

Tecnológico de Costa Rica  
Escuela de Ingeniería Electrónica



## Reconocimiento automático de señales peatonales accesibles usando un enfoque adaptativo

Documento de tesis sometido a consideración para optar por el grado académico de  
Maestría en Electrónica con Énfasis en Procesamiento Digital de Señales

Juan Manuel Fonseca Solís

Cartago, 26 de abril de 2018

Esta obra está bajo una licencia [Creative Commons](#)  
“Reconocimiento-NoCommercial-SinObraDerivada 4.0 Inter-  
nacional”.





Declaro que el presente documento de tesis ha sido realizado enteramente por mi persona, utilizando y aplicando literatura referente al tema e introduciendo conocimientos y resultados experimentales propios.

En los casos en que he utilizado bibliografía he procedido a indicar las fuentes mediante las respectivas citas bibliográficas. En consecuencia, asumo la responsabilidad total por el trabajo de tesis realizado y por el contenido del presente documento.

Juan Manuel Fonseca Solís

Cartago, 26 de abril de 2018

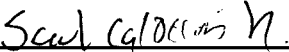
Céd: 4-0202-0807

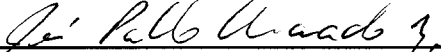


**Instituto Tecnológico de Costa Rica**  
**Escuela de Ingeniería Electrónica**  
**Tesis de Maestría**  
**Tribunal evaluador**

Tesis de maestría defendida ante el presente Tribunal Evaluador como requisito para optar por el grado académico de maestría, del Instituto Tecnológico de Costa Rica.

Miembros del Tribunal

  
M.Sc. Saúl Calderón Ramírez  
**Profesor lector**

  
Dr.-Ing. Pablo Alvarado Moya  
**Profesor lector**

  
Dr. Arturo Camacho Lozano  
**Director de Tesis**

Los miembros de este Tribunal dan fe de que la presente tesis ha sido aprobada y cumple con las normas establecidas por la Escuela de Ingeniería Electrónica.

Cartago, 26 de abril de 2018

**Instituto Tecnológico de Costa Rica**  
**Escuela de Ingeniería Electrónica**  
**Tesis de Maestría**  
**Tribunal evaluador**

**ACTA DE EVALUACIÓN**

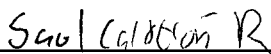
Tesis de maestría defendida ante el presente Tribunal Evaluador como requisito para optar por el grado académico de maestría, del Instituto Tecnológico de Costa Rica.

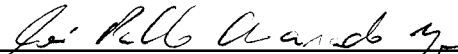
Estudiante: **Ing. Juan Manuel Fonseca Solís**

Nombre del Proyecto:

**“Reconocimiento de señales peatonales audibles usando un enfoque adaptativo”**

Miembros del Tribunal

  
M.Sc. Saúl Calderón Ramírez  
**Profesor lector**

  
Dr.-Ing. Pablo Alvarado Moya  
**Profesor lector**

  
Dr. Arturo Camacho Lozano  
**Director de Tesis**

Los miembros de este Tribunal dan fe de que la presente Tesis de Maestría para optar por el grado académico de Máster en Electrónica con Énfasis en Procesamiento Digital de Imágenes y cumple con las normas establecidas por la Escuela de Ingeniería Electrónica.

Nota Final de Tesis: -95-

Cartago, 26 de abril de 2018

# Resumen

Se presenta un algoritmo de reconocimiento de señales peatonales accesibles (APS, por sus siglas en inglés), un tipo de sonido emitido por los semáforos peatonales para habilitar el paso en los cruces peatonales. La detección automática de estas modulaciones de frecuencia es de interés para los encargados del control del tráfico porque permitiría realizar el reconocimiento de otros tonos audibles, como: las bocinas del tren, las sirenas de ambulancias y las alarmas de patrullas policiales, entre otros. Hasta ahora, autores previos han logrado reconocer los APS con éxito parcial, pues los diseños expuestos han presentado núcleos de reconocimiento musical subóptimos, umbrales de tono fijos incapaces de adaptarse al nivel cambiante de ruido de la calle y un procesamiento separado de contornos musicales continuos y discontinuos. Para resolver estos problemas se presenta un algoritmo que utiliza un diseño de núcleo de reconocimiento musical de tres armónicas con decaimiento proporcional a  $1/k^2$ , dos algoritmos de estimación dinámica del umbral de tono que varían según la relación señal-ruido (TS2Means y la media móvil exponencial) y la distancia de Mahalanobis con matrices de covarianza modeladas según los contornos musicales APS para soportar los momentos de ruido. Las mejores tasas de detección alcanzadas fueron de 93% de precisión, 89% de especificidad, 92% de sensibilidad, 92% de medida F y 80% del coeficiente de correlación de Matthew.

**Palabras clave:** procesamiento digital de señales, acústica, señales peatonales accesibles, reconocimiento de altura musical, distancia de Mahalanobis





# Abstract

An algorithm for the recognition of accessible pedestrian signals (APS), a type of sound emitted by pedestrian traffic lights to enable passage at pedestrian crossings, is presented. The automatic detection of these frequency modulations is of interest for traffic controllers because it allows the recognition of other audible tones, such as: train horns, ambulance sirens and police patrol alarms, among others. So far, previous authors have managed to recognize APSs with partial success, since the exposed designs have presented suboptimal musical recognition kernels, fixed tone thresholds unable to adapt to the changing level of street noise and a separate processing of continuous and discontinuous musical contours. To solve these problems, we present an algorithm that uses a three-harmonics musical recognition kernel design with a decay proportional to  $1/k^2$ , two algorithms for the dynamic estimation of the tone threshold, that vary according to the signal-to-noise ratio (TS2Means and the leaky integrator), and the Mahalanobis distance with covariance matrices modeled according to the APS musical contours for noise robustness. The best detection rates reached were 93% precision, 89% specificity, 92% recall, 92% F-score, and 80% Matthew's correlation coefficient.

**Keywords:** digital signal processing, acoustics, accessible pedestrian signals, pitch recognition, Mahalanobis distance



*A Dios, por haberme permitido llegar hasta este punto y haberme dado salud para lograr mis objetivos, además de brindarme su infinita gracia.*

*A mis padres, Juan y Cecilia, por su apoyo incondicional y por la paciencia que tuvieron conmigo durante el desarrollo de este trabajo.*

*A mis hermanos José y Diego, a quien les toca cablearlo y mecanizarlo, yo ya lo programé.*

*Y a los investigadores de este país, con la esperanza de que se generen nuevas empresas y puestos de trabajo en el campo del procesamiento de audio, y que las invenciones y vocaciones científicas no se reserven solo a los países desarrollados.*



# Agradecimientos

Agradezco a los ingenieros Mario Monge y Sharon Bejarano por haber compartido en el 2013 su estudio preliminar de los APS ante la clase TEIA-2600; a los investigadores Sebastián Ruiz y Arturo Camacho por haberme brindado la oportunidad de trabajar en el proyecto de *Reconocimiento de semáforos peatonales para ser usados en dispositivos móviles* (B6146); al Centro de Investigaciones en Tecnologías de la Información (CITIC) por haberme aceptado como investigador en dicho proyecto; a los profesores Yadira Solano, Marta Calderón, Kryscia Ramírez, Arturo Camacho y Luis Quesada por haberme recomendado para realizar estudios de posgrado; a la Escuela de Electrónica del TEC, por haberme admitido en su programa de maestría, aún sabiendo que mi formación inicial no es en ingeniería electrónica sino en computación e informática; a los profesores Pablo Alvarado y Saúl Calderón, por las correcciones realizadas en este trabajo y por la motivación transmitida para mejorar la calidad y la cantidad de las explicaciones —fue un honor tenerlos como miembros del tribunal evaluador y les agradezco haber compartido muchos de los métodos numéricos aquí empleados— y, por último, al profesor Arturo Camacho por haberme iniciado en el estudio del procesamiento digital de audio, por su orientación como director de tesis, por su rigurosidad en la buena redacción y por su guía en el uso correcto de los métodos del estado del arte.

“Nunca recibas una cruz sin besarla humildemente con agradecimiento” - S. L. Griñón de Monfort

Juan Manuel Fonseca Solís

Cartago, 7 de mayo de 2018



# Índice general

Índice de figuras	v
Índice de tablas	ix
<b>1 Introducción</b>	<b>1</b>
1.1 Trabajos previos	3
1.1.1 Crosswatch: detección de cruces peatonales usando celulares y visión por computadora	3
1.1.2 Zebralocalizer: identificación y localización de cruces peatonales	5
1.1.3 Monitoreo de comunidades de ranas: una aplicación del aprendizaje automático	6
1.1.4 Detección automática de vocalizaciones de la rana <i>Diasporus hylaeformis</i> en grabaciones de audio	7
1.1.5 Método para detectar silbidos, gemidos y otros contornos musicales	8
1.1.6 Algoritmo RASP	8
1.2 Definición del problema a resolver	10
1.2.1 Reconocimiento del tono musical	10
1.2.2 Umbrales de detección	14
1.2.3 Segmentos nulos en las plantillas	18
1.2.4 Robustez contra el ruido	21
1.3 Objetivos y estructura del documento	22
<b>2 Marco teórico</b>	<b>27</b>
2.1 Tamaño mínimo del inventariado	28
2.1.1 Ventana rectangular	28
2.1.2 Ventanas de Hann y de Hamming	29
2.2 Reconocimiento de la altura musical	31
2.2.1 Altura musical	31
2.2.2 Escala ERB	32
2.2.3 Reconocimiento de la altura musical	33
2.2.4 Umbralización dinámica en el tiempo	36
2.3 Media móvil exponencial	38
2.3.1 Costo computacional	41
2.4 Medidas de similitud	43

2.4.1	Definición de distancia . . . . .	44
2.4.2	Distancia euclidiana . . . . .	46
2.4.3	Distancia coseno . . . . .	46
2.4.4	Vectores aleatorios y matriz de covarianza . . . . .	46
2.4.5	Distancia de Mahalanobis . . . . .	48
2.4.6	Evaluación de las distancias analizadas . . . . .	48
2.4.7	Pseudoinversa de una matriz . . . . .	50
2.4.8	Análisis de componentes principales . . . . .	51
2.5	Tasas de detección . . . . .	52
2.6	Redes neuronales convolucionales para procesar sonido . . . . .	54
<b>3</b>	<b>Reconocimiento automático de señales peatonales accesibles usando un enfoque adaptativo</b>	<b>55</b>
3.1	Rediseño del núcleo . . . . .	56
3.2	Distancia de Mahalanobis para contornos musicales . . . . .	58
3.2.1	Recorte de las grabaciones de interés . . . . .	58
3.2.2	Cálculo de los contornos musicales . . . . .	60
3.2.3	Distribución de los contornos musicales . . . . .	60
3.2.4	Cálculo de las matrices de covarianza . . . . .	61
3.2.5	Matrices de covarianza sintéticas . . . . .	63
3.2.6	Tratamiento de la salida de la distancia de Mahalanobis . . . . .	68
3.2.7	Uso de la distancia de Mahalanobis en contornos musicales . . . . .	68
3.3	Umbral adaptativo de tono . . . . .	69
3.4	Implementación de la solución . . . . .	69
<b>4</b>	<b>Resultados y análisis</b>	<b>73</b>
4.1	Metodología de evaluación . . . . .	73
4.1.1	Pertinencia de la metodología existente . . . . .	74
4.1.2	Mejoramiento de los conjuntos de evaluación . . . . .	75
4.2	Matrices de puntajes del núcleo propuesto . . . . .	75
4.3	Escenarios de prueba . . . . .	77
4.3.1	Escenario 1: configuración original . . . . .	81
4.3.2	Escenario 2: metodología de evaluación mejorada . . . . .	82
4.3.3	Escenario 3: núcleo propuesto . . . . .	86
4.3.4	Escenario 4: núcleo y banco de núcleos propuestos . . . . .	91
4.3.5	Escenario 5: matrices de covarianza reales . . . . .	94
4.3.6	Escenario 6: matrices de covarianza sintéticas . . . . .	97
4.3.7	Escenario 7: TS2Means . . . . .	100
4.3.8	Escenario 8: EMA . . . . .	103
4.3.9	Escenario 9: las mejores combinaciones . . . . .	106
<b>5</b>	<b>Conclusiones</b>	<b>109</b>
	<b>Bibliografía</b>	<b>111</b>



---

<b>A</b>	<b>Datos usados en la comparación de distancias</b>	<b>117</b>
<b>B</b>	<b>Demostración de la equivalencia entre la distancia euclidiana sobre datos decorrelacionados y la distancia de Mahalanobis</b>	<b>119</b>
<b>C</b>	<b>Demostración de la definición de distancia de Mahalanobis</b>	<b>121</b>
C.1	Norma de Mahalanobis . . . . .	123



# Índice de figuras

1.1	Ejemplo de un dispositivo Novax DS100. . . . .	1
1.2	Cálculo del descriptor de vídeo de tasa de profundidad. . . . .	4
1.3	Núcleo musical para identificar la altura musical de un tono de 2 kHz. . . . .	8
1.4	Espectrograma del cucú. . . . .	12
1.5	Espectrograma del chirrido alto. . . . .	12
1.6	Espectrograma del chirrido bajo. . . . .	13
1.7	Núcleo armónico de 4 armónicas con $f_0 = 0.9$ kHz. . . . .	13
1.8	Núcleo uniarmónico con $f_0 = 2$ kHz. . . . .	13
1.9	Núcleo armónico (de 3 armónicas) impares con $f_0 = 1.750$ kHz. . . . .	14
1.10	Matriz de puntajes del evaluador cucú para distintas grabaciones. . . . .	15
1.11	Matriz de puntajes del evaluador chirrido alto para distintas grabaciones. . . . .	16
1.12	Matriz de puntajes del evaluador chirrido bajo para distintas grabaciones. . . . .	17
1.13	Diagrama del funcionamiento del algoritmo RASP. . . . .	19
1.14	Distribución de los 22 filtros del algoritmo RASP. . . . .	19
1.15	Ejemplo de los contornos de las plantillas APS. . . . .	19
1.16	Ejemplo de un tono cucú con un segmento de ruido en lugar de silencio. . . . .	21
1.17	Ejemplo del análisis de periodicidad de dos contornos musicales. . . . .	23
1.18	Espectrograma y contorno musical de un chirrido bajo ruidoso. . . . .	24
1.19	Espectrograma y contorno musical de un chirrido bajo limpio. . . . .	24
2.1	Ejemplo de una ventana rectangular, Hann y Hamming. . . . .	29
2.2	Lóbulos espectrales pertenecientes a una ventana rectangular de ancho $T_v$ . . . . .	30
2.3	Lóbulos espectrales pertenecientes a una ventana Hann de ancho $T_v$ . . . . .	31
2.4	Ejemplo de la magnitud espectral de la nota LA4 (440 Hz). . . . .	32
2.5	Ejemplo de tonos musicales distintos con una altura musical similar . . . . .	32
2.6	Distribución de frecuencias por posición en la cóclea, según la escala ERB. . . . .	33
2.7	Ilustración de la cóclea. . . . .	34
2.8	Ejemplo de núcleos SWIPE aplicados a la magnitud espectral de un tono. . . . .	35
2.9	Sonoridad de un seno cardinal. . . . .	36
2.10	Curvas isofónicas del oído humano . . . . .	37
2.11	Diagrama de Bode y diagrama de ceros de la media móvil. . . . .	40
2.12	Diagrama de Bode y diagrama de ceros de la EMA. . . . .	41
2.13	Retardos de grupo para la media móvil e EMA. . . . .	42
2.14	Comparación de los ordenes de duración lineales, logarítmicos y cuadráticos. . . . .	44

2.15	Corrida de la función <i>w2means</i> del algoritmo TS2Means. . . . .	45
2.16	Proyección ortogonal del vector $x$ sobre el vector $y$ . . . . .	47
2.17	Datos del viento y de la lluvia acumulada en el Aeropuerto Juan Santamaría. . . . .	49
3.1	Diagrama sobre el funcionamiento del algoritmo RASP mejorado. . . . .	55
3.2	Núcleo de tres armónicas propuesto para $f_0 = 1$ kHz. . . . .	57
3.3	Núcleo de siete armónicas propuesto para $f_0 = 1$ kHz. . . . .	57
3.4	Distribución de los 22 filtros del banco de filtros propuesto. . . . .	57
3.5	Ejemplos de desalineación de las anotaciones originales. . . . .	59
3.6	Cálculo del contorno musical de un chirrido alto. . . . .	60
3.7	Histograma de una sola grabación de chirrido bajo. . . . .	61
3.8	Histograma de todas las grabaciones cucú. . . . .	62
3.9	Histograma de todas las grabaciones de chirrido alto. . . . .	62
3.10	Histograma de todas las grabaciones de chirrido bajo. . . . .	63
3.11	Matriz de covarianza cucú obtenida al analizar las grabaciones. . . . .	64
3.12	Matriz de covarianza del chirrido alto obtenida al analizar las grabaciones. . . . .	64
3.13	Matriz de covarianza del chirrido bajo obtenida al analizar las grabaciones. . . . .	65
3.14	Diagonal de la matriz de covarianza del APS cucú. . . . .	66
3.15	Diagonal de la matriz de covarianza del APS chirrido alto. . . . .	67
3.16	Diagonal de la matriz de covarianza del APS chirrido bajo. . . . .	67
3.17	Matrices de covarianza sintéticas cucú, chirrido alto y chirrido bajo. . . . .	68
3.18	Normalización de la distancia de Mahalanobis. . . . .	68
3.19	Umbrales $\alpha$ calculados por TS2Means. . . . .	70
3.20	Umbrales $\alpha$ calculados por la EMA usando $\lambda = 0.99$ . . . . .	70
4.1	Anotaciones manuales y automáticas de una grabación de chirrido alto. . . . .	74
4.2	Señal de alerta de un chirrido alto tomada del estudio de Ruiz <i>et al.</i> . . . . .	75
4.3	Ejemplo de los conjuntos $A$ y $B$ calculados con el enfoque original. . . . .	76
4.4	Ejemplo de los conjuntos $A$ y $B$ calculados por el enfoque propuesto. . . . .	76
4.5	Puntajes de los núcleos propuestos de 3 armónicas para el cucú. . . . .	77
4.6	Puntajes de los núcleos propuestos de 3 armónicas para el chirrido alto. . . . .	78
4.7	Puntajes de los núcleos propuestos de 3 armónicas para el chirrido bajo. . . . .	78
4.8	Puntajes de los núcleos propuestos de 7 armónicas para el cucú. . . . .	79
4.9	Puntajes de los núcleos propuestos de 7 armónicas para el chirrido alto. . . . .	79
4.10	Puntajes de los núcleos propuestos de 7 armónicas para el chirrido bajo. . . . .	80
4.11	Contornos musicales y señales de alerta del escenario 1. . . . .	84
4.12	Espectrogramas para las grabaciones de cucú, chirrido alto y chirrido bajo. . . . .	85
4.13	Señales de alerta reportadas en el estudio de Ruiz <i>et al.</i> . . . . .	85
4.14	Contornos musicales y señales de alerta del escenario 2. . . . .	87
4.15	Contornos musicales y señales de alerta del escenario 3. . . . .	90
4.16	Contornos musicales y señales de alerta del escenario 4. . . . .	93
4.17	Contornos musicales y señales de alerta del escenario 5. . . . .	96
4.18	Contornos musicales y señales de alerta del escenario 6. . . . .	99
4.19	Contornos musicales y señales de alerta del escenario 7. . . . .	102

---

4.20 Contornos musicales y señales de alerta del escenario 8. . . . .	105
4.21 Contornos musicales y señales de alerta del escenario 9. . . . .	108



# Índice de tablas

1.1	Bancos de filtros RASP para cada tipo de APS. . . . .	9
1.2	Velocidad de desplazamiento lineal promedio de un adulto mayor. . . . .	10
1.3	Umbral fijos de tono del prototipo original de RASP. . . . .	18
1.4	Periodo detectado en distintas grabaciones según el umbral de tono. . . . .	22
2.1	Comparación de distancias euclidiana, coseno y de Mahalanobis . . . . .	50
2.2	Matriz de confusión de un clasificador binario. . . . .	52
3.1	Cantidad de muestras de entrenamiento usadas por tipo de APS. . . . .	59
4.1	Resumen de los escenarios propuestos para evaluar las mejoras propuestas. . . . .	81
4.2	Parámetros del escenario 1. . . . .	82
4.3	Tasas originales vs. escenario 1. . . . .	83
4.4	Tasas obtenidas en el escenario 1 con $q = 2$ para los chirridos. . . . .	83
4.5	Tasas de rendimiento promedio obtenidas en el escenario 2. . . . .	86
4.6	Parámetros del escenario 3. . . . .	87
4.7	Rendimiento del cucú en escenario 3 para 3 armónicas. . . . .	88
4.8	Rendimiento del chirrido alto en escenario 3 para 3 armónicas. . . . .	88
4.9	Rendimiento del chirrido bajo en escenario 3 para 3 armónicas. . . . .	88
4.10	Rendimiento del cucú en escenario 3 para 7 armónicas. . . . .	89
4.11	Rendimiento del chirrido alto en escenario 3 para 7 armónicas. . . . .	89
4.12	Rendimiento del chirrido bajo en escenario 3 para 7 armónicas. . . . .	89
4.13	Mejores tasas de rendimiento obtenidas en el escenario 3. . . . .	90
4.14	Parámetros del escenario 4. . . . .	91
4.15	Rendimiento del cucú en escenario 4. . . . .	91
4.16	Rendimiento del chirrido alto en escenario 4. . . . .	92
4.17	Rendimiento del chirrido bajo en escenario 4. . . . .	92
4.18	Mejores tasas de rendimiento obtenidas en el escenario 4. . . . .	92
4.19	Parámetros del escenario 5. . . . .	94
4.20	Rendimiento del cucú en escenario 5. . . . .	94
4.21	Rendimiento del chirrido alto en escenario 5. . . . .	95
4.22	Rendimiento del chirrido bajo en escenario 5. . . . .	95
4.23	Mejores tasas de rendimiento obtenidas en el escenario 5. . . . .	95
4.24	Parámetros del escenario 6. . . . .	97
4.25	Rendimiento del cucú en escenario 6. . . . .	97

4.26 Rendimiento del chirrido alto en escenario 6. . . . .	98
4.27 Rendimiento del chirrido bajo en escenario 6. . . . .	98
4.28 Mejores tasas de rendimiento obtenidas en el escenario 6. . . . .	98
4.29 Parámetros del escenario 7. . . . .	100
4.30 Rendimiento del cucú en escenario 7. . . . .	100
4.31 Rendimiento del chirrido alto en escenario 7. . . . .	101
4.32 Rendimiento del chirrido bajo en escenario 7. . . . .	101
4.33 Mejores tasas de rendimiento obtenidas en el escenario 7. . . . .	101
4.34 Parámetros del escenario 8. . . . .	103
4.35 Rendimiento del cucú en escenario 8. . . . .	103
4.36 Rendimiento del chirrido alto en escenario 8. . . . .	104
4.37 Rendimiento del chirrido bajo en escenario 8. . . . .	104
4.38 Mejores tasas de rendimiento obtenidas en el escenario 8. . . . .	104
4.39 Parámetros del escenario 9. . . . .	106
4.40 Rendimiento del cucú en escenario 9. . . . .	106
4.41 Rendimiento del chirrido alto en escenario 9. . . . .	107
4.42 Rendimiento del chirrido bajo en escenario 9. . . . .	107
4.43 Mejores tasas de rendimiento obtenidas en el escenario 9. . . . .	107
4.44 Resumen del rendimiento general obtenido en cada escenario de pruebas. . . . .	108
A.1 Datos del Aeropuerto Juan Santamaría desde 1990 hasta 2016. . . . .	118

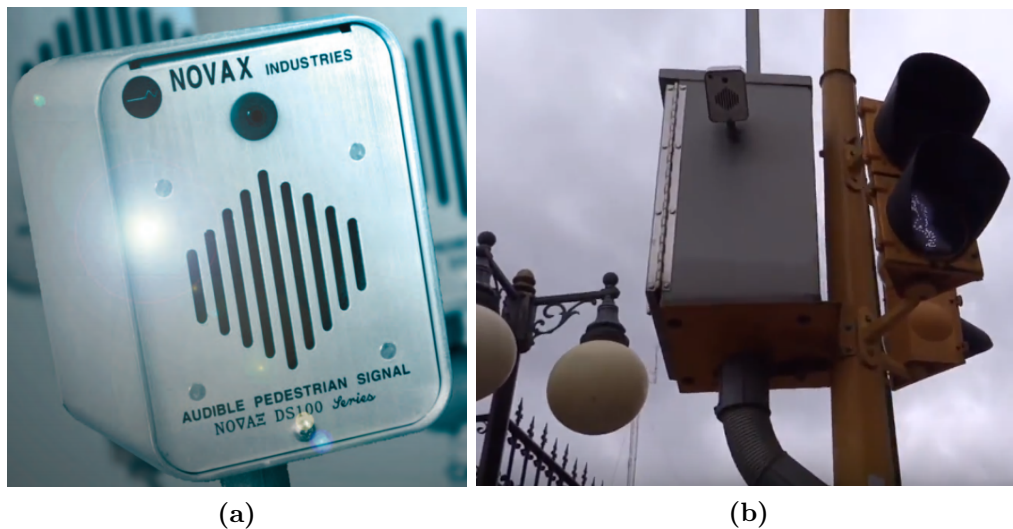


# Capítulo 1

## Introducción

Desde la década de los 90 es común encontrar en Costa Rica semáforos peatonales junto a los cruces peatonales ubicados en las rutas más transitadas de la GAM (Gran Área Metropolitana) y los centros de población de las zonas rurales. Estos semáforos emiten señales lumínicas y acústicas, conocidas como *señales peatonales accesibles* (APS, por sus siglas en inglés), que son producidas por unidades electrónicas autoajustables. La figura 1.1 presenta una de estas unidades de modelo Novax DS100. Estos dispositivos siguen estándares internacionales de calidad como el *Manual de administración uniforme de dispositivos del control de tráfico* (MUTCD, por sus siglas en inglés) y la *Guía para garantizar la accesibilidad en el derecho al tránsito* (PROWAG, por sus siglas en inglés) [1, 2].

Aunque los APS benefician especialmente a personas con discapacidades auditivas y vi-



**Figura 1.1:** (1.1a) Fotografía de un dispositivo Novax DS100 (tomada de la página del fabricante) y (1.1b) semáforo tipo cucú localizado cerca del Parque Central de San José [3].

suales, también son usados diariamente por el resto de la población, entre ellos los adultos mayores y los niños en edad escolar. Estos últimos también son vulnerables al cruzar la calle, pues aunque no padecen de ceguera y sordera permanente, pueden sufrir desórdenes temporales o problemas de atención, que limitan su percepción de los estímulos externos. Según datos del Programa de la Sociedad de la Información y el Conocimiento (PRO-SIC) del 2011 y del Instituto Nacional de Estadísticas y Censos (INEC), la población con discapacidades, junto con la población adulta mayor y la población de niños en edad escolar, representa el 32% del total de habitantes del país [4]. Este es un porcentaje alto que tiende a aumentar debido al envejecimiento de la generación del *baby boom*. Este fenómeno de envejecimiento se padece igualmente a nivel global, lo que ha impulsado en otros países el diseño de ciudades accesibles que faciliten la movilidad de sus habitantes. En EE.UU e Italia, por ejemplo, se han usado teléfonos inteligentes para emitir alertas vibratorias en presencia de cruces peatonales, y en Chile se han instalado *semáforos de suelo* cuyas luces alertan a los peatones que no levantan la vista al usar su celular [5, 6, 7]. En Costa Rica, también se han realizado esfuerzos, entre ellos el desarrollo de la aplicación móvil DesplazaTEC diseñada para indicar a los usuarios, mediante comandos de voz, la ubicación de distintos puntos de interés en el campus universitario del TEC [8]. Además, se han abierto conferencias internacionales sobre ciudades accesibles, como la Smart City Expo World Congress 2018 celebrada en Barcelona, la Urban Future Conference 2018 de Vienna y la Crowdsourcing the City 2018 llevada a cabo en el Reino Unido. En estas conferencias el tema de los teléfonos inteligentes es recurrente, en particular, las aplicaciones para la detección de APS en tiempo real, un tipo de problema hasta ahora reservado para sistemas operativos de tiempo real.<sup>1</sup> Existen varias razones que lo explican, entre ellas, recientemente se ha incorporado a estos dispositivos una cantidad abundante de sensores (acelerómetros, giroscopios, cámaras fotográficas, micrófonos y sensores de movimiento), una capacidad de procesamiento paralela, la disponibilidad de aceleradores de *hardware*, un nivel de consumo de potencia bajo (gracias a la arquitectura ARM), una amplia disponibilidad de bibliotecas de procesamiento digital de señales (DSP) y un precio accesible para el público. Solo en Costa Rica, la cantidad de líneas celulares promedio por persona en el 2013 fue de 1.5 [9]. La detección de APS usando dispositivos móviles podría beneficiar no solo a las personas con discapacidades auditivas y visuales, sino también a los conductores de vehículos para reconocer sonidos de tránsito como las sirenas de las ambulancias, las alarmas de patrullas policiales o las bocinas del tren. Los conocimientos derivados de la detección de APS también podrían aplicarse en otras áreas, como por ejemplo, la recuperación de información musical, los sistemas de vigilancia automática y la detección de los cantos animales en grabaciones hechas en la naturaleza. Estas últimas son difíciles de procesar, pues requieren de un experto humano que escuche los audios grabados durante semanas o meses y anote los eventos de interés contenidos en ellos. Aunque la cantidad de aplicaciones es mayor a las mencionadas, y se quisiera abarcar todas, en este trabajo se estudia el reconocimiento de las APS por los sonidos que emiten. Seguidamente, la sección 1.1 resume los trabajos previos que se han realizado para la

---

<sup>1</sup>Ni Android ni iOS son sistemas operativos de tiempo real porque no poseen calendarizadores que permitan asignar la prioridad máxima de ejecución a una aplicación que no sea del sistema.

detección de los cruces peatonales, la sección 1.2 explica los problemas encontrados para uno de estos algoritmos, llamado RASP, y la sección 1.3 presenta los objetivos planteados para resolver las deficiencias de RASP, junto con la estructura del resto del documento.

## 1.1 Trabajos previos

El reconocimiento de APS ha sido estudiado previamente por Ivanchenko *et al.* y Ahmetovic *et al.* usando un enfoque de visión por computadora [5, 6]. Los esfuerzos en esta área se han concentrado en identificar los cruces peatonales dibujados en la calle, como lo explican las secciones 1.1.1 y 1.1.2. Sin embargo, en toda la literatura revisada, solamente Ruíz *et al.* han realizado un esfuerzo para reconocer los APS por los patrones acústicos que emiten [10]. Su método, basado en estudios para reconocer cantos de ranas, ballenas y aves, se explica en la sección 1.1.6. Las secciones 1.1.3, 1.1.4 y 1.1.5 resumen algunos estudios en el campo de la biología necesarios para entender la solución propuesta usando sonido.

### 1.1.1 Crosswatch: detección de cruces peatonales usando celulares y visión por computadora

Ivanchenko *et al.* diseñaron un algoritmo para reconocer cruces peatonales en fotografías de  $320 \times 240$  píxeles, tomadas en un celular Nokia. El algoritmo fue implementado en tiempo real usando Symbian C++. El enfoque propuesto se dividió en dos etapas: la primera se encargó de calcular los descriptores de vídeo y la segunda de incorporar esos descriptores de vídeo a un *algoritmo de propagación de creencias* (BP, por sus siglas en inglés) para la toma de decisiones. Los autores explican que la detección de los cruces peatonales se realizó identificando los segmentos correspondientes a los bordes superiores e inferiores de cada rectángulo mediante la conversión a escala de grises, la aplicación de un núcleo de difuminación de la forma  $\langle 1, 2, 1 \rangle$ , la aplicación de un núcleo de derivación de la forma  $\langle -1, 0, 1 \rangle$ , la conversión al valor absoluto de la imagen resultante, la *supresión de no máximos*, la umbralización del valor de los píxeles, operaciones morfológicas de erosión y dilatación, y el descarte de los segmentos menores a 20 píxeles de largo. La lista de descriptores del vídeo es la siguiente:

**Largo de los segmentos:** qué tan largos son los bordes de los rectángulos más cercanos del cruce peatonal en comparación a los del fondo de la imagen.

**Magnitud del gradiente de los bordes:** la magnitud del gradiente, es decir, el cambio del color o intensidad en una imagen que tiende a ser más fuerte en los rectángulos del cruce peatonal que en el fondo.

**Paralelismo entre los segmentos vecinos:** el nivel de paralelismo entre los rectángulos del cruce peatonal.

**Traslape horizontal:** el número de columnas de píxeles compartidas por cada par de rectángulos del cruce peatonal.

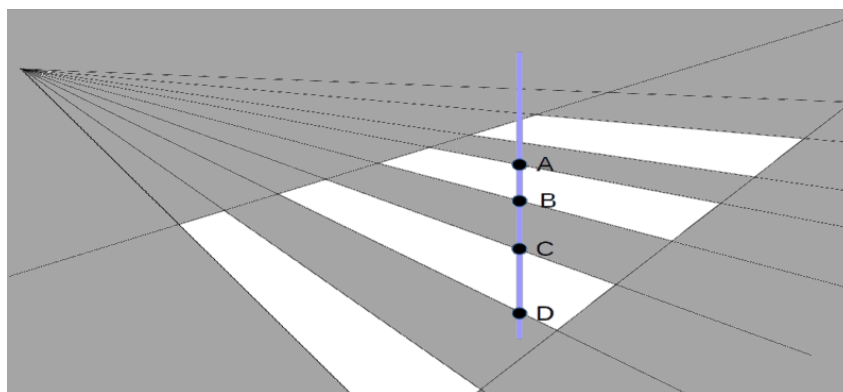
**Consistencia del color:** uniformidad en el color de los píxeles de los rectángulos del cruce peatonal, que, a diferencia del valor absoluto del color, no varía dependiendo de la cantidad de luz que incide en los objetos.

**Uniformidad del ancho de las rayas:** ancho de los rectángulos del cruce peatonal, que tiende a ser más pequeño en la parte superior de las imágenes y más grande en la parte inferior, como consecuencia de la profundidad.

**Tasa de profundidad:** similar al descriptor anterior, con la diferencia de que mide la tasa de reducción de la “altura” de los rectángulos mediante una línea vertical que interseca cuatro segmentos del cruce, es decir:

$$\frac{AC \cdot BD}{BC \cdot AD} = \frac{3}{4},$$

donde los vértices A, B, C y D se definen en la figura 1.2.



**Figura 1.2:** Cálculo del descriptor de vídeo de tasa de profundidad.

Los autores no explican la forma de calcular cada descriptor, pero dan un ejemplo con el descriptor de paralelismo. Para este descriptor fue necesario calcular la magnitud de la diferencia de la pendiente entre cada par de segmentos de línea  $i, j$ . La pendiente de los segmentos es obtenida respecto la inclinación con el eje horizontal de la imagen, y las diferencias entre pendientes de segmentos se agrupan en una secuencia  $C = \{C_{i,j}\}$  con el fin de encontrar la secuencia de estados frente-fondo  $X = x_1, x_2, \dots, x_n$  que maximice la probabilidad *a posteriori* del teorema de Bayes. Esta probabilidad se encuentra asumiendo que la inclinación de los segmentos es un proceso *independiente e idénticamente distribuido* (*i.i.d*) como sigue:<sup>2</sup>

$$P(X|C) = \frac{P(C|X)P(X)}{P(C)} \propto P(X) \prod_{ij} P(C_{ij}|x_i, x_j).$$

<sup>2</sup>El concepto de *i.i.d* se explica con detalle en la sección 2.4.4.

De anotaciones manuales se determinó que el valor inicial recomendado de estado frente-fondo fue  $P_i(x_i = 0) = 0.65$ , donde  $x_i = 0$  significa que el segmento corresponde al fondo. Al analizar un conjunto de 90 imágenes en las que 30 eran cruces peatonales y el resto ruido, se determinó que el prototipo tiene una sensibilidad del 72% y una especificidad del 95%. Al evaluar el uso de la implementación de Symbian C++ para Nokia en un ambiente real, se determinó que el sistema es siempre preciso y robusto contra variaciones de brillo y cambios de perspectiva. El ambiente real consistió en un recorrido por la ciudad a través de 15 intersecciones, con un 50% de probabilidad de encontrarse con un cruce peatonal. En esta prueba participaron dos usuarios no videntes.

### 1.1.2 ZebraLocalizer: identificación y localización de cruces peatonales

Ahmetovic *et al.* propusieron una mejora del método de Ivanchenko *et al.* empleando un acelerómetro para encontrar el horizonte de la imagen, lo que permitió detectar cruces peatonales aún cuando el teléfono estuviera inclinado [5]. En este proyecto se desarrollaron dos soluciones: ZebraRecognizer, una librería de *software* para calcular la posición relativa del usuario con respecto al cruce peatonal, y ZebraLocalizer, una aplicación de iOS que interactúa con el usuario para ubicarlo frente al cruce peatonal. Las órdenes de alineamiento se dieron por voz y vibraciones. Las de voz fueron de tres tipos: rotación, desplazamiento y acercamiento. El algoritmo ZebraRecognizer funciona como sigue: se aplica una media móvil para atenuar el ruido del sensor, se localiza el horizonte de la imagen utilizando los datos del acelerómetro, se reconocen los segmentos de línea en la imagen usando el método *line segment detector* (LSD) [11], se descartan los segmentos que no son paralelos al horizonte y cuyo largo no es suficiente, se ordenan los segmentos de mayor a menor distancia respecto la posición del usuario, y se agrupan los segmentos para identificar los rectángulos del cruce peatonal. Con base en el reconocimiento de los rectángulos se aplican descriptores de vídeo para determinar, que en efecto, se esté observando un cruce peatonal. La toma de decisiones no usó el algoritmo BP, sino otra estructura que no fue explicada. Los descriptores del vídeo empleados fueron: la consistencia del color, la uniformidad del ancho de las rayas y la tasa de profundidad (explicados en la sección 1.1.1). Los resultados de las pruebas en computadora mostraron que el ZebraRecognizer logró una precisión del 100% y una sensibilidad del 77% en 400 imágenes de prueba de  $192 \times 144$  píxeles. Estas imágenes contenían cruces peatonales y otros elementos que podían confundirse con ellas, como señalización del tranvía y escaleras. Las pruebas del ZebraLocalizer en el iPhone 4 permitieron determinar que el uso del acelerómetro ahorró un 25% del tiempo de procesamiento y que la tasa de procesamiento alcanzada fue de diez imágenes por segundo. La evaluación realizada con usuarios invidentes permitió determinar que la velocidad lineal de desplazamiento promedio de estos fue de 1.3 m/s. Entre los puntos a mejorar del ZebraLocalizer, identificadas por sus autores, están:

- Sustituir los mensajes de voz por tonos audibles, pues estos últimos se confunden

menos con el ruido ambiental.

- Procurar que los mensajes sean de corta duración para evitar distraer al usuario de otros sonidos y pistas táctiles (como bocinas de los automóviles, ruidos de motores y los relieves del cruce peatonal) que son usados para ubicarse en el ambiente.
- Cuidar el orden de las instrucciones para evitar que el peatón camine fuera de la acera, lo que puede ocurrir al transmitir las ordenes de rotación y desplazamiento laterales en el orden incorrecto.
- Reducir el número de rotaciones seguidas de desplazamientos para no demandar un grado de atención alto por parte del usuario.
- Aconsejar al usuario que porte el dispositivo en la mano contraria al muro de la cuadra, para facilitar al teléfono capturar las imágenes de la calle y evitar rotaciones innecesarias.
- Verificar que el escenario se mantenga estable al menos tres cuadros antes de emitir una instrucción, para evitar que la oclusión producida por el paso de los carros sobre el cruce peatonal confunda al algoritmo.

Como trabajo futuro para el ZebraRecognizer, Ahmetovic *et al.* comentan que se podría usar un método de filtrado de ruido basado en el filtro Kalman y que podrían combinarse los datos del acelerómetro con los datos del giroscopio para mejorar la estimación del horizonte. Una observación respecto de este artículo es que los autores parecen haber desconocido que el filtro Kalman puede realizar ambas tareas, es decir, tanto el filtrado del ruido como la fusión sensorial; lo que sería necesario es modelar el comportamiento del sistema en una matriz y el aporte de cada sensor en un vector [12].

### 1.1.3 Monitoreo de comunidades de ranas: una aplicación del aprendizaje automático

En 1996, Taylor *et al.* crearon un método para reconocer vocalizaciones de 22 especies distintas de ranas. El algoritmo propuesto empleó aprendizaje automático para interpretar un vector de 40 dimensiones, compuesto por puntos escogidos estratégicamente en el espectrograma. La plantilla diseñada para ubicar los puntos fue reutilizada para todas las especies de ranas, mediante un escalamiento temporal [13]. Como era de esperar de un enfoque de aprendizaje automático, el algoritmo requirió de una etapa extensa de recopilación de vectores de aprendizaje y del ajuste manual de los hiperparámetros de la red neuronal. Entre estos hiperparámetros estaban la tasa de aprendizaje, el número de iteraciones, el tamaño del lote y la función del error [14, 15]. Los resultados mostraron que solo el sonido de un tipo de rana fue reconocido con una precisión del 90%; el resto obtuvo resultados deficientes. Se identificó la causa del problema en la incapacidad del algoritmo para usar una sola plantilla en todos los cantos de rana. Un detalle interesante

es que el audio procesado por unidad de tiempo fue del 25%, es decir, que por cada minuto de audio grabado solo se podían procesar 15 segundos.

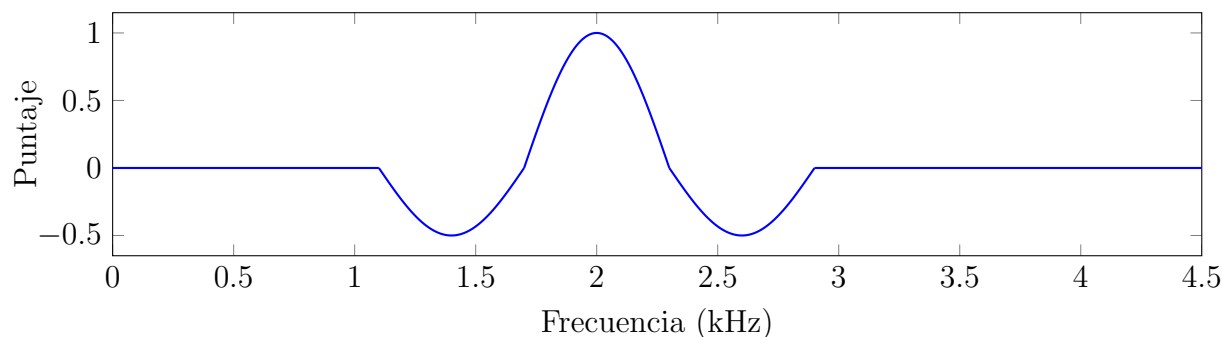
### 1.1.4 Detección automática de vocalizaciones de la rana *Diasporus hylaeformis* en grabaciones de audio

Camacho *et al.* desarrollaron un método para reconocer los cantos de la rana *Diasporus hylaeformis*, encontrada en Costa Rica. El enfoque propuesto fue usar descriptores de audio como sonoridad, timbre y altura musical. La sonoridad se empleó para seleccionar segmentos de audio que contenían comienzos de energía, lo que permitió disminuir la carga computacional descartando aquellos que no contenían eventos acústicos. El timbre y la altura musical se usaron para determinar si el audio procedía o no de una rana. Se utilizó un algoritmo de agrupamiento para clasificar los comienzos (*onsets*, en inglés) de las grabaciones en dos tipos: de primer y segundo plano. Para identificar la altura musical se empleó el algoritmo SWIPE, ajustado para trabajar en el rango de 2.5 kHz a 3.5 kHz. El núcleo unarmónico de reconocimiento de este algoritmo fue construido usando un periodo de coseno alzado como el que se muestra en la figura 1.3, y que se define como sigue [16]:

$$K(f_0, f) = \begin{cases} \cos(2\pi f/f_0), & 0 \leq |f/f_0 - 1| < \frac{1}{4} \\ \frac{1}{2} \cos(2\pi f/f_0), & \frac{1}{4} < |f/f_0 - 1| < \frac{3}{4} \\ 0, & \text{en el resto,} \end{cases} \quad (1.1)$$

donde  $f_0$  es la frecuencia fundamental del núcleo de reconocimiento y  $f$  es la frecuencia analizada (en hercios). Como puede observarse en la figura, el lóbulo principal del núcleo está ubicado en el rango  $3/4 < f/f_0 < 5/4$ , mientras que los lóbulos secundarios negativos están ubicados en los rangos:  $1/4 < f/f_0 < 3/4$  y  $5/4 < f/f_0 < 7/4$ , esto se hace para que al multiplicar el núcleo con la magnitud espectral de sonidos no armónicos se obtenga una altura musical nula. Cabe señalar que estos núcleos son herramientas para dar un puntaje a la distribución de energía de las magnitudes espectrales y no son magnitudes espectrales por sí mismas. Los puntajes positivos son aplicados a la frecuencia fundamental y sus armónicas, y los negativos al resto de frecuencias; además, la forma suavizada de la función coseno permite admitir desafinaciones en los tonos.

Los comienzos de audio identificados como cantos de ranas fueron vectorizados como tuplas de dos entradas: la similitud del timbre obtenido con el timbre de una rana y la altura musical, esto permitió obtener un conjunto de puntos que Camacho *et al.* usaron para estimar el número de individuos presentes en las grabaciones. Para ello agruparon los puntos en dos clases (rana focal y no focal), usando un algoritmo de  $k$ -medias, y posteriormente aplicaron un *análisis multivariable de prueba de varianza* (ANOVA, por sus siglas en inglés) para determinar con una probabilidad del 99% si ambas clases eran la misma. En caso positivo, el algoritmo finaliza, y, en caso negativo, se repite el procedimiento de agrupación por  $k$ -medias y ANOVA en el grupo no focal para determinar si hay



**Figura 1.3:** Núcleo musical para identificar la altura musical de un tono de 2 kHz.

más ranas presentes. El porcentaje de precisión alcanzado fue del 99%, el de sensibilidad fue del 92%, y la medida F fue del 95%. Estas métricas se lograron mantener incluso en presencia de ruido proveniente del viento, la lluvia, la manipulación del micrófono y voces humanas.

### 1.1.5 Método para detectar silbidos, gemidos y otros contornos musicales

Mellinger *et al.* publicaron un algoritmo para detectar contornos de frecuencia melódicos no identificados previamente, es decir, sin que hubiera una etapa de entrenamiento previo [17]. Este algoritmo fue utilizado para reconocer automáticamente nuevos tipos de cantos animales en grabaciones marinas, lo que permitió ahorrar horas de trabajo a los expertos humanos identificando los instantes de interés. El algoritmo partió el audio en ventanas de tiempo traslapadas, para las cuales calculó el *contorno musical*, es decir, la señal conteniendo las frecuencias fundamentales reconocidas en cada instante de tiempo, y luego buscó correspondencias de los contornos musicales entre ventanas vecinas. Los contornos musicales fueron determinados identificando el pico de frecuencia con amplitud más alta en cada subventana de la *transformada de Fourier de corto plazo* (STFT, por sus siglas en inglés). La comparación entre ventanas vecinas se logró utilizando una distancia o métrica no revelada, pero que podría haber sido la distancia euclidiana. Los autores explicaron que para atenuar los efectos del ruido en las grabaciones se utilizó una interpolación polinómica en las magnitudes espectrales. El método alcanzó una *tasa de detección de negativos* del 3% al estudiar el canto de un tipo de ballena de Hawái llamada *Balaenoptera acutorostrata*. Esta tasa se interpreta como la probabilidad de cometer al menos un error de detección en todo el cuerpo de grabaciones analizado.

### 1.1.6 Algoritmo RASP

En 2017 Ruiz *et al.* propusieron un algoritmo conocido como RASP (Reconocimiento Acústico de Semáforos Peatonales), para reconocer tres tipos de sonidos presentes en los semáforos peatonales en Costa Rica: cucú, chirrido alto y chirrido bajo [10]. El algoritmo



**Tabla 1.1:** Bancos de filtros RASP para un cucú, un chirrido alto y un chirrido bajo.

Banco	Rango (kHz)	N. <sup>o</sup> núcleos	Tipo de núcleo	$df$ (Hz)
Cucú	[0.90, 1.10]	2	Armónico	200
Ch. Alto	[2.00, 3.00]	11	Uniarmónico	100
Ch. Bajo	[0.95, 1.75]	9	Armónico impar	100

se basa en las técnicas acústicas resumidas en las secciones 1.1.3, 1.1.4 y 1.1.5, empleadas para reconocer el canto de especies animales. Al igual que la sección 1.1.4, el método RASP utiliza núcleos de reconocimiento musical basados en SWIPE, cuyas frecuencias características, decaimiento frecuencial y número de armónicas fue ajustado por tipo de sonido. Existen tres bancos de núcleos, uno por cada tipo de sonido. La cantidad de núcleos en cada banco se establece de acuerdo a la resolución frecuencial elegida, por ejemplo, para el sonido cucú se utiliza una resolución de  $df = 200$  Hz en el rango [0.9, 1.1] kHz, lo que produce dos núcleos. Un resumen de la configuración de los bancos de núcleos empleados por RASP se muestra en la tabla 1.1. En ella se ve que  $df = 100$  Hz para ambos chirridos, en lugar de 50 Hz como se hizo en el estudio de Ruiz *et. al* (de experimentos previos se sabe que el rendimiento usando ambas resoluciones es similar, pero menos costoso con 100 Hz).

El contorno musical es generado por cada banco de núcleos eligiendo la frecuencia fundamental del núcleo de reconocimiento con el puntaje más alto al multiplicarlo con la magnitud espectral. Este núcleo se usa para encontrar correspondencias con las modulaciones de frecuencia de los APS. La distancia empleada para determinar las correspondencias fue la euclidiana, con una modificación para brindar valores en el rango unitario. La señal de alerta emitida, es decir, los valores de distancia en cada instante de tiempo respecto el patrón APS buscado, fue “limpiada” mediante un control del número de eventos que explota la periodicidad de estos sonidos. Las tasas de detección obtenidas por este método al analizar 79 grabaciones recopiladas en San José, fueron de 87% de precisión, 83% de especificidad, 86% de sensibilidad y 85% de medida F. Las buenas métricas obtenidas motivaron el desarrollo de una implementación en dispositivos móviles, que fue diseñada para satisfacer requerimientos de tiempo real suave. La implementación también admitió el procesamiento de un cuarto APS llamado cucú bajo, de periodo 1200 ms y con dos tonos de 1060 Hz y 850 Hz [18].<sup>3</sup> El rendimiento de la aplicación móvil en escenarios reales logró emitir una alerta de cruce temprana (en menos de cinco picos de actividad) para el APS cucú, pero tardía para los chirridos alto y bajo. Esta métrica de cinco picos de actividad fue fijada para asegurarse que un adulto mayor con velocidad de desplazamiento lineal (alta) de 1.4 m/s pudiera cruzar la calle mientras el APS estuviera sonando. Para calcular la métrica de cinco picos se estudiaron las grabaciones recopiladas en la GAM y los semáforos que las emitieron, y con esta información se determinó que la longitud reglamentaria de los cruces peatonales es de 6.6 m y 16.5 m, para dos y cinco

<sup>3</sup>La versión beta de la aplicación se puede encontrar en el siguiente enlace de GooglePlay: <https://play.google.com/store/apps/details?id=ucr.citic.rasp>.

**Tabla 1.2:** Velocidad de desplazamiento lineal promedio de un adulto mayor.

Tipo	Velocidad (m/s)		
	Monge	Brown	Media
Baja	0.7	1.0	0.8
Media	0.9	1.4	1.1
Alta	1.1	1.7	1.4

carriles, respectivamente, y que la duración promedio de las secuencias principales APS es de  $11.4 \pm 1.4$  s y  $27.3 \pm 4.2$  s, para dos y cinco carriles, respectivamente. La velocidad promedio de desplazamiento se calculó con base en la tabla 1.2, que muestra la velocidad promedio de desplazamiento lineal reportada por dos autores distintos y que se aproxima a la calculada en la sección 1.1.2 [19, 20]. La evaluación final del algoritmo RASP determinó que la detección del sonido cucú es robusta y la del chirrido alto aceptable, pero que la del chirrido bajo es deficiente porque se confunde con el sonido de otros APS y el ruido proveniente de fuentes armónicas en la calle, como voces humanas, cantos de aves y bocinas de carros [21].

## 1.2 Definición del problema a resolver

Como se explica a continuación, la causa de los problemas del algoritmo RASP podría radicar en que el núcleo musical para el chirrido alto no fue capaz de diferenciar la frecuencia fundamental de sus armónicas y que no se estaban usando umbrales de detección de tono dinámicamente adaptables al nivel de ruido de la calle. Además, el sistema es más complicado de lo necesario, tratando al sonido cucú como un caso especial, en vez de extender el enfoque de plantillas para admitir contornos discontinuos. A continuación, la sección 1.2.1 explica el problema de reconocimiento del núcleo del chirrido alto, la sección 1.2.2 explica cómo funcionan los umbrales de tono fijos, la sección 1.2.3 explica la razón por la cual el algoritmo original no pudo manejar segmentos nulos en las plantillas musicales de las modulaciones de frecuencia y la sección 1.2.4 aborda la necesidad de usar umbrales de tono adaptables.

### 1.2.1 Reconocimiento del tono musical

El algoritmo RASP reconoce los APS mediante núcleos de reconocimiento musical de tres tipos: armónico, uniarmónico y armónico impar [10]. El núcleo armónico posee todas las armónicas y es usado por el sonido cucú; el núcleo uniarmónico posee solo una armónica y es usado por el chirrido alto; y el núcleo armónico impar posee solo las armónicas impares y es usado por el chirrido bajo.

Las figuras 1.7, 1.8 y 1.9 muestran ejemplos de estos núcleos. La forma de cada uno

fue construida con base en la distribución de energía de los espectrogramas, calculados con una ventana Hamming de 512 entradas, un traslape del 20% y una frecuencia de muestro de 11.025 kHz. Al construir los espectrogramas con una resolución mayor en frecuencia, es decir, empleando una ventana Hamming 512 entradas, un traslape del 96% y una frecuencia de muestreo de 22.050 kHz, se obtiene un resultado como el de las figuras 1.4, 1.5 y 1.6. La resolución en frecuencia es mayor porque la ventana de análisis es más larga, capturando frecuencias de longitud de onda mayor. Además, la cantidad de traslape es más extensa, contrarrestando el efecto del *principio de incertidumbre*.<sup>4</sup> Con base en los espectrogramas de mayor resolución se observa que los diseños de núcleo para ambos chirridos fueron subóptimos, a diferencia del núcleo cucú, que fue construido correctamente con 3 o 4 armónicas. Para el chirrido alto se empleó una sola armónica en lugar de las cuatro observables, y en el caso de chirrido bajo se emplearon solo las armónicas impares 1, 3 y 5 en lugar de 1, 2 y 3 [10].

Para comprobar que la cantidad de armónicas usadas en el núcleo pudiera explicar los problemas de reconocimiento, se calcularon matrices de puntajes mediante una prueba unitaria de Android sobre la aplicación. Estas matrices de puntajes muestran en el eje horizontal el tiempo de la grabación y en el eje vertical los puntajes para los núcleos del banco de núcleos. Cada columna de la matriz de puntajes puede interpretarse como un vector de puntajes  $z_{m,i}^{(t)}$  de dimensión  $I \times 1$  ( $I$  es el número de núcleos en el banco) correspondiente a la  $m$ -ésima subventana analizada de la  $t$ -ésima ventana de grabación. Estos vectores de puntajes expresan la similitud de cada núcleo  $K_i$ , de dimensión  $N \times 1$ , con la  $m$ -ésima ventana espectral  $|S^{(t)}(m, f)|$ , de dimensión  $N \times 1$ , producida por la STFT. El vector de puntajes se expresa como sigue:

$$z_{m,i}^{(t)} = K_i^T |S^{(t)}(m, f)|. \quad (1.2)$$

En las matrices de puntajes de las figuras 1.10, 1.11 y 1.12 se observa un tratamiento anómalo en el banco de núcleos del chirrido alto, pues la figura 1.11a muestra que se capturó más energía de la necesaria al analizar el sonido cucú, y la figura 1.11b muestra que se capturó menos energía de la requerida para identificar el patrón de líneas verticales de su propio sonido. Esto indica que las tasas de detección del chirrido alto podrían ser bajas porque el núcleo uniarmónico no distingue la frecuencia fundamental de sus subarmónicas. Esto se comprueba en la figura 1.11a, donde el tercer filtro interpreta erróneamente a 2700 Hz como la frecuencia fundamental, cuando es más bien la tercer armónica del tono cucú en 900 Hz.<sup>5</sup> La incorporación de un núcleo con más armónicas evitaría que 2700 Hz obtenga un puntaje alto, no porque otro núcleo en 900 Hz capture más energía que él (pues 900 Hz está fuera del rango del chirrido alto en [2, 3] kHz) sino porque entre más armónicas tenga el núcleo mayor será la penalización de la energía faltante.

<sup>4</sup>El principio dicta que a mayor resolución frecuencial menor resolución temporal, y viceversa.

<sup>5</sup>Dependiendo del contexto, la frecuencia fundamental es considerada como la primera armónica.

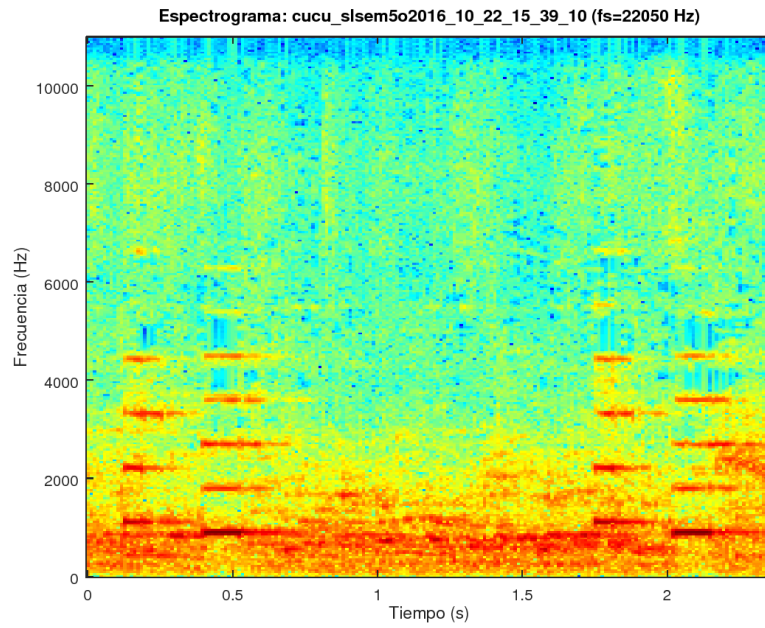


Figura 1.4: Espectrograma del cucú.

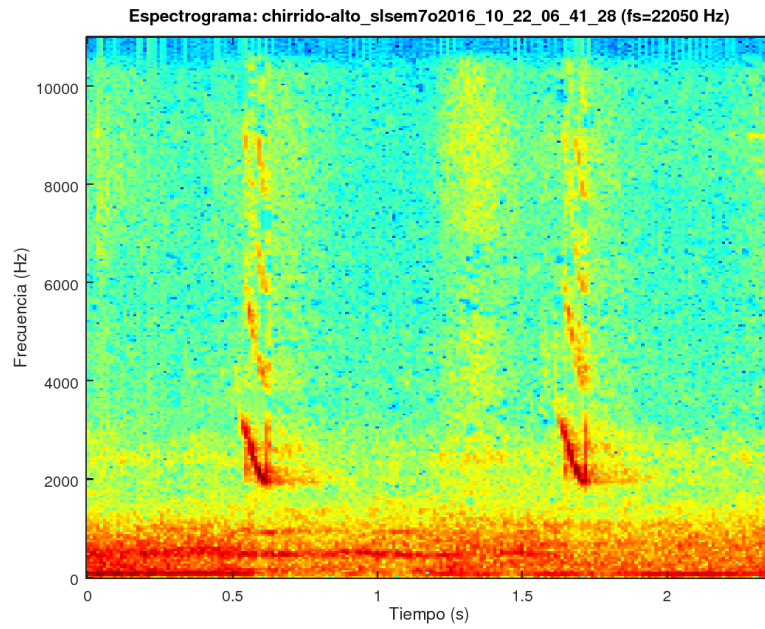
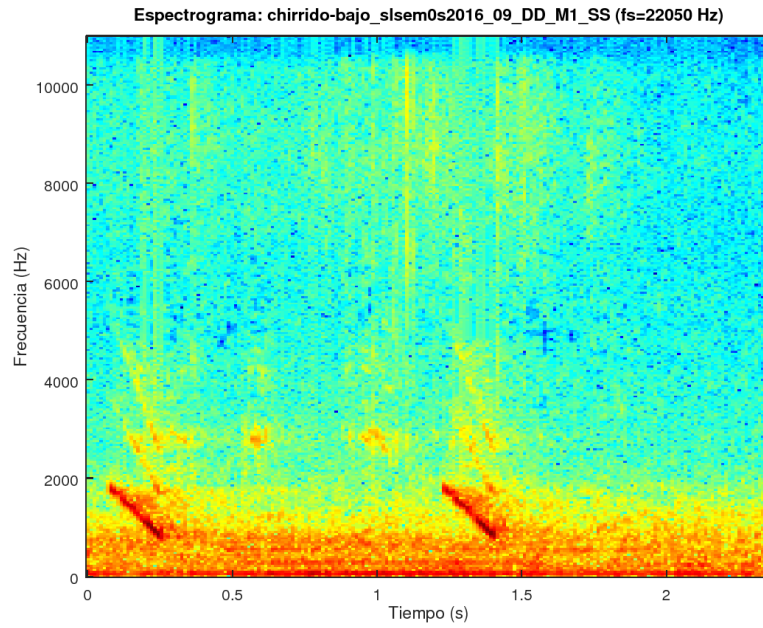
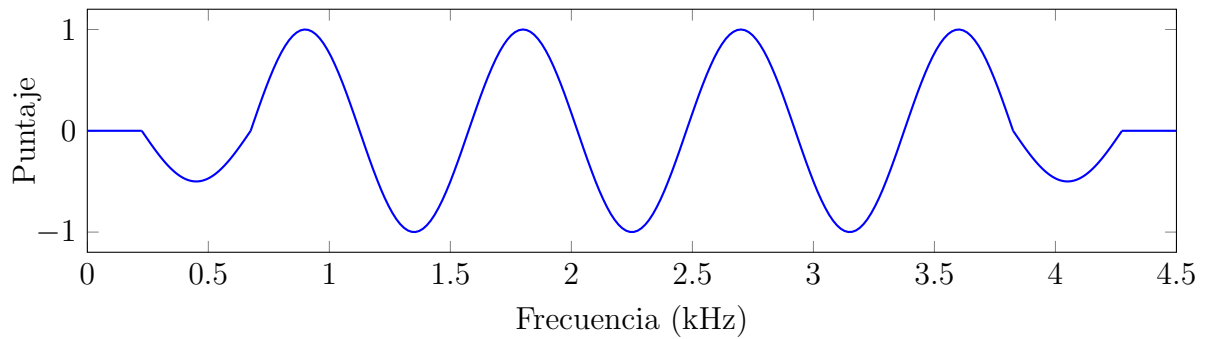


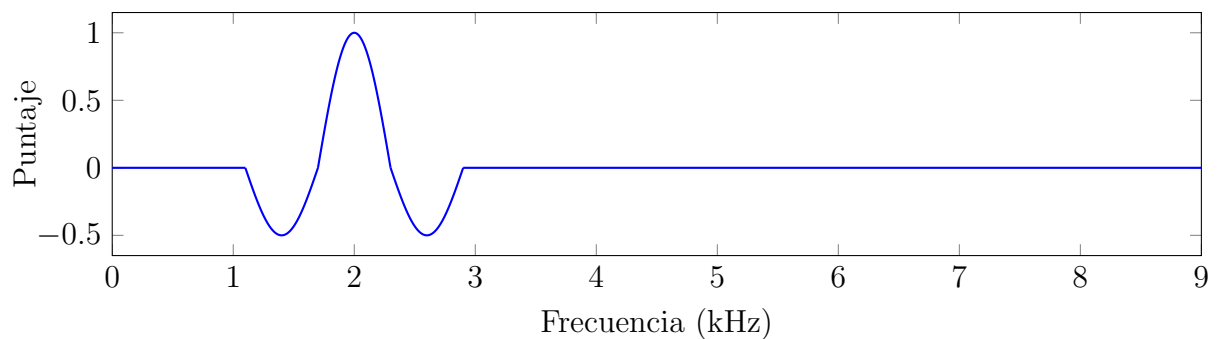
Figura 1.5: Espectrograma del chirrido alto.



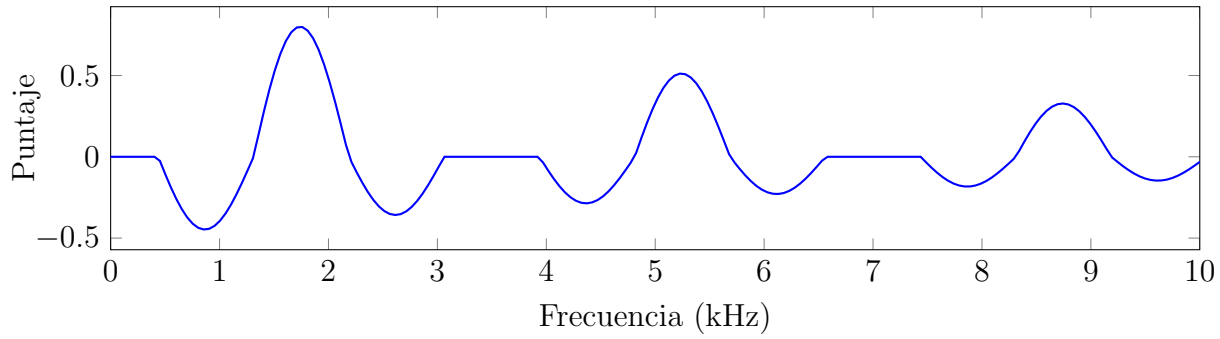
**Figura 1.6:** Espectrograma del chirrido bajo.



**Figura 1.7:** Núcleo armónico de 4 armónicas con  $f_0 = 0.9$  kHz.



**Figura 1.8:** Núcleo uniarmonico con  $f_0 = 2$  kHz.



**Figura 1.9:** Núcleo armónico (de 3 armónicas) impares con  $f_0 = 1.750$  kHz.

## 1.2.2 Umbrales de detección

El vector de puntajes, en la ecuación 1.2, se emplea para reconocer el contorno musical  $x^{(t)}$ , de dimensión  $M \times 1$ , como sigue:

$$x_m^{(t)} = c_m^{(t)} H(p_m^{(t)} - \alpha),$$

donde  $M$  es la cantidad de ventanas en el contorno musical,  $c_m^{(t)}$  es la frecuencia fundamental detectada en la  $m$ -ésima ventana de la  $t$ -ésima grabación (cuyo núcleo obtuvo el puntaje mayor, es decir,  $p_m^{(t)} = \max \{z_{m,i}^{(t)}\}$ ),  $H$  es la función escalón unitario de Heaviside y  $\alpha \in [0, 1]$  es el umbral de detección del tono. Como se comentó en la sección 1.2.1, el contorno musical y las plantillas de los APS se usan para identificar el tipo de sonido. Las segundas se definen como sigue [10]:

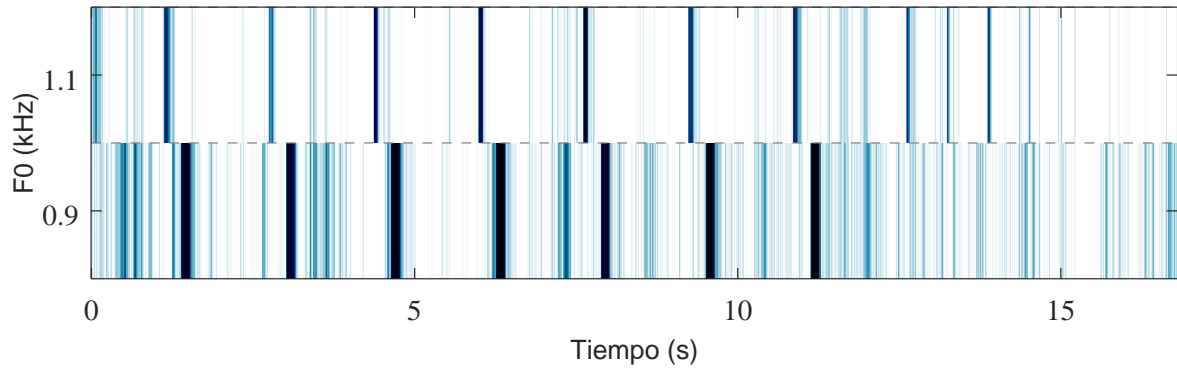
$$O^{(\text{cucú})}(t') = \begin{cases} 1100 \text{ Hz}, & t' \in [0, 70] \text{ ms} \\ 900 \text{ Hz}, & t' \in [270, 400] \text{ ms} \\ 0, & \text{en el resto,} \end{cases} \quad (1.3)$$

$$O^{(\text{calto})}(t') = \begin{cases} -10 \frac{\text{Hz}}{\text{ms}} t' + 3000 \text{ Hz}, & t' \in [0, 100] \text{ ms} \\ 0, & \text{en el resto} \end{cases} \quad (1.4)$$

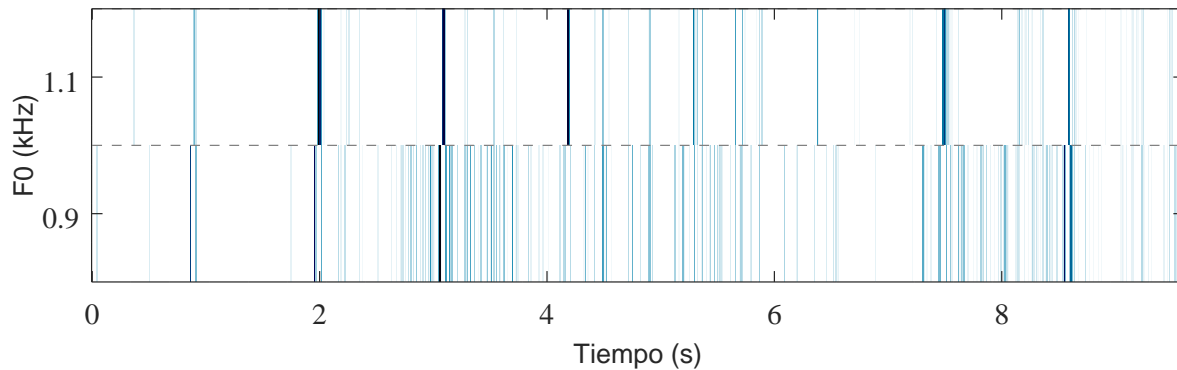
y

$$O^{(\text{cbajo})}(t') = \begin{cases} -5 \frac{\text{Hz}}{\text{ms}} t' + 1750 \text{ Hz}, & t' \in [0, 160] \text{ ms} \\ 0, & \text{en el resto,} \end{cases} \quad (1.5)$$

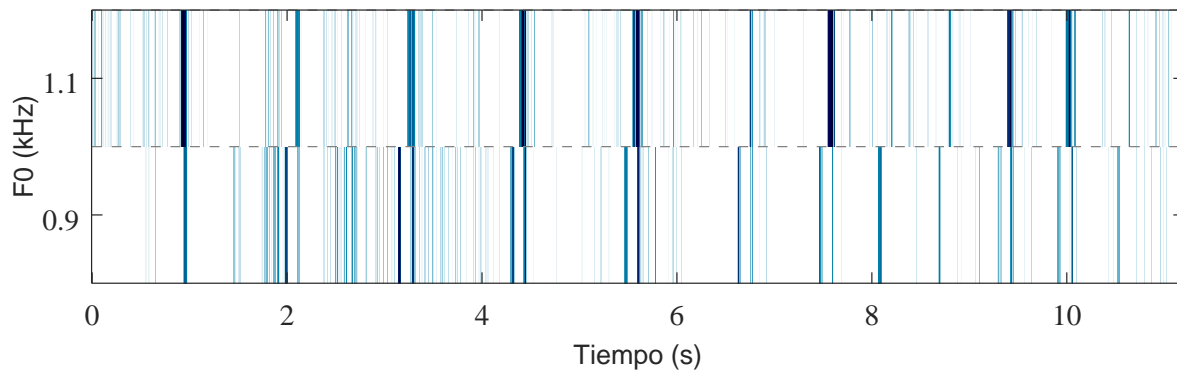
donde  $O^{(\text{cucú})}(t') = O^{(\text{cucú})}(t' + 1630 \text{ ms})$ ,  $O^{(\text{calto})}(t') = O^{(\text{calto})}(t' + 1120 \text{ ms})$  y  $O^{(\text{cbajo})}(t') = O^{(\text{cbajo})}(t' + 1110 \text{ ms})$  corresponden a los sonidos cucú, chirrido alto y chirrido bajo, respectivamente. Un ejemplo de estos contornos puede observarse en la figura 1.15 y, como se observa en las ecuaciones 1.4 y 1.5, el contorno musical de los chirridos se modeló como una recta decreciente en el tiempo en vez de una función exponencial (por simplicidad). Retomando lo mencionado en la sección 1.1.6, las plantillas y el contorno musical se



(a) Grabación cucú analizada con el banco de núcleos cucú, donde se puede observar fuertemente marcado el patrón APS en 1100 Hz y 900 Hz, como se esperaba.

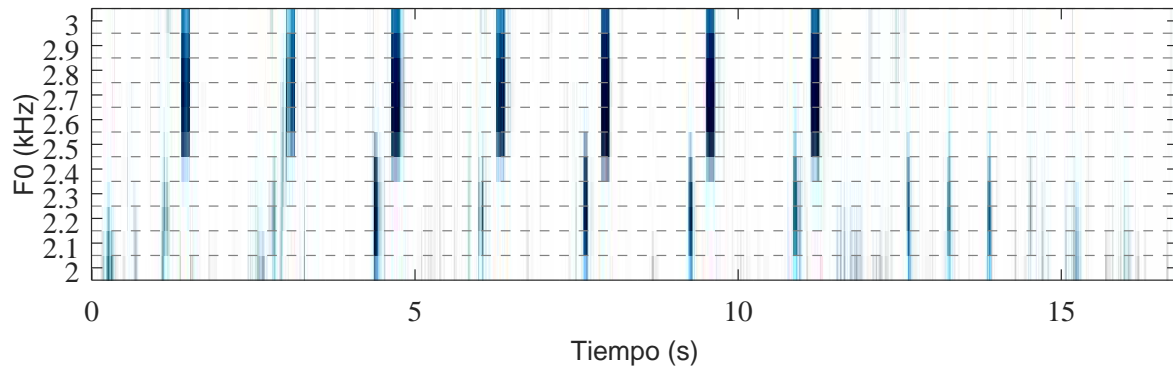


(b) Grabación chirrido alto analizada con el banco de núcleos cucú, donde la energía capturada es baja y no provoca detecciones erróneas.

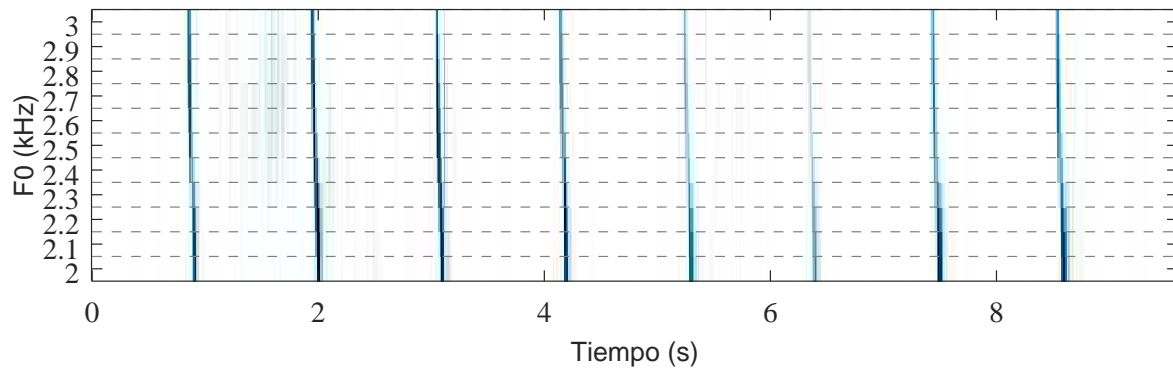


(c) Grabación chirrido bajo analizada con el banco de núcleos cucú, donde la energía capturada es baja excepto para el tono en 1100 Hz. Afortunadamente esto no provoca detecciones erróneas porque no se reconoce el patrón completo.

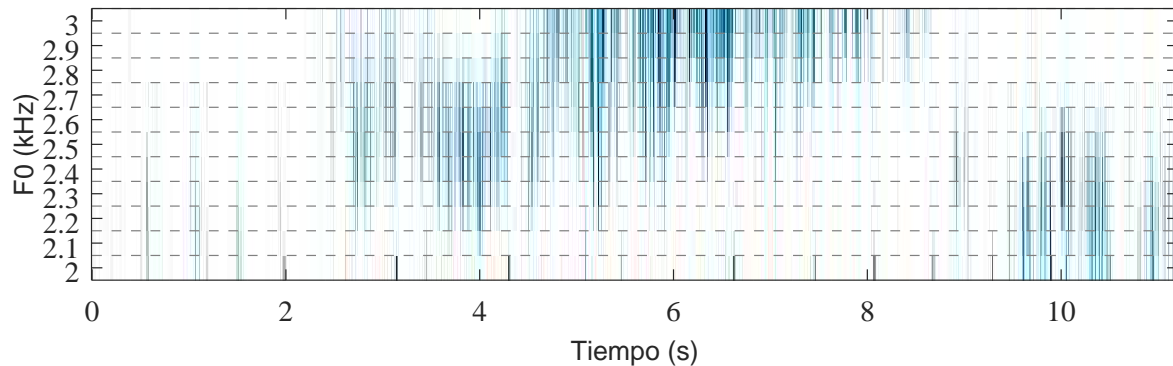
**Figura 1.10:** Puntajes de los dos núcleos armónicos del evaluador cucú (1100–900 Hz, de 3 y 4 armónicas respectivamente) para grabaciones de cucú (1.10a), chirrido alto (1.10b) y chirrido bajo (1.10c), como función del tiempo.



(a) Grabación cucú analizada con el banco de núcleos chirrido alto, donde se reconoce erróneamente un tono de 2700 Hz, que es más bien la tercer armónica de un tono de 900 Hz.



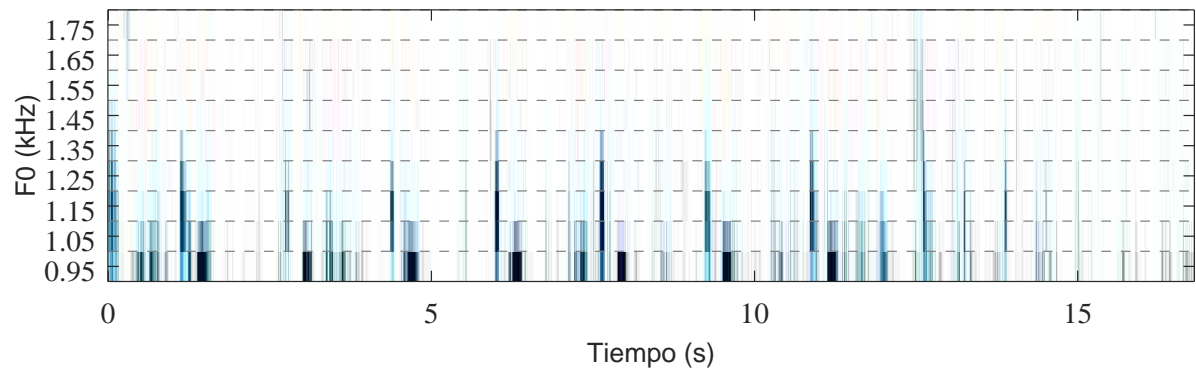
(b) Grabación chirrido alto analizada con el banco de núcleos chirrido alto, donde se captura tenuemente la energía de la modulación APS en 3–2 kHz.



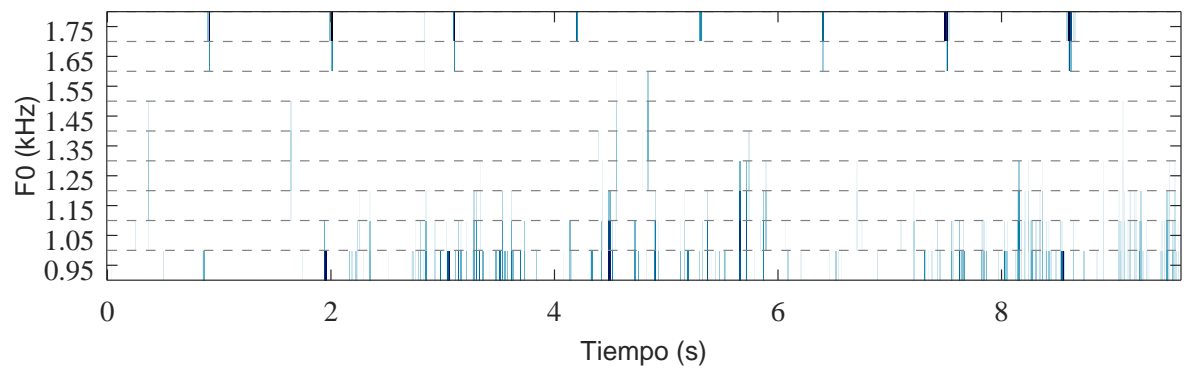
(c) Grabación chirrido bajo analizada con el banco de núcleos chirrido alto, donde la energía capturada es baja y no provoca detecciones erróneas.

**Figura 1.11:** Puntajes de los once núcleos uniarmónicos del evaluador chirrido alto (3–2 kHz) para grabaciones de cucú (1.11a), chirrido alto (1.11b) y chirrido bajo (1.11c), como función del tiempo.

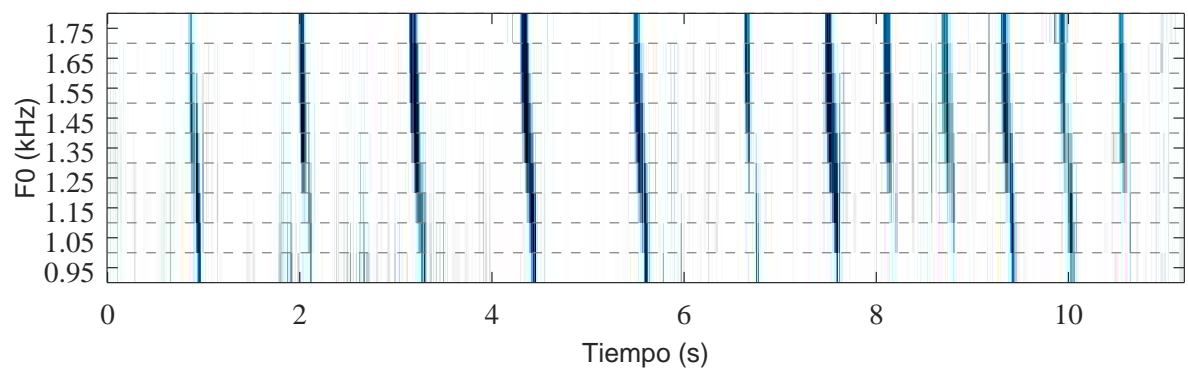




- (a) Grabación cucú analizada con el banco de núcleos chirrido bajo, donde se captura energía para el tono cucú de 1100 Hz y 900 Hz, lo cual es indeseable pero inevitable pues las modulaciones de ambos tipos de APS comparten el mismo rango de frecuencias.



- (b) Grabación chirrido alto analizada con el banco de núcleos chirrido bajo, donde la energía capturada es baja y no provoca detecciones erróneas.



- (c) Grabación chirrido bajo analizada con el banco de núcleos chirrido bajo, donde se captura correctamente la energía de la modulación APS en 1750–950 Hz.

**Figura 1.12:** Puntajes de los nueve núcleos armónico impar (de 5 armónicas) del evaluador chirrido bajo (1750–950 Hz) para grabaciones de cucú (1.12a), chirrido alto (1.12b) y chirrido bajo (1.12c), como función del tiempo.

**Tabla 1.3:** Umbrales fijos de tono del prototipo original de RASP.

APS	$\alpha$	$\beta$	N. <sup>o</sup> eventos
Cucú	0.14	0.45	1
Ch. Alto	0.07	0.30	3
Ch. Bajo	0.07	0.30	2

comparan mediante una versión modificada de la distancia euclidiana —también llamada norma L2 modificada—, definida como sigue (excepto para el sonido cucú):<sup>6</sup>

$$d(x_m^{(t)}, o) = \left( 1 - \frac{\|o - x_m^{(t)}\|_2}{\xi^{\frac{1}{2}}(o_\xi - o_1)} - y_{k,\xi} \right)^+, \quad (1.6)$$

donde  $(\cdot)^+$  es el operador de rectificación de media onda que permite eliminar valores negativos (que no pueden ser distancias),  $o$  es la versión discreta de la plantilla  $O$  conteniendo únicamente las modulaciones de frecuencia,  $\xi = \dim(o)$ ,  $(o_\xi - o_1) = f_{\max} - f_{\min}$  es el largo del rango de frecuencias, y  $y_{k,\xi} \in [0, 1]$  es una penalización cuya severidad depende de la cantidad de valores nulos  $k$  encontrados en el contorno musical. Recordando que  $t$  es el índice de la ventana de grabación analizada, se puede interpretar las similitudes de los contornos con las plantillas a través del tiempo discreto como una señal de “amplitud de alerta”  $d'(t) = d(x^{(t)}, o)$ , y con ella detectar la presencia de un APS en la grabación. Para incrementar la especificidad del método se explota la cualidad repetitiva de los APS construyendo una señal depurada de alerta  $a(t)$  como sigue [10]:

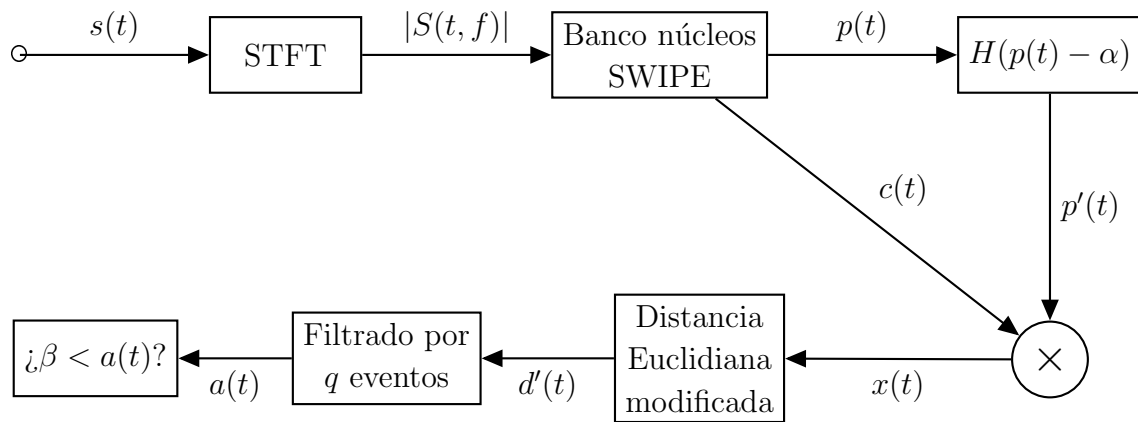
$$a(t) = \prod_{i=0}^q \max_{|\epsilon| \leq 0.025l} \{d'(t + il + \epsilon)\},$$

donde  $q$  es la cantidad de eventos a considerar en el control de periodicidad,  $l = MT$  es la distancia en número de muestras entre los eventos,  $M$  es la tasa de muestreo de la STFT (definida más adelante en la sección 1.2.3),  $T$  es el periodo en segundos del APS y  $\epsilon$  es un valor de tolerancia (no más de un 2.5% del número de muestras entre los eventos) [10]. En RASP se asigna  $q = 1$  al sonido cucú,  $q = 3$  al chirrido alto y  $q = 2$  al chirrido bajo. Posterior a este cálculo (llamada lógica de filtrado), las entradas de la señal de alerta son anuladas cuando no cumplen la condición  $\beta < a(t)$ , donde  $\beta$  y  $\alpha$  son los umbrales fijos de tono y de alerta, respectivamente. Las figuras 1.13 y 1.14 presentan diagramas que resumen gráficamente el algoritmo descrito y en la tabla 1.3 se presentan los valores de los umbrales usados en el prototipo del algoritmo RASP.

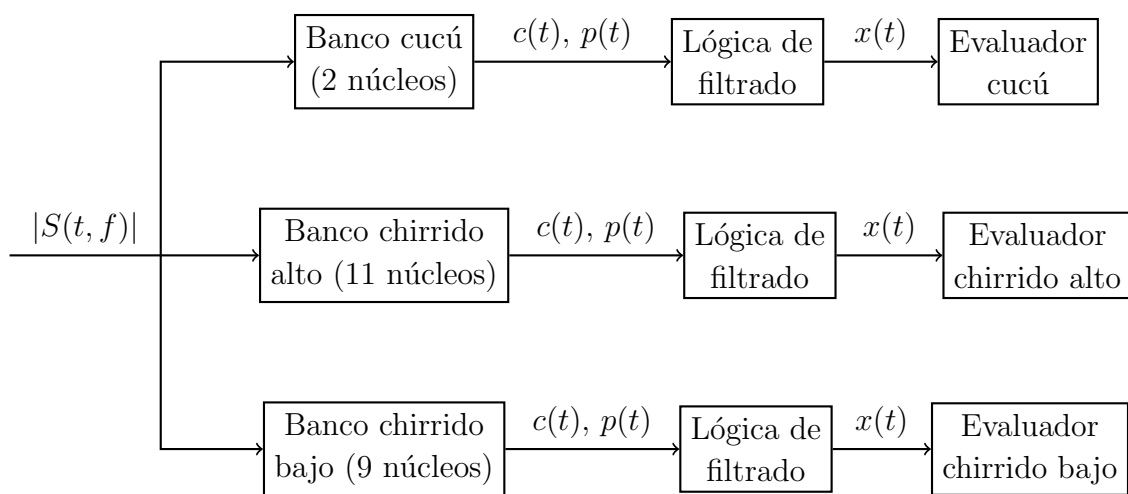
### 1.2.3 Segmentos nulos en las plantillas

El contorno musical y las plantillas utilizadas deben tener igual cantidad de puntos para que la fórmula de la distancia funcione. Para lograr esto se requiere conocer los retardos

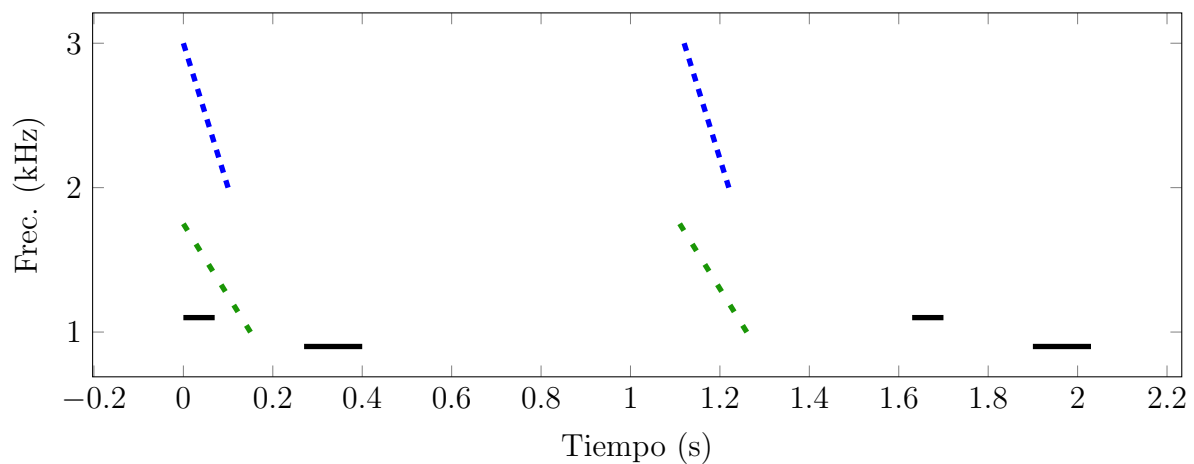
<sup>6</sup>La fórmula de la distancia euclidiana modificada analiza segmentos del contorno que contienen valores nulos y no nulos, solo que a estos primeros se los penaliza, no se los elimina.



**Figura 1.13:** Diagrama del funcionamiento del algoritmo RASP.



**Figura 1.14:** Distribución de los 22 filtros del algoritmo RASP. Tres bancos de filtros distintos generan sus propias señales  $c(t)$  y  $p(t)$ .



**Figura 1.15:** Ejemplo visual de los contornos de las plantillas APS. En negro sólido, el sonido cucú, en azul intermitente denso, el chirrido alto y en verde intermitente, el chirrido bajo.

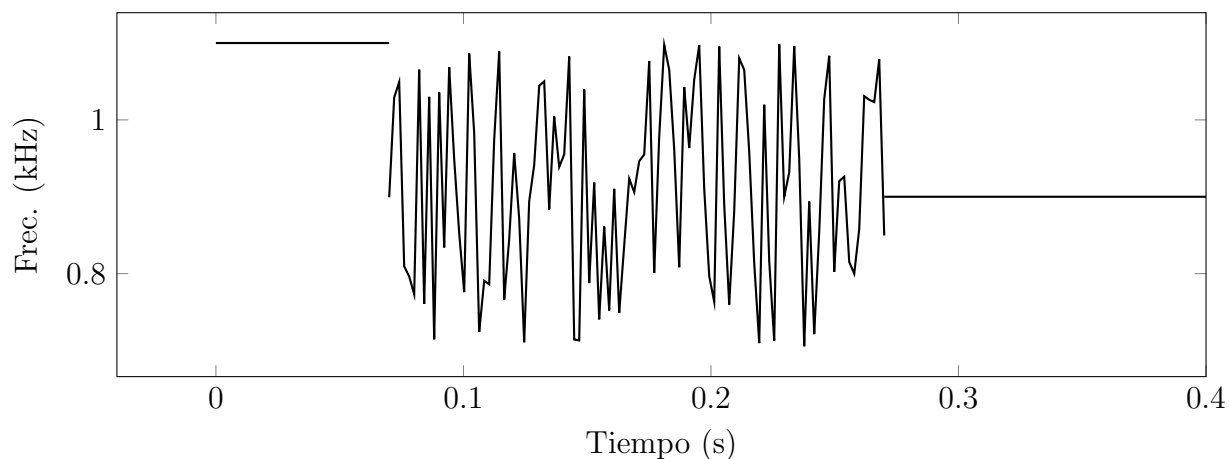
que el sistema sufre al grabar y procesar el sonido, y la tasa de ventanas espectrales generadas. Conociendo estos parámetros, el número de muestras contenidas en las plantillas musicales se puede aproximar como sigue:

$$M \approx \frac{f_s}{N_{\text{FFT}} + \tau f_s}, \quad (1.7)$$

donde  $N_{\text{FFT}} = 2^{\lceil \log_2(T_v f_s) \rceil}$  es la potencia de dos más cercana por la derecha del número de muestras de la ventana de análisis temporal  $T_v$ ,  $f_s$  es la frecuencia de muestreo de la grabación,  $T_v = 8/f_{\min}$ , y  $\tau$  es el retardo del procesamiento producido por el sistema. La ecuación separa las  $f_s$  muestras capturadas por segundo en grupos de  $N_{\text{FFT}}$ , con el fin de ser analizadas individualmente por la *transformada rápida de Fourier* (FFT, por sus siglas en inglés). Se usan ocho periodos de  $T_v$  para evitar el traslape de los lóbulos principales de los senos cardenales correspondientes a las magnitudes espectrales de las ventanas rectangulares, pero en realidad solo son necesarios dos periodos, como se verá en la sección 2.1. Por simplicidad, el número de muestras  $\tau f_s$ , requeridas por el procesamiento del sistema, puede interpretarse como un incremento en el largo de la ventana de análisis. La implementación para dispositivos móviles de RASP utiliza los valores  $f_s = 22050$  Hz,  $f_{\min} = 900$  Hz (la menor frecuencia de todos los APS) y  $\tau = 5$  ms, lo que genera  $M = 60$  FFT por segundo, pero con  $\tau = 0$  se podrían obtener hasta  $M = 86$  FFT/s.<sup>7</sup> Esta tasa de ventanas generadas por segundo determina el tamaño del búfer para cada plantilla musical como  $\lceil MT \rceil$ , donde  $T$  es el periodo de la plantilla. Esto, para el tono cucú de periodo  $T = 1.63$  s, significaría la existencia de un búfer de 140 entradas, correspondientes a 6 entradas del tono de 1100 Hz, 17 entradas de silencio, 11 entradas del tono de 900 Hz y 106 entradas de silencio. En el caso del chirrido alto (de periodo  $T = 1.12$  s), el búfer contendría 97 entradas, correspondientes a 9 entradas de la modulación de frecuencia y 88 entradas de silencio. Por último, para el chirrido bajo (de periodo  $T = 1.11$  s) el búfer contendría 96 entradas, correspondientes a 14 entradas de la modulación de frecuencia y 82 entradas de silencio.

Procesar las entradas de silencio es problemático, pues en ellas el contorno musical contiene ruido en vez de silencio, lo que produce una diferencia no nula que aumenta la distancia respecto de la plantilla musical. Un ejemplo donde sería necesario procesar estos momentos de silencio es en la figura 1.16, donde el momento de ruido ocurre en medio del contorno musical. Por esto, el algoritmo RASP descarta los segmentos nulos de las plantillas modelando únicamente rectas con pendiente decreciente, en el caso de los chirridos, y eliminando los segmentos nulos de los contornos, en el caso del cucú. Esto plantea un problema adicional, pues hay que modelar una plantilla que identifique los segmentos no nulos, y aparte, un arreglo de plantillas para cada subcontorno musical, lo que incrementa la complejidad computacional. Anecdóticamente, Ruiz *et al.* no tuvieron que usar arreglos de subplantillas para procesar el sonido cucú, pues al contener este sonido dos tonos constantes de 900 Hz y 1100 Hz, optaron por contar la proporción de tonos deseados en los instantes no nulos. Sin embargo, este enfoque no es apto para procesar barridos

<sup>7</sup>Las FFTs son de tamaño  $N_{\text{FFT}}/2$ , pues la magnitud espectral de la señal de audio es simétrica.



**Figura 1.16:** Ejemplo de un tono cucú con un segmento de ruido en lugar de silencio.

discontinuos de frecuencia y además, hace necesaria la inclusión de casos especiales, los cuales incrementan la complejidad del algoritmo.

### 1.2.4 Robustez contra el ruido

Un problema que surge al usar los umbrales fijos de la sección 1.2.2, y la simplificación de los contornos de la sección 1.2.3, es que el reconocimiento es propenso al ruido. Esto sucede porque algunas APS, como el chirrido bajo, usan una banda de frecuencia menor a 2000 Hz, contaminada por el ruido de la calle. Además, tienen una modulación de frecuencia de corta duración, que los hace fácilmente enmascarable por otros sonidos. El APS cucú, aunque usa una frecuencia de 900 Hz, no tiene este problema, porque su contorno es más largo y estable que el resto de las APS. Para mitigar la dificultad de reconocer la modulación del chirrido bajo, se podría incrementar el umbral de detección de tono para hacer el reconocimiento más estricto, pero esto tiene el inconveniente de hacer que el algoritmo pierda sensibilidad.

Para ejemplificar la variabilidad del umbral de tono  $\alpha$  se realizó un ejercicio con cinco grabaciones en el que se hizo un barrido de valores y para cada uno se calculó la periodicidad del contorno musical. Como se observa en las figuras 1.17a y 1.17b, el cálculo de la periodicidad se realizó buscando el pico de energía más alto en la magnitud espectral del contorno musical, lo que ha sido empleado en otros problemas similares, como el cálculo del ciclo de las manchas solares a partir de la señal de fluctuación de energía recibida desde la tierra [22]. La tabla 1.4 muestra los periodos obtenidos por cada valor  $\alpha$ ; aquellos periodos más cercanos a 1.11 s se resaltan con negrita. Se comprueba que el periodo deseado se obtiene en cada grabación con un umbral distinto, lo que indica que  $\alpha$  varía dependiendo de la relación señal-ruido (SNR, por sus siglas en inglés), es decir, qué tan fuerte suena el APS con respecto al sonido de fondo. Una SNR negativa implica que la intensidad del ruido es más fuerte que la de la señal, y una positiva lo contrario. Lamentablemente, al estar la señal y el ruido entremezclados, estimar la SNR es difícil,

**Tabla 1.4:** Periodo detectado en distintas grabaciones según el umbral de tono. La periodicidad correcta se resalta con negrita.

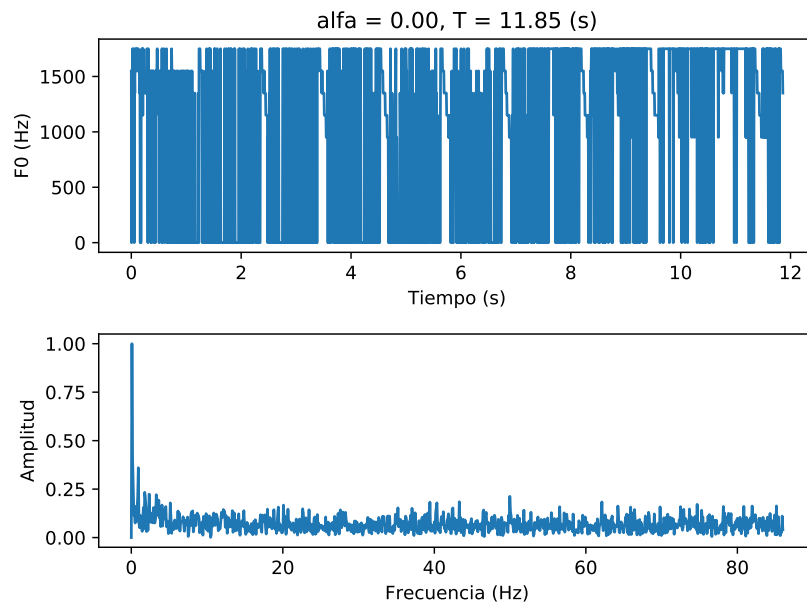
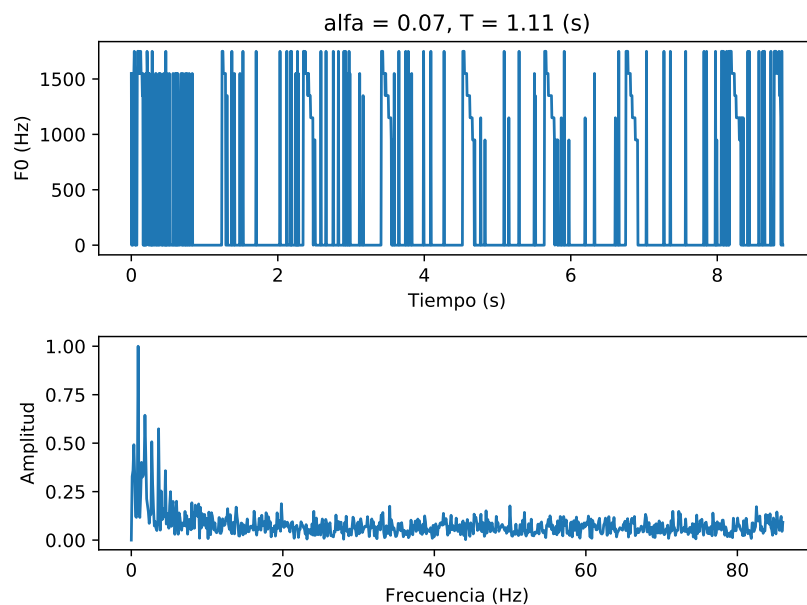
$\alpha$	ID. Grabación					
	SNR alto		SNR bajo			
	A	B	C	D	E	F
0.00	11.85	1.25	1.07	12.99	3.14	1.86
0.07	<b>1.11</b>	<b>1.13</b>	1.15	1.08	0.59	0.56
0.14	<b>1.11</b>	1.05	1.21	0.56	<b>1.06</b>	0.57
0.21	0.55	1.05	<b>1.13</b>	<b>1.09</b>	0.57	<b>1.17</b>
0.29	0.53	1.18	<b>1.13</b>	0.01	0.37	3.51
0.36	2.80	0.59	2.53	0.05	0.00	0.00
0.43	0.00	1.33	0.00	0.00	0.00	0.00
0.50	0.00	0.99	0.00	0.00	0.00	0.00
0.57	0.00	0.73	0.00	0.00	0.00	0.00
0.64	0.00	0.00	0.00	0.00	0.00	0.00
0.71	0.00	0.00	0.00	0.00	0.00	0.00
0.79	0.00	0.00	0.00	0.00	0.00	0.00
0.86	0.00	0.00	0.00	0.00	0.00	0.00
0.93	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00

por lo que se clasificó el SNR de manera subjetiva. Un ejemplo que muestra la mejoría del contorno musical al usar el umbral  $\alpha$  correcto se observa en las figuras 1.18b (grabación C) y 1.19b (grabación B): la primera con una SNR baja y la segunda con una SNR alta. En estas grabaciones las mejores estimaciones de periodicidad se lograron con un valor de  $\alpha = 0.21$  para el primer caso y  $\alpha = 0.29$  para el segundo.

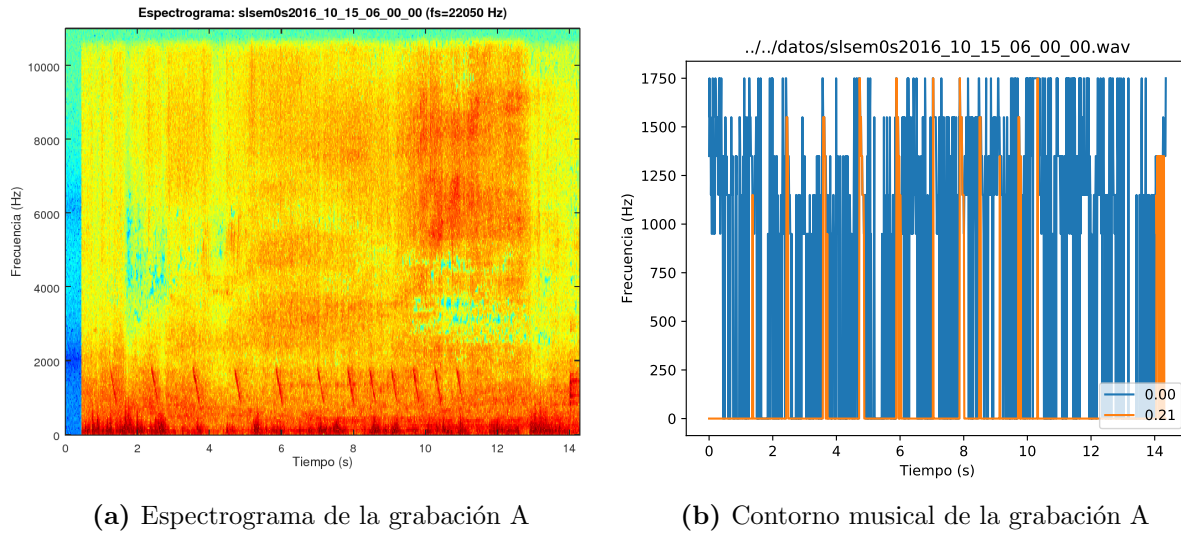
### 1.3 Objetivos y estructura del documento

Con base en el análisis realizado en la sección 1.2, donde se identificaron problemas al usar núcleos de reconocimiento uniarmónicos, tratamientos distintos para el sonido cucú y los chirridos, y el uso de un umbral de tono estático, se establece como objetivo principal el mejorar el reconocimiento de las señales peatonales accesibles creando un algoritmo que identifique el sonido de los APS y como objetivos específicos los siguientes:

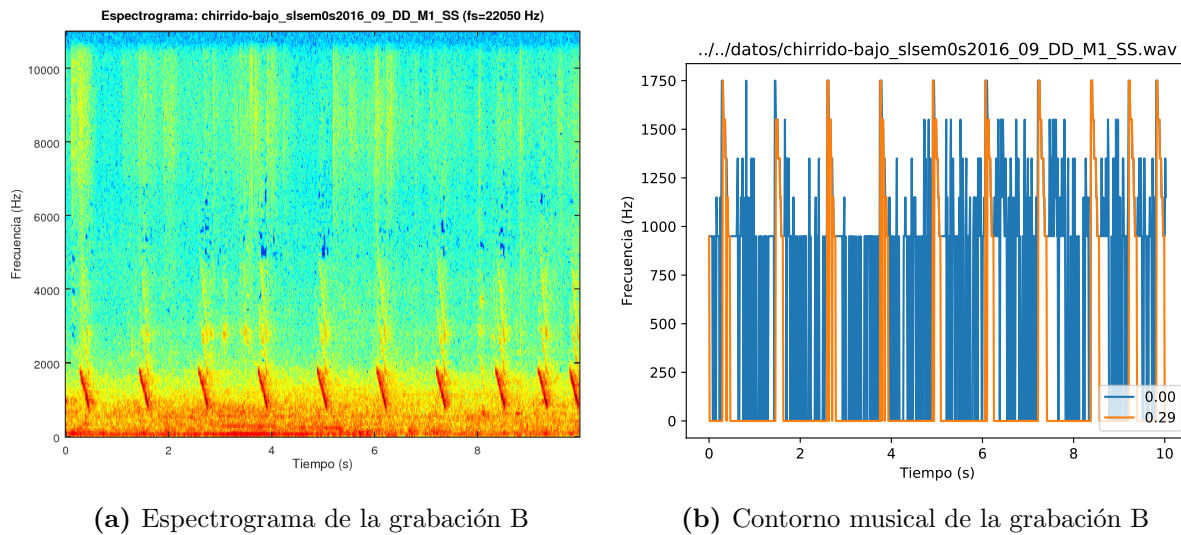
1. Crear una implementación propia del algoritmo RASP que utilice la metodología de evaluación original y verificar que reproduzca las mismas tasas de detección reportadas inicialmente.
2. Mejorar la metodología de evaluación del algoritmo RASP para que penalice los

(a)  $\alpha = 0.00$ (b)  $\alpha = 0.07$ 

**Figura 1.17:** Ejemplo del análisis de periodicidad de dos contornos musicales obtenidos de la misma grabación de chirrido bajo pero utilizando niveles distintos del umbral de tono  $\alpha$ . (1.17a) El pico más alto está en 0.0844 Hz, lo que indica un periodo erróneo. (1.17b) El pico más alto está en 0.9 Hz, lo que indica el periodo correcto de la señal APS.



**Figura 1.18:** (1.18a) Espectrograma y (1.18b) contorno musical de una grabación de chirrido bajo ruidosa para los valores de  $\alpha = 0.00$  y  $\alpha = 0.21$ .



**Figura 1.19:** (1.19a) Espectrograma y (1.19b) contorno musical de una grabación de chirrido bajo limpia para los valores de  $\alpha = 0.00$  y  $\alpha = 0.29$ .



picos de alerta faltantes en las estimaciones realizadas, aumentando, cuando corresponda, la detección de falsos negativos que ocurran dentro de la secuencia principal del APS.

3. Determinar el nuevo rendimiento alcanzando por el algoritmo RASP, mediante la metodología de evaluación propuesta, para tener un umbral inferior de referencia con el cual evaluar las mejoras siguientes.
4. Proponer un diseño de núcleo de reconocimiento de altura musical que asigne un puntaje mayor a la frecuencia fundamental y menor a sus armónicas y subarmónicas.
5. Proponer una distancia que admita los segmentos nulos (ruido o silencio) de los contornos musicales de APS, unificando el tratamiento del sonido cucú con los chirridos.
6. Utilizar un mecanismo de ajuste automático del umbral de tono que varíe proporcionalmente con la relación señal-ruido.
7. Determinar la mejor configuración obtenida para el núcleo, el banco de núcleos, la distancia, y los umbrales de detección.

Para cumplir con los objetivos propuestos, este documento se organiza de la siguiente manera. El capítulo 2 brinda el marco teórico necesario para entender el método propuesto, incluyendo las tasas de detección para medir el rendimiento alcanzado (precisión, sensibilidad, especificidad, medida F y coeficientes de correlación de Mathew), entre otros conceptos. El capítulo 3 describe la solución planteada para resolver los problemas estudiados. El capítulo 4 presenta la metodología de evaluación mejorada e identifica grabaciones con picos de alerta faltantes entre el primer y último comienzo para verificar que se aplique la penalización deseada de un periodo APS. En ese mismo capítulo se presentan las matrices de puntajes generadas por cada núcleo musical (para evaluar la efectividad del núcleo propuesto), y las tasas de detección alcanzadas al analizar las 79 grabaciones disponibles con las mejoras realizadas para los objetivos 4, 5 y 6. Finalmente el capítulo 5 presenta las conclusiones y el trabajo futuro.



# Capítulo 2

## Marco teórico

En este capítulo se explican los conceptos necesarios para entender la solución propuesta. Hasta ahora, se ha estudiado que el algoritmo RASP convierte una señal de audio  $s(t)$ , muestreada a  $f_s$  Hz, en una señal de alerta  $a(t)$ . Esta última contiene los picos de actividad APS detectados para la grabación procesada. El proceso mediante el cual se logra obtener esta señal inicia calculando la STFT de la señal  $s(t)$ . Esta transformada consiste en aplicar la FFT sobre ventanas de largo  $T_v$  segundos, formando otra señal denotada como  $|S(t, f)|$  con los componentes espectrales. La sección 2.1 explica cómo hallar el tamaño óptimo para calcular  $T_v$  según el tipo de ventana elegido, y con ello evitar distorsiones en la señal de salida. La sección 2.2 explica el algoritmo SWIPE, empleado por RASP, para estimar la altura musical. Este algoritmo analiza la señal  $|S(t, f)|$  y produce un contorno musical  $c(t)$  y un vector de puntajes  $p(t)$ . La sección 2.2 también explica los conceptos psicoacústicos necesarios para entender SWIPE y el algoritmo de ajuste automático de tono TS2Means, el cual genera una señal  $p'(t)$  a partir del vector de puntajes, y separa los sonidos armónicos de los no armónicos de forma dinámica según la SNR, resolviendo el problema de los umbrales fijos explicado en la sección 1.2.2. La sección 2.3 explica la teoría básica para entender la *media móvil exponencial*, un tipo de filtro digital pasabajas de bajo costo computacional que podría usarse como sustituto del TS2Means. La sección 2.4 introduce una serie de medidas de similitud usadas para reconocer las correspondencias del contorno musical con las plantillas musicales de los APS y la distancia de Mahalanobis, que, a diferencia de otras métricas, permite manejar segmentos ruidosos como los mencionados en la sección 1.2.3, tomando en cuenta la covarianza de los datos. La sección 2.5 muestra cómo calcular las tasas de detección (precisión, sensibilidad, especificidad, medida F y coeficiente de correlación de Matthew) usadas para cuantificar el rendimiento del método propuesto y compararlo con las soluciones existentes. Por último, la sección 2.6 explica la razón por la que las redes neuronales convolucionales no se usan en este trabajo.

## 2.1 Tamaño mínimo del inventariado

La elección de un tamaño adecuado de ventana al aplicar la STFT es importante para obtener espectros sin deformación y con una buena resolución temporal. Estos espectros afectan la calidad de las estimaciones realizadas. En esta sección se explica el tamaño mínimo que deben tener los tres tipos de ventana más utilizados en el análisis de frecuencias: rectangular, de Hann y de Hamming, para modelar una señal armónica evitando la interferencia producida por el traslape de los lóbulos principales espectrales propios de cada ventana.<sup>1</sup>

### 2.1.1 Ventana rectangular

Por la propiedad del escalamiento en el tiempo se sabe que la transformada de Fourier de una ventana rectangular de ancho  $T_v > 0$  es un seno cardinal (senc) con cruces en cero cada  $1/T_v$  Hz (el lóbulo principal abarca el rango  $[-1/T_v, 1/T_v]$ ), como se observa en la figura 2.1. Esto se representa matemáticamente en la siguiente expresión [23, 24]:

$$\square(t/T_v) \xrightarrow{\mathcal{F}} T_v \text{senc}(T_v f), \quad (2.1)$$

donde las funciones  $\square(t)$  y  $\text{senc}(f)$  se definen como sigue:

$$\square(t) = \begin{cases} 1 & |t| < 1/2 \\ 0 & \text{en el resto} \end{cases}$$

y

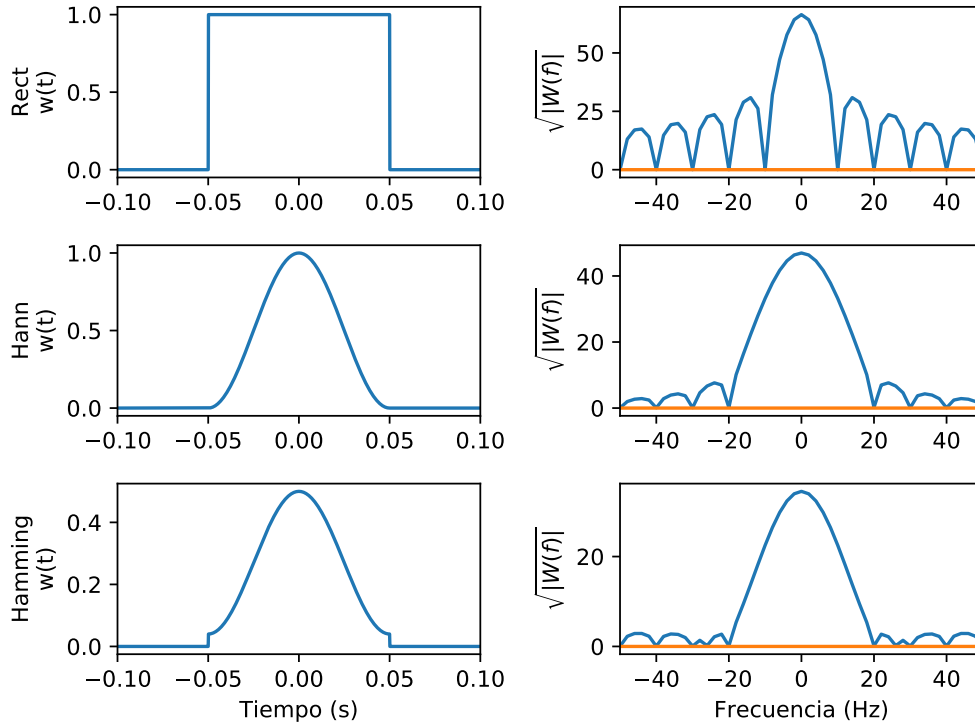
$$\text{senc}(f) = \begin{cases} \frac{\sin(\pi f)}{(\pi f)} & f \neq 0 \\ 1 & f = 0. \end{cases}$$

Al procesar un tono complejo  $x(t)$  compuesto por la suma de cosenos  $\sum_{k=0}^{\infty} \cos(2\pi k f_0 t)$ , donde  $f_0$  es la frecuencia fundamental del tono,  $k \in \mathbb{Z}^+$  y  $t$  es el tiempo discreto, la multiplicación de la ventana rectangular por cada segmento de la señal de audio produce en la magnitud del espectro una convolución de la función del seno cardinal con las armónicas presentes. Por cuestiones de simplicidad se asume que la frecuencia fundamental y las armónicas tienen amplitud unitaria. El resultado se expresa como sigue:

$$\square(t/T_v) x(t) \xrightarrow{\mathcal{F}} T_v \text{senc}(T_v f) * \frac{1}{2} \sum_{k=0}^{\infty} \delta(f \pm k f_0), \quad (2.2)$$

donde el tren de impulsos del lado derecho se deriva de la expresión  $\cos(2\pi f_0 t) \xrightarrow{\mathcal{F}} 0.5\delta(f \pm f_0)$ , y la operación de convolución puede interpretarse como un desplazamiento del seno cardinal al valor  $k f_0$  [23]. Tomando la primer armónica en  $f_0$ , y la segunda en  $2f_0$ , se tiene que los cruces por cero de los lóbulos principales de la función senc ocurren en  $f_0 \pm 1/T_v$  y

<sup>1</sup>Esta sección se hizo con base en los resultados publicados para una asignación del curso de Procesamiento de Sonido MP-6154.



**Figura 2.1:** Ejemplos de la señal en el tiempo y la frecuencia de ventanas rectangular, Hann y Hamming de duración  $T_v = 0.1$  s. Los lóbulos principales tienen sus cruces por cero en  $\pm 1/T_v$  para el caso de la ventana rectangular y  $\pm 2/T_v$  para el resto. La raíz cuadrada en la magnitud espectral es usada para visualizar mejor los lóbulos secundarios.

$2f_0 \pm 1/T_v$  (como se observa en la figura 2.2), lo que indica que para evitar que los lóbulos espectrales principales se traslapen entre sí es necesario satisfacer la siguiente condición:<sup>2</sup>

$$\forall i \in \mathbb{Z}^+ [if_0 + 1/T_v < (i+1)f_0 - 1/T_v], \quad (2.3)$$

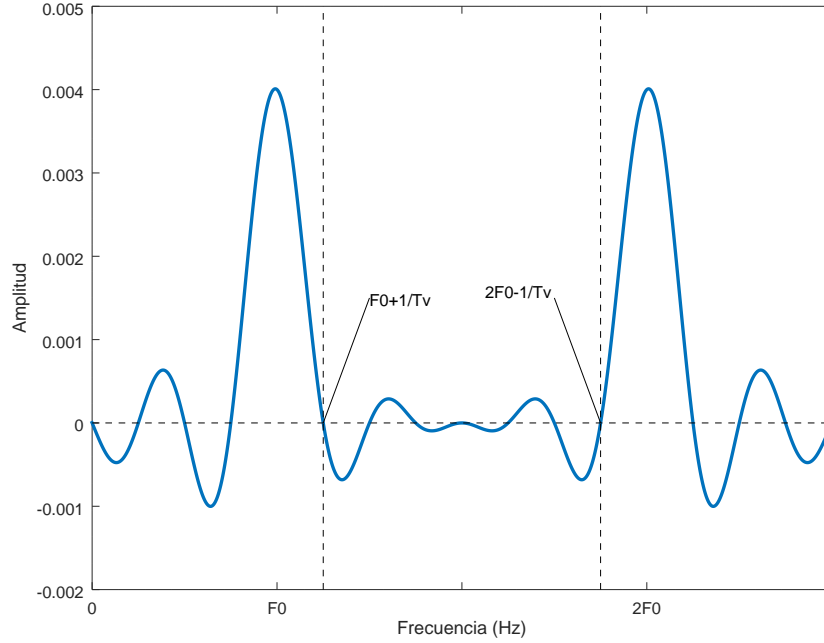
es decir, que el tamaño de la ventana de análisis rectangular debe ser igual o mayor a dos ciclos de una señal periódica ( $2T_0 < T_v$ ) para evitar que los lóbulos espectrales principales se intersequen y haya distorsiones en la STFT.

### 2.1.2 Ventanas de Hann y de Hamming

Las ventanas de Hann y de Hamming también son populares en el cálculo de la STFT. Por eso es necesario repetir el razonamiento de la sección 2.1.1 empleando la ecuación que generaliza estas dos funciones [25, 26]:

$$h(t) = \frac{1}{2} \Pi \left( \frac{t}{T_v} \right) \left[ \alpha + \beta \cos \left( \frac{2\pi t}{T_v} \right) \right], \quad (2.4)$$

<sup>2</sup>Se toman las frecuencias  $f_0$  y  $2f_0$  en lugar de  $-f_0$  y  $f_0$  porque la distancia entre las primeras es menor, es decir,  $2f_0 - f_0 < f_0 - (-f_0)$ .



**Figura 2.2:** Funciones seno cardinal localizadas en  $f_0$  y  $2f_0$  pertenecientes al espectro de una ventana rectangular de ancho  $T_v$ .

con  $\alpha = 1$  y  $\beta = 1$  para la ventana de Hann y  $\alpha = 0.54$  y  $\beta = 0.46$  para la ventana de Hamming.<sup>3</sup> Como se observa en la figura 2.1, su transformada de Fourier es una suma de tres funciones seno cardinal, que se cancelan entre sí formando un solo lóbulo espectral con cruces por cero en  $\{\pm 2/T_v, \pm 3/T_v, \pm 4/T_v, \dots\}$  (como se ve observa en la figura 2.1), lo que se expresa como sigue:

$$H(f) = \alpha \operatorname{senc}(fT_v) + \frac{\beta}{2} \operatorname{senc}\left[T_v\left(f - \frac{1}{T_v}\right)\right] + \frac{\beta}{2} \operatorname{senc}\left[T_v\left(f + \frac{1}{T_v}\right)\right].$$

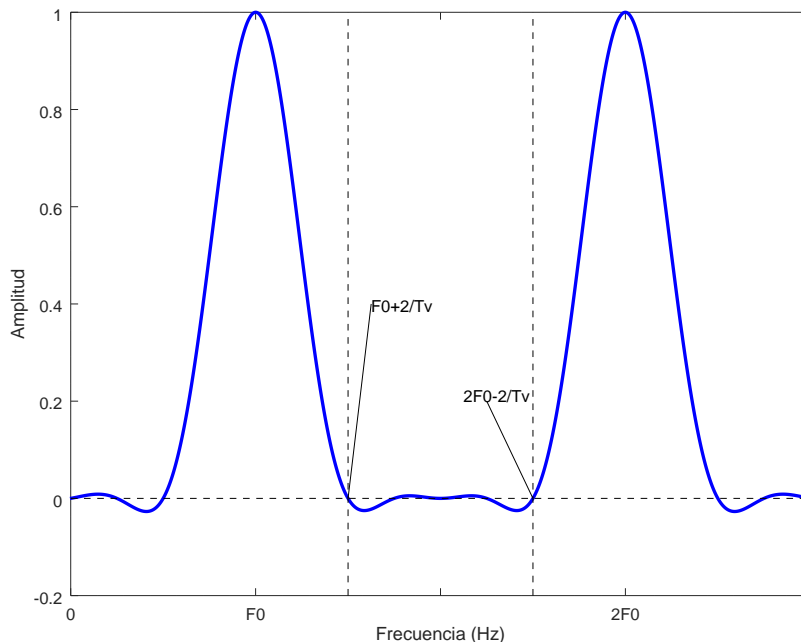
Al igual que en el caso rectangular, cuando se procesa un tono complejo  $x(t)$ , compuesto por una suma de cosenos, la multiplicación de la ventana Hann por cada segmento de la señal de audio produce en la magnitud del espectro una convolución de la función  $H(f)$  con las armónicas presentes. El resultado se expresa como sigue:

$$h(t)x(t) \xrightarrow{\mathcal{F}} H(f) * \frac{1}{2} \sum_{k=0}^{\infty} \delta(f \pm kf_0), \quad (2.5)$$

Tomando la primer armónica en  $f_0$ , y la segunda en  $2f_0$ , se tiene que los cruces por cero de los lóbulos principales de la función  $H(f)$  ocurren en  $f_0 \pm 2/T_v$  y  $2f_0 \pm 2/T_v$  (como se observa en la figura 2.3), lo que indica que para evitar que los lóbulos espectrales principales se traslapen entre sí es necesario satisfacer la siguiente condición:

$$\forall i \in \mathbb{Z}^+ [if_0 + 2/T_v < (i+1)f_0 - 2/T_v], \quad (2.6)$$

<sup>3</sup>La ventana Hamming es una optimización de la ventana Hann, que busca minimizar la amplitud de los lóbulos secundarios.



**Figura 2.3:** Dos lóbulos localizados en  $f_0$  y  $2f_0$  pertenecientes al espectro de una ventana Hann de ancho  $T_v$ . Para el caso de la ventana Hamming, la interacción es casi idéntica.

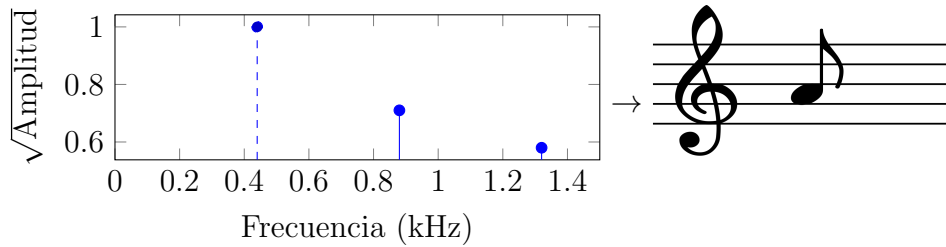
lo que significa que la ventana de análisis de Hann o Hamming debe ser igual o mayor a cuatro ciclos de una señal periódica ( $4T_0 < T_v$ ) para que los lóbulos principales no se traslapen.

## 2.2 Reconocimiento de la altura musical

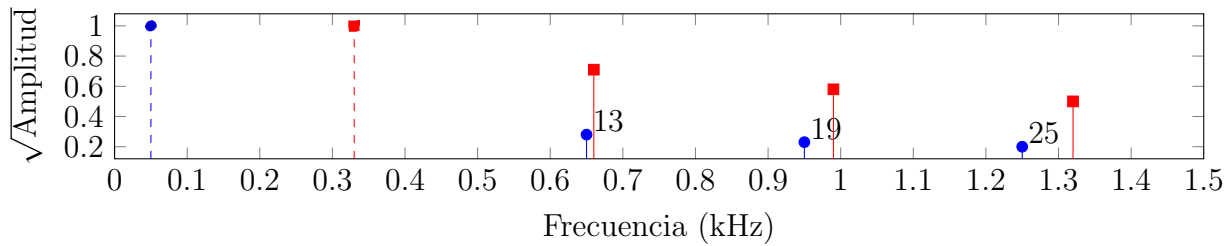
A continuación, la sección 2.2.1 explica qué es la altura musical y cómo se diferencia del tono, la sección 2.2.2 explica en qué consiste la escala psicoacústica ERB y su función en el modelado del oído humano, la sección 2.2.3 estudia un algoritmo de reconocimiento de altura musical llamado SWIPE, que utiliza la escala ERB y la sección 2.2.4 explica el algoritmo TS2Means, que procesa la señal de puntajes producida por SWIPE para separar los sonidos armónicos de los no armónicos.

### 2.2.1 Altura musical

La altura musical se define como la sensación auditiva que permite ordenar las notas musicales en una escala, de grave a agudo [23]. También es un fenómeno psicoacústico, en el que la percepción subjetiva del cerebro busca explicar la energía presente en la magnitud espectral de los sonidos percibidos, identificando la frecuencia fundamental del sonido y sus armónicas. La altura musical, al ser una valoración propia del oído humano,



**Figura 2.4:** Ejemplo de la altura musical de la nota LA4 a partir de un espectro con frecuencia fundamental faltante (línea intermitente).



**Figura 2.5:** Distribución de la energía de un tono de 50 Hz con sus armónicas 13, 19 y 25 (650 Hz, 950 Hz y 1250 Hz), representadas con marcadores circulares, y la energía de un tono de 330 Hz con sus armónicas 2, 3 y 4 (660 Hz, 990 Hz y 1320 Hz), representadas con marcadores cuadrados.

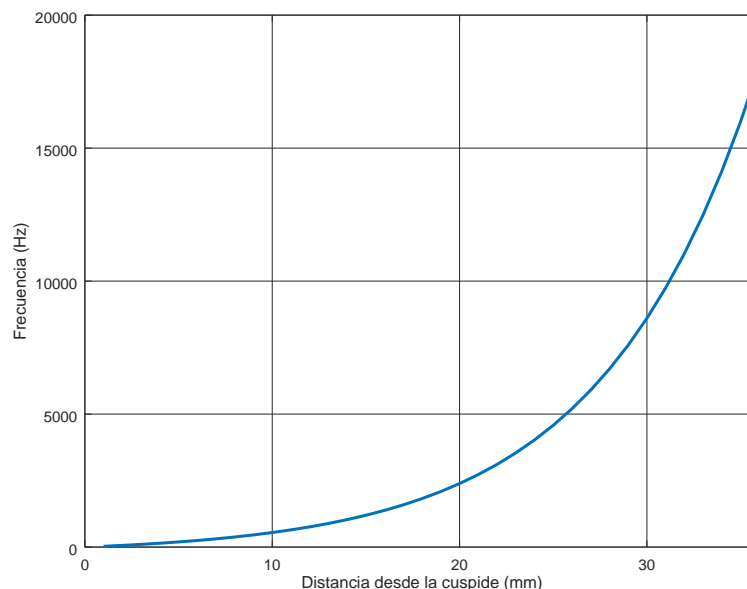
no siempre es igual al tono. Dos ejemplos que lo ilustran pueden encontrarse en las figuras 2.4 y 2.5. En el primer caso, la frecuencia fundamental está ausente y el cerebro la reconstruye calculando el máximo común denominador (MCD) de las armónicas presentes. En el segundo caso, el cerebro confunde las armónicas 13, 19 y 25 de un tono grave, con las armónicas 2, 3 y 4 de un tono más agudo desafinado [23].<sup>4</sup> La razón por la que se estudia la altura musical y no el tono, es que permite diseñar algoritmos que simulen la forma cómo las personas escuchan, es decir, contando con una mayor resolución de las frecuencias bajas y una menor resolución de las frecuencias altas, pues la altura musical está distribuida en una escala (semi) logarítmica, lo que también es aprovechado en el diseño de los sonidos APS.

## 2.2.2 Escala ERB

La teoría psicoacústica afirma que la distribución de frecuencias de la membrana basilar dentro de la cóclea (ilustrada en la figura 2.7) es aproximadamente logarítmica, es decir, que existe resolución lineal para procesar las frecuencias bajas del rango 0–230 Hz, y una resolución logarítmica para las frecuencias altas del rango 230–20 kHz [28]. El *ancho*

<sup>4</sup>Ejemplo tomado de los apuntes de Arturo Camacho [25] y originalmente descubierto por Patel *et al.* en experimentos psicoacústicos realizados en sujetos humanos [27].





**Figura 2.6:** Distribución de frecuencias por posición en la cóclea, según la escala ERB.

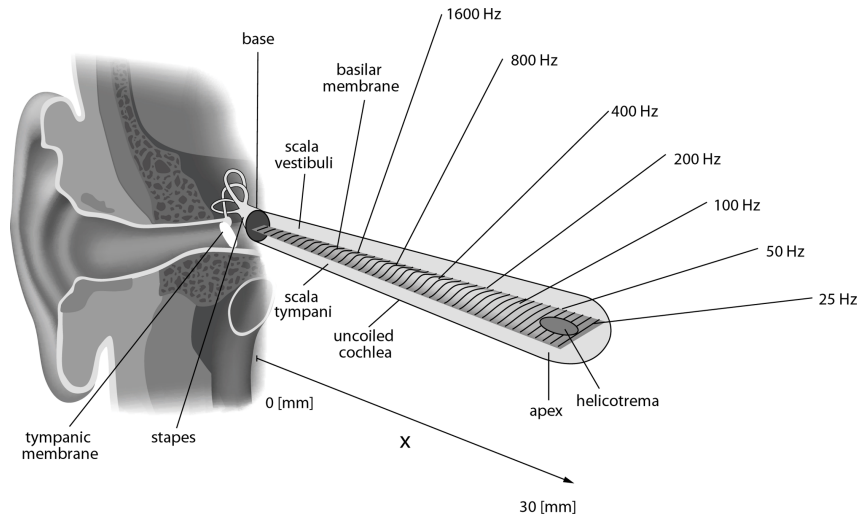
de banda rectangular equivalente (ERB, por sus siglas en inglés) modela este fenómeno mediante la siguiente expresión [29, 28, 23, 30]:

$$d(f) = 5.7\text{mm} \log_2(1 + f/230\text{Hz}),$$

donde  $0 \leq d(f) \leq 36$  es la distancia en milímetros (desde la cúspide de la cóclea hasta su base) en donde la frecuencia  $f$  Hz es reconocida. La figura 2.6 permite observar una representación gráfica de esta escala. En general, las escalas psicoacústicas han sido de gran utilidad para crear algoritmos de compresión de sonido, como el MPEG-layer-3 (MP3), pues han permitido ahorrar información del audio suprimiendo la energía de las frecuencias inaudibles. Precisamente, las *bandas críticas* determinan si dos frecuencias son lo suficientemente cercanas como para ignorar la más débil de ellas y obtener una distorsión baja del sonido, lo que se conoce como *efecto de enmascaramiento* [24].

### 2.2.3 Reconocimiento de la altura musical

Los algoritmos de detección de altura musical permiten recuperar información valiosa de la fuente del sonido, como por ejemplo: el género del hablante, la entonación de una palabra y el nombre de la nota musical. Con esto se han podido resolver problemas más complejos como la transcripción musical automática, la consulta de canciones por tarareo, la compresión de audio y la detección de desórdenes de la voz. Uno de los métodos más exitosos reconociendo la altura musical es el algoritmo *sawtooth waveform inspired pitch estimator* (SWIPE) [23], que encuentra la frecuencia fundamental buscando la onda diente de sierra cuyo espectro se asemeje más al espectro del sonido analizado. SWIPE se basó en cinco algoritmos de estimación de altura musical previos: producto de armónicas (*harmonic product spectrum*) [32], suma de armónicas (*subharmonic summation*) [33], suma



**Figura 2.7:** Ilustración de la cóclea [31]. En la imagen los autores usaron una aproximación de 30 mm y midieron la distancia desde la base hasta la cúspide. En este trabajo se usa una longitud máxima de 36 mm y la distancia se mide desde la cúspide hasta la base.

de armónicas ponderada [33], tasa de armónicas a subarmónicas (*subharmonic to harmonic ratio*) [34], y autocorrelación [35]. Además, emplea la escala ERB para implementar un decaimiento proporcional a  $1/k^2$  de las armónicas de su núcleo de reconocimiento. La fórmula original, propuesta por Camacho, se define como sigue:

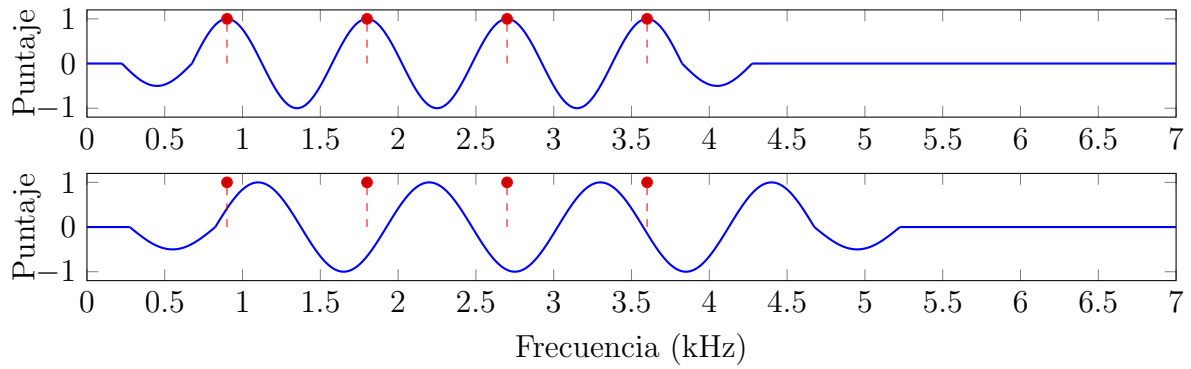
$$p(t) = \operatorname{argmax}_{f_0} \frac{\int_0^{\operatorname{ERBs}(f_{\max})} \frac{1}{\eta(\varepsilon)^{1/2}} \sqrt{|X(t, f_0, \eta(\varepsilon))|} K(f_0, \eta(\varepsilon)) d\varepsilon}{\left( \int_0^{\operatorname{ERBs}(f_{\max})} |X(t, f_0, \eta(\varepsilon))| d\varepsilon \right)^{1/2} \left( \int_0^{\operatorname{ERBs}(f_{\max})} \frac{1}{\eta(\varepsilon)} [K^+(f_0, \eta(\varepsilon))]^2 d\varepsilon \right)^{1/2}}, \quad (2.7)$$

donde  $f_{\max}$  es la frecuencia máxima a considerar,  $\varepsilon$  es la frecuencia en la escala ERB y  $(\cdot)^+ = \max\{0, \cdot\}$  es el operador de rectificación de media onda utilizado en el denominador para evitar que la suma de las entradas del núcleo musical se anule e indefina la fracción. El núcleo de reconocimiento musical  $K(f_0, f)$  (diseñado para admitir armónicas desafinadas o ausentes) se define como sigue:

$$K(f_0, f) = \begin{cases} \cos(2\pi f_0/f), & 3/4 < f_0/f < n(f_0) + 1/4 \\ \frac{1}{2} \cos(2\pi f_0/f), & 1/4 < f_0/f < 3/4 \vee n(f_0) + 1/4 < f_0/f < n(f_0) + 3/4 \\ 0 & \text{en el resto,} \end{cases} \quad (2.8)$$

donde  $n(f_0) = \lfloor f_{\max}/f_0 - 3/4 \rfloor$  es la cantidad de armónicas del núcleo que caben dentro del rango  $[0, f_{\max}]$ . En la ecuación 2.7,  $X(t, f_0, f')$  es la transformada de Fourier de corto plazo de  $x(t)$  definida como sigue:

$$X(t, f_0, f') = \int_{-\infty}^{\infty} w_{4k/f_0}(t' - t) x(t') e^{-j2\pi f' t'} dt', \quad (2.9)$$



**Figura 2.8:** Ejemplo de 2 núcleos SWIPE de 900 Hz y 1100 Hz (sin decaimiento), aplicados a la magnitud espectral de un tono de 900 Hz con 4 armónicas. El puntaje del primer núcleo es el máximo obtenible, lo que lo convierte en el núcleo ganador.

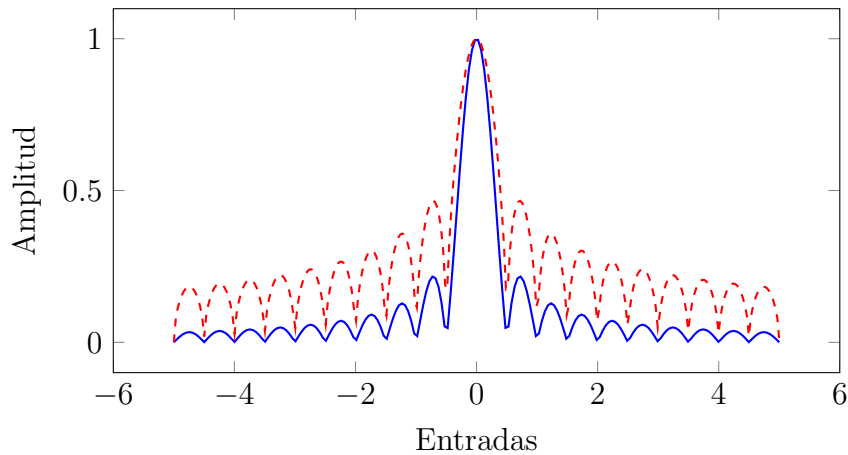
donde  $w_{4k/f_0}(t' - t)$  es una ventana de análisis en el dominio del tiempo (por ejemplo, Hamming o Hann) optimizada para analizar la frecuencia  $f_0$  y mitigar el principio de incertidumbre. La suma del núcleo musical es aproximadamente cero a propósito para asignar una altura nula al ruido (de hecho, es ligeramente negativa debido al decaimiento del núcleo). La figura 2.8 muestra un ejemplo de cómo se aplican los núcleos musicales a cada ventana  $|X(t, f_0, f)|$  de un tono musical puro de 900 Hz.

Como se explicó en las secciones 1.1.4 y 1.1.6, los núcleos musicales no son espectros de audios reales, sino puntajes que se aplican para saber si la distribución de energía se parece o no al tono deseado. El término  $\eta(\epsilon)^{-0.5}$  de la ecuación 2.7, es un decaimiento de las armónicas de los núcleos de reconocimiento que busca evitar que al analizar tonos con frecuencia fundamental ausente, alguna de sus armónicas obtenga el mismo puntaje que la frecuencia fundamental. Este decaimiento es  $p$ -armónico, pues sigue la forma  $1/k^p$ , donde  $p$  es el decaimiento elegido y  $k$  es el índice de la armónica. Además, existe otro tipo de decaimiento llamado geométrico, que sigue la forma  $r^{k-1}$ , donde  $r < 1$  es la base deseada y  $k$  es el índice de la armónica. Como ya se mencionó, el primer tipo fue empleado en la fórmula original de SWIPE, y el segundo tipo fue usado por el algoritmo RASP con un valor de  $r = 0.86$  [10, 23]. Una versión más simple de SWIPE, que no usa la escala ERB, puede formularse como sigue [10]:

$$p(t) = \operatorname{argmax}_{f_0} \frac{\int_0^{f_{\max}} \sqrt{|X(t, f_0, f)|} K(f_0, f) df}{\left( \int_0^{f_{\max}} |X(t, f_0, f)| df \right)^{1/2} \left( \int_0^{f_{\max}} [K^+(f_0, f)]^2 df \right)^{1/2}}. \quad (2.10)$$

Cabe mencionar que SWIPE evolucionó al *sawtooth waveform inspired pitch estimator prime* (SWIPEP), un algoritmo que utiliza núcleos de reconocimiento con armónicas primas para reducir la tasa de error por subarmónicos. Esto es un tipo de error producido cuando un subarmónico de  $f_0$  es identificado erróneamente como la frecuencia fundamental [23].

Como puede observarse en la figura 2.9, se utiliza la raíz cuadrada de la magnitud espectral

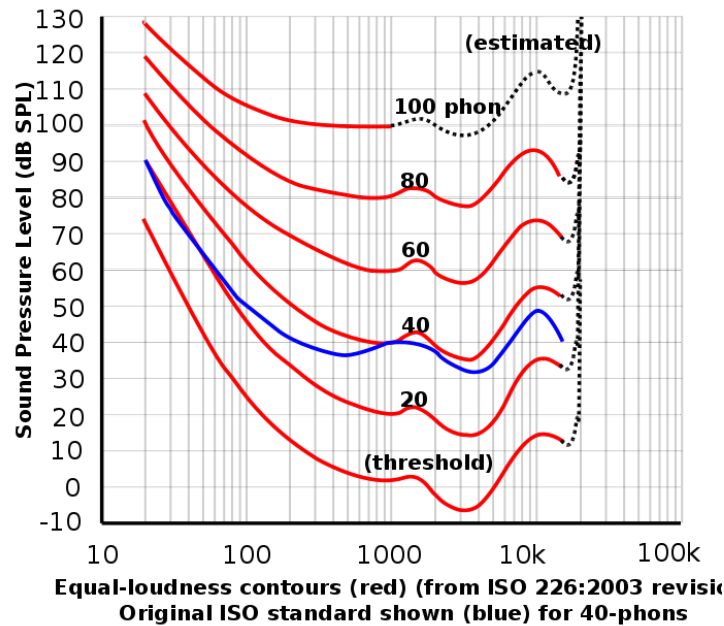


**Figura 2.9:** Sonoridad de un seno cardinal. La línea continua representa la amplitud original y la intermitente representa la aproximación de la sonoridad.

para mejorar la resolución de la amplitud de las frecuencias, aumentando las amplitudes pequeñas y manteniendo constantes a las grandes, es decir,  $S \propto A^{1/2}$  [23]. Este escalamiento también es usado como una aproximación de un término psicoacústico llamado *sonoridad* o *loudness*, que se define como la sensación auditiva que permite ordenar los sonidos en una escala de intensidad sonora, de fuerte a débil [28]. La sonoridad depende de las frecuencias presentes en los tonos, siendo los sonidos en el rango 1000 – 6000 Hz los que se escuchan más fuertemente. Esto se comprueba en las curvas isofónicas o *equal-loudness contour* de la figura 2.10, donde hay un mínimo en el rango señalado. La unidad de medida de la sonoridad es el *sonio*, que tiene un valor unitario para un tono de 1 kHz a un nivel de intensidad sonora de 40 dB. El *fonio*, que también es otra unidad de medida de la sonoridad, ha caído en desuso debido a que no es linealmente proporcional a la sonoridad [28]. Nótese que aunque el nivel de intensidad sonora y el nivel de poder sonoro se miden en  $W/m^2$  (vatios entre metros al cuadrado) y que ambos se expresan en la escala de decibelios, el nivel de poder sonoro no depende de la distancia desde la fuente del sonido, como sí depende el nivel de intensidad sonora [36].

## 2.2.4 Umbralización dinámica en el tiempo

SWIPE produce como salida el contorno musical detectado y el puntaje de las frecuencias ganadoras en cada instante. La segunda señal brinda información sobre la certeza con que se reconoció la frecuencia fundamental y permite identificar los sonidos armónicos de los no armónicos (los segundos tienen puntaje cero o negativo). El algoritmo *time-series 2-means* (TS2Means) [38] facilita la tarea de discretizar la señal de puntajes en estos dos tipos de sonidos. Esto no es trivial de lograr porque la señal de puntajes es continua, y la división entre niveles no siempre es clara. En señales ruidosas tiende a ser plana o a sufrir una modulación de baja frecuencia. El TS2Means emplea agrupación basada en el algoritmo de *k-vecinos más próximos* (*k*-nn, por sus siglas en inglés), que consiste en



**Figura 2.10:** Curvas isofónicas del oído humano [37]. Se observa que para un tono de 1 kHz la sonoridad es equivalente al nivel de intensidad sonora. Un ejemplo de como usar las curvas es el siguiente: si se quiere percibir un tono de 100 kHz igual de fuerte que un tono de 1 kHz a 0 dB, sería necesario subir el “volumen” hasta 25 dB.

asignar a cada muestra la clase más utilizada por sus  $k$  vecinos [39].<sup>5</sup> TS2Means funciona así, se define una función  $c_i(n, \mu, N)$  que determina el valor del centroide en cada instante  $n$  como sigue:

$$c_i(n, \mu, N) = \frac{\sum_{n'=-N}^N p_i s(n' + n) w_N(n')}{\sum_{n'=-N}^N p_i w_N(n')}, \quad (2.11)$$

donde  $i$  es el identificador del centroide (0 para los sonidos armónicos y 1 para los no armónicos),  $\mu(n) \in \{0, 1\}$  es una función de pertenencia de los puntajes  $s(n)$  a cada centroide (encontrada usando  $k$ -nn),  $N$  es el número de vecinos derechos e izquierdos a considerar,  $w_N(n')$  es una ventana de análisis Hann centrada en  $n$  y de tamaño  $2N + 1$ ,  $n'$  es el índice de desplazamiento dentro de la ventana, y  $p_0 = \mu(n' + n)$  y  $p_1 = 1 - \mu(n' + n)$  son valores que indican si el puntaje  $s(n + n')$  debe o no incluirse en la actualización del valor del centroide (se excluyen los valores de clases distintas) [38]. Se utiliza una ventana de Hann, en lugar de una ventana rectangular, porque se desea calcular un promedio ponderado con los  $2N + 1$  vecinos abarcados. La ponderación posee un valor máximo en el centro, un valor mínimo en los extremos, y un decaimiento exponencial en el medio, lo que otorga más peso al valor actual y menos a los valores cercanos a los extremos. Luego, se procede a calcular la combinación de parámetros  $N^*$  y  $\mu^*(n)$  que maximice la

<sup>5</sup>Los vecinos son contados incrementando el radio de una región esférica cuyo centro es la muestra que está siendo clasificada.

separación entre los centroides, según el siguiente criterio:

$$[\mu^*(n), N^*(n)] = \underset{\mu, N}{\operatorname{argmax}} \{c_1(n, \mu, N) - c_0(n, \mu, N)\}, \quad (2.12)$$

donde la optimización es realizada mediante un barrido de valores en el rango  $T_{\min} \leq N/f_s \leq T_{\max}$ , y en cada iteración se repite el cálculo de  $\mu(n)$ . El paso final consiste en clasificar definitivamente cada puntaje como armónico si cumple la condición  $s(n) \geq 0.5(c_1(n, \mu, N) + c_0(n, \mu, N))$ , y como no armónico en caso contrario [38]. A pesar de los buenos resultados reportados por TS2Means, una desventaja de este método es que al utilizar los  $M$  vecinos derechos de cada entrada de la señal de puntajes, el algoritmo no es causal y por lo tanto no puede emplearse en sistemas de tiempo real donde no se conoce la totalidad de la señal con antelación.

## 2.3 Media móvil exponencial

Como se explicó en la sección 2.2.4, el algoritmo TS2Means aplica un promedio ponderado de los  $M$  vecinos izquierdos y derechos de cada entrada de la señal de puntajes para actualizar los centroides armónicos e inarmónicos. El promedio se actualiza incluyendo a los vecinos clasificados con el mismo tipo, mediante el algoritmo  $k$ -nn, y escogiendo el mejor número de vecinos a considerar. Otro algoritmo más simple que también permite calcular el promedio local de una función  $x[n]$  analizando los  $M - 1$  vecinos de cada entrada es la *media móvil* (MA, por sus siglas en inglés) y, aunque procesa solo los vecinos izquierdos y no maneja dos centroides para cada tipo de sonido, permite determinar un punto medio entre las clases armónicas e inarmónicas de manera causal, para sistemas en tiempo real. La MA, se puede especificar mediante la siguiente convolución:

$$y_M[n] = \frac{1}{M} \sum_{k=0}^{M-1} x[n - k]. \quad (2.13)$$

Su función de transferencia, obtenida mediante la transformada Z, se define de la siguiente manera:

$$H(z) = \frac{1}{M} \sum_{k=0}^{M-1} z^{-k},$$

donde la sumatoria tiene la forma de una serie de potencias y puede reducirse a una expresión más sencilla usando la ecuación general  $\sum_{k=0}^K \alpha^k = (1 - \alpha^{K+1})/(1 - \alpha)$  con  $\alpha = z^{-1}$ . La expresión resultante es:

$$H(z) = \frac{1}{M} \frac{(1 - z^{-M})}{(1 - z^{-1})}. \quad (2.14)$$

Los ceros y polos de esta ecuación se pueden encontrar haciendo la sustitución  $z = e^{j\omega}$ , lo que devuelve la respuesta en frecuencia, y luego multiplicando la fórmula por  $(e^{j\omega(M+1)})/(e^{j\omega(M+1)}) = 1$ , lo que da como resultado:

$$H(e^{j\omega}) = \frac{1}{M} \frac{(e^{j\omega M} - 1)}{(e^{j\omega} - 1)} \frac{1}{e^{j\omega(M-1)}}, \quad (2.15)$$

donde se aprecia que hay un cero y un polo en  $\omega = 0$  rad/s, y que ambos se cancelan entre sí. El resto de ceros ocurren cuando  $e^{j\omega M} = 1$ , es decir, cuando  $j\omega M$  es múltiplo de  $\pm 2\pi$ , lo que se puede calcular variando  $k \in \mathbb{Z}$  dentro de la ecuación  $\omega = 2\pi k/M$  rad/s. Estos ceros dividen al círculo unitario en  $M$  secciones ubicadas en  $\omega = 0, 2\pi/M, \dots, 2\pi(M-1)/M$  rad/s; por ejemplo, con  $M = 30$ , se producen divisiones en el círculo unitario ubicadas en los instantes  $\omega = 0, 2\pi/30, \dots, 2\pi \cdot 29/30$  rad/s, como se observa en la figura 2.11b [24]. El primer cruce con el eje  $\omega$  distinto de cero se da en  $2\pi/30$  rad/s, lo que, según la figura 2.11a, corresponde al cruce por  $\omega$  del lóbulo central en la magnitud espectral  $|H(e^{j\omega})|$ . Como este lóbulo es más grande que los lóbulos secundarios, permite interpretar a la media móvil como un *filtro pasabajas* con frecuencia de corte  $f_c$ , definida como [24]:

$$f_c = \frac{2\pi}{30} \cdot \frac{f_s}{2\pi},$$

donde  $f_s$  es la frecuencia de muestreo. Haciendo por ejemplo  $f_s = 22050$  Hz, se obtendría un filtro que atenúa la energía de las frecuencias por encima de  $f_c = 735$  Hz. Sin embargo, el filtrado pasabajas de la media móvil tiene el inconveniente de que la banda de transición, que separa la banda amplificada de la banda atenuada no es pronunciada, lo que afecta la especificidad del filtrado. Otra interpretación de la media móvil es la de ser un filtro peine desplazado, pues los lóbulos secundarios están ubicados en múltiplos de una frecuencia en común. De hecho al hacer  $L = 1$ , la fórmula 2.13 es similar a la de los filtros peine [40]:

$$H(z) = \sum_{k=0}^{M-1} h[k]z^{-kL}.$$

Como punto final, se advierte que la aplicación de la media móvil en el dominio temporal usando la convolución 2.13 es costoso, pues requiere que el sistema promedie  $M$  entradas en cada iteración, las cuales deben ser almacenadas en un espacio en memoria de ese tamaño. Para efectos de realizar un procesamiento menos costoso, se utiliza su ecuación de diferencias, definida como sigue:

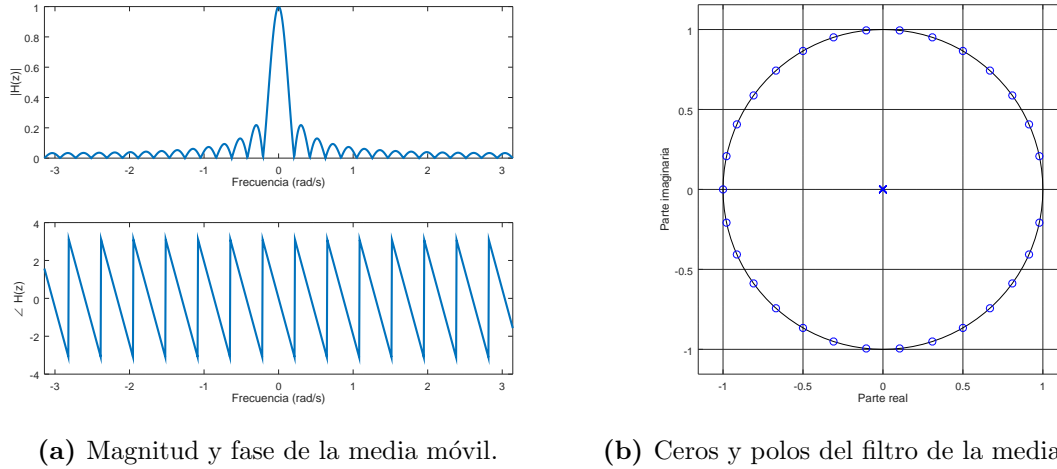
$$y_M[n] = y_M[n-1] + \frac{1}{M}(x[n] - x[n-M]), \quad (2.16)$$

la cual se deriva de la ecuación 2.14 y es recursiva. Su implementación requiere solo de dos sumas, una división y dos entradas en memoria: una para la salida anterior y otra para la  $M$ -ésima entrada pasada.

## Media móvil exponencial

Una forma de interpretar la ecuación de diferencias 2.16, para que solo dependa de la entrada  $x[n]$ , es obtenida como sigue [24]:

$$\begin{aligned} y_M[n] &= y_M[n-1] + \frac{1}{M}(x[n] - x[n-M]), \\ &= \frac{1}{M} \sum_{k=0}^{M-1} x[n-1-k] + \frac{1}{M}(x[n] - x[n-M]), \end{aligned} \quad \text{def. 2.13}$$



**Figura 2.11:** (2.11a) Diagrama de Bode y (2.11b) diagrama de ceros y polos de la media móvil con  $M = 30$ .

$$\begin{aligned}
 &= \frac{1}{M} \left[ \left( x[n-1-(M-1)] + \sum_{k=0}^{M-2} x[n-1-k] \right) + (x[n] - x[n-M]) \right], \\
 &= \frac{1}{M} \left( \sum_{k=0}^{M-2} x[n-1-k] + x[n] + (x[n-1-(M-1)] - x[n-M]) \right), \\
 &= \frac{1}{M} \left( \frac{M-1}{M-1} \sum_{k=0}^{M-2} x[n-1-k] + x[n] \right), \\
 &= \frac{1}{M} [(M-1)y_{M-1}[n-1] + x[n]], \quad \text{def. 2.13} \\
 &= \frac{M-1}{M} y_{M-1}[n-1] + \frac{1}{M} x[n].
 \end{aligned}$$

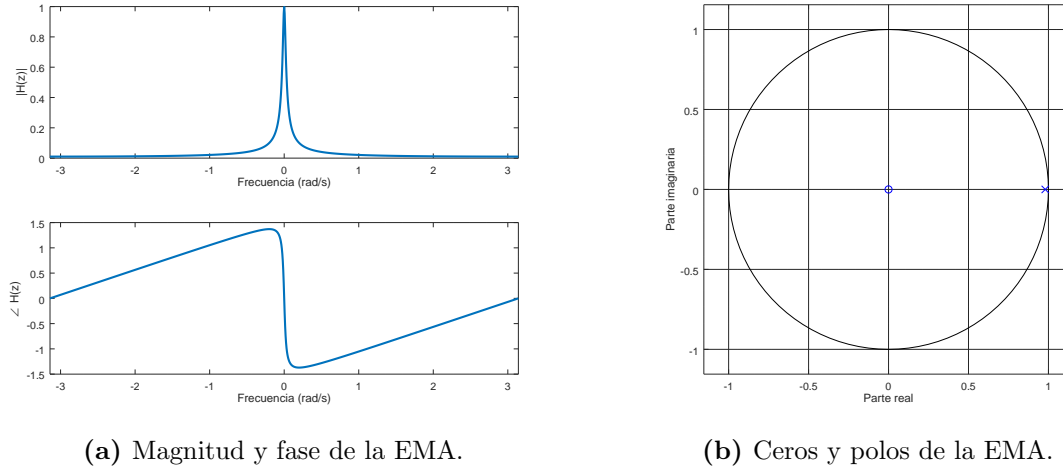
Esta ecuación se interpreta como sigue: se “deshace” el promedio del instante  $n-1$  multiplicando la salida  $y_{M-1}[n-1]$  por  $M-1$ , lo que devuelve la suma  $x[n-M] + \dots + x[n-1]$ , luego, se suma la nueva entrada  $x[n]$  y se vuelve a dividir el resultado por  $M$  para obtener el promedio final. Haciendo el cambio de variable  $\lambda = (M-1)/M$  (un parámetro denominado *factor de olvido*) se puede definir la versión recursiva de la media móvil como sigue:

$$y_M[n] = \lambda y_{M-1}[n-1] + (1-\lambda)x[n], \quad (2.17)$$

lo que se conoce como *media móvil exponencial* (EMA, por sus siglas en inglés) o *leaky integrator*. Cuando  $M \gg 0$  entonces  $\lambda \rightarrow 1$ , es decir, que entre más vecinos se consideren más cercano a la unidad será el factor de olvido. Por emplear valores de las salidas anteriores, la EMA es clasificada como un filtro IIR y, para que sea estable, debe cumplirse que  $|\lambda| < 1$ . Como puede observarse en la figura 2.12a, su respuesta en frecuencia es más suave que la de la media móvil. La ecuación que la describe es [24]:

$$H(z) = \frac{1-\lambda}{1-\lambda z^{-1}},$$





**Figura 2.12:** (2.12a) Diagrama de Bode y (2.12b) diagrama de ceros y polos de la EMA con  $\lambda = 0.97$ .

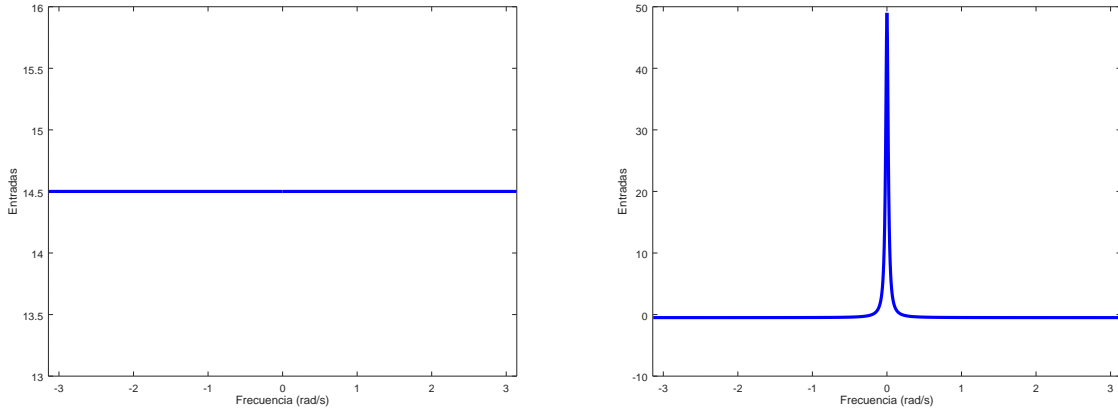
donde se observa que la función tiene un cero en  $z = 0$  y un polo en  $z = \lambda$  (visible en la figura 2.12b). El único polo está dentro del círculo unitario, lo que hace al sistema estable, y permite calcular las condiciones iniciales.

### Fase de los filtros estudiados

Hasta ahora se han analizado las magnitudes espectrales de la media móvil y la EMA, pero no sus fases:  $\angle H(e^{j\omega})$ . Estas también ofrecen información útil, como el *retardo de grupo* sufrido por cada frecuencia filtrada. En la figura 2.11a, se aprecia que la fase de la media móvil es lineal (las discontinuidades se deben a que se aplicó un módulo en  $\pi$ ), lo que indica que las frecuencias sufren un retardo constante de 14.5 entradas, lo que se comprueba en la figura 2.13a. Este valor implica que, con una frecuencia de muestreo de  $f_s = 86$  Hz (misma empleada en la sección 1.2.3), el retardo de grupo sería de 0.17 segundos. En el caso de la EMA, la figura 2.12a muestra que las frecuencias por encima de  $f_c$  no sufren un retardo de grupo, pero las que están por debajo sí —más exactamente producen un retardo de 49 entradas, o sea, 0.6 segundos—, lo que se comprueba en la figura 2.13b. Este retardo es un precio a pagar por usar la versión recursiva, que ahorra tiempo de procesamiento y memoria. La sección 3 muestra que, a pesar de todo, el retardo de grupo no representa una distorsión determinante como para afectar el desempeño del cálculo del umbral de tono [41, 40].

### 2.3.1 Costo computacional

En esta sección se analiza el orden de duración de varios tipos de filtrado, entre ellos el filtrado por convolución, el filtrado en el dominio frecuencial, el TS2Means y la EMA. El orden de duración es calculado usando la notación *O grande*, un método empleado en área



(a) Retardo de grupo de la media móvil con  $M = 30$ . (b) Retardo de grupo de la media móvil con  $\lambda = 0.97$ .

**Figura 2.13:** Retardos de grupo para la media móvil e EMA.

de las ciencias de la computación e informática para definir el tope superior del número de operaciones a realizar, considerando el peor caso en el que debe incurrir cada algoritmo [42, 43].

**Filtrado por convolución.** Por cada uno de los  $G$  valores de la señal de entrada deben promediarse  $M - 1$  vecinos izquierdos, lo que toma  $f(g) = Mg$  operaciones, haciendo que el orden de duración sea  $O(GM)$ . En el peor de los casos, sería necesario leer tantos vecinos como muestras tenga la señal de entrada ( $M = g$ ), lo que tomaría  $f(g) = g^2$  operaciones, equivalente a un orden de duración cuadrático de  $O(G^2)$ .

**Filtrado en el dominio frecuencial.** La magnitud espectral del filtro  $|H(e^{j\omega})|$  y la magnitud espectral de la señal de entrada  $|X(e^{j\omega})|$ , calculadas por FFT, requieren realizar  $f_1(g) = 2g \log g$  operaciones. Luego ambas señales son multiplicadas punto a punto, lo que toma una cantidad lineal:  $f_2(g) = g/2$  (la mitad del tamaño pues el espectro es simétrico en el rango  $[-\pi, \pi]$ ). También debe convertirse de vuelta la señal del dominio frecuencial al dominio temporal usando la *transformada inversa rápida de Fourier* (IFFT, por sus siglas en inglés), que nuevamente toma un número logarítmico:  $f_3(g) = 2g \log g$ . La suma de las funciones es equivalente a  $f(g) = 4g \log g + g/2$ . Al omitir las constantes de proporcionalidad y conservar la función de crecimiento más rápido de la suma (siguiendo las reglas de la notación *O grande*) se obtiene un orden de duración logarítmico de  $O(G \log G)$ .

**TS2Means.** Para cada entrada de la señal a procesar, la formula de optimización 2.12 realiza un barrido de valores  $N_{\min} \leq n \leq N_{\max}$ , lo que en el peor de los casos toma  $f_1(g) = g$  iteraciones, es decir, tantos vecinos como muestras tenga la señal de entrada. En cada iteración se realiza una resta de dos centroides y una corrida del algoritmo  $k$ -nn. El número de operaciones de este último puede aproximarse como  $f_2(g) = g$ , según el comportamiento lineal observado en la figura 2.15 [38].

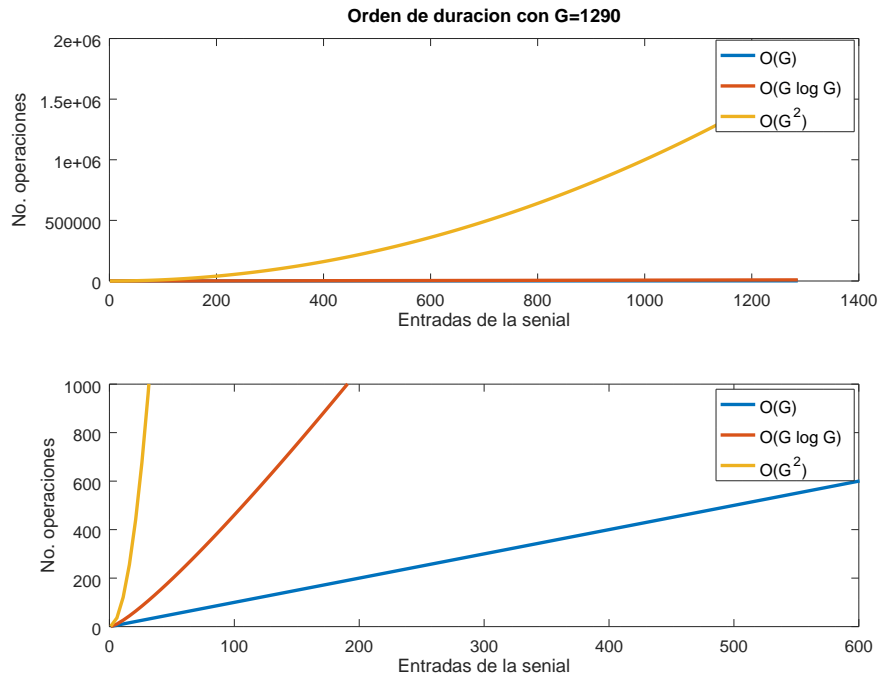
Para calcular cada centroide mediante la fórmula 2.11, el peor de los casos requiere  $2g + g$  sumas ( $2g$  en el numerador y  $g$  en el denominador) y una división, es decir,  $f_3(g) = 2g + g + 1$ . En total, el número operaciones que toma el TS2Means es de  $f(g) = f_1(g)(f_2(g) + f_3(g))$ , es decir,  $f(g) = g(g + (2g + g + 1)) = 4g^2 + g$ . Quitando las constantes de proporcionalidad y considerando el crecimiento dominante, esto da como resultado un orden de duración cuadrático:  $O(G^2)$ .

**EMA.** Por cada uno de los  $G$  valores de la señal de entrada, la ecuación de diferencias 2.17 suma el valor actual de la entrada con el valor anterior de la salida. Las multiplicaciones por las constantes  $\lambda$  y  $1 - \lambda$  añaden dos operaciones extra. El nuevo valor recursivo se almacena en memoria durante cada iteración por lo que no es necesario eliminar la recursión aplicando el criterio de la fórmula general. En resumen, la EMA toma  $f(g) = 3g$  operaciones, lo que, ignorando las constantes de proporcionalidad, permite obtener un orden de duración lineal:  $O(G)$ .

La figura 2.14 realiza un barrido de valores de  $g$  en el rango  $[0, 1290]$  entradas, con el fin de visualizar los órdenes de duración calculados. El tope máximo es equivalente a 15 s de audio, muestreado con una frecuencia de muestreo de  $f_s = 86$  Hz (usada en la sección 1.2.4). Se observa que el costo computacional más grande corresponde al filtrado por convolución y al TS2Means, el intermedio al filtrado en el dominio frecuencial, y el más bajo a la EMA por su ecuación de diferencias. Aparte de usar la EMA, se podrían haber considerado otros filtros IIR más complejos como los de Butterworth o Chebyshev, que hubieran permitido obtener un retardo de grupo menor y una banda de transición más corta, pero aún así, el retardo máximo determinado en la sección 2.3 es admisible y la complejidad computacional es baja, lo que hace a EMA una opción factible de usar.

## 2.4 Medidas de similitud

Como se explicó en la sección 1.2.2 la distancia euclidiana modificada es usada para encontrar correspondencias entre el contorno del audio analizado y las plantillas de los APS. Esta medida de similitud no es única, por ello se estudian otras alternativas con mejor reconocimiento. A continuación, la sección 2.4.1 repasa el concepto de distancia o métrica, la sección 2.4.2 analiza la definición de la distancia euclidiana, la sección 2.4.3 contempla la distancia coseno, la sección 2.4.4 brinda un repaso sobre algunos conceptos de probabilidad necesarios para entender la sección que sigue, la sección 2.4.5 estudia la distancia de Mahalanobis, y la sección 2.4.6 hace una comparación entre las distancias estudiadas para determinar la mejor, detectando patrones con varianzas distintas. Al final, la sección 2.4.7 explica cómo se calcula la pseudoinversa de una matriz y la sección 2.4.8 explica el *análisis de componentes principales* (PCA, por sus siglas en inglés), ambos son conceptos que ayudan a entender mejor la distancia de Mahalanobis.



**Figura 2.14:** Comparación de los ordenes de duración lineales, logarítmicos y cuadráticos. En la segunda figura se presenta un acercamiento para apreciar mejor los ordenes de duración más bajos.

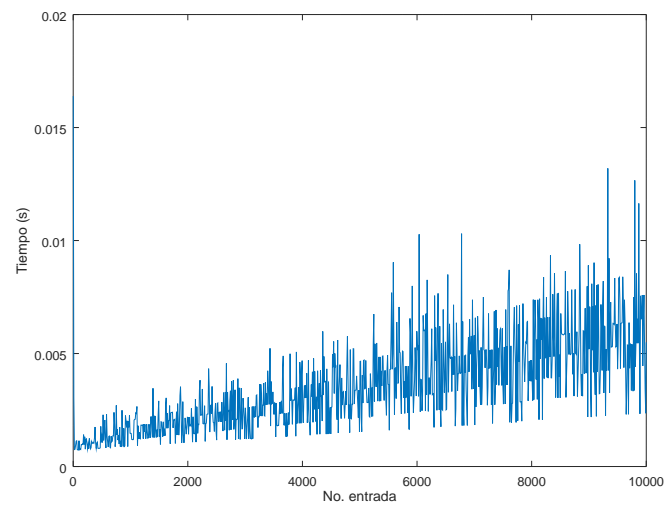
### 2.4.1 Definición de distancia

La topología, o sea, la rama de la matemática encargada de estudiar las propiedades de los cuerpos geométricos inalterados por deformaciones continuas (donde las transformaciones no agregan puntos que no estaban anteriormente o quitan los existentes), define que, para un conjunto  $E$  no vacío, una *distancia* o *métrica* es una función  $d : E \times E \rightarrow \mathbb{R}^+$  que cumple las siguientes condiciones [44, 45]:

- i **No negatividad:**  $\forall x, y \in E [0 \leq d(x, y)]$ .
- ii **Propiedad idéntica:**  $d(x, y) = 0 \iff x = y$ .
- iii **Simetría:**  $\forall x, y \in E [d(x, y) = d(y, x)]$ .
- iv **Desigualdad triangular:**  $\forall x, y, z \in E [d(x, z) \leq d(x, y) + d(y, z)]$ .

Si la distancia no cumple con la propiedad (ii) entonces se le denomina *pseudodistancia* o *pseudométrica*, y al par  $(E, d)$  se le denomina *espacio topológico*. Las propiedades i–iv se comprenden por analogía con los espacios  $\mathbb{R}^2$  y  $\mathbb{R}^3$ . La desigualdad triangular es necesaria para respetar que la distancia más corta entre todo par de puntos es la línea recta (en un espacio métrico plano).<sup>6</sup>

<sup>6</sup>En ciertos problemas de la topología los espacios métricos son más bien curvados y la distancia más corta entre dos puntos son unas líneas curvas llamadas geodésicas.



**Figura 2.15:** Corrida de la función *w2means* del algoritmo TS2Means para realizar la agrupación por *k*-nn. Se observa que el orden de duración es lineal.

### 2.4.2 Distancia euclidiana

La distancia euclidiana se define como la norma  $L_2$  de la diferencia entre los puntos  $x, y$  de un espacio de Hilbert:

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{n=0}^{N-1} |x_n - y_n|^2}, \quad (2.18)$$

donde un espacio de Hilbert  $(\mathbb{R}^N, d)$  es un espacio vectorial normado, y por tanto métrico (porque la norma induce a la métrica), donde la norma al cuadrado es igual al producto interno, es decir,  $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$  [46]. Este producto interno no siempre es igual al producto punto 2.18, sino que puede ser cualquier función que tome dos elementos del espacio vectorial y los mapee a un número complejo, cumpliendo cuatro condiciones llamadas: linealidad, simetría y ser positiva definida [46].<sup>7</sup>

### 2.4.3 Distancia coseno

El producto punto  $\langle x, y \rangle$  escalado por la norma de  $y$  establece la magnitud de la proyección ortogonal del vector  $x$  sobre el vector  $y$ , es decir [24, 47]:

$$\|\text{proy}_y x\| = \cos(\theta) \|x\| = \frac{\langle x, y \rangle}{\|y\|},$$

donde  $\theta$  es el ángulo entre los vectores  $x$  y  $y$ , como se observa en la figura 2.16. Esto da pie a la definición de la distancia coseno [39]:

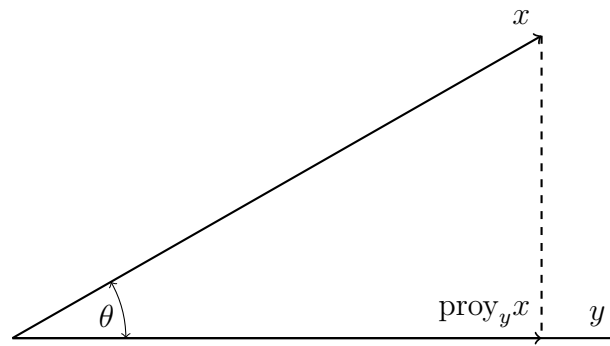
$$d(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\| \|y\|}.$$

Esta distancia tiene la particularidad de que, al normalizar el producto punto de los vectores, se descarta la diferencia entre las magnitudes y se compara únicamente la diferencia entre las direcciones. Nótese que la distancia es unitaria solo en el rango  $\theta \in [0, \pi/2]$ .

### 2.4.4 Vectores aleatorios y matriz de covarianza

Una *variable aleatoria discreta*  $X$  es una función que modela los posibles valores discretos  $x \in \mathbb{R}$  de un experimento (por ejemplo, el lanzamiento de un dado de seis caras) y que utiliza una distribución de probabilidad asociada para asignar la probabilidad de ocurrencia,  $f[X = x] = f_X(x)$ , a cada valor [48]. Para variables discretas esta función se denominada *función de masa de probabilidad* (PMF, por sus siglas en inglés). Ejemplos de distribuciones de probabilidad para variables discretas son la binomial, la Bernoulli, la Poisson, la geométrica, y la uniforme. Todas ellas cumplen dos condiciones:  $f_X(x) \geq 0$

<sup>7</sup>La distancia de Mahalanobis de la sección 2.4.5 define otro tipo de producto interno que no es igual al producto punto.



**Figura 2.16:** Proyección ortogonal del vector  $x$  sobre el vector  $y$ .

y  $\sum_x f_X(x) = 1$ . El *valor esperado* o media de una variable aleatoria discreta, denotado por  $\mu_X$ , se define como sigue:

$$\mu_X = E[X] = \sum_x x f_X(x), \quad (2.19)$$

por su parte, la *varianza*  $\sigma_X^2$  se define como el cuadrado de la desviación esperada respecto de la media, y se calcula como sigue:

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f_X(x). \quad (2.20)$$

Otra medida de utilidad es la *covarianza*  $\sigma_{XY}$  entre dos variables aleatorias  $X, Y$ , que define la variación de una con respecto de la otra ( $\sigma_{XY} = 0$  implica que ambas variables son independientes) y se calcula como sigue:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_{(x,y)} (x - \mu_X)(y - \mu_Y) f_{XY}(x, y), \quad (2.21)$$

donde  $f_{XY}(x, y)$  es la *función de masa de probabilidad conjunta*, que determina la probabilidad de ocurrencia de la combinación de valores ( $X = x, Y = y$ ). Cuando no se cuenta con la PMF de un proceso, sino solo con  $N$  muestras tomadas aleatoriamente, se recurre a la *media de la muestra*  $\bar{x}$  (es decir, el promedio) como aproximación de  $\mu_X$ , y a la *varianza de la muestra*  $s_X^2$  como aproximación de  $\sigma_X^2$ . Para ellas, se utiliza una PMF uniforme en las ecuaciones 2.19, 2.20 y 2.21, es decir,  $f_X(x) = f_{XY}(x, y) = 1/N$ . Otro concepto importante es el de *vector aleatorio*, que consiste en una secuencia de variables aleatorias  $X_0, X_1, \dots, X_{N-1}$ , que cuando  $N \rightarrow \infty$ , recibe el nombre de *proceso aleatorio* [24]. Cuando las variables de un vector aleatorio son independientes entre sí, su PMF es la multiplicación de las PMF de cada variable aleatoria, es decir:

$$f_{X_0 X_1 \dots X_{N-1}}(x_0, x_1, \dots, x_{N-1}) = f_{X_0}(x_0) \cdot f_{X_1}(x_1) \cdots f_{X_{N-1}}(x_{N-1}).$$

Si, además, todas las variables aleatorias siguen la misma distribución de probabilidad, es decir, que son *independientes e idénticamente distribuidas* (i.i.d.), la PMF del proceso aleatorio se simplifica como sigue:

$$f(x_0, x_1, \dots, x_{N-1}) = f(x_0) \cdot f(x_1) \cdots f(x_{N-1}).$$

Dos vectores aleatorios  $X = X_0, X_1, \dots, X_{N-1}$  y  $Y = Y_0, Y_1, \dots, Y_{N-1}$  permiten derivar otro concepto llamado *matriz de covarianza*, que se calcula mediante la expresión  $\Sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)^T]$ , donde  $\mu_X, \mu_Y \in \mathbb{R}^N$  y  $\Sigma_{XY}$  es una matriz  $N \times N$ . De aquí en adelante, se llamará simplemente  $\Sigma$  a la matriz  $\Sigma_{XX}$ .

### 2.4.5 Distancia de Mahalanobis

La distancia de Mahalanobis, también definida en un espacio de Hilbert, es una generalización de la distancia euclidiana, en la que se considera la covarianza de los datos de la siguiente manera [39]:

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}, \quad (2.22)$$

donde  $x, y$  son muestras del mismo vector aleatorio  $X$  y  $\Sigma^{-1}$  es la matriz de covarianza inversa. La distancia de Mahalanobis se expresa en términos del número de desviaciones estándar ( $\sigma$ ) de alejamiento respecto del vector de medias; de hecho, cuando se emplean vectores  $\mathbb{R}^1$  en la ecuación 2.22, se obtiene una fórmula básica en estadística [49]:

$$d(x, y) = \frac{x - y}{\sigma}. \quad (2.23)$$

Como se demuestra en el apéndice B, al utilizar la matriz  $\Sigma$ , se aplica implícitamente un *análisis de componentes principales* (PCA, por sus siglas en inglés) que realiza un “blanqueamiento” de los datos para eliminar su covarianza [50, 51, 52].<sup>8</sup> Además, como se demuestra en el apéndice C, la distancia de Mahalanobis cumple la definición de una métrica, solo si se logra asegurar que  $\Sigma$  es una matriz definida positiva (p.d., por sus siglas en inglés), es decir, que satisface la condición  $\forall x \neq 0 [x^T \Sigma x > 0]$  [46]. Lo que no siempre es el caso, y por eso la distancia de Mahalanobis se denomina una pseudométrica.

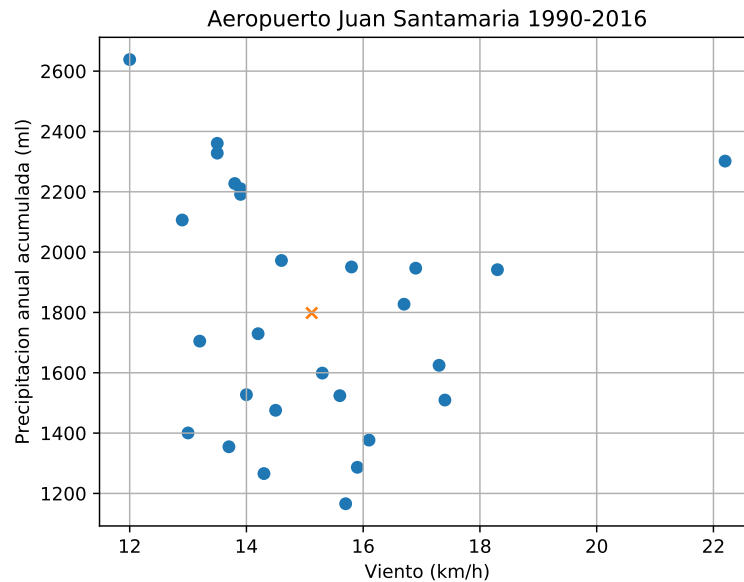
Es común trabajar con una versión aproximada de  $\Sigma$ , llamada la *matriz de covarianza de la muestra*, o  $S$ , que se obtiene al aplicar la ecuación 2.21 sobre un conjunto de  $M$  muestras del vector aleatorio  $X$ . Estas muestras también son llamadas *datos de entrenamiento*. Para trabajar con la versión aproximada, es necesario asegurarse que los datos de entrenamiento sean independientes entre sí y que  $M \geq N$  (la cantidad de datos de entrenamiento debe ser mayor o igual que el número de variables de los vectores), sino,  $S$  no es invertible y debe emplearse el método de la pseudoinversa [53].

### 2.4.6 Evaluación de las distancias analizadas

Las distancias euclidiana y coseno suponen que todas las variables de los vectores a comparar tienen la misma unidad de medida, es decir, que si alguna variable está varios órdenes de magnitud por encima de las otras, insensibiliza la distancia. Para ilustrar la conveniencia de usar la distancia de Mahalanobis respecto de las otras dos, se propone

<sup>8</sup>Otra forma común usada para entender la distancia de Mahalanobis es mediante la distancia a curvas equiprobables de distribuciones gaussianas.





**Figura 2.17:** Datos del viento y de la lluvia acumulada en el Aeropuerto Juan Santamaría en los años 1990 a 2016. El punto medio se señala con una equis.

un problema de clasificación del viento (en kilómetros por hora) y la precipitación acumulada anual (en mililitros) del Aeropuerto Juan Santamaría durante los años 1990 a 2016. Los datos de los 27 años disponibles han sido dibujados en la figura 2.17. El problema consiste en determinar qué tan típicos serían los años,  $x_1 = [15.1 \text{ km/h}, 2100 \text{ ml}]$  y  $x_2 = [5.0 \text{ km/h}, 1798.1 \text{ ml}]$ , respecto de la media  $\bar{x} = [15.1 \pm 2.1 \text{ km/h}, 1798.1 \pm 399.7 \text{ ml}]$ . Se observa que  $x_1$  contiene la media del viento y una cantidad de precipitación mayor respecto de  $\bar{x}$ , y que  $x_2$  contiene la cantidad media de precipitación y una velocidad del viento menor respecto de  $\bar{x}$ .

Antes de calcular cualquier distancia se intuye que, a diferencia de  $x_1$ , el año  $x_2$  es atípico, pues el viento difiere en 10.1 km/h respecto de la media (casi cinco desviaciones estándar, según la ecuación 2.23), mientras que  $x_1$  varía solo en 302 ml respecto de la media (menos de una desviación estándar, según la ecuación 2.23).

Como se observa en la tabla 2.1, solo la distancia de Mahalanobis y coseno pudieron reflejar este razonamiento, la primera pues  $d(x_1, \bar{x}) = 0.9 < 4.5 = d(x_2, \bar{x})$  y la segunda pues  $d(x_1, \bar{x}) = 1.0 \cdot 10^{-6} < 1.6 \cdot 10^{-5} = d(x_2, \bar{x})$ . Sin embargo, la distancia de Mahalanobis es más útil que el resto: la distancia euclidiana determinó que el estado  $x_1$  era el más atípico, lo que no es de sorprender, pues la magnitud de la precipitación está 2 ordenes de magnitud por encima de la magnitud del viento; y la distancia coseno arrojó distancias difíciles de comparar por ser casi nulas. El cálculo de la matriz de covarianza usada en la distancia de Mahalanobis se encuentra en el apéndice A.

**Tabla 2.1:** Comparación de distancias euclidiana, coseno y de Mahalanobis entre un centroide y dos muestras  $x_1$  y  $x_2$  para el ejemplo del viento y la precipitación anual del Aeropuerto Juan Santamaría.

Muestra	Distancia		
	Euclidiana	Coseno	Mahalanobis
$x_1$	302	$1.0 \cdot 10^{-6}$	0.8
$x_2$	10	$1.6 \cdot 10^{-5}$	4.9

## 2.4.7 Pseudoinversa de una matriz

El cálculo de la distancia de Mahalanobis puede verse afectado cuando la cantidad de vectores,  $M$ , usados para determinar la matriz de covarianza de la muestra,  $S$ , es menor que el número de variables del vector aleatorio  $X$ , es decir:  $M < N$ , donde  $M = \rho(S)$  es el rango de  $S$ . En ese caso  $S$  no es invertible, porque al menos uno de sus renglones es linealmente dependiente del resto (y nulo en la forma escalonada), lo que produce un determinante nulo (es decir,  $\det(S) = 0$ ) que indefinición la expresión de la matriz inversa [47, 54]:

$$S^{-1} = \frac{1}{\det(S)} \text{adj}(S),$$

donde  $\text{adj}(S)$  es la matriz adjunta de  $S$ , es decir, la matriz traspuesta de los cofactores  $C_{i,j}$  de  $S$ , los cuales se definen como  $C_{i,j} = (-1)^{i+j} |P_{i,j}|$ , siendo  $P_{i,j}$  el  $i, j$ -ésimo menor de  $S$ , o sea, la matriz  $S$  sin la  $i$ -ésima fila y la  $j$ -ésima columna. Afortunadamente existe una versión aproximada de la matriz inversa llamada la *pseudoinversa*, o  $S^+$ , que se obtiene mediante una técnica llamada *descomposición por valores singulares* (SVD, por sus siglas en inglés) [55]. La SVD permite representar la matriz  $S$  como una multiplicación de otras tres matrices,  $U$ ,  $\Sigma$  y  $V$ , como sigue [56]:<sup>9</sup>

$$S_{N \times N} = U_{N \times M} \Sigma_{M \times M} V_{M \times N}^T, \quad (2.24)$$

donde  $V$  es una matriz cuyas columnas son los autovectores con autovalor no nulo de  $S^T S$ ,  $U$  se obtiene mediante la expresión  $U = S V \Sigma^{-1}$  y  $\Sigma$  es una matriz diagonal cuyos elementos son los valores característicos no nulos de  $S^T S$ . Tanto  $V$  como  $U$  son matrices ortogonales, es decir, que su inversa es igual a su traspuesta. Los valores característicos se definen como las raíces cuadradas de los autovalores de  $S^T S$ , es decir,  $\sigma_i = \sqrt{\lambda_i}$  [57, 55, 56]. Los autovalores  $\lambda_i$  y los autovectores  $v_i$  son los términos que hacen que el sistema  $S v_i = \lambda_i v_i$ , tenga solución. Haciendo el ordenamiento  $\sigma_0 \geq \sigma_1 \dots \sigma_{M-1} > 0$  (válido cuando  $S$  es simétrica, como ocurre con la matriz de covarianza) se puede definir la matriz  $\Sigma$  como sigue:

$$\Sigma = \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_M \end{bmatrix}.$$

<sup>9</sup>A diferencia del resto del documento, esta sección no se refiere a  $\Sigma$  como la matriz de covarianza.

Con estas tres matrices, se puede aproximar la pseudoinversa de  $S$  de la siguiente manera:

$$S^+ = U\Sigma^{-1}V^T,$$

donde  $\Sigma^{-1}$  se puede calcular a partir de [47]:

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1/\sigma_M \end{bmatrix}.$$

La pseudoinversa  $S^+$  contiene los renglones de la inversa que son engendrados por el subespacio de los autovectores con autovalor no nulo, que hacen que la *norma de Frobenius* sea la menor posible [58]. En este sentido, las matrices no singulares (es decir, invertibles) cuyos autovalores son no nulos, puede interpretarse como un caso especial donde  $S^+ = S^{-1}$ . La técnica de la SVD es usada en general para solucionar sistemas de ecuaciones lineales [56].

## 2.4.8 Análisis de componentes principales

El *análisis de componentes principales* (PCA, por sus siglas en inglés), mencionado en la sección 2.4.5, es una técnica que permite reducir la dimensionalidad de un vector conservando las variables con mayor varianza y descartando el resto. El PCA es utilizado en el campo del reconocimiento automático de patrones, para mitigar la *maldición de la dimensionalidad*. Esta última ocurre por dos razones. La primera es cuando hay variables que no aportan pistas para clasificar un patrón, sino que representan ruido, haciendo más lento el aprendizaje porque obligan al clasificador a aprender a ignorar el ruido, y si no hay suficientes muestras de entrenamiento se produce una *parálisis de aprendizaje*. La segunda, es cuando la cantidad de variables a procesar excede la capacidad de generalización del clasificador, y este opta por “memorizar” los patrones (fenómeno conocido como *overfitting*), siendo incapaz de clasificar correctamente patrones nuevos [50, 51, 59].

El PCA afirma que si solo  $M$  de los  $N$  coeficientes de su transformada son conservados, entonces los coeficientes  $a_0, \dots, a_{M-1}$  minimizarán el *error medio cuadrático* (MSE, por sus siglas en inglés) entre  $\varphi(x)$ , el vector original, y su aproximación  $\hat{\varphi}(x)$ . El vector  $\hat{\varphi}(x)$  se define como:

$$\hat{\varphi} = \bar{\varphi} + \sum_{n=1}^M a_n u^{(n)}, \quad (2.25)$$

donde  $u^{(i)}u^{(j)} = \delta_{ij}$  es la base algebraica óptima para representar los vectores originales, los coeficientes  $a_n$  se obtienen mediante el producto punto  $u^{(n)}\phi^{(n)} \in \mathbb{R}$ , y los vectores  $\phi^{(n)}$  son copias de los vectores originales centrados en cero, es decir,  $\phi^{(n)} = \varphi^{(n)} - \bar{\varphi}$ . Cada  $u^{(i)}$  es en realidad un autovector de la matriz de covarianza  $S_{N \times N}$  (matriz de covarianza de la muestra) obtenida al aplicar la ecuación 2.21 sobre los datos de entrenamiento, y, al igual que los autovalores  $\lambda_k$ , resuelven la expresión  $Su^{(n)} = \lambda_n u^{(n)}$ . El valor  $M$  se determina

**Tabla 2.2:** Matriz de confusión de un clasificador binario. A diferencia del caso multi-clase, las filas y columnas no son las categorías sino la pertenencia a la única categoría.

		Predicho	
		Positivo	Negativo
Real	Positivo	VP	FN
	Negativo	FP	VN

ordenando los autovalores de mayor a menor y eligiendo el valor característico  $\sigma_i = \sqrt{\lambda_i}$  que minimiza el MSE, es decir,  $\sigma_0 \geq \sigma_1 \dots \geq \sigma_M \dots \geq 0$  [50]. Cuando se utiliza la matriz de correlación en lugar de la matriz de covarianza, el análisis PCA es conocido como la *transformada de Karhunen-Loève* (KLT, por sus siglas en inglés) [46].

## 2.5 Tasas de detección

La sección 4.1 menciona cuatro métricas denominadas verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Estas métricas se obtienen al evaluar la *matriz de confusión* de un clasificador binario que determina si un individuo pertenece o no a la clase deseada. Una descripción de la matriz de confusión se muestra en la tabla 2.2.

Dado que la señal de audio se separa en ventanas de tiempo disjuntas, los positivos se refieren a las ventanas donde el algoritmo detectó un APS y los negativos a las ventanas donde no se encontró ninguna correspondencia. Con base en estas métricas se pueden calcular otras más útiles, llamadas *tasas de detección*, entre las que destacan: *acierto*, *precisión*, *especificidad*, *sensibilidad*, *medida F* y el *coeficiente de correlación de Matthew* (MCC, por sus siglas en inglés) [10], que se definen a continuación [10, 5, 60].

**Acierto:** Proporción del número de ventanas correctamente identificadas (positivas y negativas) entre el total de ventanas. No es recomendada cuando los datos no son simétricos, es decir, cuando la proporción entre positivos y negativos está muy por debajo de la unidad  $((VP + FP)/(VN + FN) \ll 1)$ . Se calcula así:

$$a = \frac{VP + VN}{VP + VN + FP + FN}.$$

**Precisión.** Probabilidad de que una alerta emitida sea real, es decir, que al acatarla no se esté poniendo en riesgo la seguridad del usuario. Se obtiene calculando la proporción entre número de ventanas APS correctamente identificadas como positivas y el total de ventanas APS positivas detectadas (incluyendo las detecciones erróneas). Se calcula así:

$$p = \frac{VP}{VP + FP}.$$

**Exhaustividad.** Probabilidad de que una alerta de cruce sea emitida cuando el APS está sonando, o sea, la *tasa de detección* del sistema para identificar los momentos cuando el cruce peatonal puede ser transitado. Se obtiene calculando la proporción entre el número de ventanas APS correctamente identificadas como positivas y el total de ventanas APS positivas. Se calcula así:

$$r = \frac{VP}{VP + FN}.$$

**Especificidad.** Probabilidad de que la alerta emitida no sea falsa. Útil para determinar la *robustez contra el ruido* del sistema al procesar ruido del ambiente. Se obtiene calculando la proporción entre el número de ventanas APS correctamente identificadas como negativas y el total de ventanas APS negativas (también conocida como la *tasa de negativos verdaderos*). Se calcula así:

$$e = \frac{VN}{VN + FP}.$$

**Medida F.** Equilibrio entre la seguridad y la eficacia del sistema, y sustituto del acierto cuando los datos no son simétricos. Se define así:  $F = 2pr/(p + r)$ , lo que también es la media armónica entre  $p$  y  $r$ :

$$F = \frac{2}{\frac{1}{r} + \frac{1}{p}}.$$

**Coefficiente de correlación de Matthew (MCC).** Medida que toma en cuenta tanto las detecciones positivas como las negativas. Al igual que la medida F, puede emplearse aún cuando los datos no son simétricos. Al definir la variable  $d = (VP + FP)(VP + FN)(VN + FP)(VN + FN)$ , el MCC se puede expresar como sigue:

$$c = \begin{cases} 0 & d = 0 \\ \frac{VP \cdot VN - FP \cdot FN}{\sqrt{d}} & \text{en el resto,} \end{cases}$$

donde  $c = 1$  significa que el reconocimiento fue el ideal ( $FP = FN = 0$ ),  $c = 0$  que el clasificador tiene un desempeño aleatorio, y  $c = -1$  que el reconocimiento fue el peor posible ( $VP = VN = 0$ ) [61, 62].

En este estudio se prefiere la medida F como sustituto del acierto porque, como explica más adelante la sección 4.1, la secuencia principal de los APS —entre el primer y último pico de actividad— constituye los positivos del sistema y estos positivos abarcan la mayor parte de las grabaciones. Los momentos de ruido anteriores o posteriores a esa secuencia principal son menos extensos, haciendo que los datos sean asimétricos.

Al constituir un sistema de misión crítica, las tasas de precisión y especificidad son importantes para el algoritmo RASP, pues toman en cuenta la cantidad de falsos positivos que podrían provocar accidentes de tránsito al generar alertas de cruce falsas. Sin embargo, considerando que el algoritmo de reconocimiento podría emplearse también en otras áreas de menor impacto, donde se requiere evaluar todas las métricas (VP, FP, VN y FN), se usarán los MCC como métrica para realizar las optimizaciones del capítulo 4.

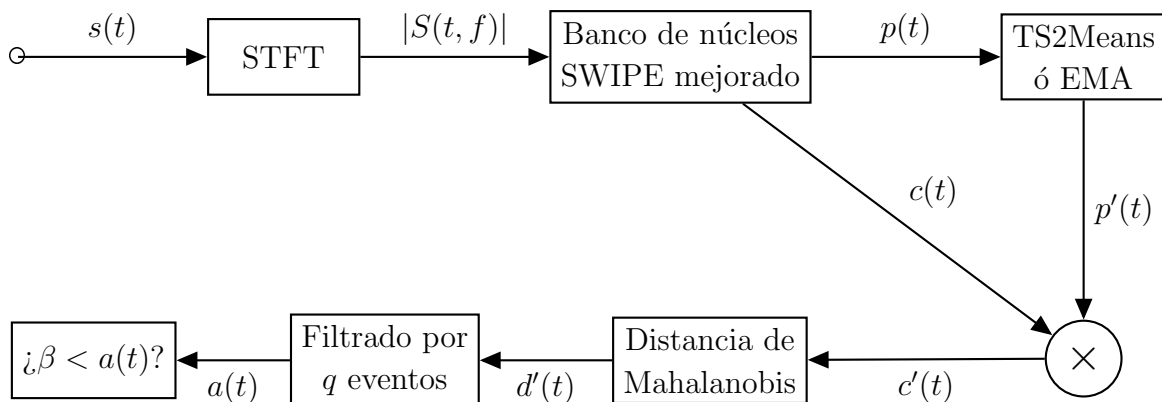
## 2.6 Redes neuronales convolucionales para procesar sonido

Es posible procesar audio usando diseños CNN similares a los publicados por Lecun *et al.* y Cirstea *et al.* [63, 64] para el reconocimiento de caracteres escritos a mano. Con ese fin, otros autores han optado por procesar una representación gráfica del sonido, por ejemplo: espectrogramas producidos por la STFT (como los de las figuras 1.4, 1.5 y 1.6) y escalogramas producidos por la transformada de ondeletas (*wavelets*) [65, 66, 67]. A pesar de los buenos resultados reportados por otros estudios al usar estas redes neuronales convolucionales (y otras adaptadas especialmente para procesar el sonido sin una representación intermedia), en este trabajo no se optó por usarlas debido a seis razones. Primero, las CNN son usadas para procesar datos crudos y realizar una extracción de características automática, lo que no tiene sentido repetir en este trabajo, pues los métodos usados ya realizan el reconocimiento de la altura musical del sonido empleando técnicas robustas del estado de la cuestión. Segundo, algunas de las mejoras aquí propuestas pretenden ser implementadas en la aplicación RASP para dispositivos móviles, la cual no consideró en su diseño la incorporación de librerías de redes neuronales convolucionales. Tercero, para cubrir la carga computacional requerida por las CNN podría ser necesario emplear *hardware* especializado, que no poseen todos los teléfonos inteligentes. Cuarto, las CNN requieren un proceso continuo de ajuste y evaluación de los hiperparámetros (tasa de aprendizaje, función del error, tamaño del lote y número de iteraciones) que demanda una cantidad considerable de trabajo, a veces equivalente a entender el problema y resolverlo con los métodos tradicionales. Quinto, para entrenar una CNN es necesario recolectar suficientes muestras de entrenamiento, al menos tantas como variables tenga el patrón a reconocer, y en el caso del APS cucú estas son insuficientes, como se estudiará en la sección 3.2.1. Sería posible elaborar muestras artificiales de sonidos APS que permitan completar la cantidad necesaria para una CNN, sin embargo, no se conoce la distribución de probabilidad del ruido de fondo de las grabaciones reales, que provienen de muchas fuentes distintas de sonido. Y aunque se conociera, al entrenar las CNN con modelos sintéticos de generación de sonido se corre el riesgo de que la red desarrolle un sesgo hacia el modelado y pierda capacidad de generalización. Y sexto, una vez entrenadas las CNN, no se tiene conocimiento de ningún método que permita extraer el conocimiento almacenado en los pesos de las capas no convolucionales, lo que impediría migrar la solución a un enfoque que no use aprendizaje profundo [59].

# Capítulo 3

## Reconocimiento automático de señales peatonales accesibles usando un enfoque adaptativo

Como se determinó en la sección 1.2 y el capítulo 2, el algoritmo RASP puede ser mejorado en al menos tres aspectos: los núcleos empleados deberían considerar más armónicas para hacer más claro el contorno musical, se debería usar la distancia de Mahalanobis para admitir plantillas con segmentos nulos, y se debería realizar un ajuste automático del umbral de tono empleando TS2Means y la EMA. Para el primer aspecto, la sección 3.1 propone usar un núcleo de tres armónicas (el máximo común divisor del número de armónicas de los APS) y un decaimiento proporcional a  $1/k^2$  de las frecuencias del espectro; para el segundo, la sección 3.2 explica cómo se puede generar la matriz de covarianza real y sintética para los tres tipos de APS a partir de los contornos musicales de las grabaciones recopiladas; y para el último, la sección 3.3 explica mediante experimentos preliminares, cómo la aplicación del TS2Means y la EMA podrían ayudar a mejorar la efectividad del algoritmo RASP. Una versión del diagrama 1.13 con las mejoras propuestas se muestra en la figura 3.1.



**Figura 3.1:** Diagrama sobre el funcionamiento del algoritmo RASP mejorado.

### 3.1 Rediseño del núcleo

El análisis realizado en la sección 1.2.1 determinó que las tres primeras armónicas de los sonidos APS están siempre presentes, por lo que se propone usar un solo diseño de núcleo, facilitando el mantenimiento del código de la solución y la ejecución de las pruebas. Se define el núcleo de reconocimiento musical como sigue:

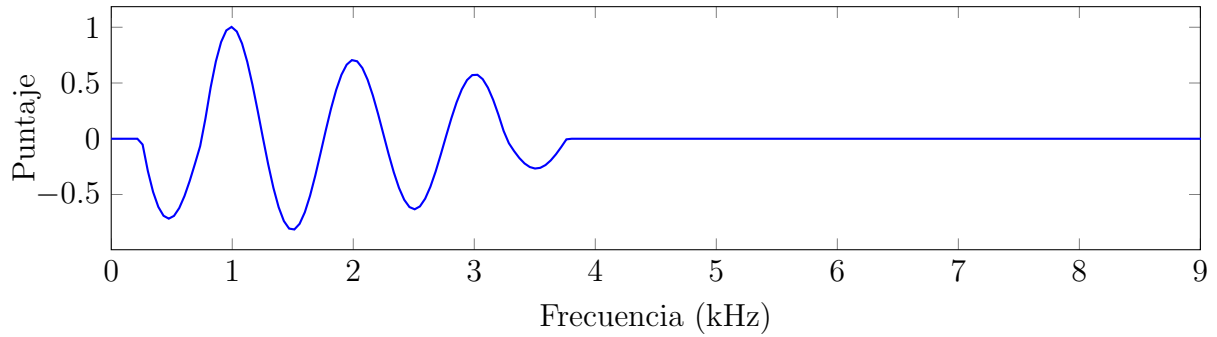
$$k(f) = \psi(f) (f/f_0)^{-1/2} \cos(2\pi f/f_0),$$

donde  $\psi(f)$  corresponde a una ponderación que asegura que el primer y último lóbulo negativo tengan un escalamiento de 0.5 y que solo las armónicas primas sean conservadas. Para un núcleo de  $A$  armónicas, la ponderación se define como sigue:

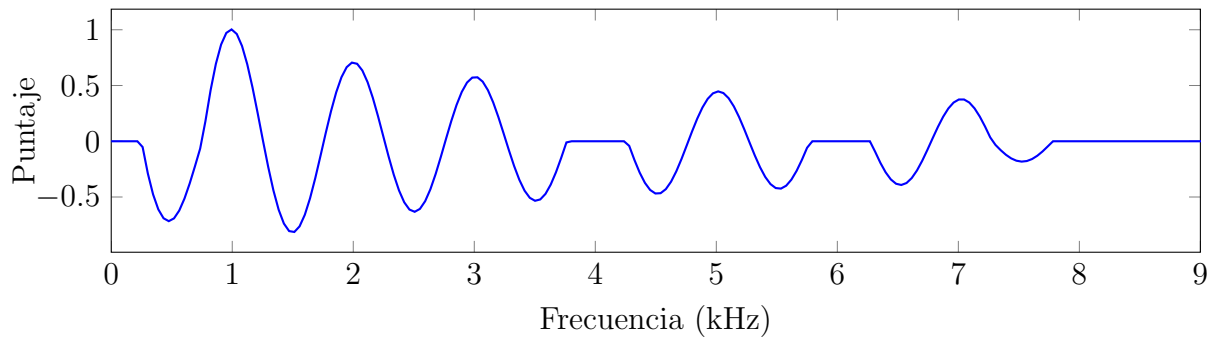
$$\psi(f) = \begin{cases} 1, & (0.75 < f/f_0 < A + 0.25) \wedge (r(f/f_0) \text{ es primo} \vee 0.25 \leq |r(f/f_0) - f/f_0|) \\ 0.5, & 0.25 < f/f_0 < 0.75 \vee A + 0.25 < f/f_0 < A + 0.75 \\ 0, & \text{en el resto,} \end{cases}$$

donde  $f_0$  es la frecuencia fundamental a procesar y  $r(\cdot)$  es la función de redondeo a la unidad más próxima. Este diseño del núcleo de reconocimiento propone conservar solo las armónicas primas para imitar la habilidad de SWIPEP de reducir el error por subarmónicos, es decir, evitar que una de las subarmónicas de la frecuencia fundamental sea confundida con la altura musical. Por otro lado, para evitar el caso de la confusión de la altura musical con alguna armónica, se propone usar el factor  $(f/f_0)^{-0.5} = 1/k^2$  (donde  $k$  es el índice de la  $k$ -ésima armónica) para aplicar un decaimiento que conserve el valor original en  $f = f_0$  y que disminuya los puntajes del núcleo conforme  $f \rightarrow \infty$ . Este decaimiento es el utilizado por SWIPE, y aunque la teoría de *series de Fourier* afirma que la cota superior del decaimiento de los armónicos de una señal discontinua es lineal, o sea,  $1/k$  (y que este decaimiento es el máximo posible) [26], como la ecuación 2.10 trabaja con una aproximación de la sonoridad igual a la raíz cuadrada de la magnitud espectral, entonces el decaimiento debe adaptarse de la misma manera:  $\sqrt{1/k} = \sqrt{(f/f_0)^{-1}} = (f/f_0)^{-0.5}$ . Un ejemplo del diseño de núcleo propuesto se muestra en la figura 3.2, para el caso de 3 armónicas, y en la figura 3.3, para el caso de 7 armónicas. Respecto del banco de filtros, se propone también usar un diseño unificado para todos los APS, es decir, que contenga la unión de sus rangos de frecuencias:  $\{900, 1100\} \cup \{2000, 2100, \dots, 3000\} \cup \{900, 1000, \dots, 1800\}$  Hz, como se observa en la figura 3.4, lo que evita la dependencia excesiva del umbral de tono  $\alpha$ , pues a mayor cantidad de frecuencias fundamentales candidatas, mayor posibilidad de obtener un puntaje de tono alto.

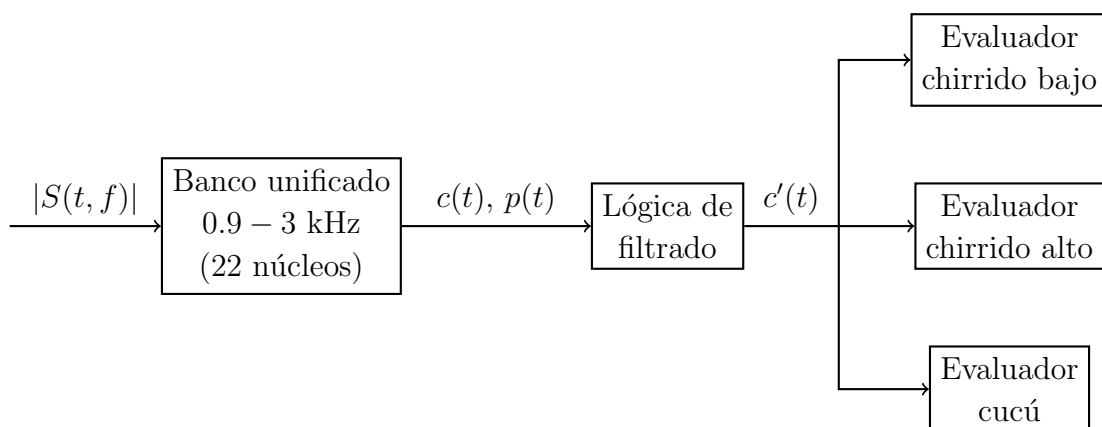




**Figura 3.2:** Núcleo de tres armónicas propuesto para  $f_0 = 1$  kHz.



**Figura 3.3:** Núcleo de siete armónicas propuesto para  $f_0 = 1$  kHz.



**Figura 3.4:** Distribución de los 22 filtros del banco de filtros propuesto. Como se observa, la misma señal de contorno,  $c(t)$ , misma señal de puntajes,  $p(t)$ , y misma lógica de filtrado es aprovechada por todos los evaluadores, reduciendo la complejidad computacional.

## 3.2 Distancia de Mahalanobis para contornos musicales

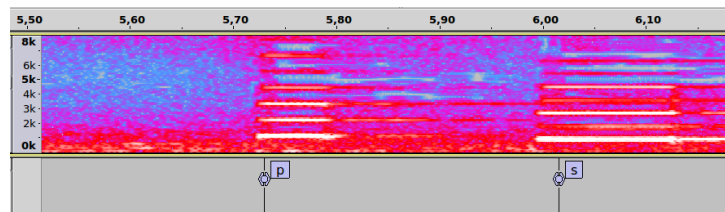
A diferencia de Ruiz *et al.*, que usaron una distancia euclidiana modificada que solo admitía contornos continuos, se propone usar una distancia de Mahalanobis que admita contornos musicales discontinuos, como los del sonido cucú. A continuación, la sección 3.2.1 explica la forma propuesta para recortar las modulaciones de frecuencia de los APS presentes en las grabaciones, la sección 3.2.2 describe el método de extracción del contorno musical utilizado para analizar los recortes, la sección 3.2.3 comprueba que los contornos musicales siguen una distribución normal, la sección 3.2.4 explica la forma empleada para entrenar las matrices de covarianza APS a partir de los contornos obtenidos, la sección 3.2.5 expone el método propuesto para obtener versiones sintéticas de las matrices de covarianza APS —permitiendo prescindir de las etapas de recopilación, recorte, preprocesamiento y entrenamiento al analizar otros tipos de sonido—, y finalmente, la sección 3.2.6 presenta un tratamiento posterior de la salida de la distancia de Mahalanobis para normalizarla al rango unitario, por cuestiones de conveniencia en la etapa de emisión de alertas.

### 3.2.1 Recorte de las grabaciones de interés

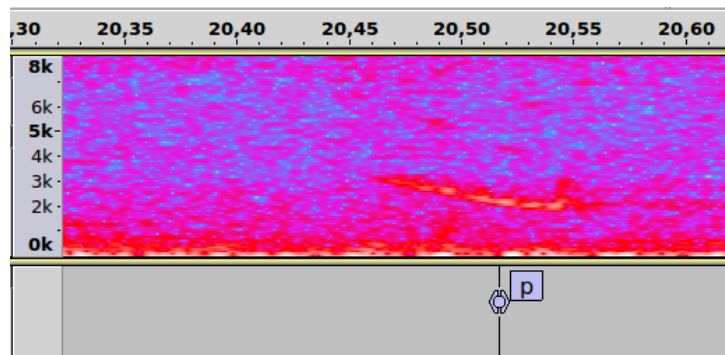
Al igual que en el estudio de Ruiz *et al.*, se cuenta con un conjunto de entrenamiento de 79 grabaciones tomadas de la GAM para los tres tipos de APS admitidos. La duración de cada fonograma es variable, pero todos contienen una secuencia de actividad completa de la señal, es decir, una secuencia principal que indica al usuario la posibilidad de cruce y una secuencia de finalización que le advierte que el tiempo se está agotando. Las modulaciones de frecuencia en la secuencia principal tienen una separación de  $T$  segundos, mientras que en la secuencia de finalización se encuentran a  $t \approx T/2$  entre sí. La secuencia final puede o no estar presente, dependiendo de la configuración de cada dispositivo, y es común que los sonidos cucú no la tengan. Las 79 grabaciones fueron anotadas manualmente por Ruiz *et al.* para indicar los comienzos de las modulaciones de frecuencia (*onsets*, en inglés) y facilitar la etapa de evaluación. En esta sección se propone usar esas anotaciones con el fin de dividir las grabaciones en subseñales de  $T$  segundos, correspondiente a un periodo de cada APS, como se explica en la sección 1.2.2. Durante la etapa de recorte, un problema detectado fue que las anotaciones están desalineadas unos milisegundos respecto al comienzo, lo que afecta la calidad de los recortes realizados, por lo que se procedió a corregirlas manualmente. Un ejemplo de esta desalineación se observa en la figura 3.5. Después de corregir las anotaciones y realizar los recortes, se obtuvo el cuerpo de entrenamiento descrito en la tabla 3.1.

**Tabla 3.1:** Cantidad de muestras de entrenamiento usadas por tipo de APS.

APS	N. <sup>o</sup> grabaciones	N. <sup>o</sup> contornos
Cucú	29	153
Ch. Alto	36	723
Ch. Bajo	14	66
Total	79	942

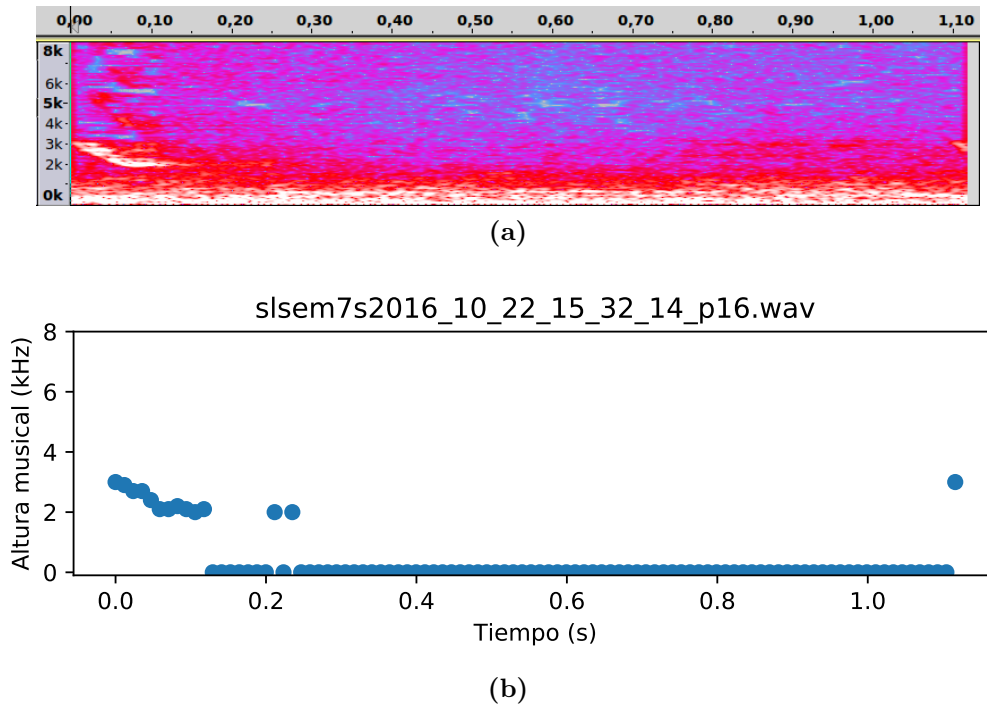


(a)



(b)

**Figura 3.5:** Ejemplos de desalineación de las anotaciones originales. (3.5a) Retraso leve de la segunda anotación del contorno cucú y (3.5b) un retraso más importante en la única anotación del contorno del chirrido bajo. El eje horizontal corresponde al tiempo (en segundos) y el eje vertical a la frecuencia (en hercios).



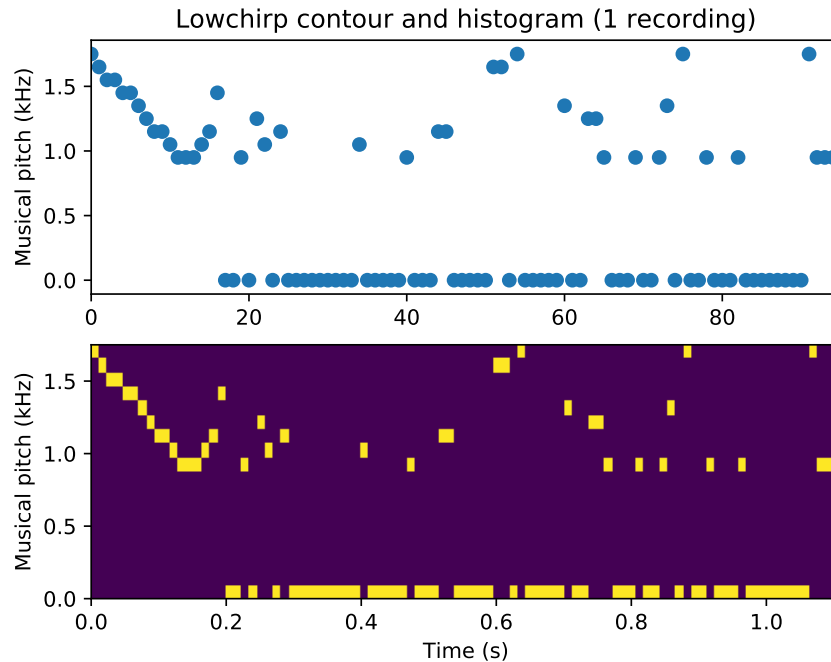
**Figura 3.6:** Ejemplo del cálculo del contorno musical para un recorte de chirrido alto calculado usando el núcleo propuesto de 3 armónicas. (3.6a) Espectrograma del recorte y (3.6b) su contorno musical.

### 3.2.2 Cálculo de los contornos musicales

Una vez obtenidos los recortes de las modulaciones de frecuencia APS, fue necesario calcular los contornos musicales usados para entrenar cada matriz de covarianza. Para ello se empleó el diseño de núcleo propuesto de tres armónicas de la sección 3.1, junto con los bancos de núcleos originales descritos en la tabla 1.1, que son [900, 1100] Hz para el cucú, [2, 3] kHz para el chirrido alto y [950, 1750] Hz para el chirrido bajo, con  $df = 200$  Hz, 100 Hz, 200 Hz, respectivamente. Como se mencionó en la introducción del capítulo, se usó el núcleo de tres armónicas, pues tres es el máximo común divisor de las armónicas de todos los APS. El umbral de tono fue nulo, es decir,  $\alpha = 0.0$ , con el fin de incluir todo el ruido posible en el cálculo de la matriz de covarianza (los puntajes negativos se descartaron). Un ejemplo de los contornos calculados se muestra en la figura 3.6, donde el procesamiento realizado al recorte n.º 16 de una grabación de APS de chirrido alto muestra que se capturó correctamente la modulación de frecuencia de 3 kHz a 2 kHz antes de los 0.15 s, y que se obtuvieron alturas nulas para el resto.

### 3.2.3 Distribución de los contornos musicales

Una condición que deben cumplir los contornos musicales de APS es ser ergódicos, o por lo menos, seguir una distribución normal. Aunque se podría suponer que esto ya lo cumplen los sonidos APS, pues los dispositivos electrónicos que los emiten tienen que apearse

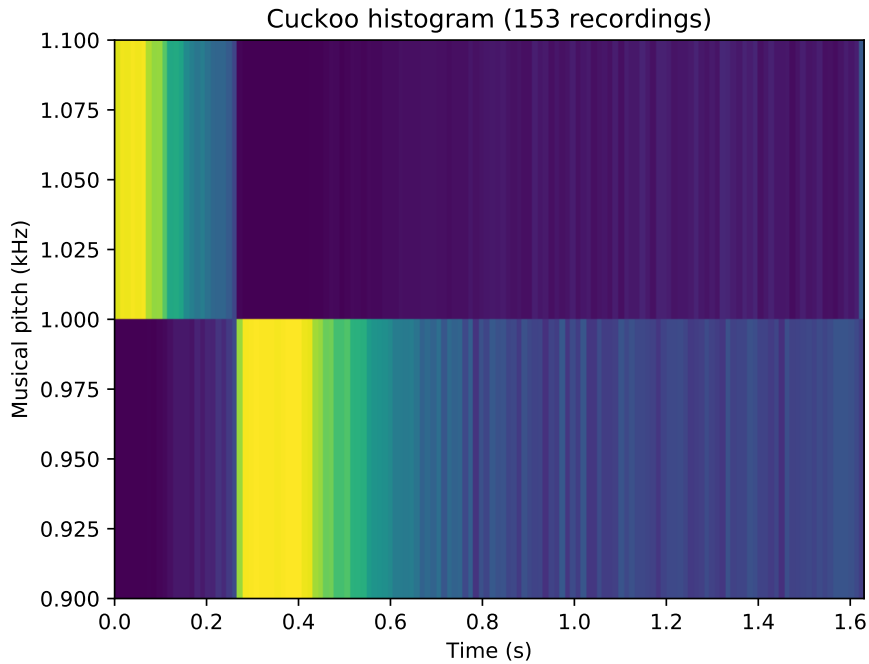


**Figura 3.7:** Histograma de una sola grabación de chirrido bajo. Se observa que, en efecto, el histograma de una sola grabación es igual a su contorno musical.

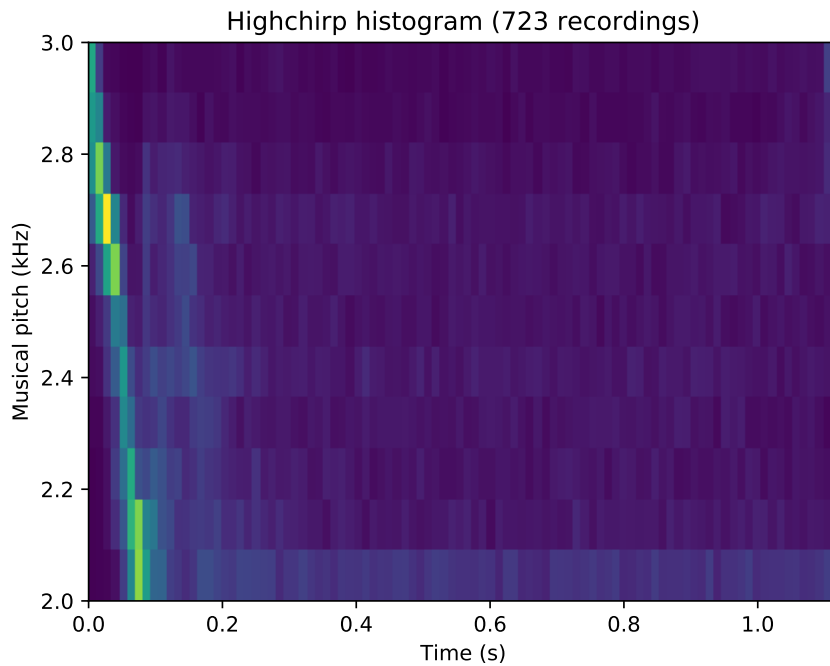
a los estándares de la MUTCD y PROWAG, es conveniente comprobarlo construyendo histogramas de alturas musicales con las grabaciones disponibles. La figura 3.7 ilustra un histograma generado al procesar un solo contorno musical de chirrido bajo, y se observa que el resultado corresponde exactamente con la señal ingresada, lo que indica que las casillas de frecuencia aumentan una unidad a la vez (como es de esperar). Conforme se analizan más grabaciones, los contadores de las casillas aumentan de valor, y si las grabaciones APS siguen la distribución normal deseada, entonces los histogramas finales deben reflejar una forma similar a la de las plantillas definidas en la sección 1.15 (los chirridos tendrían más bien un decaimiento exponencial). Se debe notar que en las regiones ruidosas posteriores a las modulaciones de frecuencia, no interesa que la altura musical sea siempre nula, pues puede haber ruido musical incluido. Las figuras 3.8, 3.9 y 3.10 muestran que, luego del análisis de todas las grabaciones de la tabla 3.1, los histogramas de frecuencias siguen conservando la forma esperadas, indicando que el cálculo de las matrices de covarianza es válido para generalizar la detección de los APS (al menos con las grabaciones disponibles).

### 3.2.4 Cálculo de las matrices de covarianza

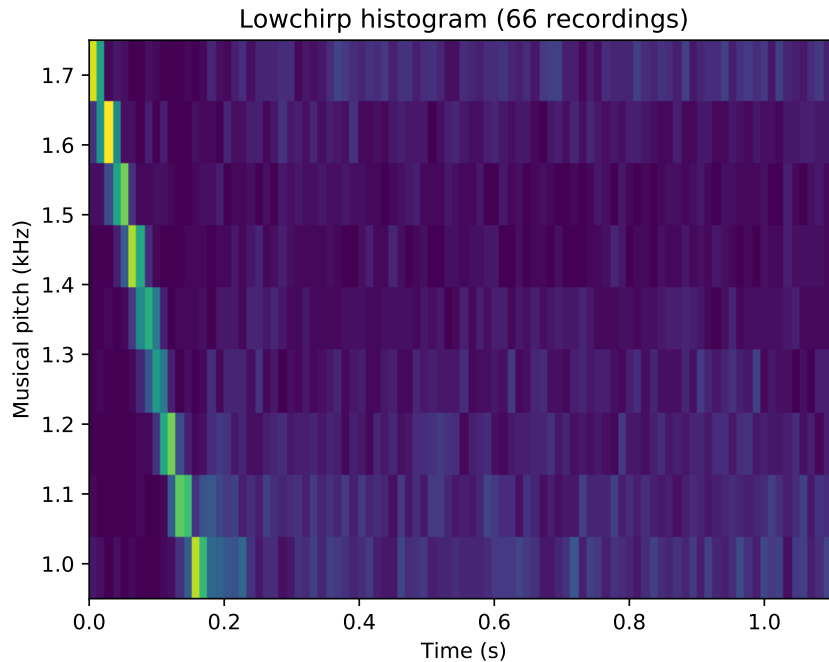
Una vez encontrados los contornos musicales para cada recorte, se calcularon las matrices de covarianza empleando el procedimiento descrito en la sección 2.4.5. Las matrices de covarianza obtenidas se muestran como imágenes en las figuras 3.11, 3.12 y 3.13, donde



**Figura 3.8:** Histograma de todas las grabaciones cucú. Se observa que, en efecto, la distribución es normal y que se aproxima al patrón deseado de un tono de 1100 Hz de 70 ms, seguido de 200 ms de silencio, seguido de un tono de 900 Hz de 130 ms.



**Figura 3.9:** Histograma de todas las grabaciones del chirrido alto. Se observa que, en efecto, la distribución es normal y que se aproxima a la forma de un barrido de frecuencias de 3 – 2 kHz en 100 ms.

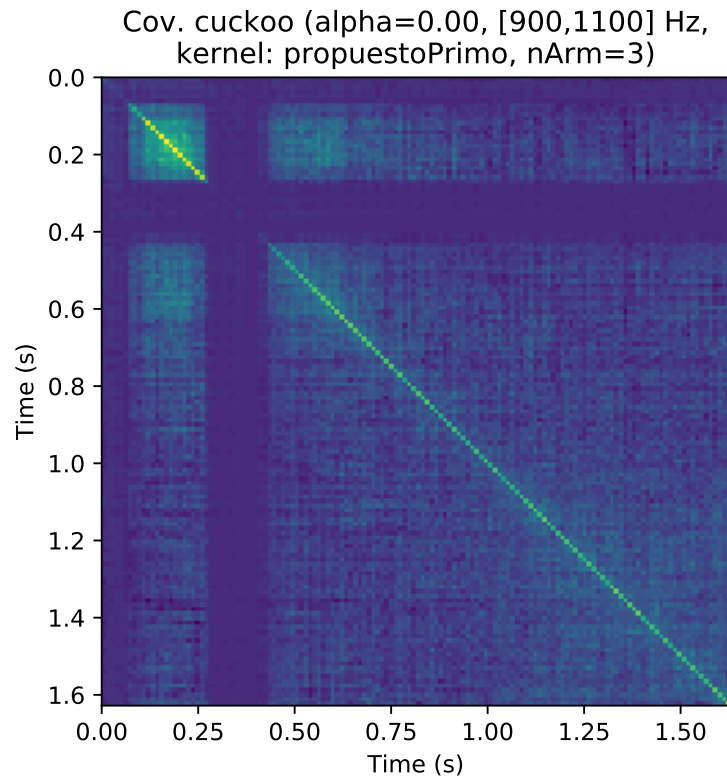


**Figura 3.10:** Histograma de todas las grabaciones de chirrido bajo. Se observa que, en efecto, la distribución es normal y que se aproxima a la forma de un barrido de frecuencias de 1750 – 950 Hz en 160 ms.

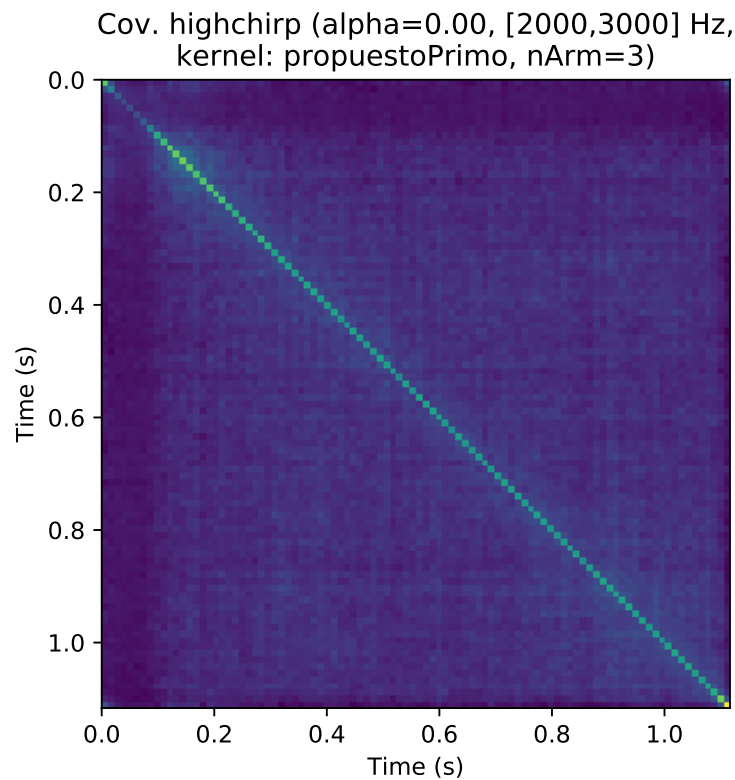
los píxeles de colores más claros corresponden a una covarianza alta, y los colores más oscuros corresponden a una covarianza baja. Salvo un breve instante de 0.3 s en la matriz de covarianza del chirrido bajo (correspondiente a un segmento fuera de la modulación de interés que se puede descartar), se observa que todas las matrices tienen una covarianza  $s_{XY}$  casi nula y una varianza alta  $s_X^2$ . Se destaca también que las varianzas disminuyen en los instantes correspondientes a las modulaciones APS y aumentan en el resto, es decir, solo son altas en los instantes de ruido. Este comportamiento es el esperado, pues en las modulaciones de frecuencia los contornos musicales siguen un patrón definido (no caótico, como en el ruido), lo que implica que poseen una varianza pequeña. La matriz de covarianza cucú, a diferencia de las de los chirridos alto y bajo, es singular, por lo que se usó el método SVD, descrito en la sección 2.4.7, para calcular su pseudoinversa. Una posible explicación de que solo 132 de los 153 contornos musicales cucú de la tabla 3.1 fueron linealmente independientes, es que el tamaño del rango de frecuencias de los chirridos alto y bajo es de 1000 Hz y 800 Hz, respectivamente, pero el del sonido cucú es apenas de 200 Hz, por lo que es más fácil para este APS repetir una señal de contorno musical.

### 3.2.5 Matrices de covarianza sintéticas

Como se determinó en la sección anterior, las matrices de covarianza de la sección 3.2.4 presentan varianzas bajas en los instantes correspondientes a las modulaciones musicales de la diagonal, y varianzas altas en los instantes de ruido. Dado que el mismo comporta-

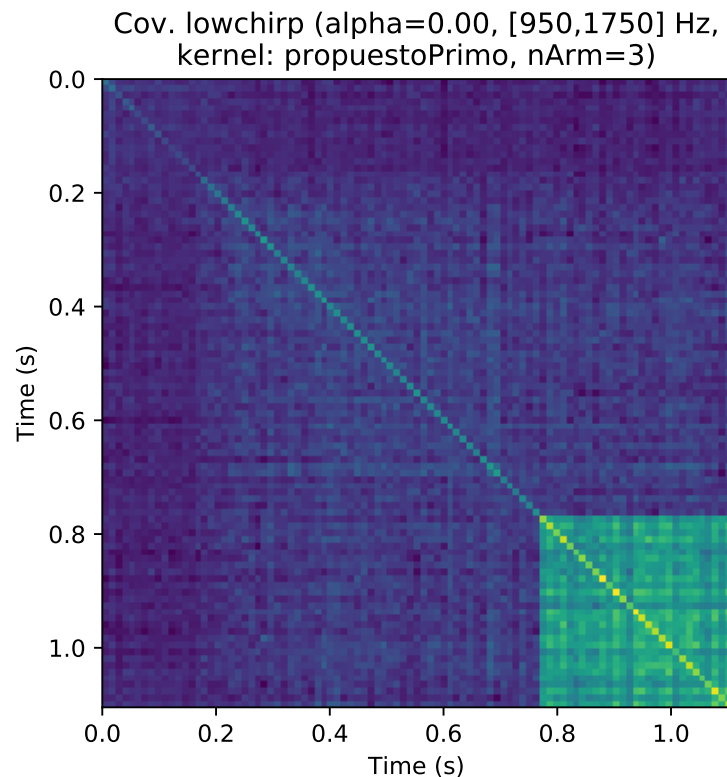


**Figura 3.11:** Matriz de covarianza cucú obtenida al analizar las grabaciones.



**Figura 3.12:** Matriz de covarianza del chirrido alto obtenida al analizar las grabaciones.



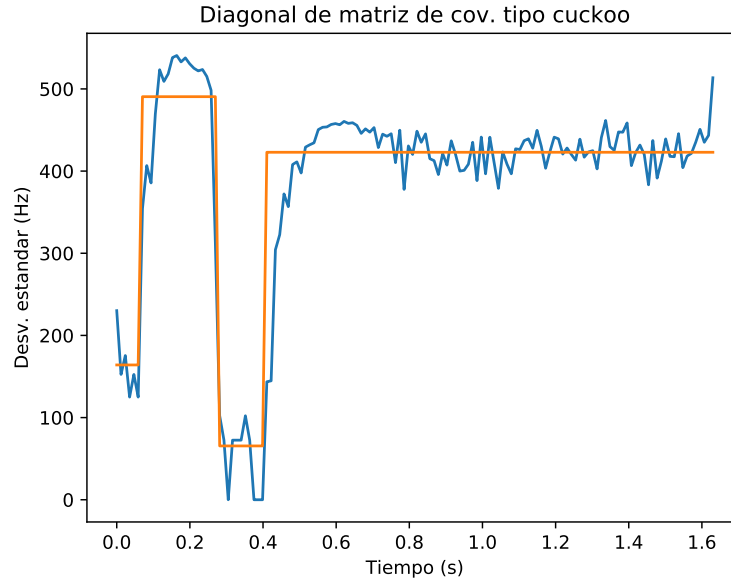


**Figura 3.13:** Matriz de covarianza del chirrido bajo obtenida al analizar las grabaciones.

miento se observa en las tres matrices de covarianza, surgió el interés por crear versiones sintéticas que permitieran procesar nuevos tipos de sonidos sin la necesidad de capturar sus grabaciones. Como lo muestran las figuras 3.14, 3.15 y 3.16, la desviación estándar “promedio” de la modulación de frecuencia del cucú es de  $s_{\min} = 115$  Hz (el promedio entre 165 Hz y 65 Hz), la del chirrido alto es de  $s_{\min} = 780$  Hz y la del chirrido bajo es de  $s_{\min} = 360$  Hz. Estas varianzas parecen ser directamente proporcionales a la resolución de frecuencias del banco de núcleos,  $df$ , y al rango de frecuencias analizado,  $F = f_{\max} - f_{\min}$ , e inversamente proporcionales a la duración del contorno musical  $L$ . Para averiguar si esta proporcionalidad es correcta se consultó la información descrita en la tabla 1.1 y la sección 1.2.2, y se construyó el siguiente sistema de ecuaciones :

$$\begin{pmatrix} df & F & L \end{pmatrix} \begin{pmatrix} b \\ 1.3937 \\ 1.0055 \\ -3648.4375 \end{pmatrix} = \begin{pmatrix} s_{\min} \\ 115 \\ 780 \\ 360 \end{pmatrix},$$

donde se observa que, en efecto,  $df$  y  $F$  son directamente proporcionales a la varianza porque sus coeficientes,  $b_0$  y  $b_1$ , son positivos y cercanos a la unidad, y que  $L$  es inversamente proporcional, porque su coeficiente  $b_2$  es negativo (aún más influyente que los otros, pues  $-3648 \ll -1$ ). Esto podría explicarse como que  $s_{\min} \propto df$  porque los saltos en frecuencia determinan la distancia mínima entre las alturas musicales candidatas y  $s_{\min} \propto F/L$  porque entre mayor es la pendiente del decaimiento frecuencial, mayor es el riesgo de que una



**Figura 3.14:** Raíz cuadrada de la diagonal de la matriz de covarianza del APS cucú. La desviación estándar del primer tono es aproximadamente 165 Hz, la desviación estándar del segundo tono es aproximadamente 65 Hz, y la desviación estándar máxima del ruido es 540 Hz.

misma ventana de análisis contenga dos tonos musicales, lo que confunde al estimador de altura musical. Empleando estas observaciones se propone modelar la varianza de los contornos musicales APS como sigue:

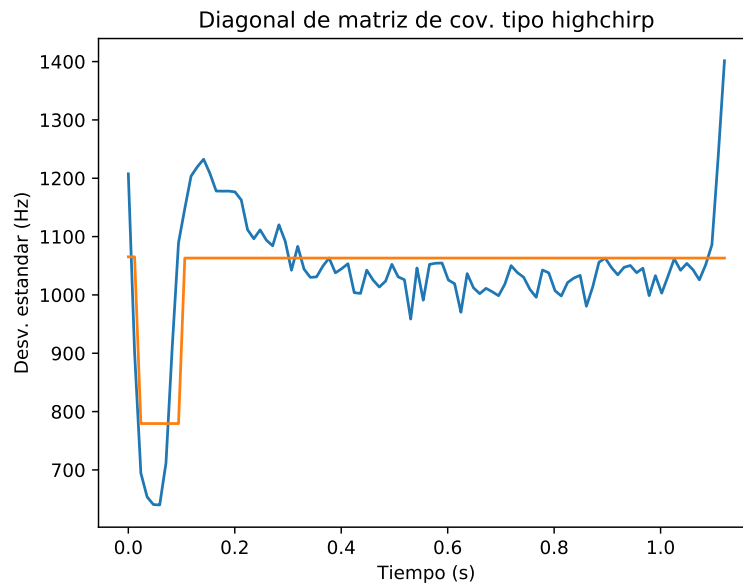
$$S_{t,t}^{(\text{cuckoo})} = \begin{cases} (115 \text{ Hz})^2, & 0s \leq t \leq 0.06s \vee 0.27s \leq t \leq 0.4s \\ (1100 \text{ Hz})^2, & \text{en el resto,} \end{cases}$$

$$S_{t,t}^{(\text{calto})} = \begin{cases} (780 \text{ Hz})^2, & 0.01s \leq t \leq 0.09s \\ (3000 \text{ Hz})^2, & \text{en el resto} \end{cases}$$

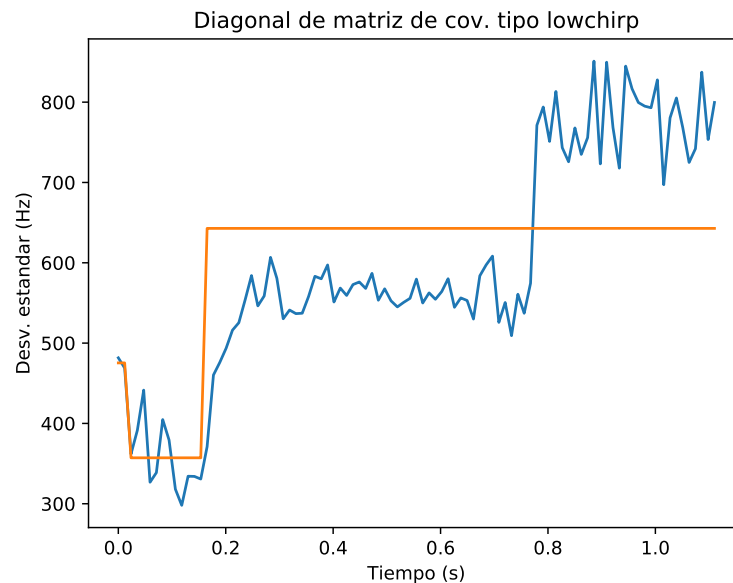
y

$$S_{t,t}^{(\text{cbajo})} = \begin{cases} (360 \text{ Hz})^2, & 0.01s \leq t \leq 0.15s \\ (1750 \text{ Hz})^2, & \text{en el resto,} \end{cases}$$

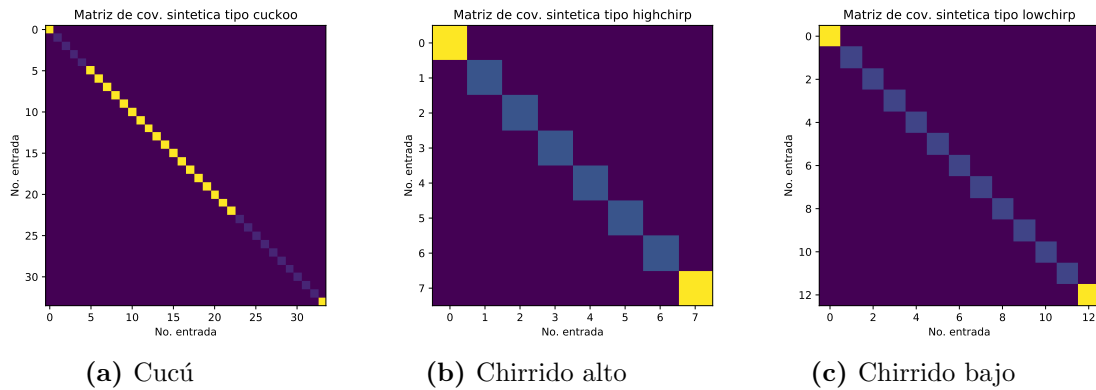
donde se eligió que la varianza máxima fuera la frecuencia máxima admitida, pues como lo muestra la figura 3.7, la altura musical del ruido oscila entre cero y la frecuencia más alta. Para mejorar la calidad de la estimación, también se propone que la primer y última entrada de cada diagonal contenga el valor de varianza más alto, pues en estos instantes las deficiencias en los recortes realizados podrían afectar la varianza. Una visualización gráfica de las matrices sintéticas propuestas para los instantes del contorno musical se muestra en la figura 3.17.



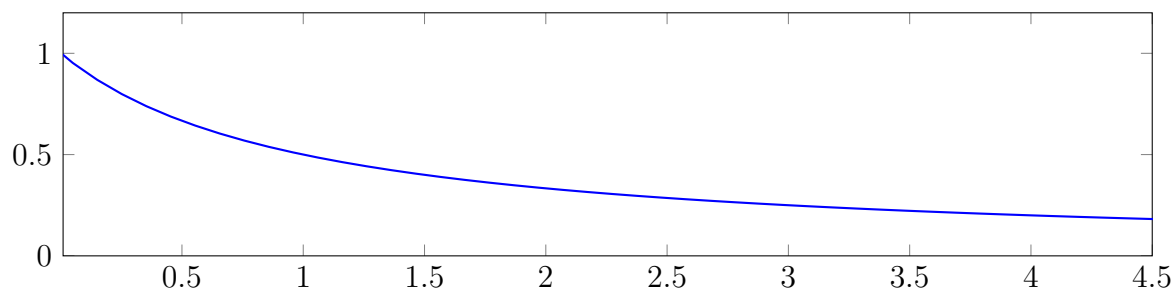
**Figura 3.15:** Raíz cuadrada de la diagonal de la matriz de covarianza del APS chirrido alto. La desviación estándar de la modulación de frecuencia es aproximadamente 780 Hz, y la desviación estándar máxima del ruido es 1400 Hz.



**Figura 3.16:** Raíz cuadrada de la diagonal de la matriz de covarianza del APS chirrido bajo. La desviación estándar de la modulación de frecuencia es aproximadamente 360 Hz, y la desviación estándar máxima del ruido es 850 Hz.



**Figura 3.17:** Matrices de covarianza sintéticas cucú, chirrido alto y chirrido bajo.



**Figura 3.18:** Gráfica de la ecuación  $y = 1/(s + 1)$ , usada para normalizar la distancia de Mahalanobis en el rango unitario.

### 3.2.6 Tratamiento de la salida de la distancia de Mahalanobis

Dado que la distancia de Mahalanobis se expresa en términos del número de desviaciones estándar respecto de la media, es necesario normalizarla en el rango unitario para hacerla compatible con la aplicación del umbral de alerta  $\beta$ . Para ello se usa la ecuación  $y = 1/(s + 1)$  graficada en la figura 3.18, que permite obtener un nivel de alerta máximo de la unidad cuando la desviación estándar es nula, es decir, cuando la correspondencia es perfecta y, por otro lado, permite obtener un valor de 0.3 cuando la desviación es de dos. Con desviaciones estándar mayores a dos, el puntaje disminuye aún más de manera geométrica, lo que resulta conveniente porque estas desviaciones están fuera del rango tolerado.

### 3.2.7 Uso de la distancia de Mahalanobis en contornos musicales

En la búsqueda realizada en revistas especializadas y bases de datos de patentes (Espacenet, Patentscope, USPTO) no se ha encontrado literatura que explique que la distancia de Mahalanobis pueda emplearse para reconocer patrones acústicos mediante el contorno musical. Tampoco se ha encontrado literatura que proponga que la matriz de covarianza pueda modelarse mediante los parámetros del estimador de altura. El concepto más cer-

cano encontrado fue desarrollado por Moh *et al.* para reconocer canciones de la colección USPOP2002, pero ellos más bien entrenaron una matriz de covarianza empleando vectores de 20 coeficientes cepstrales de mel (MFCC, por sus siglas en inglés) [53]. Con estos coeficientes modelaron el timbre de los sonidos a identificar, pero no su altura musical, como se propone en esta sección.

### 3.3 Umbral adaptativo de tono

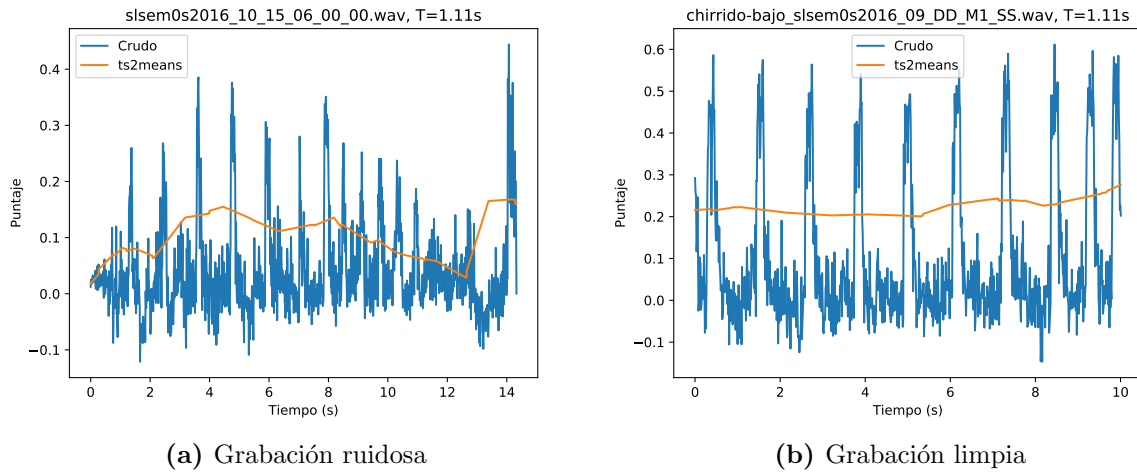
Uno de los problemas del enfoque original es que separa los puntajes de los sonidos armónicos y no armónicos usando un umbral estático, que es incapaz de adaptarse a los niveles de SNR altos y bajos. Como se dijo en la sección 2.2, esto puede corregirse procesando la salida del banco de núcleos con el algoritmo TS2Means o la EMA, pues ambos realizan un ajuste dinámico del umbral de tono. En experimentos preliminares se obtuvieron buenos resultados usando ambas técnicas. Por ejemplo, las figuras 3.19a y 3.19b muestran que TS2Means ubica al umbral de tono en el punto medio entre los picos armónicos y el tope del piso no armónico, y las figuras 3.20a y 3.20b muestran que la EMA ubica el umbral de tono al tope del piso no armónico.

Para asegurar la estabilidad del TS2Means se utiliza una ventana de análisis con duración mínima de  $T_v = T$ , es decir, un periodo de la señal APS, y una duración máxima de  $T_v = 2T$ . Esto garantiza que se cubre al menos un pico de actividad APS en la parte más significativa de la ventana de análisis [38]. Respecto a la EMA, los experimentos preliminares fueron realizados usando un valor  $\lambda = 0.99$  (correspondiente a un periodo del chirrido bajo muestreado a 86 FFT/s), y aunque la EMA no permite obtener una separación tan buena como la de TS2Means, es una alternativa atractiva por su bajo costo computacional (demostrado en la sección 2.3.1) y porque su retraso de grupo no es tan significativo. Sin embargo, una desventaja de la EMA es que  $\lambda$  podría ser otro parámetro a optimizar.

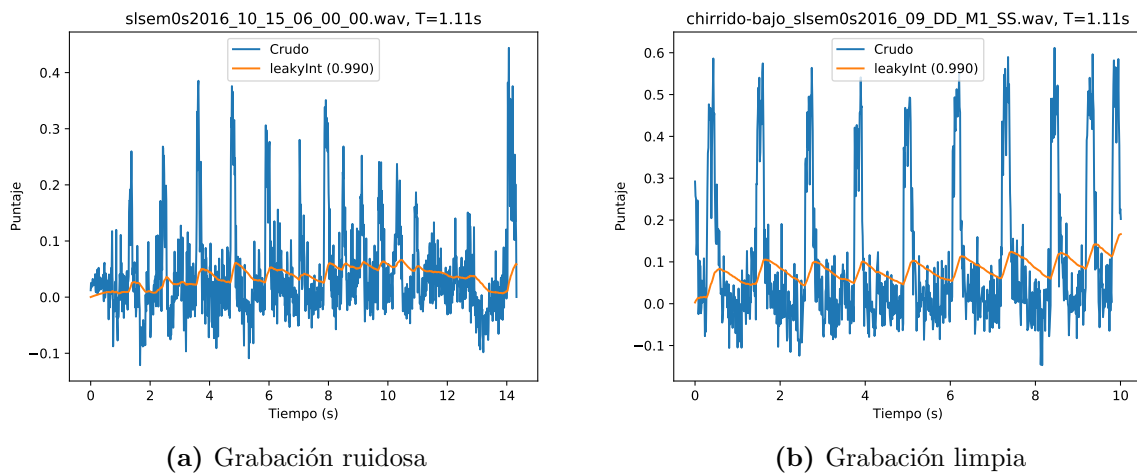
### 3.4 Implementación de la solución

Se propone implementar las mejoras propuestas usando el lenguaje de programación Python. Además de ser este un lenguaje apto para el prototipado rápido, posee bibliotecas científicas de procesamiento digital de señales (DSP, por sus siglas en inglés) como Numpy (operaciones de álgebra lineal), Scipy (cálculo de la FFT, lectura y escritura de audio) y Matplotlib (graficación de señales).

Además, Python permite tener un mayor control sobre las estructuras de datos, a diferencia de Octave y Matlab, e incluso provee interfaces bien documentadas para interactuar con código de bajo nivel, como C y C++. Python también provee módulos para invocar código Octave, como el *oct2py*, e invocar código Matlab, como el *PyMatlab*, lo que resulta útil para utilizar la implementación en Matlab del TS2Means [68]. Otra ventaja de



**Figura 3.19:** Umbrales  $\alpha$  calculados por TS2Means.



**Figura 3.20:** Umbrales  $\alpha$  calculados por la EMA usando  $\lambda = 0.99$ .

Python, es que usa precisión numérica “doble” (1 bit de signo, 11 bits de exponente y 52 bits de mantisa) en el tipo de dato *float64* de la bibliotecas Scipy, la cual cumple con el estándar IEEE 754 y permite manejar errores de redondeo bajos (de  $2^{-52} \approx 2.22 \cdot 10^{-16}$ ). Este es un factor a tomar en cuenta para evitar imprecisiones en los cálculos [69, 70]. Se eligió la versión 2.7, y no la versión 3 de Python, particularmente por la experiencia del autor con la primera.





# Capítulo 4

## Resultados y análisis

En este capítulo se explica la metodología de evaluación para medir el desempeño de la solución propuesta y luego se documentan los resultados obtenidos. Para ello, la sección 4.1 explica la metodología de evaluación a usar, que contiene una modificación respecto de la usada en el estudio de Ruiz *et al.* La sección 4.2 presenta las matrices de puntajes construidas con el banco de núcleos y el diseño de núcleo propuesto. Estas son útiles para evaluar la calidad de los núcleos y asegurarse de que la altura musical estimada sea clara, es decir, que la energía en los armónicos o subarmónicos sea al menos un 20% más baja que la energía presente en la frecuencia fundamental. Finalmente, la sección 4.3 presenta los escenarios de prueba a ejecutar para evaluar las mejoras propuestas, y los resultados obtenidos, junto con una valoración que justifica o reprueba la pertinencia de cada mejora.

### 4.1 Metodología de evaluación

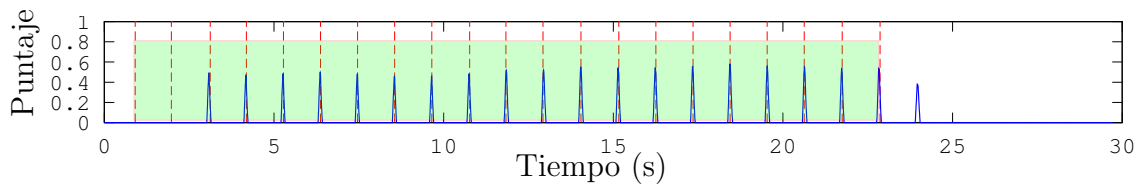
Como se comentó en la sección 3.2.1, se contó con anotaciones manuales que indican el comienzo de las modulaciones de frecuencia APS. Con esta información se determinó la cantidad de verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN), como hicieron originalmente Ruiz y sus colegas. Ellos propusieron contar estos estadísticos mediante operaciones de conjuntos, agrupando el número de ventanas en el rango establecido por el primer y último comienzo detectado, y calculando los estadísticos usando intersecciones, complementos y restas. La forma de definir los conjuntos anotados y observados es la siguiente:

$$A = \{i \in \mathbb{N} / t_{A_1} \leq i \leq t_{A_N}\} \quad (4.1)$$

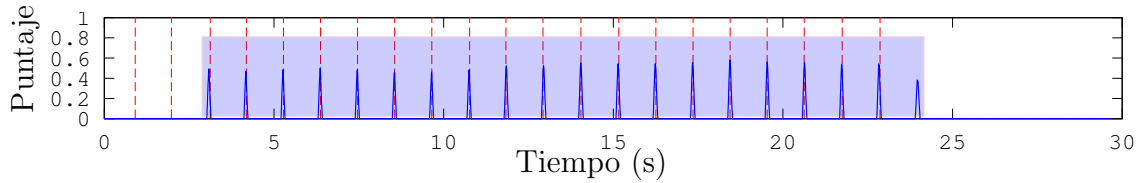
y

$$B = \{j \in \mathbb{N} / t_{B_1} \leq j \leq t_{B_N}\}, \quad (4.2)$$

donde  $t_{A_1}$  y  $t_{A_N}$  son la primer y última anotación manual, y  $t_{B_1}$  y  $t_{B_N}$  son las primer y última anotación hecha por el sistema. Estos conjuntos se muestran en la figura 4.1.



(a) Conjunto formado con las anotaciones manuales



(b) Conjunto formado con las alertas emitidas

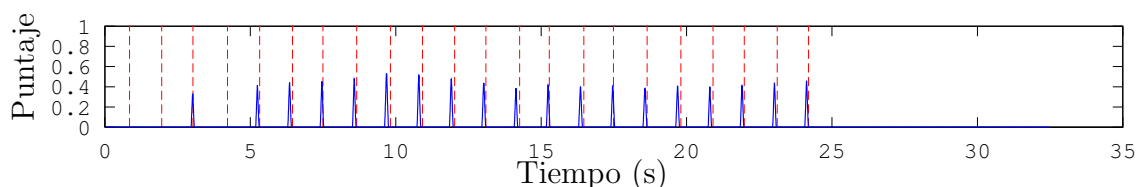
**Figura 4.1:** Anotación manual (rectángulo verde de la figura. 4.1a) y automática (rectángulo azul de la figura 4.1b) de las ventanas de una grabación de chirrido alto correspondientes a la secuencia principal de actividad. Las líneas verticales intermitentes son las anotaciones manuales, y las líneas verticales sólidas son las alertas emitidas (señal de alerta tomada de [10]).

Asimismo,  $dt$  es el recíproco de la frecuencia de muestreo de la STFT (es decir,  $dt = 1/86$ , según lo expuesto en la sección 1.2.3). Las métricas de detección deseadas se calculan como sigue:  $VP = \text{card}(A \cap B)$ ,  $FP = \text{card}(B - A)$ ,  $VN = \text{card}(A^c \cap B^c)$  y  $FN = \text{card}(A - B)$ . El operador  $\text{card}(\cdot)$  se refiere a la cardinalidad. Con base en las métricas se obtienen las tasas de detección: precisión, especificidad, sensibilidad, medida F y el coeficiente de correlación de Matthew, según lo explicado en la sección 2.5.

### 4.1.1 Pertinencia de la metodología existente

Aunque la metodología de evaluación es útil, Ruiz *et al.* asumieron con este método que las alertas emitidas dentro del intervalo delimitado por  $t_{B_1}$  y  $t_{B_N}$  corresponden, en todos los casos, con los picos de actividad del APS, es decir, que los comienzos detectados dentro del conjunto  $B$  están siempre separados por  $T$  segundos, sin faltar ninguno, y que en medio de los comienzos no hay falsos negativos [10], lo que no siempre se cumple, como se observa en la figura 4.2.

Una mejora de este método de verificación es necesaria, pero debe seguir dos criterios de evaluación del dominio de los APS. El primero es que no tiene sentido castigar las alertas que no se ubican exactamente en los comienzos posteriores a cada pico de actividad del APS, ya que la respuesta de cualquier sistema no es inmediata, sino que necesita tiempo para detectar el sonido. El segundo es que, por cuestiones del dominio del problema, anotar los falsos positivos que ocurren en medio de los picos de actividad no es importante, pues se desea que las alertas sean emitidas mientras el APS esté activo. En cambio, se considera más urgente verificar que el rango comprendido entre el primer y el último



**Figura 4.2:** Señal de alerta de un chirrido alto tomada del estudio de Ruiz *et al.* El segundo pico de actividad, ignorado erróneamente, corresponde a un falso negativo.

comienzo contenga suficientes alertas para cubrir cada pico de actividad, y que el primer comienzo detectado ocurra lo más pronto posible después del primer pico de actividad. Esto ya lo realiza el método de evaluación actual, lo que faltaba era castigar los picos ausentes en medio de la primer y última anotación automática, y una forma de hacerlo es calcular la unión del conjunto de ventanas, donde cada conjunto comprende un periodo del APS y se forma posteriormente a cada pico de alerta emitido. Esto conlleva a la mejora propuesta en la sección 4.1.2.

### 4.1.2 Mejoramiento de los conjuntos de evaluación

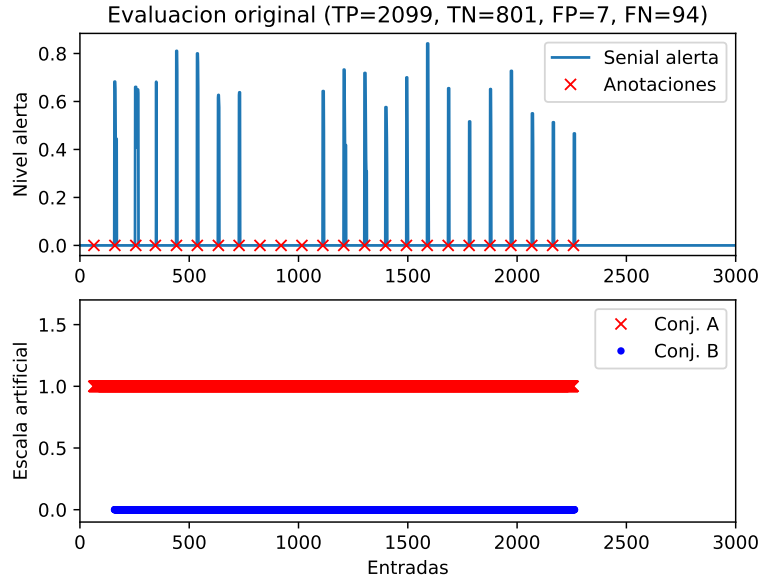
Siguiendo un enfoque similar al planteado en la literatura, se propone redefinir el conjunto  $B$ , de las ventanas anotadas, considerando una cantidad finita de ventanas por cada comienzo detectado [10], es decir, en vez de tomar todas las ventanas entre el primer y último comienzo, se propone asignar a cada pico de alerta emitido el equivalente a un periodo de actividad del APS, y luego concatenar los elementos asignados mediante el operador de unión. De esta manera se pueden castigar los picos de alerta faltantes, descritos en la sección 4.1. Formalmente esto se plantea redefiniendo el conjunto  $B$  como sigue:

$$B = \bigcup_{m=1}^M \{j \in \mathbb{N} / t_{B_m} \leq j \, dt \leq t_{B_m} + T\},$$

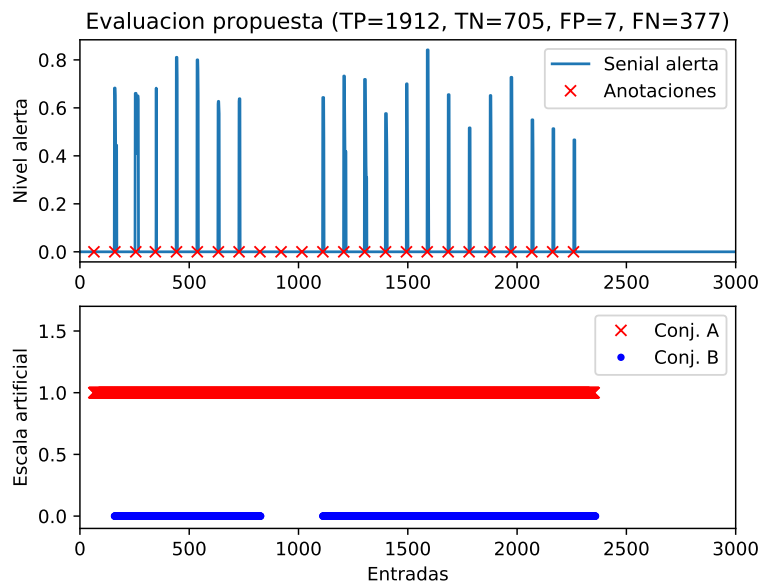
donde  $dt$  es el recíproco de la cantidad de ventanas generadas por segundo y  $\{t_{B_1}, \dots, t_{B_M}\}$  son los instantes de alerta detectados por el sistema. Esto tiene la ventaja de evitar contar como verdaderos positivos las ventanas posteriores a un pico de actividad del APS para el que no se emitió ninguna alerta. Las figuras 4.3 y 4.4 ilustran la utilidad de este enfoque, donde el método propuesto penalizó las alertas faltantes aumentando la cantidad de falsos negativos de 94 a 377, lo que permite satisfacer el segundo objetivo planteado en la sección 1.3.

## 4.2 Matrices de puntajes del núcleo propuesto

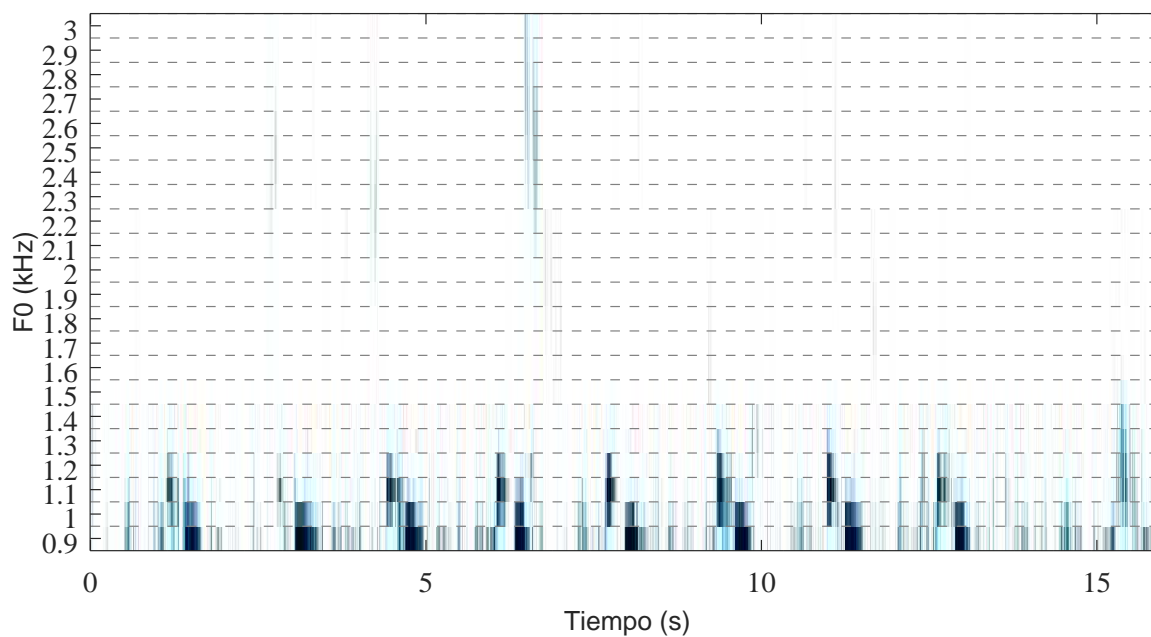
Con el fin de asegurar que los núcleos propuestos en la sección 3.1 fueran capaces de reconocer la altura musical correctamente, es decir, diferenciando la frecuencia fundamental



**Figura 4.3:** Ejemplo de los conjuntos  $A$  y  $B$  calculados por el enfoque original en una grabación real. Las entradas 750 a 1050 no se penalizan.



**Figura 4.4:** Ejemplo de los conjuntos  $A$  y  $B$  calculados por el enfoque propuesto en una grabación real. Las entradas 800 a 1050 se penalizan.



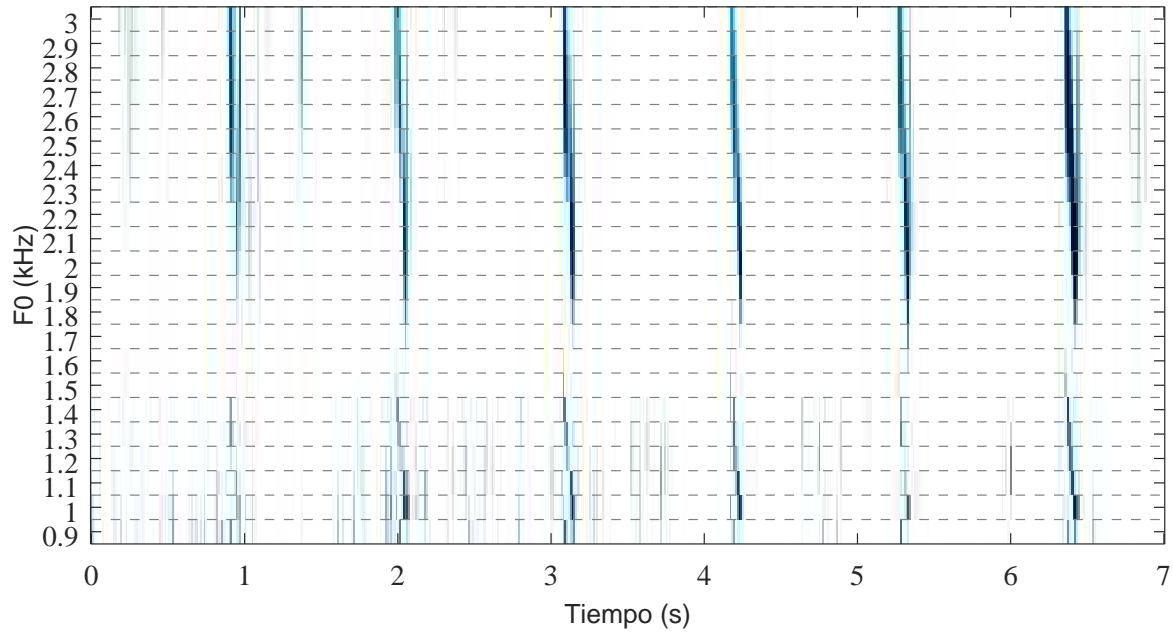
**Figura 4.5:** Puntajes de los núcleos propuestos (de 3 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del cucú. El contorno cucú está correctamente ubicado en 1.1 kHz y 0.9 kHz.

de sus armónicas o subarmónicas, se construyeron las matrices de puntajes de las figuras 4.5, 4.6 y 4.7, para el caso de 3 armónicas, y las figuras 4.8, 4.9 y 4.10, para el caso de 7 armónicas. En ellas se utilizó el banco de núcleos unificado de la sección 3.1 para tener un rango suficientemente grande de frecuencias. Se muestra que la energía de las frecuencias fundamentales es mayor respecto de sus armónicas y subarmónicas, y que el traslape entre los contornos musicales es pequeño, lo que permite afirmar que el primer objetivo de la sección 1.3 fue satisfecho y sugiere que el uso del nuevo núcleo musical unificado es beneficioso por sí solo.

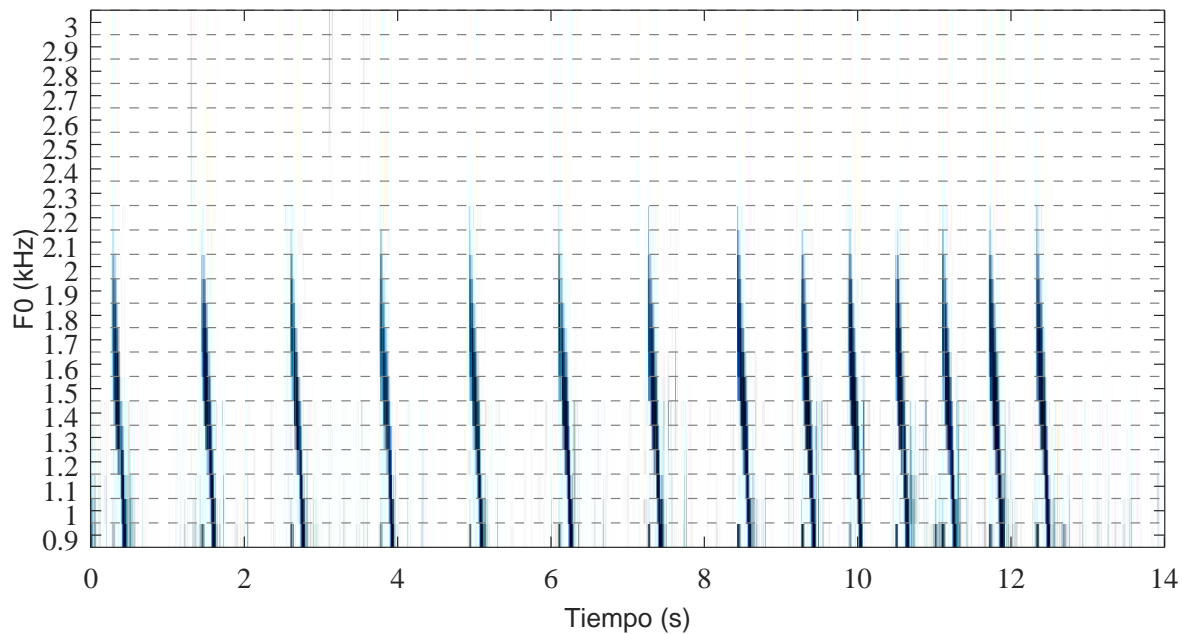
### 4.3 Escenarios de prueba

Hasta ahora, el algoritmo propuesto cuenta con tres tipos de APS a reconocer, dos tipos de núcleo (propuesto vs. mejorado), dos tipos de bancos de núcleos (original vs. unificado) tres tipos de distancias (euclidiana/proporción, Mahalanobis con matriz real y Mahalanobis con matriz sintética), tres formas de establecer el umbral de tono (fijo, TS2Means e EMA) y dos métodos de evaluación (original y propuesto). Para evaluar esa cantidad de variables de manera exhaustiva es necesario ejecutar  $3 \cdot 2 \cdot 2 \cdot 3 \cdot 3 \cdot 2 = 216$  casos de prueba.<sup>1</sup> Para reducir el esfuerzo, se propone evaluar cada mejora de manera individual, respecto

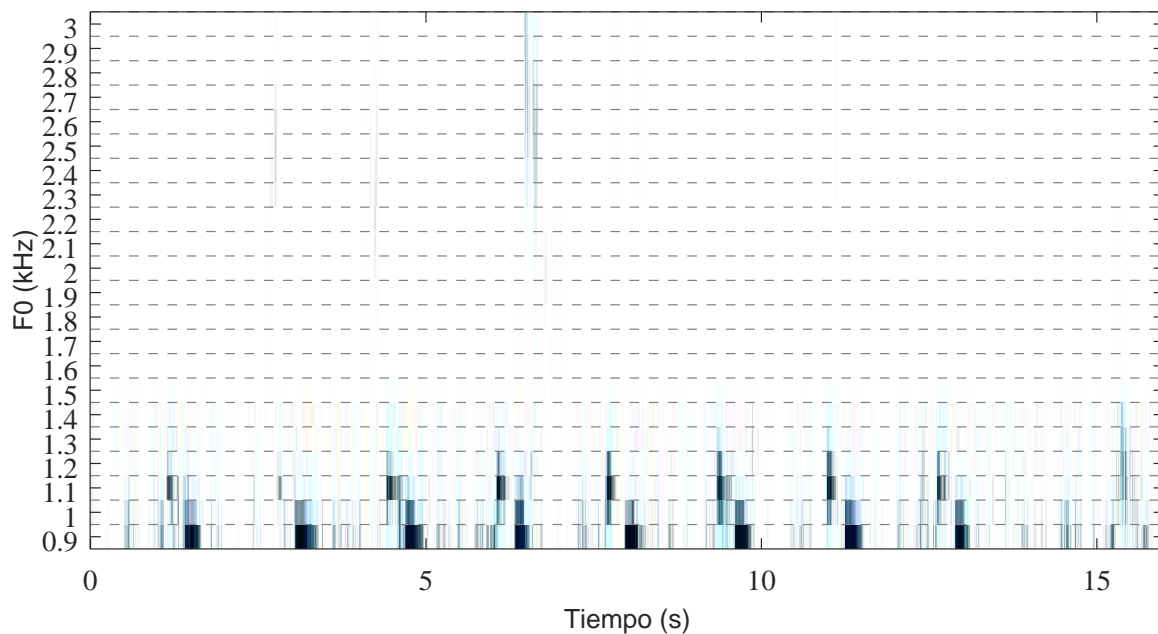
<sup>1</sup>La ejecución automática de cada caso de prueba para los tres tipos de APS duró aproximadamente una hora, usando el equipo disponible para el estudio: Intel(R) Celeron(R) CPU B820 @ 1.70GHz, dual-core, 4 GB de RAM y Xubuntu 16.04.9.



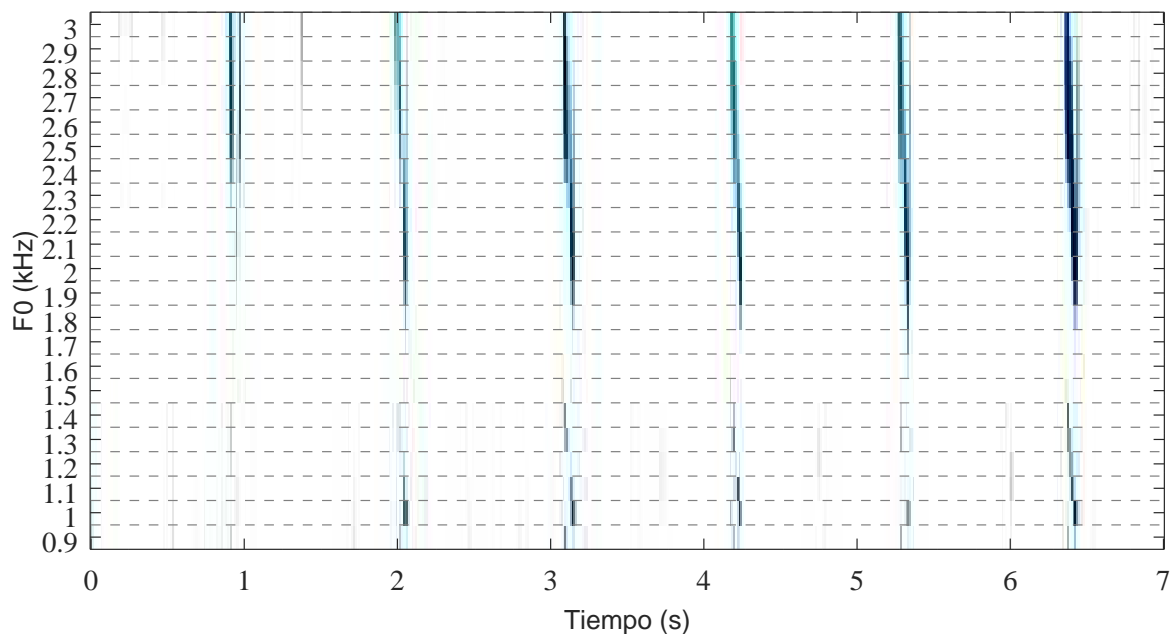
**Figura 4.6:** Puntajes de los núcleos propuestos (de 3 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del chirrido alto. El contorno del chirrido alto está correctamente ubicado en la modulación 3 kHz a 2 kHz.



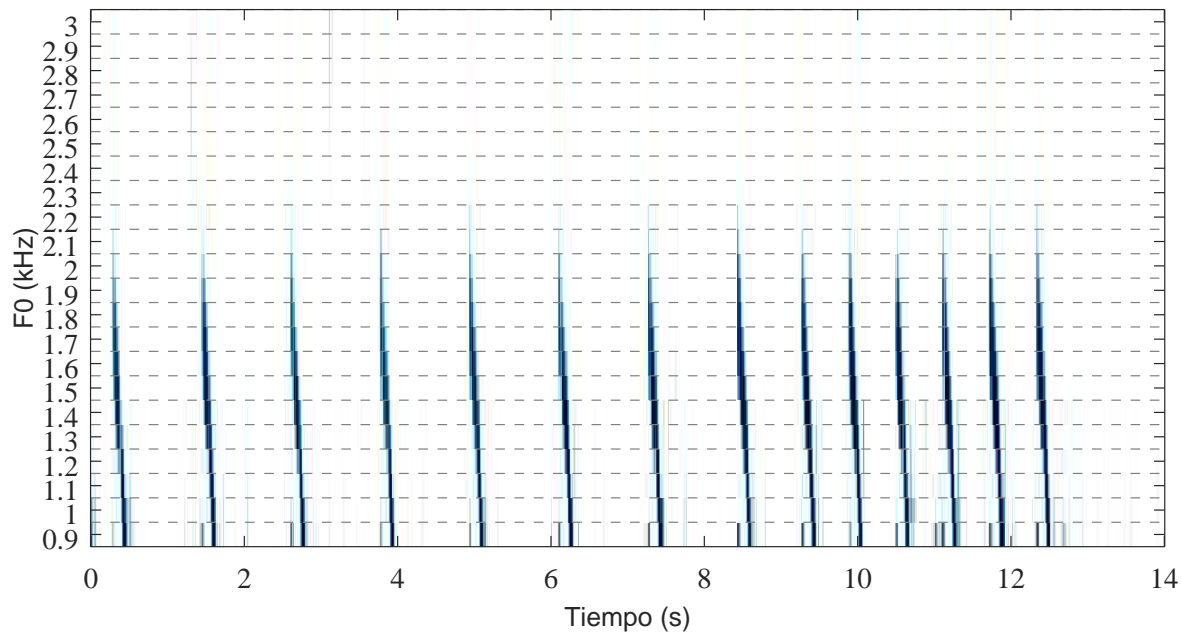
**Figura 4.7:** Puntajes de los núcleos propuestos (de 3 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del chirrido bajo. El contorno del chirrido bajo está correctamente ubicado en la modulación 1.750 kHz a 0.950 kHz.



**Figura 4.8:** Puntajes de los núcleos propuestos (de 7 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del cucú. El contorno cucú está correctamente ubicado en 1.1 kHz y 0.9 kHz.



**Figura 4.9:** Puntajes de los núcleos propuestos (de 7 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del chirrido alto. El contorno del chirrido alto está correctamente ubicado en la modulación 3 kHz a 2 kHz.



**Figura 4.10:** Puntajes de los núcleos propuestos (de 7 armónicas cada uno) del banco de núcleos compartido (0.9–3 kHz) para la grabación del chirrido bajo. El contorno del chirrido bajo está correctamente ubicado en la modulación 1.750 kHz a 0.950 kHz.

al rendimiento original, y luego montar un escenario definitivo con los componentes con el mejor rendimiento. Para ello se plantean los 9 escenarios descritos en la tabla 4.1, donde el escenario 1 reproduce las métricas obtenidas en el enfoque original para confirmar que la solución fue implementada correctamente; el escenario 2 evalúa el comportamiento del sistema original usando la metodología de evaluación mejorada (las métricas obtenidas aquí se usan como un “piso de evaluación” para evaluar el resto de mejoras propuestas); el escenario 3 evalúa la efectividad del diseño de núcleo mejorado, para los casos de 3 y 7 armónicas; el escenario 4 evalúa la efectividad del uso del banco de núcleos unificado; los escenarios 5 y 6 evalúan las mejoras alcanzadas con la distancia de Mahalanobis, uno para el caso de las matrices de covarianza reales y el otro para el caso de las sintéticas (dado que se usan umbrales de tono y alerta fijos — $\alpha$  y  $\beta$ , respectivamente— en los escenarios 3–6, se hace un barrido de valores para encontrar la mejor combinación); los escenarios 7 y 8 estudian las mejoras que provee el uso del TS2Means y la EMA (en el primer caso solo se hace un barrido con los valores de  $\beta$  y en el segundo caso se hace el barrido de valores  $\lambda$  y  $\beta$ ) y en el escenario 9 se determina cuáles son las métricas de rendimiento más altas alcanzadas usando la mejor combinación de núcleos, bancos de núcleos, distancias y umbrales. En cada escenario se evalúan los tres tipos de APS estudiados, promediando las métricas de precisión, especificidad, sensibilidad, medida F y el coeficiente de correlación de Matthew. Como criterio de aceptación, se considera que una mejora es justificable cuando la diferencia del MCC entre el escenario siendo evaluado y el escenario 2 es mayor al 5%. Como se discutió en la sección 2.5 se utiliza el MCC como medida de evalua-



**Tabla 4.1:** Resumen de los escenarios propuestos para evaluar las mejoras propuestas.

ID	Núcleo	Banco	Distancia	Umbral	Evaluación
1	original	original	euclidiana modificada	fijo	original
2	original	original	euclidiana modificada	fijo	<b>propuesta</b>
3	<b>propuesto</b>	original	euclidiana modificada	fijo	propuesta
4	<b>propuesto</b>	<b>propuesto</b>	euclidiana modificada	fijo	propuesta
5	original	original	<b>mahalanobis matriz real</b>	fijo	propuesta
6	original	original	<b>mahalanobis matriz sintética</b>	fijo	propuesta
7	original	original	euclidiana modificada	<b>TS2Means</b>	propuesta
8	original	original	euclidiana modificada	<b>EMA</b>	propuesta
9	<b>mejor</b>	<b>mejor</b>	<b>mejor</b>	<b>mejor</b>	propuesta

ción, pues abarca los cuatro tipos de estadísticos: VP, VN, FP y FN. A continuación, las secciones 4.3.1 a la 4.3.9 presentan los resultados obtenidos en cada escenario.

### 4.3.1 Escenario 1: configuración original

El escenario 1 usa los núcleos, bancos de núcleos, distancias, número de eventos a considerar y umbrales fijos de  $\alpha$  y  $\beta$  originales. La tabla 4.2 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.11 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. La figura 4.11a muestra que los umbrales del APS cucú sigue siendo válidos y que la totalidad de los picos de la secuencia principal fueron detectados, con una especificidad alta. La figura 4.11b muestra que los umbrales para el chirrido alto también son correctos, logrando detectar 19 de los 22 picos de la secuencia principal (86%). Es normal que el evaluador del chirrido alto no reconozca los dos primeros picos de actividad, pues la alerta se emite luego de encontrar tres eventos consecutivos cuya multiplicación es superior al umbral  $\beta = 0.3$ . La figura 4.11c también muestra que los umbrales para el chirrido bajo fueron correctos, logrando reconocer 7 de los 8 picos de la secuencia principal (88%). Se observa, que el valor de  $\beta = 0.3$  debería aumentar a 0.7 para mejorar la especificidad y no detectar los últimos cuatro picos de la secuencia de finalización. La figura 4.12 muestra los espectrogramas usados para determinar las cantidad de picos de las secuencias principales.

**Tabla 4.2:** Parámetros del escenario 1.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$	$\beta$
Cucú	Armónico	4	Proporción	[0.90, 1.10]	200	0.14	0.45
Ch. Alto	Armónico	1	L2 mod.	[2.00, 3.00]	100	0.07	0.3
Ch. Bajo	Arm. impar	5	L2 mod.	[0.95, 1.75]	100	0.07	0.3

Las tasas de detección alcanzadas al analizar las 79 grabaciones fueron de 87% de precisión, 92% de especificidad, 69% de sensibilidad, 71% de medida F y 61% de MCC. Lamentablemente, como lo muestra la tabla 4, solo la precisión se mantiene fiel al valor original, mientras que el resto de tasas difiere de 10% a 20%. A pesar de estas desviaciones, se considera que el escenario 1 fue implementado de manera correcta, pues la figura 4.13 muestra que el análisis de las grabaciones individuales fue similar al procesamiento original. Además, debe considerarse que hubieron diferencias con respecto a la solución de Ruiz *et al.*, pues, por cuestiones de simplicidad, el ancho de los lóbulos del núcleo del chirrido alto no fue de  $0.3f_0$  y el núcleo cucú usado fue de 4 armónicas para los tonos de 1100 Hz y 900 Hz, y no de 4 y 3 armónicas, para el primer y segundo tono, respectivamente.

Evaluando el desempeño del escenario 1 con  $q = 2$ , tanto para el chirrido alto como para el chirrido bajo, se obtuvieron los datos de la tabla 4.4. En ella se observa que el coeficiente de correlación de Matthew o MCC (explicado en la sección 2.5) del cucú fue de 92%, el del chirrido alto fue de 63% y el del chirrido bajo fue de 37%, siendo este último el más bajo de todos. El rendimiento promedio de MCC fue del 64%. En los escenarios futuros se sigue usando el parámetro  $q = 2$  para ambos chirridos, ya que el uso de  $q = 3$  no fue justificado en el estudio de Ruiz *et al.* y se considera que dos eventos son suficientes para emitir una alerta. El MCC se utiliza como medida de comparación general, en lugar del promedio de las tasas de detección, por dos razones: la primera, que permite resumir en una sola cantidad las métricas de los verdaderos negativos, verdaderos positivos, falsos positivos y falsos negativos, y la segunda, que no tiene sentido promediar todas las tasas de detección, pues la medida F ya es una media armónica entre la precisión y la especificidad. También se pudo haber calculado una media aritmética o armónica entre la sensibilidad y la medida F, pero no se encontró literatura que respalde este uso.

### 4.3.2 Escenario 2: metodología de evaluación mejorada

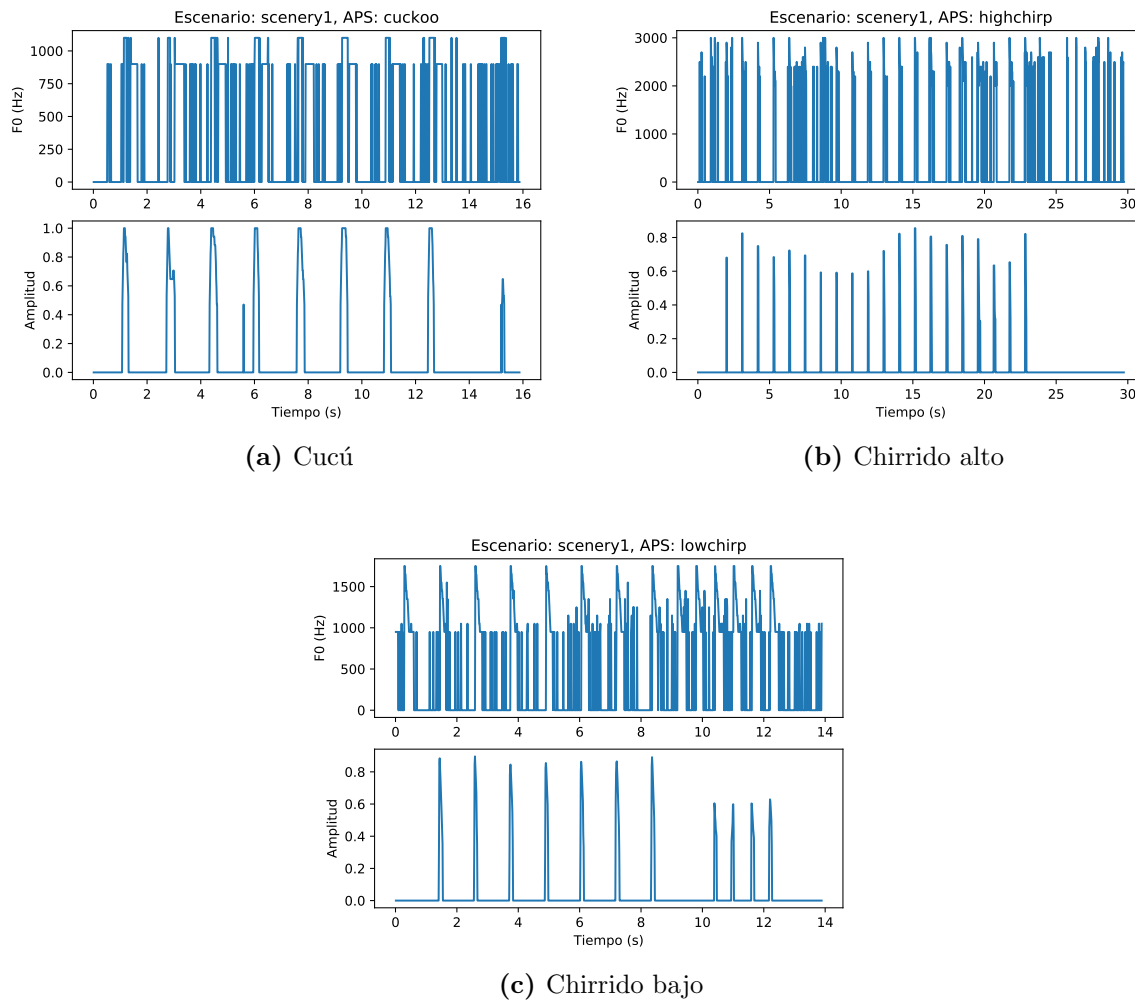
El escenario 2 utiliza la misma configuración descrita en el escenario 1, con la diferencia de que la metodología de evaluación es la empleada en la sección 4.1.2. La figura 4.14 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Las tasas de detección alcanzadas con esta nueva metodología al analizar las 79 grabaciones fue de 89% de precisión, 90% de especificidad, 69% de sensibilidad y 73% de medida F. El rendimiento general en términos del MCC fue de 61% con una desviación

**Tabla 4.3:** Tasas originales vs. escenario 1.

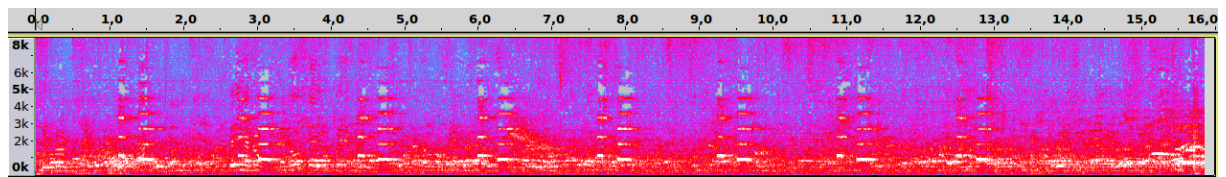
Métrica	Rendimiento (%)	
	Original	Escenario 1
Precisión	87	87 ± 11
Especificidad	83	92 ± 06
Sensibilidad	86	69 ± 24
Medida F	85	71 ± 21
MCC	–	61 ± 24

**Tabla 4.4:** Tasas individuales obtenidas en escenario 1 con  $q = 2$  para ambos chirridos y  $q = 1$  para cucú.

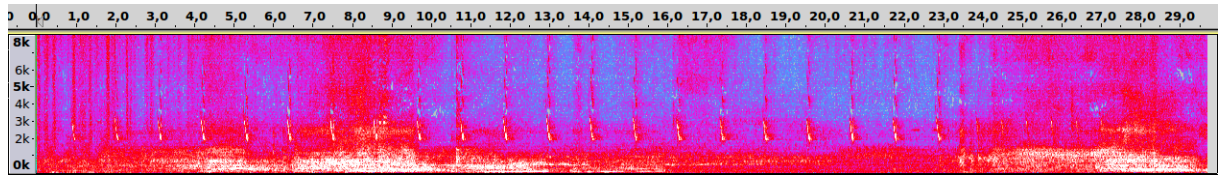
Métrica	Rendimiento (%)			
	Cucú	Chirrido alto	Chirrido bajo	General
Precisión	96 ± 06	93 ± 17	73 ± 31	87 ± 11
Especificidad	89 ± 17	95 ± 17	87 ± 26	90 ± 04
Sensibilidad	100 ± 01	70 ± 35	46 ± 37	72 ± 23
Medida F	98 ± 04	75 ± 34	48 ± 37	74 ± 21
MCC	92 ± 12	63 ± 33	37 ± 33	64 ± 23



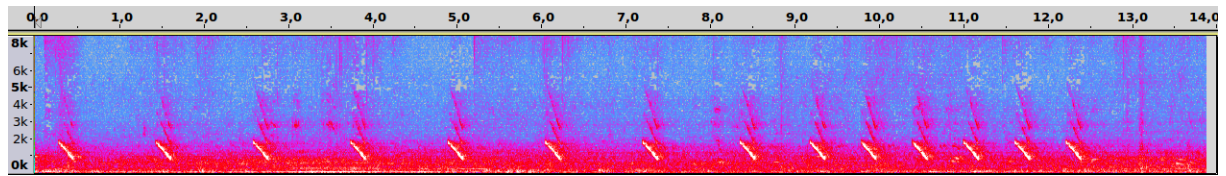
**Figura 4.11:** Señales de alerta del escenario 1 para grabaciones individuales con parámetros descritos en la tabla 4.2. En cada figura hay dos subfiguras, la superior representa el contorno musical y la inferior contiene las correspondencias con la plantilla APS filtradas de acuerdo a la cantidad de eventos  $q$ .



(a) Cucú

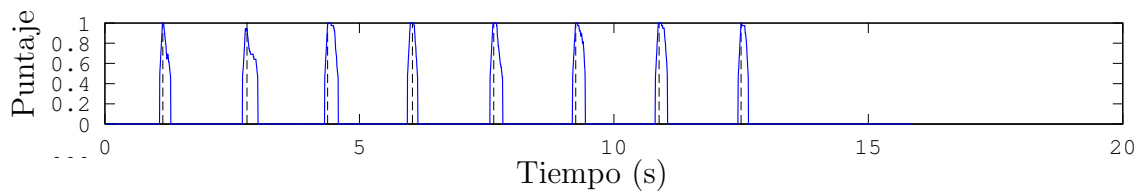


(b) Chirrido alto

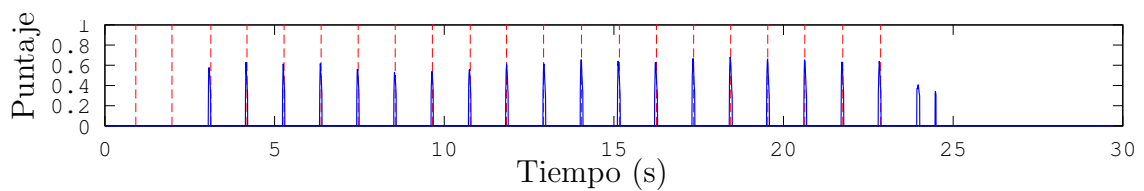


(c) Chirrido bajo

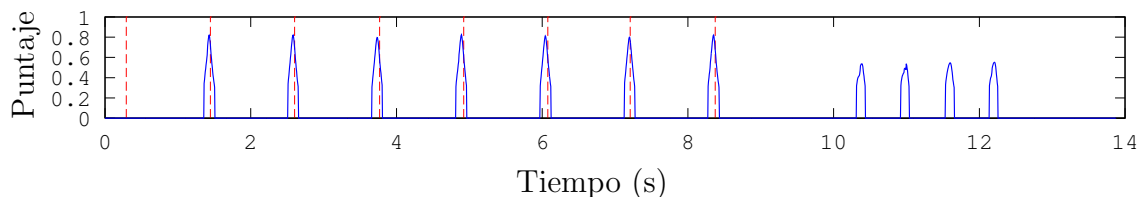
**Figura 4.12:** Espectrogramas para las grabaciones de cucú, chirrido alto y chirrido bajo analizadas individualmente por la solución. El eje horizontal corresponden al tiempo (en segundos) y el vertical a la frecuencia (en hercios).



(a) Cucú



(b) Chirrido alto



(c) Chirrido bajo

**Figura 4.13:** Señales de alerta reportadas en el estudio de Ruiz *et. al* para las grabaciones individuales, según los parámetros descritos en la tabla 4.2.

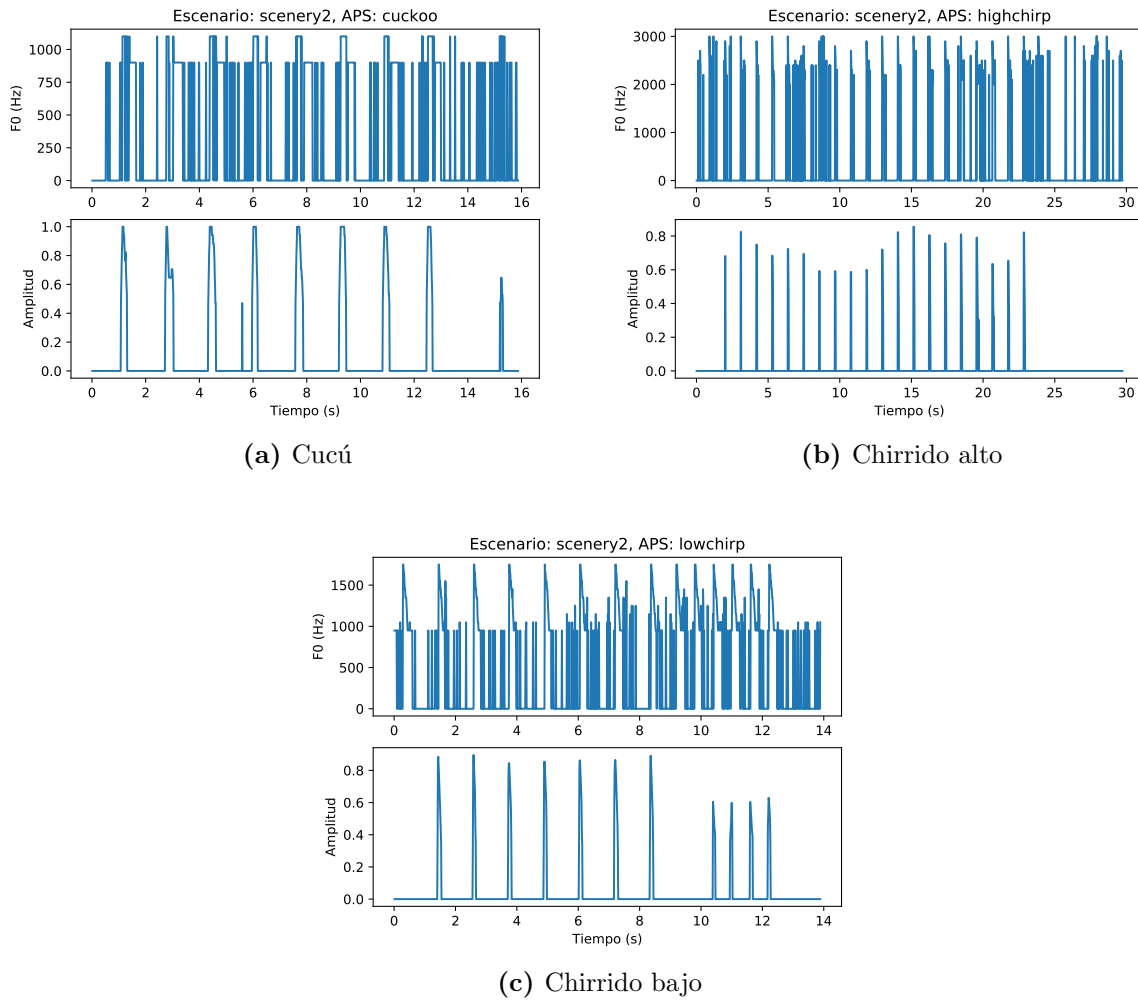
**Tabla 4.5:** Tasas de rendimiento promedio obtenidas en el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	97 ± 05	93 ± 16	77 ± 30	89 ± 09
Especificidad	84 ± 15	96 ± 15	91 ± 19	90 ± 06
Sensibilidad	100 ± 03	60 ± 34	48 ± 36	69 ± 23
Medida F	98 ± 03	67 ± 33	53 ± 37	73 ± 19
MCC	89 ± 12	52 ± 30	41 ± 35	61 ± 21

estándar de 21%, cercano al alcanzado en el escenario 1, pero con una sensibilidad disminuida en un 3%, como se observa en la tabla 4.5, lo que indica que la nueva metodología penalizó con más fuerza a la cantidad de falsos negativos, como se esperaba. El valor MCC obtenido en este escenario se usa en los siguientes escenarios como referencia para determinar si hubo una mejora o un retroceso con las propuestas planteadas.

### 4.3.3 Escenario 3: núcleo propuesto

El escenario 3 usa el tipo de núcleo propuesto, pero con los bancos de núcleos, distancias y umbrales originales. La tabla 4.6 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.15 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Manteniendo la configuración descrita, se hizo un barrido de parámetros  $\alpha$  y  $\beta$  para encontrar el coeficiente de correlación de Matthew más alto contra el cual comparar el rendimiento alcanzado en el escenario 2. Dado que en el escenario 1 los valores menores a 0.5 poseen la mayor variabilidad y considerando que el estudio de Ruiz *et. al* usa los umbrales 0.07, 0.14 y 0.45, se decidió distribuir los parámetros del barrido logarítmicamente mediante la curva  $y(x) = 0.001 \cdot 3.1^x$ . Las tablas 4.7, 4.8, 4.9, 4.10, 4.11, y 4.12 muestran el resultado de los barridos, por tipo de APS y número de armónicas. La tabla 4.13 muestra que las mejores combinaciones de los umbrales permitieron alcanzar un MCC de 89% para el cucú (con un núcleo de 3 armónicas,  $\alpha = 0.03$  y  $\beta = 0.9$ ), 57% para el chirrido alto (con un núcleo de 3 armónicas,  $\alpha = 0.001$  y  $\beta = 0.003$ ), y 41% para el chirrido bajo (con un núcleo de 3 armónicas,  $\alpha = 0.03$  y  $\beta = 0.3$ ). Las mejores métricas también se repiten en el caso del núcleo de siete armónicas, pero se prefirió el núcleo de tres, por tener un menor costo computacional. El rendimiento general MCC fue de 62%, es decir, un 1% mayor al escenario 2, y la desviación estándar fue de 20%. Solo el chirrido alto fue significativamente beneficiado, pues obtuvo una mejora del 5%, pero los sonidos cucú y chirrido bajo no (para ellos no hubo mejora). Esto indica que el núcleo del chirrido bajo podría haber sido implementado correctamente en el estudio de Ruiz *et. al*, y que, al igual que en la sección 1.1.3, la reutilización del mismo núcleo musical para procesar distintos tipos de sonido no es buena idea. Esto justifica el uso de los núcleos propuestos de 3 armónicas solamente para el chirrido alto.



**Figura 4.14:** Contornos musicales y señales de alerta del escenario 2 para grabaciones individuales con parámetros descritos en la tabla 4.2.

**Tabla 4.6:** Parámetros del escenario 3.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Propuesto	3, 7	Proporción	[0.90, 1.10]	200	Fijo
Ch. Alto	Propuesto	3, 7	L2 mod.	[2.00, 3.00]	100	Fijo
Ch. Bajo	Propuesto	3, 7	L2 mod.	[0.95, 1.75]	100	Fijo

**Tabla 4.7:** Rendimiento promedio MCC del cucú en escenario 3 para el caso de 3 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	5	5	5	5	11	47	87
0.003	5	5	5	5	12	49	87
0.010	7	7	7	7	16	56	87
0.030	10	10	10	10	24	64	<b>88</b>
0.090	31	31	31	31	49	76	81
0.300	76	76	76	76	72	69	53
0.900	0	0	0	0	0	0	0

**Tabla 4.8:** Rendimiento promedio MCC del chirrido alto en escenario 3 para el caso de 3 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	<b>56</b>	<b>56</b>	55	55	55	50	0
0.003	55	55	54	55	54	48	0
0.010	55	55	55	55	53	46	0
0.030	51	51	51	51	48	37	0
0.090	24	24	24	23	20	13	0
0.300	2	2	2	2	1	0	0
0.900	0	0	0	0	0	0	0

**Tabla 4.9:** Rendimiento promedio MCC del chirrido bajo en escenario 3 para el caso de 3 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	13	14	13	14	20	35	16
0.003	13	14	14	15	23	36	16
0.010	16	17	18	23	30	31	16
0.030	24	24	25	31	33	<b>40</b>	14
0.090	38	38	38	39	34	23	9
0.300	8	8	8	9	9	4	0
0.900	0	0	0	0	0	0	0



**Tabla 4.10:** Rendimiento promedio MCC del cucú en escenario 3 para el caso de 7 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	21	21	21	21	43	71	<b>88</b>
0.003	22	22	22	22	45	73	<b>88</b>
0.010	26	26	26	26	46	74	<b>88</b>
0.030	33	33	33	33	52	76	82
0.090	53	53	53	53	67	81	73
0.300	73	73	73	73	72	65	37
0.900	0	0	0	0	0	0	0

**Tabla 4.11:** Rendimiento promedio MCC del chirrido alto en escenario 3 para el caso de 7 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

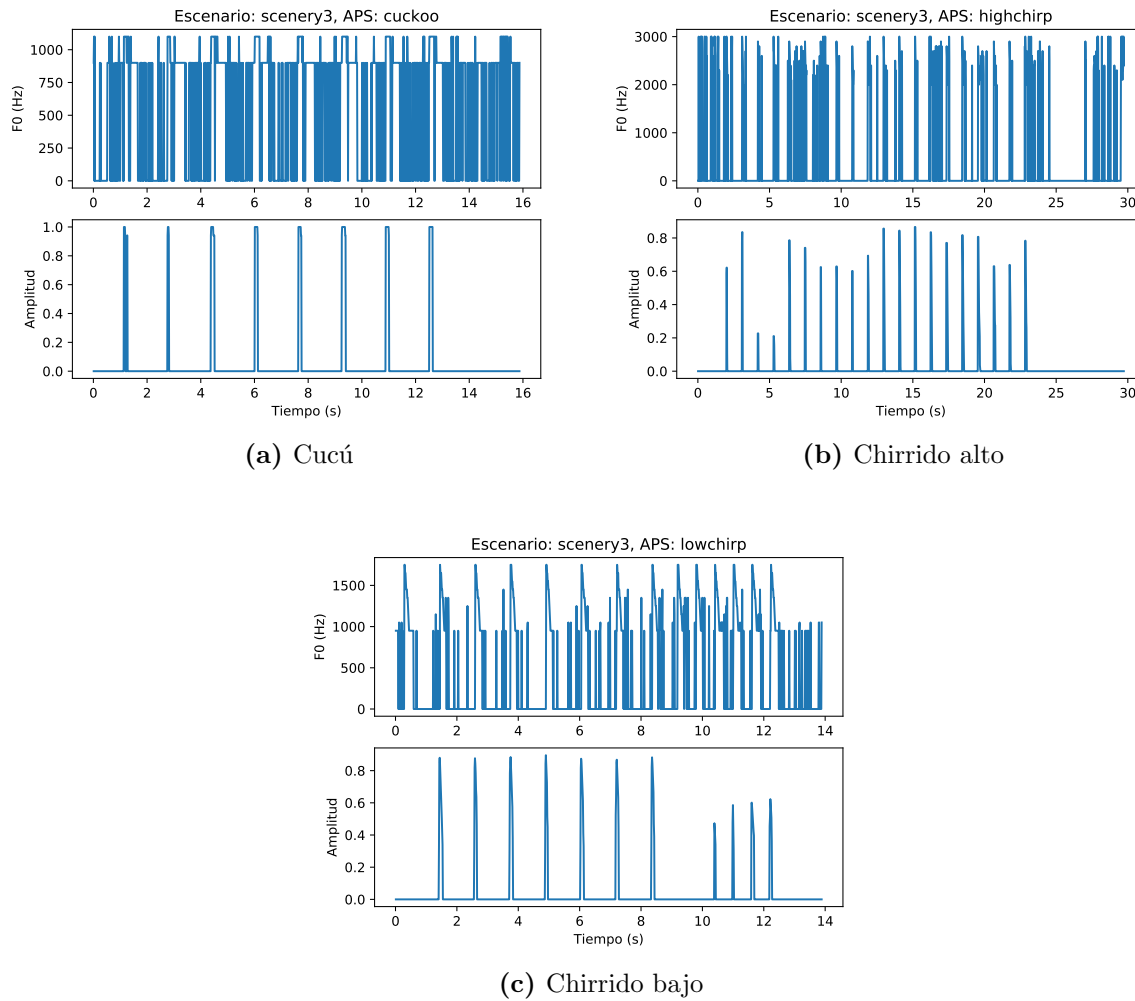
$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	<b>56</b>	<b>56</b>	55	55	54	44	0
0.003	55	55	55	55	53	42	0
0.010	53	53	53	52	49	40	0
0.030	44	44	44	42	38	28	0
0.090	18	18	18	17	15	10	0
0.300	1	1	1	1	1	0	0
0.900	0	0	0	0	0	0	0

**Tabla 4.12:** Rendimiento promedio MCC del chirrido bajo en escenario 3 para el caso de 7 armónicas y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	36	<b>37</b>	<b>37</b>	34	35	29	14
0.003	36	36	36	33	35	28	14
0.010	33	34	34	35	33	29	9
0.030	32	32	31	31	26	23	9
0.090	20	20	20	20	20	17	6
0.300	1	0	0	1	4	0	0
0.900	0	0	0	0	0	0	0

**Tabla 4.13:** Mejores tasas de rendimiento obtenidas en el escenario 3. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	98 ± 09	94 ± 12	79 ± 22	91 ± 09
Especificidad	95 ± 06	86 ± 30	89 ± 23	90 ± 04
Sensibilidad	94 ± 21	75 ± 29	49 ± 36	73 ± 19
Medida F	95 ± 19	79 ± 25	53 ± 37	76 ± 18
MCC	89 ± 22	57 ± 29	41 ± 32	62 ± 20
Diff. esc. 2	0	5	0	1



**Figura 4.15:** Contornos musicales y señales de alerta del escenario 3 para grabaciones individuales con parámetros descritos en la tabla 4.6 y la mejor combinación de valores para  $\alpha$  y  $\beta$ .

**Tabla 4.14:** Parámetros del escenario 4.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Propuesto	3	Proporción	[0.90, 3.00]	100	Fijo
Ch. Alto	Propuesto	3	L2 mod.	[0.90, 3.00]	100	Fijo
Ch. Bajo	Propuesto	3	L2 mod.	[0.90, 3.00]	100	Fijo

**Tabla 4.15:** Rendimiento promedio MCC del cucú en escenario 4 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$							
	0.001	0.003	0.01	0.03	0.09	0.3	0.9	
0.001	9	9	9	9	19	60	<b>84</b>	
0.003	9	9	9	9	19	60	<b>84</b>	
0.010	11	11	11	11	20	62	<b>84</b>	
0.030	13	13	13	13	30	69	<b>84</b>	
0.090	34	34	34	34	53	76	80	
0.300	76	76	76	76	72	69	53	
0.900	0	0	0	0	0	0	0	

#### 4.3.4 Escenario 4: núcleo y banco de núcleos propuestos

El escenario 4 usa las distancias y umbrales originales, pero con el tipo de núcleo y el banco de núcleos propuesto. La tabla 4.14 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.16 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Manteniendo la configuración descrita, se calculó el coeficiente de correlación de Matthew haciendo un barrido de valores  $\alpha$  y  $\beta$  para encontrar los mejores umbrales contra los cuales comparar el rendimiento alcanzado en el escenario 2. Al igual que en el escenario 3, los valores del barrido se distribuyeron logarítmicamente y el núcleo usado fue el mejor en común, es decir, el de 3 armónicas. Se evaluó el uso del banco de núcleos unificado junto con el diseño de núcleo propuesto para disminuir la tasa de error por subarmónicos. Las tablas 4.15, 4.16 y 4.17 muestran el resultado de los barridos, por tipo de APS. La tabla 4.18 muestra que las mejores combinaciones de los umbrales permitieron alcanzar un MCC de 85% para el cucú ( $\alpha = 0.03$ ,  $\beta = 0.9$ ), 52% para el chirrido alto ( $\alpha = 0.001$ ,  $\beta = 0.003$ ), y 40% para el chirrido bajo ( $\alpha = 0.003$ ,  $\beta = 0.3$ ). El rendimiento general MCC fue de 59%, un 3% inferior al rendimiento del escenario 2, mientras que la desviación estándar fue de 19%. Al haber un retroceso en el rendimiento, no se puede justificar el uso del banco de núcleos unificado para ningún tipo de sonido (al menos con la configuración escogida).

**Tabla 4.16:** Rendimiento promedio MCC del chirrido alto en escenario 4 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

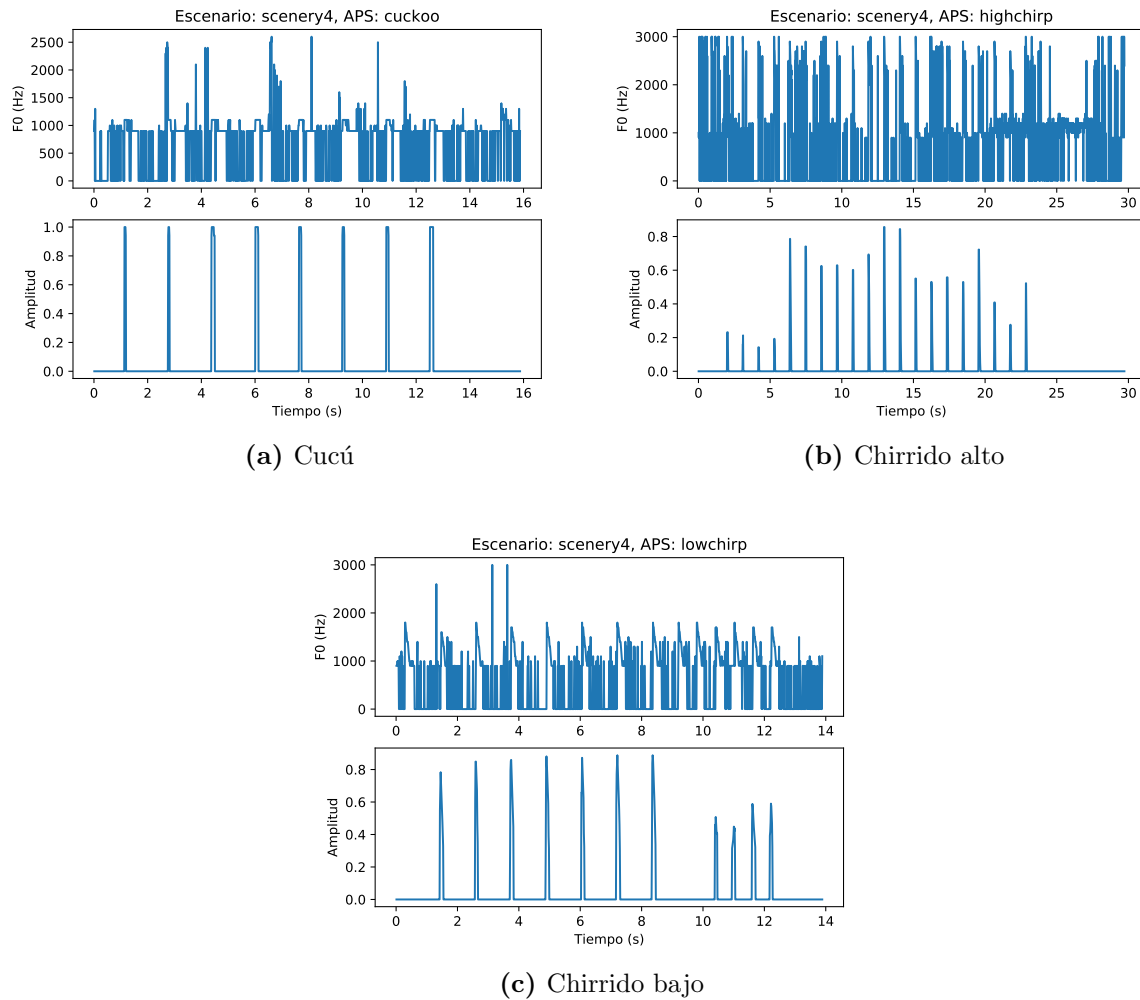
$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	<b>51</b>	<b>51</b>	50	49	44	35	0
0.003	50	50	50	49	46	35	0
0.010	<b>51</b>	50	50	<b>51</b>	47	34	0
0.030	50	50	49	48	44	30	0
0.090	24	24	24	22	19	13	0
0.300	2	2	2	2	1	1	0
0.900	0	0	0	0	0	0	0

**Tabla 4.17:** Rendimiento promedio MCC del chirrido bajo en escenario 4 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	12	11	11	12	20	37	14
0.003	14	12	13	13	22	<b>39</b>	14
0.010	14	14	15	15	25	34	14
0.030	24	24	26	28	31	38	14
0.090	36	37	37	<b>39</b>	32	25	9
0.300	7	7	7	8	8	4	0
0.900	0	0	0	0	0	0	0

**Tabla 4.18:** Mejores tasas de rendimiento obtenidas en el escenario 4. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	98 $\pm$ 09	94 $\pm$ 10	75 $\pm$ 21	89 $\pm$ 10
Especificidad	95 $\pm$ 06	81 $\pm$ 31	76 $\pm$ 31	84 $\pm$ 09
Sensibilidad	92 $\pm$ 21	74 $\pm$ 29	61 $\pm$ 33	76 $\pm$ 13
Medida F	94 $\pm$ 20	78 $\pm$ 25	61 $\pm$ 29	77 $\pm$ 14
MCC	85 $\pm$ 24	52 $\pm$ 29	40 $\pm$ 31	59 $\pm$ 19
Diff. esc. 2	-4	0	-4	-3



**Figura 4.16:** Contornos musicales y señales de alerta del escenario 4 para grabaciones individuales con parámetros descritos en la tabla 4.14 y la mejor combinación de valores para  $\alpha$  y  $\beta$ .

**Tabla 4.19:** Parámetros del escenario 5.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Armónico	4	Mahalanobis R.	[0.90, 1.10]	200	Fijo
Ch. Alto	Armónico	1	Mahalanobis R.	[2.00, 3.00]	100	Fijo
Ch. Bajo	Arm. impar	5	Mahalanobis R.	[0.95, 1.75]	100	Fijo

**Tabla 4.20:** Rendimiento promedio MCC del cucú en escenario 5 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	3	5	5	6	6	6	6
0.003	4	5	6	6	7	7	7
0.010	4	6	6	8	8	8	8
0.030	4	5	6	8	10	10	10
0.090	2	2	2	21	27	27	27
0.300	0	0	0	69	<b>76</b>	72	72
0.900	0	0	0	0	0	0	0

### 4.3.5 Escenario 5: matrices de covarianza reales

El escenario 5 usa los tipos de núcleo, bancos de núcleos y umbrales originales, pero la distancia usada es la de Mahalanobis, con las matrices de covarianza reales calculadas en la sección 3.2.4. La tabla 4.19 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.17 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Manteniendo la configuración descrita, se calculó el coeficiente de correlación de Matthew haciendo un barrido de valores  $\alpha$  y  $\beta$  para encontrar buenos umbrales contra los cuales comparar el rendimiento alcanzado en el escenario 2. Al igual que en el escenario 3, los valores del barrido se distribuyeron logarítmicamente. Las tablas 4.20, 4.21 y 4.22 muestran el resultado de los barridos, por tipo de APS. La tabla 4.23 muestra que las mejores combinaciones de los umbrales permitieron alcanzar un MCC de 77% para el cucú ( $\alpha = 0.3$ ,  $\beta = 0.09$ ), 64% para el chirrido alto ( $\alpha = 0.003$ ,  $\beta = 0.3$ ), y 48% para el chirrido bajo ( $\alpha = 0.01$ ,  $\beta = 0.09$ ). El MCC general fue de 63%, es decir, un 2% mayor al escenario 2, y la desviación estándar fue de 12%. Los chirridos altos y bajos fueron particularmente beneficiados, pues su MCC mejoró en 12% y 7%, respectivamente, pero el cucú tuvo un retroceso de  $-12\%$ . Por lo tanto, se justifica el uso de la distancia de Mahalanobis con matriz de covarianza real solamente para los chirridos.

**Tabla 4.21:** Rendimiento promedio MCC del chirrido alto en escenario 5 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

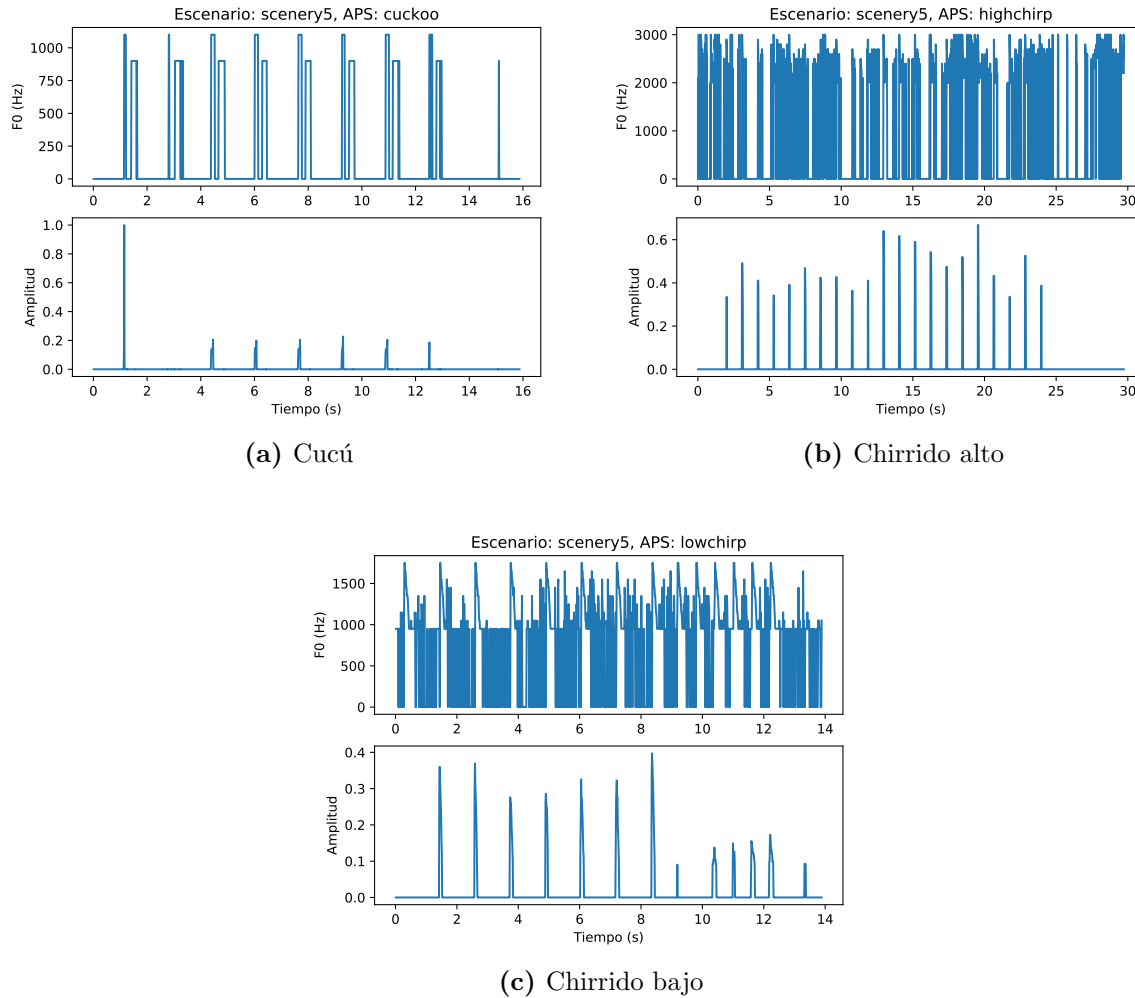
$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	17	17	17	21	35	<b>63</b>	0
0.003	17	17	17	21	36	<b>63</b>	0
0.010	17	17	17	21	40	<b>63</b>	0
0.030	17	17	17	27	52	57	0
0.090	17	17	17	47	54	34	0
0.300	17	17	17	13	5	3	0
0.900	17	17	17	0	0	0	0

**Tabla 4.22:** Rendimiento promedio MCC del chirrido bajo en escenario 5 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	2	2	2	1	45	41	0
0.003	2	2	2	2	45	41	0
0.010	2	2	2	-2	<b>47</b>	40	0
0.030	2	2	1	4	<b>47</b>	39	0
0.090	2	2	1	6	30	23	0
0.300	2	2	2	1	5	1	0
0.900	2	2	2	0	0	0	0

**Tabla 4.23:** Mejores tasas de rendimiento obtenidas en el escenario 5. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	94 $\pm$ 07	98 $\pm$ 05	77 $\pm$ 19	90 $\pm$ 10
Especificidad	70 $\pm$ 21	93 $\pm$ 15	79 $\pm$ 27	81 $\pm$ 10
Sensibilidad	99 $\pm$ 07	76 $\pm$ 26	68 $\pm$ 31	81 $\pm$ 13
Medida F	96 $\pm$ 06	82 $\pm$ 22	68 $\pm$ 30	82 $\pm$ 12
MCC	77 $\pm$ 19	64 $\pm$ 23	48 $\pm$ 36	63 $\pm$ 12
Diff. esc. 2	-12	12	7	2



**Figura 4.17:** Contornos musicales y señales de alerta del escenario 5 para grabaciones individuales con parámetros descritos en la tabla 4.19 y la mejor combinación de valores para  $\alpha$  y  $\beta$ .



**Tabla 4.24:** Parámetros del escenario 6.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Armónico	4	Mahalanobis S.	[0.90, 1.10]	200	Fijo
Ch. Alto	Armónico	1	Mahalanobis S.	[2.00, 3.00]	100	Fijo
Ch. Bajo	Arm. impar	5	Mahalanobis S.	[0.95, 1.75]	100	Fijo

**Tabla 4.25:** Rendimiento promedio MCC del cucú en escenario 6 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$							
	0.001	0.003	0.01	0.03	0.09	0.3	0.9	
0.001	0	0	0	0	54	3	0	
0.003	0	0	0	0	56	3	0	
0.010	0	0	0	0	60	5	0	
0.030	0	0	0	0	80	14	0	
0.090	0	0	0	0	<b>86</b>	37	4	
0.300	0	0	0	0	63	51	12	
0.900	0	0	0	0	0	0	0	

### 4.3.6 Escenario 6: matrices de covarianza sintéticas

El escenario 6 usa los tipos de núcleo, bancos de núcleos y umbrales originales, pero la distancia usada es la de Mahalanobis, con las matrices de covarianza sintéticas creadas en la sección 3.2.5. La tabla 4.24 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.18 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Manteniendo la configuración descrita, se calculó el coeficiente de correlación de Matthew haciendo un barrido de valores  $\alpha$  y  $\beta$  para encontrar buenos umbrales contra los cuales comparar el rendimiento alcanzado en el escenario 2. Al igual que en el escenario 3, los valores del barrido se distribuyeron logarítmicamente. Las tablas 4.25, 4.26 y 4.27 muestran el resultado de los barridos, por tipo de APS. La tabla 4.28 muestra que las mejores combinaciones de umbrales permitieron alcanzar un MCC de 87% para el cucú ( $\alpha = 0.09$ ,  $\beta = 0.09$ ), 63% para el chirrido alto ( $\alpha = 0.03$ ,  $\beta = 0.3$ ), y un 56% para el chirrido bajo ( $\alpha = 0.003$ ,  $\beta = 0.3$ ). El MCC general fue de 69%, es decir, un 8% mayor al escenario 2, y la desviación estándar fue de 14%. Los chirridos alto y bajo fueron particularmente beneficiados, con un incremento en el MCC de 11% y 15%, respectivamente, pero el cucú tuvo un 2% menos de rendimiento. Por lo tanto, se justifica el uso de la distancia de Mahalanobis con matriz de covarianza sintética solamente para los chirridos.

**Tabla 4.26:** Rendimiento promedio MCC del chirrido alto en escenario 6 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

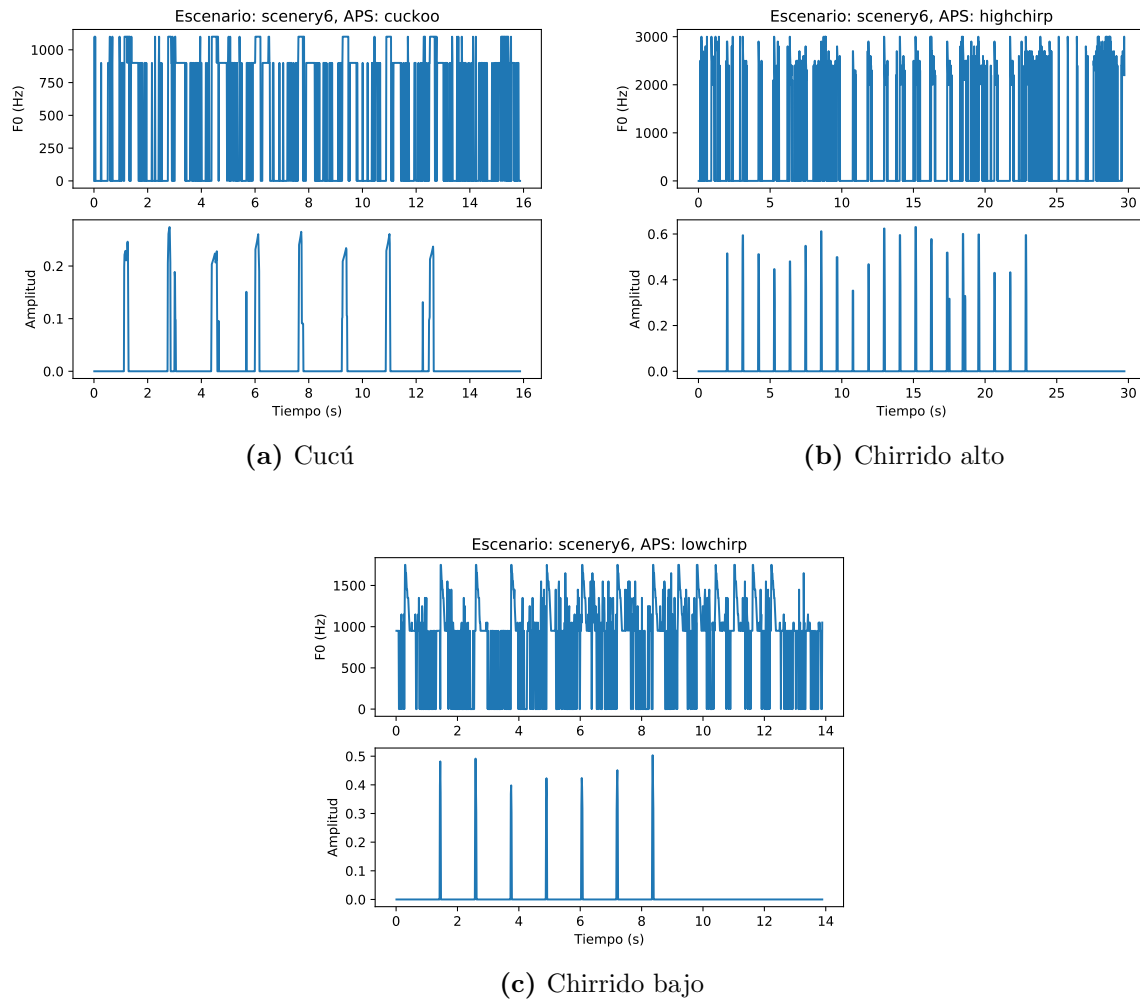
$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	17	17	17	18	27	57	0
0.003	17	17	17	18	28	57	0
0.010	17	17	17	19	32	59	0
0.030	17	17	17	24	42	<b>62</b>	0
0.090	17	17	17	45	<b>62</b>	47	0
0.300	17	17	17	18	11	4	0
0.900	17	17	17	0	0	0	0

**Tabla 4.27:** Rendimiento promedio MCC del chirrido bajo en escenario 6 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.001	2	2	2	4	11	<b>55</b>	0
0.003	2	2	2	4	11	<b>55</b>	0
0.010	2	2	2	3	14	54	0
0.030	2	2	3	4	29	51	0
0.090	2	2	-2	14	36	29	0
0.300	2	2	10	6	9	1	0
0.900	2	2	0	0	0	0	0

**Tabla 4.28:** Mejores tasas de rendimiento obtenidas en el escenario 6. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	97 $\pm$ 06	97 $\pm$ 08	86 $\pm$ 23	93 $\pm$ 06
Especificidad	87 $\pm$ 18	88 $\pm$ 28	100 $\pm$ 01	91 $\pm$ 07
Sensibilidad	98 $\pm$ 13	79 $\pm$ 27	54 $\pm$ 37	77 $\pm$ 18
Medida F	97 $\pm$ 10	83 $\pm$ 22	61 $\pm$ 40	80 $\pm$ 15
MCC	87 $\pm$ 18	63 $\pm$ 29	56 $\pm$ 37	69 $\pm$ 14
Diff. esc. 2	-2	11	15	8



**Figura 4.18:** Contornos musicales y señales de alerta del escenario 6 para grabaciones individuales con parámetros descritos en la tabla 4.24 y la mejor combinación de valores para  $\alpha$  y  $\beta$ .

**Tabla 4.29:** Parámetros del escenario 7.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Armónico	4	Proporción	[0.90, 1.10]	200	TS2Means
Ch. Alto	Armónico	1	L2 mod.	[2.00, 3.00]	100	TS2Means
Ch. Bajo	Arm. impar	5	L2 mod.	[0.95, 1.75]	100	TS2Means

**Tabla 4.30:** Rendimiento promedio MCC del cucú en escenario 7 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
Automático	38	38	38	38	42	58	<b>87</b>

### 4.3.7 Escenario 7: TS2Means

El escenario 7 usa los tipos de núcleo, bancos de núcleos y distancias originales, pero el umbral de tono  $\alpha$  es determinado dinámicamente por el TS2Means. La tabla 4.29 muestra que el detalle de las configuraciones para cada tipo de sonido y la figura 4.19 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Manteniendo la configuración descrita, se calculó el coeficiente de correlación de Matthew haciendo un barrido de valores  $\beta$  para encontrar buenos umbrales contra los cuales comparar el rendimiento alcanzado en el escenario 2. Al igual que en el escenario 3, los valores del barrido se distribuyeron logarítmicamente. Las tablas 4.30, 4.31 y 4.32 muestran el resultado de los barridos, por tipo de APS. La tabla 4.33 muestra que las mejores combinaciones de los umbrales permitieron alcanzar un MCC de 88% para el cucú ( $\beta = 0.9$ ), 53% para el chirrido alto ( $\beta = 0.003$ ), y 46% para el chirrido bajo ( $\beta = 0.03$ ). El rendimiento MCC general fue de 62%, es decir, un 1% superior al escenario 2, y la desviación estándar fue de 19%. A diferencia del chirrido bajo, que obtuvo una mejora del 5%, el cucú y el chirrido alto no fueron beneficiados significativamente, por lo que se justifica el uso del umbral dinámico TS2Means solo para el chirrido bajo. Este resultado sorprende, porque la claridad de los contornos musicales en la figura 4.19 es la mejor obtenida en este trabajo; sin embargo, podría ser que el filtrado frecuencial-temporal sea tan específico que no tolere ruido en los contornos musicales y esto haga que incremente el número de entradas nulas, lo cual es penalizado por la distancia euclidiana modificada. Las figuras 4.19a y 4.19b confirman que hay un aumento en la cantidad de entradas nulas de los contornos musicales del cucú y del chirrido alto.

**Tabla 4.31:** Rendimiento promedio MCC del chirrido alto en escenario 7 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

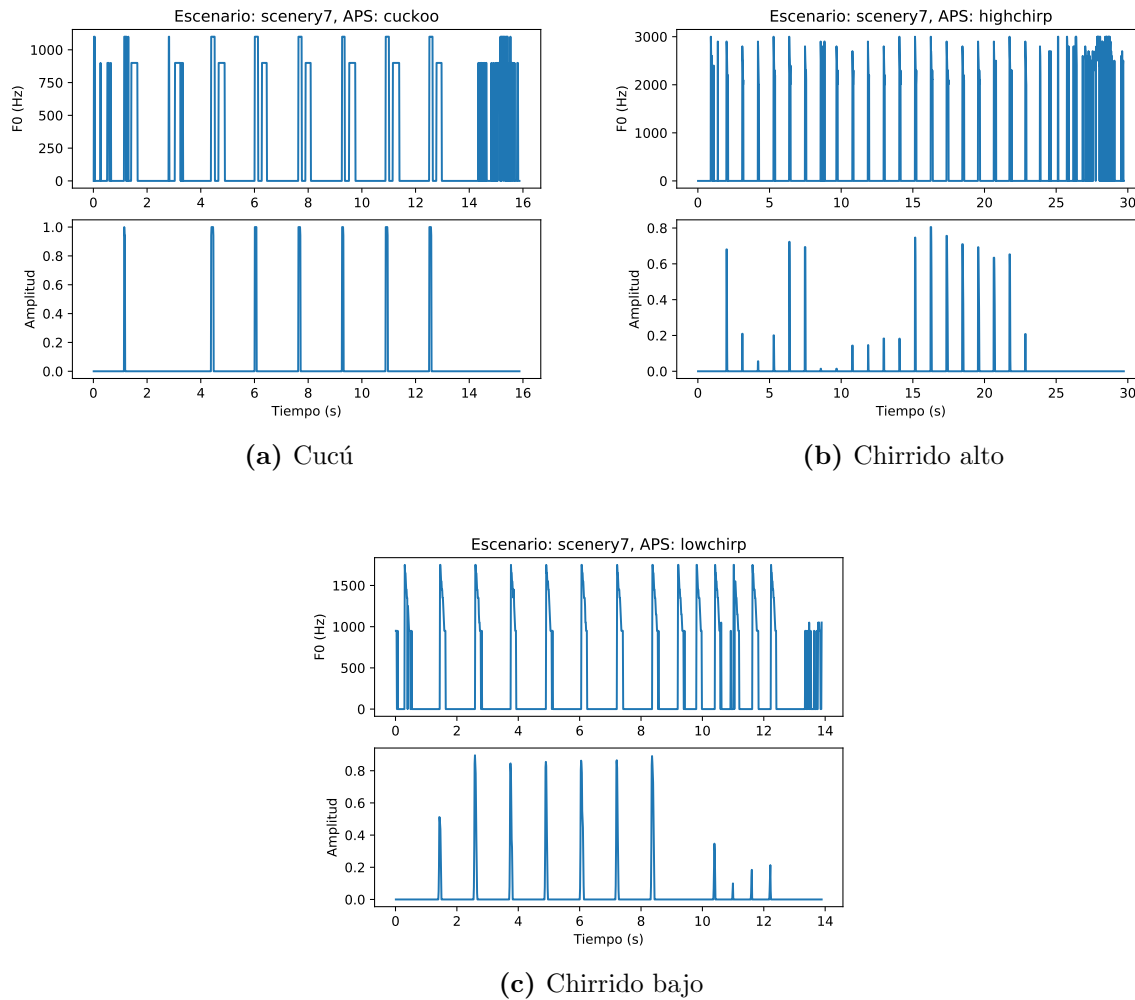
$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
Automático	<b>52</b>	<b>52</b>	<b>52</b>	50	46	31	0

**Tabla 4.32:** Rendimiento promedio MCC del chirrido bajo en escenario 7 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\alpha$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
Automático	42	42	42	<b>45</b>	39	27	1

**Tabla 4.33:** Mejores tasas de rendimiento obtenidas en el escenario 7. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	100 $\pm$ 01	97 $\pm$ 08	86 $\pm$ 26	94 $\pm$ 07
Especificidad	97 $\pm$ 04	97 $\pm$ 07	94 $\pm$ 17	96 $\pm$ 02
Sensibilidad	94 $\pm$ 14	63 $\pm$ 25	47 $\pm$ 32	68 $\pm$ 20
Medida F	96 $\pm$ 10	74 $\pm$ 22	56 $\pm$ 31	75 $\pm$ 17
MCC	88 $\pm$ 20	53 $\pm$ 25	46 $\pm$ 29	62 $\pm$ 19
Diff. esc. 2	-1	1	5	1



**Figura 4.19:** Contornos musicales y señales de alerta del escenario 7 para grabaciones individuales con parámetros descritos en la tabla 4.29 y la mejor combinación de valores para  $\beta$ .

**Tabla 4.34:** Parámetros del escenario 8.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Armónico	4	Proporción	[0.90, 1.10]	200	EMA
Ch. Alto	Armónico	1	L2 mod.	[2.00, 3.00]	100	EMA
Ch. Bajo	Arm. impar	5	L2 mod.	[0.95, 1.75]	100	EMA

**Tabla 4.35:** Rendimiento promedio MCC del cucú en escenario 8 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	0	0	0	0	0	9	83
0.900	0	0	0	0	0	10	93
0.950	0	0	0	0	0	12	<b>96</b>
0.975	0	0	0	0	0	14	95
0.987	0	0	0	0	0	25	95
0.993	1	1	1	1	5	34	94
0.996	4	4	4	4	9	40	94

### 4.3.8 Escenario 8: EMA

El escenario 8 usa los tipos de núcleo, bancos de núcleos y distancias originales, pero el umbral de tono  $\alpha$  es determinado dinámicamente por la EMA. La tabla 4.34 muestra el detalle de las configuraciones para cada tipo de sonido y la figura 4.20 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Como puede apreciarse, la claridad de los contornos musicales no es tan buena como la obtenida al usar el TS2Means. Manteniendo la configuración descrita, se calculó el coeficiente de correlación de Matthew haciendo un barrido de valores  $\lambda$  y  $\beta$  para encontrar buenos umbrales contra los cuales comparar el rendimiento alcanzado en el escenario 2. Al igual que en el escenario 3, los valores  $\beta$  del barrido se distribuyeron logarítmicamente, y en el caso de  $\lambda$ , sigue la distribución  $5 \cdot 2^x$ , que permite obtener una cantidad de vecinos incrementable como: 5, 10, 20, 40, 80, 160 y 320. Las tablas 4.35, 4.36 y 4.37 muestran el resultado de los barridos, por tipo de APS. La tabla 4.38 muestra que las mejores combinaciones de umbrales permitieron alcanzar un MCC de 97% para el cucú ( $\lambda = 0.95$ ,  $\beta = 0.9$ ), un 78% para el chirrido alto ( $\lambda = 0.987$ ,  $\beta = 0.03$ ), y un 64% para el chirrido bajo ( $\lambda = 0.996$ ,  $\beta = 0.09$ ). El rendimiento MCC general fue de 80%, es decir, un 19% mayor al escenario 2 (el mejor obtenido en el estudio), y la desviación estándar fue de 14%. Como la mejora es sustancial y generalizada, se justifica el uso de la EMA para todos los sonidos.

**Tabla 4.36:** Rendimiento promedio MCC del chirrido alto en escenario 8 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	57	57	57	55	49	31	0
0.900	67	67	68	69	67	52	0
0.950	74	73	75	<b>77</b>	74	64	0
0.975	73	74	75	76	75	69	0
0.987	75	76	<b>77</b>	<b>77</b>	75	70	0
0.993	75	75	76	76	76	70	0
0.996	74	74	75	75	76	69	0

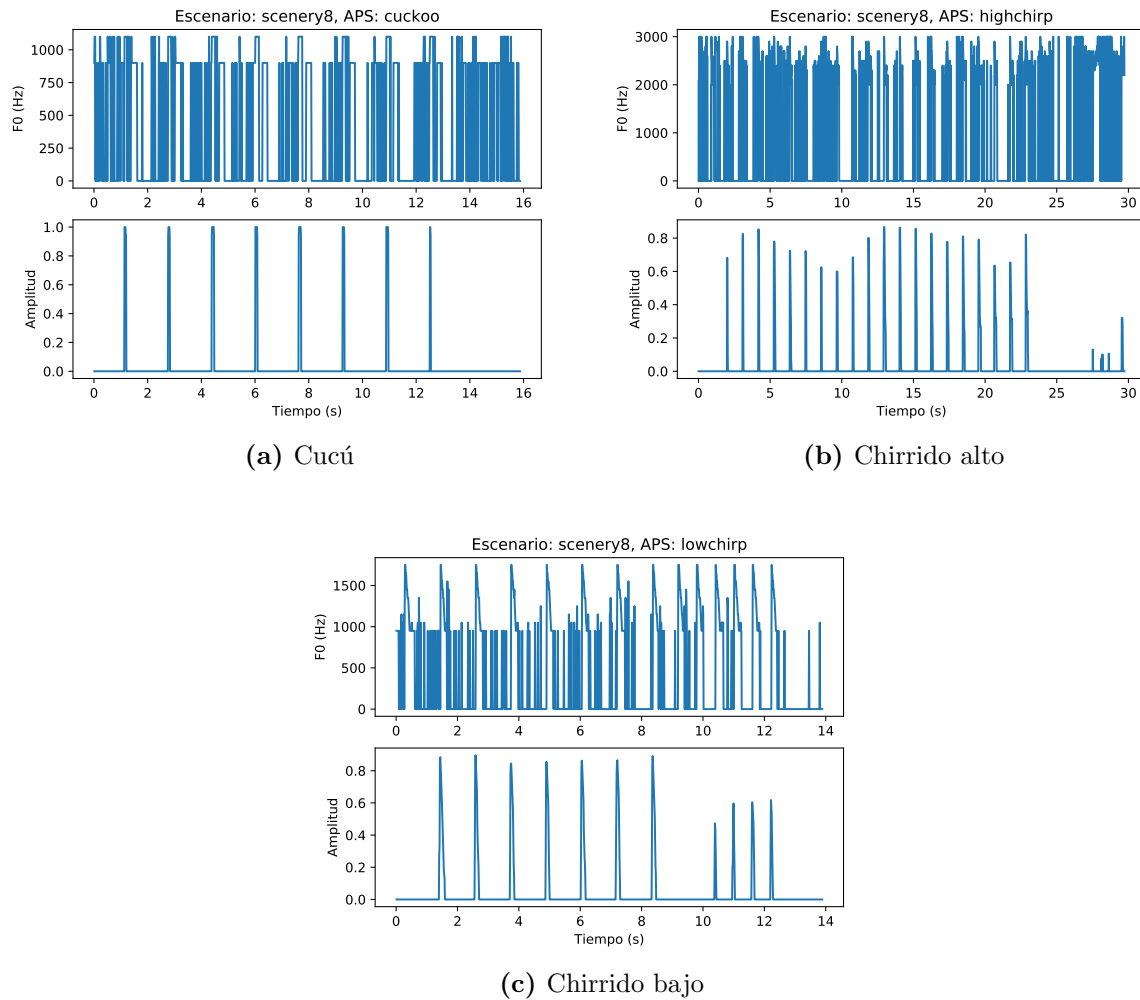
**Tabla 4.37:** Rendimiento promedio MCC del chirrido bajo en escenario 8 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	29	29	26	34	25	8	0
0.900	41	43	42	46	44	29	3
0.950	50	53	59	56	58	44	6
0.975	52	52	56	61	55	47	17
0.987	49	50	53	59	53	49	17
0.993	56	56	56	58	58	57	18
0.996	56	55	57	59	<b>63</b>	59	20

**Tabla 4.38:** Mejores tasas de rendimiento obtenidas en el escenario 8. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	100 $\pm$ 01	97 $\pm$ 05	81 $\pm$ 20	93 $\pm$ 09
Especificidad	97 $\pm$ 04	91 $\pm$ 13	79 $\pm$ 23	89 $\pm$ 08
Sensibilidad	100 $\pm$ 03	90 $\pm$ 14	85 $\pm$ 19	92 $\pm$ 07
Medida F	100 $\pm$ 02	93 $\pm$ 10	82 $\pm$ 18	92 $\pm$ 08
MCC	97 $\pm$ 09	78 $\pm$ 17	64 $\pm$ 35	80 $\pm$ 14
Diff. esc. 2	8	26	23	19





**Figura 4.20:** Contornos musicales y señales de alerta del escenario 8 para grabaciones individuales con parámetros descritos en la tabla 4.34 y la mejor combinación de valores para  $\lambda$  y  $\beta$ .

**Tabla 4.39:** Parámetros del escenario 9.

APS	Núcleo	Arm.	Distancia	$[f_{\min}, f_{\max}]$ (kHz)	$df$ (Hz)	$\alpha$
Cucú	Armónico	4	Proporción	[0.90, 1.10]	200	EMA
Ch. Alto	Propuesto	3	Mahalanobis S.	[2.00, 3.00]	100	EMA
Ch. Bajo	Arm. impar	5	Mahalanobis S.	[0.95, 1.75]	100	EMA

**Tabla 4.40:** Rendimiento promedio MCC del cucú en escenario 9 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	0	0	0	0	0	9	83
0.900	0	0	0	0	0	10	93
0.950	0	0	0	0	0	12	<b>96</b>
0.975	0	0	0	0	0	14	95
0.987	0	0	0	0	0	25	95
0.993	1	1	1	1	5	34	94
0.996	4	4	4	4	9	40	94

### 4.3.9 Escenario 9: las mejores combinaciones

Como se determinó en los escenarios anteriores, las mejoras más importantes fueron obtenidas con el uso del núcleo propuesto para el chirrido alto, el uso de la distancia de Mahalanobis con matriz de covarianza sintética para los chirridos, y el uso de la EMA para todos los sonidos. La tabla 4.39 muestra la configuración del escenario final. La figura 4.21 muestra tres demostraciones del procesamiento realizado por el sistema en grabaciones individuales. Las tablas 4.40, 4.41 y 4.42 muestran el resultado de los barridos, por tipo de APS. La tabla 4.43 muestra que las mejores combinaciones de umbrales permitieron alcanzar un MCC de 97% para el cucú ( $\lambda = 0.95$ ,  $\beta = 0.9$ ), un 68% para el chirrido alto ( $\lambda = 0.975$ ,  $\beta = 0.3$ ), y un 66% para el chirrido bajo ( $\lambda = 0.993$ ,  $\beta = 0.09$ ). Se destaca que los valores  $\lambda$  para los sonidos cucú y los chirridos son similares a los usados en el escenario 8, lo que es favorable porque significa que no siempre necesitan ser optimizados. El rendimiento general obtenido fue de 77%, un 16% más alto que el reportado en el escenario 2 (el segundo más alto en el estudio), y la desviación estándar fue de 15%. Para fines comparativos, la tabla 4.44 muestra el resumen de los resultados obtenidos en este escenario y los anteriores.

**Tabla 4.41:** Rendimiento promedio MCC del chirrido alto en escenario 9 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

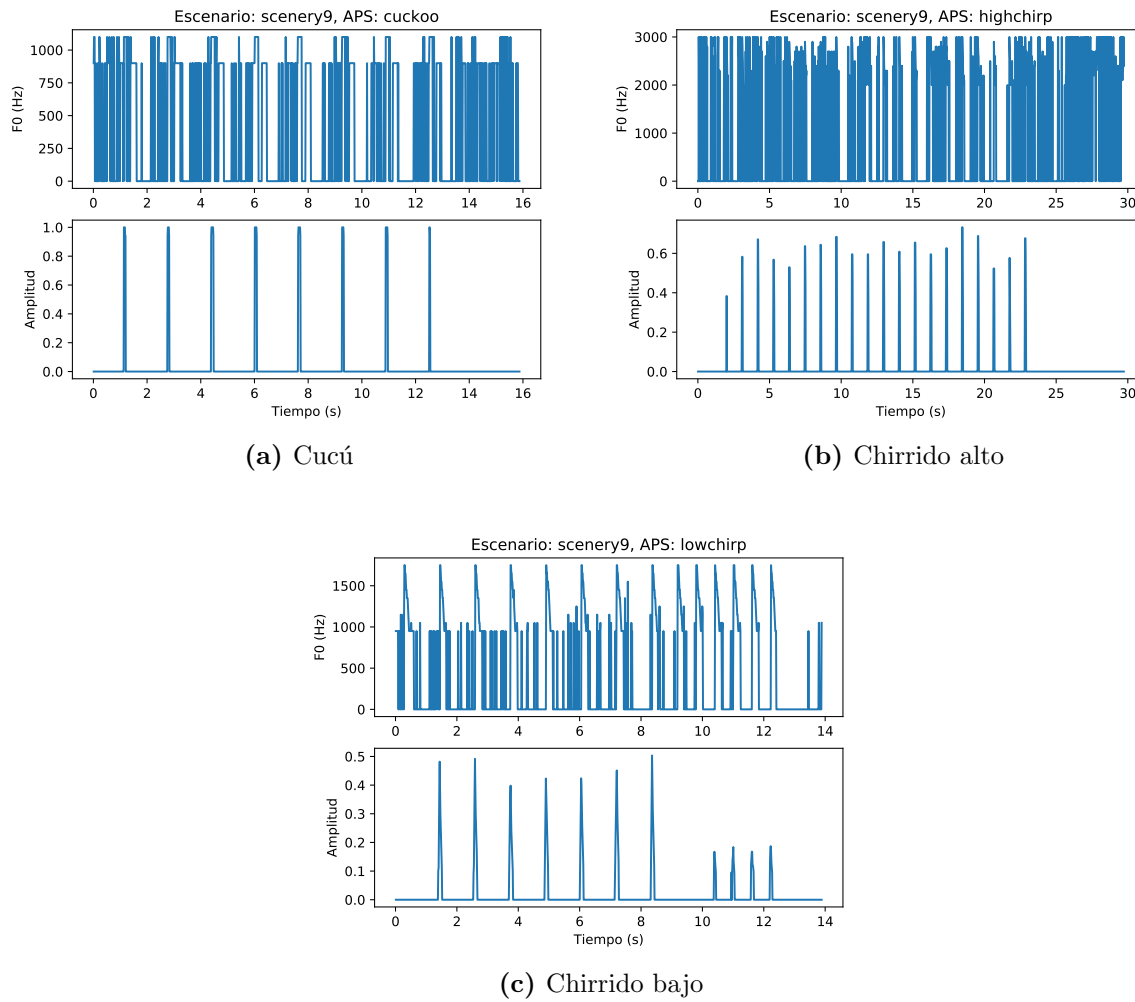
$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	17	17	17	16	46	36	0
0.900	17	17	17	16	43	57	0
0.950	17	17	17	15	43	66	0
0.975	17	17	17	15	43	<b>67</b>	0
0.987	17	17	17	15	43	<b>67</b>	0
0.993	17	17	17	14	44	66	0
0.996	17	17	17	15	46	65	0

**Tabla 4.42:** Rendimiento promedio MCC del chirrido bajo en escenario 9 y distintas combinaciones de umbrales ( $\pm 1\%$  de error).

$\lambda$	$\beta$						
	0.001	0.003	0.01	0.03	0.09	0.3	0.9
0.800	2	2	2	26	7	2	0
0.900	2	2	2	27	34	21	0
0.950	2	2	2	16	49	28	0
0.975	2	2	3	16	56	41	0
0.987	2	2	3	17	58	44	0
0.993	2	2	3	19	<b>65</b>	47	0
0.996	2	2	4	24	62	49	0

**Tabla 4.43:** Mejores tasas de rendimiento obtenidas en el escenario 9. La fila inferior muestra las diferencias con el escenario 2.

Métrica	Rendimiento (%)			
	Cucú	Ch. Alto	Ch. Bajo	General
Precisión	100 $\pm$ 01	99 $\pm$ 09	89 $\pm$ 19	96 $\pm$ 05
Especificidad	97 $\pm$ 04	99 $\pm$ 03	95 $\pm$ 17	97 $\pm$ 02
Sensibilidad	100 $\pm$ 03	76 $\pm$ 26	69 $\pm$ 33	82 $\pm$ 14
Medida F	100 $\pm$ 02	83 $\pm$ 23	74 $\pm$ 32	86 $\pm$ 11
MCC	97 $\pm$ 09	68 $\pm$ 24	66 $\pm$ 33	77 $\pm$ 15
Diff. esc. 2	8	16	25	16



**Figura 4.21:** Contornos musicales y señales de alerta del escenario 9 para grabaciones individuales con parámetros descritos en la tabla 4.39 y la mejor combinación de valores para  $\lambda$  y  $\beta$ .

**Tabla 4.44:** Resumen del rendimiento general obtenido en cada escenario de pruebas.

Métrica	Rendimiento (%)								
	1	2	3	4	5	6	7	8	9
Precisión	87	89	91	89	90	93	94	93	96
Especificidad	90	90	90	84	81	91	96	89	97
Sensibilidad	72	69	73	76	81	77	68	92	82
Medida F	74	73	76	77	82	80	75	92	86
MCC	64	61	62	59	63	<b>69</b>	62	<b>80</b>	<b>77</b>

# Capítulo 5

## Conclusiones

La implementación propia del algoritmo RASP, creada para el escenario 1, logró producir señales de alerta similares a las publicadas por el primer estudio, a pesar de la desviación del 10% al 20% en el rendimiento de las métricas de especificidad, sensibilidad y medida F. El procesamiento de grabaciones individuales de la sección 4.1.2 y la disminución de la sensibilidad reportada en el escenario 2, permitieron determinar que la metodología de evaluación mejorada logró penalizar los picos de alerta faltantes, aumentando la detección de falsos negativos cuando correspondía. Después de evaluado con la nueva metodología, se determinó que el rendimiento del algoritmo RASP original en términos del MCC fue del 62%, una tasa de detección usada como base para evaluar el resto de los escenarios de pruebas.

El diseño del núcleo musical propuesto en el escenario 3 mostró una mejora significativa para el chirrido alto, pero no para el cucú y el chirrido bajo, lo que indica que el diseño de núcleos debe hacerse de forma individual para cada sonido. En el caso del chirrido alto la sección 4.2 determinó que el diseño de núcleo propuesto logró disminuir la confusión de la frecuencia fundamental con sus armónicas y subarmónicas, pues las matrices de puntajes mostraron calificaciones altas en las regiones correspondientes a los contornos esperados y bajas en el resto. Por su parte, el uso del banco de núcleos unificado del escenario 4 no tuvo un impacto significativo para ningún tipo de APS, por lo que se descartó su uso.

El escenario 5 comprobó que la distancia de Mahalanobis es apta para admitir contornos musicales disjuntos, y las matrices de covarianza reales calculadas permitieron observar que los segmentos de ruido tienen una varianza alta y los segmentos de las modulaciones de frecuencia tienen una varianza baja. El escenario 6 demostró que se pueden generar matrices de covarianza sintéticas ignorando la covarianza y estableciendo la varianza con base en los parámetros del rango de frecuencias, duración del contorno musical y la resolución de frecuencias del banco de núcleos empleado. Los resultados obtenidos con las matrices sintéticas fueron mejores que los de las matrices reales, lo que hace factible que estas puedan ser empleadas en casos donde las grabaciones sean insuficientes respecto al número de entradas del contorno musical. Aún así, es necesario procesar nuevos tipos de sonidos con el fin de determinar si el vector de coeficientes  $b$ , usado para estimar la

varianza mínima mediante una combinación lineal, se mantiene sin cambio.

El escenario 7 determinó que el algoritmo de TS2Means no es compatible con la distancia euclidiana modificada, pues el filtrado realizado por TS2Means elimina las alturas musicales que hayan sido contaminadas con ruido, lo cual agrega ceros al contorno musical y hace que la penalización sea más severa. Además, la sección 2.2.4 determinó que este algoritmo no puede ser empleado por sistemas en tiempo real porque no es causal. Aún así no se puede despreciar la labor del TS2Means, que ha demostrado ser útil en otras tareas de recuperación de información musical. El escenario 8 determinó que la EMA, una versión simplificada del TS2Means, es capaz de realizar una limpieza causal y más tolerante al ruido, lo que permitió incrementar el rendimiento del algoritmo en un 19% (el mejor rendimiento del estudio).

El escenario 9 determinó que usando las mejoras individuales más significativas se puede incrementar el rendimiento MCC del algoritmo en un 16% (el segundo mejor rendimiento alcanzado en el estudio). Estas modificaciones consistieron en el uso del núcleo propuesto para el chirrido alto, el uso de la distancia de Mahalanobis con matriz de covarianza sintética para los chirridos y el uso de la EMA para todos los sonidos. Debido a que las tasas de detección obtenidas fueron más bajas que las del escenario 8, se comprueba que el rendimiento emergente no es igual a la suma de las mejoras por separado.

Como trabajo futuro, se proponen cinco tareas: la primera es realizar una prueba estadística que permita determinar con cierto grado de confianza que la covarianza de las matrices de la sección 3.2.4 es “despreciable”; la segunda es repetir la evaluación de los nueve escenarios de prueba usando un algoritmo genético, lo que permitiría considerar valores más amplios para  $\alpha$ ,  $\beta$  y  $\lambda$  que los establecidos con los incrementos estáticos; la tercera es incorporar el método de la validación cruzada para verificar que el sistema pueda mantener el rendimiento alcanzado aún cuando se analice un solo subconjunto del total de grabaciones APS; la cuarta es usar las matrices de covarianza sintéticas para procesar sonidos nuevos como sirenas de ambulancias, alarmas de patrullas policiales y bocinas del tren; y la quinta es usar curvas exponenciales (en lugar de rectas) para representar las modulaciones de frecuencia de los chirridos. A parte de las mejoras mencionadas, el proyecto RASP podría mejorar en dos aspectos: el primero es incrementar el desempeño de la aplicación móvil del CITIC incorporando la EMA en lugar de los umbrales estáticos, lo cual solo requeriría implementar la ecuación 2.17; y lo segundo, es acompañar a RASP de un módulo de procesamiento de vídeo (como el de la sección 1.1.2) para alinear al peatón frente a los cruces peatonales, evitando que usuarios no videntes crucen en el lugar incorrecto. Esto sigue siendo necesario de agregar pues los dispositivos electrónicos generadores instalados en el país aún no son unidireccionales. Además, el módulo de vídeo permitiría que el sistema sea usable incluso cuando no hay sonido (como se encontró en varios semáforos).

# Bibliografía

- [1] U.S. DEPARTMENT OF TRANSPORTATION. «Manual on Uniform Traffic Control Devices for Streets and Highways (MUTCD)». Federal Highway Administration (FHWA) (2009).
- [2] U.S. ACCESS BOARD. «Proposed Rights-of-Way Guidelines». Federal Register (2011).
- [3] RAINBOW GARDENS. San jose downtown costa rica hd, - central park (parque) walking around and grabbing a bite. URL: <https://www.youtube.com/watch?v=Nm-KMvid90o&t=9s> (2013). Última vez consultado el 30 de Abril de 2018.
- [4] INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS. «Costa Rica: Indicadores de Educación y de Contexto». Unicef (2014). Última vez consultado el 21 de Febrero del 2017.
- [5] D. AHMETOVIC. Zebraloocalizer: identification and localization of pedestrian crossings. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, ACM* páginas 275–284 (2011).
- [6] V. IVANCHENKO. Crosswatch: a camera phone system for orienting visually impaired pedestrians at traffic intersections. *International Conference on Computers for Handicapped Persons, Springer* páginas 1122–1128 (2008).
- [7] LA NACIÓN. Chile crea semáforos para los adictos a los smartphones. URL: <http://www.lanacion.com.py/mundo/2017/07/12/chile-crea-semaforos-para-los-adictos-a-los-smartphones/> (2017). Última vez consultado el 21 de Febrero del 2017.
- [8] NOEMI CHINCHILLA. Personas no videntes o con baja visión contarán con app para puntos de referencia. URL: <https://www.tec.ac.cr/hoyeneltec/2016/09/27/personas-no-videntes-baja-vision-contaran-app-puntos-referencia> (2016). Última vez consultado el 9 de Marzo de 2018.
- [9] CINTHYA ARIAS, WALTHER HERRERA, ERICK IRIGARAY, HANNY RODRÍGUEZ, RODOLFO RODRÍGUEZ, MAX RUIZ, ANA SEGURA y DAVID VARGAS. «Estadísticas del sector de telecomunicaciones. Informe 2010-2013». Superintendencia de Telecomunicaciones (2014). Última vez consultado el 21 de Febrero del 2017.

- [10] SEBASTIÁN RUIZ, ARTURO CAMACHO y JUAN M. FONSECA SOLÍS. Automatic recognition of accessible pedestrian signals. *Acoustical Society of America* **141**(5), 3913–3914 (2017).
- [11] R. GROMPONE VON GIOI, J. JAKUBOWICZ, J. M. MOREL y G. RANDALL. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32**(4), 722–732 (Abril 2010).
- [12] SIMON HAYKIN. «Adaptive Filter Theory». Prentice-Hall, Inc. (1996).
- [13] A. TAYLOR, G. WATSON, G. GRIGG y H. MCCALLUM. Monitoring frog communities: An application of machine learning. *In Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence* páginas 1564–1569 (1996).
- [14] CHRISTOPHER M. BISHOP. «Neural Networks for Pattern Recognition». Oxford University Press, Inc., New York, NY, EE.UU. (1995).
- [15] JAMES A. FREEMAN y DAVID M. SKAPURA. «Neural Networks: Algorithms, Applications, and Programming Techniques». Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, EE.UU. (1991).
- [16] A. CAMACHO, A. GARCIA RODRIGUEZ y F. BOLAÑOS. Automatic detection of vocalizations of the frog diasporus hylaeformis in audio recordings. *The Journal of the Acoustical Society of America* **130**(3) (2011).
- [17] D. K. MELLINGER, S. W. MARTIN, R. P. MORRISSEY, L. THOMAS y J. J. YOSCO. A method for detecting whistles, moans, and other frequency contour sounds. *The Journal of the Acoustical Society of America* **129**(6) (2011).
- [18] JUAN M. FONSECA SOLÍS, SEBASTIÁN RUIZ y ARTURO CAMACHO. Reconocimiento automático de señales accesibles en semáforos. *Jornadas Costarricenses de Investigación en Computación e Informática* (2017).
- [19] TINNETH MONGE ACUÑA y YISLEN SOLÍS JIMÉNEZ. El síndrome de caídas en personas adultos mayores y su relación con la velocidad de marcha. *Revista Médica de Costa Rica y Centroamérica* **LXXIII**(618), 91–95 (2016). Última vez consultado el 12 de Diciembre del 2017.
- [20] KRISTIN C. BROWN, HEATHER M. HANSON, FLAVIO FIRMANI, DANMEI LIU, MEGAN M. MCALLISTER, KHALIL MERALI, JOSEPH H. PUYAT y MAUREEN C. ASHE. Gait speed and variability for usual pace and pedestrian crossing conditions in older adults using the gaitrite walkway. *Gerontology and Geriatric Medicine* (2015). Última vez consultado el 12 de Diciembre del 2017.
- [21] JUAN M. FONSECA SOLÍS y ARTURO CAMACHO. Diseño, implementación y evaluación de la aplicación rasp, extensión del informe final de investigación n.º 326-b6-146.



- Centro de Investigaciones en Tecnologías de Información y Comunicación, Universidad de Costa Rica* (2017).
- [22] CLEVE MOLE. «Numerical Computing with MATLAB». The MathWorks, Inc., Natick, Massachusetts (2014). Última vez consultado el 24 de Febrero del 2017.
- [23] ARTURO CAMACHO y JOHN G. HARRIS. «A sawtooth waveform inspired pitch estimator for speech and music». Tesis Doctoral, (2008).
- [24] P. PRANDONI y M. VETTERLI. «Signal Processing for Communications». EPFL Press (2008).
- [25] ARTURO CAMACHO LOZANO. Tarea 4: Algoritmos de estimación de altura. *MP6154 – Procesamiento de Sonido, Instituto Tecnología de Costa Rica* (2017).
- [26] P. ALVARADO. «Señales y Sistemas Fundamentos Matemáticos». Instituto Tecnológico de Costa Rica (2008).
- [27] ANIRUDDH D. PATEL y EVAN BALABAN. Human pitch perception is reflected in the timing of stimulus-related cortical activity. *Nature Neuroscience* **4**, 839 (2001).
- [28] B. MOORE y J. BRILL. «An Introduction to the Psychology of Hearing». The Journal of the Acoustical Society of America, 6 edición (2013).
- [29] B.R. GLASBERG y B.C.J. MOORE. Derivation of auditory filter shapes from notched-noise data. *Hearing Research* **47**, 103–138 (1990).
- [30] J. FONSECA. Reconocimiento automático de la altura musical en la interfaz de un juego de computadora. *Ingeniería, revista de la Universidad de Costa Rica* **25**(1) (2015).
- [31] WIKIPEDIA CONTRIBUTORS. Uncoiled cochlea with basilar membrane. URL: [https://commons.wikimedia.org/wiki/File:Uncoiled\\_cochlea\\_with\\_basilar\\_membrane.png](https://commons.wikimedia.org/wiki/File:Uncoiled_cochlea_with_basilar_membrane.png) (2018). Última vez consultado el 28 de Marzo de 2018.
- [32] M. R. SCHROEDER. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America* **43**(4), 829–834 (1968).
- [33] DIK J. HERMES. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America* **83**(1), 257–264 (1988).
- [34] XUEJING SUN. A pitch determination algorithm based on subharmonic-to-harmonic ratio. *the 6th International Conference of Spoken Language Processing* páginas 676–679 (2000).
- [35] L. RABINER. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **25**(1), 24–33 (Feb 1977).

- [36] YANG-HANN KIM. «Sound Propagation: An Impedance Based Approach». Wiley (2010).
- [37] WIKIPEDIA CONTRIBUTORS. Equal-loudness contour. URL: [https://en.wikipedia.org/w/index.php?title=Equal-loudness\\_contour&oldid=816516870](https://en.wikipedia.org/w/index.php?title=Equal-loudness_contour&oldid=816516870) (2018). Última vez consultado el 26 de Febrero del 2017.
- [38] A. CAMACHO. Detection of pitched/unpitched sound using pitch strength clustering. *Proceedings of the Ninth International Conference on Music Information Retrieval* páginas 533–537 (2008).
- [39] RICHARD O. DUDA, PETER E. HART y DAVID G. STORK. «Pattern Classification». Wiley-Interscience (2000).
- [40] J. G. PROAKIS y D. G. MANOLAKIS. «Tratamiento Digital de Señales». Prentice Hall (1998).
- [41] JOSÉ PABLO ALVARADO MOYA. «Procesamiento Digital de Señales», tomo 1. Instituto Tecnológico de Costa Rica (2011).
- [42] THOMAS CORMEN, CHARLES E. LEISERSON, RONALD L. RIVEST y CLIFFORD STEIN. «Introduction to Algorithms». MIT Press and McGraw-Hill (2001).
- [43] ALFRED V. AHO. «Data Structures and Algorithms». Addison-Wesley Longman Publishing Co., Inc. Boston, MA, EE.UU. (1983).
- [44] JOHN L. KELLEY. «Topología general», tomo 27. Springer-Verlag New York (1975).
- [45] MARTA MACHO STADLER. «Topología de espacios métricos». Universidad del País Vasco (2010).
- [46] M. VETTERLI y K. KOVACEVIĆ. «Wavelets and Subband Coding». Originalmente publicado por Prentice Hall (2007).
- [47] STANLEY I. GROSSMAN. «Álgebra lineal». McGrawHill, 6 edición (2008).
- [48] JUAN CARLOS BRICEÑO LOBO. «CI1210 - Probabilidad y estadística». Escuela de Ciencias de la Computación e Informática, Universidad de Costa Rica (2009).
- [49] RICHARD G. BRERETON. The mahalanobis distance and its relationship to principal component scores. *Journal of Chemometrics* **29**(3), 143–145 (2015).
- [50] L. SIROVICH y M. KIRBY. Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A* **4**(3), 519–524 (Mar 1987).
- [51] KARL PEARSON F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572 (1901).

- [52] LINDSAY I SMITH. A tutorial on principal components analysis. *Cornell University* (2002).
- [53] YVONNE MOH y JOACHIM M. BUHMANN. Regularized online learning of pseudo-metrics. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* páginas 1990–1993 (2010).
- [54] D. C. HOYLE. Accuracy of pseudo-inverse covariance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(7), 1470–1481 (Julio 2011).
- [55] ROBERT E. HARTWIG. Singular value decomposition and the moore-penrose inverse of bordered matrices. *SIAM Journal on Applied Mathematics* **31**(1), 31–41 (1976).
- [56] RICHARD SZELISKI. Computer Vision : Algorithms and Applications. *Computer* **5**, 832 (2010).
- [57] EDEL GARCÍA. «Singular Value Decomposition (SVD). A fast track tutorial». Universidad de Lisboa (2006). Última vez consultado por el 10 de Febrero de 2018.
- [58] PABLO ALVARADO MOYA. «Descomposición en Valores Singulares». Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica (2013). Última vez consultado el 10 de Febrero de 2018.
- [59] YADIRA SOLANO. Improvement of convergence speed and generalization capability in artificial neural networks by normalizing the backpropagation algorithm. *Graduate School of Science and Technology, Chiba University* (1997).
- [60] ANDREAS DE RUITER (MICROSOFT). Performance measures in azure ml: Accuracy, precision, recall and f1 score. URL: <https://blogs.msdn.microsoft.com/andreasderuiter/2015/02/09/performance-measures-in-azure-ml-accuracy-precision-recall-and-f1-score/> (2015). Última vez consultado el 14 de Marzo de 2018.
- [61] A. ALEXANDRIDIS, E. CHONDRODIMA, G. PAIVANA, M. STOGIANNOS, E. ZOIS y H. SARIMVEIS. Music genre classification using radial basis function networks and particle swarm optimization. En «2014 6th Computer Science and Electronic Engineering Conference (CEECE)», páginas 35–40 (Sept 2014).
- [62] M. SHEPPERD. How do i know whether to trust a research result? *IEEE Software* **32**(1), 106–109 (Jan 2015).
- [63] Y. LECUN, L. BOTTOU, Y. BENGIO y P. HAFFNER. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (Nov 1998).
- [64] B. I. CÎRSTEA y L. LIKFORMAN-SULEM. Tied spatial transformer networks for digit recognition. En «2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)», páginas 524–529 (Oct 2016).

- [65] HONG SU, HUI ZHANG, XUELIANG ZHANG y GUANGLAI GAO. Convolutional neural network for robust pitch determination. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2016-May**, 579–583 (2016).
- [66] SANGEUN KUM. Classification-Based Singing Melody Extraction Using Deep Convolutional Neural Networks. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* **1**(Noviembre), 1–14 (2017).
- [67] RACHEL M BITTNER, BRIAN MCFEE, JUSTIN SALAMON, PETER LI y JUAN P BELLO. Deep Saliency Representations for F0 Estimation in Polyphonic Music. *18th International Society for Music Information Retrieval Conference, Suzhou, China, 2017* (2017).
- [68] MCKENNA R. LOVEJOY AND MARK A. WICKERT. Using the Ipython notebook as the computing for signals and systems courses. *2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)* guages páginas 289–294 (2015).
- [69] DAVID GOLDBERG. What every computer scientist should know about floating-point arithmetic. *Association for Computing Machinery, Inc* (1991).
- [70] IEEE. Standard for binary floating-point arithmetic. *ANSI/IEEE Std 754-1985* (1985).
- [71] ARIEL SOLÓRZANO. «Compendio Ambiental». Estado de la Nación (2017). Última vez consultado el 25 de Diciembre de 2017.
- [72] YU TSUMURA. Positive definite real symmetric matrix and its eigenvalues. URL: <https://yutsumura.com/positive-definite-real-symmetric-matrix-and-its-eigenvalues>. Última vez consultado el 29 de Abril de 2018.

# Apéndice A

## Datos usados en la comparación de distancias

La tabla A.1 muestra los datos utilizados para realizar la comparación entre distancias en la sección 2.4.6. Estos tratan del viento (en kilómetros por hora) y la precipitación acumulada anual (en mililitros) del Aeropuerto Juan Santamaría durante los años 1990 a 2016, y fueron tomados del Informe del Estado de la Nación [71]. La matriz de covarianza de la muestra y su matriz inversa se definen como sigue:

$$S = \begin{pmatrix} 4.3 & -74.3 \\ -74.3 & 153840.8 \end{pmatrix}$$

y

$$S^{-1} = \begin{pmatrix} 0.23322 & 0.00011 \\ 0.00011 & 0.00001 \end{pmatrix},$$

donde se observa que efectivamente, la diagonal de la matriz de covarianza corresponde con las varianzas de los datos, pues:  $\sqrt{4.3} = 2.1 = s_{0,0}$  y  $\sqrt{153840.8} = 392.2 \approx 399.7 = s_{1,1}$ . Curiosamente la covarianza es negativa, lo que indica que a mayor viento menor precipitación anual. ¿Será que los fuertes vientos alejan a las nubes cargadas de agua?

**Tabla A.1:** Velocidad promedio anual del viento y precipitación anual acumulada (PAA) del Aeropuerto Juan Santamaría desde 1990 hasta 2016.

Año	Viento (km/h)	PAA (ml)
1990	18.3	1941.7
1991	17.3	1624.7
1992	15.8	1950.7
1993	12.9	2106.4
1994	14.0	1527.3
1995	12.0	2638.4
1996	13.9	2210.1
1997	16.7	1827.3
1998	22.2	2301.7
1999	16.9	1946.7
2000	15.6	1524.1
2001	15.9	1286.6
2002	17.4	1509.6
2003	13.5	2328.1
2004	15.3	1598.9
2005	13.2	1704.9
2006	14.6	1972.2
2007	13.8	2227.4
2008	13.5	2360.6
2009	15.7	1165.8
2010	13.9	2191.4
2011	13.0	1400.5
2012	14.3	1265.8
2013	14.2	1729.5
2014	14.5	1475.6
2015	16.1	1376.8
2016	13.7	1354.7
Mínimo	12.0	1165.8
Máximo	22.2	2638.4
Promedio	15.1	1798.1
Desv. estándar	2.1	399.7

## Apéndice B

# Demostración de la equivalencia entre la distancia euclidiana sobre datos decorrelacionados y la distancia de Mahalanobis

Al calcular la distancia Mahalanobis entre  $x = [x^{(1)}, \dots, x^{(M)}]^T$ , un conjunto de  $M$  experimentos del vector aleatorio en sentido amplio  $X$  (WSS, por sus siglas en inglés), y  $\mu_X$ , se está aplicando un *análisis de componentes principales* (PCA, por sus siglas en inglés) sobre  $x$  y calculando la distancia euclidiana respecto  $\mu_X$ .<sup>1</sup>

*Demostración.* Sea  $x^{(m)} = \mu_X + T\varphi^{(m)} \in \mathbb{R}^N$  el PCA de  $x$ , y  $T = [u^{(1)}, u^{(2)}, \dots, u^{(N)}]^T$  la transformada de *Karhunen-Loève* (de dimensión  $N \times N$ ). Se debe demostrar que  $r = [(x^{(m)} - \mu_X)^T \Sigma_x^{-1} (x^{(m)} - \mu_X)]^2 = [(\varphi^{(m)} - \mu_\varphi)^T (\varphi^{(m)} - \mu_\varphi)]^2$ . Definiendo  $\Sigma_x = E[(x - \mu_X)(x - \mu_X)^T]$  como la matriz de covarianza de  $X$  (de dimensión  $N \times N$ ) y  $\Sigma u^{(n)} = \lambda_n u^{(n)}$  como la relación de  $\Sigma_x$  con sus autovalores y autovectores [46], entonces:

$$\begin{aligned}
 |r|^2 &= (x^{(m)} - \mu_X)^T \Sigma_x^{-1} (x^{(m)} - \mu_X) && x^{(m)} = \mu_X + T\varphi^{(m)} \\
 &= (T\varphi^{(m)})^T \Sigma_x^{-1} (T\varphi^{(m)}) && (AB)^T = B^T A^T \\
 &= (\varphi^{(m)T} T^T) \Sigma_x^{-1} (T\varphi^{(m)}) && \text{prop. asociatividad} \\
 &= \varphi^{(m)T} (T^T \Sigma_x^{-1} T) \varphi^{(m)} && \mu_\varphi = 0 \\
 &= (\varphi^{(m)} - \mu_\varphi)^T (T^T \Sigma_x^{-1} T) (\varphi^{(m)} - \mu_\varphi) && (AB)^{-1} = A^{-1} B^{-1} \\
 &= (\varphi^{(m)} - \mu_\varphi)^T (T^T (T^{-1} \Sigma_x^{-1})^{-1}) (\varphi^{(m)} - \mu_\varphi) && (AB)^{-1} = A^{-1} B^{-1} \\
 &= (\varphi^{(m)} - \mu_\varphi)^T ((T^{-1} \Sigma_x^{-1}) (T^T)^{-1})^{-1} (\varphi^{(m)} - \mu_\varphi) && T^T T = T T^T = I \\
 &= (\varphi^{(m)} - \mu_\varphi)^T (T \Sigma_x^{-1} T^T)^{-1} (\varphi^{(m)} - \mu_\varphi)
 \end{aligned}$$

<sup>1</sup>Es necesario usar vectores estacionarios de *sentido amplio* porque su distribución de probabilidad permanece constante durante todo el proceso, lo que permite calcular su media y varianza solo una vez. Se estudia el caso de la distancia de  $x$  con su centroide, ya que a partir de ahí se puede extender la demostración para dos procesos aleatorios.

$$\begin{aligned}
&= (\varphi^{(m)} - \mu_\varphi)^T \Sigma_\varphi^{-1} (\varphi^{(m)} - \mu_\varphi) & \Sigma_\varphi &= T \Sigma_x T^T \ (\dagger) \\
&= (\varphi^{(m)} - \mu_\varphi)^T (\varphi^{(m)} - \mu_\varphi) & \Sigma_\varphi^{-1} &= I \ (\ddagger)
\end{aligned}$$

□

(†). En la demostración se afirma que  $\Sigma_\varphi = T \Sigma_x T^T$  porque:

$$\begin{aligned}
\Sigma_\varphi &= E[\varphi \varphi^T] \\
&= E[(T(x - \mu_X))(T(x - \mu_X))^T] \\
&= E[T(x - \mu_X)(x - \mu_X)^T T^T] \\
&= T E[(x - \mu_X)(x - \mu_X)^T] T^T && T \text{ es factorizable en la fórmula 2.20.} \\
&= T \Sigma_x T^T.
\end{aligned}$$

(‡). Como los datos transformados por PCA tienen covarianza nula, si se usa la fórmula de la correlación en lugar de la covarianza (factible al usar *Karhunen-Loève*), es decir,  $\Sigma_\varphi = 1/M [(\varphi - \mu_\varphi)^T (\varphi - \mu_\varphi)]$ , entonces  $\Sigma_\varphi = I$ , pues  $\forall [i, j \in \mathbb{N} / \sigma_{i,j} = \delta_{i-j}]$ , es decir, la correlación entre variables distintas es nula y unitaria para el resto [46].



# Apéndice C

## Demostración de la definición de distancia de Mahalanobis

*Demostración.* Suponiendo que  $\Sigma$  sea una matriz definida positiva o p.d. (es decir,  $\forall x \neq \underline{0} [x^T \Sigma x > 0]$ ), se sabe que  $\Sigma^{-1}$  existe y también es p.d. [72].<sup>1</sup> La distancia de Mahalanobis:  $d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$ , es, en efecto, una distancia o métrica, porque cumple con las condiciones establecidas en la sección 2.4.1:

**No negatividad:**  $\forall x, y \in E [0 \leq d(x, y)]$

i  $0 = d(x, y)$

$$\begin{aligned} 0 &= \sqrt{(x - y)^T \Sigma^{-1} (x - y)} \\ &= \sqrt{\underline{0}^T \Sigma^{-1} \underline{0}} && x = y, \underline{0} = (0, 0, \dots, 0)_{N \times 1} \\ &= 0 \end{aligned}$$

ii  $0 < d(x, y)$

$$\begin{aligned} 0 &< \sqrt{(x - y)^T \Sigma^{-1} (x - y)} && x \neq y \\ &< (x - y)^T \Sigma^{-1} (x - y) && \Sigma^{-1} \text{ es p.d.} \end{aligned}$$

**Propiedad idéntica:**  $d(x, y) = 0 \iff x = y$

---

<sup>1</sup>La demostración de la invertibilidad de  $A$  consta de dos partes. En la primera, se parte de la definición de los autovalores y autovectores de  $A$  ( $\lambda_i$  y  $x^{(i)}$ , respectivamente), se realiza una multiplicación en ambos lados de la ecuación por el término  $x_i^T$  y se hace que  $\lambda_i \|x^{(i)}\|^2 > 0$  para indicar que  $\lambda_i > 0$  (la norma es positiva o cero). En la segunda, se utiliza la definición de matriz inversa de la sección 2.4.7.

$$\text{i } x = y \implies d(x, y) = 0$$

$$\begin{aligned} (x - y) &= \underline{0} & \underline{0} &= (0, 0, \dots, 0)_{N \times 1} \\ (x - y)^T &= \underline{0}^T \\ (x - y)^T \Sigma^{-1} &= \underline{0}^T \Sigma^{-1} & \Sigma \text{ es p.d.} &\implies \Sigma^{-1} \text{ existe} \\ (x - y)^T \Sigma^{-1} (x - y) &= \underline{0}^T \Sigma^{-1} \underline{0} \\ &= (\underline{0}^T \Sigma^{-1}) \underline{0} & \text{asociatividad multiplicativa} \\ &= \underline{0}^T \underline{0} & \underline{0} &\in \text{kernel}(E) \\ &= 0 \end{aligned}$$

$$\text{ii } d(x, y) = 0 \implies x = y$$

$$\begin{aligned} \sqrt{(x - y)^T \Sigma^{-1} (x - y)} &= 0 \\ \left( \sqrt{(x - y)^T \Sigma^{-1} (x - y)} \right)^2 &= 0^2 \\ (x - y)^T \Sigma^{-1} (x - y) &= 0 \\ \implies x - y &= \underline{0} \vee \Sigma^{-1} = 0_{N \times N} & 0_{N \times N} &: \text{matriz de ceros} \\ \implies x &= y \vee \forall x [x^T \Sigma^{-1} x = 0] \\ \implies x &= y \vee (\rightarrow \leftarrow) & \Sigma^{-1} \text{ es p.d.} &\implies \forall x \neq \underline{0} [x^T \Sigma^{-1} x > 0] \\ \implies x &= y \end{aligned}$$

**Simetría:**  $\forall x, y \in E [d(x, y) = d(y, x)]$

$$\begin{aligned} d(x, y) &= \sqrt{(x - y)^T \Sigma^{-1} (x - y)} \\ &= \sqrt{(-y + x)^T \Sigma^{-1} (-y + x)} & \text{conmutatividad de la suma} \\ &= \sqrt{(-1)(y - x)^T \Sigma^{-1} (-1)(y - x)} & (-s)^T = -s^T \\ &= \sqrt{(y - x)^T \Sigma^{-1} (y - x)} & -AB = AB(-1) \wedge (-1)(-1) = 1 \\ &= d(y, x) \end{aligned}$$

**Desigualdad triangular:**  $\forall x, y, z \in E [d(x, z) \leq d(x, y) + d(y, z)]$

Sea  $\|x - y\|_m$  la norma de Mahalanobis (demostrada en la sección que sigue) y  $d(x, y)$  su distancia (tal que  $\|x - y\|_m = d(x, y)$ ) entonces:<sup>2</sup>

$$\begin{aligned} d(x, z) &\leq d(x, y) + d(y, z) \\ \|x - z\|_m &\leq \|x - y\|_m + \|y - z\|_m \\ \|x - y + y - z\|_m &\leq \|x - y\|_m + \|y - z\|_m \\ \|(x - y) + (y - z)\|_m &\leq \|x - y\|_m + \|y - z\|_m \quad \text{def. norma, cumple desigualdad triangular} \end{aligned}$$

□

<sup>2</sup>La idea de usar la norma de Mahalanobis para demostrar la propiedad de la subaditividad fue sugerida por Amey Joshi en el foro: <https://math.stackexchange.com/questions/528318/how-can-one-prove-that-mahalanobis-distance-is-a-metric>.

## C.1 Norma de Mahalanobis

*Demostración.* El producto interno de Mahalanobis  $\langle x, y \rangle_m$  engendra la norma de Mahalanobis  $\|x - y\|_m$  de la siguiente manera:  $\|x - y\|_m^2 = \langle x - y, x - y \rangle_m$ . La norma de Mahalanobis engendra a su vez a la distancia de Mahalanobis  $d(x, y)$  como sigue:  $d(x, y) = \|x - y\|_m$ . Y la distancia de Mahalanobis se define así:  $d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$ . Estos términos se relacionan en la expresión:  $\sqrt{\langle x - y, x - y \rangle_m} = \|x - y\|_m = d(x - y, 0) = d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}$ . Para demostrar que la norma de Mahalanobis, en efecto, cumple la definición de una norma, es necesario probar las siguientes propiedades [45]:

**Absolutamente escalable:**  $\|ax\|_m = |a|\|x\|_m$ , con  $a \in \mathbb{R}$

$$\begin{aligned} \|ax\|_m &= \sqrt{(ax)^T \Sigma^{-1} (ax)} \\ &= \sqrt{a^2 (x^T \Sigma^{-1} x)} \\ &= |a| \sqrt{x^T \Sigma^{-1} x} \\ &= |a| \|x\|_m \end{aligned}$$

**Definida para el vector cero:**  $\|x\|_m = 0 \iff x = \underline{0}$

i  $x = \underline{0} \implies \|x\|_m = 0$

$$\begin{aligned} \|x\|_m &= \sqrt{x^T \Sigma^{-1} x} \\ &= \sqrt{\underline{0}^T \Sigma^{-1} \underline{0}} \\ &= 0 \end{aligned}$$

$$\underline{0} = (0, 0, \dots, 0)_{N \times 1}$$

ii  $\|x\|_m = 0 \implies x = \underline{0}$

$$\begin{aligned} \|x\|_m &= 0 \\ x^T \Sigma^{-1} x &= 0 \end{aligned}$$

$$\implies x = \underline{0} \vee \Sigma^{-1} = 0_{N \times N}$$

$0_{N \times N}$  : matriz de ceros

$$\implies x = \underline{0} \vee \forall x [x^T \Sigma^{-1} x = 0]$$

$$\implies x = \underline{0} \vee (\rightarrow \leftarrow)$$

$\Sigma^{-1}$  es p.d.  $\implies \forall x \neq \underline{0} [x^T \Sigma^{-1} x > 0]$

$$\implies x = \underline{0}$$

**No negatividad:**  $0 \leq \|x\|_m$

i  $0 = \|x\|_m$

$$\begin{aligned} 0 &= \sqrt{x^T \Sigma x} \\ &= \sqrt{\underline{0}^T \Sigma \underline{0}} \\ &= 0 \end{aligned}$$

$$x = \underline{0} = (0, 0, \dots, 0)_{N \times 1}$$

$$\text{ii } 0 < \|x\|_m$$

$$0 < \sqrt{x^T \Sigma^{-1} x} \\ < x^T \Sigma^{-1} x$$

$\Sigma^{-1}$  es p.d.

**Subaditividad:**  $\|x + y\|_m \leq \|x\| + \|y\|_m$

$$\begin{aligned} \|x + y\|_m^2 &\leq (\|x\| + \|y\|)^2 \\ (\sqrt{(x + y)^T \Sigma^{-1} (x + y)})^2 &\leq (\sqrt{x^T \Sigma^{-1} x} + \sqrt{y^T \Sigma^{-1} y})^2 \\ (x + y)^T \Sigma^{-1} (x + y) &\leq x^T \Sigma^{-1} x + 2\sqrt{x^T \Sigma^{-1} x} \sqrt{y^T \Sigma^{-1} y} + y^T \Sigma^{-1} y \\ (x^T + y^T) \Sigma^{-1} (x + y) &\leq \text{idem} \\ (x^T \Sigma^{-1} + y^T \Sigma^{-1}) (x + y) &\leq \text{idem} \\ x^T \Sigma^{-1} x + y^T \Sigma^{-1} x + x^T \Sigma^{-1} y + y^T \Sigma^{-1} y &\leq x^T \Sigma^{-1} x + 2\sqrt{x^T \Sigma^{-1} x} \sqrt{y^T \Sigma^{-1} y} + y^T \Sigma^{-1} y \\ y^T \Sigma^{-1} x + x^T \Sigma^{-1} y &\leq 2\sqrt{x^T \Sigma^{-1} x} \sqrt{y^T \Sigma^{-1} y} \\ x^T \Sigma^{-1} y + x^T \Sigma^{-1} y &\leq \text{idem} \\ 2x^T \Sigma^{-1} y &\leq 2\sqrt{x^T \Sigma^{-1} x} \sqrt{y^T \Sigma^{-1} y} \\ x^T \Sigma^{-1} y &\leq \sqrt{x^T \Sigma^{-1} x} \sqrt{y^T \Sigma^{-1} y} \\ \langle x, y \rangle_m &\leq \|x\|_m \|y\|_m \end{aligned} \quad (\dagger)$$

(†). Queda demostrado al haber llegado a la desigualdad de Holder.

□