entre los puntos más influyentes en el modelado de voz por fase glótica, se tiene la detección de los instantes en los que se produce el cierre glótico, que es el que marca la frontera entre la fase abierta y la fase cerrada. Hay dos formas de llevar a cabo dicha detección, a partir de la señal de voz o utilizando señales complementarias. Empleando la señal acústica se han desarrollado métodos que identifican el cierre glótico a través del error de predicción, [27, 28], también desde el retardo de grupo [33], otros desde transformadas tiempo-frecuencia como [34], algunos desde la pendiente de fase [35] y más efectivamente usando resonadores centrados en cero [36]. Una comparación se ofrece en [37] . Una detección más precisa se obtiene sin embargo a costa de señales complementarias, como el electroglotograma usado en [38] . Parece claro que la estimación será sensible a la exactitud de la determinación de estos instantes, por ello en esta tesis vamos a utilizar información directa de la glotis, asociadas a los registros con los que se trabaje.

Aún así, en esta tesis utilizamos un método sencillo de obtención de los ICG [1] que parte de una estimación de la frecuencia de vibración de las cuerdas vocales. Para ello, hacemos un uso conjunto de las autocorrelaciones y las diferencias de módulo en el dominio temporal., basado en las autocorrelaciones y las diferencias de módulo en el dominio temporal. Con este método pretendemos trabajar con un sistema de reconocimiento de un solo canal, el de voz, a la vez que aprovechamos los cálculos de autocorrelación que se tienen que hacer en los algoritmos de estimación de modelos lineales que se van a proponer. El método propuesto brota del detector de pitch mostrado en [10], que emplea la función de autocorrelación ponderada con la inversa de las diferencias de amplitud en magnitud normalizadas y que dice que son robustas en entornos ruidosos. Nosotros le hemos añadido unos umbrales en ambas medidas para detectar la sonoridad, cosa que no sostenía el detector citado. El detector de ICG se implementaría entonces a partir del pitch obtenido en las tramas sonoras, buscando la amplitud máxima de la trama (se supone que el cierre glótico está en las cercanías de los máximos absolutos) y con el valor de pitch rastreando los máximos cuasiperiódicos a un lado y a otro de aquel.

### 3.4.3. Estimación de los modelos de fases síncrona con el pitch

A diferencia del proceso de estimación asíncrona, en el que se emplea un tamaño de análisis fijo, la estimación por fase tiene que ser síncrona con el pitch. Los instantes de cierre glótico serán los que marcarán las referencias para situar la ventana de análisis. El intervalo de análisis viene definido por el tamaño de las fases, considerándose que pudiera tomar un tamaño variable y teniendo cuidado de no solaparse. En la determinación de las fases se debe considerar el proceso físico subyacente, refiriéndose al conocimiento de saber a que fase pertenece principalmente, según esté más abierta que cerrada o viceversa. De cualquier forma se establecerá que su comportamiento en ambas fases sea estacionario.

Lo que queda en evidencia en todos los trabajos es el pequeño tamaño que tienen la fases y la problemática que ello supone en el proceso de estimación, por el hecho de poseer pocos datos. Se puede decir que la fase cerrada abarca un 30 % del periodo , lo que equivale a intervalos menores a unos 3 ms, en comparación con los 20 ms de la estimación asíncrona. Este efecto se agrava en los casos con pitch alto, como las voces femeninas o de niños. Para superar este tipo de escollos se han propuesto métodos que reúnen una mayor cantidad de datos con las mismas características. La forma de llevarlo a cabo es tomando conjuntamente datos de varias fases [30, 7], correspondientes a periodos de pitch consecutivos, donde se considera que las características de la fonación no varían. Más aún pudiera ocurrir que la glotis no se llegara a cerrar completamente [39], en cuyo caso el uso de modelos AR quedaría en cuestión. En cualquier caso el número de datos necesarios en cada fase debería ser suficientemente grande para obtener resultados fiables, por lo menos el doble del orden del modelo de cada una de ellas, o viceversa.

La manera de enfrentarse a las estimaciones en cada fase, en los que se emplean modelos AR o ARMA para representar las resonancias y antiresonancias del sistema, también podría verse no solo actuando sobre cada fase independientemente, como en [7], sino de manera conjunta. En [4], por ejemplo, se consideran las dos fases en paralelo, con un modelo MA para la fase abierta, que incluyera los ceros subglóticos y un modelo ARMA para la fase cerrada, que aporte los polos supraglóticos y retardando los ceros supraglóticos para que no se solapen temporalmente con la anterior. La excitación en cada uno de esos casos deberá de ser debidamente establecida. Si se considera la estimación sobre el periodo completo de un tramo sonoro, parece adecuado señalar un pulso como excitación. Esto se deriva de utilizar un tren de pulsos como fuente de los sonidos sonoros. En el caso de que la estimación fuera sobre la fase cerrada, sería acertado tomar una excitación nula, si se supone que la fuente está desacoplada del tracto. Por el contrario, la estimación en fase abierta conlleva el desconocimiento de la excitación. Para sortear este problema [7] propone ignorarla, por desconocida, o reducir su efecto, trabajando con un modelo en el que la acción de la radiación se traslada a la excitación compensando el impacto de la glotis.

Es generalmente aceptado que trabajando sobre las fases se consigue una caracterización más precisa del sistema, a costa de una mayor carga computacional. Aumentar el número de datos con muestras de fases consecutivas, si se tratan las fases independientemente, o con los datos de las dos fases en un periodo de pitch, si se tratan conjuntamente, mejoran la fiabilidad de las estimaciones. Aún así creemos que se puede seguir mejorando, y no solo en fiabilidad, sino en robustez y precisión. En el caso de emplear datos de la misma fase consecutivamente, la robustez y la precisión podrían aumentar si la estimación se hiciera simultáneamente sobre todos sus parámetros, evitando que alguno de ellos no se detectara en alguna fase por cancelación polo-cero o por unas condiciones de ambiente no controladas. En el otro caso, el uso de datos de ambas fases sobre un solo periodo de pitch, maneja adecuadamente la estimación simultánea para determinar los diferentes parámetros, pero no aprovecha la capacidad de mejorar la fiabilidad aumentando el número de datos con periodos sucesivos. Vamos a hacer diferentes propuestas para avanzar por esas ideas aportadas. A la vez queremos introducir una formulación

compacta que nos permita definir los métodos de estimación de manera unificada y que podamos aplicar sobre ellos, de forma simplificada, posibles transformaciones que mejoren las propiedades de la caracterización en aplicaciones de reconocimiento.

### 3.4.4. Estimación de los modelos de fases con polos comunes

Para llevar a cabo la estimación de los modelos de fase, decidimos trabajar con modelos ARMA en ambas fases. Con ello permitiremos una representación más flexible de todos los sonidos, que pueda llegar a ser más precisa. En fase cerrada intentando caracterizar mejor, sobretodo, los sonidos nasales y en fase abierta para llegar a parametrizar las características subglotales.

Definamos *y(n,k)* como las muestras de una fase cualquiera (abierta o cerrada), correspondientes al periodo n-ésimo de una señal de voz *s(n,k)*, en la que el índice $k$ representa la posición de la muestra dentro del periodo. El modelo ARMA, que parametriza a cualquiera de las dos fases, vendrá dado por la siguiente expresión:

$$y(n,k) = -\sum_{i=1}^{p} a_{n,i} y(n,k-i) + \sum_{i=0}^{q} b_{n,i} u(n,k-i) \tag{3.1}$$

donde *u(n,k)* es la señal de excitación en el periodo n-ésimo, sobre el intervalo $\{k=0,\cdots,N_n\text{-}1\}$, y $\{a_{n,i},\ i=1,\cdots,p\}$ y $\{b_{n,i},\ i=1,\cdots,q\}$, son los coeficientes AR y MA, de órdenes *(p,q)* respectivamente. En el dominio Z, los polos y los ceros del filtro acústico estarán representados por las raíces de los polinomios

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \quad y \quad B(z) = \sum_{i=0}^{q} b_i z^{-i}$$

Conociendo, o suponiendo conocidas, las señales de entrada y salida, un método idóneo para la estimación de los parámetros, como ya se indicó en 2.5.3, es el de mínimos cuadrados, comúnmente empleado en los modelos de fases. La estimación de mínimos cuadrados se obtiene minimizando una función de coste que se define con el error cuadrático medio como sigue,

$$C_1(n) = \sum_{k=K_1}^{K_2} e^2(n,k) \tag{3.2}$$

siendo dicho error en este caso, el error de reconstrucción, dado por

$$e(n,k) = y(n,k) + \sum_{i=1}^{p} a_{n,i} y(n,k-i) + \sum_{i=0}^{q} b_{n,i} u(n,k-i) \tag{3.3}$$

Usando notación matricial y suponiendo que *u(n,k)* es conocida o que puede ser adecuadamente estimada como la señal de error obtenida de un modelado AR($\infty$), la 3.3 se podrá expresar, sobre el intervalo *{k=p,$\cdots$,$N_n$-1}*, de la siguiente manera:

$$
\begin{aligned}
\boldsymbol{e}_n &= \boldsymbol{y}_n - \left[\begin{array}{cc} \boldsymbol{Y}_n & \boldsymbol{U}_n \end{array}\right] \boldsymbol{h}_n = \boldsymbol{y}_n - \boldsymbol{H}_n \boldsymbol{h}_n \qquad (3.4)\\
\boldsymbol{e}_n &= \left[\begin{array}{ccc} e(n,p) & \cdots & e(n,N_n-1) \end{array}\right]^T \\
\boldsymbol{y}_n &= \left[\begin{array}{ccc} y(n,p) & \cdots & y(n,N_n-1) \end{array}\right]^T \\
\boldsymbol{h}_n &= \left[\begin{array}{ccccc} a_{n,1} & \cdots & a_{n,p} & b_{n,0} & \cdots & b_{n,q} \end{array}\right]^T = \left[\begin{array}{ccc} \boldsymbol{a}_n & \cdots & \boldsymbol{b}_n \end{array}\right]^T \\
\boldsymbol{Y}_n &= \left[\begin{array}{cccc} y(n,p-1) & y(n,p-2) & \cdots & y(n,0) \\ \vdots & \vdots & \ddots & \vdots \\ y(n,N_n-2) & y(n,N_n-3) & \cdots & y(n,N_n-p-1) \end{array}\right] \\
\boldsymbol{U}_n &= \left[\begin{array}{cccc} u(n,q) & u(n,q-1) & \cdots & u(n,0) \\ \vdots & \vdots & \ddots & \vdots \\ u(n,N_n-1) & u(n,N_n-2) & \cdots & u(n,N_n-q-1) \end{array}\right]
\end{aligned}
$$

La solución de mínimos cuadrados vendrá dada entonces por el método de las covarianzas, con una expresión del tipo de 2.38, debido a la similitud del error definido con 2.26 y 2.36. Por ello los coeficientes que minimizan $C_1(n)$ serán obtenidos a partir de:

$$\boldsymbol{h}_n = \left(\boldsymbol{H}_n^T \boldsymbol{H}_n\right)^{-1} \boldsymbol{H}_n^T \boldsymbol{y}_n \qquad (3.5)$$

Poniendo atención a la 3.4, nos podemos dar cuenta de que $\boldsymbol{e}_n$ está definido sobre el intervalo *{k=p,$\cdots$,$N_n$-1}*. El motivo es que los datos empleados en la estimación pertenezcan todos ellos a la fase en estudio. Aún siendo éste el método más común, nosotros preferimos ampliar el número de ecuaciones en el sistema para hacerlo más robusto, a costa de introducir el efecto del enventanamiento. Suponiendo que fuera de la fase la señal es nula, o sea enventanando, es fácil ver como el intervalo sobre el que se minimiza la función de coste puede ser redefinido sobre un rango mayor, *{k=0,$\cdots$,$N_n$+p-1}*, debido a que las matrices de señal, $\boldsymbol{Y}_n$ y de excitación, $\boldsymbol{U}_n$, puede ser extendidas por la repuesta al impulso infinita inherente al modelo ARMA. Por tanto, podemos pensar en sacar ventaja del uso del intervalo completo sobre el que el cual el error es distinto de cero. Con esta idea en mente, se propone volver a definir las matrices de la siguiente manera:

$$
\begin{aligned}
\boldsymbol{e}_n &= \left[\begin{array}{cccc} e(n,0) & \cdots & e(n,N_n+p-1) \end{array}\right]^T \\
\boldsymbol{y}_n &= \left[\begin{array}{ccccccc} y(n,0) & \cdots & y(n,N_n-1) & 0 & \cdots & 0 \end{array}\right]^T \\
\boldsymbol{h}_n &= \left[\begin{array}{ccccccc} a_{n,1} & \cdots & a_{n,p} & b_{n,0} & \cdots & b_{n,q} \end{array}\right]^T = \left[\begin{array}{ccc} \boldsymbol{a}_n & \cdots & \boldsymbol{b}_n \end{array}\right]^T \\
\boldsymbol{Y}_n &= \left[\begin{array}{cccc}
0 & 0 & \cdots & 0 \\
y(n,0) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
y(n,N_n-2) & y(n,N_n-3) & \cdots & y(n,N_n-p-1) \\
y(n,N_n-1) & y(n,N_n-2) & \cdots & y(n,N_n-p) \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & y(n,N_n-1)
\end{array}\right] \\
\boldsymbol{U}_n &= \left[\begin{array}{cccc}
u(n,0) & 0 & \cdots & 0 \\
u(n,1) & u(n,0) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
u(n,N_n-1) & u(n,N_n-2) & \cdots & u(n,N_n-q-1) \\
0 & u(n,N_n-1) & \cdots & u(n,N_n-q) \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0
\end{array}\right]
\end{aligned}
\tag{3.6}
$$

La 3.6 es válida para *(q+1)*$<p$. En el caso de que *(q+1)*$\geq p$ esta ecuación puede ser fácilmente arreglada y tomar una forma similar. Las ecuaciones en 3.6 nos llevan al método de las autocorrelaciones, que se pueden ver como una solución de mínimos cuadrados extendida. Una consideración a tener en cuenta es que este método garantiza la estabilidad de los modelos estimados, a diferencia de la solución por el método de las covarianzas.

Independientemente de la formulación de las matrices $\boldsymbol{Y}_n$ y $\boldsymbol{U}_n$, en la 3.4 y 3.6, la función de coste sigue siendo la misma, variando únicamente los límites de la sumatoria. Las expresiones de ambas soluciones serán por tanto análogas. Con ellas vamos a ver la robustez y la precisión de las estimaciones.

Adicionalmente y de manera similar se podría aplicar el método de las covarianzas modificado, persiguiendo una mayor resolución, lo cual conllevaría únicamente modificar los elementos de las matrices $\boldsymbol{H}_n^T\boldsymbol{H}_n$ y $\boldsymbol{H}_n^T\boldsymbol{y}_n$ en 3.5, por sus equivalente según 2.43.

Algo que queremos resaltar aquí es que la estimación, con cualesquiera de los métodos anteriores, actúa solamente sobre una fase, sea cual fuere, y que si ésta tiene pocos datos en relación al orden del modelo, las estimaciones no serían del todo fiables. Además, no se puede extender el intervalo de análisis sobre las muestras adyacentes, ya que podría sobrepasar el intervalo definido por la fase bajo estudio, incluyendo en ese caso datos de la otra fase.

Una forma de actuar para superar esos problemas de fiabilidad, se comentó que fue

aumentar el número de muestras utilizando datos de la misma fase pero de periodos consecutivos, como se han empleado en el llamado método multiciclo [30, 7]. El uso de periodos consecutivos se justifica por la lenta variación de las resonancias y antiresonancias en relación al pitch. En esos trabajos la solución de mínimos cuadrados se aplica únicamente a la parte AR del modelo y lo hace de una forma similar a 3.5, de la siguiente manera:

$$\boldsymbol{a}_n = \left( \sum_{k=n-K}^{n+K} \boldsymbol{Y}_k^T \boldsymbol{Y}_k \right)^{-1} \left( \sum_{k=n-K}^{n+K} \boldsymbol{Y}_k^T \boldsymbol{y}_k \right) \tag{3.7}$$

Efectivamente se ha aumentado el número de datos para hacer la estimación, haciendo un promediado sobre las covarianzas de diferentes periodos de pitch consecutivos con similares características. Otros trabajos también emplean promediado, como [39], pero sobre fases alineadas en el dominio temporal. En este caso el objetivo es, sin embargo, la robustez frente al ruido. Por otro lado, en ninguno de estos métodos hay una intención distinta de alcanzar unas estimaciones más fiables y robustas. A este promediado sobre K periodos lo nombraremos como AMCCK.

Nosotros proponemos, además de las metas anteriores, perseguir estimaciones más precisas, involucrando a varias fases de periodos consecutivos para llevar a cabo un modelado simultáneo en todas ellas y conseguir una reconstrucción temporal ajustada a la señal bajo estudio. La propuesta es utilizar la idea de modelado de polos acústicos comunes y ceros en salas acústicas [14]. En ella se describe una formulación específica a través del análisis conjunto de diferentes trayectorias, con modelos lineales 'fuente-receptor' que comparten los polos acústicos y se diferencian en los ceros. A partir de esta idea habría que identificar las características comunes de voz, que podrían ser las resonancias del tracto vocal y dejar las anti-resonancias para caracterizar las diferencias de cada conjunto de datos de los diferentes periodos. Con esta forma de proceder se intenta que el objetivo sea la reconstrucción precisa mediante el modelado simultáneo de periodos con características comunes, aumentando el número de datos empleado y consiguiendo paralelamente mejorar la fiabilidad y la robustez. El modelado que subyace en la propuesta se basa en modelos ARMA, totalmente idóneos con los modelos de fase en los que estamos interesados, para lo cual vamos a desarrollar una formulación compacta [5, 13].

#### 3.4.4.1. Estimación basada en polos comunes y ceros particulares

Supongamos que, independientemente de la fase considerada, y para $M$ periodos consecutivos de voz, la estructura de los ceros (resonancias y antiresonancias subglotales o comportamientos no ideales en fase cerrada) varía ligeramente de periodo en periodo y la estructura de los polos (resonancias y antiresoancias del tracto vocal, nasal y radiación) se mantiene razonablemente constante, ya que su lenta y ligera variación no es significativa. En este caso es posible redefinir la 3.4 y la 3.6 para hacer la estimación de los coeficientes simultáneamente a todos los periodos. Ahora la ecuación del error la

podemos definir de la forma siguiente:

$$
\begin{aligned}
\boldsymbol{e}_{n,M} &= \boldsymbol{y}_{n,M} - \boldsymbol{H}_{n;M}\,\boldsymbol{h}_{n,M} \\
\boldsymbol{H}_{n,M} &= \begin{bmatrix} \boldsymbol{Y}_{n+0} & \boldsymbol{U}_{n+0} & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{Y}_{n+1} & \mathbf{0} & \boldsymbol{U}_{n+1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Y}_{n+M-1} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{U}_{n+M-1} \end{bmatrix} \\
\boldsymbol{e}_{n,M} &= \begin{bmatrix} \boldsymbol{e}_{n+0} & \cdots & \boldsymbol{e}_{n+M-1} \end{bmatrix}^T \\
\boldsymbol{y}_{n,M} &= \begin{bmatrix} \boldsymbol{y}_{n+0} & \cdots & \boldsymbol{y}_{n+M-1} \end{bmatrix}^T \\
\boldsymbol{h}_{n,M} &= \begin{bmatrix} \boldsymbol{a}_n & \boldsymbol{b}_{n+0} & \cdots & \boldsymbol{b}_{n+M-1} \end{bmatrix}^T \\
\boldsymbol{a}_n &= \begin{bmatrix} a_{n,1} & \cdots & a_{n,p} \end{bmatrix}^T \\
\boldsymbol{b}_j &= \begin{bmatrix} b_{j,0} & \cdots & b_{j,q} \end{bmatrix}^T
\end{aligned}
\tag{3.8}
$$

donde $\boldsymbol{Y}_{n+j}$ y $\boldsymbol{U}_{n+j}$ , *{j=0,⋯,M-1}*, son las matrices de señal y de excitación correspondiente a la fase dada *y(n+j,k)* del periodo *s(n+j,k)* . Los vectores de señal $\boldsymbol{y}_j$ y de error $\boldsymbol{e}_j$ son similares a los de la 3.4. El vector de coeficientes $\boldsymbol{h}_{n,M}$ tiene ahora *p+Mx(q+1)* elementos. Los primeros, $\{\boldsymbol{a}_i\}$, corresponden a la estructura de polos comunes y los restantes, $\{\boldsymbol{b}_i\}$, corresponden a su estructura de ceros particulares. Esta ecuación nos lleva a lo que llamamos Modelo (Extendido) de Polos Comunes y Ceros particulares sobre *M(>1)* periodos, (E)CPPZM.

Al enfrentarnos a la estimación de $\boldsymbol{h}_{n,M}$, nuestra ecuación de error en la 3.8 es similar a la de [6], donde los autores estimaban los polos acústicos comunes y los ceros de varias funciones de transferencia relacionadas con la cabeza. A pesar de las diferencias en las aplicaciones, ambas situaciones comparten una formulación matricial similar, que lleva a una solución usando el método de mínimos cuadrados. La función de coste es ahora definida como la suma cuadrática del error de reconstrucción sobre el índice temporal $k$ de la señal de una fase dada, dentro de $M$ periodos consecutivos, empezando por el n-ésimo.

$$
C_M(n) = \sum_{j=0}^{M-1} \sum_{k=0}^{L} e^2(n+j,k) \tag{3.9}
$$

Los coeficientes $\boldsymbol{h}_{n,M}$ que minimizan $C_M(n)$ en la 3.8, usando el método de mínimos cuadráticos, pueden ser representados ahora en forma vectorial, como sigue:

$$
\boldsymbol{h}_{n,M} = \left(\boldsymbol{H}_{n,M}^T \boldsymbol{H}_{n,M}\right)^{-1} \boldsymbol{H}_{n,M}^T \boldsymbol{y}_{n,M} \tag{3.10}
$$

Los autores en [6] demostraron que esta solución es adecuada para la estimación

del orden óptimo, *(p,q)* en experimentos de identificación de características en salas acústicas. En cambio, nosotros adoptamos esta solución para estudiar la validez de la 3.5 y la 3.10, como estimaciones de las características del modelo de producción de voz. Para ello realizaremos algunos experimentos de seguimiento de las variaciones naturales del tracto vocal.

#### 3.4.4.2. Estimación basada en polos y ceros comunes

El modelado se puede simplificar en algunos casos asumiendo que ambas estructuras, la de polos y la de ceros, se mantienen constantes durante los $M$ periodos consecutivos. Esto puede ocurrir en algunos casos, como en la fase cerrada de sonidos vocálicos. También se puede aplicar cuando el objetivo sea mejorar la robustez frente al ruido. La 3.8 necesitará por tanto de algunas modificaciones para tomar en cuenta el hecho de que los nuevos coeficientes $\{b_{i,j}\}$ son comunes a todos los periodos bajo análisis.

$$\boldsymbol{H}_{n,M} = \begin{bmatrix} \boldsymbol{Y}_{n+0} & \boldsymbol{U}_{n+0} \\ \boldsymbol{Y}_{n+1} & \boldsymbol{U}_{n+1} \\ \vdots & \vdots \\ \boldsymbol{Y}_{n+M-1} & \boldsymbol{U}_{n+M-1} \end{bmatrix} \tag{3.11}$$

$$\boldsymbol{h}_{n,M} = \begin{bmatrix} \boldsymbol{a}_n & \boldsymbol{b}_n \end{bmatrix}^T$$

Esta ecuación nos lleva a lo que llamamos Modelo (Extendido) de Polo Comunes y Ceros Comunes sobre *M(>1)* periodos, (E)CPCZM.

La estimación del modelo vendrá dada por 3.10, con las matrices de datos y parámetros determinados por 3.11.

## 3.5. Las características instantáneas de voz sorda

Así como en la voz sonora encontramos que las cuerdas vocales se cierran y se abren alternativamente, teniendo que modelar ambas fase de la glotis, con los sonidos sordos, la glotis se mantiene abierta, por lo que se propone modelar dichas tramas con un modelo ARMA, similar al de la fase abierta de los tramos sonoros, pero con una excitación adecuada. Para la excitación podemos elegir tanto una señal aleatoria de igual potencia que la señal de voz, como el residuo obtenido con un filtrado inverso de orden suficientemente alto, con el objetivo de utilizar un señal similar a la real.

La estimación de los parámetros se llevaría a cabo de forma asíncrona, durante los intervalos de duración de dichos sonidos y con los métodos de mínimos cuadrados vistos en las secciones anteriores.

## 3.6. Experimentos y resultados

Vamos ahora a evaluar las características de los algoritmos de estimación propuestos, comparándolos con los métodos más usados en las mismas tareas, o con aquellos que sean

más competitivos en ellas. Plantearemos una serie de experimentos para ir apreciando las distintas aportaciones. El foco de atención lo pondremos sobre tres aspectos básicos en los métodos de estimación, como son la consistencia, la fiabilidad y la robustez. Los resultados nos servirán para tener un primer indicio de su utilidad en sistemas de reconocimiento automático del habla, donde se pretenden usar y determinar las condiciones óptimas de uso para poder sacarle el máximo provecho.

Para llevar a cabo estos esperimentos se ha usado la Base de Datos de voz Keele (anexo A), debido a que proporciona información de los instantes de cierre glótico (ICG), sonoridad y pich. Los registros empleados comprenden las locuciones de 5 hombres y 5 mujeres, de unos 40 segundos de duración de un conocido cuento inglés y con frecuencia de muestreo de 20 kHz.. La causa principal de su elección ha sido por que entonces podremos probar las características de los estimadores sin la influencia de errores provenientes de la determinación de los ICG.

A partir de los ICG las fases abierta y cerrada se tienen que definir. En la literatura se utilizan porcentajes para la fase cerrada que van desde el 30 % al 50 %, ya que como se indicó anteriormente no es un intervalo inambiguo. La mayoría de los autores suelen utilizar un intervalo para la fase cerrada del 40 % del tamaño del periodo, empezando en el ICG. Ahora la fase abierta podría ser representada por un porcentaje algo menor a la totalidad restante, comenzando a partir del final de la fase cerrada y terminando antes del ICG, para evitar los intervalos de mayor turbulencias que ocasionaran no linealidades en zonas anteriores a dicho instante. Ello también ayudaría a trabajar en presencia de errores producidos por la determinación de los mismos. Nosotros tomaremos el 48 % para la fase abierta.

### 3.6.1. Efecto del intervalo de análisis de la función de coste

El primer objetivo es ver la influencia que tiene utilizar el mayor intervalo posible sobre el que minimizar la función de coste 3.2 para obtener las estimaciones. Eso es lo mismo que considerar el número máximo de ecuaciones del modelo sobre los datos de la fase cerrada o abierta en un solo periodo. Nos estamos refiriendo a usar como método de estimación el método de las autocorrelaciones 3.53.4, que defendemos como más apropiado en el análisis localizado en las fases glóticas [5], frente al de las covarianzas 3.53.6, usado en la mayoría de los trabajos tanto de filtrado inverso como en reconocimiento con parámetros de fase glótica. El método de las autocorrelaciones nosotros lo llamamos extendido, y a él nos referiremos con el prefijo (E) en los acrónimos de los estimadores que emplearemos, para hacer hincapié en que el rango de errores sobre el que se calcula la función de coste se extiende al máximo posible (ver [17] para apreciar los distintos conjuntos de ecuaciones que se pueden tomar). La extensión conlleva el enventanado de los datos al intervalo de la fase glótica, ya que las muestras contiguas de la otra fase pertenecen a otro modelo, no pudiéndose utilizar en aquel, por lo que habría que suponerlos cero. Con esto se puede perder resolución, pero nuestra intención es ganar en consistencia y fiabilidad. Adicionalmente se asegura la estabilidad de las estimaciones debido a las características inherentes, comentadas en el capítulo anterior, que posee el método de las autocorrelaciones.

Como hemos comentado anteriormente los experimentos que vamos a realizar a continuación utilizan Keele, por lo que disponen de los registros de voz y la información de los ICG obtenidas por laringograma. Las estimaciones se realizan tanto sobre la fase cerrada como sobre la fase abierta de los voces sonoras.

En un primer experimento se va a considerar la adecuación de los métodos de estimación extendidos frente a la variación del orden de los modelos. Los modelos de las fases serán de tipo ARMA(p,q), con la condición de que q<=p. Teniendo en cuenta la frecuencia de muestreo de la base de datos empleada, la variación del orden será entre 1 y 20, tanto para p como para q. El tamaño para la fase cerrada del 40 % del periodo y del 48 % para la fase abierta, contigua a la anterior. Los métodos usados fueron: el defendido en este trabajo, el método de mínimos cuadrados extendido, (E)MCE, o método de las autocorrelaciones, frente al método usado mayoritariamente, el de mínimos cuadrados, MCE, o métodos de las covarianzas. Los resultados muestran el error de reconstrucción sobre todas las fases, considerando un solo periodo de análisis, frente a la variación del orden de los modelos y se representan en la figura 3.1.

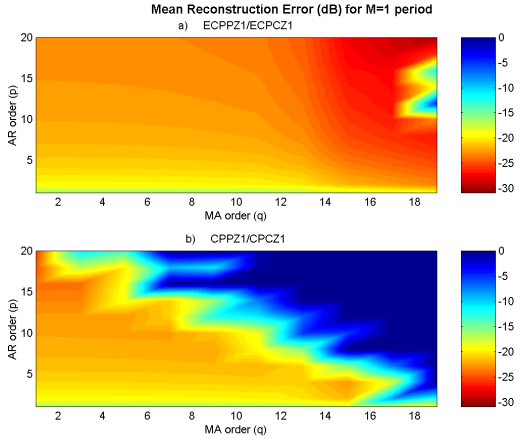



Figura 3.1: Error de reconstrucción (dB) para el método MCE (b) y el método extendido (E)MCE (a) para 1 periodo.

En dicha figura se aprecia como el uso del método extendido supera en funcionamiento al MCE, permitiendo una mayor flexibilidad a la hora de trabajar con distintos y mayores órdenes del modelo ARMA, respecto al otro método. Esto puede ser debido a la ventaja que supone aumentar el número de ecuaciones cuando las tramas de análisis son cortas en exceso, como en el análisis localizado. Estos resultados solo confirman las propiedades de consistencia que tienen estos métodos.

Un segundo experimento nos permitirá evaluar la mejora anterior en función del tamaño de la trama de análisis. El intervalo de análisis se fue considerando desde el mínimo al máximo encontrado en los registros. Los modelos empleados fueron ARMA, con el orden peor posible, o sea con el orden máximo en tramos tan cortos. Teniendo en cuenta la frecuencia de muestreo de la base de datos empleada, los modelos fueron

de tipo ARMA(20,19). Los resultados muestran el error de reconstrucción frente a la variación del intervalo de análisis y se representan en la figura 3.2.

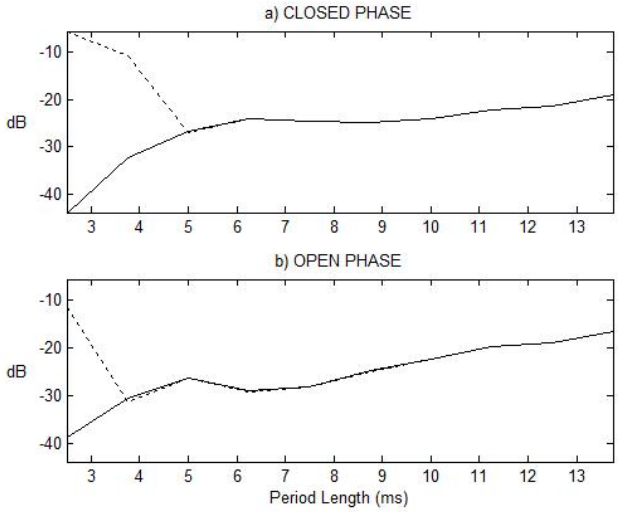



Figura 3.2: Error de reconstrucción (dB) para el método MCE(- -) y el método extendido (E)MCE (-) para 1 periodo.

Las gráficas evidencian los mejores resultados del método extendido, situando los mayores aciertos precisamente en las tramas de análisis cortas, en concreto menores a 5ms. Con esto queremos reforzar nuestra apuesta por el método de las autocorrelaciones en el análisis localizado, aunque pueda perder resolución como consecuencia del enventanado. Tanto mayor es la mejora cuanto menor es el tamaño de la trama de análisis. Sin embargo dicha mejora no se mantiene en tramas mayores que 5 ms, en tramas grandes, como puedan ser las usadas en el análisis asíncrono, con tramas que abarcan varios periodos.

### 3.6.2. Efecto del modelado de fase sobre periodos contiguos

El segundo objetivo es comprobar el efecto de aumentar el número de datos con que se trabaja, debido al corto tamaño de las tramas del análisis localizado. Se emplearán los métodos propuestos de polos comunes, que recopilarán muestras de la misma fase pero de periodos contiguos, en lo que se suponen que las características del sistema no varían. Lo que se pretende es mejorar la fiabilidad de las estimaciones, a costa de que el error de reconstrucción pueda aumentar ligeramente porque la suposición anterior pueda no ser cierta en algunas pocas ocasiones. También aquí introduciremos la formulación extendida en los diferentes métodos de estimación.

En este experimento, igual que en apartado anterior, consideraremos primero la adecuación de los métodos de estimación sobre periodos contiguos frente a la variación del orden de los modelos. El número de periodos contiguos lo hemos limitado a tres, que es

el equivalente típico al tamaño de trama de análisis asíncrono. Los modelos de las fases son ARMA(p,q), variando p y q entre 1 y 20 y siendo q<=p. Las fases cerrada y abierta son del 40 % y 48 % respectivamente. Los métodos empleados fueron: el de polos comunes y ceros particulares, 3.103.8, que llamamos (E)CPPZ3 y CPPZ3; y con polos y ceros comunes, 3.103.11, que llamamos (E)CPCZ3 y CPCZ3. El prefijo (E) hace referencia a que usa la extensión del rango en la función de coste y el número final indica el número de periodos contiguos. Nos podemos dar cuenta que los métodos (E)MCE y MCE, son equivalentes a (E)CPPZ1 y CPPZ1, respectivamente, o a (E)CPCZ1 y CPCZ1. Los resultados muestran el error de reconstrucción sobre todas las fases, considerando 3 periodos de análisis, frente a la variación del orden de los modelos y se representan en la figura 3.3.

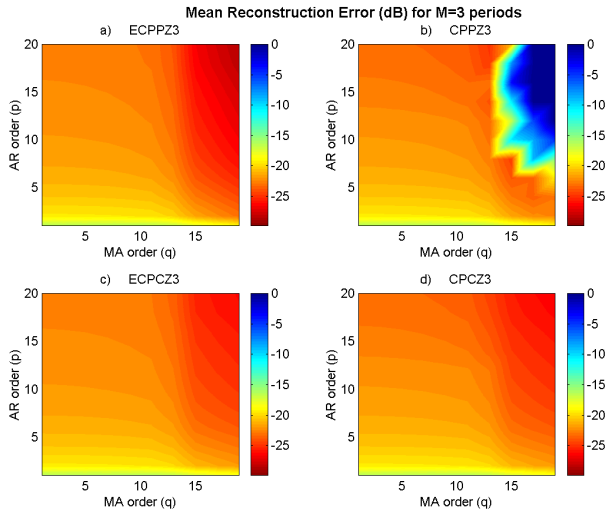



Figura 3.3: Error de reconstrucción (dB) para los métodos (a) (E)CPPZ1 y (b) CPPZ1 y para los métodos (c) (E)CPCZ3 y (d) CPCZ3.

En las gráficas se confirman los mejores resultados de los métodos extendidos. Además se aprecia un comportamiento algo más adecuado cuando se toman polos comunes y ceros particulares. Lo primero puede ser consecuencia de lo ya comentado del mayor número de ecuaciones, a la vez que testimonia que también con la inclusión de diferentes periodos los métodos siguen siendo consistentes. Lo segundo puede deducirse de que el uso de un mayor número de datos ofrecen una mayor flexibilidad en el modelado.

Siguiendo los mismos pasos que en el apartado anterior, en este otro experimento indagaremos en la mejora de los métodos que emplean periodos contiguos analizando el error de reconstrucción sobre el tamaño de las tramas. El intervalo de análisis considerado fue desde el mínimo al máximo encontrado en los registros. Los modelos empleados fueron ARMA, con el orden peor posible, ARMA(20,19). Los resultados del error de reconstrucción frente a la variación del intervalo de análisis y se representan en la figura 3.4

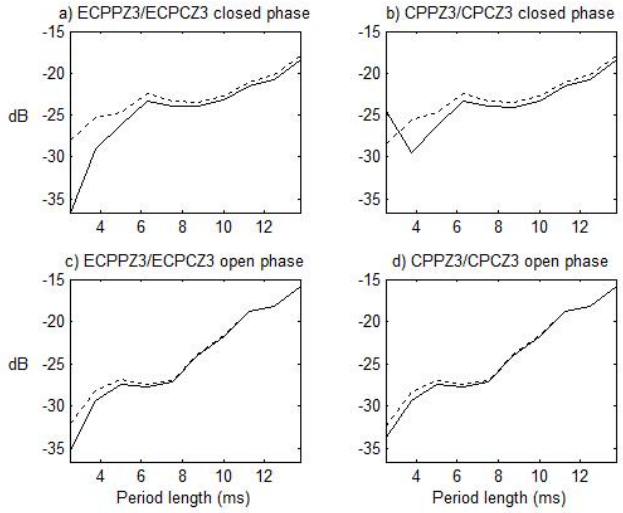



Figura 3.4: Error de reconstrucción (dB) para fase cerrada (a) – (b) y fase abierta (c) – (d) usando los clásicos CPPZ3 (-) - CPCZ3(- -) y los métodos extendidos (E)CPPZ3 (-) - (E)CPCZ3 (- -).

En dichos resultados se sigue manteniendo la ventaja de los métodos extendidos para tramas cortas, ahora menores a 8 ms, aunque con menor eficacia que con los métodos que solo emplean un periodo. La razón podría estar en que ya se está aumentando el número de errores que tiene en cuenta la función de coste al considerar varios periodos, por lo que no es tan determinante como en aquel. El efecto, sin embargo, se sigue notando, lo que avala la consistencia del método y por tanto su preferencia. El error de reconstrucción no es menor que cuando se toma un solo periodo, como podría parecer apelando a la consistencia, por el hecho de que el sistema de producción no se mantiene inalterado durante los varios periodos de análisis que se tomaran.

### 3.6.3. Efecto en el seguimiento de las características fonéticas en voz sonora

El tercer objetivo es evaluar la capacidad de los métodos propuestos en la caracterización efectiva de la voz, como soporte básico en los sistemas de reconocimiento automático del habla. Vamos a usarlos en una aplicación que consista en el seguimiento de las características fonéticas en voz sonora por medio del análisis localizado y nos fijaremos en la fiabilidad de los resultados. Los datos utilizados serán algunas de las transiciones fonéticas sacadas de la base de datos keele y el propósito no es tanto hacer un estudio exhaustivo sobre todas las transiciones como determinar el potencial de los métodos propuestos en la tarea de parametrización de la voz.

Se han diseñado varios experimentos, cada uno con una transición fonética diferente. Los modelos empleados son de tipo ARMA (12,11), habiendo elegido dicho orden por considerarlo suficiente para representar las voces y resaltar los resultados. El seguimiento

se hará sobre los polos de los modelos, que representan a los formantes en cada fase. Los formantes son obtenidos de las raíces del polinomio AR de cada modelo, a los que se le imponen algunas restricciones realistas, por su situación en frecuencia y sus anchos de banda. Las condiciones que los formantes deberán cumplir son que se correspondan con frecuencias mayores de 200Hz o menores de 9.8 kHz y que tenga un módulo mayor que 0.8. Las fases cerrada y abierta se toman del 40 % y del 48 % del tamaño del periodo respectivamente, a partir del ICG. Los métodos empleados serán los propuestos sobre 1 y sobre 3 periodos consecutivos, (E)CPCZ1 y (E)CPCZ3, frente al método multiciclo, 3.7, la única alternativa que también usa periodos consecutivos y empleada en otros trabajos de investigación y el espectrograma. De los métodos propuesto solo evaluamos el más simple, el de polos y ceros comunes, para evidenciar la potencialidad de las propuestas.

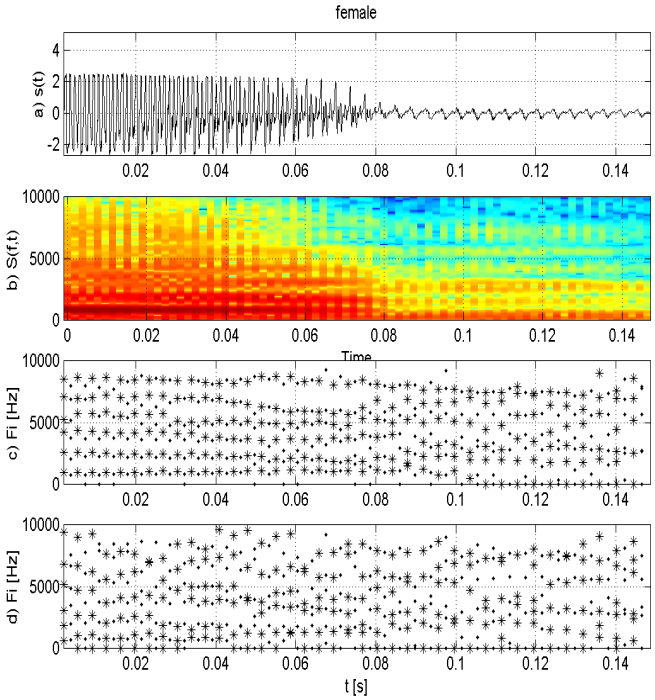



Figura 3.5: a) Registro fonético /an/ de mujer. b) Espectrograma. c) Estimación (E)CPCZ1. d) Estimación CPCZ1.

Los experimentos se realizaron sobre los segmentos siguientes: el segmento de voz /an/ pronunciada por una mujer, que contiene una transición vocal-consonante; el segmento de voz /ndmei/ pronunciado por un hombre, que contiene transiciones consonante-consonante, consonante-vocal y vocal-vocal; y el segmento de voz /aveller/ pronunciado por un hombre y compuesto por transiciones vocal-consonante y consonante-vocal. Los resultados se muestran en las figuras 3.5 y 3.6 para la primera transición, y en 3.7 y 3.8

para los dos segmentos siguientes, respectivamente. Los formantes de la fase cerrada los representaremos con símbolos '*', los de fase abierta con el símbolo '.' y los mostraremos en la misma gráfica para comprobar como siguen el comportamiento espectral del segmento de voz y observar las diferencias en las características de cada fase.

En la figura 3.5 se muestra la transición /an/ estimadas sobre un solo periodo, que equivale a los métodos clásicos, MCE (fig. 3.5.c) y MCC (fig. 3.5.d). Se observa una clara irregularidad en el seguimiento de los formantes, aunque el método extendido para un periodo, MCE, mantiene una mayor fiabilidad que el de covarianzas, MCC.

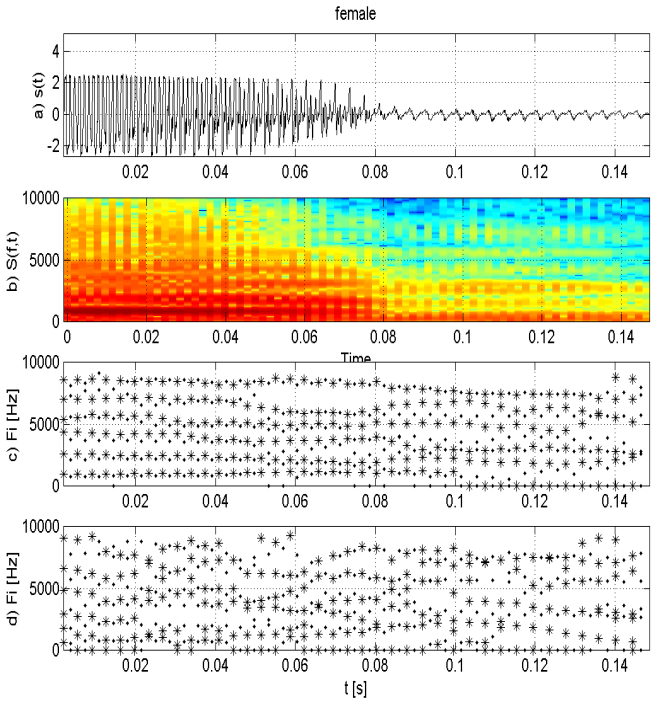



Figura 3.6: a) Registro fonético /an/ de mujer. b) Espectrograma. c) Estimación (E)CPCZ3. d) Estimación AMCC3.

En la figura 3.6 se muestra de nuevo la transición vocal-consonante /an/ pero ahora utilizando varios periodos. En la figura 3.6.c se contempla la variación de los formantes estimados con el método propuesto usando 3 periodos, CPCZ3. Se observa que los formantes siguen de forma precisa las frecuencias del espectrograma (fig. 3.6.b) y lo hacen de manera más regular en comparación con la que se observa en el método de promediado de covarianzas sobre 3 periodos, AMCC3, (fig. 3.6.d). En cualquier caso siempre presentan un mejor comportamiento que los métodos que usan 1 solo periodo. Los formantes de la fase cerrada y abierta siguen el espectrograma con las diferencias propias de cada fase.

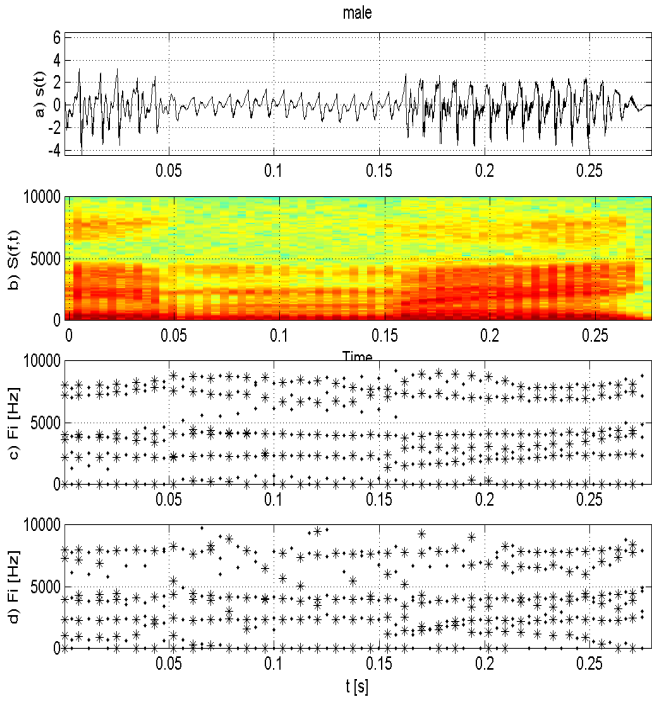



Figura 3.7: a) Registro sonoro /ndmei/ de hombre. b) Espectrograma. c) Estimación (E)CPCZ3. d) Estimación AMCC3.

En la figura 3.7 se muestra el segmento de voz /ndmei/, que contiene transiciones consonante-consonante, consonante-vocal y vocal-vocal. En la figura 3.7.c se contempla la variación de los formantes estimados con el método propuesto usando 3 periodos, CPCZ3. Se observa un seguimiento eficaz de los formantes comparado con el espectrograma (fig. 3.7.b) y una buena consistencia, tanto para las transiciones como para el diptongo. Se demuestra un mejor comportamiento que las estimaciones realizadas con el método clásico promediado en la matriz de covarianzas sobre 3 periodos, AMCC3 (fig. 3.7.d).

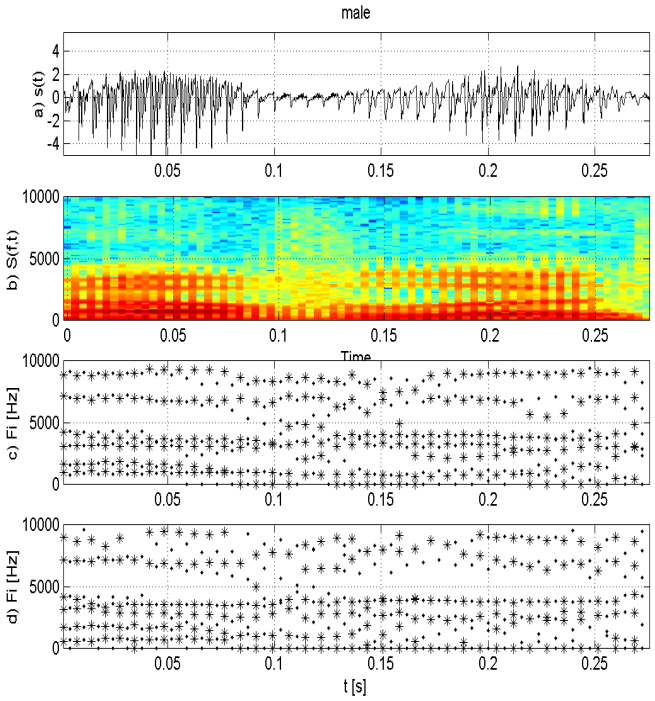



Figura 3.8: a) Registro fonético /aveller/ de hombre. b) Espectrograma. c) Estimación (E)CPCZ3. d) Estimación AMCC3.

En la figura 3.8.a se muestra el segmento de voz /aveller/, compuesto por varias transiciones, vocal-consonante y consonante-vocal. En la figura 3.8.c se contempla la variación de los formantes estimados con el método propuesto usando 3 periodos, CPCZ3. También en este tipo de transiciones se constata la mayor eficacia y consistencia del método propuesto en relación al espectrogramam (fig. 3.8.b) y en comparación a las estimaciones realizadas con el método clásico promediado en la matriz de covarianzas sobre 3 periodos, AMCC3 (fig. 3.8.d).

## 3.7. Conclusiones

Hemos abordado el problema de estimar parámetros comunes de voz en periodos consecutivos. La formulación que adoptamos es un buen marco para definir diferentes aproximaciones de estimación de coeficientes asociados a la estructura de polos y ceros. Es válido tanto para fase abierta como cerrada. Los experimentos han mostrado que la formulación extendida mejora la clásica en términos del error de reconstrucción para un orden del modelo dado. También, que la dependencia con el tamaño del periodo de análisis es mejor manejada por estos métodos extendidos y usando la información de varios periodos. Y finalmente, que integrar la información de varios periodos mejora el

comportamiento y proporciona mayor fiabilidad, tanto para hombres como para mujeres, así como en transición fonéticas de diferente naturaleza, en el seguimiento de las variaciones fonéticas naturales del habla.

## Bibliografía

[1] I. Pérez-Castellano, P. J. Quintana-Morales, and J. L. Navarro-Mesa, "Clasificacion de voz patologica mediante modelado ARMA comun a varios periodos," in *URSI*, 2004.

[2] J. Holmes, "Formant excitation before and after glottal closure," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '76.*, vol. 1, pp. 39 – 42, Apr. 1976.

[3] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Prentice-Hall, 1978.

[4] H. Kamata, H. Oka, and Y. Ishida, "Estimation of vocal tract transfer function considering the glottis open and close characteristics," in *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, vol. 1, pp. 137 –140 vol.1, May 1993.

[5] P. J. Quintana-Morales and J. L. Navarro-Mesa, "An approach to common acoustical pole and zero modeling of consecutive periods of voiced speech," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, ISCA, 2003.

[6] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki, "Common-acoustical-pole and zero modeling of head-related transfer functions," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 188 –196, Mar. 1999.

[7] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 313 –327, July 1998.

[8] G. Fant, *Acoustic Theory of Speech Production.* Mouton De Gruyter, revised ed., Jan. 1971.

[9] D. O'Shaughnessy, "Improving analysis techniques for automatic speech recognition," in *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, vol. 3, pp. III–65 – III–68 vol.3, Aug. 2002.

[10] H. Kobayashi and T. Shimamura, "A weighted autocorrelation method for pitch extraction of noisy speech," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1307 –1310 vol.3, 2000.

[11] J. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, pp. 1215 –1247, Sept. 1993.

[12] L. Wood and D. Pearce, "Excitation synchronous formant analysis," *Communications, Speech and Vision, IEE Proceedings I*, vol. 136, pp. 110 – 118, Apr. 1989.

[13] J. L. Navarro-Mesa and P. J. Quintana-Morales, "Seguimiento de las variaciones naturales de la voz sonora en fase abierta y cerrada," in *URSI*, 2003.

[14] Y. Haneda, S. Makino, and Y. Kaneda, "Common acoustical pole and zero modeling of room transfer functions," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 320 –328, Apr. 1994.

[15] J. L. Flanagan, *Speech analysis; synthesis and perception.* Springer-Verlag, 1972.

[16] S. M. Kay, *Modern spectral estimation: theory and application.* Prentice Hall, 1988.

[17] J. Marple, S.L., "A tutorial overview of modern spectral estimation," in *, 1989 International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89*, pp. 2152 –2157 vol.4, May 1989.

[18] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencies," *Electr. and Commun. in Japan*, vol. 52-A, pp. 36–43, 1970.

[19] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[20] J. Markel, "Digital inverse filtering-a new tool for formant trajectory estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 2, pp. 129–137, 1972.

[21] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 3, pp. 229–234, 1977.

[22] B. S. Atal and M. R. Schroeder, "Linear prediction analysis of speech based on a pole-zero representation," *The Journal of the Acoustical Society of America*, vol. 64, pp. 1310–1318, Nov. 1978. PMID: 744832.

[23] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, 1968.

[24] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*, vol. 14, pp. 99–145, June 2011.

[25] S. Kay and J. Marple, S.L., "Spectrum analysis - a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380–1419, 1981.

[26] D. G. Childers and C.-F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 7, pp. 663–671, 1994.

[27] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *The Journal of the Acoustical Society of America*, vol. 56, no. 5, pp. 1625–1629, 1974.

[28] D. Wong, J. Markel, and J. Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[29] 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, pp. 623–650, Oct. 2011.

[30] D. Brookes and D. Chan, "Speaker characteristics from a glottal airflow model using robust inverse filtering," *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, vol. 16, pp. 501–501, 1994.

[31] X. Chi and M. Sonderegger, "Subglottal coupling and its influence on vowel formants," *The Journal of the Acoustical Society of America*, vol. 122, p. 1735, Sept. 2007. PMID: 17927433.

[32] S. M. Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, no. 1, pp. 20–32, 2010.

[33] B. Yegnanarayana and R. L. H. M. Smits, "A robust method for determining instants of major excitations in voiced speech," in *, 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95*, vol. 1, pp. 776–779 vol.1, 1995.

[34] J. Navarro-Mesa, E. Lleida-Solano, and A. Moreno-Bilbao, "A new method for epoch detection based on the cohen's class of time frequency representations," *IEEE Signal Processing Letters*, vol. 8, no. 8, pp. 225–227, 2001.

[35] A. Kounoudes, P. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. I–349–I–352, 2002.

[36] K. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[37] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, pp. 651–697, Nov. 2011.

[38] A. Krishnamurthy, "Two channel (speech and egg) analysis for formant and glottal inverse filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, vol. 9, pp. 60–63, 1984.

[39] B. Cranen and L. Boves, "On subglottal formant analysis," *The Journal of the Acoustical Society of America*, vol. 81, pp. 734–746, Mar. 1987. PMID: 3584682.

Capítulo 4

# TRANSFORMACIÓN PERCEPTUAL EN EL MODELADO LINEAL LOCALIZADO

## 4.1. Introducción

El análisis de la señal de voz necesita que el modelado que represente el proceso físico en el que está inmerso, que no solo incluya diferentes características del proceso de producción de voz, sino también el proceso de percepción. Este último tiene su importancia en la medida en que la voz se produce para que pueda ser escuchada, o al menos esa es su intención.

En el capítulo anterior hemos analizado la señal de voz desde el punto de vista del proceso físico de producción. En este capítulo ampliamos nuestro enfoque introduciendo también elementos psicoacústicos. Para ello modificaremos el modelo mediante la introducción de células paso-todo que mejoren la capacidad perceptual, en el sentido de aproximarlo frecuencialmente al comportamiento no lineal que presenta el oido. Traba-

jaremos a nivel de fase glótica. Estableceremos el modelo transformado y obtendremos los métodos de estimación correspondientes.

## 4.2. El proceso de la percepción

Un aspecto destacado a considerar en los parámetros de modelado del habla, aparte de que representen adecuadamente las características acústicas de la voz, es su relación con las propiedades perceptuales del odio. La percepción forma parte indisoluble del sistema natural del que proviene, el de la comunicación, y por aplicación, al caso que nos ocupa del reconocimiento automático del habla, y con su consideración se busca una mayor eficacia y robustez en la parametrización.

El proceso de percepción es un mecanismo complejo que está ligado a las características físicas del oído. De forma simplificada se puede decir que acondiciona la onda sonora en el oído externo, la adapta con el oído medio y la transforma en el oído interno. El oído externo dirige la onda sonora al tímpano, en el oído medio, protegiendo también al sistema auditivo de sonidos intensos y prolongados y teniendo una respuesta con una resonancia entorno a los 3 kHz. El oído medio se encarga de adaptar acústicamente el oído externo con el oído interno, transmitiendo de forma mecánica y eficaz las vibraciones sonoras detectadas en el tímpano y conduciéndolas a la ventana oval en el oído interno. En el oído interno se encuentra la cóclea, rellena de líquido y que contiene la membrana basilar y las células pilosas, los principales elementos de percepción. La membrana basilar vibra con sonidos periódicos, como consecuencia de las ondas viajeras producidas en el líquido coclear por las perturbaciones percibidas a través de la ventana oval, generando resonancias en puntos de la misma que dependen de la frecuencia de dichos sonidos. El proceso de percepción es algo más complicado cuando se trata de sonidos complejos pudiendose destacar, en cualquier caso, la relación no lineal entre la resolución y la frecuencia, que es no lineal. Se podría decir que es aproximadamente lineal hasta 1kHz y logarítmica a partir de dicha frecuencia.

## 4.3. Modelado de los aspectos perceptuales

Uno de los principios psicoacústicos más utilizado es el comportamiento no lineal del sistema auditivo señalado, que ha sido modelado con bandas críticas, dispuestas de manera no uniforme y más estrechas en las frecuencias bajas que en las altas [10]. Las técnicas que lo introducen realizan una modificación de la información espectral, buscando una mayor resolución en baja frecuencia, para aproximarla a las escalas Bark o Mel . Varias son las aplicaciones en las que se ha incorporado esta propiedad perceptual con un reflejo positivo en sus resultados, por lo que proponemos estudiar su inclusión en la formulación investigada [11, 7, 8]. La modificación espectral se puede llevar a cabo desde dos puntos de vista, según se pretenda transformar la señal a analizar o el modelo de análisis. En el primer caso la señal se puede transformar por medio de métodos basados en banco de filtros o en la transformada de Fourier, dejándola preparada para su análisis posterior [2, 3]. En el segundo caso, la modificación se realiza sobre el modelo

de análisis, que incorpora implícitamente la nueva distribución espectral y se realiza por medio de la transformación bilineal [4, 11, 9]. Es este segundo enfoque el que nos interesa, y que proponemos para aplicarlo a nuestro modelado ARMA de fases síncrona con el pitch [12].

## 4.4. Modelado perceptual de fases sobre periodos consecutivos

Vamos a dar un paso más en nuestro enfoque del modelado lineal localizado desarrollado en el capítulo anterior, introduciendo el punto de vista de la percepción auditiva. Esto lo vamos a llevar acabo mediante la aplicación de una función de modificación espectral que controla la resolución de frecuencia de un modo psicoacústico.

### 4.4.1. Modelado perceptual mediante secciones paso-todo

El análisis se basa en la idea de utilizar secciones paso-todo, dispersivas y de primer orden, en lugar de las unidades de retardo que aparecen en las ecuaciones en diferencia de un modelo ARMA, definidas en 3.1. Esta noción ha sido introducida en codificación predictiva lineal modificada espectralmente (por ejemplo, [7, 6]), en el contexto de un análisis típico trama a trama, asíncrono. En su lugar, nosotros proponemos desarrollar esta idea en un modo síncrono con el pitch para la obtención de un modelo modificado espectralmente de polos comunes y ceros particulares (WCPPZ) sobre varios períodos adyacentes.

Una sección paso-todo, dispersiva y de primer orden, puede ser vista como un elemento de retardo dependiente de la frecuencia y se define de la siguiente forma, [7].

$$\tilde{z}^{-1} = D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \tag{4.1}$$

donde $-1 < \lambda < 1$. La respuesta en fase de la célula paso-todo D(z) es una función no lineal que depende del ajuste del factor de modificación espectral, o warping, $\lambda$. Realmente, el mapeo desde una escala uniforme de frecuencia hacia una escala de frecuencias modificadas espectralmente es gobernada por la respuesta en fase de D(z), la cual es dada por:

$$\tilde{\omega} = \arg(D(e^{-j\omega})) = \omega + 2\arctan\left(\frac{\lambda \sin(\omega)}{1 - \lambda\cos(\omega)}\right) \tag{4.2}$$

La respuesta en fase determina la transformación que tiene lugar en la sección paso-todo. Si el parámetro $\lambda$ es 0, la función de transferencia de la sección se reduce a una simple unidad de retardo, con fase lineal y retardo de grupo constante. Para valores positivos de $\lambda$ se incrementa la resolución en las frecuencias bajas. Por otro el contrario,

valores negativos producen una mayor resolución en altas frecuencias. La elección del valor adecuado de $\lambda$ dependerá por tanto de la aplicación que se esté considerando. Con un valor apropiado, la escala de frecuencia modificadas espectralmente podría mostrar una buena similitud con la escala Bark, pudiendo optimizar la resolución en frecuencia desde el punto de vista de la percepción acústica. Por ejemplo, con una velocidad de muestreo de 20 KHz, la escala Bark es obtenida usando $\lambda \approx 0,65$ (e.g., [5]).

Un parámetro importante es el punto intermedio de la modificación, o turning point, $f_{tp}$, que define la frecuencia para la cual la modificación espectral no afecta a la resolución en frecuencia, esto es, donde el retardo de grupo es 1. La expresión que define dicho parámetro, $f_{tp}$, en función del parámetro de warping y de la frecuencia de muestreo viene dada por:

$$f_{tp} = \pm \frac{f_s}{2\pi} \arccos(\lambda) \tag{4.3}$$

La resolución del sistema modificado espectralmente con $\lambda \geq 0$ será por tanto mayor por debajo (y menor por encima) del turning point, que en un sistema convencional, con resolución uniforme de frecuencia. A altas velocidades de muestreo (por ejemplo $f_s > 8\,KHz$) la diferencia entre la escala modificada espectralmente y la escala lineal es significativa.

Un punto clave, que a menudo no se le da la importancia debida en la modificación espectral de la escala de frecuencia, es la respuesta al impulso de las secciones paso-todo, cuya expresión se puede deducir fácilemte de 4.1 y es la siguiente:

$$d(n) = -\lambda\delta(n) + (1-\lambda^2)\sum_{i=1}^{\infty} \lambda^{i-1}\delta(n-1) \tag{4.4}$$

Se puede apreciar que ésta es una respuesta al impulso infinita y que realiza una integración temporal infinita. De esta forma, *d(n)* manifiesta un efecto tanto de retardo como de dispersión. Combinando dos o más secciones de retardo elementales en serie, podemos incrementar, si se quiere dramáticamente, la extensión de la respuesta al impulso resultante, debido a que cada muestra generada por el primer filtro , *d(n)*, generará un conjunto de muestras en el segundo, *d(n)*d(n)*, y así sucesivamente. Cualquier número de secciones paso-todo pueden ser conectadas en serie, que la respuesta combinada será igualmente paso-todo. El incremento de la extensión de la respuesta al impulso puede ser visto también desde el incremento de la fase, $\tilde{\omega}$ y el correspondiente retardo de grupo.

### 4.4.2. Modelado perceptual de la fase abierta y cerrada

Volviendo al modelo paramétrico que caracteriza cada una de las fases de nuestro sistema, en el dominio Z, la expresión de 3.1 se transforma en:

$$Y_n(z) = -\sum_{i=1}^{p} a_i^n z^{-i} Y_n(z) + \sum_{i=0}^{q} b_i^n z^{-i} U(z) \tag{4.5}$$

Cuando las unidades de retardo $z^{-1}$ son sustituidas por secciones paso-todo de primer orden del tipo de la 4.1, obtenemos

$$Y_n(z) = -\sum_{i=1}^{p} a_i^n D^i(z) Y_n(z) + \sum_{i=0}^{q} b_i^n D^i(z) U(z) \tag{4.6}$$

donde $D^i(z)$ es un operador de retardo generalizado en el dominio Z. En el dominio temporal, este operador aplicado sobre una señal dada *x(n)* puede ser definido como:

$$d_i[x(n)] = \underbrace{d(n) * d(n) * \cdots * d(n)}_{i-veces} * x(n) \tag{4.7}$$

Transformando 4.6 al dominio del tiempo y aplicando 4.7, obtenemos la expresión extendida de 3.3, caracterizada por la siguiente expresión

$$y(n,k) = -\sum_{i=1}^{p} a_i^n d_i(y(n,k)) + \sum_{i=0}^{q} b_i^n d_i(u(n,k)) \tag{4.8}$$

### 4.4.3. Estimación de los modelos de fases con polos comunes modificados perceptualmente

Para llevar a cabo la estimación de los parámetros del modelo modificado perceptualmente, definimos el error de reconstrucción de la señal de forma similar a como lo habíamos hecho en la capítulo anterior, pero incorporando los operadores de modificación paso-todo sobre las muestras de señal, de la siguiente manera:

$$e(n,k) = y(n,k) + \sum_{i=1}^{p} a_i^n d_i(y(n,k)) - \sum_{i=0}^{q} b_i^n d_i(u(n,k)) \tag{4.9}$$

Usando notación matricial y suponiendo que *u(n,k)* es conocida o puede ser adecuadamente estimada, la 4.9 se podrá poner como sigue:

$$\boldsymbol{e}_n = \boldsymbol{y}_n - \left[\begin{array}{cc} \boldsymbol{Y}_n & \boldsymbol{U}_n \end{array}\right] \boldsymbol{h}_n = \boldsymbol{y}_n - \boldsymbol{H}_n \boldsymbol{h}_n \tag{4.10}$$

$$\boldsymbol{Y}_n = \left[\begin{array}{cccc} d_1(y(n,0)) & d_2(y(n,0)) & \cdots & d_p(y(n,0)) \\ d_1(y(n,1)) & d_2(y(n,1)) & \cdots & d_p(y(n,1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_1(y(n,N_n-1)) & d_2(y(n,N_n-1)) & \cdots & d_p(y(n,N_n-1)) \\ \vdots & \vdots & \cdots & \vdots \\ d_1(y(n,N_n+N'-1)) & d_2(y(n,N_n+N'-1)) & \cdots & d_p(y(n,N_n+N'-1)) \end{array}\right]$$

$$\boldsymbol{e}_n = \left[\begin{array}{ccc} e(n,0) & \cdots & e(n,N_n+N'-1) \end{array}\right]^T$$

$$\boldsymbol{y}_n = \left[\begin{array}{cccccc} y(n,0) & \cdots & y(n,N_n-1) & 0 & \cdots & 0 \end{array}\right]^T$$

$$\boldsymbol{h}_n = \left[\begin{array}{cccccc} a_1^n & \cdots & a_p^n & b_0^n & \cdots & b_q^n \end{array}\right]^T$$

$$\boldsymbol{U}_n = \left[\begin{array}{cccc} d_0(u(n,0)) & d_1(u(n,0)) & \cdots & d_q(u(n,0)) \\ d_0(u(n,1)) & d_1(u(n,1)) & \cdots & d_q(u(n,1)) \\ \vdots & \vdots & \ddots & \vdots \\ d_0(u(n,N_n-1)) & d_1(u(n,N_n-1)) & \cdots & d_q(u(n,N_n-1)) \\ \vdots & \vdots & \cdots & \vdots \\ d_0(u(n,N_n+N'-1)) & d_1(u(n,N_n+N'-1)) & \cdots & d_q(u(n,N_n+N'-1)) \end{array}\right]$$

Donde $\boldsymbol{h}_n$, $\boldsymbol{y}_n$ y $\boldsymbol{e}_n$ son los vectores de coeficientes, de señal y de error, respectivamente. E $\boldsymbol{Y}_n$ y $\boldsymbol{U}_n$ son las matrices de señal y excitación, respectivamente.

Hemos introducido un parámetro de extensión, *N'*, que controla el intervalo de muestras donde el error va a ser definido. Este parámetro fue previamente introducido en [1], asociado a la respuesta al impulso infinita de la parte AR en 3.1. Esto nos lleva a lo que habíamos llamado Modelo Extendido de Polo Comunes y Ceros Particulares. Ahora, este parámetro adquiere un nuevo significado, porque también está asociado a 4.4.

Las estimaciones de los parámetros modificados perceptualmente se lleva a cabo a través de los métodos de estimación desarrollados en el capítulo anterior, sin más que sustituir las matrices modificadas perceptualmente por sus homólogas. Por ello podremos obtener estimaciones tanto de modelos de polos comunes y ceros particulares como de modelos de polos y ceros comunes.

En la sección de experimentos y resultados estudiaremos hasta qué punto nos ayuda la nueva formulación presentada en 4.10. Asumimos que debido al efecto de integración temporal de 4.4 además de la modificación de la escala espectral producirá pequeños errores simultáneamente en tiempo y frecuencia.

## 4.5. Experimentos y resultados

Evaluaremos ahora las características que aportan los algoritmos de estimación de los modelos transformados perceptualmente, comparándolos con los modelos no transformados . Plantearemos una serie de experimentos para ir apreciando las distintas aportaciones, poniendo atención en los mismos aspectos considerados anteriormente, la consistencia, la fiabilidad y la robustez. Los resultados nos servirán para ver su potencialidad en sistemas de reconocimiento automático del habla, donde se pretenden usar y determinar las condiciones óptimas de uso para poder sacarle el máximo provecho.

Estos experimentos se han desarrollado bajo las mismas condiciones que los experimentos del capítulo anterior, para poder comparar con los resultados anteriores.

Se ha utilizado la Base de Datos Keele, con locuciones de 5 hombres y 5 mujeres, de unos 40 segundos de duración, se han empleado un intervalo para la fase cerrada del 40 % del tamaño del periodo, empezando en el ICG y para la fase abierta el 48 % desde el final de la anterior. Al introducir una nueva variable, el factor de warping, hemos decidido fijar el modelo de trabajo a uno fijo ARMA(p,q) que pueda representar adecuadamente las voces de la Base de Datos muestreadas a 20 kHz.

### 4.5.1. Efecto del intervalo de análisis de la función de coste

El primer objetivo es ver la influencia que tiene utilizar el parámetro de extensión, *N'*, que controla el intervalo de muestras donde el error va a ser definido y que ahora puede ser definido de manera arbitraria, al contrario del modelo no transformado que estaba limitado a p. Estamos considerando el número máximo de de ecuaciones del modelo sobre los datos de la fase cerrada o abierta.

El conjunto de experimentos consistió en variar el parámetro de extensión, *N'*, entre 0 y 100, para un modeloa ARMA(16,15) con valor de $\lambda = 0{,}75$, que equivale a un $f_{tp} = 2{,}3$ kHz. Los resultados muestran la relación señal a error de reconstrucción, SRR, (en dB) evaluados sobre todos los periodos y se representan en la figura 4.1.

Los métodos empleados fueron: el de polos comunes y ceros particulares, definidos con las equaciones 3.10 y 3.8, que llamamos (E)CPPZ3 y su modelo modificado en frecuencia, utilizando las matrices modificadas en frecuencia, establecidas en 4.10, (W)CPPZ3. El prefijo (E) hace referencia a que usa la extensión del rango en la función de coste, el prefijo (W) indica que el modelo corresponde al modificado en frecuencia y el número final en ambos casos indica el número de periodos contiguos. En la figura 4.1 se observa claramente como el modelo transformado en frecuencia tiene un comportamiento superior al no transformado, o uniforme en frecuencia. Además se deduce que a partir de una extensión de 20 muestras, la estimación no mejora en ninguna de las fases. Este resultado es similar con cualquier orden del modelo ARMA(q,p) por lo que deducimos que no depende de él. Esto contrasta con los resultados anteriores obtenidos en el caso de los modelos uniformes en frecuencia, en los que si que dependía del orden p de la parte AR. Puede decirse que la estimación en el caso tranformado es más consistente.

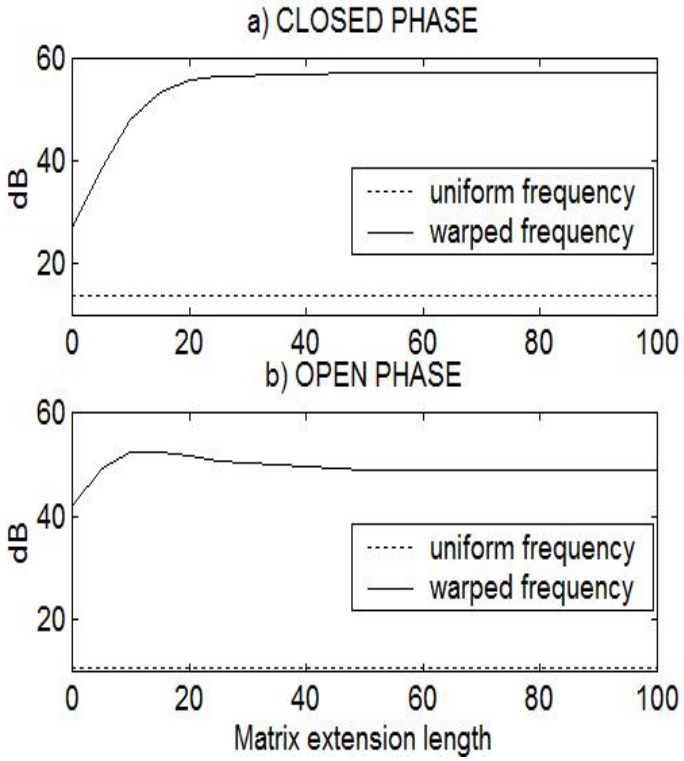



Figura 4.1: Evolución del SRR con el parámetro de extensión, *N'*.

### 4.5.2. Efecto de la modificación de la resolución en frecuencia

El segundo objetivo es comprobar como afecta el factor de warping, o sea, como varía la resolución en frecuencia, a las estimaciones y si se puede deducir algún valor óptimo de trabajo. En cualquier caso ello dependerá de la frecuencia de muestreo utilizada.

En estos experimentos se utilizaron modelos iguales en cada fase, un ARMA(16,15) con un parámetro de extensión de 20 muestras y el factor de warping $\lambda$ se varió entre 0 y 1 en ambos casos. Los resultados muestran la relación señal a error de reconstrucción, SRR, (en dB) evaluados sobre todos los periodos y se representan en la figura 4.2. El método de estimación empleado fue el (W)CPPZ3.

Las gráficas evidencian que no hay efecto hasta $\lambda = 0{,}2$, que la mejora es limitada a partir de $\lambda = 0{,}4$ y que hay un valor óptimo. Este factor óptimo no es igual para la fase cerrada que para la fase abierta y esto puede ser debido a que los intervalos de análisis en ambas fases no coincide. De cualquier forma, el punto óptimo está en torno a $\lambda = 0{,}75$, que equivale a $f_{tp} = 2{,}3$ kHz y que coincide con el sugerido en [5]. En cualquier caso lo que nos está indicando dicho valor es donde está llevando su mayor esfuerzo el método de estimación, aumentando la resolución en frecuencia por debajo de aquel valor y reduciendo allí el error de reconstrucción.

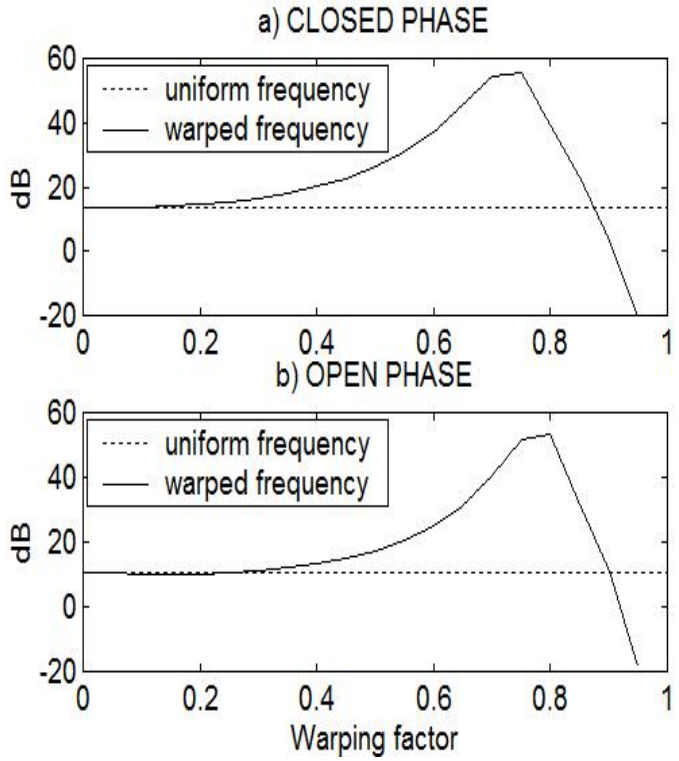



Figura 4.2: Evolución del SRR con el factor de warping.

Para ver el comportamiento de la modificación de la resolución en frecuencia en el dominio frecuencial, vamos a llevar a cabo un experimento que represente el espectro del error en la fase cerrada y en la fase abierta para un intervalo de análisis dado.

Empleando el modelo definido anterior, con el factor de warping óptimo, calculamos los espectros del error producido en la fase cerrada y en la fase abierta a partir de los datos obtenidos con los métodos de estimación (E)CPPZ3 y (W)CPPZ3 y los representamos en la figura 4.3.

Aquí nos podemos fijar en dos aspectos. Primero, la estimación de los métodos transformados supera el comportamiento correspondiente al de los métodos no transformados, como se aprecia en el menor nivel de potencia del error de reconstrucción con los métodos transformados. Segundo, nos sorprende que el espectro del error en los métodos transformados sea bastante similar en frecuencia, ya que se esperaba en las frecuencias bajas, debido al factor de warping empleado. Este comportamiento refuerza las potencialidad de este tipo de modelado.

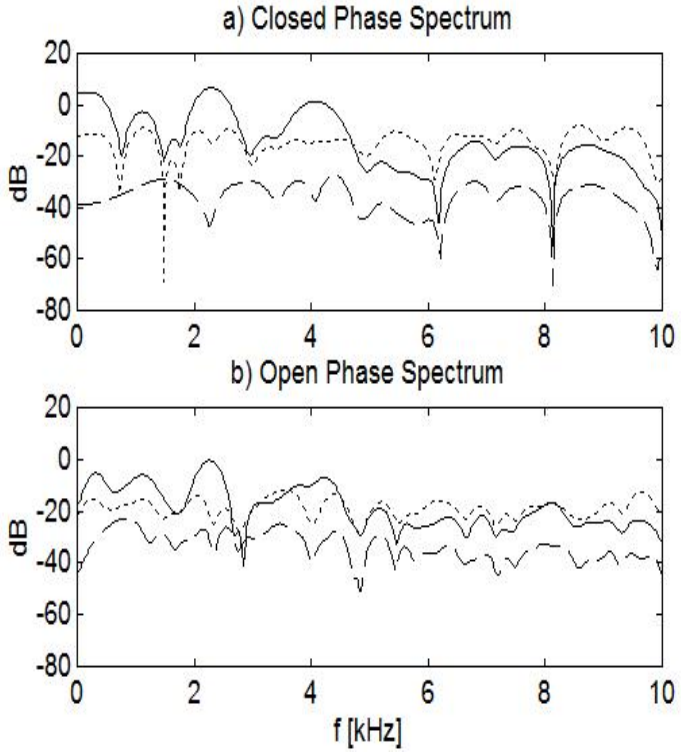



Figura 4.3: Error spectral para los métodos (W)CPPZ3 y (E)CCPZ3.

### 4.5.3. Efecto del modelado de fases sobre periodos contiguos

El tercer objetivo es seguir indagando en la mejora de los métodos de estimación que emplean periodos contiguos analizando su comportamiento sobre el tamaño de los intervalos de análisis.

Los experimentos desarrollados en este apartado, utilizaron un modelo ARMA(16,15) en ambas fases, con un parámetros de extensión de 20 muestras y un factor de warping de 0.75 en cada caso y se analizó la relación señal a error de reconstrucción, SSR, (en dB), en función del tamaño de las tramas de análisis y se representan en la figura 4.4. El intervalo de análisis considerado fue desde el mínimo al máximo encontrado en los registros. Los métodos empleados fueron (E)CPPZ3 y (W)CPPZ3.

Las curvas muestran un mejor comportamiento de los métodos transformados en frecuencia sobre los no transformados. Además indican que los métodos transformados trabajan mejor con tramas de mayor tamaño que con las más cortas. El comportamiento es más o menos estable hasta 10 ms y mejora considerablemte para intervalos de análisis superiores a 12 ms.

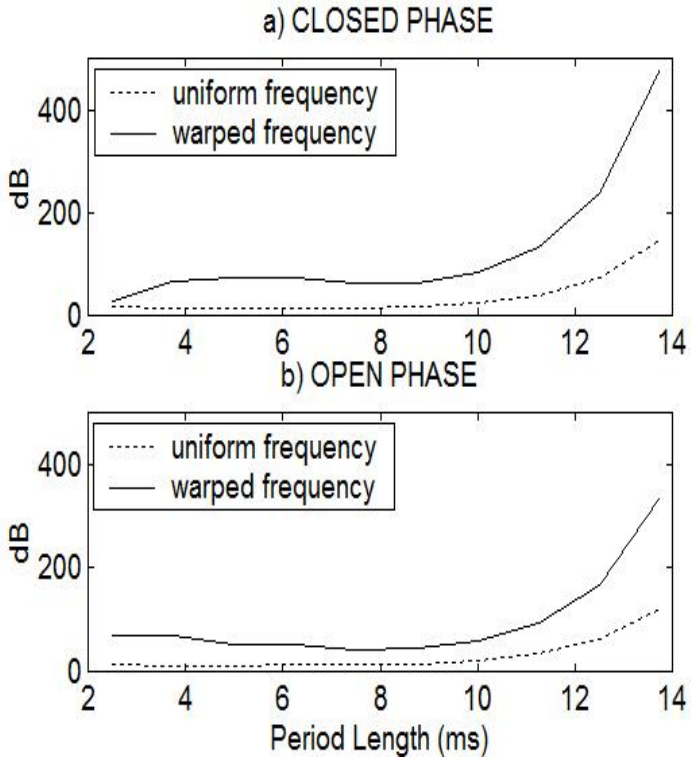



Figura 4.4: Evolución del SRR con el tamaño del periodo.

### 4.5.4. Efecto en el seguimiento de las características fonéticas en voz sonora

El cuarto objetivo es evaluar la capacidad de los métodos transformados en la caracterización efectiva de la voz, como soporte básico en los sistemas de reconocimiento automático del habla. Vamos a usarlos en una aplicación mimética a la del capítulo anterior que consistía en el seguimiento de las características fonéticas en voz sonora por medio del análisis localizado y nos fijarmos igualmente en la fiabilidad de los resultados. Los datos utilizados serán las mismas transiciones fonéticas de aquel experimento y el propósito no es tanto hacer un estudio exhaustivo sobre todas las transiciones como determinar el potencial de los métodos transformados en la tarea de parametrización de la voz.

Los modelos empleados son de tipo ARMA (12,11). El seguimiento se hará sobre los polos de los modelos, que representan a los formantes en cada fase. Los formantes son obtenidos de las raíces del polinomio AR de cada modelo, a los que se le imponen las mismas condiciones del apartado 3.5.3. Las fases cerrada y abierta se toman del 40 % y del 48 % del tamaño del periodo respectivamente, a partir del ICG. Los métodos empleados serán el AMCC3, el (E)CPCZ3 y el (W)CPCZ3, con los que pretendemos evidenciar la potencialidad de las propuestas transformadas.

Los experimentos se realizaron sobre los segmentos siguientes: el segmento de voz

/aveller/ pronunciado por un hombre y compuesto por transiciones vocal-consonante y consonante-vocal; el segmento de voz /an/ pronunciada por una mujer, que contiene una transición vocal-consonante; y el segmento de voz /ndmei/ pronunciado por un hombre, que contiene transiciones consonante-consonante, consonante-vocal y vocal-vocal. En las figuras los formantes de la fase cerrada los representaremos con símbolos '*', los de fase abierta con el símbolo '.' y los mostraremos en la misma gráfica para comprobar como siguen el comportamiento espectral del segmento de voz y observar las diferencias en las características de cada fase.

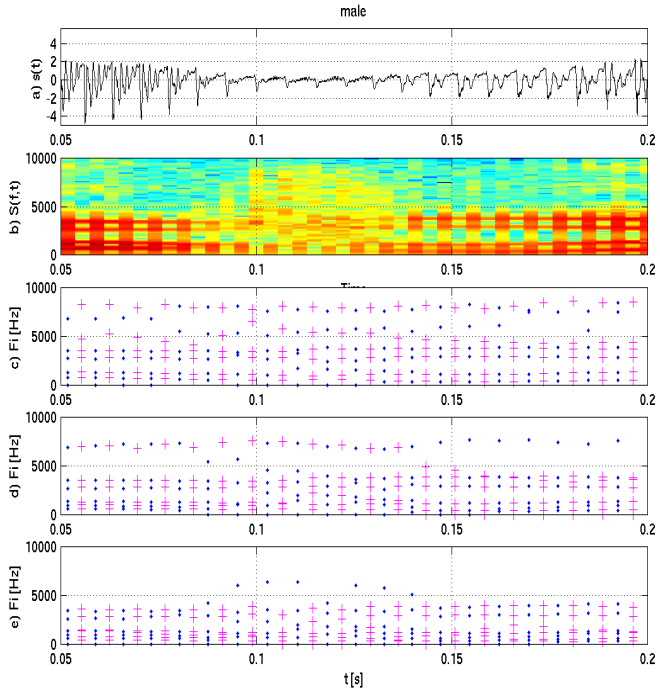



Figura 4.5: a) Registro sonoro /aveller/ de hombre. b) Espectrograma. c) (W)CPCZ3 con $\lambda$=0.25. d) (W)CPCZ3 con $\lambda$=0.45. e) (W)CPCZ3 con $\lambda$=0.65.

En la figura 4.5 se representa la transición /aveller/, en la cual vamos a ver como influye el factor de warping en el seguimiento de los formantes. Para ello hemos realizado un experimento variando el factor de warping entre los valores más representativos, que son 0.25, 0.45 y 0.65 y que se corresponderían con las siguientes frecuencias de turning de 4.2 kHz, 3.5 kHz y 2.75 kHz respectivamente. En la gráfica se observa como con el factor de warping mayor, 0.65, que tiene la mejor resolución en baja frecuencia, las más importantes perceptualmente, quedan pocos formantes para seguir las frecuencias altas. El método con el factor de warping intermedio, 0.45, parace un buen compromiso para seguir las buenas caracterísctcas perceptuales de las bajas frecuencias y al mismo tiempo ser capaz de seguir las características de las frecuencias altas, buenas para la inteligibi-

lidad . El efecto con 0.25 parece que no beneficia la captación de las características de baja frecuencia, distribuyéndose la representación de los formantes de forma uniforme a lo largo de ancho de banda de la señal.

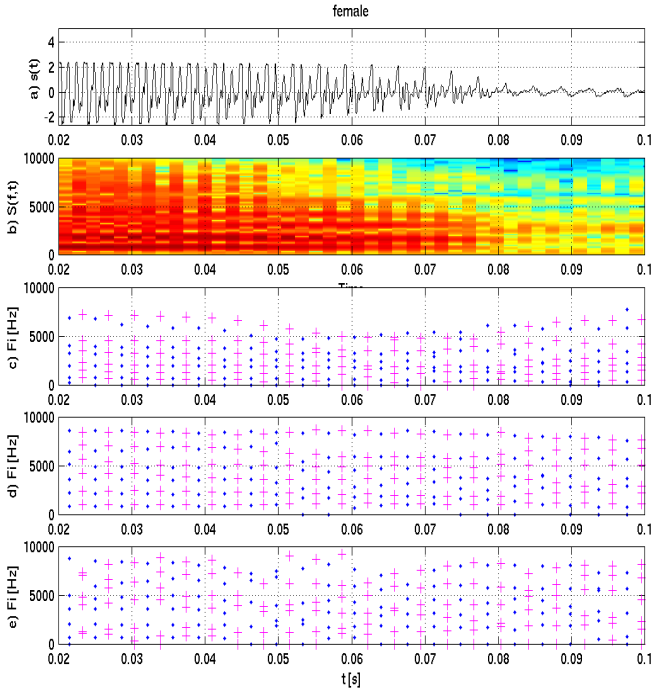



Figura 4.6: a) Registro sonoro /an/ de mujer. b) Espectrograma. c) Estimación (W)CPCZ3 con $\lambda$=0.45. d) Estimación (E)CPCZ3. e) Estimación AMCC3.

En la figura 4.6 se muestra la transición /an/ con la que vamos a comparar los métodos transformados frente a los no transformados y los clásicos. Se puede apreciar como con el método transformado (W)CPCZ3, con $\lambda$=0.45, se obtiene una mejor resolución en bajas frecuencias comparado con el (E)CPCZ3, a la vez que parece que conlleva un contenido aceptable de información de las frecuencias altas y por supuesto un comportamiento más regular que el AMCC3. Los formantes de las fases cerrada y abierta siguen el espectrograma con sus propias diferencias y son capaces de definir la transición claramente.

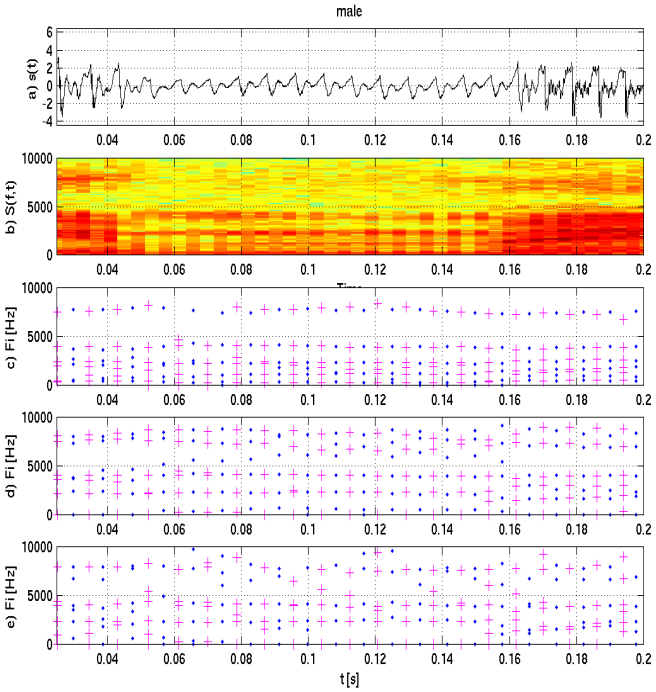



Figura 4.7: a) Registro sonoro /ndmei/ de hombre. b) Espectrograma. c) Estimación (W)CPCZ3 con $\lambda$=0.45. d) Estimación (E)CPCZ3. e) Estimación AMCC3.

En la figura 4.7 se representa la transición /ndmei/, con un mayor número de transiciones respecto de las otras. Se observa la eficiencia del estimador transformado, (W)CPCZ3 y su buena consistencia en las transiciones, así como su regularidad a los largo del espectrograma.

Finalmente quisimos comprobar el comportamiento del método de estimación en un entorno adverso, con presencia de ruido moderadamente bajo, a 20 dB de relación señal a ruido y con un error aleatorio en las marcas de ICG, de 3 muestras como máximo. La figura 4.8 representa el experimento con el segmento /aveller/ y en ella se puede apreciar la buena respuesta de nuestro estimador (W)CPCZ3 inmerso en un entorno adverso. Las estimaciones son similares a las alcanzadas en el caso de la voz limpia de la figura 4.5 y superiore sen cualquier caso a los métodos no transformados. Creemos que la mejor resoluciones de los métodos transformados en las bandas bajas es la razón de su robustez.

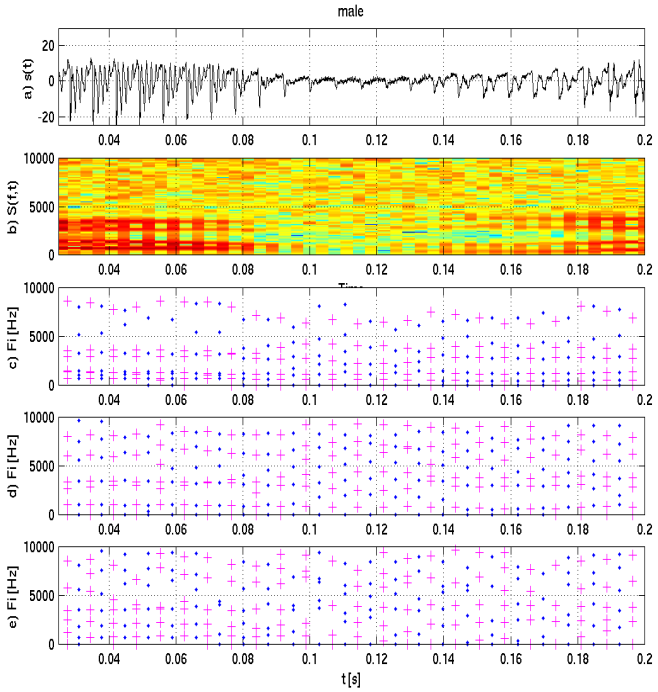



Figura 4.8: a) Registro sonoro ruidoso /aveller/ de hombre. b) Espectrograma. c) Estimación (W)CPCZ3 con $\lambda$=0.45. d) Estimación (E)CPCZ. e) Estimación AMCC3.

## 4.6. Conclusiones

En este capítulo hemos presentado un marco general de procesado localizado apropiado para el seguimiento de las variaciones naturales de la voz sonora, a la vez que se incorporan las características perceptuales que son uso común. Desde el punto de vista de la implementabilidad, hemos adaptado la formulación matricial para la estimación de los coeficientes de dichos filtros en el sentido de mínimos cuadrados. Con esta formulación hemos abordado el problema de seguimiento de las características de polos y ceros de la voz sonora.

## Bibliografía

[1] P. J. Quintana-Morales and J. L. Navarro-Mesa, "An approach to common acoustical pole and zero modeling of consecutive periods of voiced speech," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, ISCA, 2003.

[2] A. Oppenheim, D. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast fourier transform," *Proceedings of the IEEE*, vol. 59, pp. 299 – 301, Feb. 1971.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, Apr. 1990. PMID: 2341679.

[4] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[5] I. Smith, J.O. and J. Abel, "The bark bilinear transform," in *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pp. 202 –205, Oct. 1995.

[6] K. Roth, I. Kauppinen, P. Esquef, and V. Valimaki, "Frequency warped burg's method for AR-modeling," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pp. 5 – 8, Oct. 2003.

[7] A. Harma and U. Laine, "A comparison of warped and conventional linear predictive coding," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, pp. 579 –588, July 2001.

[8] H. Matsumoto and M. Moroto, "Evaluation of mel-LPC cepstrum in a large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, pp. 117 –120 vol.1, 2001.

[9] A. Harma, "Implementation of frequency-warped recursive filters," *Signal Processing*, vol. 80, pp. 543–548, Mar. 2000.

[10] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, pp. 451 –515, Apr. 2000.

[11] M. Karjalainen, A. Harma, and U. Laine, "Realizable warped IIR filters and their properties," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 3, pp. 2205 –2208 vol.3, Apr. 1997.

[12] P. J. Quintana-Morales and J. L. Navarro-Mesa, "Frequency warped ARMA analysis of the closed and the open phase of voiced speech," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*, ISCA, 2004.

Capítulo 5

# FUNCIÓN DENSIDAD DE PROBABILIDAD MEDIANTE MODELOS PARAMÉTRICOS LINEALES

## 5.1. Introducción

El modelado estadístico de procesos o series temporales plantea la necesidad de buscar los estadísticos que mejor los representen para cada aplicación. La función densidad de probabilidad es, sin duda, muy atractiva porque de ella podremos generar el resto de estadísticos. En el caso del procesado de la señal de voz nos podría servir para modelar estadísticamente tanto las muestras de voz como los parámetros que extraemos para su caracterización.

En este capítulo trabajaremos en la definición de funciones de densidad de probabilidad a partir de modelos paramétricos lineales, y analizaremos las modificaciones

que deben sufrir éstos últimos para que puedan ser consideradas como las primeras. Inicialmente veremos diferentes métodos de estimación e introduciremos uno basado en el error de reconstrucción, al estilo de los desarrollados en el capítulo 3. Adicionalmente extenderemos el trabajo sobre mezclas de funciones de densidad de probabilidad y desarrollaremos algunos métodos de estimación basados en la maximización de la verosimilitud logarítmica, a través del algoritmo de expectation-maximization (EM).

## 5.2. Funciones de densidad de probabilidad paramétricas frente a no paramétricas

El conocimiento estadístico de un fenómeno aleatorio podemos decir que está condensado en su función densidad de probabilidad (PDF, por sus siglas en inglés). Se entiende por ello la importancia que tiene y la necesidad que nos mueve a obtenerla. Diversos autores que trabajan en aplicaciones de todo tipo se esfuerzan por estimarla a partir de los datos observados en campos tan dispares como economía, biología o física, [12, 11, 10], siendo el reconocimiento de modelos donde más nos interesa su aplicación.

Los métodos de estimación de PDF a partir de un conjunto finito de datos se pueden dividir entre paramétricos y no paramétricos, atendiendo a si tienen, o no, un modelo subyacente en el que apoyarse. En los no paramétricos, no se tiene información acerca del proceso que lo está generando, o es demasiado complejo para tener una expresión analítica. Estos servirán, por tanto, para definir cualquier fenómeno pero sus resultados dependerán de la cantidad de datos para caracterizar el fenómeno completamente. Entre ellos el más popular está basado en la técnica de núcleos (técnicas de kernel, como el histograma) [9, 8] . Por el contrario, los métodos paramétricos definen explícitamente el fenómeno en el que se basan y controlan con un número relativamente pequeño de parámetros las características del proceso. Son más eficientes y robustos. Necesitan un número menor de datos para estimar los parámetros y obtienen una PDF que se ajustará a todo el proceso en mayor o menor medida según haya sido más o menos acertada la definición del modelo. Entre ellos encontramos PDF de procesos Uniformes, Gaussianos, Laplacianos, etc. [7]. En esta tesis hacemos aportaciones sobre los métodos de estimación basados en el modelado ARMA. Si bien este enfoque no es nuevo, sí que hacemos aportaciones sobre los métodos de estimación. Además introducimos la formulación para la estimación de mezclas de densidad tipo AR de orden 2, similar a la mezcla de Gaussianas.

Cuando el fenómeno estudiado es desconocido o más complejo que el producido por un proceso específico simple y, por tanto, no definible por un solo modelo, podemos hacer uso de una alternativa como es la mezcla finita de PDF [20, 14]. Esta es útil en dos casos cuanto menos. Por un lado, cuando el proceso complejo es la mezcla de varios procesos conocidos. Y por otro, cuando se quiere aproximar su distribución de datos con una mezcla de funciones. En el primer caso solo tendríamos que ajustarnos a cada proceso. En el segundo deberíamos definir la función a emplear y estimar los parámetros de la mezcla. Como ejemplo de éste último caso tenemos la mezcla de gaussianas como modelo, ya que una PDF puede ser adecuadamente especificada por una mezcla finita

de PDF normales [5]. De cualquiera de las dos formas debemos manejar las PDF que componen la mezcla con formas adecuadas a los estadísticos locales.

## 5.3. Aproximación basada en modelos paramétricos de densidad espectral de potencia

Entre las PDF paramétricas tenemos una familia basada en los modelos paramétricos de la densidad espectral de potencia (PSD). Usadas estas porque ambos tipos de densidades son funciones finitas, reales y positivas. Ambas poseen propiedades muy interesantes para caracterizar procesos estocásticos [16, 23] y disponen de las conocidas técnicas de estimación espectral.

Varios autores han trabajado con este tipo de aproximaciones. Por ejemplo, [4] estudió la aproximación con modelos AR que explota su representación con las Series de Fourier, al igual que lo hizo [3] con modelos ARMA. En ambos casos se demuestra la potencia de los modelos para aproximar cualquier tipo de función, aunque señalan la débil consistencia de los estimadores. Desde otros puntos de vista también se ha abordado el problema, donde se ha trabajado principalmente con modelos AR, [15] desde la teoría de sistemas no lineales y [21] desde la teoría espectral, el cual remarcaba la ventaja que podría tener el modelo en la definición de las colas en problemas no gaussianos. Otros autores también han empleado los modelos AR en diversos problemas. Así, en [13] se usan en estimación de entropía o [2] en estimación ciega de canal, donde no solo se insiste en las propiedades de aproximación, sino que lo pone en práctica.

En dichos trabajos, a partir de las similitudes entre la función característica y la de autocorrelación [16], se establece el cálculo de una función densidad de probabilidad a través de una densidad espectral de potencia. En ellos se realiza un cálculo elegante y eficaz de una función densidad de probabilidad paramétrica de tipo AR, utilizando una función densidad espectral de potencia asociada a un nuevo proceso, este debidamente relacionado con aquel del que se quiere obtener la función de probabilidad. Esto mantiene abierto el camino para plantear el uso de otras densidades paramétricas lineales y el desarrollo de métodos de estimación.

A partir de lo anterior, en esta tesis se proponen inicialmente dos caminos para la definición de una PDF basada en PSD autorregresivo de media móvil (ARMA). Ambos caminos se introducen para caracterizar mejor las funciones que se pretenden representar, utilizando tanto polos como ceros. Por un lado, evolucionaremos desde el trabajo de Kay [21] para añadir al modelo AR una parte MA. Por otro lado, plantearemos un nueva forma de trabajar en el modelado probabilístico con el modelo ARMA basada en el error de reconstrucción, al estilo de los desarrollados en el capítulo 3 [19]. Este último desarrollo parte de la idea de generar un nuevo proceso muestral, a partir del original, que tuviera como PSD la PDF que se propone. Minimizando el error de reconstrucción se esperan obtener estimaciones más suaves, menos sensibles al error de orden y con menos picos espurios.

Antes de continuar con la definición de los modelos vamos establecer la relación entre las funciones PSD y PDF y a tratar con dos aspectos que afectan a todas las

aproximaciones basadas en funciones de densidad de potencia. Lo primero que nos encontramos es la necesidad de tener que truncar el espacio muestral, ya que los modelos de PSD están definidos en el eje de abscisas entre $[-\pi\ \pi]$ y lo segundo es con la necesidad de normalizar a 1 el área bajo la aproximación PDF, para que tenga características de función de probabilidad. El desarrollo que presentamos a continuación lo haremos con un modelo AR, pero será extensible al caso ARMA.

### 5.3.1. Ajuste de parametros para la aproximación basadas en PSD

Para utlizar las funciones PSD de tipo AR como PDF, vamos a definir primero el proceso de la señal en la variable de frecuencia, $w$ y después vamos a identificar dicha variable con un proceso muestral $\{x\}$. Sea pues $S(w)$ la PSD de un modelo AR de orden $p$, con $w = 2\pi x$.

$$S(w)\mid_{w=2\pi x} = \frac{\eta^2}{\left|1+\sum_{i=1}^{P} a_i e^{-j2\pi xi}\right|^2} = \frac{\eta^2}{\left|1+\sum_{i=1}^{P} e^{-\frac{Bi}{2}+jw_i} e^{-j2\pi x}\right|^2} \tag{5.1}$$

donde la definición dada de $w$ permite trabajar en radianes a partir de la variable aleatoria x, para la cual queremos estimar su PDF.

$S(w)$ es una función todo-polos, que está caracterizada por picos espectrales controlados por los polos, siendo $w_i$ la posición de cada pico o frecuencia de resonancia y $B_i$ su ancho de banda. También ocurre que la función tiende a cero cuando $w$ se aleja de los picos. Estas funciones dependen del orden $p$ y de los parámetros, que pueden ser de varios tipos: los coeficientes del polinomio, los polos, etc.

Esta función de PSD es una función real y positiva, por lo que podría describir a una PDF cuando la variable $w$ describa un proceso muestral $\{x\}$ a través de la definición de la variable, $w = 2\pi x$. En este caso, su transformada de Fourier inversa, que representa la autocorrelación del proceso, $R(k)$, se puede identificar con la función característica, $\Psi(k)$, de $S(x)$ , con $x = w/2\pi$, cuando ésta es utilizada como PDF [17].

$$R(k) = \mathcal{F}^{-1}\left[S(w)\right] = \frac{1}{2\pi}\int_{-1/2}^{1/2} S(x)e^{j2\pi xk}dx = E\left\{e^{j2\pi xk}\right\} = \Psi(k) \tag{5.2}$$

Las muestras de la función de autocorrelación llegan, por tanto, a ser muestras de la función característica.

Encontramos, sin embargo, dos diferencias específicas entre las funciones PDF y PSD, y que tendremos que superar. La primera es que, siguiendo la teoría espectral, la función $S(w)$ cuando tiene coeficientes $a_i$ reales es par y periódica en $w$, con periodicidad $2\pi$ radianes, quedando definida completamente en $w \in [0\ \pi]$ radianes. Así pues, la función $S(w)$ usada como PDF, que por su naturaleza no es periódica, tendrá que definirse también completamente sobre$[0\ \pi]$, y en consecuencia la variable $x = w/2\pi$, que caracteriza al conjunto de muestras, tendrá que concentrarse en el intervalo $x \in [0\ 1/2]$.

Por comodidad vamos a renombrar la variable que representa el espacio muestral original y pasarlo a denominarla *{y}*, para hacer una traslación al eje positivo y normalizarlo respecto a un valor que nos asegure que esté en el intervalo. Para ello habrá que hacer una previsión del espacio muestral.

Una manera práctica de implementar la normalización podría ser por medio de una ponderación de con el valor máximo del espacio previsto [15], o de su desviación típica [21]. En distribuciones no acotadas se tendría que definir un intervalo finito en el que estuviera representado un porcentaje suficientemente alto del proceso. En el caso del intervalo finito, las muestras transformadas, que llamamos ahora *{x}*, se pueden obtener de la siguiente forma:

$$x = \frac{y - \tilde{y}_{min}}{(\tilde{y}_{max} - \tilde{y}_{min})} \tag{5.3}$$

siendo $\tilde{y}_{max}$ e $\tilde{y}_{min}$ los valores máximo y mínimo de $y$, respectivamente. En la práctica estos valores se han de aproximar lo más fielmente posible. Ésta no debe ser una gran preocupación, debido a que la mayoría de las variables aleatorias de procesos estocásticos reales están acotadas.

La segunda diferencia destacada es que, por definición, dada una PDF, la probabilidad total tiene que ser igual a 1. Para asegurarnos el cumplimiento de dicha propiedad en el espacio muestral original tendremos que ajustar el valor de $\eta^2$ en la PSD para que se cumpla 5.4 en el intervalo adecuado del espacio muestral transformado,

$$\int_0^{1/2} \frac{\eta^2}{\left|1 + \sum_{i=1}^{P} a_i e^{-j2\pi xi}\right|^2} dx = \frac{2\pi}{(\tilde{y}_{max} - \tilde{y}_{min})} \tag{5.4}$$

Para encontrar una solución exacta a la condición de área unidad también podemos aplicar el teorema de los residuos de Cauchy [6] sobre la PSD:

$$\int_{-\pi}^{\pi} S(w)dw = \int_{-1/2}^{1/2} \left|H(e^{j2\pi x})\right|^2 d(2\pi x) = \frac{1}{j}\oint_{|z|=1} z^{-1}H(z)H(-z)dz = \eta^2 \sum_{i=1}^{2P} 2\pi A_i = 1 \tag{5.5}$$

donde hemos usado la simetría par de *S(w)* alrededor de *w=0*, $\{A_i\}$ representa los residuos de los polos de $z^{-1}H(z)H(-z)$ en el interior del circulo unidad y $\sum_{i=1}^{2P} 2\pi A_i \in \Re$, ya que los $A_i$ aparecen en pares complejos conjugados. Así, como esperábamos, la expresión de $\eta^2$ definida en 5.1 y que obtenemos a partir de 5.5 es un número real, asegura la normalización al área unidad y es compacta.

### 5.3.2. Estimación de la PDF basada en Modelos AR

Definida la aproximación PDF como una PSD de modelo AR, los parámetros del modelo pueden estimarse a partir de las ecuaciones de Yule-Walker como expusimos en el capítulo 2 y como hace [21]. La solución al sistema de ecuaciones $\boldsymbol{R}_y\boldsymbol{a} = -\boldsymbol{r}_y$, donde $\boldsymbol{a} = \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix}^T$ es el vector de parámetros a estimar, $\boldsymbol{R}_y$ la matriz de correlaciones y $\boldsymbol{r}_y$ el vector de correlaciones, se construirá desde sus estimaciones, teniendo en cuenta su relación con la función característica $\Psi_y(i)$, expresadas en 5.2.

Ahora, $\boldsymbol{R}_y$ y $\boldsymbol{r}_y$ serán la matriz y el vector de funciones características de nuestro proceso y sus estimaciones estimaciones serán $\left(\widehat{\boldsymbol{R}}_y\right)_{1\leq i,j\leq p} = \widehat{\Psi}_y(i-j)$ y $(\widehat{\boldsymbol{r}}_y)_{1\leq i\leq p} = \widehat{\Psi}_y(i)$ respectivamente. Las estimaciones de dicha función se pueden obtener por el método de los momentos, como hace [21]. Cualquier método de estimación de parámetros AR puede ser usado para su resolución. Y una vez hayan sido encontrados estos coeficientes, podremos deducir la constante del numerador del modelo como $\sigma^2 = \widehat{R}_y(0) + \sum_{i=1}^{P} a(i)\widehat{R}_y(-i)$.

### 5.3.3. Aproximación basada en Modelos ARMA(p,2)

A partir de la definición de la PDF basada en modelos AR dada por [21] y teniendo en cuenta que el espacio muestral transformado tiene que estar en el intervalo definido por la PSD original, en $w \in [0\ \pi]$, en esta tesis proponemos trabajar con un modelo ARMA(p,2) que lo fuerce a ello por encima de todo. A la parte AR(p) definida anteriormente le añadiríamos una parte MA de orden 2, que ponga un cero en la PSD en 0 y otro en pi. El nuevo modelo quedaría definido de la siguiente forma:

$$S(w)\mid_{w=2\pi x} = \frac{\eta^2 \left|1 - e^{-j4\pi xi}\right|^2}{\left|1 + \sum_{i=1}^{P} a_i e^{-j2\pi xi}\right|^2} \tag{5.6}$$

### 5.3.4. Estimación de la PDF basada en la aproximación con Modelos ARMA(p,2)

Definida la aproximación PDF como una PSD de modelo ARMA, los parámetros del modelo pueden estimarse a partir de las ecuaciones de Yule-Walker extendido, como expusimos en el capítulo 2. Solo emplearemos las ecuaciones para obtener los parámetros de la parte AR(p) del proceso ya que los coeficientes de la parte MA están fijados a 1.

La solución al sistema tiene que resolver el siguiente sistema de ecuaciones:

$$R_y(m) = -\sum_{k=1}^{p} a_k R_y(m-k) \qquad m \geq q+1 \tag{5.7}$$

donde podemos definir el número de ecuaciones con las que trabajar. Las muestras de la función de autocorrelación las obtendremos desde sus estimaciones, que serán las

estimaciones de la función característica, $\widehat{R}_y(m) = \widehat{\Psi}_y(m)$ .

### 5.3.5. Aproximación basada en Modelos ARMA(p,q)

En esta sección vamos a proponer la aproximación de PDF con modelos de PSD de tipo ARMA [19] desde una perspectiva un poco diferente a la anterior. Para empezar nuestra aproximación preguntarnos algo que también subyace en AR y ARMA(p,2). Esto es, si podríamos si se puede encontrar un nuevo proceso $z_y(n)$ cuyo espectro sea precisamente la PDF del proceso original *{y}* [13]y nos encontraríamos con una respuesta afitrmativa. El proceso sería $z_y(n) = e^{jny+\varphi}$ [15], que tiene dicha propiedad si *{y}* es un proceso muestral y $\varphi$ está uniformemente distribuido sobre [0 $2\pi$] y es independiente de *{y}*. En este caso la variable $n$ juega el mismo papel que lo haría la variable tiempo. Por simplicidad omitiremos $\varphi$ en nuestra formulación. Se podría demostrar que la función de autocorrelación $z_y(n)$ depende solo del intervalo de tiempo y que es también igual a la función característica de *{y}*, $\Psi_y(i)$. Así, hemos creado un nuevo proceso, $z_y(n)$, cuya función de autocorrelación es a la vez la función característica del proceso original *{y}* y su función de PSD igual a la PDF de *{y}*.

Siendo nuestro proceso de tipo ARMA, la señal de salida, $z_y(n)$, vendrá dada por la siguiente ecuación en diferencias:

$$z_y(n) = -\sum_{i=1}^{p} a_i z_y(n-i) + \sum_{j=0}^{q} b_j z_u(n-j) \tag{5.8}$$

donde $z_u(n) = e^{jnu}$ es la señal de excitación construida desde el proceso de excitación *{u}*, *{n=0,...,N-1}*, $\{a_i\}$ y $\{b_j\}$ son los coeficientes de la parte AR y MA, respectivamente. Y la PDF en el dominio de la frecuencia podrá expresarse como:

$$p(y) = K\frac{\left|b_0 + b_1 e^{-jy} + \cdots + b_q e^{-jyq}\right|^2}{\left|1 + a_1 e^{-jy} + \cdots + a_p e^{-jyp}\right|^2} \tag{5.9}$$

donde como antes $-1/2 \leq y \leq 1/2$ y K será elegido para que se cumpla la condición de normalización sobre le espacio muestral. Si aseguramos que la PDF sea simétrica alrededor de *y=0*, entonces los coeficientes $\{a_i,\ b_j\}$ serán reales. Como en el caso de los métodos basados en AR, debemos asegurar que el proceso aleatorio estará contenido entre -$\frac{1}{2}$ y $\frac{1}{2}$.

Ahora procederemos con la estimación de la PDF, donde aplicaremos un método que minimiza la señal del error de reconstrucción, en la línea del capítulo 3.

### 5.3.6. Estimación de los parámetros del modelo ARMA(p,q)

Recordando la ecuación 5.8 podemos definir la señal del error de reconstrucción como sigue:

$$z_e(n) = z_y(n) + \sum_{i=1}^{p} a_i z_y(n-i) + \sum_{j=0}^{q} b_j z_u(n-j) \tag{5.10}$$

Usando notación matricial y suponiendo que $z_u(n)$ es conocido o apropiadamente estimado (por ejemplo como la salida de un filtro de error de predicción de orden elevado), la 5.10 se convierte en la siguiente:

$$\begin{aligned} \boldsymbol{z_e} &= \boldsymbol{z_y} - \left[\begin{array}{cc} \boldsymbol{Y} & \boldsymbol{U} \end{array}\right] \boldsymbol{h} = \boldsymbol{z_y} - \boldsymbol{H}\boldsymbol{h} \\ \boldsymbol{z_y} &= \left[\begin{array}{ccccccc} z_y(0) & \cdots & z_y(N-1) & 0 & \cdots & 0 \end{array}\right]^T \\ \boldsymbol{z_e} &= \left[\begin{array}{ccc} z_e(0) & \cdots & z_e(N+p-1) \end{array}\right]^T \\ \boldsymbol{h} &= \left[\begin{array}{ccccccc} a_1 & \cdots & a_p & b_0 & \cdots & b_q \end{array}\right]^T \end{aligned} \tag{5.11}$$

$$\boldsymbol{Y} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ z_y(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ z_y(N-2) & z_y(N-3) & \cdots & z_y(N-p-1) \\ z_y(N-1) & z_y(N-2) & \cdots & z_y(N-p) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_y(N-1) \end{bmatrix}$$

$$\boldsymbol{U}_n = \begin{bmatrix} z_e(0) & 0 & \cdots & 0 \\ z_e(1) & z_e(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ z_e(N-1) & z_e(N-2) & \cdots & z_e(N-q-1) \\ 0 & z_e(N-1) & \cdots & z_e(N-q) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

donde el intervalo *{n=0,...,N+p-1}*, en el cual el error está definido, es debido a la longitud infinita inherente a la respuesta al impulso de la parte AR del modelo. El vector de coeficientes $\boldsymbol{h}$ que minimiza el error cuadrático medio puede ser obtenido en sentido

de mínimos cuadrados llevando a la siguiente solución:

$$\boldsymbol{z_y} = \boldsymbol{H}^T\boldsymbol{h} \rightarrow \mathbf{h} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{z_y} \tag{5.12}$$

Es interesante notar que $(\boldsymbol{H}^T\boldsymbol{H})^{-1}$y $\boldsymbol{H}^T\boldsymbol{z_y}$en 5.12 son la matriz y el vector de correlaciones, respectivamente, cuyos elementos son estimaciones de la función característica obtenidas con el estimador de los momentos.

Antes de seguir sería interesante hacer algunas observaciones. Para distribuciones de colas largas, el factor de escala debe de ser elegido con especial cuidado. Por ejemplo, si el factor de multiplicación comprime en exceso la distribución alrededor de cero, entonces necesitaremos incrementar la resolución de frecuencia en las bandas bajas para obtener una buena descripción para densidades multimodales. Esto puede causar la necesidad de elevar los órdenes del modelo.

## 5.4. Modelos de mezclas de PDF basadas en PSD AR(2)

Siguiendo el camino marcado inicialmente, proponemos ahora introducir las PDF basadas en PSD en una mezcla de densidades. Definiremos entonces un modelo de mezclas de PSD lineales y desarrollaremos un método para estimar los parámetros basados en la técnica EM [18](bilmes98). Trabajando con funciones AR de orden 2, pretendemos simplificar los cálculos del modelado y seguir abarcando las estadísticas globales del proceso caracterizando las estadísticas locales con cada una de las componentes AR de la mezcla. El desarrollo del método EM para nuestro caso particular de funciones AR nos proporcionará las fórmulas de recurrencia para estimar los parámetros.

### 5.4.1. Mezclas de densidades AR(2)

Construyamos ahora una mezcla finita de PDF con modelos AR(2) para obtener un modelo más potente y flexible. Para ello definimos una distribución *p(x)*, con un conjunto de parámetros $\Theta$, como una combinación lineal de PDF componentes [20], que en nuestro caso serán de tipo AR(2), $S_m(x)$.

$$p(x|\Theta) = \sum_{m=1}^{M} \alpha_m S_m(x|\theta_m) \tag{5.13}$$

siendo $M$ el número de componentes de la mezcla, $\alpha_m$ el factor de ponderación de PDF componente $S_m(x)$, con parámetros $\theta_m$ y cumpliendo que

$$0 \leq \alpha_m \leq 1 \qquad y \qquad \sum_{m=1}^{M} \alpha_m = 1 \tag{5.14}$$

Las PDF constitutivas de la mezcla que nos van a interesar serán las más simples, o sea, aquellas caracterizadas por un solo máximo en x$\in$ [0 1/2]. Esta simplificación matemática se puede aplicar sin disminuir la potencia de la mezcla de la 5.13 siempre y cuando se utilice un número de componentes suficientemente alto para poder representar todos los máximos o modos del proceso. Las componentes, $S_m(x|\theta_m)$, serán AR(2) con la frecuencia de resonancia y el ancho de banda del par de polos complejos conjugados como parámetros.

$$S_m(x|\boldsymbol{\theta}_m) = \frac{\eta_m^2}{|1 - 2e^{\sigma_m}\cos(w_m)e^{-j2\pi x} + e^{2\sigma_m}e^{-j4\kappa x}|^2} = \frac{\eta_m^2}{|A_m(j2\pi x)|^2} \tag{5.15}$$

siendo $\sigma_m = \frac{B_m}{2}$, $\boldsymbol{\theta}_m = \{\sigma_m,\ w_m\}$ el conjunto de parámetros y $\eta_m^2$ la constante que normaliza a 1 el área bajo la componente $m$ de la mezcla .

Con el fin de justificar el uso de la 5.13 y la 5.15 como un modelo para la estimación de PDF hay que tener en cuenta lo que simboliza una resonancia en la teoría de sistemas. Ésta representa la frecuencia, $w_m$, en la cual la respuesta del sistema es localmente máxima, siendo el valor en aquel máximo dependiente de $\sigma_m$. Al extender este concepto a PDF hay que tener presente que la función de densidad puede tener un máximo local ubicado en diferentes frecuencias. Esto se puede ver como un modo en la densidad probabilística. Además, los coeficientes de mezcla, $\alpha_m$, pueden ser vistos como los pesos de un sistema lineal. Así, en 5.13 ellos representan la contribución de cada componente de la mezcla. Por lo tanto, podemos esperar que el modelo de mezcla propuesto sea capaz de aproximar cualquier PDF multimodal arbitraria. Esta expectativa está en acuerdo con el hecho bien conocido (por ejemplo, [22, 13]) de que, para órdenes suficientemente grandes, un modelo AR puede representar cualquier proceso. En nuestro caso el número de componentes de la mezcla, $M$, también debe ser suficientemente grande.

Una vez definido el modelo de mezcla autorregresivo debemos especificar un algoritmo especialmente adecuado para entrenar a este tipo de funciones de densidad. En este trabajo se adopta el expectation-maximitation (EM) [18], que ha demostrado ser adecuado para una amplia variedad de aplicaciones y garantizar la convergencia a un máximo local.

### 5.4.2. Estimación de los parámetros del modelo de mezclas AR(2)

El algorítmo EM es una generalización del estimador de máxima verosimilitud (ML) desarrollado para simplificar la optimización de funciones intratables, como la considera, suponiendo que hay información adicional oculta al observador o datos incompletos.

Trabajando con un conjunto de datos observados de $N$ muestras definidas en $x_n \in \left[-\frac{1}{2},\ \frac{1}{2}\right]$, el algoritmo EM intenta, iterativamente, encontrar los parámetros del modelo, $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}_m,\ \boldsymbol{\sigma}_m,\ \boldsymbol{w}_m\}$ que maximicen la verosimilitud logarítmica esperada, como

desarrollada en

$$Q(d\boldsymbol{\Theta},\, \boldsymbol{\Theta}^g) = \sum_{m=1}^{M}\sum_{n=1}^{N} \log(\alpha_m) p(m \,|\, x_n,\, \boldsymbol{\Theta}^g) + \sum_{m=1}^{M}\sum_{n=1}^{N} \log(p_m(x_n \,|\, \boldsymbol{\theta}_n)) p(m \,|\, x_n,\, \boldsymbol{\Theta}^g) \tag{5.16}$$

Esta función tiene información oculta acerca de la probabilidad de la componente *m-ésima* de la mezcla que ha generado cada observación, $p(m \,|\, x_n,\, \boldsymbol{\Theta}^g)$ y se actualiza en cada iteración, desde los parámetros calculados en la iteración anterior, $\boldsymbol{\Theta}^g$. Como puede verse, hay que inicializar los parámetros al comienzo del proceso.

Ahora se puede proceder a obtener las expresiones para estimar el modelo, diferenciando la 5.16 con respecto a los parámetros e igualando a cero para obtener una forma cerrada para las expresiones de formación. Esto se realiza en tres pasos.

En primer lugar, se maximiza la función respecto de los coeficientes de mezcla. Introduciendo el multiplicador de Lagrange, con la limitación de la 5.14 y diferenciando respecto a un coeficiente de mezcla dada, $\alpha_i$, los coeficientes óptimos se consiguen en una forma cerrada como sigue:

$$\frac{d}{d\alpha_i}\left[Q + \lambda\left(\sum_{m=1}^{M}\alpha_m - 1\right)\right] = 0 \tag{5.17}$$

$$\alpha_i = \frac{1}{N}\sum_{n=1}^{N} p\left(i \,|\, x_n,\, \boldsymbol{\Theta}^g\right),\ i = 1, \cdots, M \tag{5.18}$$

donde la probabilidad de la componente *i-ésima* es dado por:

$$p(i \,|\, x_n,\, \boldsymbol{\Theta}^g) = \frac{\alpha_i p_i(x_n|\boldsymbol{\theta}_i^g)}{\sum_{m=1}^{M}\alpha_m p_m(x_n|\boldsymbol{\theta}_m^g)} \tag{5.19}$$

En segundo lugar, obtenemos las frecuencias de resonancia estimadas derivando respecto a $w_i$ e igualando a cero de la siguiente manera:

$$\frac{dQ}{dw_i} = -\sum_{n=1}^{N} p_i(x_n|\boldsymbol{\theta}_i)[\cos(2\pi x_n) - 2e^{\sigma_i}\cos(w_i) + e^{2\sigma_i}\cos(2\pi x_n)] p(i|x_n, \boldsymbol{\Theta}^g) = 0 \tag{5.20}$$

Esta es una función intratable debido a que el parámetro que se quiere estimar está incluido en el término $p_i(x|\theta_i)$ y no podemos obtener una forma cerrada como desea-

mos. Para hacer frente a este problema, simplificamos el desarrollo mediante el uso de los parámetros obtenidos, $\theta_i^g$,en la iteración anterior, $p_i(x|\theta_i^g)$. En aras de mantener la simplicidad, damos una expresión basada en $\cos(w_i)$ en lugar de $w_i$ porque es más fácilmente manejable y, por lo tanto, más adecuado para algunas consideraciones que haremos más adelante. Después de algunos arreglos, obtenemos la expresión cerrada siguiente:

$$\cos(w_i) = \frac{e^{\sigma_i} + e^{-\sigma_i}}{2} \frac{\sum_{n=1}^{N} \cos(2\pi x_n) p_i(x_n|\boldsymbol{\theta}_i^g) p(i|x_n, \boldsymbol{\Theta}^g)}{\sum_{n=1}^{N} p_i(x_n|\boldsymbol{\theta}_i^g) p(i|x_n, \boldsymbol{\Theta}^g)} \tag{5.21}$$

Una inspección intencionada de la 5.21 revela que hemos logrado una expresión coherente para el coseno, $\cos(w_i)$, de la frecuencia de resonancia de cada componente. Esta expresión representa un promedio ponderado de cosenos afectados por un factor que depende de los anchos de banda. Los pesos se dan en términos de la probabilidad marginal de la componente bajo estudio.

Los pesos pueden verse de forma generalizada como sigue:

$$k_i(x_n, \boldsymbol{\Theta}^g) = \frac{p_i(x_n|\boldsymbol{\theta}_i^g) p(i|x_n, \boldsymbol{\Theta}^g)}{\sum_{n=1}^{N} p_i(x_n|\boldsymbol{\theta}_i^g) p(i|x_n, \boldsymbol{\Theta}^g)} \tag{5.22}$$

siendo una expresión apropiada y coherente ya que cumplen con la restricción de normalización, es decir, $\sum_n k_i(x_n, \boldsymbol{\Theta}^g) = 1$. El factor que acompaña a la suma de cosenos ponderados puede ser identificado como el coseno hiperbólico de $\sigma_i$ y es necesario para garantizar la estabilidad de los modelos estimados como se verá en más adelante.

Por razones de conveniencia, definimos una función $R_i(k)$ como

$$R_i(k) = \sum_{n=1}^{N} \cos(2\pi x_n k) p_i(x_n|\boldsymbol{\theta}_i^g) p(i|x_n, \boldsymbol{\Theta}^g) \tag{5.23}$$

Más adelante haremos algunas consideraciones acerca de esta función. Ahora, continuamos nuestra argumentación mediante la simplificación de la 5.21 en una forma compacta.

$$\cos(w_i) = \cosh(\sigma_i) \frac{R_i(1)}{R_i(0)} \tag{5.24}$$

Y en tercer lugar, procedemos a maximizar la 5.16 con respecto a $\sigma_i$. Tomando la derivada de la 5.16 con respecto al $\sigma_i$ e igualando a cero, después de algunos arreglos, podemos obtener

$$\frac{dQ}{d\sigma_i} = -\sum_{n=1}^{N} p_i(x_n|\boldsymbol{\theta}_i)\{-2e^{\sigma_i}\cos(w_i)[\cos(2\pi x_n)- 2e^{\sigma_i}\cos(w_i) + e^{2\sigma_i}\cos(2\pi x_n)]+ 2e^{2\sigma_i}[\cos(4\pi x_n) - 2e^{\sigma_i}\cos(w_i)\cos(2\pi x_n)+ e^{2\sigma_i}]\}p(i|x_n, \boldsymbol{\Theta}^g) = 0 \tag{5.25}$$

Teniendo en cuenta los mismos comentarios apuntados sobre la 5.20 acerca del término $p_i(x_n|\boldsymbol{\theta}_i)$ y su sustitución por $p_i(x_n|\boldsymbol{\theta}_i^g)$, y observando que la 5.20 se encuentra entre los primeros corchetes del segundo miembro de la 5.25 con valor cero, podemos obtener una gran simplificación para el desarrollo la 5.25. A continuación sacamos el término $e^{2\sigma_i}$, al igual que en la 5.15, y usando la 5.23 llegamos a la siguiente expresión compacta

$$e^{2\sigma_i} = \frac{R_i^2(1) - R_i(0)R_i(2)}{R_i^2(0) - R_i^2(1)} \quad , \quad i = 1, \cdots, M \tag{5.26}$$

### 5.4.3. Algunas consideraciones teóricas

Hemos conseguido un estimador ML con fórmulas compactas, la 5.24 y la 5.26. Estas fórmulas se basan en la función $R_i(k)$. Comenzamos nuestras consideraciones al señalar que esta expresión se asemeja mucho a una función de autocorrelación. Para ver si $R_i(k)$ es una definición válida de una función de autocorrelación, debemos hacer dos observaciones.

En primer lugar, recordando que $x_n \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, podemos remplazar la función del coseno en la 5.23 por la función exponencial, $z = e^{j2\pi x}$. Y en segundo lugar, se observa que el producto $p_i(x_n|\boldsymbol{\theta}_i^g)p(i|x_n, \boldsymbol{\Theta}^g)$ representa algo parecido a una densidad de probabilidad marginal *i-ésima* de la mezcla. Por lo tanto, la 5.23 puede rescribirse como:

$$R_i(k) = \sum_{x_n \in \{-\frac{1}{2}, \frac{1}{2}\}} \exp(2\pi x_n k) p_i(x_n|\boldsymbol{\theta}_i^g)p(i|x_n, \boldsymbol{\Theta}^g) = \Psi_i(k) \tag{5.27}$$

la cual también pueden ser identificada como la función característica, $\Psi_i(k)$, asociada a la PDF "marginal" *i-ésima*. Teniendo en cuenta las equivalencias entre la autocorrelación y la función característica derivada de la identificación entre el PSD y PDF que hicimos en la sección anterior, concluimos que la 5.23 es una función de autocorrelación de hecho. Por todo lo anterior llamaremos a $R_i(k)$, función de autocorrelación "marginal" asociada a la componente *i-ésima*. Ahora, después de haber asumido que la 5.23 es una función de autocorrelación válida, procedemos a identificar los $\boldsymbol{a}_i = \{a_{1i}, a_{2i}\}$ en la 5.15. Así, $a_{1i} = e^{\sigma_i}\cos(w_i)$ y $a_{2i} = e^{2\sigma_i}$. Si usamos estas dos expresiones para los coeficientes de la 5.24 y la 5.26, después de algunos arreglos, se obtiene la expresión

para $\boldsymbol{a}_i$ como sigue

$$a_{1i} = \frac{R_i(0)R_i(1) - R_i(1)R_i(2)}{R_i^2(0) - R_i^2(1)} \tag{5.28}$$

$$a_{2i} = \frac{R_i^2(1) - R_i(0)R_i(2)}{R_i^2(0) - R_i^2(1)} \tag{5.29}$$

que son las mismas expresiones dadas por las ecuaciones de YW, que garantiza la estabilidad de los modelos estimados [22]. Este es un resultado importante porque demuestra la validez de nuestro algoritmo de entrenamiento para los componentes de la mezcla.

## 5.5. Experimentos y resultados

A continuación evaluaremos los modelos propuestos y sus métodos de estimación. Los siguientes experimentos tienen 3 objetivos sobre los modelos paramétricos basados en PSD y de mezclas de PSD, comprobar que los algoritmos de estimación desarrollados son robustos y convergen, que son capaces de aproximar de manera aceptable una distribución arbitraria, compitiendo con las gaussianas y que lo hacen satisfactoriamente tanto cuando usamos el orden correcto, como cuando se subdimensiona o sobredimensiona respecto el orden correcto.

En el primer experimento vamos a comprobar que el método propuesto ARMA(p,q) basado en el error de reconstrucciónde es capaz de aproximar funciones de densidad de forma apropiada. Para ello vamos a comparar el método propuesto de estimación de la PDF con modelo ARMA(p,q), PDF-ARMApq, con el método de estimación de la PDF con modelo AR propuesto en [21], PDF-AR, sobre el proceso de mezcla de Gaussianas utilizando en su trabajo y dado por la siguiente expresión,

$$p(y) = \frac{1-\varepsilon}{\sqrt{2\pi\sigma_1^2}} exp\left[-\frac{1}{2}\left(\frac{y^2}{\sigma_1^2}\right)\right] + \frac{\varepsilon}{\sqrt{2\pi\sigma_2^2}} exp\left[-\frac{1}{2}\left(\frac{y^2}{\sigma_2^2}\right)\right] \tag{5.30}$$

donde $\varepsilon$ = 0,1, $\sigma_1 = 1$ y $\sigma_2 = 10$ . Para un conjunto de datos del proceso $\{y\}$ con 1000 muestras promediando lasestimaciones sobre 50 realizaciones independients, usando un modelo ARMA(12,4) para la arpoximación de polos comunes y un modelo AR(12) para la de Kay, los resultados de la simulación se muestran en la figura 5.1. Queremos contrastar también las estimaciones paramétricas con las no paramétricas, por lo que hemos incluido un histograma, PDF-HIST, con 100 intervalos. Todas las PDF están normalizadas. Los datos originales has sido apropiadamente escalados para asegurar que están en el intervalo [-1/2,1/2].

Como podemos ver la PDF-ARMApq(12,4) muestra una estimación correcta respecto

al histograma y más suave que la PDF-AR(12), en la que se pueden observar algunos picos espurios. Este es un problema relacionado tanto con el método en si mismo como con la selección del orden del modelo. Parece que nuestro método es menos sensible al error producido por la mala selección del orden. Por otro lado vemos como el histograma tiene intervalos que no están definidos, con lo cual es muy difícil trabajar con él en sistemas de procesamiento complejos.

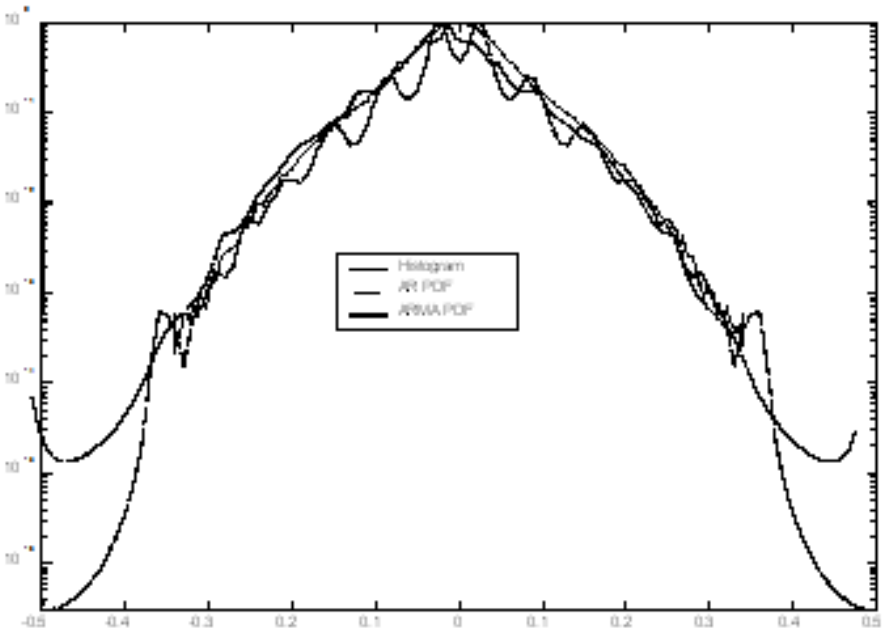



Figura 5.1: PDF-HIST (-), PDF-AR(12) (–) y PDF-ARMApq(12,4) (-.)

En los tres siguientes experimentos se va a comprobar la robustez del método propuesto de mezcla respecto del número de componentes de la mezcla. Para ello vamos a comparar el método propuesto de estimación de la PDF de mezclas de AR, PDF-MAR, el método de estimación de la PDF con modelo AR, PDF-AR, el histograma, PDF-HIST y un modelo de mezcla de Gaussianas, GMM, cuando la selección del orden de la PDF a estimar es el correcto, está subdimensionado o sobredimensionado. El orden de los dos modelos de mezclas será el mismo, PDF-MAR(p) y GMM(p) y el del método PDF-AR será el doble, PDF-AR(2p), ya que cada 2 polos definen una resonancia y por tanto equivalen a un modo y además así todos trabajan con el mísmo número de parámetros.

El proceso a estimar una mezcla de Gausianas con 3 componentes en la mezcla, tiene

500 muestras y viene dado por la siguiente expresión:

$$p(x) = 0{,}3 * N(-3, 1) + 0{,}2 * N(2, 2) + 0{,}5 * N(4, 0{,}3) \tag{5.31}$$

donde $N(\mu, \sigma^2)$ representa una densidad Gaussiana monomodal con media $\mu$ y varianza $\sigma^2$.

Primeramente estudiamos el comportamiento de los métodos de estimación cuando tienen información del proceso verdadero, eligiendo un número de componentes de la mezcla igual al original, en este caso 3. En la figura 5.2 se representa la estimación del proceso 5.31 con los modelos PDF-MAR(3), GMM(3) y PDF-AR(6). Podemos ver como nuestro modelo PDF-MAR(3) se ajusta bastante bien al verdadero, el GMM(3), mientras que el PDF-AR(6) presenta un sesgo en la localización de las resonancias y los anchos de banda.

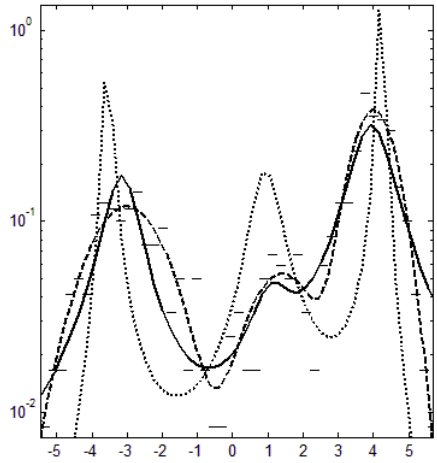



Figura 5.2: PDF-HIST (-), PDF-MAR(3) (- -), GMM(3) (–) y PDF-AR(6) (..).

Acontinuación queremos ver el comportamiento cuando se subdimensionan los modelos. En la figura 5.3 se muestran los resultados cuando se emplea un orden 2 para la mezcla, y utilizando los métodos de estimación PDF-MAR(2), GMM(2) y PDF-AR(4). Se comprueba que los tres métodos se aproximan bien a los 2 modos principales, aunque el PDF-AR(4) vuelve a mostrar un ligero sesgo y un ancho de banda algo estrecho. El método propuesto PDF-MAR(2) parece que tiende a modelar los anchos de banda también algo más pequeños que el GMM(2).

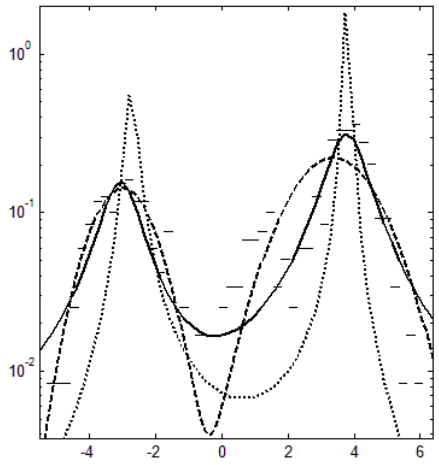



Figura 5.3: PDF-HIST (-), PDF-MAR(2) (- -), GMM(2) (–) y PDF-AR(4) (..).

Finalmente vamos a analizar lo que sucede cuando la estimación se sobredimensiona, utilizando un orden 10 para la mezcla. La figura 5.4 muestra los resultados de las estimaciones de PDF-MAR(10), GMM(10) y PDF-AR(20). El modelo AR genera muchos pico espurios, mientras que los modelos de mezcla ajustan mucho mejor su aproximación y tienden a situar las componentes de la mezcla extra alrededor de los modos pincipales de la PDF verdadera. Particularmente, el modelo de mezclas AR parece que tiene tendencia a resolver mejor las estructuras finas que la mezcla de gaussianas mientras evita espurios. Véase el espurio de la mezcla de gaussianas alrededor de x=-0.5 que no muestra la mezcla de AR.

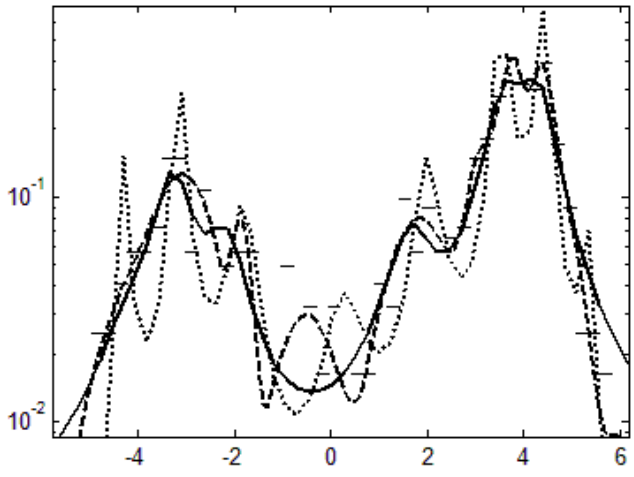



Figura 5.4: PDF-HIST (-), PDF-MAR(10) (- -), GMM(10) (–) y PDF-AR(20) (..).

Por útimo vamos a ver el comportamiento del método propuesto de estimación de la PDF con modelo ARMA(p,2), PDF-ARMAp sobre una muestra de datos con pdf desconcida [1] y la vamos a comparar al resto de los métodos propuestos y a la mezcla de Gaussianas. Podemos comprobar el buen comportamiento que ofrecen los modelos tanto el PDF-MAR(10) como el PDF-ARMAp(10,2) y que parece que no son peores que el GMM(10), cuando los datos no son Gaussianos, por lo que parece una buena opción para trabajar con datos reales.

## 5.6. Conclusiones

Hemos hecho aprotaciones novedosas a partir de un nuevo enfoque basado en el modelo de estimación de probabilidad función de densidad que explota el paralelismo existente entre el PDF y la función de densidad espectral de potencia. Primeramente se ha introducido una extensión y la formulación sobre los modelos AR incluyendo una parte MA fija, para hacerlos más robustos en general. Por otro lado proponemos un método que se inicia desde el uso de un proceso de nueva señal que permite la introducción de métodos de estimación espectral basada en la minimización de la señal de la reconstrucción. Las simulaciones muestran que nuestro método es menos sensible a la selección de órdenes y tiene picos de menos espurios que el método basado en AR. Hemos presentado la formulación matemática en forma de matriz. Esta formulación se puede extender a la incorporación de varias realizaciones en un intento de obtener una mejor estimación que con promedios.

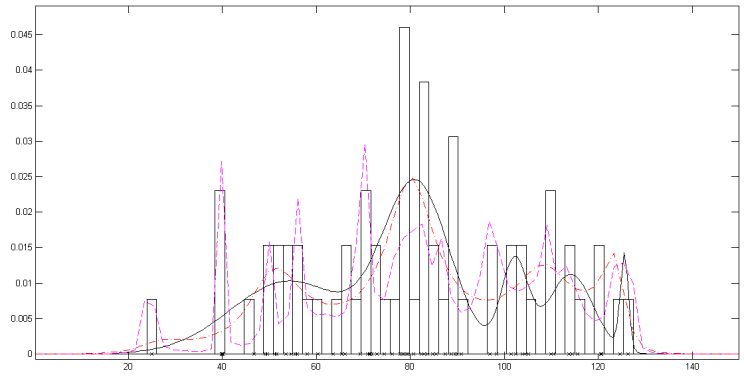



Figura 5.5: PDF-HIST (-), PDF-MAR(10) (- -), GMM(5) (–) y PDF-ARMAp(10,2) (..).

Hemos propuesto un nuevo PDF sobre la base de mezclas de PSD de tipo AR de orden mínimo, que tiene una visión local del problema y reducen la complejidad. Hemos desarrollado un estimador de máxima verosimilitud para sus parámetros desde el enfoque de EM enfoques y logrado fórmulas recurrentes de actualización. El comportamiento ha sido probado en experimentos diferentes en comparación con el modelo de mezcla de gaussiana, modelo AR y el histograma.

Se demuestra que no es absolutamente necesario para conocer con exactitud la orden de proceso debido a los resultados satisfactorios con órdenes moderadamente más pequeños y más grandes que los correctos. Este resultado sugiere que nuestro modelo de densidad es adecuado para adaptarse a diferentes distribuciones multimodales.

Podemos concluir que el modelo propuesto mezcla de ARMM es nuevo, atractivo y eficiente. Particularmente, nuestro ARMM se compara satisfactoriamente con el GMM, ampliamente utilizado.

## Bibliografía

[1] T. Fushiki, S. Horiuchi, and T. Tsuchiya, "A new computational approach to density estimation with semidefinite programming," in *Research Memorandum 898, The Institute of Statistical Mathematics*, 2003.

[2] J. Via, I. Santamaria, and M. Lázaro, "Blind restoration of binary signals using a line spectrum fitting approach," in *Signal Processing Conference, 2004 12th European*, pp. 461–464, IEEE, 2004.

[3] J. D. Hart and H. L. Gray, "On arma probability density estimation.," tech. rep., DTIC Document, 1981.

[4] J.-P. Carmichael, "The autoregressive method: a method of approximating and estimating positive functions," tech. rep., DTIC Document, 1976.

[5] X. Zhuang, Y. Huang, K. Palaniappan, and Y. Zhao, "Gaussian mixture density modeling, decomposition, and applications," *Image Processing, IEEE Transactions on*, vol. 5, no. 9, pp. 1293–1302, 1996.

[6] B. P. P. Rodríguez, *An introduction to complex function theory*. 1991.

[7] M. K. Varanasi and B. Aazhang, "Parametric generalized gaussian density estimation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404–1415, 1989.

[8] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[9] M. Rosenblatt *et al.*, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[10] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 1. Elsevier, 1992.

[11] H. G. Othmer, S. R. Dunbar, and W. Alt, "Models of dispersal in biological systems," *Journal of mathematical biology*, vol. 26, no. 3, pp. 263–298, 1988.

[12] J. Kmenta, *Elements of econometrics.* Macmillan New York, 1971.

[13] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *Signal Processing, IEEE Transactions on*, vol. 48, pp. 1687 –1694, June 2000.

[14] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical analysis of finite mixture distributions.* Wiley, 1985.

[15] A. Pagés-Zamora and M. A. Lagunas, "New approaches in non-linear signal processing: Estimation of the probability density function by spectral estimation methods," in *IEEE-ATHOS workshop on higher-order statistics*, 1995.

[16] M. B. Priestley, *Spectral analysis and time series.* Academic Press, 1981.

[17] A. Papoulis, *Probability, random variables, and stochastic processes.* McGraw-Hill, 1984.

[18] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep. TR-97-021, ICSI, 1997.

[19] J. L. Navarro-Mesa and P. J. Quintana-Morales, "An approach to ARMA modelling of probability density functions," in *Learning Workshop*, 2004.

[20] G. McLachlan and D. Peel, *Finite Mixture Models.* Wiley-Interscience, 1 ed., Oct. 2000.

[21] S. Kay, "Model-based probability density function estimation," *Signal Processing Letters, IEEE*, vol. 5, pp. 318 –320, Dec. 1998.

[22] S. M. Kay, *Modern spectral estimation: theory and application.* Prentice Hall, 1988.

[23] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Transactions on Instrumentation and Measurement*, vol. 51, no. 2, pp. 211–216, 2002.

Capítulo 6

# INTEGRACIÓN DEL MODELADO PARAMÉTRICO EN SISTEMA DE RAH BASADO EN MODELOS OCULTOS DE MARKOV

## 6.1. Introducción

El reconocimiento automático del habla es una potente herramienta de procesado de voz que permite la interacción hombre-máquina de manera cómoda y eficaz a través de la comunicación oral, habiendo algunos aspectos que todavía siguen siendo un desafío para su plena superación. El proceso de extracción de caraterísticas relevantes y su uso en aplicaciones son ejemlos claros. Destacamos, por ejemplo, la clasificación de voces patológicas. En cualquier caso, el proceso de reconocimiento necesita de un proceso de análisis de señal que extraiga sus características seguido de otro más general de decodificación, que suele contar con un bloque de modelado acústico en el que, de

hecho, se integra dicha extracción.

En este capítulo se abordará la integración de los trabajos de parametrización localizada y de modelado probabilístico de series temporales desarrollados en los capítulos anteriores, en algunas aplicaciones de reconocimiento automático del habla sobre voz patológica basada en modelos ocultos de Markov y se evaluará su comportamiento en relación con algunos de los métodos clásicos.

## 6.2. Complejidad de los sistemas de RAH

Las aplicaciones de reconocimiento automático del habla se basan en el proceso humano al que intentan representar, en el cual intervienen diferentes niveles de comprensión, además del que tiene que ver con la señal acústica y que interactúan entre si. Una forma de abordar el problema, que se advierte tan complejo, es dividirlo en niveles que podamos identificar como que intervienen en dicho proceso y a continuación tratar con cada uno de ellos de forma individual o combinándolos de forma progresiva. Cuanto más niveles utilizemos de forma combianda mejor serán los resultados. En el caso del proceso de la comunicación humana podemos identificar los siguientes niveles y sus funciones esenciales:

- **Nivel acústico**: Se analizan las características físicas de la señal vocal y se extraen sus parámetros con la información más relevante. Representa el nivel más básico con capacidad de clasificación y proporciona información sobre los objetos sonoros elementales. En sí ya constituye un herramienta de reconocimiento muy potente y es la base sobre la que se sustente el armazón general.
- **Nivel léxico**: Se incoporan hipótesis de palabras en función de las hipótesis de unidades subléxicas aportadas por el nivel acústico.
- **Nivel sintáctico**: Se generan hipótesis de frases aplicando reglas gramaticales basadas en el uso y normalización del lenguaje tomando como fuente la información porporcionada por el nivel léxico.
- **Nivel semántico**: Se trabaja con hipótesis sobre el significado de las frases reconocidas y comprobando la coherencia del mensaje.
- **Nivel pragmático**: Representa el nivel más alto de abstracción y en él se examina la relación entre los símbolos, los usuarios que los producen y la aplicación que se desarrolla.

Además de la señal que lleva el mensaje, el sistema puede manejar otro tipo de fuentes de conocimiento, que se relacionan con los niveles reflejados anteriormente. Según las características y especialización de cada sistema, dichas fuentes de conocimiento pueden estar simplificadas en mayor o menor grado o incluso estar prácticamente ausentes. Por ejemplo, en sistemas de reconocimiento de palabras aisladas, tan sólo las fuentes de conocimiento acústico y léxico son necesarias, mientras que en sistemas de reconocimiento

de palabras conectadas es necesaria además una tercera fuente, que normalmente reúne los conocimientos sintácticos, semánticos y pragmáticos.

Es por lo anterior que se debería introducir el concepto de modelo de lenguaje, que tiene que ver con el nivel léxico o sintáctico y es el que restringe las secuencias de unidades subléxicas a reconocer.

Aparte de lo dicho anteriormente, se pueden añadir otros problemas que nos encontramos en las aplicaciones de reconocimiento de voz y que dificultan su tarea, entre los que destacamos los siguientes: la **coarticulación**, tiene que ver con la influencia que tienen entre si los fonemas adyacentes y que dificultan la caracterización de ellos por verse afectadas; la **segmentación**, que influye en la identificación de las unidades léxicas y subléxicas al no haber una segmentación clara entre ellas que no ofrezca confusión; la **variabilidad** inter e intra-locutor, que presenta diferencias en la locución no solo entre locutores sino en el mismo locutor y que dificulta la caracterización de las unidades de reconocimiento; la **espontaneidad** de la locución donde la articulación del sonido se define mejor o peor; el **entorno** de uso que condiciona la eficacia cuando las condiciones son adversas; y el **tipo de habla** si el locutor presenta problemas en la comunicación oral asociado con alguna patología y se dificulta el proceso de reconocimiento.

## 6.3. Modelos ocultos de Markov en los sistemas de RAH

En esta tesis nos centramos en las tareas del nivel acústico, en el que encontramos el bloque de parametrización, el bloque de modelado de las unidades fonéticas y el bloque de clasificación. Los sistemas de reconocimiento automático del habla empezaron como sistemas deterministas, en el que una medida de distancia indicaba su desviación respecto a la señal que se esperaba. Eran sistemas expertos intuitivos que utilizaban diferentes conjuntos de parámetros, como formantes o LPC y distintas técnicas, como DTW o cuantificación vectorial (VQ), para tratar de compensar el alineamiento temporal de alguna manera. Fueron útiles y eficaces hasta que el modelado estocástico entró en escena y a partir de su desarrollo y se hizo popular, findamentalmente debido a sus mejores tasas de reconocimiento respecto de los métodos de clasificación precedentes. Este considera la señal como una observación de un proceso aleatorio y, por tanto, caracterizado por una función densidad de probabilidad. El modelo estocástico más popular actualmente es el Modelo Oculto de Markov (Hidden Markov Model, HMM) por su versatilidad, flexibilidad y sus buenos resultados [16, 13, 9, 14]. Muchos trabajos han empleado las redes neuronales en reconocimiento, tanto como modelos deterministas, así como estocásticos y también embebido en algún bloque de ambos modelos, sin que se hayan impuesto a los HMM [11, 10], al menos no de forma clara salvo en alguna aplicación muy específica.

Los HMM deben principalmente su potencia a que las observaciones se pueden caracterizar mediante funciones de probabilidad asociadas a un estado, en el que se encuentra, que permanece oculto al observador, puesto que desconoce el proceso, siendo éste, a su vez, caracterizado en su dinámica temporal según la estructura de estados del modelo. O sea, son capaces de capturar simultáneamente la variabilidad temporal y la naturaleza

estadística de los procesos que modelan. Las estructuras de estado típicas en reconocimiento suelen ser de izquierda a derecha, por similitud con el proceso de producción de voz y de primer orden, por simplicidad. Los HMM se definen principalmente por la terna de probabilidades iniciales de estados, de transición entre estados y de observación en cada estado. Las probabilidades de observación pueden ser discretas o continuas según se defina el conjunto de observaciones. En el caso que nos ocupa de voz, una definición discreta implica que las observaciones se tienen que cuantificar en una librería del estilo VQ, con la consiguiente pérdida en la caracterización probabilística, por lo que la mayoría de los estudios, sean cualesquiera de las representaciones de la señal de voz que usen, siendo continuas, emplean funciones de densidad continua para obtener los mejores resultados. Todos estos conceptos se pueden revisar de manera más profunda y formal en [16].

Si bien las técnicas de RAH han progresado enormemente y alcanzado las tasas de reconocimiento más altas con los sistemas basados en modelos ocultos de Markov (HMM), mayores al 95 %, lo que no ha alcanzado tanta madurez es la aplicación de estas técnicas para el caso particular de personas con problemas en la comunicación oral.

De cualquier forma, diferentes aproximaciones se han propuesto en el problema de reconocimiento de voz con patología, utilizando tanto HMM discretos como continuos sobre distintos conjuntos de parámetros que caracterizan a las señales de voz. Mayoritariamente, los mejores resultados se han conseguido usando como base los parámetros MFCC (mel frequency cepstral coefficients), debido al realce efectivo de la información fonética que realiza y al alto grado de decorrelación que alcanza.

En esta tesis vamos a avanzar hacia la solución del problema de reconocimineto de voz por dos vías. Por un lado, la información fonética se va a representar de manera detallada tomándola en intervalos de pitch o menores, como pueden ser las fases cerradas y abiertas de las cuerdas vocales en los tramos sonoros [8]. Las señales en estos intervalos llevan información muy fina y localiazada en el tiempo relativa al proceso de producción de voz. En consecuencia, reflejan bien la existencia o no de problemas en dicho proceso. Esto nos lleva a plantear la utilización de una parametrización específica para cada intervalo, síncrona con el periodo de pitch, dentro del bloque de extracción de características de nuestro sistema. Y por otro lado, atendiendo a caracterización probabilística del modelado del habla, en esta tesis consideraremos la alternativa de los modelos paramétricos lineales frente a las mezclas de Gaussianas como funciones de densidad de probabilidad por sus propiedades de aproximación [6]. En los apartados siguientes desarrollaremos estos aspectos.

## 6.4. Introducción de la parametrización basada en modelos de fases con polos comunes en el sistema RAH

El modelado de la señal voz es el primer paso que tenemos que dar ya que nos proporciona una representación variante en el tiempo de la señal que nos ayudará a

caracterizar de manera correcta las pronunciaciones.

Proponemos que la información fonética sea representada de manera eficiente teniendo en cuenta la sonoridad y, dentro de la voz Sonora, distinguimos la fase cerrada y abierta de las cuerdas vocales. La señal en estas fases lleva información relativa al proceso de producción, reflejando la existencia o no de deficiencias en dicho proceso. La idea que proponemos es utilizar una parametrización específica para cada fase dentro del bloque de extracción de características de nuestro sistema, siendo esto una nueva forma de estimar las características de la voz. Esto es, lo clásico es tomar la señal por tramas y hacer un análisis de su conjunto. Nuestra aproximación trabajamos igualmente por tramas pero se hacen dos tratamientos en cada una. En el primero, se determinan los instantes de cierre glótico para delimitar los periodos. Y en el segundo, se procede a la extracción de características comunes a los periodos comprendidos en la trama.

Trabajamos con los parámetros que se han deducido de los procesos de estimación vistos en capítulos anteriores, que pueden ser de dos tipos, modelado de polos comunes a varios periodos consecutivos y ceros particulares (CPPZ), modelado de polos y ceros comunes a varios periodos consecutivos (CPCZ) y modelado dichos parámetros transformados en frecuencia (FWCPPZ) y (FWCPCZ). De estos modelos ya vimos su fiabilidad y consistencia analizando la capacidad de seguimiento fonético.

Para el problema de reconocimiento de voz con y sin patologías, una buena parametrización a priori sería la de voz sonora sobre periodos consecutivos con polos y ceros comunes debido a que muestran una buena capacidad para la caracterización fonética [7] además de permitir la inclusión de los vectores en un sistema de reconocimiento clásico. Para hacer esto tenemos varias opciones. Fundamentalmente, vinculadas a qué tipo de información tomar y de qué fase. Así, por ejemplo, podemos tomar información de las dos fases o de una dada y, en cualquier caso, tomar información de la parte AR o MA. Podríamos tomar la fase cerrada por considerarla estable al ser la respuesta del tracto vocal y de ella extraer información sobre patologías vinculadas al tracto. Por otro lado, podemos tomar la fase abierta que estaría acoplada a un órgano como la tráquea y puede reflejar otras patologías vinculadas a dicho órgano. La parametrización de la voz sorda se realiza de forma eficiente con el modelo AR clásico.

Para acometer la parametrización propuesta se necesita un detector de sonoridad así como un detector de instantes de cierre glótico (ICG), siendo la efectividad de éstos responsables directos de la superioridad del modelado propuesto sobre los clásicos, AR y ARMA. También habría que decidir que tanto por ciento del periodo corresponde a fase cerrada y cuanto a fase abierta.

El detector de sonoridad propuesto está basado en el detector de pitch de [15] que emplean la función de autocorrelación ponderada con la inversa de la función de diferencias de amplitud en magnitud normalizada (AMDF) y del que dice que es más robusto que los convencionales en entornos ruidosos. Esta última característica pudiera ser interesante para el uso con determinadas patologías. Un problema del detector de pitch propuesto es que no da indicación explícita de la sonoridad. Es por ello que fue necesario incluir una serie de umbrales que garantizara la correcta detección en la gran mayoría de los casos. Solo se tomaría el valor de pitch como periodicidad de una trama sonora

si la autocorrelación normalizada (AUTOC) superara un umbral (p.e. 0.7) al mismo tiempo que la inversa normalizada de la AMDF fuera inferior a otro umbral (p.e. 0.8). El primer umbral, suficientemente elevado, sirve para garantizar que solo tramas 'muy' periódicas (en el interior de un fonema sonoro) pueden ser detectadas como sonoras. El segundo umbral, suficientemente grande, sirve para garantizar que solo tramas de amplitud medianamente grande pueden ser detectadas como sonoras (las tramas sordas suelen ser de baja amplitud). Las tramas de transición sonoro-sordo, sordo-sonoro y sonoro-sonoro, presentan una autocorrelación periódica cuyo primer pico puede no ser suficientemente grande para superar el primer umbral, pero la componente sonora le confiere una amplitud moderadamente grande a la trama, por lo que se detecta disminuyendo el umbral de la autocorrelación siempre y cuando la inversa de la AMDF normalizada fuera suficientemente pequeña, o sea disminuyendo dicho umbral.

El detector de ICG se implementa a partir del pitch obtenido en las tramas sonoras. Se busca la amplitud máxima de la trama (se supone que el cierre glótico está en las cercanías de los máximos absolutos) y con el valor de pitch se rastrean los máximos cuasiperiódicos a un lado y a otro de aquel.

Uno de los mejores métodos de parametrización para reconocimiento es el basado en los MFCC [9], debido al realce efectivo de la información fonética que realiza y al alto grado de decorrelación que alcanza. Otro buen método de parametrización, que a su vez es directamente comparable a los que proponemos en este trabajo, es el LPC clásico. Por estos motivos se tomarán como parametrizaciones de referencia para evaluar las nuestras.

## 6.5. Introducción de la PDF basadas en modelos paramétricos de la PSD tipo ARMA y AR en el sistema RAH

En el modelado HMM, la función de densidad de probabilidad de observaciones puede ser paramétrica o no, según se fije el modelo que ha generado a las observaciones o no. Una PDF paramétrica útil para alguno de los principales conjuntos de parámetros de voz es la basada en una gaussiana, por su tratabilidad matemática y sus buenas prestaciones. Sin embargo, debido a la diversidad de aspectos de las pronunciaciones, es usual considerar a dichas densidades como multimodales y una buena aproximación a ellas, comúnmente usada, es la mezcla de funciones [16]. Entre todas las mezclas, la mezcla de Gaussianas es la más popular, por sus expresiones, buena formulación, buenos procedimientos de entrenamiento, basados en el algoritmo de EM y eficacia [9]. También por sus resultados, de forma general, aunque su comportamiento sea subóptimo, esto es, no sea el esperado cuando los modos sean muy amplios o muy estrechos.

Una forma alternativa de representar las densidades de observación multimodales se ha obtenido al emplear las muestras de voz como observaciones y por tanto caracterizar cada estado como un proceso AR conducido por un ruido gaussiano. Diferentes trabajos han proporcionando expresiones compactas para el entrenamiento a partir de

las características de la voz y basadas en la estimación de Baum-Welch [5, 17]. A pesar de lo desarrollado del aparato matemático asociado y la elegancia de la formualción planteada por los autores, los resultados de clasificación no han resultado competitivos por lo que pronto cayeron en desuso. En este punto proponemos incorporar los modelos paramétricos lineales desarrollados en los capítulos anteriores, ARMA(p,2), ARMA(p,q) y mezclas de AR(2), como caracterización de las densidades de probabilidad en cada estado, debido a las consideraciones ya expresadas y a los resultados vistos en el desarrollo de los métodos de análisis. Adaptamos las expresiones matemáticas de estimación desarrolladas para dichos modelos y las empleamos en la obtención de las probabilidades de observación [4], basándonos en los trabajos de Baum-Welch (p.e., [12]).

El entrenamiento estaría dirigido por las ecuaciones clásicas del modelado de Markov conjuntamente con las ecuaciones de estimación adaptadas de los parámetros de las PDF propuestas en los capítulos anteriores y se compararían con los sistemas que usan mezclas de gaussianas.

## 6.6. Experimentos y resultados

El objetivo fundamental de nuestros experimentos ha sido comprobar la competitividad de las aportaciones propuestas del modelado paramétrico lineal en los diferentes bloques del nivel acústico de un reconocedor automático del habla. Primero veremos el efecto de utilizar los modelos paramétricos localizados en el bloque paramétrico y a continuación analizaremos el comportamiento al introducir los modelos paramétricos de las PDF en el bloque de modelado probabilístico del habla.

### 6.6.1. Experimientos de Reconocimiento de palabra sobre la Base de Datos HADECO

En este experimento se lleva a cabo la implementación de un reconocedor de palabras aisladas sobre la base de datos HADECO, realizado por el Grupo de Ingeniería Acústica de la ULPGC [2] y que consta de 57 palabras pertenecientes a un código fonológico inducido. La base de datos contiene un conjunto de datos de voces patológicas y no patológicas con 3543 registros, de las cuales 1077 se corresponden con pacientes sanos y 2466 a pacientes con patologías. Las características de la base de datos se pueden ver en el anexo A. El reconocedor del habla es semiespontáneo e independiente de locutor, siendo el tamaño del vocabulario pequeño. El reconocedor está basado en HMM con mezclas de gaussianas.

El objetivo del sistema es reconocer si una palabra ha sido pronunciada por un paciente con patología o sin ella. Así pues, necesitamos para cada palabra del código un par de HMM, uno para caracterizar las pronunciaciones con patologías y otro sin patologías.

Cada HMM tiene como características principales la topología, el número de estados y el número de gaussianas por estado. La topología es de izquierda a derecha, sin saltos de más de un estado.

Para elegir el número de estados tomamos como referencia el número de fonemas de

cada modelo y sobre este valor estudiamos un margen plausible que queda optimizado en función de la tasa de clasificación. El número de gaussianas tiene que ver con las distintas características que se podrían observar en diferentes pronunciaciones de una misma palabra, así que su elección final es también empírica [9]. El entrenamiento de los HMM se hace mediante el algoritmo de Baum-Welch.

En el proceso de clasificación hemos partido del conocimiento de la palabra bajo prueba. De esta forma estamos ante un test de hipótesis binario, buena o mala pronunciación, en el que cada clase tiene su propio modelo. Para los test de hipótesis sabemos en cada momento qué palabra se está pronunciando y el grado de bondad (patología o no) de la pronunciación será la incógnita. Así, por ejemplo, si estamos con la palabra 'palmera' tomaremos una de las pronunciaciones y trataremos de ver si la pronunciación corresponde a una voz con o sin patología.

La forma de abordar este problema consiste en tomar cada pronunciación, calcular la probabilidad de que los HMM correspondientes con y sin patología genere esa pronunciación y comparar las dos probabilidades. El modelo que dé lugar a la mayor probabilidad indicará si la pronunciación ha sido buena o mala.

El análisis se ha hecho por tramas de 30 mseg con un desplazamiento entre tramas de 10 mseg. A cada trama se le aplica una ventana de Hamming. Antes de cualquier procesado se ha hecho un preénfasis con un coeficiente de 0'95.

Se han probado los modelos de parametrización basado en polos comunes, tanto con ceros particulares, CPPZ, como con ceros comunes, CPCZ. Igualmente se han utilizado las versiones transformadas en frecuencia de ambos, (W)CPPZ y (W)CPCZ, respectivamente. Los resultados obtenidos en ambos casos son prácticamente iguales si bien el de polos y ceros comunes es ligeramente mejor. Por este motivo se exponen los resultados de las técnicas síncronas con el periodo cuando éstos están basados en características de polos y ceros comunes.

Las parametrizaciones finalmente utilizadas fueron por tanto CPCZ en voz sonora, sobre periodos consecutivos con polos y ceros comunes y LPC en voz sorda. En la voz sonora, la fase cerrada de cada periodo empieza en el máximo y acaba al 40 % de dicho periodo mientras que la fase abierta comienza al terminar la fase cerrada y acaba a un 12 % del final del periodo. Como periodos consecutivos en el modelado CPCZ se emplearon todos los de la trama pues se asumen características acústicas comunes. En todos los casos, los parámetros empleados fueron los de la parte AR. El modelo se aplicó sobre la fase cerrada, CPCZfc, sobre la fase abierta, CPCZfa, utilizando los parámetros de la fase cerrada y de la fase abeierta, CPCZac y sobre el periodo completo, CPCZpc.

Se crearon entonces 57 HMM para las palabras del código con patologías y 57 HMM para las sin patologías dando lugar a un total de 114 modelos. Para cada modelo se utilizaron el 75 % de los datos de cada palabra para entrenamiento y el 25 % para test. Los experimentos se repitieron 50 veces eligiendo aleatoriamente los datos de los conjunto en cada iteración y los resultados se promediaron. Los resultados de la tabla 6.1 muestran la tasa de aciertos de cada parametrización para los registros con patologías (CP) y sin patologías (SP), obtenidas con el óptimo de cada modelo (número de estados, número de gaussianas por estado, etc.). Los modelos propuestos se compararon con los

| Parámetros | % CP | % SP | % Total |
|---|---|---|---|
| MFCC | 96.58 | 94.05 | 95.85 |
| LPC | 93.80 | 93.50 | 93.71 |
| CPCZpc | 95.79 | 96.62 | 96.03 |
| CPCZfc | 95.59 | 94.70 | 95.33 |
| CPCZfa | 95.00 | 96.30 | 95.38 |
| CPCZac | 94.66 | 95.50 | 94.90 |
| WCPCZfc | 97.13 | 96.49 | 96.94 |

Cuadro 6.1: Resultados de la clasificación para voz con patologías (CP), sin patología (SP) y total (Total) en función del tipo de parámetros para la base de datos HADECO.

clásicos MFCC y LPC. Los resultados del diseño preliminar de las condiciones en que se realizaron estos experimentos se muestra en el anexo B donde se ha entrenado y reconocido con toda la base de datos.

Como se puede observar, los resultados obtenidos con las nuevas parametrizaciones son del mismo orden que con MFCC y ligeramente superiores a LPC. Varias puntualizaciones han de hacerse. En primer lugar, con las nuevas parametrizaciones obtenemos mejores resultados que con la LPC clásica. Por otro lado, las parametrizaciones transformadas ofrecen un comportamiento ligeramente superior a sus homólogas. Además, no parece que hacer una distinción de la fase en cada periodo dé mejor resultado que utilizar el periodo completo (CPCZpc). No obstante, las diferencias de resultados no son especialmente significativas. Sí es de destacar que la fase abierta (CPCZfa) parece ofrecer mejores resultados que la cerrada. Hemos de llamar la atención sobre CPCZpc frente a MFCC pues, en general, parece aceptado que una transformación cepstral es 'casi' una garantía de éxito frente a otras formas de parametrizar. De nuestros resultados esto no queda tan claro.

Como conclusión de los resultados obtenidos podemos decir que los modelos de fases con polos comunes son competitivos frente a los métodos clásicos en determinadas cirscunstancias. Dentro de ellos el de modelado transformado de fases con polos comunes se apunta como la mejor opción, quizás por tener información más detallada en las bajas frecuencias. Por otro lado, se puede destacar que trabajar con la fase cerrada o abierta por separado o conjunta no aporta mucha ventaja respecto a trabajar a nivel de intervalos de ICG. Esta forma de trabajar difiere de cualquier manera respecto de los métodos más usados que lo hacen asíncronamente.

### 6.6.2. Experimientos de Reconocimiento de palabra sobre la Base de Datos KAY ELEMETRIC

En este experimento se lleva a cabo la implementación de un reconocedor de patología sobre la base de datos KAY ELEMETRIC [3], del que sacamos un conjunto de registros de la /ah/ sostenida, 159 pertenecientes a pacientes sanos y 654 que corresponden a

pacientes con diversas patologías, con una duración de 1 segundo cada registro. Las características de la base de datos se pueden ver en el anexo A. El reconocedor del habla que diseñamos es semiespontáneo e independiente de locutor, siendo el tamaño del vocabulario de un solo fonema.

En estos experimentos tomamos una sola características, la entropía de permutación (Permutation Entropy, PE). Esta entropía se presenta como una medida útil para analizar la complejidad de la señal de voz . Las medidas de entropía, las dimensiones fractales y los exponentes de Lyapunov han sido tradicionalmente los parámetros de complejidad más ampliamente utilizados. La mayoría de los métodos clásicos ignoran el orden de los valores de una determinada serie temporal. En el año 2002 Bandt y Pompe plantearon la Entropía de Permutación [1], que combinaba los conceptos de entropía y de dinámica simbólica con el objetivo de crear una nueva medida de complejidad. La PE además de ser una caracetrística robusta y sencilla, se presenta como una medida adecuada para el estudio de series temporales caóticas y tienen en cuenta el orden temporal de los valores de la señal.

Planteamos, por tanto, el uso de esta medida para estudiar los cambios en las dinámicas de la voz y buscar de esta manera mejoras en la detección de patologías. De esta forma, aún con la ausencia de caracterización del sistema de producción de voz por parte de la PE al estilo del modelo paramétrico, lo cierto es que hemos visto experimentalmente que estas medidas permiten seguir muy bien los cambios en los patrones de la señal de voz y, por tanto, nos pueden ser útiles en la detección de patologías. Para una visión más detallada de esta característica consultar el anexo C.

La parametrización de PE se hace sobre tramas de la señal de voz de 30 mseg desplazadas cada 10 mseg, lo que daría un conjunto de parámetros igual al número de tramas y de dimensión 1.

El reconocedor que utilizamos está basado en una PDF para cada clase. Sería el caso más sencillo de HMM con un solo estado. La parametrización es de una dimensión basada en una medida de la entropía de permutación. El objetivo que nos planteamos es doble. Por un lado, ver la potencialidad de las aproximaciones basadas en modelos paramétricos de PSD cuando se utilizan como PDF en un sistema de reconocimiento de voz. Y por otro, ver las prestaciones de PE como característica relevante. Para ello hemos diseño este experimento con características especiales y es que la dimensionalidad de la parametrización sea 1. De aquí nace la importancia de tratar con una parametrización unidimensional potente que sirvs para nuestros intereses.

Para las PDF se han utlizado las aproximaciones basadas en modelos ARMA(p,2), ARMA(p,q) y mezclas de AR(2) vistas en el capítulo 5. La adaptación de los métodos de estimación alli desarrollados es directa en este caso, ya que estamos considerando un solo estado. Así, el modelo ARMA(p,2) se estimaría con las ecuaciones de Yule_Walker extendido 5.7 sobre los parámetros de la parte AR, ya que los de la parte MA ya vienen dados en la definición del modelo. Por otro lado, el modelo ARMA(p,q) se obtendría como solución de mínimos cuadrados sobre el error de reconstrucción del modelo no lineal creado a partir de los datos 5.12. Finalmente, el modelo de mezclas de AR(2) se estimaría a través del algoritmo de estimación basado en EM desarrollado 5.28 y 5.29.

En el entrenamiento de los modelos se utilizaron el 75 % de los datos de cada tipo de voz, y el 25 % se dejaron para test. Los experimentos se realizaron mediante una validación cruzada iterando 50 veces. En cada iteración se eligieron aleatoriamente los datos de los conjuntos, y los resultados de clasificación se promediaron. Los resultados de la tabla 6.2 muestran la tasa de aciertos de cada aproximación PDF para los registros con patología (CP) y sin patologías (SP). Los modelos propuestos se compararon con el clásico de mezcla de gaussianas (GMM) y con la aproximación de modelado AR(p) propuesta en [18].

Para analizar los resultados debemos tener en cuenta que las mezclas de Gaussianas y de AR(2) se nombran como GMM(x) y MAR(x) respectivamente, donde la x indica el número de componentes de la mezcla y por tanto pueden representar los modos de la densidad. Por otra parte los modelos ARMA(p,2), ARMA(p,q) y AR(p) indican entre paréntesis su orden y de alguna manera el parámetro p, que indica el número de máximos espectrales de la PSD, representa los modos de la PDF. Por todo ello podemos decir que los modelos GMM(n) y MAR(n) son equivalentes a los modelos ARMA(2n,2), ARMA(2n,q) y AR(2n), por representar el mismo número de modos de la PDF.

Examinando los resultados vemos que el modelado de PDF con modelos de PSD paramétricos es competitivo con la mezcla de gaussianas en todos los casos, siendo el modelo ARMA(p,2) el que mejores tasas de resonocimiento alcanza, lo que nos informa positivamente sobre su capacidad de aproximar distribuciones complejas unidimensionales. Estas presentan mejores prestaciones que el modelo AR(p) utilizado en [18]. Por otro lado, comprobamos que la mezcla de modelos AR(2) se comportan de manera similar.

El comportamiento alcanzado por los modelos ARMA parece que tienen más capacidad de aproximación que el AR presentado, quizás por que el método de estimación tiene menos variabilidad. Por otro lado, una comparación de ARMA con GMM nos hace ver que pueden haber situaciones en las que la aproximación paramétrica de variables aleatorias vale la pena explorarla.

Para el estudio de los resultados también deberíamos fijarnos en la sensibilidad, tasa de aciertos con patologías y en la especificidad, tasa de aciertos sin patologías. Siendo ligeramente superiores las prestaciones obtenidas por nuestras aproximaciones con respecto al modelado clásico y al AR.

## 6.7. Conclusiones

Desde el punto de vista de la extracción de características en todo momento subyace la hipótesis de que si hay diferencias perceptibles entre voz (particularmente la de tipo sonoro) con patología y sin ella podremos extraer estas diferencias. Nos hemos aproximado a esto con un planteamiento basado en distinguir entre fases abierta y cerrada.

Otra cosa es cómo hacer esa extracción. Desde el punto de la caracterización estadística de las variables, las formas que hemos estudiado partían de una experiencia previa sobre estimación óptima de los coeficientes ARMA según un criterio de error cuadrático medio mínimo. Faltaba estudiar si este 'óptimo' también se manifestaba en un sistema de clasificación-reconocimiento de voz. Un test de hipótesis binario nos ha servido de

| Modelos | % CP | % SP | % Total |
|---|---|---|---|
| AR(2) | 77.68 ± 2.89 | 86 ± 6.07 | 79,31±2.36 |
| AR(6) | 70.24 ± 4.42 | 87.49 ± 7.28 | 73,61±2.62 |
| AR(10) | 80.6 ± 3.76 | 69.9487 ± 8.2 | 78,52±2.85 |
| AR(16) | 82.34 ± 6.55 | 71.54 ± 16.7 | **80,23±1.45** |
| AR(20) | 79.43 ± 7.47 | 76.56 ± 19.79 | 78,87±2.22 |
| ARMA(2,1) | 94.47±6.57 | 32.1±30.36 | **82,27±2.35** |
| ARMA(6,5) | 75.02±6.89 | 81.18±17.56 | 76,22±3.96 |
| ARMA(10,9) | 74.86±12.13 | 80.4±17.69 | 75,94±2.89 |
| ARMA(16,15) | 74.84±8.12 | 79.9±18.81 | 75,83±7.69 |
| ARMA(20,19) | 76.3±9.39 | 80.67±14.17 | 77,15±8.05 |
| ARMA(2,2) | 78.16±2.46 | 85.64±5.04 | 79,62±2.58 |
| ARMA(6,2) | 76.27±3.85 | 88.15±4.36 | 78,59 ±1.77 |
| ARMA(10,2) | 76.86±3.70 | 87.02±6.11 | 78,85±2.16 |
| ARMA(16,2) | 79.16±3.22 | 84.36±5.54 | 80,18±3.09 |
| ARMA(20,2) | 80.28±2.98 | 86.21±5.02 | **81,44±2.30** |
| GMM(1) | 77.68±4.19 | 88±5.54 | 79,70±2.48 |
| GMM(3) | 78.39±3.7 | 86±4.61 | **79,88±2.48** |
| GMM(5) | 76.78±4.05 | 87.23±4.76 | 78,82±2.99 |
| GMM(8) | 77.58±3.93 | 86.82±5.36 | 79,39±3.64 |
| GMM(10) | 77.96±3.4 | 87.33±4.93 | 79,79±2.28 |
| MAR(1) | 80.08±2.51 | 74.77±7.74 | 79,04±2.05 |
| MAR(3) | 79.18±3.02 | 76.72±8.09 | 78,70±3.15 |
| MAR(5) | 79.44±3.33 | 77.9±6.75 | 79,14±6.55 |
| MAR(8) | 79.75±2.98 | 75.13±7.26 | 78,85±4.84 |
| MAR(10) | 80.01 ± 3.06 | 76.1±6.16 | **79,25±5.04** |

Cuadro 6.2: Resultados de la clasificación para voz con patologías (CP), sin patología (SP) y total (Total) en función del tipo de parámetros para la base de datos KAY ELEMETRIC.

marco sobre el que trabajar.

Una vez desarrollado ese marco faltaba volver sobre la idea de distinguir las fases de voz sonora. No podemos extraer resultados concluyentes sobre qué fase es mejor para clasificar. Sí podemos decir que se abre una vía que consideramos prometedora.

En cuanto a la parametrización basada en métodos lineales que proponemos, si bien los resultados no son marcadamente mejores que con las clásicas, lo cierto que es que los resultados son muy prometedores llegando a conseguir mejorar MFCC para el caso de tomar el periodo completo sin distinguir fases.

Un caso especial de parametrización es el de la entropía de permutación (PE). Su uso muestra resultados muy prometedores, que conjuntados los las PDF que proponemos resultan aún más prometedores.

Desde el punto de vista del modelado probabilistico la idea era ver si los modelos propuestos eran capaces de alcanzar el comportamiento del clásico de mezcla de Gaussiana y hemos comprobado que lo son. Así, hemos encontrado que los modelos basados en modelos ARMA ofrecen un comportamiento comparable al GMM y ligeramente superiores al AR.

## Bibliografía

[1] C. Bandt and B. Pompe, "Permutation entropy: a natural complexity measure for time series," *Physical review letters*, vol. 88, no. 17, p. 174102, 2002.

[2] N. M. J. Espinosa Yánez, J., "Herramienta de ayuda para deficiencias en la comunicación oral (hadeco)," in *Proyecto Fin de Carrera*, 2001.

[3] "Massachusetts eye and ear infirmary, voice disorders database, version 1.03. CDROM," 1994.

[4] I. Pérez-Castellano, P. J. Quintana-Morales, and J. L. Navarro-Mesa, "Clasificacion de voz patologica mediante modelado ARMA comun a varios periodos," in *URSI*, 2004.

[5] A. Poritz, "Linear predictive hidden markov models and the speech signal," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, pp. 1291 – 1294, May 1982.

[6] J.-F. Bercher and C. Vignat, "Estimating the entropy of a signal with applications," *Signal Processing, IEEE Transactions on*, vol. 48, pp. 1687 –1694, June 2000.

[7] P. J. Quintana-Morales and J. L. Navarro-Mesa, "An approach to common acoustical pole and zero modeling of consecutive periods of voiced speech," in *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, ISCA, 2003.

[8] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 313 –327, July 1998.

[9] C. Becchetti and L. P. Ricotti, *Speech recognition: theory and C++ implementation.* Wiley, June 1999.

[10] N. Morgan and H. Bourlard, "Continuous speech recognition," *Signal Processing Magazine, IEEE*, vol. 12, pp. 24 –42, May 1995.

[11] R. P. Lippmann, "Review of neural networks for speech recognition," *Neural Computation*, vol. 1, no. 1, pp. 1–38, 1989.

[12] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep. TR-97-021, ICSI, 1997.

[13] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437 –1462, Sept. 1997.

[14] D. O'Shaughnessy, "Improving analysis techniques for automatic speech recognition," in *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, vol. 3, pp. III–65 – III–68 vol.3, Aug. 2002.

[15] H. Kobayashi and T. Shimamura, "A weighted autocorrelation method for pitch extraction of noisy speech," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, pp. 1307 –1310 vol.3, 2000.

[16] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* PTR Prentice Hall, 1993.

[17] B.-H. Juang and L. Rabiner, "Mixture autoregressive hidden markov models for speech signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, pp. 1404 – 1413, Dec. 1985.

[18] S. Kay, "Model-based probability density function estimation," *Signal Processing Letters, IEEE*, vol. 5, pp. 318 –320, Dec. 1998.

Capítulo 7

# CONCLUSIONES Y LÍNEAS FUTURAS

## 7.1. Introducción

En este capítulo se expondrán las conclusiones del trabajo y las principales contribuciones. Así mismo se señalarán las líneas que deja abiertas el trabajo y que pueden servir en el futuro para mejorarlo.

## 7.2. Conclusiones y principales contribuciones

Este trabajo de investigación se ha planteado sobre la contribución que tiene la introducción del modelado paramétrico lineal, tendiendo a una visión unificada, en los sistemas de reconocimiento automático del habla, y concretamente en los bloques de extracción de características y de modelado probabilístico del nivel acústico.

Desde el punto de vista unificado, se ha particularizado en el modelado de series temporales sobre las fases glóticas de señal de voz, por un lado, y al modelado de funciones de densidad de probabilidad desde la definición de la densidad espectral de

potencia de los modelos paramétricos lineales, por el otro.

Para el bloque de extracción de características del RAH, se definió un marco de actuación localizado en las fases glóticas de la señal de voz y se desarrolló una formulación compacta de intervalos estacionarios sobre tramas adyacentes. Esta es una contribución novedosa, en tanto en cuanto engloba la caracterización común y particular de diferentes tramas, pertenecientes a un mismo intervalo estacionario y que están directamente relacionadas con el proceso físico de producción de voz sonora.

Con dicha formulación se establecieron diferentes modelos relacionados con los modos de actuación. Así, tenemos el caso particular clásico de trabajar con un solo periodo, o lo más avanzados de hacerlo con las características de resonancia común a dichas tramas y las características de antiresonancia pudiendo ser común, por simplicidad, o particular, para una caracterización física completa.

A partir de ambos modelos, se dedujeron los correspondientes algoritmos de estimación de parámetros y con ellos se comprobó que con dichos modelos se obtenía un error de reconstrucción menor, que con los modelos clásicos y una información más detallada en aplicaciones de seguimiento de formantes.

Evolucionando en el modelado por fases glóticas, se incorporó información psicoacústica, en una búsqueda de mayor capacidad discriminante. Esta actuación constituye otra contribución novedosa, debido a que no había sido añadido esta información perceptual a la estimación por fases glóticas. Con esta evolución el modelo incorpora una mayor resolución espectral perceptualmente destacable a la resolución temporal propia del modelo por fases.

Con este nuevo modelado se dedujeron los algoritmos correspondientes de estimación de parámetros y se obtuvieron mejoras sobre los resultados alcanzados con los anteriores.

Un caso singular, si bien en su concepción apartado del modelado paramétrico lineal, es el que se da con la entropía de permutación como técncia de extracción de características. Los resultados obtenidos resultan muy prometedores por las mejoras que hemos visto en nuestros experimentos.

Para el bloque de modelado probabilístico, se extendió el estudio de funciones de probabilidad derivadas de funciones de densidad espectral de potencia de modelos paramétricos desde el modelado AR hacia el modelado ARMA, por un lado, y hacia la mezcla de funciones de densidad basadas en modelos AR, por el otro lado. Tanto una como la otra son contribuciones novedosas que han aportado una mayor flexibilidad para la definición de las funciones de densidad de probabilidad desde la densidad espectral de potencia, la primera desde el espacio completo, con un mayor detalle y la segunda desde el análisis localizado, con una mayor aproximación. Los resultados alcanzados se pueden considerar competitivos respecto de los métodos generales de aproximación de funciones de densidad de probabilidad (p.e., GMM).

Por último, y como demostración de las contribuciones que se pueden conseguir con el modelado paramétrico lineal, en un marco unificado, sobre un sistema de RAH, se probaron con diferentes bases de datos.

Por una parte se obtuvo una mejora con la utilización del modelado paramétrico lineal localizado y modificado perceptualmente, sobre una aplicación de reconocimiento con

voz patológica, sobre parametrizaciones clásicas, por su mayor detalle de representación.

Por otra parte se consiguió un resultado competitivo con la utilización del modelado probabilístico con modelos ARMA y con mezclas de AR, sobre algunas aplicaciones de reconocimiento de voz patológica, respecto al modelado clásico, por su capacidad de aproximación.

## 7.3. Líneas Futuras

Como líneas abiertas que pudieran en el futuro mejorar el trabajo presente podemos citar las siguientes:

- La determinación del orden óptimo de los modelos de características.
- La exploración de la excitación más adecuada a los modelos de características.
- La integración de los tramos sonoro y sordo en el modelado, y extensión en la definición de las estructuras.
- La utilización de modelos AR de orden alto en mezclas de densidades de probabilidad en el modelado probabilístico.
- La combinación de modelos AR y MA en mezclas de densidades de probabilidad en el modelado probabilístico.
- La utilización de modelos ARMA en mezclas de densidades de probabilidad en el modelado probabilístico.
- Alternativas al Expectation-Maximization del criterio de entrenamiento.
- Exploración detallada de la entropía de permutación como técnica de extracción de características.

# Apéndice A
# Bases de Datos

En este anexo vamos a relacionar las Bases de Datos (BBDD) usadas en esta memoria, las características que tienen y como las emplearemos. Nos interesará trabajar con BBDD internacionales y que se utilicen casi como estandar, para que la comparación tenga más confianza, pero no desdeñamos otras BBDD de uso más restringido o locales, por su capacidad de utilizar datos específicos. Las BBDD que presentamos son las siguientes: KAY Elemetric, Timit, Keele y Hadeco.

## A.1. Base de Datos keele

La base de datos para nuestros experimentos sobre detección de Instantes de Cierre Glótico (ICG) ha sido proporcionada por el Departamento de Comunicaciones y Neurociencia de la Universidad de Keele, en Staffordshire, Reino Unido [2].

La base está compuesta registros de voz y laringográficos de una duración aproximada de 40 segundos de voz muestreada a 20KHz y codificada con 16 bits. Hay cinco hombres y cinco mujeres. El texto que pronuncia cada locutor es el mismo; “the northwind and

the sun were disputing which was the stronger when a traveller came along, wrapped in a warm cloak. They agreed that the one who first succeded in making the traveller take his cloak off should be considered stronger than the other. Then the northwind blew as hard he could but the more he blew the more closely did the traveller fold his cloak around him and at last the northwind gave up the attempt. Then the sun shone warmly out and immediately the traveller took off his cloak and so the northwind was obliged to confess that the sun was the stronger of the two".

Cada registro original de señal ha sido dividido en cuatro partes de aproximadamente la misma duración con puntos de corte en zona de silencio. Los nuevos registros se marcaron manualmente en una de tres categorías; silencio, voz sorda y voz sonora. En los segmentos de voz sonora se utilizó la señal laringográfica para determinar los instantes de cierre glótico (y, además, periodo y frecuencia fundamental). La información con que trabajamos es; señal de voz, información de sonoridad e instantes de cierre glótico.

El número total de instantes de cierre glótico a detectar es de 18511 para las cinco mujeres y 10781 para los cinco hombres, que conjuntamente hacen un total de 29292.

## A.2. Base de Datos Hadeco

La Base de Datos Hadeco [1] es un conjunto de datos de voz que ha sido recopilado por el Grupo de Ingeniería Acústica de la ULPGC en diferentes centros médicos de la isla de Gran Canaria. La principal característica es que contiene voces patológicas producidas por personas con problemas en la comunicación oral, tanto a nivel físico como psíquico.

De forma más particular, esta BBDD contiene un corpus independiente de locutor, estando integrado por locuciones producidas por personas de ambos sexos y de todas las edades a partir de 6 años. El corpus está separado por sexos y por gupos de edades, de 6 a 12 años, de 12 a 15, de 15 a 35, de 35 a 60 y mayores de 60. El idioma del conjunto de datos es el español y el dialecto el canario.

El estilo del habla es semiespontáneo, ya que se trata de un corpus fonológico inducido, utilizado en el trabajo de los logopedas. La BBDD está compuesta por palabras pertenecientes a un código de 57 palabras, pronunciadas tanto por pacientes sanos como patológicos. Se dispone de 28 registros por palabra en el caso de pacientes con patologías y de 11 en el caso de sanos. El número de locuciones es de 3543, de las cuales 1077 se corresponden con pacientes sanos y 2466 a pacientes con patologías.

Las grabaciones está realizadas en un entorno de bajo nivel de ruido, recogidas a 44.1 kHz y procesadas para ponerlas disponibles en la BBDD con una frecuencia de muestreo de 22.05 kHz.

## A.3. Base de Datos Kay Elemetrics

La Base de Datos de Kay Elemetrics [3] es un conjunto de registros de la vocal sostenida /a/ prounciados por locutores sanos y con patologías,. La BBDD contiene

locuciones de pacientes de ambos sexos, con edades comprendidas entre 5 y 93 años, con diferentes idiomas y origen, mayoritariamente ingles y blanco no hispano. Asociado a cada registro tiene información sobre las característcas de la fonación, la patología, el pitch, el jitter, el shimmer, etc y sus estadísticos. El número de registros es de 53 sanos y 657 con patologías. La frecuencia de muestreo la especifica en cada caso, pudiendo ser de 25 o 50 kHz. Igualmente la duración es variable, siendo de alrededor de 3 segundos para los pacientes sanos y de 1 segundo para los patológicos.

## Bibliografía

[1] N. M. J. Espinosa Yánez, J., "Herramienta de ayuda para deficiencias en la comunicación oral (hadeco)," in *Proyecto Fin de Carrera*, 2001.

[2] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *ESCA Eurospeech 1995*, vol. 8, pp. 837–840, ESCA, 1995.

[3] "Massachusetts eye and ear infirmary, voice disorders database, version 1.03. CDROM," 1994.

## Apéndice B

# Resultados de Reconocimiento sobre la base de datos Hadeco

En este anexo se relacionan los resultados de los experimentos realizados sobre la base de batos Hadeco, en el que se emplearon las parametrizaciones propuestas de modelos de fases con polos comunes frente a los métodos clásicos. Los métodos localizados utilizaron el modelo con polos comunes y ceros particulares analizados sobre 3 ciclos consecutivos, se obtuvo la parametrización para la fase abierta, CFA, para la fase cerrada, CFC, para ambas fases conjuntamente CFS, para el periodo completo, CPC y el modelo paramétrico transformado, MTF. Los métodos clásicos empleados fueron la parametrización MFCC, LPC y LPC-Cepstrum.

En las tablas de resultados se dan las tasas de acierto por el tipo de parametrización, la palabra sobre la que se efectúa el reconocimiento y los parámetros que se emplearon para conseguirlo. Los parámetros que se usaron fueron: el número de parámetros, P, el uso o no (1 o 0 respectivamente) de la información dinámica (1ª derivada), Delta, el número de estados para el modelo no patológico, Nsp, el número de estados para el modelo patológico, Ncp y el parámetro de transformación en frecuencia, $\lambda$, cuando se

use. La medida empleada para calificar el modelo fue la tasa de acierto, medida como el número de aciertos sobre el total de intentos. En las tablas se indican las tasas de acierto para las voces son patología como %SP y la tasa de aciertos para las voces con patología como %CP.

## B.1. Resultados de la parametrización LPC

Los resultados del experiemento que utiliza la parametrización LPC se presentan en la tabla B.2 donde se indican las tasas de acierto en el caso óptimo.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 16 | 1 | 5 | 5 | 100 | 89.29 |
| Boca | 16 | 1 | 4 | 5 | 100 | 85.19 |
| Bruja | 14 | 1 | 3 | 5 | 100 | 92.59 |
| Cabra | 14 | 1 | 3 | 5 | 100 | 100 |
| Campana | 10 | 1 | 4 | 5 | 100 | 92.59 |
| Caramelo | 10 | 1 | 4 | 4 | 100 | 96.30 |
| Casa | 12 | 0 | 3 | 5 | 100 | 100 |
| Clavo | 10 | 0 | 4 | 4 | 100 | 96 |
| Cuchara | 10 | 1 | 3 | 4 | 100 | 96.15 |
| Dedo | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Ducha | 10 | 0 | 3 | 4 | 100 | 92.59 |
| Escoba | 10 | 1 | 3 | 4 | 90.91 | 88.89 |
| Flan | 10 | 1 | 4 | 5 | 100 | 100 |
| Fresa | 10 | 0 | 3 | 5 | 100 | 100 |
| Fuma | 10 | 0 | 4 | 5 | 90.91 | 100 |
| Gafa | 12 | 0 | 4 | 5 | 100 | 100 |
| Globo | 18 | 1 | 3 | 4 | 100 | 92.59 |
| Gorro | 14 | 0 | 3 | 5 | 90.91 | 96.30 |
| Grifo | 12 | 0 | 3 | 4 | 100 | 96.30 |
| Indio | 10 | 1 | 5 | 4 | 100 | 92.59 |
| Jarra | 10 | 1 | 5 | 5 | 100 | 100 |
| Jaula | 10 | 1 | 3 | 3 | 100 | 96.30 |
| Lápiz | 14 | 0 | 5 | 4 | 90.91 | 92.59 |
| Lavadora | 10 | 1 | 3 | 3 | 100 | 92.59 |
| Luna | 12 | 0 | 3 | 3 | 100 | 92.59 |
| Llave | 10 | 0 | 5 | 5 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Mariposa | 10 | 1 | 3 | 5 | 90.91 | 96.30 |
| Moto | 14 | 0 | 3 | 4 | 90.91 | 88.89 |
| Niño | 12 | 0 | 3 | 4 | 90.91 | 92.59 |
| Ojo | 10 | 1 | 4 | 5 | 100 | 100 |
| Pala | 14 | 1 | 4 | 5 | 90.91 | 92.59 |
| Palmera | 10 | 1 | 4 | 5 | 100 | 100 |
| Pan | 14 | 1 | 3 | 5 | 90.91 | 96.30 |
| Peine | 18 | 1 | 4 | 4 | 100 | 88.89 |
| Periódico | 10 | 0 | 3 | 5 | 100 | 100 |
| Pez | 14 | 1 | 4 | 4 | 100 | 96.15 |
| Piano | 16 | 0 | 4 | 4 | 100 | 92.59 |
| Pie | 10 | 1 | 4 | 5 | 90.91 | 96.15 |
| Piña | 10 | 1 | 3 | 5 | 90.91 | 96.30 |
| Pistola | 12 | 1 | 4 | 5 | 100 | 96.30 |
| Plátano | 14 | 0 | 3 | 5 | 100 | 100 |
| Playa | 16 | 1 | 4 | 4 | 100 | 96.30 |
| Preso | 18 | 0 | 4 | 5 | 100 | 96 |
| Pueblo | 12 | 1 | 5 | 5 | 100 | 88.46 |
| Puerta | 12 | 1 | 3 | 5 | 90.91 | 100 |
| Ratón | 18 | 1 | 3 | 5 | 90.91 | 96.15 |
| Semáforo | 10 | 1 | 3 | 5 | 100 | 96.15 |
| Silla | 18 | 0 | 3 | 5 | 90 | 100 |
| Sol | 18 | 0 | 4 | 3 | 100 | 92.59 |
| Tambor | 10 | 1 | 5 | 4 | 100 | 81.48 |
| Taza | 12 | 1 | 5 | 5 | 100 | 96.30 |
| Teléfono | 10 | 0 | 3 | 5 | 100 | 92.59 |
| Toalla | 12 | 0 | 5 | 5 | 90.91 | 96.30 |
| Toro | 18 | 1 | 4 | 3 | 100 | 96.30 |
| Tortuga | 16 | 0 | 5 | 5 | 90.91 | 92.59 |
| Tren | 14 | 0 | 3 | 5 | 100 | 92.31 |
| Zapato | 10 | 1 | 3 | 4 | 100 | 100 |
| TOTAL | | | | | 97.43 | 95.07 |

Cuadro B.2: Resultados óptimos en Hadeco utilizando LPC

Examinando los resultados se puede ver que existen 11 palabras con éxito absoluto: cabra, casa, flan, fresa, gafa, jarra, ojo, palmera, periódico, plátano y zapato. También se observa que la tasa de acierto mínima para las señales patológicas es del 81.48 % y corresponde a la palabra tambor, mientras que para señales no patológicas es del 90 % y corresponde a la palabra silla. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 41 palabras y para señales patológicas dicha tasa de acierto se da para 14 palabras. A partir de las tasas globales de acierto de señales sin patologías y con patologías se obtiene la tasa global de este sistema que es del 95.75 %.

## B.2. Resultados de la parametrización MFCC

Los resultados del experiemento que utiliza la parametrización MFCC se presentan en la tabla B.4 donde se indican las tasas de acierto en el caso óptimo.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 14 | 1 | 5 | 5 | 100 | 100 |
| Boca | 10 | 1 | 5 | 4 | 100 | 92.59 |
| Bruja | 12 | 1 | 4 | 4 | 90.91 | 96.30 |
| Cabra | 10 | 1 | 4 | 4 | 90.91 | 100 |
| Campana | 14 | 0 | 5 | 5 | 100 | 92.59 |
| Caramelo | 12 | 1 | 5 | 3 | 100 | 100 |
| Casa | 10 | 1 | 5 | 5 | 90.91 | 92.59 |
| Clavo | 14 | 1 | 3 | 3 | 100 | 100 |
| Cuchara | 18 | 0 | 4 | 5 | 100 | 100 |
| Dedo | 12 | 1 | 4 | 4 | 100 | 100 |
| Ducha | 16 | 0 | 5 | 4 | 90.91 | 100 |
| Escoba | 14 | 1 | 3 | 5 | 90.91 | 92.59 |
| Flan | 10 | 1 | 5 | 4 | 100 | 100 |
| Fresa | 18 | 1 | 5 | 5 | 90.91 | 100 |
| Fuma | 12 | 0 | 4 | 5 | 90.91 | 96.30 |
| Gafa | 14 | 0 | 5 | 4 | 81.82 | 96.30 |
| Globo | 12 | 0 | 5 | 5 | 90.91 | 100 |
| Gorro | 10 | 1 | 5 | 4 | 100 | 96.30 |
| Grifo | 18 | 1 | 5 | 5 | 100 | 96.30 |
| Indio | 10 | 1 | 3 | 5 | 90.91 | 96.30 |
| Jarra | 18 | 1 | 4 | 5 | 100 | 100 |
| Jaula | 10 | 0 | 5 | 4 | 90.91 | 96.30 |
| Lápiz | 18 | 1 | 3 | 3 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Lavadora | 16 | 0 | 3 | 4 | 90.91 | 92.59 |
| Luna | 14 | 1 | 5 | 5 | 90.91 | 100 |
| Llave | 10 | 1 | 4 | 4 | 90.91 | 100 |
| Mariposa | 10 | 1 | 4 | 3 | 100 | 92.59 |
| Moto | 14 | 1 | 5 | 4 | 100 | 100 |
| Niño | 12 | 0 | 3 | 3 | 90.91 | 96.30 |
| Ojo | 12 | 1 | 3 | 5 | 90.91 | 100 |
| Pala | 16 | 0 | 5 | 3 | 100 | 100 |
| Palmera | 12 | 1 | 3 | 5 | 90.91 | 92.31 |
| Pan | 16 | 1 | 5 | 4 | 100 | 100 |
| Peine | 12 | 1 | 5 | 4 | 100 | 96.30 |
| Periódico | 16 | 1 | 3 | 4 | 100 | 96 |
| Pez | 10 | 0 | 3 | 5 | 100 | 100 |
| Piano | 12 | 1 | 5 | 5 | 90.91 | 100 |
| Pie | 14 | 0 | 5 | 3 | 81.82 | 92.31 |
| Piña | 14 | 1 | 5 | 4 | 100 | 96.30 |
| Pistola | 16 | 0 | 5 | 5 | 100 | 96.30 |
| Plátano | 12 | 1 | 3 | 5 | 100 | 100 |
| Playa | 10 | 1 | 5 | 5 | 100 | 100 |
| Preso | 14 | 0 | 3 | 3 | 100 | 92 |
| Pueblo | 10 | 0 | 5 | 5 | 100 | 100 |
| Puerta | 14 | 1 | 4 | 4 | 100 | 100 |
| Ratón | 12 | 1 | 4 | 4 | 90.91 | 100 |
| Semáforo | 12 | 1 | 4 | 3 | 100 | 96.15 |
| Silla | 16 | 0 | 3 | 3 | 90 | 100 |
| Sol | 12 | 1 | 3 | 3 | 100 | 100 |
| Tambor | 16 | 1 | 5 | 4 | 90.91 | 96.30 |
| Taza | 10 | 1 | 4 | 4 | 90.91 | 96.30 |
| Teléfono | 16 | 1 | 4 | 5 | 90.91 | 100 |
| Toalla | 10 | 0 | 3 | 3 | 90.91 | 92.59 |
| Toro | 10 | 1 | 3 | 5 | 100 | 100 |
| Tortuga | 10 | 1 | 5 | 4 | 90.91 | 92.59 |
| Tren | 10 | 0 | 5 | 3 | 100 | 92.31 |
| Zapato | 10 | 1 | 5 | 5 | 100 | 100 |
| Total | | | | | 95.50 | 97.37 |

Cuadro B.4: Resultados óptimos en Hadeco utilizando MFCC

Examinando los resultados observamos que hay 18 palabras para las que el sistema siempre decide corretamente. También se observa que la tasa de acierto mínima para las señales patológicas es del 92 % y corresponde a la palabra preso mientras que para señales no patológicas es del 81.82 % y corresponde a las palabras gafa y pie. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 31 palabras y para señales patológicas dicha tasa de acierto se da para 29 palabras. La tasa de acierto global del sistema es del 96.83 %.

## B.3. Resultados de la parametrización CPC

Los resultados del experiemento que utiliza la parametrización del modelo de fases con polos comunes sobre el periodo completo se presentan en la tabla B.6 donde se indican las tasas de acierto en el caso óptimo.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 10 | 0 | 3 | 5 | 100 | 96.42 |
| Boca | 16 | 1 | 3 | 5 | 100 | 88.89 |
| Bruja | 10 | 0 | 3 | 4 | 100 | 92.59 |
| Cabra | 10 | 0 | 5 | 5 | 90.91 | 92.30 |
| Campana | 12 | 0 | 3 | 4 | 90.91 | 96.30 |
| Caramelo | 16 | 0 | 3 | 4 | 90.91 | 92.59 |
| Casa | 12 | 0 | 4 | 5 | 100 | 100 |
| Clavo | 10 | 0 | 3 | 5 | 100 | 100 |
| Cuchara | 18 | 0 | 4 | 5 | 100 | 100 |
| Dedo | 18 | 0 | 4 | 3 | 90.91 | 100 |
| Ducha | 12 | 0 | 3 | 3 | 100 | 92.59 |
| Escoba | 12 | 0 | 5 | 5 | 100 | 88.89 |
| Flan | 16 | 0 | 4 | 5 | 100 | 96.30 |
| Fresa | 18 | 0 | 3 | 5 | 90.91 | 100 |
| Fuma | 10 | 0 | 4 | 4 | 100 | 96.30 |
| Gafa | 16 | 0 | 4 | 5 | 100 | 100 |
| Globo | 12 | 0 | 3 | 4 | 100 | 92.59 |
| Gorro | 12 | 0 | 3 | 5 | 100 | 92.59 |
| Grifo | 16 | 0 | 4 | 5 | 100 | 85.19 |
| Indio | 10 | 0 | 3 | 3 | 100 | 92.59 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Jarra | 14 | 0 | 3 | 5 | 90 | 100 |
| Jaula | 10 | 0 | 5 | 5 | 90.91 | 100 |
| Lápiz | 16 | 0 | 3 | 5 | 90.91 | 96.30 |
| Lavadora | 10 | 0 | 3 | 4 | 100 | 92.59 |
| Luna | 10 | 1 | 5 | 4 | 90.91 | 88.89 |
| Llave | 10 | 1 | 5 | 5 | 90.91 | 96.30 |
| Mariposa | 12 | 0 | 5 | 4 | 90.91 | 96.30 |
| Moto | 12 | 0 | 5 | 5 | 90.91 | 92.59 |
| Niño | 14 | 0 | 4 | 4 | 100 | 100 |
| Ojo | 10 | 1 | 3 | 5 | 100 | 92.59 |
| Pala | 10 | 0 | 5 | 5 | 90.91 | 92.59 |
| Palmera | 10 | 1 | 4 | 5 | 100 | 100 |
| Pan | 12 | 0 | 3 | 4 | 100 | 88.89 |
| Peine | 10 | 0 | 4 | 4 | 100 | 92.59 |
| Periódico | 14 | 0 | 4 | 5 | 100 | 96 |
| Pez | 16 | 0 | 5 | 5 | 100 | 100 |
| Piano | 14 | 0 | 3 | 5 | 90.91 | 96.30 |
| Pie | 16 | 0 | 3 | 4 | 100 | 96.15 |
| Piña | 16 | 0 | 4 | 5 | 100 | 96.30 |
| Pistola | 16 | 0 | 4 | 5 | 100 | 96.30 |
| Plátano | 16 | 0 | 3 | 5 | 100 | 100 |
| Playa | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Preso | 14 | 0 | 3 | 4 | 100 | 96 |
| Pueblo | 10 | 0 | 5 | 5 | 90.91 | 88.46 |
| Puerta | 10 | 0 | 5 | 5 | 100 | 96.30 |
| Ratón | 18 | 0 | 3 | 4 | 90.91 | 96.15 |
| Semáforo | 18 | 0 | 3 | 5 | 90.91 | 92.31 |
| Silla | 18 | 0 | 3 | 5 | 100 | 96.30 |
| Sol | 18 | 0 | 3 | 5 | 90.91 | 92.59 |
| Tambor | 14 | 1 | 5 | 4 | 90.91 | 88.89 |
| Taza | 10 | 0 | 3 | 3 | 100 | 92.59 |
| Teléfono | 10 | 0 | 3 | 5 | 100 | 92.59 |
| Toalla | 14 | 1 | 4 | 5 | 90.91 | 96.30 |
| Toro | 12 | 0 | 3 | 5 | 90.91 | 96.30 |
| Tortuga | 12 | 0 | 3 | 4 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Tren | 12 | 0 | 3 | 5 | 100 | 96.15 |
| Zapato | 10 | 1 | 3 | 3 | 100 | 100 |
| TOTAL | | | | | 96.63 | 95.07 |

Cuadro B.6: Resultados óptimos en Hadeco utilizando Coeficientes del periodo completo

Examinando los resultados comprobamos que hay 9 palabras para las que el sistema siempre tiene éxito. También se observa que la tasa de acierto mínima para las señales con patologías es del 85.19 % y corresponde a la palabra grifo mientras que para señales no patológicas es del 90 % y corresponde a la palabra jarra. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 36 palabras y para señales patológicas dicha tasa de acierto se da para 13 palabras. Se destaca que en este caso la tasa de acierto para señales patológicas es más baja que para MFCC, Banco de Filtros, LPC-Cepstrum y que para la Transformación Matricial Dependiente del Modelo, mientras que para señales no patológicas es similar a la obtenida con LPC-Cpestrum. La tasa de acierto global del sistema es, en este caso, del 95.52 %.

## B.4. Resultados de la parametrización CFS

Los resultados del experiemento que utiliza la parametrización conjunta del modelo de fases con polos comunes sobre las fases abierta y cerrada se presentan en la tabla B.8 donde se indican las tasas de acierto en el caso óptimo.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 10 | 1 | 3 | 4 | 100 | 85.71 |
| Boca | 16 | 1 | 4 | 5 | 90.91 | 96.30 |
| Bruja | 10 | 0 | 5 | 4 | 100 | 92.59 |
| Cabra | 12 | 0 | 5 | 5 | 81.82 | 96.15 |
| Campana | 10 | 0 | 4 | 4 | 90.91 | 96.30 |
| Caramelo | 18 | 0 | 3 | 4 | 100 | 88.89 |
| Casa | 14 | 0 | 4 | 4 | 90.91 | 100 |
| Clavo | 10 | 0 | 4 | 5 | 100 | 92 |
| Cuchara | 12 | 0 | 3 | 4 | 100 | 96.15 |
| Dedo | 10 | 0 | 4 | 5 | 90.91 | 96.30 |
| Ducha | 10 | 0 | 4 | 5 | 90.91 | 96.30 |
| Escoba | 10 | 0 | 3 | 4 | 90.91 | 92.59 |
| Flan | 10 | 0 | 3 | 5 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Fresa | 10 | 0 | 3 | 5 | 100 | 100 |
| Fuma | 16 | 0 | 5 | 5 | 90.91 | 92.59 |
| Gafa | 12 | 0 | 5 | 5 | 100 | 96.30 |
| Globo | 10 | 0 | 4 | 5 | 100 | 92.59 |
| Gorro | 10 | 0 | 3 | 5 | 100 | 85.18 |
| Grifo | 10 | 0 | 5 | 5 | 100 | 81.48 |
| Indio | 18 | 0 | 3 | 3 | 100 | 96.30 |
| Jarra | 10 | 0 | 4 | 4 | 100 | 100 |
| Jaula | 14 | 0 | 3 | 3 | 100 | 92.59 |
| Lápiz | 12 | 0 | 3 | 5 | 90.91 | 92.59 |
| Lavadora | 12 | 0 | 5 | 5 | 90.91 | 96.30 |
| Luna | 12 | 0 | 4 | 5 | 100 | 92.59 |
| Llave | 12 | 0 | 4 | 5 | 90.91 | 92.59 |
| Mariposa | 12 | 1 | 3 | 5 | 81.82 | 96.30 |
| Moto | 14 | 0 | 3 | 4 | 90.91 | 96.30 |
| Niño | 18 | 0 | 4 | 5 | 90.91 | 92.59 |
| Ojo | 10 | 0 | 3 | 5 | 90.91 | 100 |
| Pala | 12 | 1 | 4 | 4 | 90.91 | 96.30 |
| Palmera | 12 | 0 | 3 | 3 | 90.91 | 100 |
| Pan | 10 | 1 | 5 | 4 | 100 | 100 |
| Peine | 10 | 0 | 3 | 3 | 100 | 88.89 |
| Periódico | 10 | 0 | 5 | 4 | 100 | 92 |
| Pez | 10 | 0 | 3 | 5 | 100 | 96.15 |
| Piano | 18 | 0 | 5 | 5 | 90.91 | 96.30 |
| Pie | 16 | 0 | 3 | 4 | 100 | 84.62 |
| Piña | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Pistola | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Plátano | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Playa | 12 | 0 | 4 | 4 | 100 | 92.59 |
| Preso | 10 | 0 | 4 | 5 | 100 | 96 |
| Pueblo | 14 | 0 | 5 | 5 | 100 | 88.46 |
| Puerta | 16 | 1 | 5 | 4 | 100 | 96.30 |
| Ratón | 12 | 0 | 5 | 5 | 100 | 100 |
| Semáforo | 12 | 0 | 5 | 5 | 100 | 92.31 |
| Silla | 16 | 0 | 4 | 5 | 90 | 100 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Sol | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Tambor | 14 | 0 | 4 | 5 | 100 | 81.48 |
| Taza | 10 | 0 | 5 | 5 | 100 | 92.59 |
| Teléfono | 16 | 0 | 4 | 5 | 100 | 100 |
| Toalla | 12 | 0 | 3 | 4 | 90.91 | 92.59 |
| Toro | 16 | 1 | 3 | 3 | 100 | 92.59 |
| Tortuga | 16 | 0 | 4 | 4 | 90.91 | 88.89 |
| Tren | 12 | 0 | 4 | 5 | 100 | 96.15 |
| Zapato | 16 | 0 | 4 | 4 | 90.91 | 100 |
| TOTAL | | | | | 95.98 | 94.15 |

Cuadro B.8: Resultados óptimos en Hadeco utilizando Coeficientes de las fases abierta y cerrada conjuntamente

Examinando los resultados podemos ver que hay 5 palabras para las que el sistema siempre decide adecuadamente. También se observa que la tasa de acierto mínima para las señales con patologías es del 81.48 % y corresponde a las palabras grifo y tambor mientras que para señales no patológicas es del 81.82 % y corresponde a las palabras cabra y mariposa. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 34 palabras y para señales patológicas dicha tasa de acierto se da para 10 palabras. Se observa que la tasa de acierto para señales patológicas es menor que la obtenida en casos anteriores. La tasa de acierto global del sistema es del 94.68 %.

## B.5. Resultados de la parametrización CFA

Los resultados del experiemento que utiliza la parametrización del modelo de fases con polos comunes sobre la fases abierta se presentan en la tabla B.10 donde se indican las tasas de acierto en el caso óptimo.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 14 | 0 | 3 | 5 | 100 | 96.43 |
| Boca | 14 | 1 | 3 | 5 | 90.91 | 100 |
| Bruja | 10 | 0 | 3 | 5 | 100 | 88.89 |
| Cabra | 18 | 0 | 3 | 5 | 81.82 | 92.31 |
| Campana | 12 | 0 | 3 | 4 | 90.91 | 92.59 |
| Caramelo | 10 | 0 | 3 | 5 | 81.82 | 96.30 |
| Casa | 10 | 0 | 3 | 3 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Clavo | 18 | 0 | 5 | 5 | 100 | 92 |
| Cuchara | 10 | 0 | 4 | 5 | 90.91 | 100 |
| Dedo | 10 | 0 | 4 | 4 | 100 | 96.30 |
| Ducha | 10 | 0 | 3 | 3 | 100 | 88.89 |
| Escoba | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Flan | 16 | 0 | 3 | 5 | 100 | 100 |
| Fresa | 10 | 0 | 3 | 5 | 100 | 100 |
| Fuma | 10 | 0 | 4 | 4 | 100 | 96.30 |
| Gafa | 10 | 0 | 4 | 4 | 100 | 100 |
| Globo | 14 | 1 | 5 | 5 | 100 | 92.59 |
| Gorro | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Grifo | 10 | 0 | 3 | 5 | 90 | 88.89 |
| Indio | 16 | 1 | 3 | 3 | 100 | 92.59 |
| Jarra | 18 | 0 | 3 | 5 | 100 | 100 |
| Jaula | 10 | 0 | 3 | 4 | 100 | 96.30 |
| Lápiz | 18 | 1 | 4 | 3 | 100 | 96.30 |
| Lavadora | 10 | 0 | 5 | 4 | 90.91 | 92.59 |
| Luna | 12 | 0 | 4 | 5 | 100 | 92.59 |
| Llave | 12 | 1 | 4 | 5 | 90.91 | 96.30 |
| Mariposa | 10 | 1 | 5 | 4 | 90.91 | 92.59 |
| Moto | 18 | 0 | 3 | 4 | 81.82 | 96.30 |
| Niño | 12 | 0 | 4 | 4 | 100 | 92.59 |
| Ojo | 10 | 1 | 3 | 5 | 100 | 100 |
| Pala | 10 | 1 | 5 | 4 | 100 | 96.30 |
| Palmera | 12 | 0 | 4 | 5 | 100 | 100 |
| Pan | 12 | 0 | 3 | 5 | 100 | 96.30 |
| Peine | 10 | 0 | 3 | 3 | 100 | 92.59 |
| Periódico | 16 | 0 | 5 | 5 | 100 | 96 |
| Pez | 12 | 0 | 3 | 5 | 100 | 96.15 |
| Piano | 14 | 0 | 3 | 4 | 100 | 96.30 |
| Pie | 10 | 1 | 3 | 4 | 100 | 96.15 |
| Piña | 18 | 1 | 4 | 4 | 100 | 85.18 |
| Pistola | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Plátano | 14 | 0 | 4 | 4 | 100 | 96.30 |
| Playa | 10 | 0 | 4 | 4 | 100 | 92.59 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Preso | 14 | 0 | 3 | 3 | 90.91 | 96 |
| Pueblo | 16 | 0 | 5 | 4 | 100 | 88.46 |
| Puerta | 10 | 0 | 5 | 5 | 100 | 100 |
| Ratón | 10 | 0 | 3 | 5 | 90.91 | 96.15 |
| Semáforo | 12 | 0 | 3 | 4 | 90.91 | 92.31 |
| Silla | 12 | 1 | 5 | 3 | 100 | 92.59 |
| Sol | 16 | 0 | 4 | 5 | 90.91 | 92.59 |
| Tambor | 10 | 0 | 3 | 5 | 90.91 | 85.18 |
| Taza | 12 | 1 | 3 | 5 | 100 | 88.89 |
| Teléfono | 10 | 0 | 4 | 5 | 100 | 92.59 |
| Toalla | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Toro | 18 | 1 | 3 | 5 | 90.91 | 96.30 |
| Tortuga | 18 | 0 | 3 | 4 | 100 | 85.18 |
| Tren | 16 | 0 | 4 | 5 | 100 | 100 |
| Zapato | 10 | 0 | 4 | 4 | 100 | 100 |
| TOTAL | | | | | 96.47 | 94.55 |

Cuadro B.10: Resultados óptimos en Hadeco utilizando Coeficientes de la fase abierta

Examinando los resultados nos damos cuenta que hay 9 palabras para las que el sistema siempre decide bien. También se observa que la tasa de acierto mínima para las señales con patologías es del 85.18 % y corresponde a las palabras piña, tambor y tortuga mientras que para señales no patológicas es del 81.82 % y corresponde a las palabras cabra, caramelo y moto. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 38 palabras y para señales patológicas dicha tasa de acierto se da para 11 palabras. Se observa que las tasas de acierto son similares a cuando utilizábamos las dos fases, consiguiéndose una tasa de acierto global del 95.10 %.

## B.6. Resultados de la parametrización CFC

Los resultados del experiemento que utiliza la parametrización del modelo de fases con polos comunes sobre la fases cerrada se presentan en la tabla B.12 donde se indican las tasas de acierto en el caso óptimo

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Árbol | 14 | 0 | 3 | 5 | 100 | 96.43 |
| Boca | 14 | 0 | 3 | 3 | 81.82 | 92.59 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Bruja | 10 | 0 | 3 | 4 | 100 | 88.89 |
| Cabra | 18 | 0 | 3 | 5 | 81.82 | 92.31 |
| Campana | 12 | 0 | 3 | 4 | 90.91 | 92.59 |
| Caramelo | 10 | 0 | 3 | 5 | 81.82 | 96.30 |
| Casa | 18 | 0 | 3 | 5 | 90.91 | 100 |
| Clavo | 12 | 0 | 3 | 4 | 90.91 | 96 |
| Cuchara | 10 | 0 | 4 | 5 | 90.91 | 100 |
| Dedo | 10 | 0 | 4 | 4 | 100 | 96.30 |
| Ducha | 10 | 0 | 3 | 3 | 100 | 88.89 |
| Escoba | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Flan | 16 | 0 | 3 | 5 | 100 | 100 |
| Fresa | 10 | 0 | 3 | 4 | 100 | 100 |
| Fuma | 10 | 0 | 4 | 3 | 90.91 | 96.30 |
| Gafa | 10 | 0 | 4 | 4 | 100 | 100 |
| Globo | 10 | 0 | 4 | 5 | 100 | 92.59 |
| Gorro | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Grifo | 10 | 0 | 3 | 5 | 90 | 88.89 |
| Indio | 16 | 1 | 3 | 3 | 100 | 92.59 |
| Jarra | 18 | 0 | 3 | 5 | 100 | 100 |
| Jaula | 10 | 0 | 3 | 4 | 100 | 96.30 |
| Lápiz | 18 | 1 | 4 | 3 | 100 | 96.30 |
| Lavadora | 10 | 0 | 5 | 4 | 90.91 | 92.59 |
| Luna | 12 | 0 | 4 | 5 | 100 | 92.59 |
| Llave | 12 | 1 | 4 | 5 | 90.91 | 96.30 |
| Mariposa | 16 | 1 | 4 | 4 | 90.91 | 96.30 |
| Moto | 18 | 0 | 3 | 4 | 81.82 | 96.30 |
| Niño | 16 | 0 | 4 | 5 | 90.91 | 100 |
| Ojo | 10 | 1 | 3 | 5 | 100 | 100 |
| Pala | 10 | 1 | 5 | 4 | 100 | 96.30 |
| Palmera | 12 | 0 | 4 | 5 | 100 | 100 |
| Pan | 12 | 0 | 3 | 5 | 100 | 96.30 |
| Peine | 10 | 0 | 3 | 3 | 100 | 92.59 |
| Periódico | 16 | 0 | 5 | 5 | 100 | 96 |
| Pez | 12 | 0 | 3 | 5 | 100 | 96.15 |
| Piano | 14 | 0 | 3 | 4 | 100 | 96.30 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP |
|---|---|---|---|---|---|---|
| Pie | 10 | 1 | 3 | 4 | 100 | 96.15 |
| Piña | 16 | 0 | 4 | 4 | 100 | 96.30 |
| Pistola | 10 | 0 | 4 | 5 | 100 | 96.30 |
| Plátano | 12 | 0 | 3 | 5 | 90.91 | 100 |
| Playa | 10 | 0 | 4 | 4 | 100 | 92.59 |
| Preso | 14 | 0 | 3 | 3 | 90.91 | 96 |
| Pueblo | 16 | 0 | 5 | 4 | 90.91 | 92.31 |
| Puerta | 10 | 0 | 5 | 5 | 100 | 100 |
| Ratón | 10 | 0 | 3 | 5 | 90.91 | 96.15 |
| Semáforo | 12 | 0 | 3 | 5 | 90.91 | 92.31 |
| Silla | 18 | 1 | 5 | 4 | 100 | 96.30 |
| Sol | 16 | 0 | 4 | 5 | 90.91 | 92.59 |
| Tambor | 10 | 0 | 3 | 5 | 90.91 | 85.18 |
| Taza | 10 | 0 | 4 | 4 | 100 | 92.59 |
| Teléfono | 10 | 0 | 4 | 5 | 100 | 92.59 |
| Toalla | 10 | 0 | 3 | 5 | 90.91 | 92.59 |
| Toro | 18 | 1 | 3 | 5 | 90.91 | 96.30 |
| Tortuga | 14 | 1 | 3 | 4 | 90.91 | 92.59 |
| Tren | 16 | 0 | 4 | 5 | 100 | 100 |
| Zapato | 10 | 0 | 4 | 4 | 100 | 100 |
| TOTAL | | | | | 95.18 | 95.33 |

Cuadro B.12: Resultados óptimos en Hadeco utilizando Coeficientes de la fase cerrada

Examinando los resultados detectamos que hay 9 palabras para las que el sistema siempre decide corretamente. También se observa que la tasa de acierto mínima para las señales con patologías es del 85.18 % que en este caso corresponde a la palabra tambor. Para señales no patológicas la tasa de acierto mínima es del 81.82 % y corresponde a las palabras boca, cabra, caramelo y moto. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 31 palabras y para señales patológicas dicha tasa de acierto se da para 13 palabras. Se detecta que la tasa de acierto para señales no patológicas es menor a la obtenida utilizando los coeficientes de la fase abierta, mientras que para señales con patologías es mejor. En este caso se consigue una tasa de acierto global del 95.29 %.

## B.7. Resultados de la parametrización MTF

Los resultados del experiemento que utiliza la parametrización del modelo de fases con polos comunes transformado en frecuencia sobre la fases cerrada se presentan en la tabla B.14 donde se indican las tasas de acierto en el caso óptimoTasas de acierto para Coeficientes de fase modificados en trecuencia.

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP | $\lambda$ |
|---|---|---|---|---|---|---|---|
| Árbol | 10 | 0 | 5 | 5 | 100 | 96.42 | 0.6 |
| Boca | 10 | 0 | 5 | 4 | 100 | 96.30 | 0.8 |
| Bruja | 16 | 0 | 4 | 4 | 100 | 96.30 | 0.6 |
| Cabra | 14 | 0 | 3 | 4 | 100 | 96.15 | 0.5 |
| Campana | 18 | 0 | 5 | 5 | 100 | 100 | 0.8 |
| Caramelo | 12 | 0 | 3 | 4 | 100 | 100 | 0.7 |
| Casa | 16 | 0 | 5 | 5 | 100 | 100 | 0.4 |
| Clavo | 12 | 0 | 3 | 4 | 90.91 | 100 | 0.8 |
| Cuchara | 10 | 1 | 3 | 5 | 100 | 96.15 | 0.4 |
| Dedo | 14 | 0 | 3 | 3 | 90.91 | 100 | 0.7 |
| Ducha | 16 | 0 | 5 | 5 | 90.91 | 100 | 0.6 |
| Escoba | 18 | 0 | 4 | 5 | 90.91 | 96.30 | 0.4 |
| Flan | 12 | 0 | 4 | 5 | 100 | 100 | 0.6 |
| Fresa | 10 | 0 | 4 | 5 | 90.91 | 100 | 0.6 |
| Fuma | 12 | 0 | 5 | 5 | 90.91 | 96.30 | 0.4 |
| Gafa | 16 | 1 | 4 | 5 | 100 | 100 | 0.4 |
| Globo | 14 | 0 | 5 | 5 | 100 | 96.30 | 0.4 |
| Gorro | 18 | 0 | 4 | 5 | 100 | 96.30 | 0.5 |
| Grifo | 12 | 0 | 3 | 5 | 90 | 96.30 | 0.4 |
| Indio | 18 | 1 | 3 | 5 | 100 | 100 | 0.8 |
| Jarra | 14 | 0 | 5 | 5 | 100 | 96.15 | 0.6 |
| Jaula | 14 | 0 | 4 | 5 | 100 | 100 | 0.4 |
| Lápiz | 16 | 1 | 4 | 4 | 100 | 96.30 | 0.4 |
| Lavadora | 18 | 0 | 3 | 4 | 100 | 100 | 0.7 |
| Luna | 16 | 0 | 5 | 5 | 100 | 96.30 | 0.7 |
| Llave | 12 | 1 | 4 | 5 | 90.91 | 96.30 | 0.4 |
| Mariposa | 16 | 0 | 5 | 5 | 100 | 96.30 | 0.4 |
| Moto | 14 | 0 | 3 | 4 | 90.91 | 100 | 0.7 |
| Niño | 16 | 0 | 4 | 5 | 100 | 96.30 | 0.4 |

| Palabra | P | Delta | Nsp | Ncp | %SP | %CP | $\lambda$ |
|---|---|---|---|---|---|---|---|
| Ojo | 16 | 0 | 3 | 5 | 100 | 100 | 0.5 |
| Pala | 14 | 1 | 4 | 4 | 100 | 96.30 | 0.4 |
| Palmera | 12 | 0 | 5 | 4 | 100 | 100 | 0.7 |
| Pan | 12 | 1 | 4 | 4 | 100 | 100 | 0.5 |
| Peine | 12 | 1 | 4 | 5 | 100 | 96.30 | 0.8 |
| Periódico | 12 | 0 | 3 | 4 | 100 | 100 | 0.8 |
| Pez | 10 | 0 | 3 | 5 | 100 | 100 | 0.7 |
| Piano | 10 | 1 | 5 | 5 | 90.91 | 96.30 | 0.5 |
| Pie | 16 | 0 | 4 | 5 | 100 | 100 | 0.4 |
| Piña | 10 | 0 | 4 | 5 | 100 | 100 | 0.5 |
| Pistola | 12 | 1 | 3 | 4 | 100 | 96.30 | 0.5 |
| Plátano | 14 | 0 | 3 | 5 | 100 | 100 | 0.7 |
| Playa | 12 | 0 | 3 | 5 | 90.91 | 100 | 0.8 |
| Preso | 12 | 0 | 4 | 5 | 90.91 | 100 | 0.6 |
| Pueblo | 14 | 0 | 5 | 5 | 100 | 100 | 0.7 |
| Puerta | 14 | 0 | 5 | 5 | 100 | 96.30 | 0.4 |
| Ratón | 10 | 0 | 4 | 5 | 100 | 100 | 0.6 |
| Semáforo | 12 | 0 | 3 | 4 | 100 | 96.15 | 0.8 |
| Silla | 16 | 0 | 5 | 5 | 90 | 100 | 0.8 |
| Sol | 10 | 0 | 5 | 5 | 100 | 96.30 | 0.6 |
| Tambor | 14 | 0 | 4 | 5 | 90.91 | 96.30 | 0.8 |
| Taza | 10 | 0 | 3 | 4 | 100 | 100 | 0.5 |
| Teléfono | 16 | 0 | 3 | 5 | 100 | 100 | 0.6 |
| Toalla | 18 | 1 | 5 | 4 | 100 | 96.30 | 0.6 |
| Toro | 14 | 0 | 3 | 4 | 100 | 96.30 | 0.5 |
| Tortuga | 16 | 0 | 5 | 4 | 100 | 100 | 0.8 |
| Tren | 10 | 0 | 4 | 5 | 90.91 | 96.15 | 0.7 |
| Zapato | 18 | 0 | 5 | 4 | 100 | 100 | 0.4 |
| TOTAL | | | | | 97.59 | 98.23 | |

Cuadro B.14: Resultados óptimos en Hadeco utilizando Coeficientes de fase modificados en frecuencia

Examinando los resultados podemos darnos cuenta que los resultados obtenidos con esta parametrización son mejores que los obtenidos con las otras técnicas síncronas con

el periodo. En este caso, hay 22 palabras para las que el sistema siempre acierta. Se puede observar también que la mínima tasa de acierto para las señales con patologías es del 96.15 % y corresponde a las palabras cabra, cuchara, jarra, semáforo y tren. Para señales no patológicas la tasa de acierto mínima es del 90 % y corresponde a las palabras grifo y silla. Con esta parametrización la tasa de acierto para señales sin patologías es del 100 % para 42 palabras y para señales patológicas dicha tasa de acierto se da para 30 palabras. En este caso se consigue una tasa de acierto global del 98.04 %.

Apéndice C

# Entropía de permutación

En este anexo vamos a revisar someramente los conceptos de energía de permuatción y dinámica simbólica.

## C.1. Dinámica simbólica y entropía de permutación

Cuando se trata de cuantificar la complejidad de una determinada serie temporal, las entropías suelen ser a menudo la primera elección donde los métodos clásicos como la transformada de Fourier fallan. Si además se quiere tener en cuenta el orden temporal de los valores de la serie, entonces las series temporales pueden ser codificadas mediante secuencia de símbolos basados en la teoría de la dinámica simbólica.

Las entropía de permutación explora esa representación simbólica de una serie temporal. A pesar de que existe una considerable reducción de información, estas medidas son capaces de extraer información relevante de la señal de voz. En esta tesis, se ha planteado el uso de esta entropía descrita en estudios recientes [1] y que viene definida

por la expresión que se indica a continuación para diferentes secuencias de longitud $n$:

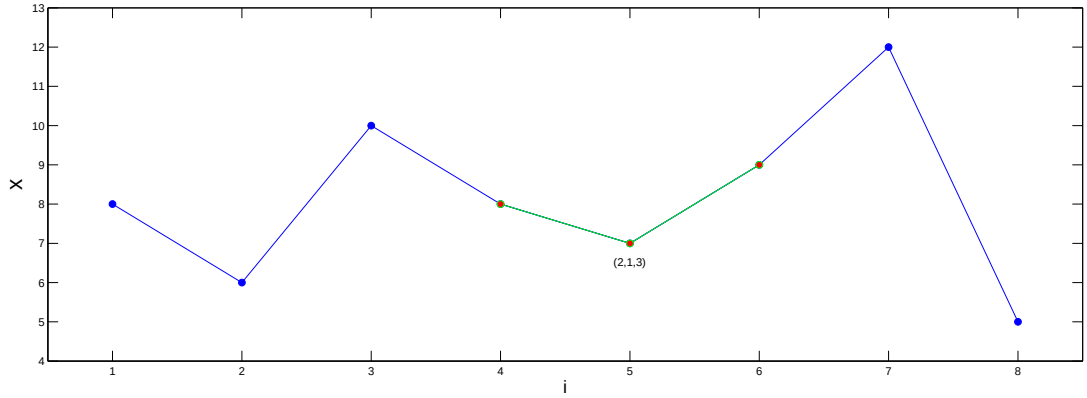
$$H_n = -\sum_{j=1}^{n!} p'_j log_2(p'_j) \tag{C.1}$$

donde $p'_j$ representa las frecuencias relativas de las secuencias de patrones de símbolos posibles.

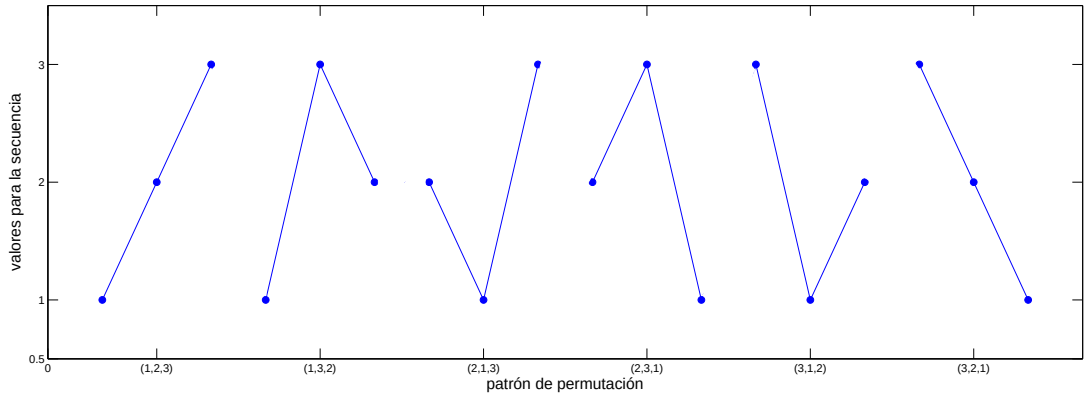
La entropía de permutación por símbolo puede ser definida de la sigueinte manera

$$h_n = -\frac{1}{(n-1)} \sum_{j=1}^{n!} p'_j log_2(p'_j) \tag{C.2}$$

El ejemplo de la figura C.1 permite observar cómo la entropía de permuación puede ser aplicada a cualquier serie de datos como lo puede ser la señal de voz.




(a) Serie de tiempo




(b) Codificación

Figura C.1: Secuencia de símbolos

En el caso del ejemplo, el orden utilizado ha sido de $n = 3$, por lo tanto, hay $n! \rightarrow 3! = 6$ secuencias de símbolo posibles. Utilizando ahora la expresión C.1, se calcula el valor de entropía para orden 3 de la secuencia [2] .

$$H_3 = -(1/6 * log_2(1/6) + 1/6 * log_2(1/6) + 1/6 * log_2(1/6) + 2/6 * log_2(2/6) + 1/6 * log_2(1/6)) \approx -(-2{,}2516) \approx 2{,}2516$$

Para el cálculo de la entropía por símbolo llegamos a la siguiente expresión:
$h_n = \frac{H_n}{2} \approx \frac{2{,}2516}{2} \approx 1{,}1258.$

## C.2. Caracterización de la complejidad de la señal de voz en el contexto de las patologías de la voz

Para estudiar el cambio en la señal de voz debidas a patologías vocales se evalúa la entropía de permutación en el caso tanto de voces sanas como con patología. Esta medida de complejidad no lineal puede distinguir entre comportamiento regular o complejo de la señal de voz y por ello puede ser útil para nuestro propósito.

Las complejidades en el caso de las patologías de la voz aparecen debido a las dinámicas no lineales de las cuerdas vocales [3] y es de esperar que la PE sea de mayor magnitud en presencia de esos comportamientos.

El cálculo de la PE se basa en la comparación de valores adyacentes desde el punto de vista del orden temporal de la señal de voz. Dicha señal de voz como sistema dinámico puede ser mapeado a una secuencia de símbolos.

## Bibliografía

[1] M. Riedl, A. Müller, and N. Wessel, "Practical considerations of permutation entropy," *The European Physical Journal Special Topics*, vol. 222, no. 2, pp. 249–262, 2013.

[2] R. G. A. N. M. J. Casanova Blancas, U., "Sistema de diagnostico aplicado a la deteccion de a apnea obstructiva del sueño mediante poligrafia," in *Proyecto Fin de Carrera*, 2014.

[3] N. Usha, V. Narayanan Namboothiri, and V. Narayanan Nampoori, *Permutation Entropy Based Analysis of Complex Signals for Characterising Change in System Dynamics*. PhD thesis, Cochin University of Science and Technology, 2008.