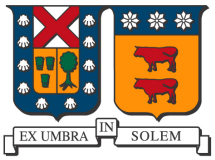
Universidad Técnica Federico Santa María

DEPARTAMENTO DE ELECTRÓNICA

VALPARAÍSO - CHILE



UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA



"Implementación en hardware de sintetizador de voz para el estudio de la calidad vocal"

Diego Agustín Ignacio Muñoz Fuentes

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRÓNICO CON MENCIÓN TELECOMUNICACIONES, SUBMENCIÓN COMPUTADORES

Profesores guía y correferente:

Dr. Matías Zañartu

Dr Juan Pablo Cortés

Fecha : 30 de octubre de 2026

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☑ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Implementación en hardware de sintetizador de voz para el estudio de la calidad vocal

Nombre del candidato(a): Diego Muñoz

Carrera / Grado: Ingeniería Civil Electrónica

Campus: CCSV **Departamento:** ELO

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Matías Zañartu, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☑ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

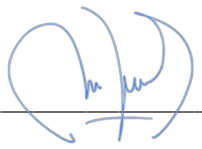
☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

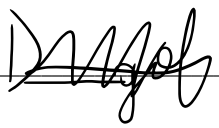
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 30/10/2026 **Firma:** 

Estudiante o Candidato(a):

Fecha: 30/10/2026 **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

Con este hito culmina mi formación superior (por ahora?), y quisiera agradecer profundamente a todas las personas que hicieron estos 7 años mas llevaderos.

A mis papá Pedro, por enseñarme a pensar, a darme cuenta que nada es tan terrible. Gracias por permitirme equivocarme y atreverme a soñar.

A mi mamá Celinda, por ser la piedra angular de la persona que soy hoy en día. Gracias por tu resiliencia, bondad y cariño que no conoce límites ni condiciones.

A mi hermana Jimena, mi compañera de maldades y la luz de mis ojos, por ser muchas veces la razón de por qué seguí adelante, aunque no lo sepas.

A mi hermana Anita, por ser mi ejemplo, mi fortaleza fuera de casa. Gracias por siempre protegerme y escucharme cuando sentía que no tenía a nadie.

Mi familia es la razón por la que pude lograr todo lo que he logrado en la vida, llevo con orgullo a mis MF a donde sea que vaya, y espero estar por siempre a la altura.

A propósito de familia, los amigos son la familia que uno mismo elige, y por lo mismo quería agradecer a todos mis amigos durante esta etapa: Seba, Felipe, Maxi, Victor, David, Esteban, Cristian, Vicente, Emilio, y más que seguro se me están pasando. Gracias por las infinitas risas, los completos, cafecitos y asados que compartimos que sin duda hicieron más amenos estos años en la U, lejos de casa.

Por último pero no menos importante, a Josefa, mi yiyis. Eres la persona más maravillosa que pude haber soñado conocer en mi etapa universitaria, gracias por tu compañía y apoyo incondicional, tu comprensión, y tu inconmensurable inteligencia. Llegaste a mi vida en el momento más difícil, para iluminarla con tu esencia y ser la mejor partner (y compañera de grupo en todos los ramos) que pude tener. Gracias por tus apuntes también, me carrearon toda la U.

Finalmente, gracias todos quienes fueron parte de mi etapa universitaria, sepan que todos fueron una pieza fundamental en llegar a este punto.

Resumen

La calidad vocal comprende los atributos perceptuales que distinguen una voz de otra más allá de su altura tonal y su intensidad. Su evaluación resulta relevante en el ámbito clínico, donde las alteraciones de la calidad vocal constituyen indicadores de patología laríngea, y en la investigación fonética, donde se requieren estímulos con atributos controlados.

Este trabajo presenta la implementación, sobre hardware de procesamiento digital de señales de bajo costo, de un sistema capaz de extraer parámetros acústicos asociados a la calidad vocal y de modificarlos de manera controlada. El sistema se desarrolló sobre la tarjeta LCDK de Texas Instruments, empleando un sintetizador de voz basado en el modelo fuente-filtro con análisis por predicción lineal de orden 18 y excitación mixta.

Se implementaron dos modificaciones correspondientes a mecanismos de fonación bien caracterizados en la literatura: la voz soplada, asociada a cierre glótico incompleto, y la voz tensa, asociada a mayor tensión laríngea. La caracterización se realizó mediante métricas acústicas calculadas con Praat sobre cuatro señales de prueba, tres vocales sostenidas y una frase de habla continua, complementada con una evaluación perceptual a ciegas.

Los resultados establecen que ambas modificaciones producen cambios medibles, consistentes entre vocales y selectivos: cada una altera principalmente la métrica asociada al mecanismo que emula. La voz tensa presenta un control lineal sobre la pendiente espectral, con coeficientes de determinación superiores a 0,999 y una dispersión entre vocales inferior al 4 %. La voz soplada presenta un control logarítmico saturante sobre la prominencia del pico cepstral, con un recorrido de hasta 8,3 dB en vocales sostenidas.

La comparación con una implementación alternativa bajo el paradigma de efecto de audio, que aplica las mismas modificaciones sin etapa de síntesis, permite establecer criterios de diseño: para modificaciones de naturaleza espectral ambos paradigmas resultan equivalentes, con una ventaja de cinco veces en costo computacional para el efecto directo, mientras que para modificaciones de la fuente glótica la síntesis paramétrica conserva un rango de control 4,8 veces mayor.

Palabras clave: calidad vocal, síntesis de voz, predicción lineal, procesamiento digital de señales, sistemas embebidos, voz soplada, voz tensa.

Abstract

Voice quality comprises the perceptual attributes that distinguish one voice from another beyond pitch and loudness. Its assessment is relevant in clinical settings, where voice quality alterations serve as indicators of laryngeal pathology, and in phonetic research, where stimuli with controlled attributes are required.

This work presents the implementation, on low-cost digital signal processing hardware, of a system capable of extracting acoustic parameters associated with voice quality and modifying them in a controlled manner. The system was developed on the Texas Instruments LCDK board, using a source-filter voice synthesizer with 18th-order linear predictive analysis and mixed excitation.

Two modifications were implemented, corresponding to phonation mechanisms well characterized in the literature: breathy voice, associated with incomplete glottal closure, and pressed voice, associated with increased laryngeal tension. Characterization was performed using acoustic metrics computed with Praat over four test signals, three sustained vowels and one continuous speech utterance, complemented by a blind perceptual evaluation.

Results establish that both modifications produce measurable, consistent and selective changes: each primarily alters the metric associated with the mechanism it emulates. Pressed voice exhibits linear control over spectral tilt, with coefficients of determination above 0,999 and inter-vowel dispersion below 4 %. Breathy voice exhibits saturating logarithmic control over cepstral peak prominence, spanning up to 8,3 dB in sustained vowels.

Comparison with an alternative implementation under the audio effect paradigm, applying the same modifications without a synthesis stage, allows design criteria to be established: for spectral modifications both paradigms are equivalent, with a fivefold computational advantage for the direct effect, whereas for glottal source modifications parametric synthesis retains a 4,8-fold larger control range.

Keywords: voice quality, speech synthesis, linear prediction, digital signal processing, embedded systems, breathy voice, pressed voice.

Glosario de siglas

Sigla	Significado
ADC	*Analog-to-Digital Converter*, conversor analógico-digital
APLP	*Adaptive Pre-emphasis Linear Prediction*
CCS	*Code Composer Studio*, entorno de desarrollo de Texas Instruments
CPP	*Cepstral Peak Prominence*, prominencia del pico cepstral
CPPS	*Cepstral Peak Prominence, Smoothed*, variante suavizada del CPP
DAC	*Digital-to-Analog Converter*, conversor digital-analógico
dBFS	Decibelios relativos a fondo de escala (*full scale*)
DF2T	*Direct Form II Transposed*, forma directa II transpuesta
DSP	*Digital Signal Processor*, procesador digital de señales
EDMA	Enhanced Direct Memory Access, acceso directo a memoria mejorado
EGG	Electroglotografía
GEL	General Extension Language, formato de script de configuración de CCS
GNE	*Glottal to Noise Excitation ratio*
HNR	*Harmonics-to-Noise Ratio*, relación armónicos-ruido
LCDK	*Low Cost Development Kit*, plataforma de desarrollo empleada
LF	Liljencrants-Fant, modelo de flujo glótico
LFSD	*Low Frequency Spectral Density*
LFSR	*Linear Feedback Shift Register*, generador de ruido pseudoaleatorio
LPC	*Linear Predictive Coding*, codificación por predicción lineal
LSD	*Log-Spectral Distortion*, distorsión log-espectral
McASP	Multichannel Audio Serial Port, puerto serie de audio multicanal
MELP	*Mixed Excitation Linear Prediction*
NSA	*Neck Surface Accelerometer*, acelerómetro de superficie cervical
OQ	*Open Quotient*, cociente de apertura glótica
PARCOR	*Partial Correlation*, coeficientes de reflexión
PGA	Programmable Gain Amplifier, amplificador de ganancia programable
QCP	*Quasi-Closed Phase*, método de filtrado inverso glótico
RBH	*Rauigkeit, Behauchtheit, Heiserkeit*: escala perceptual de aspereza, soplo y ronquera
SNR	*Signal-to-Noise Ratio*, relación señal-ruido
ZFF	*Zero Frequency Filtering*, filtrado de frecuencia cero

Notación

Símbolo	Significado
$s[n]$	señal de voz
$u[n]$	señal de excitación
$e[n]$	residual de predicción
$A(z)$	polinomio de predicción lineal
a_k	coeficientes de predicción lineal
k_i	coeficientes de reflexión (PARCOR)
p	orden del modelo de predicción lineal
G	ganancia del filtro de síntesis
F_0	frecuencia fundamental
T_0	período fundamental, en muestras
f_s	frecuencia de muestreo
$R(\tau)$	función de autocorrelación
$\rho(\tau)$	autocorrelación normalizada
$c[q]$	cepstrum real
q	*quefrency*
α	coeficiente de pre-énfasis
γ	factor de ensanchamiento espectral
β	coeficiente de mezcla de la excitación
k	parámetro de intensidad del efecto
OQ	cociente de apertura
N	longitud de la ventana de análisis, en muestras

Índice

Resumen **3**

Abstract **4**

Glosario de siglas **5**

Notación **6**

Índice de figuras **12**

Índice de tablas **13**

1. Capítulo 1: Introducción y Objetivos **14**
1.1. Resumen . . . 14
1.2. Motivación . . . 14
1.3. Objetivos . . . 14
1.3.1. Objetivo general . . . 14
1.3.2. Objetivos específicos . . . 15
1.4. Alcances y limitaciones . . . 15
1.5. Estructura del documento . . . 15

2. Capítulo 2: Estado del Arte **16**
2.1. El modelo fuente-filtro . . . 16
2.1.1. Modelos de pulso glótico . . . 16
2.1.2. El cociente de apertura como parámetro de control . . . 16
2.2. Calidad vocal y sus correlatos acústicos . . . 17
2.2.1. Voz soplada . . . 17
2.2.2. Voz tensa . . . 17
2.2.3. Medidas cepstrales . . . 18
2.2.4. Modelos psicoacústicos integrados . . . 18
2.3. Estimación de la fuente glótica . . . 19
2.3.1. Clasificación automática de tipos de fonación . . . 19
2.3.2. Comparación de aproximaciones y criterio de selección . . . 19
2.4. Síntesis y modificación de la calidad vocal . . . 21
2.4.1. Vocoders LPC clásicos y el problema del *buzziness* . . . 21
2.4.2. Excitación mixta . . . 21
2.4.3. Vocoders modernos . . . 21
2.4.4. Modificación deliberada de la calidad vocal . . . 22
2.5. Análisis por predicción lineal . . . 22
2.6. Implementaciones en hardware dedicado . . . 22
2.7. Síntesis del estado del arte . . . 23

3. Capítulo 3: Marco Teórico **24**
3.1. Modelo de producción de voz . . . 24
3.1.1. Formulación del modelo fuente-filtro . . . 24
3.1.2. Excitación glótica . . . 24
3.2. Análisis por predicción lineal . . . 25
3.2.1. Formulación . . . 25
3.2.2. Ecuaciones normales . . . 25
3.2.3. Recursión de Levinson-Durbin . . . 25

3.2.4. Estabilidad y ancho de banda de los polos . . . 26
3.2.5. Pre-énfasis . . . 26
3.3. Estimación de la frecuencia fundamental . . . 26
3.3.1. Método de autocorrelación . . . 26
3.3.2. Normalización y sesgo . . . 27
3.3.3. Errores de octava . . . 27
3.4. Modelos de excitación para síntesis . . . 28
3.4.1. Tren de impulsos . . . 28
3.4.2. Excitación mixta . . . 28
3.5. Métricas acústicas de calidad vocal . . . 29
3.5.1. Cepstrum . . . 29
3.5.2. Prominencia del pico cepstral . . . 29
3.5.3. Relación armónicos-ruido . . . 30
3.5.4. Pendiente espectral . . . 30
3.5.5. Diferencia de amplitud entre los primeros armónicos . . . 30
3.5.6. Medidas de perturbación . . . 31
3.6. Distorsión log-espectral . . . 31
3.7. Filtrado digital . . . 31
3.7.1. Tipos de filtro empleados . . . 31
3.7.2. Estructura de realización . . . 33
3.7.3. Filtro todo-polos de orden arbitrario . . . 33
3.7.4. Compresión de rango dinámico . . . 33

4. Capítulo 4: Diseño e Implementación 34
4.1. Plataforma de hardware . . . 34
4.1.1. Selección de la plataforma . . . 34
4.1.2. Características del hardware . . . 35
4.1.3. Entorno de desarrollo . . . 35
4.1.4. Bibliotecas de apoyo . . . 36
4.2. Flujo de datos y cadena de procesamiento . . . 37
4.2.1. Cadena de adquisición . . . 37
4.2.2. Configuración del códec mediante I^2C . . . 37
4.2.3. Modo de transferencia y atención de muestras . . . 38
4.2.4. Operaciones realizadas en la rutina de interrupción . . . 38
4.2.5. Medición del tiempo de ejecución . . . 39
4.3. Arquitectura del sistema . . . 39
4.3.1. Decisión de diseño: procesamiento por buffer . . . 39
4.3.2. Organización modular del código . . . 40
4.4. Etapa de análisis . . . 40
4.4.1. Normalización y pre-énfasis . . . 40
4.4.2. Estimación de la frecuencia fundamental . . . 41
4.4.3. Análisis por predicción lineal . . . 41
4.5. Etapa de síntesis . . . 42
4.5.1. Modelo de excitación . . . 42
4.5.2. Filtro de síntesis . . . 43
4.6. Modificaciones de calidad vocal . . . 43
4.6.1. Voz soplada . . . 43
4.6.2. Voz tensa . . . 43
4.6.3. Parámetro de intensidad . . . 44
4.7. Implementación alternativa mediante efectos de audio . . . 44

4.8. Metodología de medición . . . 45
4.8.1. Material de prueba . . . 45
4.8.2. Métricas acústicas . . . 45
4.8.3. Verificación cruzada de la implementación . . . 46
4.8.4. Evaluación perceptual . . . 46

5. Capítulo 5: Resultados **47**
5.1. Verificación funcional del sintetizador . . . 48
5.1.1. Validación de la cadena de análisis-síntesis . . . 48
5.1.2. Comportamiento con habla continua . . . 48
5.2. Modificación de la calidad vocal . . . 49
5.2.1. Selectividad del control . . . 50
5.2.2. Diferencia de amplitud entre los primeros armónicos . . . 50
5.3. Caracterización del control paramétrico . . . 51
5.3.1. Voz tensa: control lineal . . . 51
5.3.2. Breathiness: control logarítmico saturante . . . 52
5.3.3. Dependencia del material fonético . . . 53
5.4. Evaluación perceptual . . . 53
5.4.1. Limitaciones de la evaluación . . . 53
5.5. Costo computacional y comparación entre paradigmas . . . 54
5.5.1. Descomposición del costo de análisis . . . 54
5.5.2. Comparación entre paradigmas de implementación . . . 54
5.5.3. Equivalencia y diferencias en el control . . . 55
5.5.4. Comparación con un vocoder de referencia . . . 56

6. Capítulo 6: Conclusiones **57**
6.1. Cumplimiento de los objetivos . . . 57
6.1.1. Objetivo específico 1 . . . 57
6.1.2. Objetivo específico 2 . . . 57
6.1.3. Objetivo general . . . 57
6.2. Principales hallazgos . . . 57
6.2.1. Las dos modificaciones siguen leyes de control distintas . . . 57
6.2.2. El rango de control depende del material fonético . . . 58
6.2.3. Divergencia entre medición objetiva y percepción . . . 58
6.2.4. Criterios para la elección del paradigma de implementación . . . 58
6.2.5. Modelo de costo computacional . . . 59
6.3. Limitaciones . . . 59
6.3.1. Calidad de la síntesis . . . 59
6.3.2. Alcance de la evaluación perceptual . . . 59
6.3.3. Material de prueba . . . 59
6.3.4. Operación fuera de tiempo real . . . 60
6.4. Trabajo futuro . . . 60
6.4.1. Excitación mixta multibanda . . . 60
6.4.2. Optimización hacia tiempo real . . . 60
6.4.3. Ampliación de la evaluación perceptual . . . 60
6.4.4. Incorporación de otras dimensiones de calidad vocal . . . 60
6.4.5. Validación con material clínico . . . 60
6.4.6. Migración a una plataforma de mayor capacidad de cómputo . . . 61
6.5. Comentarios finales . . . 61

Anexos **65**
Anexo A: Parámetros de configuración 65
Anexo B: Script de análisis acústico 66
Anexo C: Resultados del barrido 67
Anexo D: Evaluación perceptual 68
Anexo E: Señales temporales y espectros 68

Índice de figuras

1. Modelo fuente-filtro de la producción de voz. Los tres bloques de la derecha se agrupan en un único filtro todo-polos durante el análisis por predicción lineal. 24
2. Estimación de F_0 sobre habla continua. Al tomar el máximo global del intervalo de búsqueda, las tramas de baja periodicidad reciben estimaciones espurias en el extremo del rango. La restricción a máximos locales estrictos las descarta correctamente. 27
3. Componentes de la excitación mixta. El tren de pulsos (a) se compensa en continua; el ruido (b) se escala para igualar su energía por período a la del pulso; la mezcla (c) se pondera según el grado de periodicidad medido en la trama. 29
4. Montaje experimental. El computador programa y depura la LCDK a través del emulador XDS110 mediante USB, y reproduce las señales de prueba hacia la entrada de línea de la placa mediante un cable de audio. 35
5. Flujo de datos en la plataforma. La configuración del códec se realiza una única vez mediante I^2C; la transferencia de muestras ocurre de forma continua a través del McASP, y cada muestra genera una interrupción atendida por el procesador. 37
6. Máquina de estados que gobierna la captura y el procesamiento. Las transiciones se disparan por pulsador, salvo la de captura completa, que se produce al llenarse el buffer. 39
7. Arquitectura del sistema. La etapa de análisis extrae los parámetros de la señal capturada; la de síntesis los emplea para reconstruirla, aplicando a continuación la modificación de calidad vocal seleccionada. 40
8. Envolvente espectral estimada por el filtro LPC de orden 18 sobre una trama de la vocal /a/, superpuesta al espectro de la trama. 42
9. Arquitectura del paradigma de efecto de audio. A diferencia del sintetizador, no existe etapa de análisis ni de resíntesis: la forma de onda original se conserva y solo se modifica. 45
10. Esquema del flujo experimental completo, desde la captura de las señales de prueba hasta la caracterización final del sistema. 47
11. Comparación de las formas de onda de entrada y sintetizada sobre un mismo tramo de la vocal /a/. 48
12. Espectrogramas de la señal de entrada y de las tres condiciones sintetizadas para la vocal /a/. 49
13. Envolvente espectral promediada de cada condición. La modificación *pressed* desplaza energía hacia las frecuencias altas, mientras que *breathy* produce el efecto opuesto. 50
14. Variación de la pendiente espectral en función de la intensidad del efecto de voz tensa. El comportamiento es lineal en las tres vocales. 52
15. Variación del CPPS en función de la intensidad del efecto de breathiness. El control satura progresivamente. 52
16. Costo computacional de ambos paradigmas. 55
17. Comparación del control paramétrico entre el sintetizador (línea continua) y el efecto directo (línea discontinua). 55
18. Comparación de periodicidad entre la señal de entrada y las reconstrucciones de ambos sistemas. 56

19. Vocal /a/ procesada mediante el sintetizador. De arriba abajo: señal de entrada, reconstrucción sin modificación, modificación de voz soplada y modificación de voz tensa. 69
20. Vocal /i/ procesada mediante el sintetizador. 70
21. Vocal /u/ procesada mediante el sintetizador. 71
22. Frase de habla continua procesada mediante el sintetizador. 72
23. Vocal /a/ procesada mediante efectos aplicados directamente sobre la señal capturada. 73
24. Vocal /i/ procesada mediante efectos directos. 73
25. Vocal /u/ procesada mediante efectos directos. 74
26. Frase de habla continua procesada mediante efectos directos. 74

Índice de cuadros

1. Comparación de las aproximaciones a la síntesis de voz consideradas. . 20
2. Características de la plataforma de desarrollo. 35
3. Registros del códec configurados por el sistema. 37
4. Métricas de calidad vocal a intensidad nominal ($k = 1$). 49
5. Diferencia H1–H2 (dB) y variación respecto de la condición neutral. . . 50
6. Tiempo total de procesamiento por paradigma (C6748 a 300 MHz, capturas de 5 segundos). 54
7. Comparación con WORLD sobre las mismas señales de entrada. 56
8. Configuración del códec y de la plataforma. 65
9. Parámetros del análisis y la síntesis. 65
10. Parámetros del efecto de voz soplada. 65
11. Parámetros del efecto de voz tensa. 66
12. CPPS (dB) por nivel de intensidad, modificación de voz soplada, sintetizador. 67
13. Pendiente espectral (dB/oct) por nivel de intensidad, modificación de voz tensa, sintetizador. 67
14. Desglose del tiempo de procesamiento por etapa (ms). Capturas de 5 segundos sobre C6748 a 300 MHz. 67
15. Tiempo de procesamiento del paradigma de efecto directo (ms). 67
16. Resultados de la prueba de escucha a ciegas. Los estímulos se presentaron en orden aleatorio; la columna *condición* indica la correspondencia real, desconocida por el evaluador durante la prueba. 68

1. Capítulo 1: Introducción y Objetivos

1.1. Resumen

Este trabajo aborda la implementación en hardware de un sistema capaz de extraer parámetros acústicos asociados a la calidad vocal de una señal de voz y de modificarlos de manera controlada. El sistema se desarrolló sobre una plataforma de procesamiento digital de señales de bajo costo, la tarjeta LCDK de Texas Instruments, empleando un sintetizador de voz basado en el modelo fuente-filtro.

Se implementaron dos modificaciones de calidad vocal, correspondientes a dos mecanismos de fonación bien caracterizados en la literatura: la voz soplada, asociada a un cierre glótico incompleto que introduce ruido de aspiración, y la voz tensa, asociada a un aumento de la tensión laríngea que aplana la pendiente espectral.

Los resultados muestran que ambas modificaciones producen cambios medibles, consistentes entre vocales y selectivos, en el sentido de que cada una altera principalmente la métrica acústica asociada al mecanismo que emula sin afectar significativamente a la otra. La caracterización del control paramétrico reveló además que ambos efectos siguen leyes de control distintas: lineal en el caso de la voz tensa y logarítmica saturante en el de la voz soplada.

Como contribución adicional, se desarrolló una segunda implementación bajo el paradigma de efecto de audio, que aplica las mismas modificaciones directamente sobre la señal capturada sin etapa de síntesis. La comparación entre ambos paradigmas permite establecer criterios de diseño respecto de cuándo la síntesis paramétrica resulta necesaria y cuándo un efecto de audio es suficiente.

1.2. Motivación

La calidad vocal, entendida como el conjunto de atributos perceptuales que distinguen una voz de otra más allá de su altura tonal y su intensidad, constituye un objeto de estudio relevante tanto en fonética como en el ámbito clínico. Alteraciones en la calidad vocal son indicadores de patologías laríngeas, y su evaluación forma parte del protocolo habitual de diagnóstico.

La evaluación perceptual de la calidad vocal presenta, sin embargo, una variabilidad considerable entre evaluadores [1], lo que ha motivado la búsqueda de correlatos acústicos objetivos y de herramientas capaces de generar estímulos controlados para la investigación perceptual.

La síntesis de voz con calidad vocal controlada responde a esta segunda necesidad: permite generar estímulos en los que un único atributo varía de forma conocida y graduada, manteniendo constantes los demás. La implementación de estas capacidades sobre hardware dedicado, en lugar de software de propósito general, abre la posibilidad de sistemas portátiles y de baja latencia aplicables en contextos clínicos o experimentales.

1.3. Objetivos

1.3.1. Objetivo general

Implementar en hardware algoritmos de procesamiento de señal a una señal de audio, en particular la voz humana, que permitan obtener parámetros propios de la calidad vocal y modificarlos deliberadamente para alterar la calidad vocal de la señal.

1.3.2. Objetivos específicos

OE1 Implementar en hardware algoritmos de procesamiento digital de señal para extraer parámetros relevantes de la calidad vocal a partir de una señal de voz.

OE2 Implementar algoritmos que permitan modificar de manera controlada dichos parámetros para alterar suavemente la calidad vocal de la señal reconstruida.

1.4. Alcances y limitaciones

El sistema desarrollado opera sobre señales de voz muestreadas a 16 kHz y procesa fragmentos de hasta 5 segundos de duración. No opera en tiempo real: el procesamiento se realiza sobre buffers completos, decisión de diseño cuya justificación se desarrolla en la Sección 4.3 y se sustenta cuantitativamente en los resultados de costo computacional de la Sección 5.5.

Se implementaron dos de las dimensiones de calidad vocal descritas en la literatura, voz soplada y voz tensa, seleccionadas por contar con correlatos acústicos bien establecidos y mecanismos de producción diferenciados.

La evaluación perceptual realizada tiene carácter exploratorio: involucró un único evaluador sin repeticiones, por lo que sus resultados se presentan como indicativos de tendencias y no como conclusiones estadísticamente sustentadas.

1.5. Estructura del documento

El Capítulo 2 revisa los fundamentos de la producción de voz, los correlatos acústicos de la calidad vocal y las técnicas de síntesis relevantes para este trabajo. El Capítulo 3 desarrolla el marco teórico: el modelo fuente-filtro, el análisis por predicción lineal, la estimación de la frecuencia fundamental, los modelos de excitación, las métricas acústicas empleadas y las estructuras de filtrado digital sobre las que se construyen las modificaciones. El Capítulo 4 describe la plataforma empleada, la arquitectura del sistema, los algoritmos implementados y la metodología de medición. El Capítulo 5 presenta los resultados de la caracterización del sistema, incluyendo la verificación funcional, la caracterización del control paramétrico, la evaluación perceptual y la comparación entre paradigmas de implementación. El Capítulo 6 sintetiza los principales hallazgos, discute las limitaciones del trabajo y propone líneas de trabajo futuro.

2. Capítulo 2: Estado del Arte

Este capítulo revisa los antecedentes sobre los que se sustenta el trabajo. Se organiza en cinco líneas: el modelo de producción de voz adoptado, la caracterización acústica de la calidad vocal, las técnicas de estimación de la fuente glótica, las aproximaciones a la síntesis y modificación de voz, y las consideraciones propias de la implementación en hardware dedicado.

2.1. El modelo fuente-filtro

La producción de voz se describe habitualmente mediante el modelo fuente-filtro, que separa el proceso en dos componentes acústicamente independientes: una fuente de excitación generada en la laringe y un filtro resonante constituido por el tracto vocal [2].

Durante la fonación sonora, el flujo de aire procedente de los pulmones hace vibrar los pliegues vocales, produciendo un tren de pulsos de flujo glótico cuya frecuencia de repetición corresponde a la frecuencia fundamental F_0. Este flujo se filtra al atravesar la cavidad faringo-oral, cuyas resonancias, denominadas formantes, determinan el timbre vocálico. Finalmente, la radiación en los labios introduce un realce adicional de aproximadamente +6 dB por octava.

La separación entre fuente y filtro es la que hace posible modificar la calidad vocal sin alterar el contenido fonético: los atributos de calidad vocal residen predominantemente en las características de la fuente, mientras que la identidad vocálica reside en el filtro.

2.1.1. Modelos de pulso glótico

La forma de onda del pulso glótico no es impulsiva sino que presenta una estructura temporal característica, con una fase de apertura gradual y una fase de cierre más abrupta. Rosenberg [3] demostró experimentalmente que la forma del pulso glótico afecta de manera perceptualmente significativa a la calidad de las vocales sintetizadas, y propuso modelos paramétricos para representarla.

El modelo de Liljencrants-Fant [4], comúnmente denominado modelo LF, constituye la formulación más extendida. Describe la derivada del flujo glótico mediante cuatro parámetros que controlan la duración de la fase abierta, la abruptez del cierre y la fase de retorno, permitiendo representar de forma continua el espacio entre fonaciones sopladas y tensas.

2.1.2. El cociente de apertura como parámetro de control

El cociente de apertura (*open quotient*, OQ), definido como la fracción del ciclo glótico durante la cual los pliegues permanecen separados, ocupa un lugar central en la descripción de los tipos de fonación: la voz soplada se asocia a valores elevados de OQ y la voz tensa a valores reducidos.

Sin embargo, la relación entre este parámetro fisiológico y sus manifestaciones acústicas no es tan directa como suele suponerse. Kreiman y colaboradores [5] examinaron empíricamente la relación entre el OQ, la diferencia de amplitud entre los dos primeros armónicos (H1*–H2*) y la calidad soplada percibida, midiendo simultáneamente la señal acústica e imágenes laríngeas de alta velocidad en hablantes sanos que variaban la frecuencia fundamental y la calidad vocal de forma cuasi-ortogonal. Sus

resultados indican que el respaldo empírico para la cadena causal habitualmente asumida entre estos tres niveles descriptivos es más limitado de lo que la literatura suele reflejar, y que las relaciones reportadas presentan una variabilidad considerable.

2.2. Calidad vocal y sus correlatos acústicos

La calidad vocal comprende los atributos perceptuales que distinguen a una voz de otra manteniendo constantes su altura tonal e intensidad. Constituye un rasgo tan distintivo del individuo como el rostro o las huellas dactilares, y a partir de ella se realizan inferencias paralingüísticas que abarcan desde rasgos de personalidad hasta condiciones de salud [6].

Su caracterización presenta, no obstante, una dificultad de fondo: no existe una definición objetiva universalmente aceptada. La calidad vocal se describe habitualmente mediante descriptores subjetivos como *áspera* o *soplada* [7], y su evaluación clínica se realiza mediante escalas perceptuales cuya fiabilidad entre evaluadores presenta una variabilidad ampliamente documentada [1]. Esta situación ha motivado la búsqueda sistemática de correlatos acústicos objetivos, y revisiones recientes recopilan y comparan las medidas propuestas con esa finalidad [6, 7].

2.2.1. Voz soplada

La voz soplada resulta de un cierre glótico incompleto que permite un flujo de aire turbulento durante todo el ciclo. Klatt y Klatt [8] caracterizaron sus correlatos acústicos mediante experimentos de análisis, síntesis y percepción sobre imitaciones reiterantes producidas por diez hablantes femeninas y seis masculinos. Identificaron como claves acústicas potenciales el incremento de la amplitud relativa del primer armónico conforme aumenta el cociente de apertura, el aumento del ruido de aspiración que reemplaza a los armónicos de frecuencias altas a medida que los aritenoides se separan, y modificaciones en la inclinación espectral.

Hillenbrand y colaboradores [9] evaluaron sistemáticamente qué medidas acústicas predicen mejor las valoraciones perceptuales de calidad soplada. Emplearon grabaciones de ocho hombres y siete mujeres sin patología vocal produciendo vocales sostenidas en tres grados de soplo, valoradas por veinte oyentes mediante estimación directa de magnitud. Compararon medidas de periodicidad de la señal, de amplitud del primer armónico y de inclinación espectral, concluyendo que las medidas de periodicidad ofrecen las predicciones más precisas, explicando aproximadamente el 80 % de la varianza de las valoraciones perceptuales. Trabajos posteriores extendieron el análisis a voces disfónicas y a habla continua [10].

Fröhlich y colaboradores [11] abordaron el problema desde la perspectiva de la discriminación entre mecanismos de fonación patológicos. Compararon siete medidas de calidad soplada descritas en la literatura junto con una medida propia, la razón de excitación glotal a ruido (GNE), y encontraron que únicamente esta última permitía distinguir entre todos los grupos patológicos considerados y los grupos de referencia normal y afónico.

2.2.2. Voz tensa

La voz tensa se asocia a un aumento de la tensión laríngea y de la presión subglótica, que produce un cierre glótico más abrupto y completo. Acústicamente se manifiesta principalmente como un aplanamiento de la pendiente espectral, con redistribución de energía hacia las frecuencias altas.

Gowda y Kurimo [12] propusieron un enfoque complementario basado en la densidad espectral de baja frecuencia (LFSD). Su razonamiento parte de que la fonación soplada, al presentar un cociente de apertura elevado, produce una mayor influencia de la cavidad subglótica sobre el espectro estimado, lo que se refleja como un incremento de la densidad espectral por debajo de la primera resonancia del tracto vocal. La fonación tensa, con menor influencia subglótica, presenta en cambio una densidad reducida en esa región. Emplearon esta observación como característica para el análisis y la clasificación de los tres tipos de fonación.

Garellek y colaboradores [13] propusieron un modelo psicoacústico de la fuente vocal caracterizado por cuatro parámetros de pendiente espectral: H1–H2, H2–H4, H4–2 kHz y 2 kHz–5 kHz. Mediante experimentos de ajuste del modelo a un conjunto amplio de fuentes vocales establecieron la independencia acústica y perceptual de estos parámetros, lo que los habilita como dimensiones de control separables.

2.2.3. Medidas cepstrales

De Krom [14] propuso un método para calcular la relación armónicos-ruido espectral mediante una operación de *comb-liftering* en el dominio cepstral, evaluando su sensibilidad frente al ruido aditivo y al *jitter* sobre señales sintéticas de tipo vocálico generadas a distintas frecuencias fundamentales, y analizando el efecto de la longitud de la ventana de análisis sobre los valores estimados.

De esta línea de trabajo deriva la familia de medidas de prominencia del pico cepstral. Estas cuantifican cuánto sobresale el pico cepstral correspondiente al período fundamental respecto de la tendencia general del cepstrum: una señal fuertemente periódica produce un pico prominente, mientras que la presencia de ruido lo atenúa.

La variante suavizada, denominada CPPS, se ha consolidado como el correlato acústico más robusto de la calidad soplada y de la severidad de la disfonía. Lopes y colaboradores [15] analizaron su asociación con la presencia, intensidad y tipo de alteración vocal sobre 376 muestras de individuos con quejas vocales, empleando análisis auditivo-perceptual mediante escala analógico-visual para el grado global de disfonía y para los grados de aspereza, soplo y tensión. Szklanny y colaboradores [16] estudiaron la correspondencia entre parámetros acústicos y electroglotográficos y la escala perceptual RBH en población infantil con nódulos vocales, construyendo para ello una base de datos dedicada.

2.2.4. Modelos psicoacústicos integrados

Kreiman y colaboradores [17] abordaron la ausencia de un método consensuado para la medición objetiva de la calidad vocal percibida mediante la validación de un modelo psicoacústico integral. El modelo incorpora parámetros que caracterizan las fuentes vocales armónica e inarmónica, la función de transferencia del tracto vocal, la frecuencia fundamental y la amplitud, con el propósito de cuantificar de forma completa el sonido integral de una muestra de voz. En su validación, doscientas voces con y sin patología diagnosticada fueron ajustadas al modelo mediante análisis por síntesis.

Esta línea de trabajo reconoce que ninguna medida acústica aislada captura completamente la percepción de calidad vocal. En la misma dirección, Klasmeyer [18] estudió la importancia perceptual relativa de distintos parámetros de calidad vocal, centrándose en el esfuerzo vocal y el tipo de fonación, y señalando que estos perceptos resultan de la combinación de varios parámetros acústicos: la forma del pulso glótico en el dominio temporal o el amortiguamiento de los armónicos en el dominio frecuencial, junto con la distribución espectral.

2.3. Estimación de la fuente glótica

La descomposición fuente-tracto, o estimación del flujo glótico, constituye uno de los problemas básicos del procesamiento de voz. Se han propuesto múltiples técnicas, si bien los estudios comparativos entre ellas son escasos y se han realizado sistemáticamente sobre voz sintética o vocales sostenidas.

Drugman y colaboradores [19] compararon tres de los métodos representativos del estado del arte: el filtrado inverso de fase cerrada, el filtrado inverso iterativo y adaptativo, y la descomposición de fase mixta. Sometieron estas técnicas a evaluación objetiva sobre señales sintéticas y analizaron su sensibilidad frente a diversos factores.

Las características derivadas de la fuente glótica han demostrado utilidad para la detección automática de patología vocal. Kadiri y Alku [20] realizaron un análisis sistemático de estas características, extrayéndolas mediante filtrado inverso glótico de fase cuasi-cerrada (QCP), mediante filtrado de frecuencia cero (ZFF), y directamente de la señal acústica, proponiendo además derivar coeficientes cepstrales en escala mel a partir del flujo glótico estimado.

2.3.1. Clasificación automática de tipos de fonación

La clasificación automática de calidad vocal ha empleado tradicionalmente máquinas de vectores de soporte sobre características glóticas diseñadas manualmente. Kadiri y colaboradores [21] examinaron la clasificación de las tres calidades vocales de interés en este trabajo —soplada, modal y tensa— empleando redes neuronales convolucionales sobre señales de voz y de acelerómetro de superficie cervical registradas simultáneamente. Sus resultados indican que la señal del acelerómetro resulta más informativa que la señal acústica para esta tarea, y que los clasificadores basados en aprendizaje profundo superan con margen amplio a los basados en vectores de soporte.

La existencia de esta línea de trabajo resulta pertinente para el presente desarrollo en la medida en que un sistema capaz de generar estímulos con calidad vocal controlada proporciona material sintético de propiedades conocidas, potencialmente útil para el entrenamiento y la validación de tales clasificadores.

2.3.2. Comparación de aproximaciones y criterio de selección

Las aproximaciones revisadas presentan compromisos distintos entre calidad de reconstrucción, control paramétrico y costo computacional. La Tabla 1 resume esta comparación.

Cuadro 1: Comparación de las aproximaciones a la síntesis de voz consideradas.

Aproximación	Ventajas	Limitaciones	Control de calidad vocal
Concatenativa	Calidad natural, sin degradación de resíntesis	Requiere corpus extenso; no permite modificar la fuente	Nulo: solo puede seleccionar entre unidades existentes
Formántica (tipo Klatt)	Control explícito de todos los parámetros del modelo	Requiere estimar decenas de parámetros; la calidad depende críticamente de esa estimación	Alto, pero la estimación automática desde señal real es poco robusta
LPC con excitación simple	Formulación cerrada y eficiente; separación fuente-filtro directa	Carácter mecánico (*buzziness*); modelo de excitación pobre	Directo sobre la excitación y sobre la envolvente espectral
LPC con excitación mixta (tipo MELP)	Mitiga el *buzziness* manteniendo la estructura del LPC	Mayor costo; requiere estimar la periodicidad por banda	Directo, con control adicional sobre el balance periódico/aperiódico
Vocoders de análisis-síntesis (tipo WORLD)	Máxima fidelidad de reconstrucción	Costo computacional y de memoria elevado; múltiples subsistemas	Indirecto: requiere manipular las representaciones intermedias
Neuronales	Calidad cercana a la natural	Requieren entrenamiento y recursos de cómputo incompatibles con hardware embebido	Indirecto y dependiente del espacio latente aprendido

La selección de la aproximación para este trabajo atendió a tres criterios, derivados de los objetivos planteados y de las restricciones de la plataforma.

Acceso paramétrico a la fuente glótica. El objetivo específico OE2 requiere modificar de manera controlada los parámetros de calidad vocal. Dado que dichos atributos residen predominantemente en la fuente (Sección 2.1), la aproximación elegida debe exponer la excitación como una señal manipulable de forma explícita. Esta condición descarta las aproximaciones concatenativa y neuronal, y favorece a las basadas en el modelo fuente-filtro, en las que la excitación es un objeto de primera clase.

Viabilidad sobre hardware embebido. La plataforma dispone de 256 kB de memoria interna y un procesador a 300 MHz, sin bibliotecas de cálculo numérico. Ello descarta las aproximaciones neuronales y hace costosa la implementación completa de un vocoder de análisis-síntesis moderno, que requiere múltiples transformadas rápidas de Fourier por trama y aritmética de doble precisión. La predicción lineal, en cambio, se resuelve mediante una recursión de complejidad $O(p^2)$ sobre un orden reducido, y su filtro de síntesis es una simple estructura recursiva.

Robustez de la estimación automática. La síntesis formántica ofrece un control paramétrico superior, pero requiere estimar automáticamente las frecuencias, anchos de banda y amplitudes de cada formante a partir de la señal, problema cuya solución robusta excede el alcance de este trabajo. La predicción lineal entrega una estimación

del filtro del tracto vocal mediante una formulación cerrada y numéricamente estable, sin requerir la identificación individual de las resonancias.

De la combinación de estos criterios se desprende la elección adoptada: un **modelo fuente-filtro con análisis por predicción lineal y excitación mixta**. Esta aproximación proporciona acceso directo a la excitación, que es donde se aplican las modificaciones de calidad vocal; resulta implementable sobre la plataforma disponible; y su principal limitación conocida, el carácter mecánico de la excitación por tren de impulsos, admite mitigación mediante la incorporación de una componente aperiódica ponderada, tal como se detalla en la Sección 3.4.

La contrapartida de esta elección, una fidelidad de reconstrucción inferior a la de un vocoder de análisis-síntesis moderno, se cuantifica explícitamente en la Sección 5.5 mediante la comparación con WORLD.

2.4. Síntesis y modificación de la calidad vocal

2.4.1. Vocoders LPC clásicos y el problema del *buzziness*

Los vocoders basados en predicción lineal reconstruyen la señal excitando el filtro estimado con un tren de impulsos periódico durante los segmentos sonoros y con ruido blanco durante los sordos. Este esquema produce habla inteligible pero de calidad marcadamente mecánica, fenómeno conocido en la literatura como *buzziness*.

Sambur y colaboradores [22] estudiaron específicamente este problema y demostraron que el empleo de una fuente no impulsiva, con parámetros temporales escalados en proporción al período fundamental, produce una preferencia perceptual significativa respecto de la síntesis LPC estándar.

2.4.2. Excitación mixta

La aproximación que se consolidó para mitigar el *buzziness* consiste en combinar componentes periódicas y aperiódicas en la señal de excitación. McCree y Barnwell [23] desarrollaron el modelo de excitación mixta que dio origen al vocoder MELP, posteriormente adoptado como estándar federal de codificación de voz a 2400 bits por segundo.

MELP incorpora cinco elementos sobre el modelo LPC tradicional: excitación mixta calculada independientemente en cinco bandas de frecuencia, pulsos aperiódicos, realce espectral adaptativo, dispersión de pulso y modelado de las magnitudes de Fourier del residual. El elemento determinante para la reducción del carácter mecánico es la excitación mixta multibanda, que reconoce que el grado de periodicidad de la voz real varía con la frecuencia.

2.4.3. Vocoders modernos

WORLD [24] representa el estado del arte en vocoders de alta calidad basados en análisis-síntesis. Fue desarrollado con el propósito explícito de mejorar la calidad sonora en aplicaciones de tiempo real, dado que los sistemas de alta calidad previos presentaban dificultades para el procesamiento en tiempo real debido a su elevado costo computacional. Descompone la señal mediante tres algoritmos de análisis y uno de síntesis: DIO para la estimación de la frecuencia fundamental, CheapTrick para la envolvente espectral y D4C para la aperiodicidad por banda de frecuencia.

La disponibilidad de una estimación de aperiodicidad por banda constituye la diferencia estructural más relevante respecto de los vocoders LPC clásicos, y es la que

le permite reproducir con mayor fidelidad el componente aperiódico presente en la voz natural.

2.4.4. Modificación deliberada de la calidad vocal

El uso de estas técnicas para modificar deliberadamente la calidad vocal, y no solo para reconstruirla, ha sido explorado en distintos marcos.

Nordstrom y colaboradores [25] propusieron transformar voces de alto esfuerzo vocal en voces sopladas mediante predicción lineal con pre-énfasis adaptativo (APLP). La técnica estima un filtro de énfasis espectral que permite manipular el esfuerzo vocal percibido, y adicionalmente proporciona una estimación del filtro de formantes más consistente entre distintas calidades vocales. Su argumento parte de observar que la predicción lineal con pre-énfasis constante estima una fuente vocal de envolvente espectral fija, cuando la envolvente de la fuente real varía en el tiempo.

Esta observación resulta directamente pertinente para el presente trabajo, en el que el pre-énfasis se emplea con un coeficiente constante, tal como se detalla en la Sección 4.4.

En el ámbito local, Opazo [26] abordó la modificación de calidad vocal empleando el vocoder WORLD, trabajo que constituye un antecedente directo del presente desarrollo y que motiva la comparación con dicho sistema presentada en la Sección 5.5.

2.5. Análisis por predicción lineal

La predicción lineal constituye la técnica más extendida para estimar los parámetros del filtro del tracto vocal a partir de la señal de voz. Su formulación parte de suponer que cada muestra puede aproximarse como una combinación lineal de las muestras anteriores, y que el error de esta predicción corresponde a la señal de excitación [27].

El sistema de ecuaciones resultante se resuelve eficientemente mediante la recursión de Levinson-Durbin, que además garantiza la estabilidad del filtro obtenido cuando los coeficientes de reflexión se mantienen acotados en magnitud.

La aplicación de un filtro de pre-énfasis previo al análisis constituye práctica habitual [28]: compensa la inclinación espectral natural de la voz, permitiendo que el filtro estimado modele preferentemente las resonancias del tracto vocal en lugar de destinar grados de libertad a representar dicha inclinación.

2.6. Implementaciones en hardware dedicado

La implementación de algoritmos de procesamiento de voz sobre procesadores digitales de señal dedicados presenta restricciones que no se manifiestan en implementaciones sobre computadores de propósito general: memoria limitada, ausencia de bibliotecas de cálculo numérico, y necesidad de considerar explícitamente la precisión aritmética disponible.

Estas restricciones condicionan las decisiones de diseño. La aritmética de doble precisión, disponible por defecto en entornos de escritorio, se emula en software en muchos procesadores embebidos con el consiguiente costo computacional. La memoria dinámica, empleada habitualmente por las implementaciones de referencia de los algoritmos de síntesis, resulta inviable o desaconsejable en sistemas embebidos, lo que obliga a dimensionar estáticamente todas las estructuras de datos.

En contrapartida, la implementación sobre hardware dedicado ofrece determinismo temporal y la posibilidad de construir sistemas autónomos y portátiles, características relevantes para aplicaciones clínicas o experimentales fuera del laboratorio.

2.7. Síntesis del estado del arte

De la revisión anterior se desprenden tres observaciones que enmarcan el presente trabajo.

En primer lugar, existe consenso respecto de los correlatos acústicos asociados a la voz soplada y a la voz tensa: medidas de periodicidad, particularmente las de base cepstral, para la primera; medidas de distribución espectral para la segunda. Estas medidas cuentan con implementaciones disponibles y validadas clínicamente.

En segundo lugar, la relación entre los parámetros fisiológicos, sus correlatos acústicos y la percepción resultante presenta una variabilidad que la literatura reconoce explícitamente. Ello sugiere que la caracterización de un sistema de modificación de calidad vocal debe incorporar tanto medición objetiva como evaluación perceptual, sin asumir que una predice a la otra.

En tercer lugar, los trabajos de síntesis y modificación de calidad vocal revisados se han desarrollado sobre computadores de propósito general. La implementación sobre hardware dedicado, con las restricciones que ello impone, constituye el espacio en el que se sitúa esta memoria.

3. Capítulo 3: Marco Teórico

Este capítulo desarrolla los fundamentos matemáticos sobre los que se construye el sistema implementado: el modelo de producción de voz, las técnicas de análisis empleadas para estimar sus parámetros, los modelos de excitación disponibles para la síntesis, las métricas acústicas utilizadas en la caracterización, y las estructuras de filtrado digital sobre las que se implementan las modificaciones.

3.1. Modelo de producción de voz

3.1.1. Formulación del modelo fuente-filtro

El modelo fuente-filtro representa la señal de voz $s[n]$ como el resultado de excitar un sistema lineal invariante en el tiempo con una señal de excitación $u[n]$. En el dominio de la transformada Z:

$$S(z) = U(z)\,G(z)\,V(z)\,R(z) \tag{1}$$

donde $U(z)$ representa la excitación glótica, $G(z)$ la conformación espectral introducida por la forma del pulso glótico, $V(z)$ la función de transferencia del tracto vocal y $R(z)$ el efecto de la radiación en los labios.

La radiación labial se modela habitualmente como un diferenciador de primer orden, $R(z) = 1 - z^{-1}$, que introduce un realce de aproximadamente +6 dB por octava. La conformación glótica $G(z)$ presenta, por su parte, una caída del orden de −12 dB por octava. La combinación de ambos efectos produce la inclinación espectral neta de aproximadamente −6 dB/octava característica de la voz.

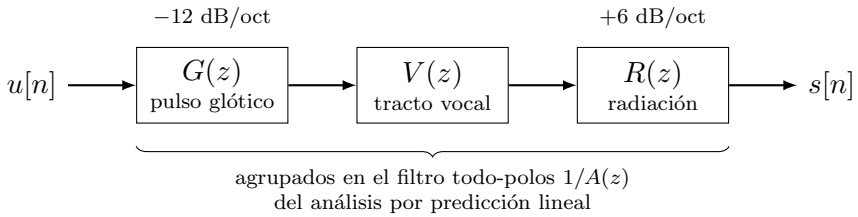


Figura 1: Modelo fuente-filtro de la producción de voz. Los tres bloques de la derecha se agrupan en un único filtro todo-polos durante el análisis por predicción lineal.

En la práctica del análisis por predicción lineal, los tres factores $G(z)$, $V(z)$ y $R(z)$ se agrupan en un único filtro todo-polos:

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

donde p es el orden del modelo, $\{a_k\}$ son los coeficientes de predicción y G es una ganancia. Esta agrupación es la que motiva el uso del pre-énfasis: al compensar previamente la inclinación de −6 dB/octava, el filtro estimado modela preferentemente las resonancias del tracto vocal.

3.1.2. Excitación glótica

Durante la fonación sonora, la excitación $u[n]$ corresponde a un tren cuasi-periódico de pulsos de flujo glótico de período $T_0 = f_s/F_0$, donde f_s es la frecuencia de muestreo

y F_0 la frecuencia fundamental. Durante los segmentos sordos, la excitación es de naturaleza aperiódica.

El cociente de apertura se define como

$$OQ = \frac{T_{\text{abierto}}}{T_0} \tag{3}$$

siendo T_{abierto} la duración de la fase en que los pliegues vocales permanecen separados. Valores elevados de OQ se asocian a la fonación soplada y valores reducidos a la fonación tensa.

3.2. Análisis por predicción lineal

3.2.1. Formulación

La predicción lineal supone que cada muestra de la señal puede aproximarse mediante una combinación lineal de las p muestras precedentes:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k \, s[n-k] \tag{4}$$

El error de predicción, denominado residual, es

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{p} a_k \, s[n-k] \tag{5}$$

Comparando con la ecuación (2) se observa que el residual corresponde a la señal filtrada por $A(z)$, esto es, a una estimación de la señal de excitación. Esta correspondencia es la que fundamenta el uso del filtrado inverso para estimar la fuente glótica.

3.2.2. Ecuaciones normales

Los coeficientes $\{a_k\}$ se determinan minimizando la energía del residual sobre la trama de análisis. Derivando respecto de cada coeficiente e igualando a cero se obtiene el sistema de ecuaciones normales:

$$\sum_{k=1}^{p} a_k \, R(|i-k|) = R(i), \qquad i = 1, \ldots, p \tag{6}$$

donde $R(\tau)$ denota la función de autocorrelación de la trama ventaneada:

$$R(\tau) = \sum_{n=\tau}^{N-1} s[n] \, s[n-\tau] \tag{7}$$

La matriz del sistema resultante es de Toeplitz simétrica, propiedad que permite su resolución eficiente [27].

3.2.3. Recursión de Levinson-Durbin

La recursión de Levinson-Durbin resuelve el sistema en $O(p^2)$ operaciones en lugar de las $O(p^3)$ que requeriría una inversión matricial directa. Partiendo de $E^{(0)} = R(0)$, para cada orden $i = 1, \ldots, p$ se calcula:

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E^{(i-1)}} \tag{8}$$

$$a_i^{(i)} = k_i, \qquad a_j^{(i)} = a_j^{(i-1)} - k_i\, a_{i-j}^{(i-1)}, \quad j = 1, \ldots, i-1 \tag{9}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{10}$$

Los coeficientes k_i se denominan coeficientes de reflexión o PARCOR, y admiten interpretación física como coeficientes de reflexión acústica en un modelo de tubos concatenados del tracto vocal.

3.2.4. Estabilidad y ancho de banda de los polos

La recursión garantiza que el filtro $1/A(z)$ resulte estable, es decir, que todos sus polos queden dentro del círculo unitario, si y solo si

$$|k_i| < 1, \qquad i = 1, \ldots, p \tag{11}$$

Esta propiedad proporciona un criterio de verificación directo en la implementación: basta acotar los coeficientes de reflexión durante la recursión para garantizar un filtro de síntesis estable.

Una técnica habitual para controlar el ancho de banda de las resonancias es el ensanchamiento espectral, que consiste en escalar los coeficientes por un factor $\gamma < 1$:

$$\tilde{a}_k = \gamma^k a_k \tag{12}$$

Esta operación desplaza cada polo de radio r a radio γr, sin alterar su frecuencia, ensanchando en consecuencia el ancho de banda de todas las resonancias.

3.2.5. Pre-énfasis

El filtro de pre-énfasis

$$H_{\text{pre}}(z) = 1 - \alpha z^{-1}, \qquad \alpha \approx 0{,}97 \tag{13}$$

aproxima un diferenciador de primer orden y compensa la inclinación espectral de la voz. Su aplicación previa al análisis LPC evita que el modelo destine polos a representar dicha inclinación en lugar de las resonancias del tracto vocal [28]. El filtro inverso, denominado de-énfasis, restituye la inclinación tras la síntesis.

3.3. Estimación de la frecuencia fundamental

3.3.1. Método de autocorrelación

La estimación de F_0 por autocorrelación se fundamenta en que una señal periódica de período T_0 presenta máximos locales de su función de autocorrelación en los retardos múltiplos de dicho período. El período se estima entonces como

$$\hat{T}_0 = \arg \max_{\tau \in [\tau_{\text{mín}}, \tau_{\text{máx}}]} R(\tau) \tag{14}$$

con los límites de búsqueda determinados por el rango de frecuencias fundamentales esperado, $\tau_{\text{mín}} = f_s/F_{0,\text{máx}}$ y $\tau_{\text{máx}} = f_s/F_{0,\text{mín}}$.

3.3.2. Normalización y sesgo

La expresión (7) presenta un sesgo sistemático: la suma en el retardo τ comprende $N - \tau$ términos, frente a los N del retardo nulo. En consecuencia, la autocorrelación normalizada

$$\rho(\tau) = \frac{R(\tau)}{R(0)} \tag{15}$$

presenta un techo inferior a la unidad incluso para señales perfectamente periódicas. La corrección por cobertura

$$\tilde{\rho}(\tau) = \frac{R(\tau)}{R(0)} \cdot \frac{N}{N - \tau} \tag{16}$$

compensa este sesgo y permite emplear $\tilde{\rho}$ como estimador del grado de periodicidad de la trama.

3.3.3. Errores de octava

El método de autocorrelación es susceptible a dos modos de error característicos. El primero consiste en seleccionar un múltiplo del período verdadero, dado que la autocorrelación presenta máximos en todos ellos. El segundo, menos evidente, se produce cuando la función decrece de forma monótona en el intervalo de búsqueda sin presentar máximo local alguno: en tal caso el máximo global coincide con un extremo del intervalo, sin que dicho retardo corresponda a periodicidad real de la señal.

La Figura 2 ilustra este segundo modo de error sobre una señal de habla continua.

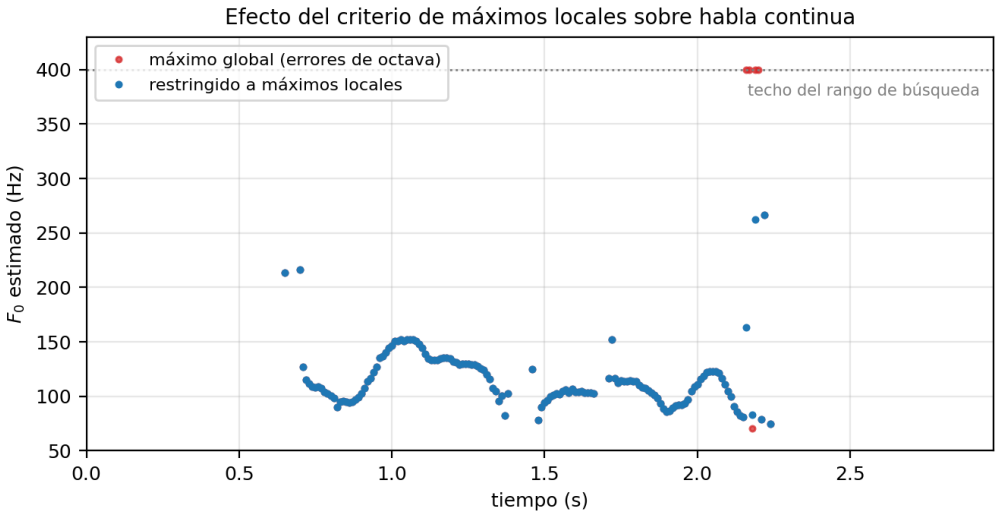


Figura 2: Estimación de F_0 sobre habla continua. Al tomar el máximo global del intervalo de búsqueda, las tramas de baja periodicidad reciben estimaciones espurias en el extremo del rango. La restricción a máximos locales estrictos las descarta correctamente.

La exigencia de que el retardo seleccionado constituya un máximo local estricto,

$$R(\hat{\tau}) > R(\hat{\tau} - 1) \quad \text{y} \quad R(\hat{\tau}) > R(\hat{\tau} + 1) \tag{17}$$

descarta el segundo modo de error y permite clasificar la trama como sorda cuando no existe candidato válido.

3.4. Modelos de excitación para síntesis

3.4.1. Tren de impulsos

El modelo de excitación más simple para segmentos sonoros es el tren de impulsos de período T_0:

$$u[n] = A \sum_m \delta[n - mT_0] \tag{18}$$

Su espectro es a su vez un tren de impulsos en frecuencia, con armónicos de igual amplitud espaciados F_0. Esta uniformidad espectral, unida a la coincidencia de fase de todos los armónicos, produce el carácter mecánico conocido como *buzziness* [22].

Dos consideraciones resultan relevantes para su implementación. La primera es que la energía media de la excitación depende del período: con amplitud A fija, la potencia media es A^2/T_0, de modo que tonos agudos reciben mayor energía. La normalización $A = \sqrt{T_0}$ hace la potencia media independiente de F_0. La segunda es que un tren unipolar presenta valor medio no nulo, $\bar{u} = A/T_0$, componente que el filtro de síntesis amplifica considerablemente dada su elevada ganancia en continua.

3.4.2. Excitación mixta

La excitación mixta combina una componente periódica y una aperiódica en proporción determinada por el grado de periodicidad de la trama:

$$u[n] = (1 - \beta)\, u_p[n] + \beta\, u_n[n] \tag{19}$$

donde β es el coeficiente de mezcla, $u_p[n]$ la componente de pulso y $u_n[n]$ la de ruido.

El coeficiente β se obtiene directamente del coeficiente de periodicidad estimado en cada trama, $\tilde{\rho}(\hat{\tau})$ (ecuación (16)), mediante

$$\beta = 1 - \tilde{\rho}(\hat{\tau}) \tag{20}$$

de modo que una trama fuertemente periódica ($\tilde{\rho} \to 1$) concentra la excitación en el tren de pulsos ($\beta \to 0$), mientras que una trama de baja periodicidad ($\tilde{\rho} \to 0$) aproxima la excitación a ruido puro ($\beta \to 1$). El valor de β se acota al intervalo $[0, 1]$ para prevenir excursiones fuera de rango ante estimaciones de periodicidad anómalas.

Para que β tenga el significado de proporción de energía, ambas componentes deben normalizarse a igual energía por período. El pulso concentra su energía en una única muestra, mientras que el ruido la distribuye entre las T_0 muestras del período; si u_n es ruido uniforme en $[-1, 1]$, de varianza $1/3$, la igualdad de energías

$$T_0 \cdot \frac{B^2}{3} = A^2 = T_0 \tag{21}$$

conduce a una amplitud de ruido $B = \sqrt{3}$.

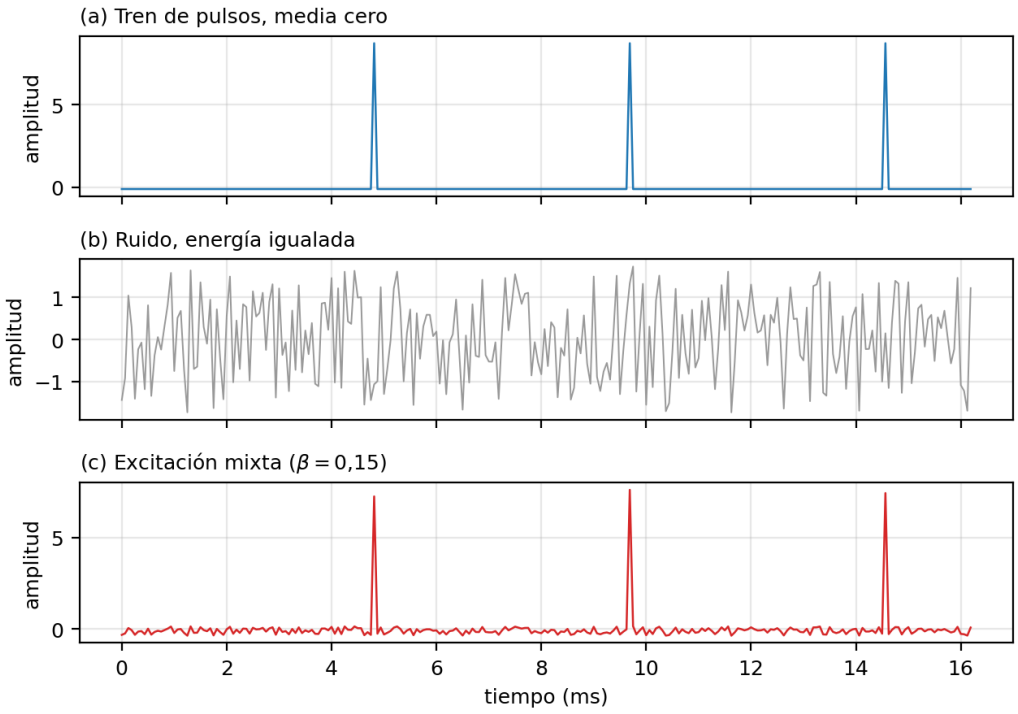


Figura 3: Componentes de la excitación mixta. El tren de pulsos (a) se compensa en continua; el ruido (b) se escala para igualar su energía por período a la del pulso; la mezcla (c) se pondera según el grado de periodicidad medido en la trama.

Los vocoders de excitación mixta más elaborados, como MELP [23], generalizan este esquema calculando un coeficiente de mezcla independiente para cada banda de frecuencia, en reconocimiento de que el grado de periodicidad de la voz real varía con la frecuencia.

3.5. Métricas acústicas de calidad vocal

3.5.1. Cepstrum

El cepstrum real de una señal se define como la transformada inversa de Fourier del logaritmo de la magnitud de su espectro:

$$c[q] = \mathcal{F}^{-1}\left\{\log\left|S(\omega)\right|\right\} \tag{22}$$

La variable independiente q tiene dimensión temporal y se denomina *quefrency*. La operación logarítmica convierte el producto de la ecuación (1) en una suma, lo que separa en el dominio cepstral la contribución de la envolvente espectral, concentrada en *quefrencies* bajas, de la estructura armónica fina, que produce un pico en la *quefrency* correspondiente al período fundamental.

3.5.2. Prominencia del pico cepstral

La prominencia del pico cepstral (CPP) cuantifica cuánto sobresale dicho pico respecto de la tendencia general del cepstrum. Se calcula ajustando una recta de regresión $\hat{c}[q]$ sobre el cepstrum en el rango de *quefrencies* correspondiente al rango de F_0 esperado, y midiendo la diferencia en el máximo:

$$\mathrm{CPP} = c[\hat{q}] - \hat{c}[\hat{q}], \qquad \hat{q} = \arg\max_{q} c[q] \tag{23}$$

La variante suavizada, CPPS, promedia el cepstrograma en las dimensiones de *quefrency* y de tiempo antes del cálculo, lo que reduce la varianza del estimador.

La interpretación es directa: una señal fuertemente periódica concentra su energía en los armónicos, produciendo un pico cepstral prominente; la presencia de ruido aperiódico rellena los valles entre armónicos y atenúa el pico. El CPPS constituye por ello un correlato acústico de la calidad soplada [9, 14].

3.5.3. Relación armónicos-ruido

La relación armónicos-ruido (HNR) cuantifica la proporción entre la energía periódica y la aperiódica:

$$\mathrm{HNR}_{\mathrm{dB}} = 10 \log_{10} \frac{E_{\text{armónica}}}{E_{\text{ruido}}} \tag{24}$$

Su estimación puede realizarse en el dominio cepstral mediante una operación de *comb-liftering* que separa ambas contribuciones [14].

3.5.4. Pendiente espectral

La pendiente espectral describe la tasa de decaimiento de la energía con la frecuencia. Se obtiene mediante ajuste lineal por mínimos cuadrados de la magnitud espectral en decibelios frente al logaritmo en base dos de la frecuencia:

$$20 \log_{10} |S(f)| \approx m \log_2 f + b \tag{25}$$

El coeficiente m, expresado en dB/octava, constituye la métrica de interés. Valores más negativos indican mayor concentración de energía en las bajas frecuencias, característica de la fonación soplada; valores menos negativos o positivos indican un espectro más plano, característico de la fonación tensa [13].

3.5.5. Diferencia de amplitud entre los primeros armónicos

La diferencia de amplitud entre el primer y el segundo armónico, habitualmente denotada H1–H2, constituye el correlato acústico más directamente ligado al cociente de apertura glótica. Se define como

$$\mathrm{H1}-\mathrm{H2} = 20 \log_{10} \frac{|S(F_0)|}{|S(2F_0)|} \tag{26}$$

donde $|S(F_0)|$ y $|S(2F_0)|$ denotan las amplitudes espectrales de los dos primeros armónicos.

Su fundamento reside en que un pulso glótico de fase abierta prolongada concentra su energía en las componentes de menor frecuencia, elevando la amplitud relativa del primer armónico. Klatt y Klatt [8] identificaron esta relación como una de las claves acústicas de la calidad soplada, y la señalaron como consecuencia directa del incremento del cociente de apertura.

Conviene, no obstante, considerar dos limitaciones. La primera es que la medida no compensa la influencia del primer formante: cuando F_1 se sitúa próximo a $2F_0$, la resonancia realza el segundo armónico y reduce artificialmente el valor de H1–H2. Existen correcciones que estiman y descuentan dicha influencia, dando lugar a la variante denotada H1*–H2*. La segunda es que el respaldo empírico de la cadena causal que vincula el cociente de apertura, la medida H1–H2 y la calidad soplada percibida resulta

más débil de lo que suele asumirse, según muestran las mediciones simultáneas de señal acústica e imagen laríngea reportadas por Kreiman y colaboradores [5].

3.5.6. Medidas de perturbación

El *jitter* y el *shimmer* cuantifican la variabilidad ciclo a ciclo del período y de la amplitud respectivamente. En su formulación local:

$$\text{jitter} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}|T_i - T_{i+1}|}{\frac{1}{M}\sum_{i=1}^{M} T_i} \tag{27}$$

$$\text{shimmer} = \frac{\frac{1}{M-1}\sum_{i=1}^{M-1}|A_i - A_{i+1}|}{\frac{1}{M}\sum_{i=1}^{M} A_i} \tag{28}$$

donde T_i y A_i denotan el período y la amplitud de pico del ciclo i-ésimo, y M el número de ciclos analizados.

3.6. Distorsión log-espectral

La distorsión log-espectral (LSD) cuantifica la diferencia entre las envolventes espectrales de dos señales, típicamente una señal original y su reconstrucción. Se define como

$$\text{LSD} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left[20\log_{10}|S(k)| - 20\log_{10}|\hat{S}(k)|\right]^2} \tag{29}$$

donde $S(k)$ y $\hat{S}(k)$ denotan las magnitudes espectrales de la señal de referencia y de la señal reconstruida en el bin de frecuencia k, y K el número de bins considerados. Valores menores indican mayor fidelidad espectral de la reconstrucción respecto de la señal original.

3.7. Filtrado digital

Todas las modificaciones de calidad vocal implementadas se construyen a partir de un conjunto acotado de filtros digitales. Esta sección desarrolla los tipos empleados, sus estructuras de realización y los criterios de diseño de sus coeficientes.

3.7.1. Tipos de filtro empleados

El sistema utiliza cinco clases de filtro, cada una asociada a una función específica dentro de la cadena de procesamiento:

Tipo	Orden	Función en el sistema
Bloqueador de continua	1	eliminar el desplazamiento del ADC
Pre-énfasis / de-énfasis	1	compensar la inclinación espectral
Pasabanda	2	conformar el ruido de aspiración
Shelving (bajo y alto)	2	modificar el balance espectral
Todo-polos	$p = 18$	modelar el tracto vocal

Bloqueador de continua. La señal entregada por el conversor analógico-digital presenta habitualmente un desplazamiento de continua que debe eliminarse antes del análisis, dado que el filtro de síntesis presenta ganancia elevada en dicha frecuencia. Se emplea un filtro pasa-altos de primer orden con polo próximo al círculo unitario:

$$H_{\mathrm{dc}}(z) = \frac{1 - z^{-1}}{1 - r z^{-1}}, \qquad r = 0{,}995 \tag{30}$$

El cero en $z = 1$ anula la componente continua, mientras que el polo en $z = r$ mantiene la respuesta aproximadamente plana por encima de unos pocos hercios. Su implementación requiere una única muestra de memoria por cada uno de los dos estados.

Filtros de primer orden para énfasis. El pre-énfasis y su inverso, descritos en la ecuación (13), son filtros de primer orden de coeficiente único. El de pre-énfasis es de respuesta finita al impulso; el de de-énfasis, de respuesta infinita:

$$y[n] = x[n] - \alpha\, x[n-1] \qquad \text{y} \qquad y[n] = x[n] + \alpha\, y[n-1] \tag{31}$$

Filtro pasabanda. El ruido de aspiración empleado en la modificación de voz soplada se conforma mediante un filtro pasabanda de segundo orden centrado en la región del tercer formante. Su función de transferencia, en la parametrización del *Audio EQ Cookbook*, es

$$H_{\mathrm{bp}}(z) = \frac{\alpha(1 - z^{-2})}{(1+\alpha) - 2\cos(\omega_0) z^{-1} + (1-\alpha) z^{-2}} \tag{32}$$

con $\omega_0 = 2\pi f_0 / f_s$, $\alpha = \sin(\omega_0)/(2Q)$ y factor de calidad $Q = f_0/\mathrm{BW}$, siendo BW el ancho de banda deseado.

Filtros *shelving*. Los filtros de tipo *shelving* aplican una ganancia constante por encima o por debajo de una frecuencia de transición, dejando la banda complementaria inalterada. Constituyen la herramienta natural para modificar el balance espectral sin introducir resonancias, razón por la cual se emplean en ambas modificaciones de calidad vocal. Definiendo $A_{\mathrm{sh}} = 10^{G/40}$ con G la ganancia en decibelios, y S_{sh} el parámetro de pendiente, los coeficientes del high-shelf son

$$\begin{aligned}
b_0 &= A_{\mathrm{sh}} \left[(A_{\mathrm{sh}}+1) + (A_{\mathrm{sh}}-1)\cos\omega_0 + 2\sqrt{A_{\mathrm{sh}}}\,\alpha \right] \\
b_1 &= -2A_{\mathrm{sh}} \left[(A_{\mathrm{sh}}-1) + (A_{\mathrm{sh}}+1)\cos\omega_0 \right] \\
b_2 &= A_{\mathrm{sh}} \left[(A_{\mathrm{sh}}+1) + (A_{\mathrm{sh}}-1)\cos\omega_0 - 2\sqrt{A_{\mathrm{sh}}}\,\alpha \right] \\
a_0 &= (A_{\mathrm{sh}}+1) - (A_{\mathrm{sh}}-1)\cos\omega_0 + 2\sqrt{A_{\mathrm{sh}}}\,\alpha \\
a_1 &= 2\left[(A_{\mathrm{sh}}-1) - (A_{\mathrm{sh}}+1)\cos\omega_0 \right] \\
a_2 &= (A_{\mathrm{sh}}+1) - (A_{\mathrm{sh}}-1)\cos\omega_0 - 2\sqrt{A_{\mathrm{sh}}}\,\alpha
\end{aligned} \tag{33}$$

con $\alpha = \tfrac{1}{2}\sin\omega_0 \sqrt{(A_{\mathrm{sh}} + 1/A_{\mathrm{sh}})(1/S_{\mathrm{sh}} - 1) + 2}$. Los coeficientes del *low-shelf* se obtienen mediante las expresiones análogas con los signos de los términos en $\cos\omega_0$ invertidos.

3.7.2. Estructura de realización

Todos los filtros de segundo orden se realizan mediante secciones *biquad* en forma directa II transpuesta. La función de transferencia de una sección es

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \tag{34}$$

y sus ecuaciones de estado

$$\begin{aligned} y[n] &= b_0 x[n] + z_1[n-1] \\ z_1[n] &= b_1 x[n] - a_1 y[n] + z_2[n-1] \\ z_2[n] &= b_2 x[n] - a_2 y[n] \end{aligned} \tag{35}$$

La elección de esta estructura responde a dos consideraciones. La primera es de eficiencia: requiere únicamente dos elementos de memoria por sección, frente a los cuatro de la forma directa I. La segunda es numérica: la forma transpuesta presenta menor sensibilidad al redondeo de los estados intermedios, propiedad relevante en implementaciones de precisión finita.

3.7.3. Filtro todo-polos de orden arbitrario

El filtro de síntesis LPC, de orden p, se implementa como una generalización de la estructura anterior con p elementos de memoria:

$$y[n] = G\,x[n] - \sum_{k=1}^{p} a_k\, y[n-k] \tag{36}$$

Dado que los coeficientes se actualizan en cada trama de análisis, la conservación del estado del filtro entre tramas consecutivas resulta necesaria para evitar discontinuidades en la señal sintetizada. Su estabilidad queda garantizada por el criterio de la ecuación (11) sobre los coeficientes de reflexión.

3.7.4. Compresión de rango dinámico

El compresor empleado en la modificación de voz tensa no es un filtro lineal sino un sistema de ganancia variable controlado por el nivel de la señal. Su detector de nivel emplea constantes de tiempo independientes para el ataque y la liberación:

$$\ell[n] = \begin{cases} \alpha_a\, \ell[n-1] + (1-\alpha_a)\,|x[n]| & \text{si } |x[n]| > \ell[n-1] \\ \alpha_r\, \ell[n-1] + (1-\alpha_r)\,|x[n]| & \text{en caso contrario} \end{cases} \tag{37}$$

con $\alpha_a = e^{-1/(\tau_a f_s)}$ y $\alpha_r = e^{-1/(\tau_r f_s)}$. La ganancia aplicada, expresada en decibelios, es

$$g_{dB}[n] = \begin{cases} -(L[n] - L_T)\left(1 - \dfrac{1}{R_c}\right) & \text{si } L[n] > L_T \\ 0 & \text{en caso contrario} \end{cases} \tag{38}$$

donde $L[n] = 20\log_{10}\ell[n]$, L_T es el umbral y R_c la razón de compresión. Un tiempo de ataque reducido permite al compresor actuar sobre las variaciones de amplitud entre ciclos glóticos consecutivos, reduciendo en consecuencia el *shimmer* de la señal procesada.

4. Capítulo 4: Diseño e Implementación

Este capítulo describe la plataforma empleada, la arquitectura del sistema desarrollado, los algoritmos implementados y la metodología de medición utilizada para caracterizarlo.

4.1. Plataforma de hardware

4.1.1. Selección de la plataforma

La implementación se realizó sobre la tarjeta *Low Cost Development Kit* (LCDK) de Texas Instruments [29], plataforma orientada al desarrollo y la docencia en procesamiento digital de señales. Los criterios que motivaron su elección fueron los siguientes:

- **Unidad de punto flotante en hardware.** El procesador C674x combina núcleos de punto fijo y de punto flotante, lo que permitió desarrollar la totalidad del sistema en aritmética de punto flotante sin recurrir a representaciones de punto fijo. Esta característica resultó determinante para la implementación de la recursión de Levinson-Durbin, numéricamente sensible.
- **Códec de audio integrado.** La tarjeta incorpora un códec estéreo de calidad suficiente para señales de voz, con entradas de línea y de micrófono seleccionables por software.
- **Memoria externa amplia.** Los 128 MB de memoria DDR permiten almacenar buffers de varios segundos de audio, requisito del esquema de procesamiento adoptado.
- **Disponibilidad de material de apoyo.** La plataforma cuenta con documentación abundante y con el repositorio de bibliotecas desarrollado para el curso de laboratorio de la universidad [30], descrito más adelante.

La elección de la LCDK no fue la primera aproximación. El desarrollo se inició sobre la plataforma Jacinto 7, de mayor capacidad de cómputo, pero el proceso se vio interrumpido por un error no resuelto (`MCASP_EBADARGS`) durante la creación del canal de transmisión del periférico McASP, que impidió completar incluso la inicialización básica del hardware de audio. Ante la magnitud de esta dificultad y la escasa documentación disponible para resolverla en un plazo razonable, se optó por migrar el desarrollo a la LCDK, plataforma con documentación abundante y con la cual se contaba con experiencia previa de uso. La plataforma Jacinto 7 no se descarta para trabajos futuros, tal como se discute en la Sección 6.4.

4.1.2. Características del hardware

Cuadro 2: Características de la plataforma de desarrollo.

Componente	Especificación
Procesador	Texas Instruments TMS320C6748 [31]
Arquitectura	C674x, VLIW de punto fijo y flotante
Frecuencia de reloj	300 MHz (verificada desde el archivo GEL)
Memoria interna	32 kB L1P, 32 kB L1D, 256 kB L2
Memoria externa	128 MB DDR2/mDDR a 150 MHz
Códec de audio	TLV320AIC3106 [32], estéreo
Resolución del códec	16 bits
Emulador	XDS110 externo, interfaz JTAG sobre USB
Entradas de audio	Línea estéreo (J10) y micrófono (J11)
Salida de audio	Línea estéreo (J12)
Periféricos usados	4 LEDs, 4 pulsadores, 8 interruptores DIP

El emulador XDS110 externo conectado a la tarjeta permite la programación y la depuración a través de una única conexión USB. Proporciona acceso a los registros del procesador, la posibilidad de establecer puntos de ruptura, y la lectura y escritura directa de la memoria, funcionalidad esta última que se empleó sistemáticamente para exportar los buffers de señal hacia el computador para su análisis posterior.

La Figura 4 resume el montaje experimental completo. El mismo computador cumple dos roles independientes: mediante la conexión USB al emulador XDS110 se programa y depura la LCDK, y se exportan los buffers de señal capturados; mediante un cable de audio hacia la entrada de línea de la placa se reproducen las señales de prueba descritas en la Sección 4.8.

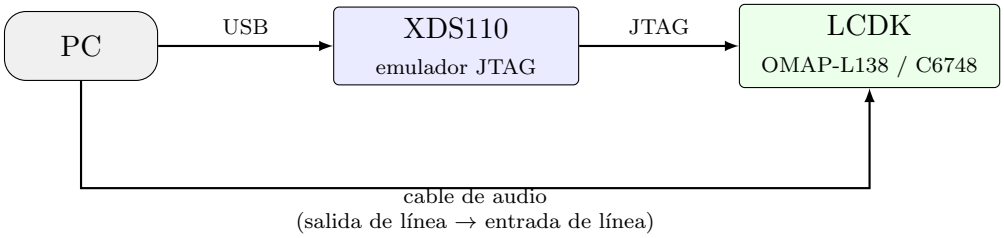


Figura 4: Montaje experimental. El computador programa y depura la LCDK a través del emulador XDS110 mediante USB, y reproduce las señales de prueba hacia la entrada de línea de la placa mediante un cable de audio.

4.1.3. Entorno de desarrollo

El sistema se desarrolló íntegramente en lenguaje C, bajo el paradigma *bare-metal*: el código se ejecuta directamente sobre el procesador sin sistema operativo ni planificador de tareas intermedios. Esta aproximación otorga control completo sobre el comportamiento temporal del sistema, característica necesaria para el manejo de la rutina de interrupción de audio, y elimina la sobrecarga que introduciría un sistema operativo de tiempo real.

El entorno de desarrollo empleado fue Code Composer Studio, en su versión basada en Theia, ejecutado sobre macOS. La compilación se realizó con las herramientas de generación de código de Texas Instruments para la arquitectura C6000.

4.1.4. Bibliotecas de apoyo

El manejo de bajo nivel de la plataforma se apoyó en el repositorio de bibliotecas desarrollado por el Dr. Gonzalo Carrasco para el curso de Laboratorio de Procesamiento Digital de Señales de la universidad [30]. Dicho repositorio proporciona la capa de abstracción sobre la que se construyó el sistema, y comprende:

- La rutina de inicialización del códec y del periférico de audio (`L138_LCDK_aic3106_init`), que configura los registros del AIC3106 y establece la cadena de transferencia de muestras.
- Una biblioteca de utilidades (`dsp_lab_utils`) que encapsula el acceso a los periféricos de la tarjeta: lectura de pulsadores e interruptores, control de los LEDs, y las funciones de medición de tiempo empleadas para la caracterización del costo computacional.
- El archivo de definición de vectores de interrupción y el *script* de configuración del reloj del sistema.

La disponibilidad de esta capa permitió concentrar el desarrollo en los algoritmos de procesamiento de voz, objeto de este trabajo, en lugar de en la configuración de bajo nivel de los periféricos.

4.2. Flujo de datos y cadena de procesamiento

Esta sección describe el recorrido de la señal desde su ingreso analógico hasta su almacenamiento en memoria, y las opciones de configuración disponibles en cada etapa.

4.2.1. Cadena de adquisición

La Figura 5 presenta el flujo de datos completo.

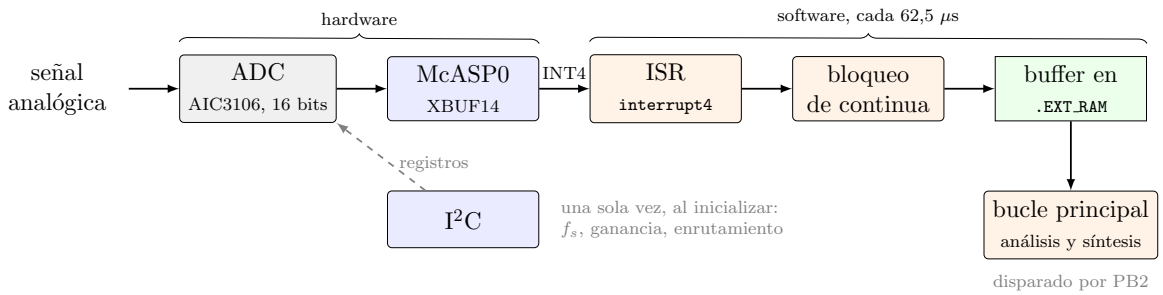


Figura 5: Flujo de datos en la plataforma. La configuración del códec se realiza una única vez mediante I²C; la transferencia de muestras ocurre de forma continua a través del McASP, y cada muestra genera una interrupción atendida por el procesador.

La señal analógica ingresa por el conector de línea y es convertida por el ADC del códec AIC3106 a 16 bits de resolución. Las muestras se transfieren al procesador mediante el periférico McASP (*Multichannel Audio Serial Port*), configurado en modo serie sincrónico. La lectura y la escritura se realizan sobre los registros `XBUF14` y `XBUF13` respectivamente, empleando un formato de palabra de 32 bits que agrupa los dos canales estéreo.

4.2.2. Configuración del códec mediante I²C

El códec AIC3106 se configura mediante escritura de registros a través del bus I²C, operación que se realiza una única vez durante la inicialización del sistema. Cada registro es de 8 bits, y la biblioteca empleada expone la función `AIC3106_writeRegister` para su acceso.

Los parámetros de configuración relevantes para este trabajo, con sus registros asociados, son los siguientes:

Cuadro 3: Registros del códec configurados por el sistema.

Registro	Función	Configuración empleada
1	Reinicio del dispositivo	valor 0x80 al inicio de la secuencia
2	Frecuencia de muestreo	0x44 para 16 kHz
7	Ruta de datos del códec	DAC izquierdo a canal izquierdo, DAC derecho a canal derecho
15, 16	Ganancia del PGA de entrada	36 dB en ambos canales
19, 22	Alimentación de los ADC	encendido de ambos conversores
21, 24	Enrutamiento de entrada de línea	intercambio de canales para compensar el cruce de pistas del circuito impreso de la tarjeta
43, 44	Atenuación digital de salida	0 dB

La selección entre entrada de línea y entrada de micrófono se realiza mediante una constante de compilación. En modo micrófono se habilita adicionalmente la polarización *MICBIAS* a 2 V mediante el registro 25, necesaria para micrófonos de condensador. El sistema desarrollado emplea la entrada de línea.

La frecuencia de muestreo se selecciona igualmente por constante, entre los valores soportados por el códec que van de 8 a 48 kHz. La elección de 16 kHz responde a que el ancho de banda resultante de 8 kHz cubre holgadamente el rango relevante de la señal de voz, manteniendo acotado el número de muestras a procesar.

4.2.3. Modo de transferencia y atención de muestras

La biblioteca de apoyo ofrece tres modos de operación para la transferencia de muestras entre el códec y el procesador:

- **Sondeo** (*polling*): el procesador consulta repetidamente el estado del periférico. Simple pero ineficiente, ya que consume ciclos en la espera.
- **Interrupción**: cada muestra disponible genera una interrupción que el procesador atiende. Permite procesamiento muestra a muestra con latencia mínima.
- **EDMA** (*Enhanced Direct Memory Access*): el controlador de acceso directo a memoria transfiere bloques completos de muestras sin intervención del procesador, empleando un esquema de doble buffer (*ping-pong*). Reduce la sobrecarga por muestra a costa de introducir una latencia igual a la duración del bloque.

El sistema desarrollado emplea el modo de **interrupción**. Esta elección responde a que la etapa de captura debe almacenar cada muestra individualmente en el buffer de análisis, sin requerir el procesamiento por bloques que justificaría el uso de EDMA, y a que el procesamiento pesado no ocurre en la rutina de interrupción sino en el bucle principal, como se detalla en la Sección 4.3.

La configuración de la interrupción asocia el evento del periférico McASP0, identificado con el número 61, a la línea de interrupción 4 del procesador, mediante la escritura de dicho valor en el registro `INTMUX1`. La rutina de servicio correspondiente, declarada como `interrupt4`, se ejecuta por tanto a la frecuencia de muestreo, es decir, cada 62,5 μs a 16 kHz.

4.2.4. Operaciones realizadas en la rutina de interrupción

La rutina de interrupción ejecuta un número acotado de operaciones, dado que dispone de un presupuesto de tiempo estricto:

1. Lectura de las muestras de entrada desde el registro del McASP.
2. Filtrado de bloqueo de continua, según la ecuación (30).
3. Almacenamiento de la muestra en el buffer de captura, si el sistema se encuentra en estado de adquisición.
4. Escritura de la muestra de salida al registro del McASP.

En el sistema desarrollado la salida se mantiene en silencio durante la captura, dado que el procesamiento no ocurre en tiempo real y no existe señal que reproducir de forma simultánea. La escritura se realiza de todas formas en cada ciclo para mantener el periférico correctamente alimentado de datos.

4.2.5. Medición del tiempo de ejecución

La caracterización del costo computacional se realizó mediante el contador de ciclos de 64 bits del procesador, accesible a través de los registros `TSCL` y `TSCH`. La biblioteca de apoyo expone este mecanismo mediante las funciones `DLU_tic` y `DLU_toc`, que registran respectivamente el valor inicial y final del contador, dejando la diferencia en cuentas disponible para su lectura.

La conversión de cuentas a unidades de tiempo requiere conocer la frecuencia del procesador. Ésta se verificó a partir del *script* de configuración GEL empleado por el entorno de desarrollo, que establece el reloj del núcleo en 300 MHz al conectar el emulador. La resolución resultante de la medición es de 3,33 ns, holgadamente suficiente para los tiempos del orden de segundos que se caracterizaron.

4.3. Arquitectura del sistema

4.3.1. Decisión de diseño: procesamiento por buffer

El sistema no opera en tiempo real. Esta decisión fue deliberada y responde a la naturaleza del algoritmo implementado: el análisis por predicción lineal requiere disponer de una ventana completa de señal antes de poder estimar los coeficientes del filtro, y la estimación de la frecuencia fundamental por autocorrelación tiene un costo computacional que, como se demuestra en la Sección 5.5, sitúa a la cadena completa en el límite de la viabilidad para operación muestra a muestra.

En su lugar, el sistema opera sobre buffers completos mediante una máquina de estados controlada por pulsadores:

1. **Captura**: el pulsador PB1 inicia la adquisición. La rutina de interrupción de audio almacena las muestras en un buffer de 5 segundos residente en memoria externa.

2. **Procesamiento**: el pulsador PB2 dispara el análisis y la síntesis, que se ejecutan en el bucle principal.

3. **Exportación**: los buffers de entrada y salida se vuelcan a archivo desde el entorno de desarrollo para su análisis posterior.

Los buffers de gran tamaño se ubicaron en memoria externa mediante directivas `DATA_SECTION` del compilador, dado que exceden holgadamente la capacidad de la RAM interna. El uso total de memoria externa del sistema es de aproximadamente 3,6 MB, un 2,8 % de la disponible.

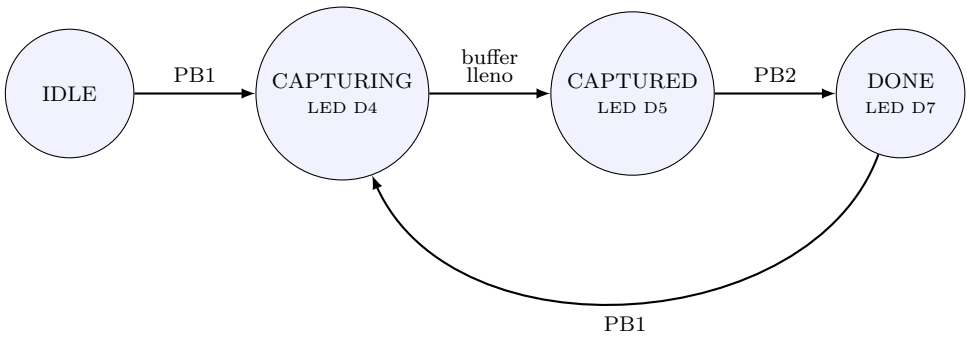


Figura 6: Máquina de estados que gobierna la captura y el procesamiento. Las transiciones se disparan por pulsador, salvo la de captura completa, que se produce al llenarse el buffer.

4.3.2. Organización modular del código

La implementación se estructuró en módulos independientes, cada uno con su interfaz declarada en un archivo de cabecera:

Módulo	Responsabilidad
`vq_config.h`	Parámetros de sintonía centralizados
`vq_biquad`	Filtros biquad y diseño de coeficientes
`vq_synth`	Análisis LPC/F_0 y síntesis
`vq_breathiness`	Efecto de voz soplada
`vq_pressed`	Efecto de voz tensa

La centralización de los parámetros en un único archivo de cabecera permitió modificar la sintonía del sistema sin alterar la lógica de los módulos, lo que resultó determinante durante la fase de ajuste experimental.

4.4. Etapa de análisis

La Figura 7 presenta la arquitectura completa del sistema.

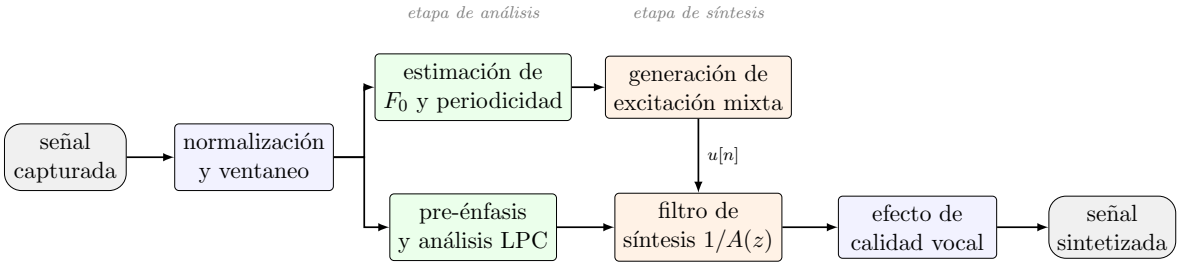


Figura 7: Arquitectura del sistema. La etapa de análisis extrae los parámetros de la señal capturada; la de síntesis los emplea para reconstruirla, aplicando a continuación la modificación de calidad vocal seleccionada.

El análisis extrae, para cada trama de la señal capturada, tres conjuntos de parámetros: la frecuencia fundamental, un coeficiente de periodicidad y los coeficientes del filtro de predicción lineal.

Se emplearon ventanas de Hamming de 25 ms (400 muestras) con un paso de 10 ms (160 muestras), valores habituales en el procesamiento de voz que equilibran la resolución temporal y la estacionariedad requerida por el análisis LPC.

4.4.1. Normalización y pre-énfasis

La señal capturada se normaliza en amplitud a un pico de 0,9 antes del análisis, lo que independiza el comportamiento del sistema del nivel de grabación.

A continuación se generan dos versiones de la señal que alimentan caminos de análisis distintos:

- Para el **análisis LPC** se aplica un filtro de pre-énfasis $H(z) = 1 - 0{,}97z^{-1}$. La voz presenta una inclinación espectral natural de aproximadamente -6 dB/octava, resultante de la combinación de la fuente glótica (-12 dB/oct) y la radiación labial ($+6$ dB/oct). Sin pre-énfasis, el filtro LPC destina parte de sus polos a modelar dicha inclinación en lugar de los formantes, con el consiguiente deterioro del ajuste en altas frecuencias. La inclinación se restituye al final de la síntesis mediante el filtro inverso.

- Para la **estimación de** F_0 se emplea la señal sin pre-énfasis. El pre-énfasis realza las frecuencias altas y atenúa las bajas, que es precisamente donde reside la información de periodicidad; su aplicación previa a la detección de *pitch* degradó la estabilidad de la estimación en un factor cercano a dos.

4.4.2. Estimación de la frecuencia fundamental

La frecuencia fundamental se estima por autocorrelación en el rango de 70 a 400 Hz, intervalo que cubre ampliamente las frecuencias fundamentales típicas de voz adulta –tanto masculina como femenina– incluyendo margen suficiente para la variación prosódica natural del habla continua. La implementación incorpora tres criterios que resultaron necesarios para obtener estimaciones robustas.

Umbral de sonoridad relativo. La decisión sonoro/sordo emplea un umbral de energía correspondiente al 2 % de la energía máxima entre todas las tramas de la captura, en lugar de una constante absoluta. Un umbral absoluto es superado por las tramas de ataque de baja energía, a las que se asigna entonces un valor espurio de F_0.

Normalización por cobertura. La autocorrelación evaluada en un retardo τ suma únicamente $N - \tau$ productos, frente a los N del retardo nulo. Sin corregir este sesgo, el coeficiente de periodicidad presenta un techo inferior a la unidad incluso para señales perfectamente periódicas. El valor se normaliza por el factor de cobertura $(N - \tau)/N$.

Restricción a máximos locales. El retardo seleccionado debe corresponder a un máximo local de la función de autocorrelación. Tomar simplemente el máximo global del rango de búsqueda permite que resulte seleccionado el extremo del intervalo cuando la función decrece de forma monótona sin presentar pico alguno, situación que se produce en tramas de baja periodicidad y que genera errores de octava. Este criterio redujo la desviación estándar de F_0 sobre habla continua de 55,7 a 17,9 Hz.

4.4.3. Análisis por predicción lineal

Los coeficientes del filtro se obtienen mediante el método de autocorrelación resuelto por la recursión de Levinson-Durbin [27]. Se empleó un orden de 18, siguiendo la regla habitual $f_s/1000 + 2$ a 4 para una frecuencia de muestreo de 16 kHz.

La Figura 8 muestra el ajuste obtenido sobre una trama de señal real.

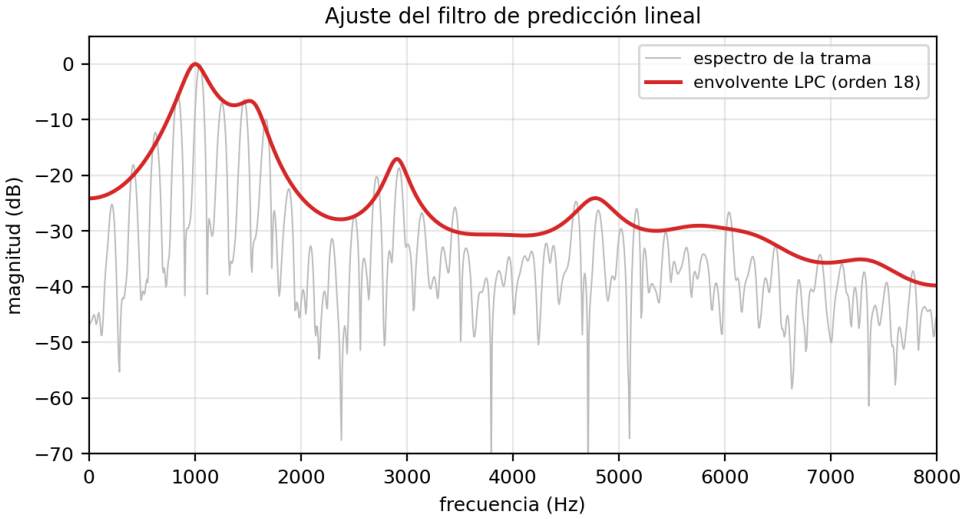


Figura 8: Envolvente espectral estimada por el filtro LPC de orden 18 sobre una trama de la vocal /a/, superpuesta al espectro de la trama.

La recursión se implementó en doble precisión internamente, convirtiendo a precisión simple únicamente el resultado final. Se incorporaron además salvaguardas numéricas: limitación de los coeficientes de reflexión al intervalo $(-1, 1)$ para garantizar la estabilidad del filtro, regularización de la autocorrelación de retardo nulo, y verificación explícita de valores no finitos.

4.5. Etapa de síntesis

La síntesis reconstruye la señal mediante el modelo fuente-filtro clásico: una señal de excitación generada a partir de F_0 y del coeficiente de periodicidad, filtrada por el modelo todo-polos del tracto vocal obtenido en el análisis.

4.5.1. Modelo de excitación

La literatura documenta que la excitación por tren de impulsos puro produce habla inteligible pero de calidad mecánica, fenómeno conocido como *buzziness* [22]. Para mitigarlo se implementó una excitación mixta que combina una componente periódica y una aperiódica, siguiendo el principio de los vocoders de excitación mixta [23], en proporción determinada por el coeficiente de periodicidad medido en cada trama.

Tres aspectos de la implementación resultaron determinantes para el comportamiento del sistema.

Normalización energética del pulso. La amplitud del impulso se escala por $\sqrt{T_0}$, donde T_0 es el período en muestras. Sin esta corrección, la energía media de la excitación depende de la frecuencia fundamental, dado que un tono agudo concentra más impulsos por unidad de tiempo.

Compensación de continua. Un tren de impulsos unipolar presenta un valor medio no nulo. Puesto que el filtro de síntesis $1/A(z)$ exhibe ganancia elevada en continua, dicha componente se amplifica y aporta energía espuria en las bandas graves. La excitación glótica real, entendida como la derivada del flujo, tiene valor medio nulo; se sustrae por tanto el valor medio del tren de impulsos.

Equilibrio energético entre componentes. El pulso concentra su energía en una única muestra por período, mientras que el ruido la distribuye entre las T_0 muestras del período. Asignar la misma amplitud a ambas componentes otorga al ruido una energía superior en un factor $T_0/3$. La amplitud del ruido se fija por tanto en $\sqrt{3}$, valor que iguala la energía de ambas componentes.

4.5.2. Filtro de síntesis

El filtrado se realiza mediante una estructura todo-polos en forma directa II transpuesta. El estado del filtro se conserva entre tramas consecutivas, lo que evita las discontinuidades periódicas que produciría su reinicialización en cada actualización de coeficientes.

Tras el filtrado se aplica el de-énfasis, filtro inverso del pre-énfasis empleado en el análisis, y una normalización final a un pico de 0,9.

4.6. Modificaciones de calidad vocal

4.6.1. Voz soplada

La implementación sigue el modelo descrito por Klatt y Klatt [8], según el cual la voz soplada se caracteriza por la presencia de ruido de aspiración sincronizado con el ciclo glótico, que reemplaza a los armónicos débiles en las frecuencias altas, junto con un aumento de la inclinación espectral.

El procesamiento consta de cuatro etapas:

1. Generación de ruido blanco mediante un registro de desplazamiento con realimentación lineal (LFSR) de 32 bits.

2. Filtrado pasabanda centrado en 3 kHz con ancho de banda de 2 kHz, correspondiente a la región del tercer formante.

3. Modulación por una envolvente rápida del valor absoluto de la señal, de constante de tiempo 0,6 ms, que sincroniza el ruido con los pulsos glóticos.

4. Mezcla aditiva con control de relación señal-ruido mediante estimadores de potencia corrida, fijando el nivel del ruido en -22 dB respecto de la voz.

Finalmente se aplica una ecualización de tipo *shelving*: realce de $+4$ dB por debajo de 400 Hz y atenuación de -6 dB por encima de 3 kHz.

4.6.2. Voz tensa

La voz tensa se caracteriza por un aplanamiento de la pendiente espectral y una reducción de la variabilidad ciclo a ciclo [13]. La implementación combina:

- Un filtro *high-shelf* con realce de $+6$ dB por encima de 2,5 kHz, que aplana la pendiente espectral.

- Un compresor de rango dinámico con detector de pico, umbral de -25 dBFS, razón de compresión 4:1, tiempo de ataque de 0,8 ms y de liberación de 50 ms. El ataque rápido reduce la variabilidad de amplitud entre ciclos consecutivos.

4.6.3. Parámetro de intensidad

Para caracterizar el control paramétrico se introdujo un parámetro único de intensidad k por efecto, que escala de forma conjunta todos los componentes del efecto correspondiente. El diseño garantiza que $k = 0$ equivalga exactamente a la condición sin modificar y que $k = 1$ corresponda a la sintonía nominal.

La alternativa de barrer un único parámetro manteniendo los restantes fijos fue descartada porque el extremo inferior del barrido no correspondería a la condición neutral, lo que impediría interpretar el umbral medido. En el caso de la voz tensa, la razón de compresión no se escala directamente sino que se interpola desde 1:1, dado que el valor neutro de un compresor es una razón unitaria y no nula.

El barrido se realizó sobre siete valores de intensidad: $k \in \{0;\ 0{,}25;\ 0{,}5;\ 0{,}75;\ 1;\ 1{,}5;\ 2\}$. La densidad de muestreo es mayor en el intervalo $[0, 1]$, donde se concentra el comportamiento de mayor interés –la transición desde la condición neutral hasta la sintonía nominal–, mientras que los valores $k = 1{,}5$ y $k = 2$ exploran el comportamiento del sistema más allá del punto de operación de diseño, permitiendo verificar la extensión de la linealidad observada en la voz tensa y de la saturación observada en la voz soplada (Sección 5.3).

Formalmente, cada parámetro p_i del efecto correspondiente, de valor nominal $p_i^{(1)}$, se escala como

$$p_i(k) = k \cdot p_i^{(1)} \tag{39}$$

con la excepción de la razón de compresión R del efecto de voz tensa, para la cual se interpola desde la razón unitaria en lugar de escalarse desde cero:

$$R(k) = 1 + k \cdot \left(R^{(1)} - 1\right) \tag{40}$$

de modo que $R(0) = 1$ (sin compresión) y $R(1) = R^{(1)} = 4{:}1$, la razón nominal reportada en el Cuadro de parámetros del Anexo A.

Los valores nominales de cada efecto se determinaron mediante un proceso de ajuste iterativo sobre las tres vocales sostenidas. En cada iteración se evaluó el efecto de la configuración propuesta sobre el CPPS (para la voz soplada) o la pendiente espectral (para la voz tensa), buscando maximizar la magnitud del cambio en la métrica objetivo sin introducir artefactos perceptualmente evidentes de distorsión o saturación. El criterio de detención combinó ambas fuentes de evidencia: se aceptó una configuración cuando producía un cambio medible y consistente entre las tres vocales, y cuando la señal resultante era percibida como una modificación creíble de calidad vocal en escucha informal.

4.7. Implementación alternativa mediante efectos de audio

Con el objeto de evaluar si la síntesis paramétrica resulta necesaria para lograr la modificación de calidad vocal perseguida, se desarrolló una segunda implementación bajo el paradigma de efecto de audio. Ésta aplica los mismos algoritmos de modificación directamente sobre la señal capturada, sin etapa de análisis ni de resíntesis.

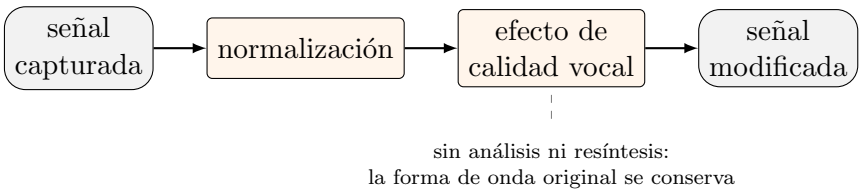


Figura 9: Arquitectura del paradigma de efecto de audio. A diferencia del sintetizador, no existe etapa de análisis ni de resíntesis: la forma de onda original se conserva y solo se modifica.

Ambas implementaciones comparten los módulos de efectos y de filtrado, de modo que las diferencias observadas son atribuibles exclusivamente al paradigma de procesamiento y no a diferencias de implementación de los efectos.

4.8. Metodología de medición

4.8.1. Material de prueba

Se emplearon cuatro señales: las vocales sostenidas /a/, /i/ y /u/ de una hablante femenina, proporcionadas por el Laboratorio de Producción de Voz (VPLab) del Advanced Center for Electrical and Electronic Engineering (AC3E) de la universidad, y una frase de habla continua extraída del repositorio de apoyo del curso de Laboratorio de Procesamiento Digital de Señales [30]. La frase fue producida por un hablante distinto del correspondiente a las vocales sostenidas, dado que ambos materiales provienen de fuentes independientes. Las tres vocales cubren los extremos del espacio articulatorio: /a/ abierta, /i/ con segundo formante elevado y /u/ con segundo formante reducido.

La inclusión de tres vocales responde a una consideración metodológica: el ajuste de parámetros sobre una única señal condujo, durante el desarrollo, a configuraciones que resultaban óptimas para esa señal pero deficientes para las restantes.

4.8.2. Métricas acústicas

Las métricas de calidad vocal se calcularon con Praat [33] versión 6.4.06 sobre las señales exportadas desde la memoria del DSP.

La métrica principal para la voz soplada es el CPPS (*Cepstral Peak Prominence, Smoothed*), que cuantifica la prominencia del pico cepstral respecto de la tendencia del cepstrum y constituye el correlato acústico más robusto de la calidad soplada [9, 14]. Se empleó la implementación oficial de Praat sobre un *power cepstrogram*, con los siguientes parámetros: piso de *pitch* 60 Hz, paso temporal 0,002 s, frecuencia máxima 5000 Hz, rango de búsqueda 60–330 Hz, interpolación parabólica, y sustracción de tendencia por decaimiento exponencial mediante ajuste robusto.

Se registraron adicionalmente la relación armónicos-ruido (HNR) por autocorrelación cruzada, el *jitter* local, el *shimmer* local y las estadísticas de F_0.

La pendiente espectral, métrica principal para la voz tensa, se calculó como el ajuste lineal por mínimos cuadrados de la magnitud espectral en decibelios frente al logaritmo en base dos de la frecuencia, en el rango de 100 Hz a 5 kHz, expresándose en dB/octava.

La diferencia de amplitud entre los dos primeros armónicos, definida en la ecuación (26), se calculó mediante una implementación propia. El procedimiento estima primero la frecuencia fundamental por autocorrelación sobre el segmento central de la señal, y mide a continuación la amplitud espectral máxima en sendas ventanas de

tolerancia del 12 % centradas en F_0 y en $2F_0$, sobre un espectro calculado con relleno de ceros para aumentar la resolución en frecuencia.

La estimación de F_0 empleada en este cálculo incorpora una salvaguarda frente a errores de subarmónico: dado que la autocorrelación de una señal periódica presenta máximos de magnitud comparable en todos los múltiplos del período, se selecciona el máximo local de menor retardo entre aquellos que alcanzan el 88 % del valor máximo. Sin esta precaución, la estimación puede seleccionar el doble del período verdadero y desplazar en consecuencia la medición de ambos armónicos.

La implementación se validó sobre señales sintéticas construidas con una diferencia H1–H2 impuesta, en el rango de -6 a $+14$ dB. El error máximo de estimación resultó de 0,03 dB en todo el rango, incluyendo los valores negativos en que el segundo armónico supera al primero.

4.8.3. Verificación cruzada de la implementación

Se desarrolló un procedimiento de verificación que resultó determinante para la depuración del sistema. El código C del proyecto se compiló también para computador de escritorio, procesando las mismas capturas exportadas desde el DSP.

La correlación entre la salida obtenida en el computador y la obtenida en el hardware resultó de 1,0000, con diferencia máxima muestra a muestra nula. Esta equivalencia numérica exacta permitió analizar el comportamiento del sistema en un entorno de desarrollo convencional con la certeza de estar observando exactamente el mismo procesamiento que ejecuta el DSP, y descartar de plano las diferencias de compilador o de hardware como origen de los comportamientos observados.

4.8.4. Evaluación perceptual

Se diseñó una prueba de escucha a ciegas con 17 estímulos de 2 segundos, presentados en orden aleatorio. Para cada uno de los dos efectos se generaron seis niveles de intensidad no nulos ($k = 0{,}25$ a $k = 2{,}00$), lo que aporta 12 estímulos modificados; a estos se sumaron cinco muestras adicionales en condición neutral ($k = 0$, idéntica para ambos efectos) como control de falsas alarmas, completando los 17 estímulos evaluados. El evaluador desconocía la correspondencia entre estímulo y condición.

La inclusión de estímulos de control responde a que la escucha no ciega realizada durante el desarrollo mostró una tendencia a anticipar el efecto esperado, y permite estimar la tasa de falsas alarmas necesaria para interpretar correctamente los aciertos obtenidos.

5. Capítulo 5: Resultados

Este capítulo presenta los resultados obtenidos con el sistema implementado sobre la plataforma LCDK. Se organizan en cinco partes: la verificación funcional del sintetizador (Sección 5.1), la caracterización cuantitativa de las modificaciones de calidad vocal (Sección 5.2), la caracterización del control paramétrico mediante barrido de intensidad (Sección 5.3), la evaluación perceptual (Sección 5.4) y la caracterización del costo computacional junto con la comparación entre paradigmas de implementación (Sección 5.5).

Todas las mediciones acústicas se realizaron con Praat 6.4.06 sobre las señales exportadas desde la memoria del DSP, empleando la implementación oficial de CPPS de Praat [33] sobre un *power cepstrogram* (piso de *pitch* 60 Hz, paso 0,002 s, tope 5000 Hz, sustracción de tendencia por decaimiento exponencial). Salvo indicación contraria, las señales de entrada corresponden a vocales sostenidas /a/, /i/ y /u/ de una hablante femenina y, cuando se indica explícitamente, a una frase de habla continua producida por un hablante distinto (Sección 4.8), ambas reproducidas hacia la entrada de línea de la LCDK y capturadas a 16 kHz.

La Figura 10 resume el flujo experimental completo, desde la captura de las señales de prueba hasta la caracterización final del sistema presentada en este capítulo.

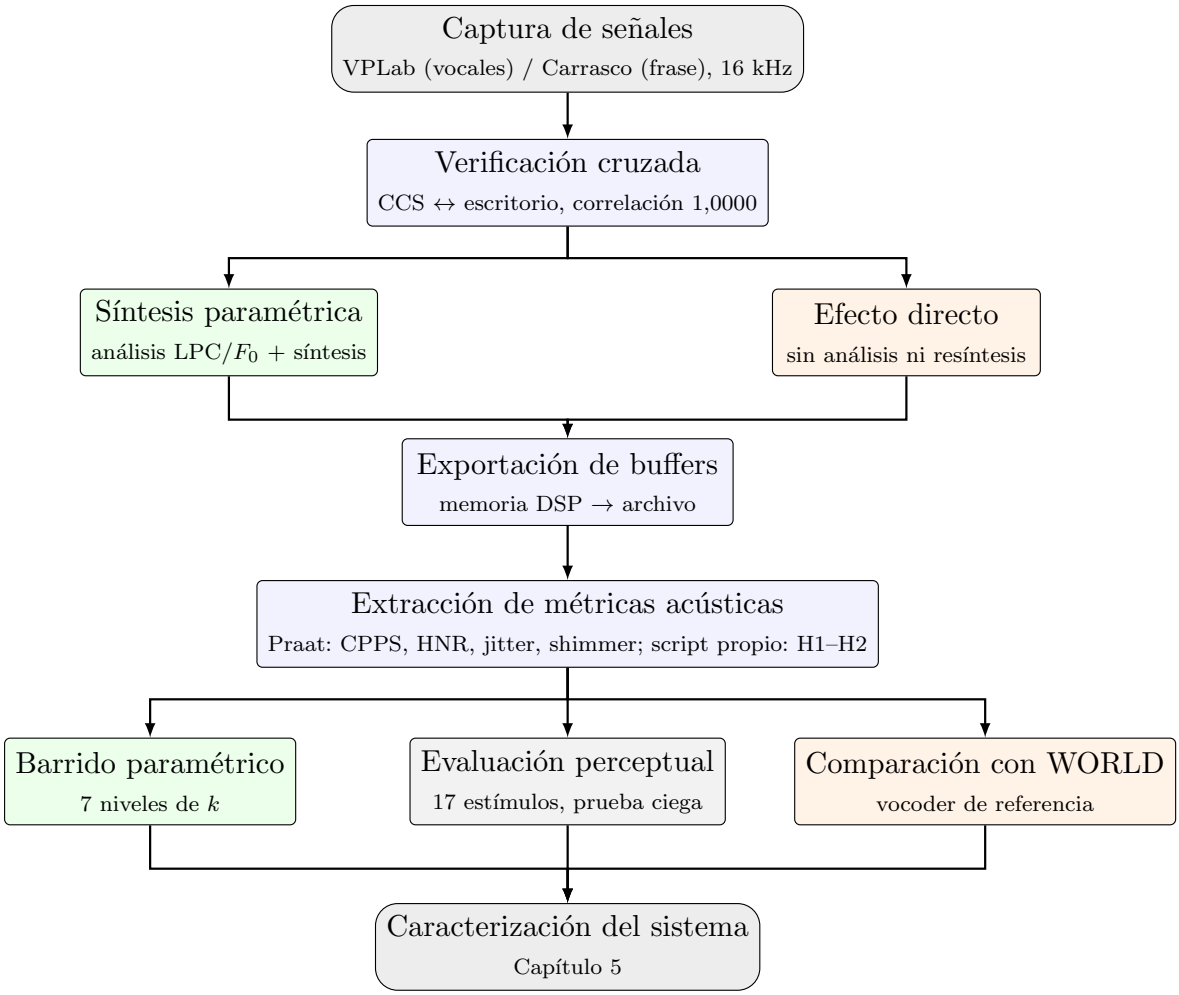


Figura 10: Esquema del flujo experimental completo, desde la captura de las señales de prueba hasta la caracterización final del sistema.

5.1. Verificación funcional del sintetizador

El sintetizador fuente-filtro implementado reconstruye la señal de voz a partir de tres conjuntos de parámetros extraídos del análisis: la frecuencia fundamental F_0, el grado de periodicidad y los coeficientes del filtro LPC de orden 18. Para verificar que la salida corresponde efectivamente a una señal sintetizada y no a una versión filtrada de la entrada, se calculó la correlación cruzada entre ambas formas de onda.

La correlación resultó de $-0{,}11$ y el error relativo de 1,45, valores que descartan cualquier relación de copia entre ambas señales (una reconstrucción idéntica daría correlación unitaria y error nulo). La inspección de las formas de onda, presentada en la Figura 11, confirma esta observación: mientras la señal original presenta la asimetría y el detalle irregular característicos de la excitación glótica natural, la señal sintetizada exhibe el patrón regular propio de una excitación por tren de pulsos.

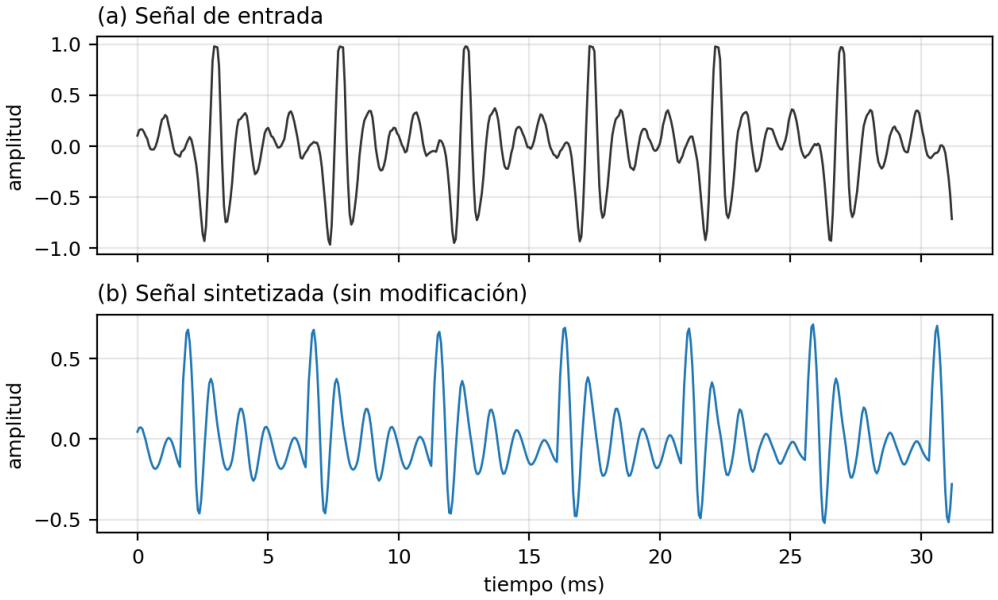


Figura 11: Comparación de las formas de onda de entrada y sintetizada sobre un mismo tramo de la vocal /a/.

5.1.1. Validación de la cadena de análisis-síntesis

Como verificación independiente de la corrección del filtro LPC, se implementó una prueba de filtrado inverso: la señal de entrada se filtró con $A(z)$ para obtener el residual, y éste se reinyectó en el filtro de síntesis $1/A(z)$. El error relativo de reconstrucción resultó de 0,000000, lo que confirma que el análisis LPC, el ventaneo, la ganancia y el filtro de síntesis operan correctamente. Esta prueba permitió establecer que las limitaciones de calidad observadas se originan exclusivamente en el modelo de excitación y no en la estimación del filtro del tracto vocal.

5.1.2. Comportamiento con habla continua

El sistema se validó adicionalmente con una frase completa, un caso sustancialmente más exigente que la vocal sostenida por incorporar segmentos sordos, transiciones formánticas rápidas, silencios y variación real de F_0. Sobre este material, el estimador de F_0 presentó una desviación estándar de 17,9 Hz con un máximo de 152 Hz, valores consistentes con la prosodia natural del hablante.

5.2. Modificación de la calidad vocal

La Tabla 4 presenta las métricas de calidad vocal obtenidas para las tres vocales, comparando la señal de entrada, la reconstrucción sin modificación (neutral) y las dos modificaciones implementadas.

Cuadro 4: Métricas de calidad vocal a intensidad nominal ($k = 1$).

Vocal	Condición	CPPS (dB)	HNR (dB)	Pendiente (dB/oct)
/a/	entrada	10,70	23,04	−2,61
	neutral	15,25	25,80	−2,84
	breathy	9,99	21,70	−2,25
	pressed	15,30	25,19	−1,14
/i/	entrada	5,72	16,91	−1,34
	neutral	12,91	22,75	−1,45
	breathy	7,88	22,50	−1,97
	pressed	12,91	19,49	+0,23
/u/	entrada	5,50	16,17	−3,42
	neutral	14,99	23,66	−3,55
	breathy	8,74	22,56	−3,88
	pressed	15,03	21,18	−1,76

Las Figuras 12 y 13 presentan la comparación en el dominio tiempo-frecuencia y la envolvente espectral promediada de cada condición.

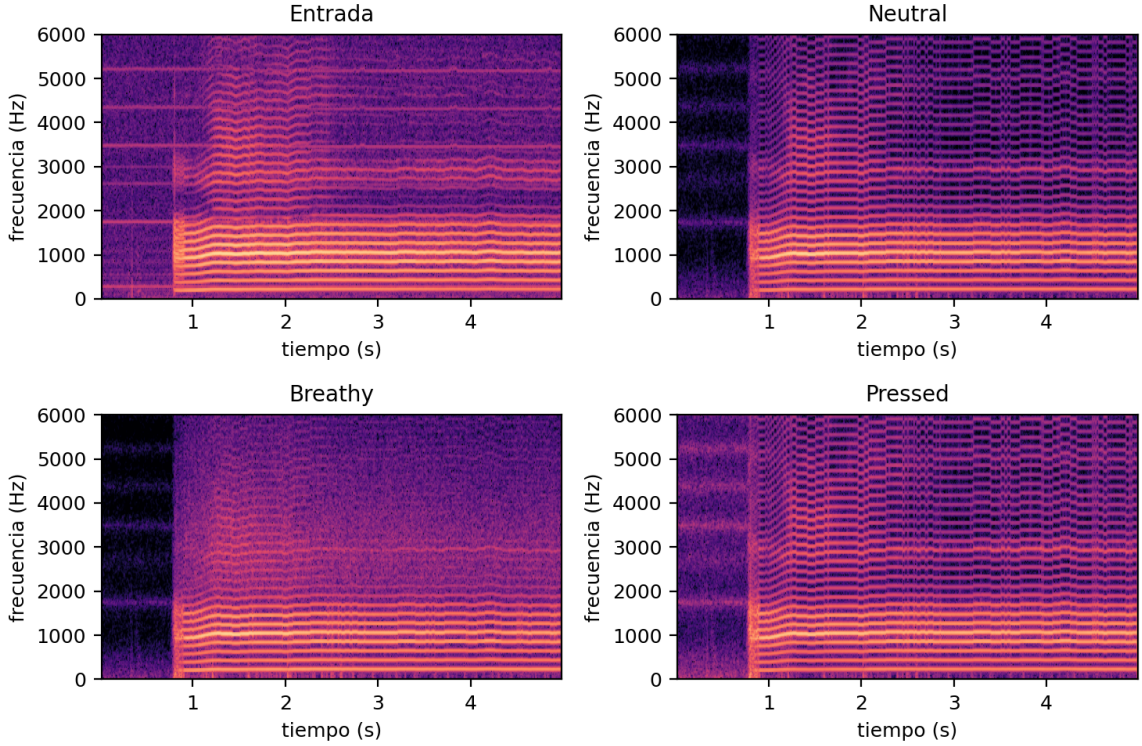


Figura 12: Espectrogramas de la señal de entrada y de las tres condiciones sintetizadas para la vocal /a/.

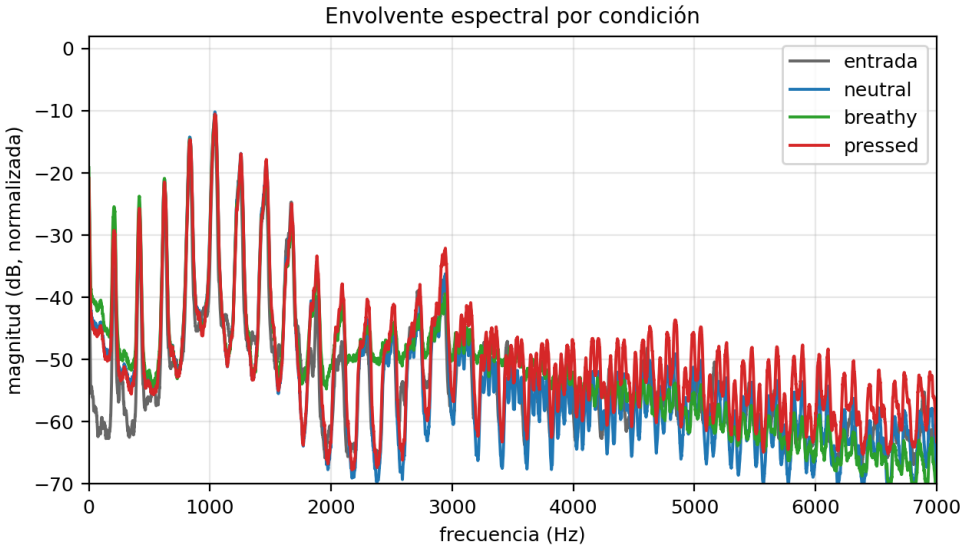


Figura 13: Envolvente espectral promediada de cada condición. La modificación *pressed* desplaza energía hacia las frecuencias altas, mientras que *breathy* produce el efecto opuesto.

5.2.1. Selectividad del control

El resultado más relevante de la Tabla 4 es que cada modificación afecta principalmente a la métrica asociada al mecanismo fisiológico que emula, sin alterar significativamente la otra.

La modificación *breathy* reduce el CPPS entre 5,03 y 6,25 dB respecto de la condición neutral, de forma consistente en las tres vocales. Esta reducción es coherente con lo predicho por la literatura [9, 14]: el ruido aspirado disminuye la prominencia del pico cepstral al reducir la proporción de energía armónica respecto de la aperiódica.

La modificación *pressed*, en cambio, deja el CPPS prácticamente inalterado (variaciones de hasta 0,05 dB) y actúa sobre la pendiente espectral, que se aplana en aproximadamente 1,7 dB/oct. Este comportamiento también es el esperado: la voz tensa no altera la periodicidad de la señal sino la distribución de energía entre las bandas graves y agudas [13].

Esta separación entre los efectos constituye evidencia de control paramétrico selectivo, que es el objetivo planteado en OE2.

5.2.2. Diferencia de amplitud entre los primeros armónicos

La Tabla 5 presenta la medida H1–H2 para las cuatro señales de prueba, que constituye una verificación independiente del mecanismo fisiológico emulado por la modificación de voz soplada.

Cuadro 5: Diferencia H1–H2 (dB) y variación respecto de la condición neutral.

Señal	Entrada	Neutral	Breathy	Δ	Pressed
/a/	0,08	5,17	7,06	+1,90	5,07
/i/	8,65	9,79	11,98	+2,19	9,91
/u/	2,71	2,01	4,10	+2,09	2,32
frase	3,81	6,37	6,80	+0,43	6,83

La modificación de voz soplada eleva H1–H2 de forma consistente en las tres vocales sostenidas, con incrementos entre 1,90 y 2,19 dB y una dispersión inferior a 0,3 dB. La modificación de voz tensa, en cambio, deja la medida prácticamente inalterada, con variaciones entre $-0{,}09$ y $+0{,}47$ dB.

Este resultado refuerza la evidencia de selectividad presentada anteriormente, y lo hace desde una métrica de naturaleza distinta a las ya consideradas: mientras el CPPS cuantifica el grado de periodicidad global y la pendiente espectral describe la distribución de energía en todo el espectro, H1–H2 mide específicamente la relación entre los dos primeros armónicos, magnitud ligada al cociente de apertura glótica [8]. Que las tres métricas apunten en la misma dirección sugiere que la modificación implementada reproduce el mecanismo fisiológico pretendido y no únicamente su manifestación en una métrica particular.

Sobre la frase de habla continua, el incremento resulta considerablemente menor ($+0{,}43$ dB). Esta observación es coherente con la reducción del rango de control reportada para el CPPS sobre el mismo material, y admite la misma explicación: el habla continua presenta segmentos sordos y transiciones en los que la estructura armónica está menos definida.

En el paradigma de efecto directo se observa el mismo patrón, con un incremento medio de 1,63 dB para la modificación de voz soplada frente a 0,49 dB para la de voz tensa, valores comparables a los del sintetizador.

5.3. Caracterización del control paramétrico

Para determinar desde qué punto cada modificación comienza a afectar la calidad vocal, se implementó un barrido paramétrico que genera siete niveles de intensidad de cada efecto sobre el mismo segmento de audio en una única pasada de procesamiento. El parámetro de intensidad k escala en conjunto todos los componentes de cada efecto, de modo que $k = 0$ equivale exactamente a la condición neutral, $k = 1$ corresponde a la intensidad nominal y $k = 2$ a una versión exagerada.

La equivalencia entre $k = 0$ y la condición neutral se verificó numéricamente: la diferencia muestra a muestra entre ambas señales resultó idénticamente nula, lo que garantiza que el origen de la escala es un cero verdadero y que los umbrales medidos son interpretables.

5.3.1. Voz tensa: control lineal

La Figura 14 muestra la variación de la pendiente espectral en función de la intensidad. El comportamiento es lineal en las tres vocales, con coeficientes de determinación superiores a 0,999:

Vocal	Pendiente (dB/oct por unidad de k)	R^2
/a/	+1,613	0,9993
/i/	+1,636	0,9998
/u/	+1,675	0,9995

Las tres pendientes se agrupan en un rango de apenas 0,063 dB/oct, lo que representa una dispersión inferior al 4 %. El control resulta por tanto esencialmente independiente de la vocal procesada.

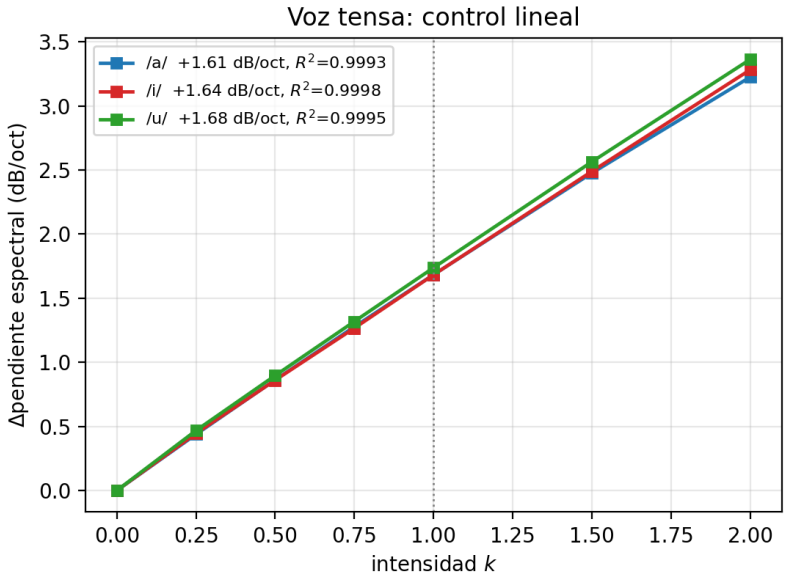


Figura 14: Variación de la pendiente espectral en función de la intensidad del efecto de voz tensa. El comportamiento es lineal en las tres vocales.

5.3.2. Breathiness: control logarítmico saturante

El comportamiento de la modificación *breathy* es cualitativamente distinto, como muestra la Figura 15. La reducción del CPPS no es proporcional a la intensidad sino que satura progresivamente: en la vocal /a/, el 47 % de la caída total ocurre en el primer tramo del barrido (k entre 0 y 0,25), mientras que la sensibilidad decae desde $-14{,}3$ dB por unidad de k al inicio hasta $-1{,}2$ al final del rango.

Los datos se ajustan a una ley logarítmica de la forma $\Delta\text{CPPS} = -a\ln(1+bk)$, con coeficientes de determinación entre 0,97 y 1,00 según la vocal. Este comportamiento es coherente con la naturaleza logarítmica de la escala en decibelios sobre la que se define el CPPS.

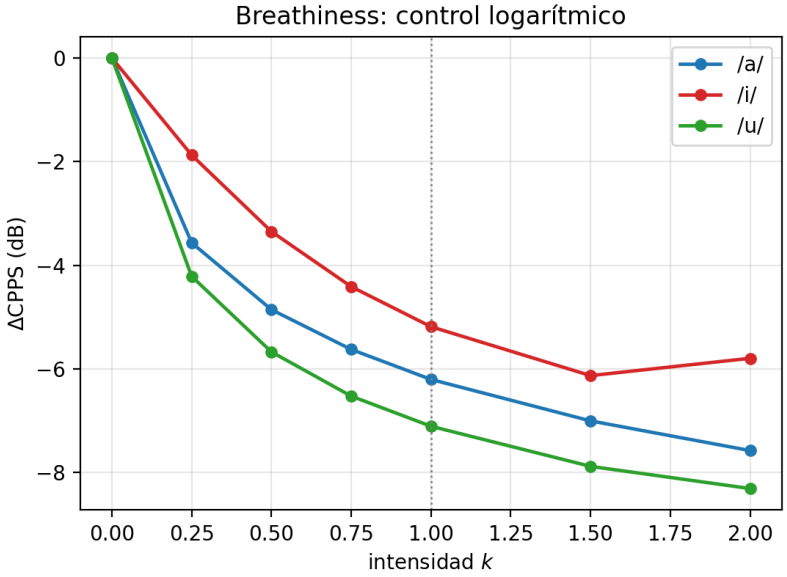


Figura 15: Variación del CPPS en función de la intensidad del efecto de breathiness. El control satura progresivamente.

La caída total de CPPS a lo largo del barrido fue de −7,58 dB en /a/, −6,13 dB en /i/ y −8,31 dB en /u/.

5.3.3. Dependencia del material fonético

El barrido aplicado sobre habla continua reveló una limitación relevante del control de breathiness. Mientras en vocales sostenidas el recorrido del CPPS alcanza entre 6 y 8 dB, sobre la frase se reduce a 1,42 dB, con saturación completa a partir de $k = 1$.

La explicación reside en el valor de partida de la métrica: el CPPS de la frase sin modificar es de 6,32 dB, frente a valores entre 13 y 18 dB en las vocales sostenidas, y su HNR es de apenas 4,97 dB frente a 25–28 dB. El habla continua es intrínsecamente menos periódica por incorporar segmentos sordos y transiciones [10], de modo que el margen disponible para reducir adicionalmente la periodicidad es considerablemente menor.

Cabe notar que la frase de habla continua fue producida por un hablante distinto de las vocales sostenidas (Sección 4.8), por lo que parte de la diferencia observada podría estar influida por características propias del hablante además de por la naturaleza del material fonético. Esta limitación se retoma en la Sección 6.3.

5.4. Evaluación perceptual

Se realizó una evaluación perceptual preliminar mediante una prueba a ciegas. Se generaron 17 estímulos de 2 segundos, presentados en orden aleatorio: 12 correspondientes a los seis niveles no nulos de intensidad de cada efecto ($k = 0{,}25$ a $k = 2{,}00$), más cinco muestras adicionales en condición neutral ($k = 0$) como control de falsas alarmas. El evaluador debía clasificar cada estímulo como *neutral*, *breathy* o *pressed*, sin conocer la correspondencia.

La tasa global de aciertos fue de 13 de 17 estímulos (76 %). Los resultados por condición se resumen a continuación:

- **Breathiness**: identificada correctamente en los niveles $k = 0{,}50$ a $k = 2{,}00$ (cinco de cinco estímulos), fallando únicamente en el nivel mínimo $k = 0{,}25$. El umbral de detección se sitúa por tanto entre $k = 0{,}25$ y $k = 0{,}50$.
- **Voz tensa**: la detección resultó inconsistente. Fue identificada en $k = 0{,}25$ pero no en $k = 1{,}00$, con respuestas dudosas en los niveles intermedios.
- **Controles neutrales**: dos de las cinco muestras sin modificación fueron reportadas como *pressed*, constituyendo falsos positivos.

La combinación de detección errática y tasa de falsa alarma no despreciable impide establecer un umbral perceptual confiable para la voz tensa a partir de estos datos. Este resultado contrasta con la caracterización objetiva del mismo efecto, y se retoma en la discusión.

5.4.1. Limitaciones de la evaluación

Esta evaluación involucró un único evaluador, sin repeticiones, con cinco a seis estímulos por condición. No permite por tanto establecer conclusiones estadísticas y debe interpretarse como una exploración preliminar orientada a identificar tendencias, no como un test perceptual formal.

5.5. Costo computacional y comparación entre paradigmas

5.5.1. Descomposición del costo de análisis

La instrumentación del código mediante los contadores de ciclos del DSP permitió desglosar el tiempo de procesamiento por etapa. Considerando únicamente la etapa de análisis y una síntesis (es decir, excluyendo el barrido paramétrico), el análisis –estimación de F0 y cálculo de coeficientes LPC– concentra aproximadamente el 72 % del tiempo de esa porción del procesamiento.

El análisis de los tiempos obtenidos para las cuatro señales evaluadas permitió construir un modelo de costo. Dado que el análisis LPC se ejecuta en todas las tramas del buffer mientras que la búsqueda de F_0 sólo se realiza en las tramas que superan el umbral de sonoridad, el tiempo total de análisis T_{an} se describe como:

$$T_{\text{an}} = N_{\text{total}} \cdot c_{\text{LPC}} + N_{\text{sonoras}} \cdot c_{F_0} \tag{41}$$

El ajuste por mínimos cuadrados sobre los cuatro casos entrega $c_{\text{LPC}} = 2{,}83$ ms por trama y $c_{F_0} = 6{,}80$ ms por trama sonora, con $R^2 = 0{,}9982$. La búsqueda de F_0 resulta por tanto 2,4 veces más costosa por trama que el análisis LPC completo de orden 18, y representa entre el 60 % y el 70 % del tiempo de análisis. Este resultado identifica la estimación de F_0 como el objetivo prioritario de optimización en caso de requerirse operación en tiempo real.

5.5.2. Comparación entre paradigmas de implementación

Con el fin de evaluar si la síntesis paramétrica es necesaria para lograr la modificación de calidad vocal perseguida, se implementó una segunda versión del sistema bajo el paradigma de efecto de audio: los mismos algoritmos de modificación aplicados directamente sobre la señal capturada, sin análisis ni resíntesis. El barrido paramétrico descrito en la Sección 5.3 se aplicó de forma idéntica sobre ambas implementaciones: mismos siete valores de intensidad k, misma señal de entrada y misma pasada de procesamiento, de modo que las diferencias en el control observadas entre paradigmas, presentadas más adelante en esta misma sección, son atribuibles exclusivamente a la arquitectura de procesamiento y no a diferencias metodológicas en la caracterización.

La Tabla 6 compara el tiempo total de procesamiento de ambas implementaciones, y la Figura 16 lo representa gráficamente.

Cuadro 6: Tiempo total de procesamiento por paradigma (C6748 a 300 MHz, capturas de 5 segundos).

Señal	Síntesis (ms)	Efecto directo (ms)	Factor
/a/	9333	1686	5,5×
/i/	9499	1697	5,6×
/u/	9586	1696	5,7×
frase	7248	1597	4,5×

Más relevante que el factor global es el costo normalizado respecto de la duración del audio. La aplicación de un efecto sobre 5 segundos de señal requiere aproximadamente 250 ms, es decir, 0,05 veces el tiempo real. La cadena de análisis más una síntesis, en cambio, requiere 4,8 segundos para procesar 5 segundos de audio (0,97 veces el tiempo real), quedando en el límite de la viabilidad para operación en tiempo real.

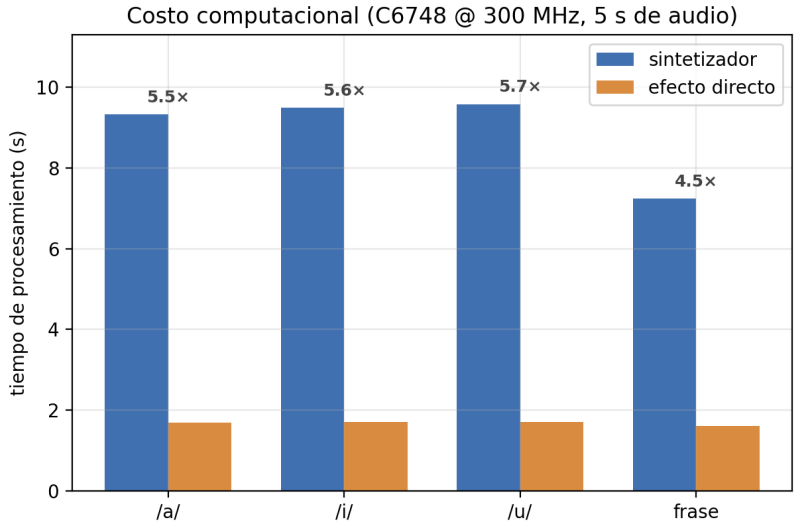


Figura 16: Costo computacional de ambos paradigmas.

5.5.3. Equivalencia y diferencias en el control

La Figura 17 compara el control paramétrico obtenido con ambos paradigmas.

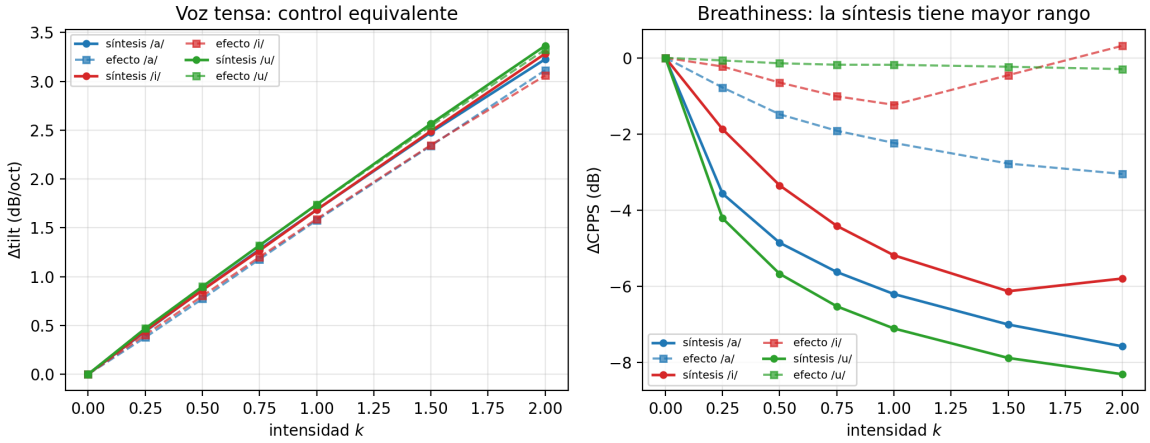


Figura 17: Comparación del control paramétrico entre el sintetizador (línea continua) y el efecto directo (línea discontinua).

Para la voz tensa, ambos paradigmas entregan un control lineal de calidad equivalente: las pendientes obtenidas con el efecto directo (+1,53 a +1,65 dB/oct por unidad de k) son estadísticamente indistinguibles de las del sintetizador (+1,61 a +1,68), con coeficientes de determinación comparables.

Para la breathiness, en cambio, el sintetizador presenta un rango de control considerablemente mayor: la caída media de CPPS es de 7,34 dB frente a 1,53 dB del efecto directo, un factor de 4,8×. La diferencia se atribuye al punto de inyección del ruido aspirado: en el sintetizador éste se incorpora a la señal de excitación, antes del filtro que modela el tracto vocal, mientras que en el efecto directo se suma a la señal ya conformada.

5.5.4. Comparación con un vocoder de referencia

Como referencia externa se procesaron las mismas señales de entrada con WORLD [24], sistema ampliamente utilizado en la literatura de síntesis y modificación de voz. La Tabla 7 resume la comparación.

Cuadro 7: Comparación con WORLD sobre las mismas señales de entrada.

Métrica	Sintetizador LCDK	WORLD	Diferencia
Distorsión log-espectral media (dB)	11,59	8,81	2,78
Desviación media de CPPS (dB)	7,08	4,49	2,59

WORLD alcanza mayor fidelidad respecto de la señal original, con una distorsión log-espectral 2,78 dB menor de forma consistente en las tres vocales. Ambos sistemas, no obstante, sobreestiman la periodicidad de la señal: como muestra la Figura 18, el CPPS de las señales reconstruidas supera al de las señales originales en ambos casos, aunque la desviación del sintetizador implementado (7,08 dB) es mayor que la de WORLD (4,49 dB).

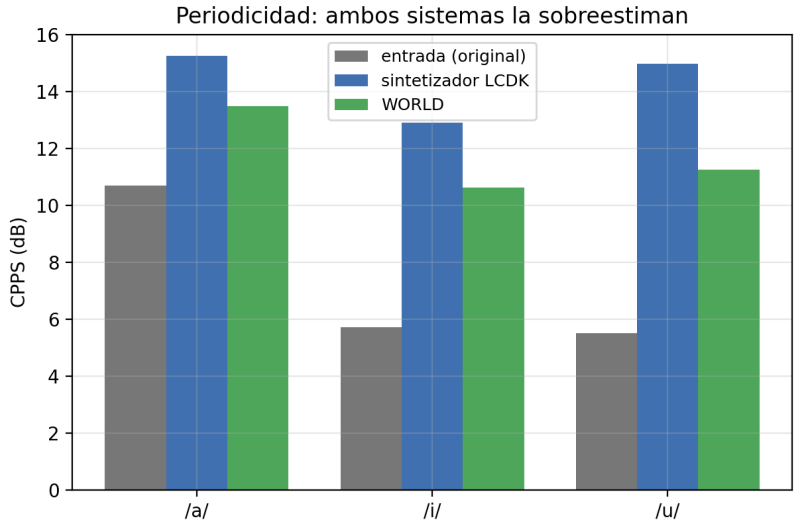


Figura 18: Comparación de periodicidad entre la señal de entrada y las reconstrucciones de ambos sistemas.

Esta diferencia es atribuible a la sofisticación del modelo de excitación: WORLD estima la aperiodicidad por banda de frecuencia mediante el algoritmo D4C, mientras que la implementación desarrollada emplea un único coeficiente de periodicidad global por trama, siguiendo el enfoque de excitación mixta de los vocoders LPC clásicos [23].

6. Capítulo 6: Conclusiones

6.1. Cumplimiento de los objetivos

6.1.1. Objetivo específico 1

El primer objetivo específico planteaba implementar en hardware algoritmos de procesamiento digital de señal para extraer parámetros relevantes de la calidad vocal a partir de una señal de voz.

Este objetivo se cumplió mediante la implementación, sobre la plataforma LCDK, de una cadena de análisis que extrae por trama la frecuencia fundamental, un coeficiente de periodicidad y los coeficientes del filtro de predicción lineal de orden 18. La corrección de la implementación se verificó mediante una prueba de filtrado inverso que arrojó un error de reconstrucción nulo, lo que confirma que el análisis captura efectivamente la información necesaria para reconstruir la señal.

La estimación resultó robusta frente a los cuatro materiales de prueba empleados, incluyendo habla continua, tras la incorporación de tres criterios que se identificaron como necesarios durante el desarrollo: el umbral de sonoridad relativo al nivel de la captura, la normalización de la autocorrelación por cobertura, y la restricción de la búsqueda a máximos locales estrictos.

6.1.2. Objetivo específico 2

El segundo objetivo planteaba implementar algoritmos que permitieran modificar de manera controlada dichos parámetros para alterar suavemente la calidad vocal de la señal reconstruida.

Este objetivo también se cumplió, y los resultados permiten precisar en qué sentido el control obtenido es efectivamente controlado y suave.

Respecto del carácter *controlado*, se demostró que cada modificación afecta principalmente a la métrica acústica asociada al mecanismo fisiológico que emula. La modificación de voz tensa altera la pendiente espectral en más de 3 dB/octava a lo largo del barrido mientras el CPPS varía menos de 0,06 dB, lo que constituye evidencia de control paramétrico selectivo.

Respecto del carácter *suave*, la caracterización mediante barrido de intensidad estableció que ambas modificaciones responden de forma monótona y continua al parámetro de control, si bien siguiendo leyes distintas: lineal para la voz tensa, con coeficientes de determinación superiores a 0,999, y logarítmica saturante para la voz soplada.

6.1.3. Objetivo general

El objetivo general se alcanzó: se implementó sobre hardware un sistema capaz de obtener parámetros propios de la calidad vocal y de modificarlos deliberadamente. El sistema opera de forma autónoma sobre la plataforma LCDK, empleando un 2,8 % de la memoria externa disponible y procesando capturas de 5 segundos en menos de 10 segundos.

6.2. Principales hallazgos

6.2.1. Las dos modificaciones siguen leyes de control distintas

El hallazgo más relevante de la caracterización paramétrica es que ambas modificaciones, pese a implementarse mediante mecanismos análogos y controlarse mediante

un parámetro de idéntica definición, responden de forma cualitativamente distinta.

La voz tensa presenta un control esencialmente lineal, con una pendiente de aproximadamente 1,64 dB/octava por unidad de intensidad y una dispersión entre vocales inferior al 4 %. La voz soplada, en cambio, satura: cerca de la mitad del recorrido total del CPPS se produce en el primer cuarto del rango de intensidad.

Este comportamiento admite explicación: el CPPS se define sobre una escala logarítmica y cuantifica una relación entre energías, de modo que su respuesta a la adición progresiva de ruido es necesariamente comprimida. La consecuencia práctica es de diseño: una escala lineal de control de breathiness resulta inadecuada para un usuario, y convendría una distribución logarítmica del parámetro.

6.2.2. El rango de control depende del material fonético

El recorrido efectivo del control de breathiness depende sustancialmente del material sobre el que se aplica: entre 6 y 8 dB de CPPS en vocales sostenidas, frente a 1,42 dB en habla continua.

La causa reside en el punto de partida de la métrica. El habla continua es intrínsecamente menos periódica por incorporar segmentos sordos y transiciones, de modo que su CPPS inicial es considerablemente menor y el margen disponible para reducirlo es más estrecho. Esta observación es coherente con la práctica establecida en la literatura clínica de emplear vocales sostenidas para la evaluación de calidad vocal.

6.2.3. Divergencia entre medición objetiva y percepción

La comparación entre la caracterización objetiva y la evaluación perceptual reveló divergencias en ambos sentidos, lo que constituye posiblemente el resultado de mayor interés conceptual del trabajo.

Para la voz tensa, la métrica objetiva muestra un control excelente —lineal, consistente entre vocales, con coeficientes de determinación superiores a 0,999— mientras que la detección perceptual en condiciones ciegas resultó errática y acompañada de falsos positivos. La métrica captura un cambio que el oído no registra de forma fiable.

Para la voz soplada en habla continua se observó la situación inversa: una caída de CPPS de apenas 0,91 dB, marginal desde el punto de vista objetivo, resultó perceptualmente evidente.

Ambas observaciones apuntan en la misma dirección: ninguna de las dos formas de evaluación caracteriza por sí sola el comportamiento del sistema, y su divergencia es informativa. Este resultado es consistente con la literatura que documenta la variabilidad de las relaciones entre parámetros fisiológicos, correlatos acústicos y percepción [5].

6.2.4. Criterios para la elección del paradigma de implementación

La comparación entre la síntesis paramétrica y el efecto de audio aplicado directamente permite establecer criterios de diseño concretos.

Para modificaciones de naturaleza espectral, como la voz tensa, ambos paradigmas entregan un control equivalente. Dado que el efecto directo requiere aproximadamente la quinta parte del tiempo de procesamiento y no introduce la degradación propia de la resíntesis, resulta preferible en este caso.

Para modificaciones que alteran la naturaleza de la fuente, como la voz soplada, la síntesis paramétrica conserva una ventaja sustancial: un rango de control 4,8 veces mayor. La diferencia se explica por el punto de inyección del ruido, que en la síntesis se incorpora a la excitación, antes del filtro que modela el tracto vocal.

El criterio de diseño que se desprende es que la elección del paradigma debe atender a la naturaleza de la modificación perseguida, y no establecerse a priori.

6.2.5. Modelo de costo computacional

La instrumentación del sistema permitió construir un modelo de costo que descompone el tiempo de análisis en un término proporcional al número total de tramas y otro proporcional al número de tramas sonoras, con un coeficiente de determinación de 0,9982 sobre los cuatro casos evaluados.

El modelo identifica la estimación de la frecuencia fundamental como el componente dominante: resulta 2,4 veces más costosa por trama que el análisis LPC completo de orden 18, y representa entre el 60 % y el 70 % del tiempo de análisis. Este resultado señala con precisión dónde debería concentrarse el esfuerzo de optimización si se persiguiera operación en tiempo real.

6.3. Limitaciones

6.3.1. Calidad de la síntesis

La comparación con WORLD [24] estableció que el sistema implementado presenta una distorsión log-espectral 2,78 dB superior. Esta diferencia es atribuible a la sofisticación del modelo de excitación: WORLD estima la aperiodicidad de forma independiente para cada banda de frecuencia, mientras que la implementación desarrollada emplea un único coeficiente global por trama.

Ambos sistemas comparten, no obstante, una limitación común: sobreestiman la periodicidad de la señal reconstruida. Es una característica conocida de la síntesis paramétrica basada en excitación determinista.

6.3.2. Alcance de la evaluación perceptual

La evaluación perceptual realizada involucró un único evaluador, sin repeticiones y con cinco a seis estímulos por condición. No permite establecer conclusiones estadísticamente sustentadas, y sus resultados deben interpretarse como indicativos de tendencias.

La inclusión de estímulos de control permitió, sin embargo, detectar una tasa de falsos positivos no despreciable para la voz tensa, información que una evaluación sin controles no habría proporcionado y que resultó determinante para interpretar correctamente los aciertos.

6.3.3. Material de prueba

Las señales empleadas corresponden a una única hablante para las vocales sostenidas y a un hablante distinto para el habla continua, dado que ambos materiales provienen de fuentes independientes (VPLab del AC3E y repositorio del curso de Laboratorio de Procesamiento Digital de Señales, respectivamente). Esta diferencia introduce una fuente adicional de variabilidad entre ambos materiales, que se suma a la distinción de naturaleza fonética discutida en la Sección 5.3: las diferencias observadas en el rango de control entre vocales sostenidas y habla continua no pueden atribuirse de forma inequívoca al tipo de material, dado que también podrían estar parcialmente asociadas a las características vocales de cada hablante.

La consistencia observada entre las tres vocales, producidas por la misma hablante, sugiere que el comportamiento del sistema no depende críticamente del material fonético en sí, pero la generalización a distintos hablantes, rangos de frecuencia fundamental y condiciones de grabación queda fuera del alcance de lo verificado.

6.3.4. Operación fuera de tiempo real

El sistema procesa buffers completos y no opera muestra a muestra. Los resultados de costo computacional sitúan la cadena de análisis y una síntesis en 0,97 veces el tiempo real, es decir, en el límite de la viabilidad, sin margen para las operaciones adicionales que una implementación en tiempo real requeriría.

6.4. Trabajo futuro

6.4.1. Excitación mixta multibanda

La brecha de calidad respecto de WORLD se atribuye al modelo de excitación de banda única. La incorporación de un coeficiente de mezcla independiente por banda de frecuencia, siguiendo el esquema de MELP [23], constituye la mejora de mayor impacto esperado sobre la calidad de la síntesis, y su costo computacional adicional es acotado.

6.4.2. Optimización hacia tiempo real

El modelo de costo identifica la estimación de F_0 como el objetivo prioritario. Entre las estrategias aplicables se cuentan la reducción del rango de búsqueda mediante seguimiento del valor estimado en tramas previas, el cálculo de la autocorrelación mediante transformada rápida de Fourier, y el submuestreo de la señal para la etapa de detección de *pitch*, dado que la información de periodicidad reside en las bajas frecuencias.

6.4.3. Ampliación de la evaluación perceptual

Una evaluación con múltiples oyentes, repeticiones de cada estímulo y un diseño de comparación por pares permitiría establecer umbrales perceptuales con respaldo estadístico. La divergencia observada entre métrica y percepción justifica particularmente este esfuerzo.

6.4.4. Incorporación de otras dimensiones de calidad vocal

El sistema implementa dos de las dimensiones descritas en la literatura. La incorporación de la aspereza, asociada a irregularidad ciclo a ciclo y caracterizada mediante *jitter* y *shimmer*, completaría las tres dimensiones de la escala perceptual RBH [16], ampliando la utilidad del sistema como generador de estímulos controlados.

6.4.5. Validación con material clínico

La aplicación del sistema sobre grabaciones de voces patológicas, y la comparación de las métricas obtenidas con valoraciones perceptuales de especialistas, permitiría evaluar su utilidad efectiva en el contexto clínico que motiva este tipo de desarrollos.

6.4.6. Migración a una plataforma de mayor capacidad de cómputo

La plataforma Jacinto 7, descartada durante el desarrollo de este trabajo por dificultades de inicialización del hardware de audio (Sección 4.1), no se descarta como plataforma de destino para trabajos futuros. Su capacidad de cómputo considerablemente mayor a la de la LCDK la posiciona como una opción atractiva para una eventual migración del sintetizador –o de los efectos de audio bajo el paradigma directo– hacia una implementación en tiempo real, objetivo que en la plataforma actual se encuentra en el límite de la viabilidad (Sección 5.5).

6.5. Comentarios finales

El trabajo demuestra que es posible implementar sobre hardware de bajo costo un sistema de análisis y modificación de calidad vocal con control paramétrico caracterizado y medible. El sistema no alcanza la calidad de reconstrucción de un vocoder del estado del arte, pero ofrece control selectivo sobre dos dimensiones de calidad vocal con un costo computacional compatible con la plataforma empleada.

Más allá del sistema en sí, la caracterización realizada aporta observaciones de interés: la existencia de leyes de control distintas para cada modificación, la dependencia del rango efectivo respecto del material fonético, y la divergencia entre medición objetiva y percepción. Estas observaciones son consecuencia de haber caracterizado el sistema mediante barrido paramétrico y evaluación perceptual con controles, en lugar de limitarse a verificar su funcionamiento en un punto de operación único.

Referencias

[1] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice Hall, 1978.

[3] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 583–590, 1971.

[4] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[5] J. Kreiman, Y.-L. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan, "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," *The Journal of the Acoustical Society of America*, vol. 132, pp. 2625–2632, 10 2012.

[6] S. A. Memon, "Acoustic correlates of the voice qualifiers: A survey," 2020.

[7] H. Dhamyal and R. Singh, "Objective measurements of voice quality," 2024.

[8] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[9] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech and Hearing Research*, vol. 37, pp. 769–778, 1994.

[10] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech," *Journal of Speech and Hearing Research*, vol. 39, no. 2, pp. 311–321, 1996.

[11] M. Fröhlich, D. Michaelis, and H. W. Strube, "Acoustic breathiness measures in the description of pathologic voices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1998, pp. 937–940.

[12] D. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density," in *Analysis of breathy, modal and pressed phonation based on low frequency spectral density*. International Speech Communication Association, 2013, pp. 3206–3210.

[13] M. Garellek, R. Samlan, B. R. Gerratt, and J. Kreiman, "Modeling the voice source in terms of spectral slopes," *The Journal of the Acoustical Society of America*, vol. 139, pp. 1404–1410, 3 2016.

[14] G. D. Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech and Hearing Research*, vol. 36, pp. 254–266, 1993.

[15] L. W. Lopes, E. S. da Silva Sousa, A. C. F. da Silva, I. M. da Silva, M. A. A. de Paiva, V. J. D. Vieira, and A. A. Almeida, "Cepstral measures in the assessment of severity of voice disorders," *CODAS*, vol. 31, 2019.

[16] K. Szklanny and P. Wrzeciono, "Relation of rbh auditory-perceptual scale to acoustic and electroglottographic voice analysis in children with vocal nodules," *IEEE Access*, vol. 7, pp. 41 647–41 658, 2019.

[17] J. Kreiman, Y. Lee, M. Garellek, R. Samlan, and B. R. Gerratt, "Validating a psychoacoustic model of voice quality," *The Journal of the Acoustical Society of America*, vol. 149, pp. 457–465, 1 2021.

[18] G. Klasmeyer, "The perceptual importance of selected voice quality parameters," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 1997, pp. 1615–1618.

[19] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[20] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal on Selected Topics in Signal Processing*, vol. 14, pp. 367–379, 2 2020.

[21] S. R. Kadiri, F. Javanmardi, and P. Alku, "Convolutional neural networks for classification of voice qualities from speech and neck surface accelerometer signals," in *Convolutional Neural Networks for Classification of Voice Qualities from Speech and Neck Surface Accelerometer Signals*, vol. 2022-September. International Speech Communication Association, 2022, pp. 5253–5257.

[22] M. Sambur, A. Rosenberg, L. Rabiner, and C. McGonegal, "On reducing the buzz in LPC synthesis," *The Journal of the Acoustical Society of America*, vol. 63, no. 3, pp. 918–924, 1978.

[23] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[24] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99D, pp. 1877–1884, 7 2016.

[25] K. I. Nordstrom, G. Tzanetakis, and P. F. Driessen, "Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, pp. 1087–1096, 2008.

[26] B. Opazo Campusano, "Voice quality modification using the WORLD vocoder," Master's thesis, Universidad Técnica Federico Santa María, 2022.

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

[28] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall, 2001.

[29] *OMAP-L138 LCDK User's Guide*, Texas Instruments, 2013, sPRUH80.

[30] G. Carrasco, "Laboratorio de procesamiento digital de señales elo314: bibliotecas para LCDK OMAP-L138," Repositorio Bitbucket, 2024, https://bitbucket.org/Gonzalo_AC3E/labdsp_elo314_utfsm.

[31] *TMS320C6748 Fixed- and Floating-Point Digital Signal Processor*, Texas Instruments, 2014, sPRS590.

[32] *TLV320AIC3106 Low-Power Stereo Audio Codec*, Texas Instruments, 2008, sLAS509.

[33] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Version 6.4.06, 2024, https://www.praat.org.

Anexos

Anexo A: Parámetros de configuración del sistema

Todos los parámetros de sintonía del sistema se encuentran centralizados en el archivo de cabecera `vq_config.h`. Las tablas siguientes recogen los valores empleados en las mediciones reportadas.

Cuadro 8: Configuración del códec y de la plataforma.

Parámetro	Valor	Descripción
`CODEC_FS`	16000 Hz	frecuencia de muestreo
`CODEC_INPUT_CFG`	`LINE_INPUT`	entrada de línea
`CODEC_ADC_GAIN`	36 dB	ganancia del amplificador de entrada
`CODEC_DAC_ATTEN`	0 dB	atenuación de salida
Reloj del núcleo	300 MHz	confirmado desde el archivo GEL

Cuadro 9: Parámetros del análisis y la síntesis.

Parámetro	Valor	Descripción
`VQ_SYNTH_BUF_SECONDS`	5	duración máxima de captura, s
`VQ_SYNTH_FRAME_MS`	25,0	ventana de análisis, ms
`VQ_SYNTH_HOP_MS`	10,0	paso entre ventanas, ms
`VQ_SYNTH_LPC_ORDER`	18	orden del filtro LPC
`VQ_SYNTH_PREEMPH`	0,97	coeficiente de pre-énfasis
`VQ_SYNTH_GAMMA_BW`	1,00	ensanchamiento espectral (desactivado)
`VQ_SYNTH_F0_MIN_HZ`	70	límite inferior de búsqueda de F_0
`VQ_SYNTH_F0_MAX_HZ`	400	límite superior de búsqueda de F_0
`VQ_SYNTH_F0_ENERGY_REL`	0,02	umbral de sonoridad, relativo al pico
`VQ_SYNTH_F0_PEAK_TH`	0,30	umbral de periodicidad

Cuadro 10: Parámetros del efecto de voz soplada.

Parámetro	Valor	Descripción
`VQ_B_NOISE_DB`	-22 dB	nivel del ruido relativo a la voz
`VQ_B_NOISE_FC`	3000 Hz	centro de la banda de ruido
`VQ_B_NOISE_BW`	2000 Hz	ancho de banda del ruido
`VQ_B_ENV_TAU`	0,6 ms	constante de la envolvente rápida
`VQ_B_LS_FC`	400 Hz	*low-shelf*: frecuencia
`VQ_B_LS_DB`	+4 dB	*low-shelf*: realce
`VQ_B_HS_FC`	3000 Hz	*high-shelf*: frecuencia
`VQ_B_HS_DB`	-6 dB	*high-shelf*: atenuación
`VQ_B_RMS_TAU`	50 ms	ventana de RMS corrido para el SNR

Cuadro 11: Parámetros del efecto de voz tensa.

Parámetro	Valor	Descripción
`VQ_P_HS_FC`	2500 Hz	*high-shelf*: frecuencia
`VQ_P_HS_DB`	+6 dB	*high-shelf*: realce
`VQ_P_COMP_THRESH_DB`	−25 dBFS	umbral del compresor
`VQ_P_COMP_RATIO`	4:1	razón de compresión
`VQ_P_COMP_ATTACK`	0,8 ms	tiempo de ataque
`VQ_P_COMP_RELEASE`	50 ms	tiempo de liberación
`VQ_P_COMP_MAKEUP_DB`	+3 dB	ganancia de compensación

Anexo B: Script de análisis acústico

Las métricas acústicas se calcularon mediante el siguiente script de Praat, que procesa un archivo de audio y entrega el CPPS, la relación armónicos-ruido, las estadísticas de F_0 y las medidas de perturbación.

```
form Analisis
    sentence archivo
endform

Read from file: archivo$

# CPPS: implementacion oficial de Praat sobre PowerCepstrogram
To PowerCepstrogram: 60, 0.002, 5000, 50
cpps = Get CPPS: "yes", 0.01, 0.001, 60, 330, 0.05,
    ... "Parabolic", 0.001, 0.05, "Exponential decay", "Robust slow"
Remove

selectObject: 1
To Pitch: 0.0, 60, 350
pitchMean = Get mean: 0, 0, "Hertz"
pitchSD = Get standard deviation: 0, 0, "Hertz"
Remove

selectObject: 1
To Harmonicity (cc): 0.01, 60, 0.1, 1.0
hnr = Get mean: 0, 0
Remove

selectObject: 1
To PointProcess (periodic, cc): 60, 350
jitter = Get jitter (local): 0, 0, 0.0001, 0.02, 1.3
plusObject: 1
shimmer = Get shimmer (local): 0, 0, 0.0001, 0.02, 1.3, 1.6

writeInfoLine: "CPPS_dB ", fixed$(cpps, 3)
appendInfoLine: "HNR_dB ", fixed$(hnr, 3)
appendInfoLine: "F0_mean_Hz ", fixed$(pitchMean, 2)
appendInfoLine: "F0_sd_Hz ", fixed$(pitchSD, 3)
appendInfoLine: "jitter_local_pct ", fixed$(jitter*100, 4)
```

```
appendInfoLine: "shimmer_local_pct ", fixed$(shimmer*100, 4)
```

Anexo C: Resultados completos del barrido paramétrico

Cuadro 12: CPPS (dB) por nivel de intensidad, modificación de voz soplada, sintetizador.

k	/a/	/i/	/u/	frase
0,00	17,77	13,59	15,95	7,99
0,25	14,20	11,72	11,74	7,45
0,50	12,91	10,24	10,28	6,84
0,75	12,14	9,18	9,42	6,48
1,00	11,56	8,41	8,84	6,25
1,50	10,77	7,46	8,07	5,93
2,00	10,19	7,79	7,64	5,67

Cuadro 13: Pendiente espectral (dB/oct) por nivel de intensidad, modificación de voz tensa, sintetizador.

k	/a/	/i/	/u/
0,00	−2,79	−1,40	−3,51
0,25	−2,35	−0,95	−3,05
0,50	−1,93	−0,54	−2,62
0,75	−1,51	−0,13	−2,19
1,00	−1,11	+0,28	−1,78
1,50	−0,31	+1,09	−0,95
2,00	+0,44	+1,88	−0,15

Cuadro 14: Desglose del tiempo de procesamiento por etapa (ms). Capturas de 5 segundos sobre C6748 a 300 MHz.

Señal	Análisis	Neutral	Breathy	Pressed	Barrido	Total
/a/	4324,7	417,1	672,3	624,3	3294,7	9333,1
/i/	4493,2	416,7	671,5	628,4	3289,0	9498,8
/u/	4567,7	418,4	673,5	632,0	3294,9	9586,4
frase	2447,1	407,3	662,5	566,8	3164,3	7248,1

Cuadro 15: Tiempo de procesamiento del paradigma de efecto directo (ms).

Señal	Breathy	Pressed	Barrido	Total
/a/	250,6	208,5	1227,2	1686,3
/i/	250,6	218,9	1227,2	1696,7
/u/	250,6	218,9	1226,6	1696,1
frase	250,6	168,8	1177,9	1597,3

Anexo D: Resultados de la evaluación perceptual

Cuadro 16: Resultados de la prueba de escucha a ciegas. Los estímulos se presentaron en orden aleatorio; la columna *condición* indica la correspondencia real, desconocida por el evaluador durante la prueba.

Estímulo	Condición real	Respuesta	Resultado
1	neutral	pressed	incorrecto
2	breathy $k = 1{,}50$	breathy	correcto
3	breathy $k = 2{,}00$	breathy	correcto
4	neutral	pressed (duda)	incorrecto
5	breathy $k = 0{,}75$	breathy	correcto
6	breathy $k = 0{,}25$	pressed	incorrecto
7	breathy $k = 0{,}50$	breathy	correcto
8	pressed $k = 1{,}00$	neutral	incorrecto
9	pressed $k = 0{,}25$	pressed	correcto
10	pressed $k = 0{,}75$	pressed (duda)	correcto
11	pressed $k = 0{,}50$	pressed (duda)	correcto
12	pressed $k = 2{,}00$	pressed	correcto
13	neutral	neutral	correcto
14	neutral	neutral	correcto
15	pressed $k = 1{,}50$	pressed	correcto
16	breathy $k = 1{,}00$	breathy	correcto
17	neutral	neutral	correcto
Aciertos: 13 de 17 (76 %)			
Falsos positivos sobre controles neutrales: 2 de 5			

Anexo E: Señales temporales y espectros

Este anexo recoge, para cada señal de prueba y cada paradigma de implementación, la forma de onda temporal y el espectro promediado de todas las condiciones evaluadas. Los tramos temporales corresponden a un mismo instante de cada señal, de modo que las formas de onda son directamente comparables entre condiciones. Los espectros se calcularon sobre el segmento central de cada señal y se normalizaron a su máximo.

E.1 Paradigma de síntesis paramétrica

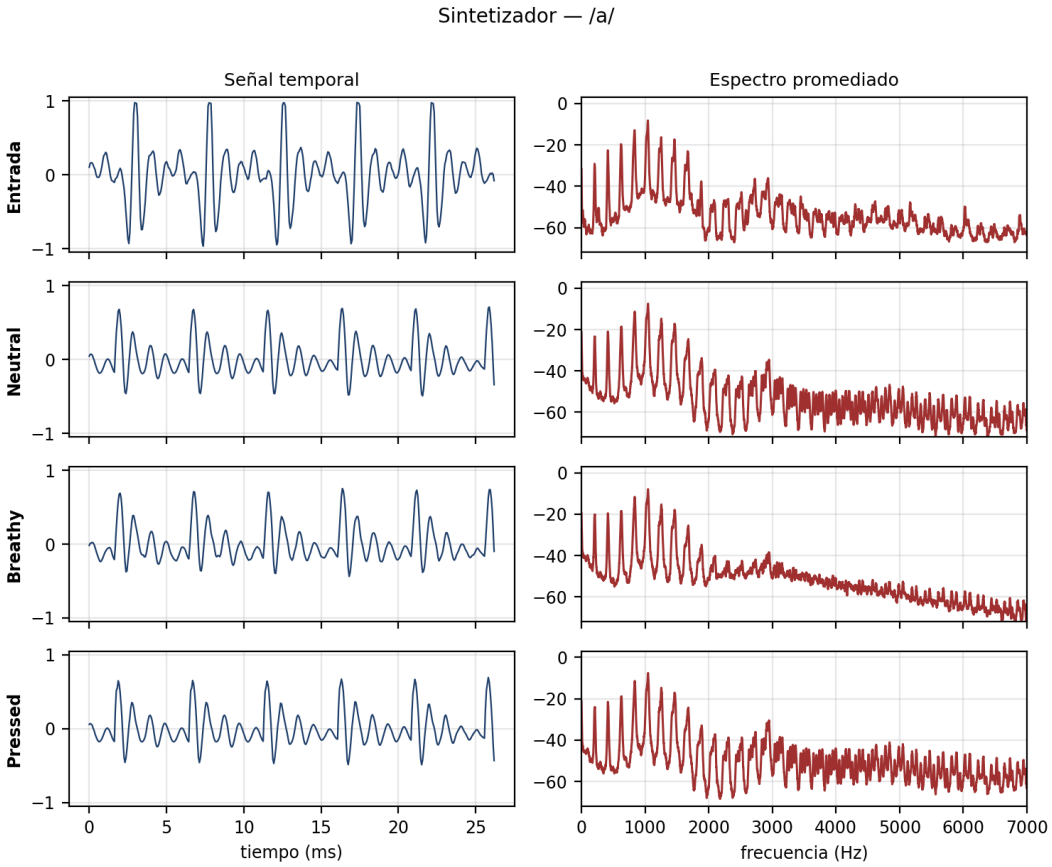


Figura 19: Vocal /a/ procesada mediante el sintetizador. De arriba abajo: señal de entrada, reconstrucción sin modificación, modificación de voz soplada y modificación de voz tensa.

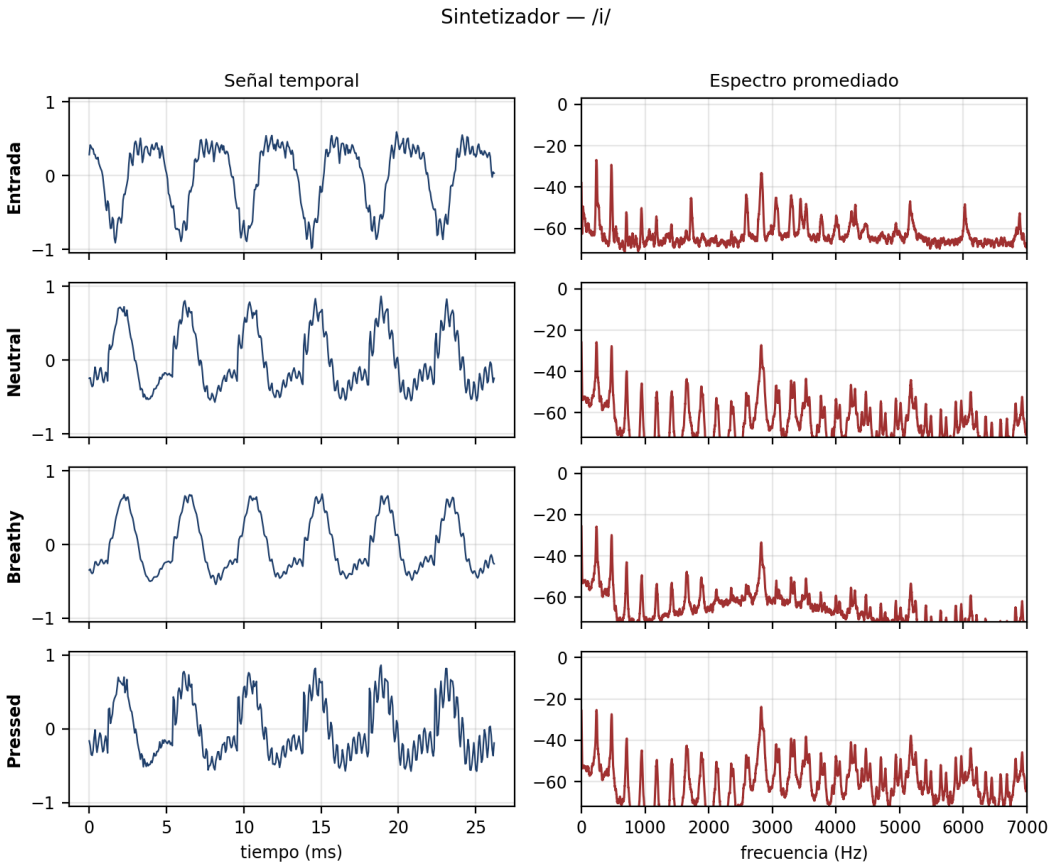


Figura 20: Vocal /i/ procesada mediante el sintetizador.

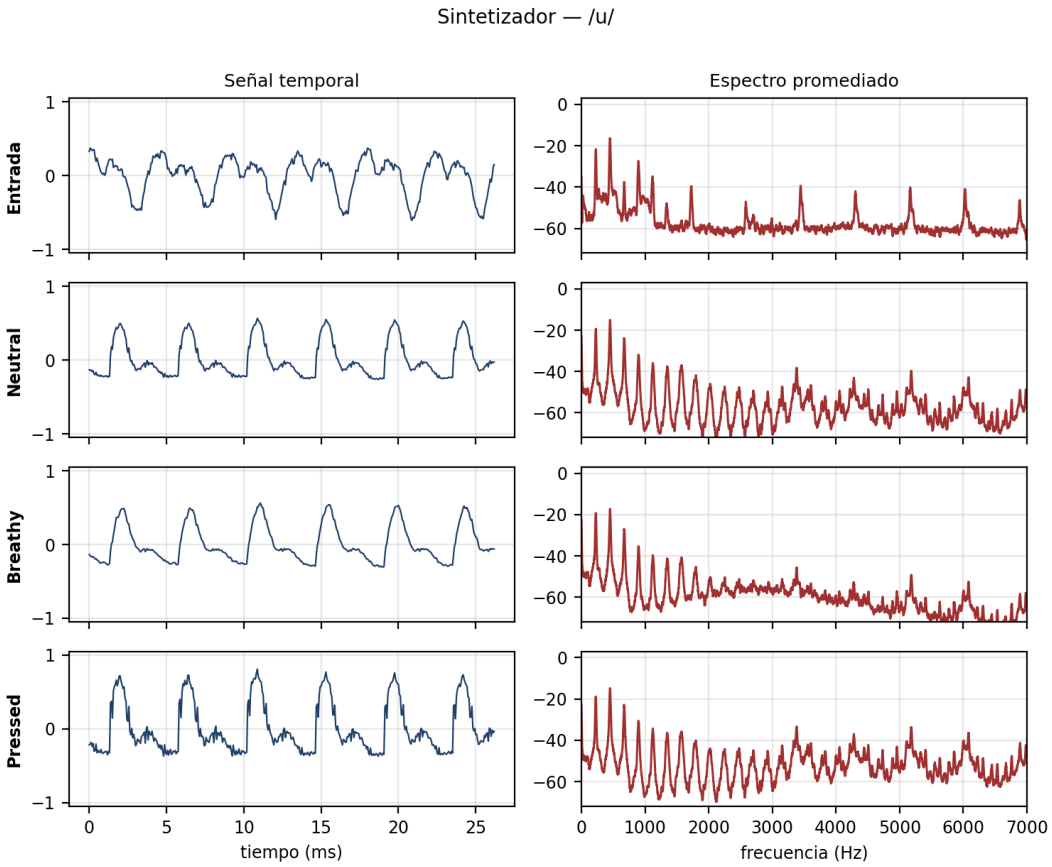


Figura 21: Vocal /u/ procesada mediante el sintetizador.

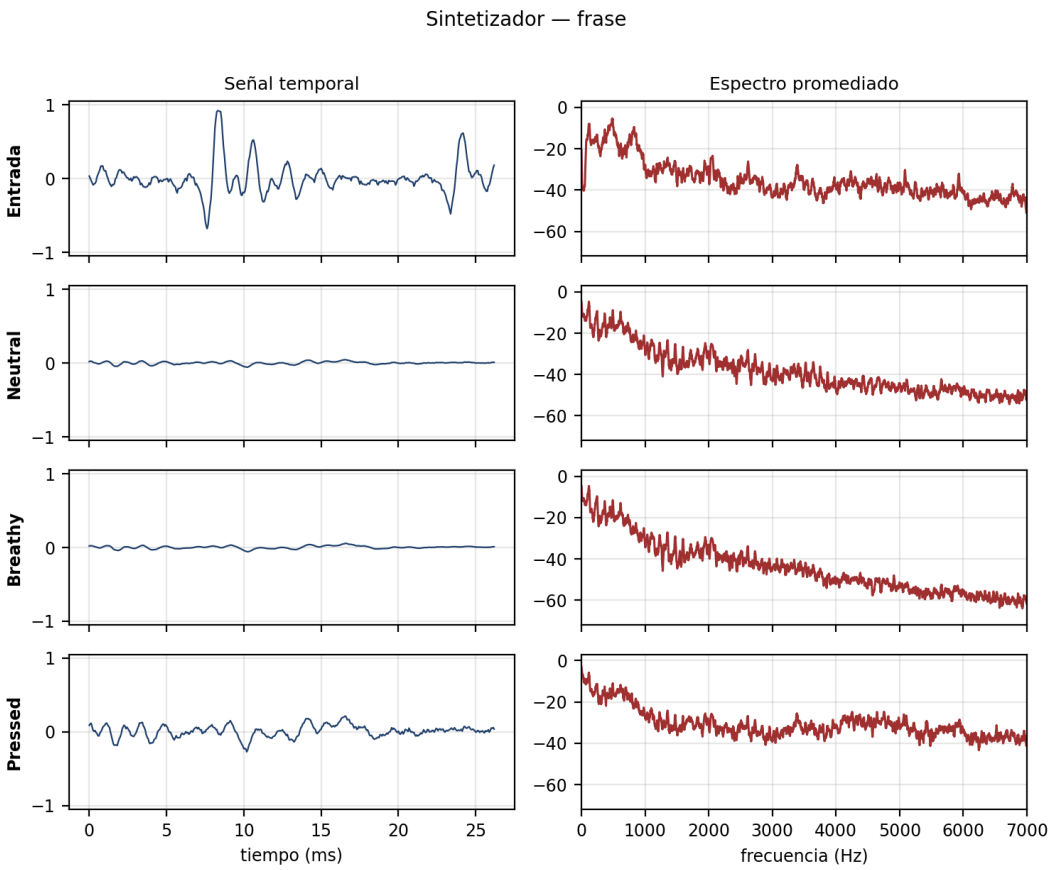


Figura 22: Frase de habla continua procesada mediante el sintetizador.

E.2 Paradigma de efecto de audio

En este paradigma no existe condición *neutral* diferenciada: al no haber etapa de resíntesis, la señal sin modificar es la propia señal de entrada.

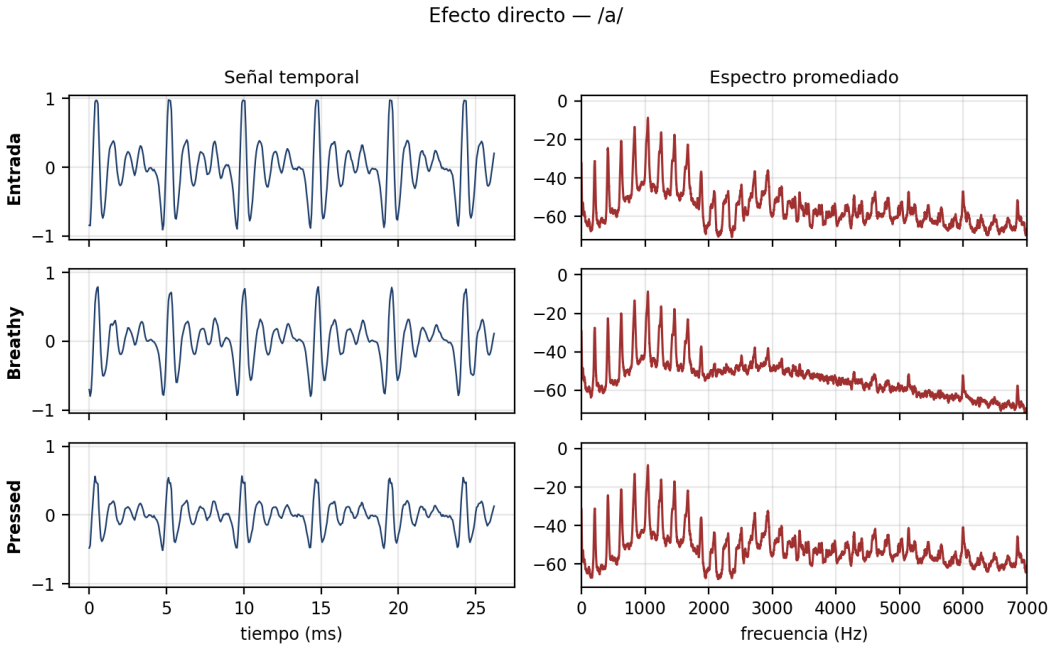


Figura 23: Vocal /a/ procesada mediante efectos aplicados directamente sobre la señal capturada.

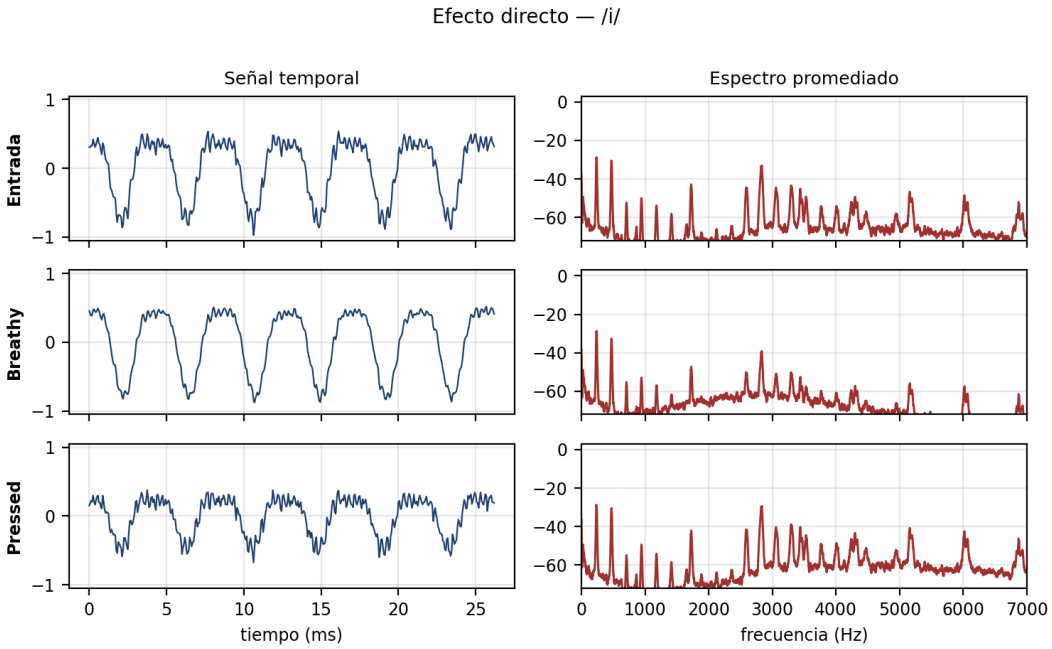


Figura 24: Vocal /i/ procesada mediante efectos directos.

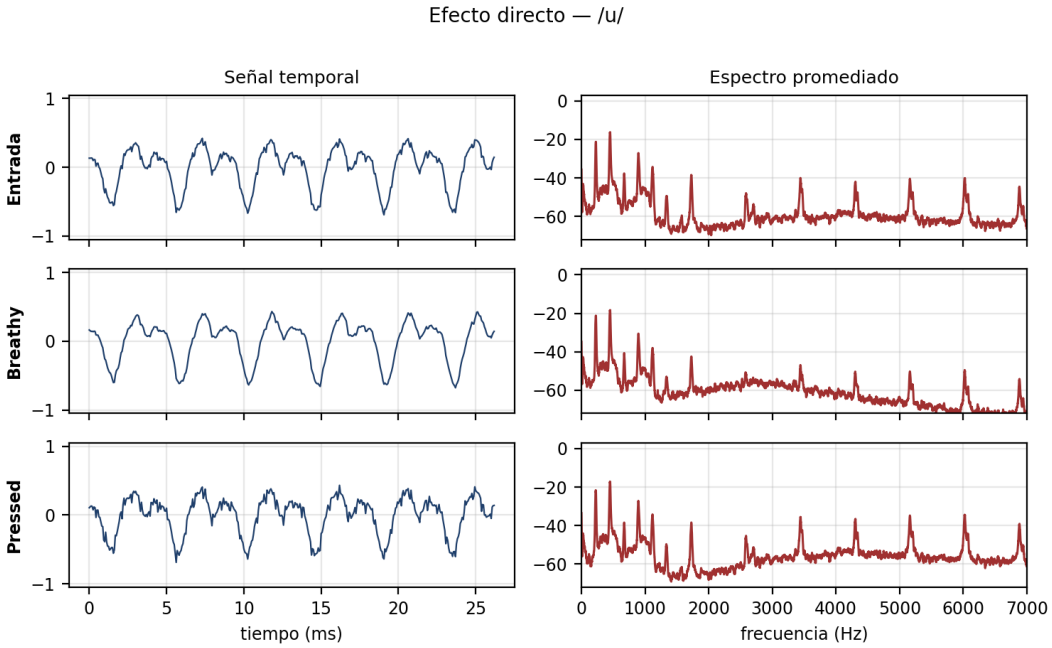


Figura 25: Vocal /u/ procesada mediante efectos directos.

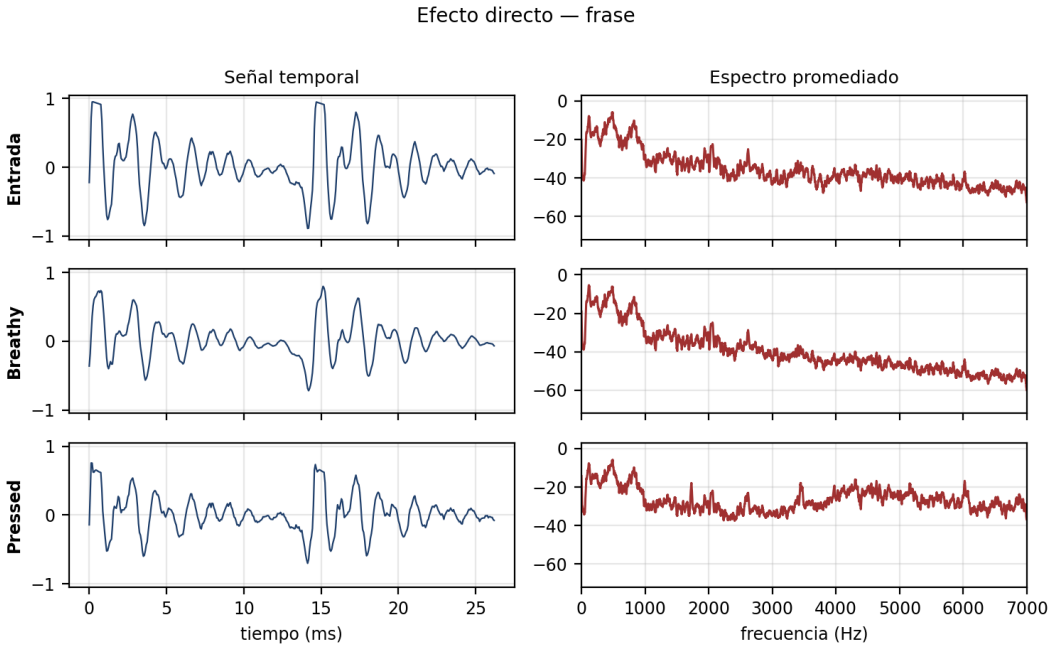


Figura 26: Frase de habla continua procesada mediante efectos directos.