UNIVERSIDAD DE SEVILLA
ESCUELA SUPERIOR DE INGENIEROS
INGENIERÍA DE TELECOMUNICACIÓN

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
ÁREA DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

PROYECTO FIN DE CARRERA

ALGORITMO AUTOMÁTICO DE SEGMENTACIÓN DE ESPECTROGRAMAS

Autor: Francisco Romero Rodríguez
Tutora: Begoña Acha Piñero
Junio 2003

PROYECTO FIN DE CARRERA:
ALGORITMO AUTOMÁTICO DE SEGMENTACION DE ESPECTROGRAMAS

Autor: Francisco Romero Rodríguez
Tutora: Begoña Acha Piñero

El tribunal nombrado para juzgar el Proyecto Fin de Carrera arriba citado, compuesto por:

Presidente:

Vocal:

Vocal secretario:

Acuerda otorgarle la calificación de:

Sevilla, a de de 2003.

AGRADECIMIENTOS

Son muchísimas las personas a las que quiero agradecer tanto la realización de este proyecto como la consecución de mi carrera. Es una tarea difícil debido a que a veces no se encuentran las palabras exactas para expresar todo lo que se siente, pero aún así es mi voluntad la de resaltar la labor y el apoyo de las siguientes personas.

En primer lugar tengo que mencionar el soporte constante de los tutores del presente Proyecto. Primeramente mis supervisores en Swansea, Dr. John Mason y Dr. Nick Evans, sin cuyo apoyo, confianza y entrega diaria no habría disfrutado ni sacado tanto partido a mi estancia académica en el año Erasmus. Seguidamente, mi mayor gratitud va dirigida a la Dra. Begoña Acha Piñero, por su asesoramiento, sus sabios consejos y amistad no sólo en este año, sino a lo largo de toda la carrera. He tenido la enorme suerte de disfrutar de la colaboración de los mejores profesionales.

A continuación tengo que destacar a mi familia. A mis queridísimos padres, Francisco y Amalia, y a mis adoradas hermanas Nieves y María. A ellos les agradezco de corazón todo lo que me han brindado siempre: amor, confianza, apoyo, consejos... y todo lo necesario para aprender y valorar la importancia de la familia a lo largo de las distintas etapas de la vida. Os llevo siempre dentro de mí.

Es el turno de mis amigos y me resulta imposible nombrar a todos los que han contribuido en mis años de la carrera. Es mi obligación la de destacar al menos la amistad, en el verdadero significado de la palabra, de las siguientes personas:

Mis amigos de siempre: Antonio, Carlos, Amaru, Jose Valencia, Mariví, Sylvia, Rosa, Chelo, Maleni, Chío, David, Jorge, Angel y Amélie son sólo los primeros nombres que me vienen a la cabeza. Ellos saben que siempre habrá algo muy especial entre todos nosotros.

Mis compañeros de carrera, destacando la intensa amistad de personas como Fernando, Jesús Ruiz, María, Rafa, Carlos, Julio, Fran, Noelia, Rocío, Jesús Velarde, Raúl, Ana y Gema. Nunca olvidaré todos lo buenos momentos que hemos pasado juntos, y ojalá no perdamos el contacto nunca, o al menos en muchísimo tiempo.

Mis “amigos Erasmus” y mucho más: Carol, Curro, Tania, Jose, Gustavo, Jasmine, y muchísimas otras personas que han contribuido en la más enriquecedora e intensa experiencia que podía imaginar. Lo vivido en los últimos nueve meses ha cambiado muchas cosas dentro de nosotros, ¿verdad? Gracias por ser así, y que continúe siempre nuestro viaje.

Este final de carrera es por tanto también mérito vuestro. Una vez más, muchísimas gracias. Rodeado de todos vosotros ha sido para mí un verdadero y auténtico placer.

ÍNDICE

CAPÍTULO 1

INTRODUCCIÓN

La tarea de separar el lenguaje del ruido ha sido un desafío particular durante muchos años. Tanto si el habla va destinada a un sistema de reconocimiento de voz automático como a una persona, el objetivo final de la tarea es esencialmente el mismo y puede ser resumido en la extracción de la señal de voz que conduce a la mejora en el reconocimiento. En este sentido, el temprano trabajo de Boll [7] es generalmente reconocido como el precursor de las investigaciones experimentales de distintas variantes bajo el título genérico de la substracción espectral. Típicamente, estas estimaciones hacen referencia a cortas transformaciones discretas de Fourier de regiones, y en los primeros acercamientos fueron derivados de intervalos sin lenguaje. A partir de ahí, diversos procedimientos han sido llevados a cabo, incluyendo acercamientos como son los ejemplos de el uso de cuantiles, o el uso de regiones locales de tiempo y frecuencia para la derivación de estimaciones de ruido. Claramente estos acercamientos tienen el potencial de la derivación de mejores estimaciones de ruido al hacer uso no sólo de todo el recorrido temporal, sino también de la posibilidad de facilitar estimaciones dentro de la misma ventana instantánea que está sujeta a la compensación, una característica particularmente beneficiosa para diferentes ruidos no-estacionarios.

Este Proyecto considera otro acercamiento que posee estos beneficios en el intento de separación de señal de ruido. Las Representaciones Tiempo-Frecuencia (RTF) son herramientas satisfactorias para el análisis de señales no-estacionarias, pero su lectura no es

siempre sencilla para la tarea de interpretación de la señal. Es por tanto necesaria la investigación del uso de las propiedades estadísticas de las RTF para el objetivo de la clasificación o el reconocimiento. En concreto, en el desarrollo del presente Proyecto ha sido elegida la RTF más usada: el espectrograma. Realmente, el proceso es el de segmentación y clasificación de regiones de un espectrograma obtenido de una señal limitada en el tiempo y formada tanto por señal como por ruido. El proceso está basado en las propiedades estadísticas de los espectrogramas de tiempo limitado y corto, así como en el filtrado morfológico. Las regiones en el espectrograma son identificadas como señal o como ruido, y posteriormente las regiones de señal son agrandadas a través del proceso del filtrado morfológico. Se ha publicado muy poco trabajo sobre el filtrado morfológico de la voz. La motivación del trabajo aquí presentado nace del reciente trabajo de Hory *et al* [1,2] en el campo de la segmentación de señal.

El acercamiento presentado en estos artículos trata sobre la segmentación de señal en condiciones de ruido aditivo, utilizando características estadísticas obtenidas de sub-unidades del espectrograma. En concreto, las estadísticas usadas son la media y la desviación estándar de las pequeñas sub-unidades, llamadas células, asociadas a cada píxel del espectrograma. Estas características estadísticas de cada célula permitirán la elaboración de un *Espacio Característico* paralelo al espectrograma. En él, y gracias a unas estimaciones de niveles de ruido obtenidas de forma cada vez más precisa en cada iteración, irá realizándose la segmentación. La técnica consiste en la selección de unas semillas que serán extraídas, y provocarán la extracción de sus respectivos píxeles vecinos en el espectrograma, utilizando por tanto un algoritmo de región creciente. Asimismo, las anteriores estimaciones de ruido servirán para determinar la condición de parada del algoritmo logrando la extracción de la mayor cantidad de señal posible. Es importante señalar que la segmentación es llevada a cabo por un algoritmo que no requiere como parámetro de entrada conocimiento previo alguno de la señal no-estacionaria. El resultado es una extracción automática de subconjuntos de la RTF, relevante para la comprensión de la señal.

El trabajo original considera señales sintéticas chirp (señales sinusoidales de frecuencia linealmente creciente) inmersas en ruido aditivo Gausiano blanco. Este Proyecto consiste en primer lugar en la implementación independiente de este algoritmo, comprobando sus características y resultados. En segundo lugar, se persigue la extensión del trabajo, realizando un estudio sobre similares señales chirps pero a distintas SNRs, y posteriormente buscando su aplicación en el contexto del reconocimiento de voz.

Los resultados de la evaluación y los experimentos obtenidos prueban que el algoritmo es desarrollado correctamente. También se demuestra que es robusto en condiciones de señal a ruido muy pobres, al conseguir una segmentación muy aceptable para valores de SNR incluso negativas. Por último, los resultados preliminares de experimentos de mejora de voz

con esta técnica, son más que satisfactorios. Los experimentos de reconocimiento de voz son evaluados en la base TIDigits con ruido añadido bajo el estándar de la configuración AURORA 2 [9]. Estas señales fueron transformadas a sus espectrogramas, los cuales fueron posteriormente segmentados. A continuación se eliminó la información de amplitud de las regiones de voz extraídas. Se observó un incremento en la precisión del reconocimiento de palabra de un 16% a un 56% ante una SNR de 5 dB, ilustrando que la forma espectral por sí sola lleva a quizás sorprendentes resultados en el Reconocimiento de Voz Automático (ASR: Automatic Speech Recognition) bajo condiciones de ruido. Estos resultados muestran por tanto el camino para continuar la investigación. El siguiente paso es incluir la información de amplitud con apropiadas técnicas de compensación de ruido, así como la mejora del algoritmo para los distintos tipos de ruido existentes.

CAPÍTULO 2

PROCESAMIENTO DE VOZ E IMÁGENES

Este capítulo pretende dar una visión global del procesamiento de la voz y de la imagen obtenida a partir de las señales de voz. En primer lugar se realiza una descripción de la producción de la voz en el ser humano, seguido de las líneas principales del tratamiento para su reconocimiento. A continuación se introduce su papel en las comunicaciones así como las distintas aplicaciones del procesamiento de la voz. Posteriormente se describe el proceso de obtención de una imagen representativa de la voz muy utilizada: el espectrograma. Esta imagen es la realmente tratada por los muy diversos algoritmos existentes en búsqueda de la mejora de la señal de audio, entre ellos el algoritmo mostrado en el presente Proyecto. Después, se realiza una descripción de las cualidades de la voz en el espectrograma, para finalizar explicando la situación del algoritmo desarrollado en los siguientes capítulos dentro del campo de la segmentación de las imágenes.

2.1 PRODUCCIÓN DE LA VOZ

La voz constituye el elemento principal de comunicación entre los seres humanos. Entender su formación y aprender a tratarla como una señal constituye uno de los campos fundamentales de la Ciencia hoy en día. Para conseguir el objetivo de un excelente tratamiento de la voz en todos los niveles y aplicaciones existentes, es necesario conocer primero todos los detalles referentes a ella, junto a los medios actuales de representación.

2.1.1 Órganos de la voz

Podemos dividir los órganos vocales en tres principales subsistemas: pulmones y tráquea, laringe y tracto vocal. Los pulmones y la tráquea son los suministradores de potencia del sistema; la laringe contiene el principal mecanismo de generación de sonido; y el tracto vocal modula el sonido resultante.

2.1.2 Producción de la voz

Físicamente la voz es producida siguiendo el siguiente proceso. El aire entra en los pulmones por el mecanismo normal de la respiración. Cuando el aire es expulsado desde los pulmones a través de la tráquea, se produce la vibración de las cuerdas vocales dentro de la laringe. El flujo de aire es troceado en pulsos cuasi-periódicos que son entonces modulados en frecuencia al pasar por la faringe, la cavidad bucal, y posiblemente la nasal. Dependiendo de la posición de las diversas articulaciones, como la lengua, los labios, la mandíbula o la boca, se producen distintos sonidos.

La operación del sistema en su conjunto es sobre todo dividida a nivel interpretativo en dos funciones: *excitación* y *modulación*, tal y como se muestra esquemáticamente en la Figura 2.1. La excitación tiene lugar principalmente en la glotis pero también en otros puntos. La modulación es realizada por los diversos órganos del tracto vocal.

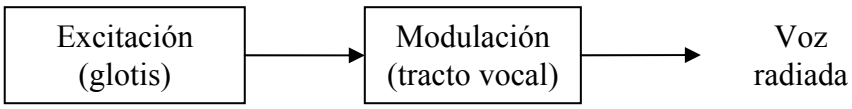


Figura 2.1 : Funciones en la generación de la voz.

La excitación es realizada de distintas maneras: compresión, susurro, fricación, fonación y vibración. En la modulación se añade la información a la salida glotal. Acústicamente, el principal medio de modulación es la operación de filtrado. La forma de

onda glotal es muy rica en harmónicos, y el tracto vocal, al igual que un tubo acústico, tiene frecuencias naturales en función de su forma. Estas resonancias naturales son llamadas *formants,* y son las responsables de la modulación de la voz. Tienen además influencia en todas las vocales y algunas consonantes, y también suministran información crucial del resto de las consonantes. Otros tipos de modulación son las distintas interrupciones y los pequeños estallidos de ruido añadidos que constituyen las consonantes. Vemos por tanto las diversas maneras de producir sonidos de las que es capaz el ser humano.

2.1.2.1 Modelo de Producción de voz

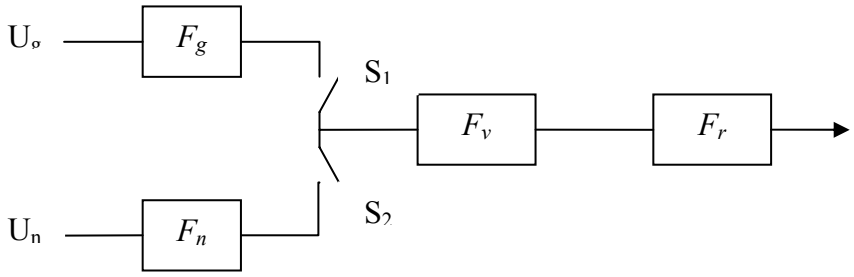


Figura 2.2 : Modelo del Tracto Vocal

La Figura 2.2 muestra un modelo simplificado del Tracto Vocal. En ella, U_g es la excitación glotal, U_n la fuente de ruido; F_g y F_n son filtros de forma. Los interruptores S_1 y S_2 permiten la selección de los vocoides vocales o de los conptoides no vocales. F_v es el Tracto Vocal, y F_r se encarga de la impedancia de radiación.

2.2 RECONOCIMIENTO DE VOZ

2.2.1 Planteamiento

El sueño de los años cincuenta era la "máquina de escribir fonética". Conociendo los formants, las transiciones y las acústicas de las consonantes, los investigadores creían que el proceso de trascripción fonética podría ser mecanizado. Esta ilusión aún no ha podido ser realizada. Actualmente existe hardware aproximado comercialmente disponible, pero no así un transcriptor fonético. Las razones de porqué no ha sido aún posible su implementación suministran una buena introducción a las realidades del reconocimiento de voz.

2.2.2 Problemas del Reconocimiento de voz

Los problemas encontrados en un reconocimiento de voz práctico pueden ser resumidos:

1. *Variaciones del hablante*. No existen dos personas cuyos sonidos suenen igual, y por tanto, la señal de voz contiene en el fondo tanto variables como información fonética dependiente de la persona que habla. Su separación es en el fondo complicada.
2. *Ambigüedad*. Las variables acústicas no pueden ser mapeadas una a una en las variables fonéticas. El cerebro humano trata este dilema aproximando, basándose en el conocimiento del lenguaje hablado en ese momento y en el tema de conversación.
3. *Variaciones en la voz individual*. Esta característica está presente incluso en la lectura cuidadosa de hablantes entrenados.
4. *Ruido e interferencias*. Los humanos pueden reconocer voz en condiciones de señal a ruido muy malas e incluso en presencia de otras voces interferentes. Una buena parte de esta capacidad depende de la habilidad del ser humano de tratar dos señales acústicas distintas a la vez, proceso que aún no se comprende perfectamente. El ruido degrada la actuación de cualquier programa de reconocimiento de voz. Las máscaras de ruido debilitan los fricativos, provocan que los predictores lineales produzcan estimaciones de parámetros incorrectas, y hace más difícil la identificación del comienzo y final de las palabras. La mayoría de los programas de reconocimiento de voz están diseñados para usarlos en condiciones muy limpias, sin apenas ruido contaminante.

Los tipos de programas destinados al reconocimiento de voz se dividen en las siguientes categorías según el problema destinado a resolver. La siguiente secuencia queda planteada en orden de dificultad creciente.

1. Reconocimiento de palabras en solitario. Es el reconocimiento de palabras separadas por pausas.
2. Detección del número de apariciones de una palabra específica en un discurso.
3. Reconocimiento de palabras sin pausas entre ellas.
4. Interpretación de la voz.

En todos estos tipos, el sistema puede ser dependiente o independiente del hablante. A cualquier nivel, los sistemas independientes de la persona que pronuncia las palabras son mucho más difíciles de implementar y muestran una actuación más pobre que los sistemas dependientes. En todos los casos, la actuación se deteriora con el crecimiento del vocabulario.

A continuación se describen algunas técnicas usadas en el primer grupo. El algoritmo automático presentado en este Proyecto es catalogado dentro de este primer caso, siendo dicho conjunto de vital importancia ya que la mayoría de los algoritmos de los grupos posteriores se basan en extensiones o en sucesivas aplicaciones del usado en el primero.

2.2.2.1 Reconocimiento de palabras en solitario

Este tipo de reconocimiento es el más sencillo, ya que las pausas entre palabras simplifican el reconocimiento al ser relativamente fácil la implementación de un detector de inicio y final de palabra. Para esta detección suelen escogerse una serie de C*aracterísticas* dependiendo del sistema. Las más usadas son:

1. Amplitud (o potencia) frente al Tiempo
2. Tasa de cruces por cero.
3. Balance genérico espectral (alta frecuencia frente a baja frecuencia)
4. Detalles espectrales específicos en alguna otra forma. Por ejemplo:
    a) DFT Espectrograma
    b) F_1, F_2 , F_3
    c) Parámetros LPC (Linear Predictor Coefficients)
    d) Salidas de bancos de filtros (populares debido a su bajo coste computacional y rapidez)

La amplitud es la principal fuente de información para la detección del comienzo y final de una palabra. También es útil para la discriminación de vocales/consonantes. La tasa de cruces por cero ayuda a caracterizar los fricativos y silbantes, al hacer un balance espectral. La información espectral de alta resolución (por ejemplo la obtenida a partir los parámetros LPC o las salidas del banco de filtros) suministra información de formant y de transiciones de formants o equivalentes.

2.3 COMUNICACIONES Y PROCESAMIENTO DE VOZ

2.3.1 Tipos de Comunicaciones

Existen tres tipos de comunicaciones de voz:

- Persona a Persona.
- Persona a Máquina. Este caso es estudiado en dos campos: Reconocimiento de voz, y Reconocimiento del hablante.
- Máquina a Persona: la técnica tratada es denominada Síntesis de voz.

El primer caso tiene la variante que las personas se encuentren separadas por una larga distancia. La inclusión de algún medio de transmisión de la voz se hace por tanto indispensable, formando el llamado 'link electrónico'. Dicha transmisión consta necesariamente de las siguientes etapas:

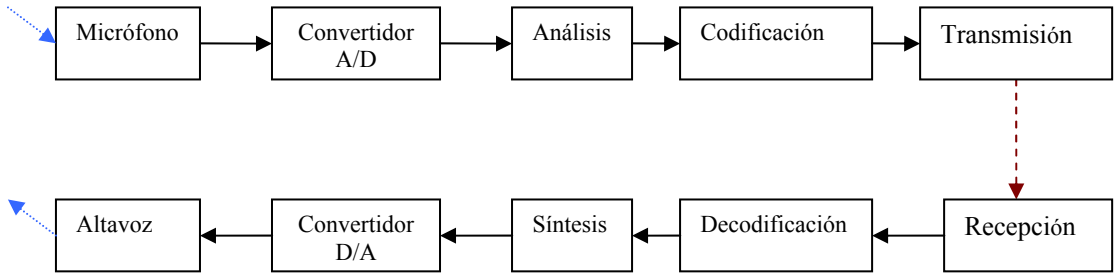


Figura 2.3 : Etapas de la Transmisión de voz en largas distancias.

2.3.2 Aplicaciones del Procesamiento de voz

Entre las muchas aplicaciones del Procesamiento de la voz, podemos destacar las siguientes:

- Comunicaciones de voz (link electrónico)
- Síntesis de voz.
- Reconocimiento automático, tanto de la voz como de la persona que habla.

La manera de llevarlas a cabo también es muy diversa, pudiendo incidir principalmente en el "Acercamiento de sistemas", pero sobre todo, en el basado en ventanas o particiones, como es el caso del espectrograma.

2.4 EL ESPECTROGRAMA

2.4.1 Definición

Un espectrograma es una visión en 3 dimensiones de una señal, siendo sus ejes el tiempo (eje horizontal), la frecuencia (eje vertical) y la intensidad o potencia (eje vertical o color).

2.4.2 Obtención del espectrograma

El espectrograma es computado a partir del dominio temporal de la señal. La Transformada de Fourier Discreta (DFT: Discrete Fourier Transform) es aplicada a la forma de onda de la señal para determinar qué frecuencias están presentes en cada momento. El resultado de la DFT es el espectro. Usando una ventana deslizante que se traslada a lo largo de la forma de onda, podemos ir creando el espectro juntando las ventanas (son computadas una a una) hasta el final de la onda. Todos los cálculos son conectados para producir el espectrograma, tal y como muestra la Figura 2.4.

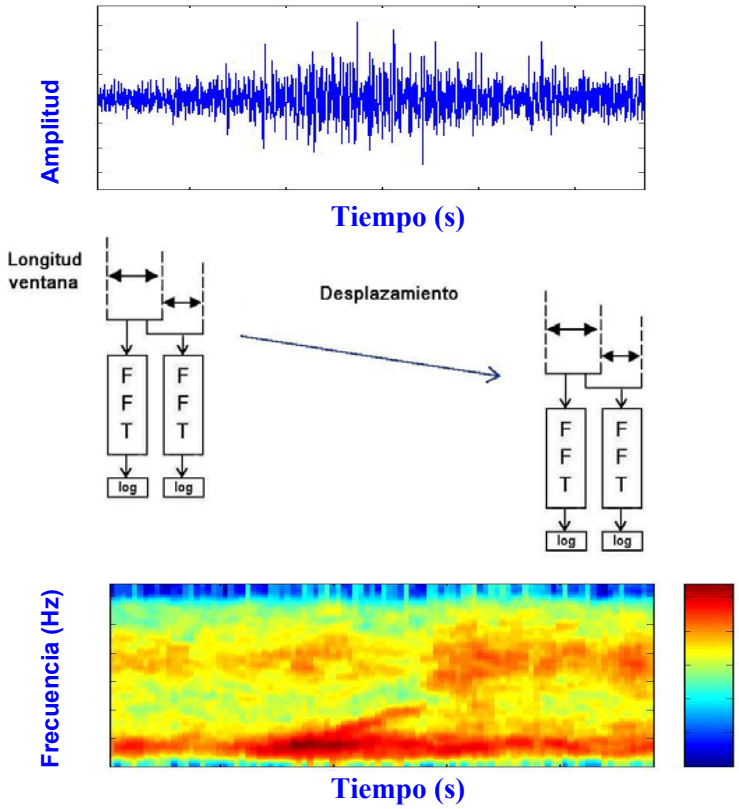


Figura 2.4 : Conversión Dominio Temporal – Espectrograma.

El resultado de la transición Dominio Temporal – Espectrograma depende fuertemente de los parámetros de cálculo usados. Otros parámetros podrían dar resultados más óptimos, pero la búsqueda de dichos valores exactos no ha sido el objetivo de este Proyecto. Los parámetros usados para el cálculo del espectrograma a lo largo de este Proyecto han sido los más utilizados comúnmente:

1. Ventana: Hanning
2. Longitud de la ventana: 256
3. Número de puntos usados en la DFT: 256
4. Desplazamiento de la ventana: 64

Los colores del espectrograma señalan las zonas importantes, representando en rojo las zonas de energía más alta. A partir de ahí, y en orden decreciente: naranja, amarillo, verde, cian, azul y magenta. Puede decirse que las regiones con mayor proximidad a las zonas de más alta energía son las regiones con mayores probabilidades de contener señal determinista, y viceversa.

El objetivo final buscado es el de la implementación de un algoritmo capaz de detectar las regiones del espectrograma que contienen señal determinista. La tarea queda dificultada por la presencia de ruido que contamina la señal, no siendo siempre obvia la segmentación de dicha imagen en función de únicamente el nivel de energía.

En este Proyecto se ha realizado el desarrollo del algoritmo automático de segmentación de señal principalmente sobre dos tipos de espectrogramas procedentes de distintas señales:

1. Espectrogramas sintéticos: chirps.
2. Espectrogramas de voz auténticos.

El primer tipo es usado para asegurar la correcta implementación del algoritmo, al ser el ejemplo de señal usado en [1] y [2]; el segundo tipo es utilizado para evaluar el potencial y la aplicabilidad del algoritmo en el procesamiento de la voz. A pesar de que en el Capítulo 5 se realiza una mayor descripción de la obtención y tratamiento de estas señales, a continuación referimos y mostramos brevemente ambos tipos.

2.4.2.1 Adquisición de espectrogramas de señales sintéticas chirp

Una señal chirp no es más que una señal sinusoidal de amplitud constante cuya frecuencia se incrementa linealmente a lo largo del tiempo. A esta señal se le añade en el dominio temporal Ruido Gausiano Blanco[1] (WGN: White Gaussian Noise) de diferentes niveles para crear una base de señales prueba con Relaciones de Señal a Ruido (SNR: Signal to Noise Ratio) en el rango -5 dB a 20 dB, asumiendo un incremento intermedio de 5 dB.

La Figura 2.5 (b) muestra un chirp ruidoso creado a partir del chirp original sin ruido (a). Ambas señales son transformadas al espectrograma (c) y (d). Nótese que todas las señales chirp de este Proyecto tienen las mismas dimensiones.

[1] Ruido con característica espectral continúa y uniforme.

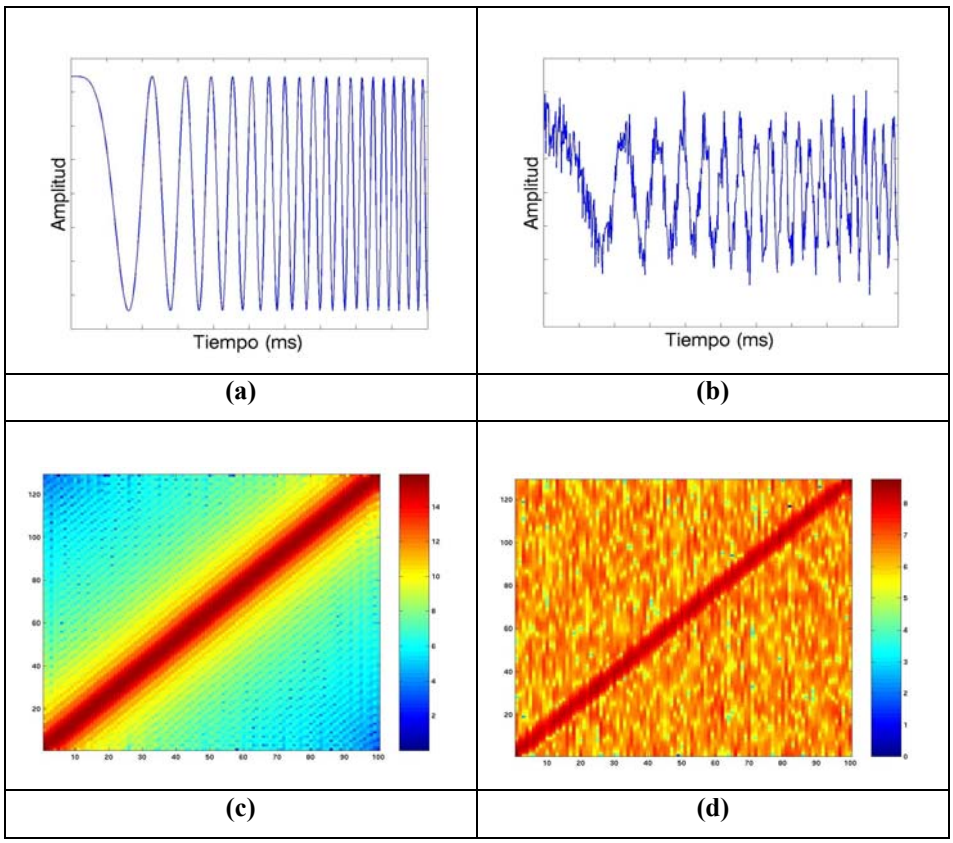


Figura 2.5 : (a) Forma de onda temporal de un chirp limpio. (b) Chirp con ruido WGN añadido. (c) Espectrograma del chirp limpio. (d) Espectrograma de la señal *(b)*

2.4.2.2 Adquisición de espectrogramas de voz real

Las muestras de voz son obtenidas de una base de datos europea. Una mayor descripción sobre esta adquisición es realizada en la Sección 5.2.1.1. Estas muestras se encuentran en el dominio temporal, y son por tanto transformadas a espectrogramas de manera análoga a la descrita anteriormente. La Figura 2.6 (b) muestra un ejemplo de espectrograma obtenido a partir de la forma de onda en (a).

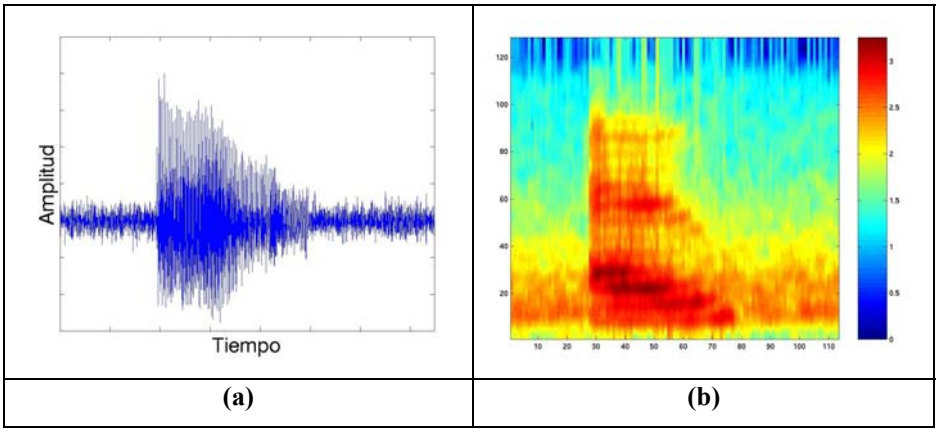


Figura 2.6 : (a) Forma de onda y (b) Espectrograma de voz contaminados por el ruido existente en una estación de tren.

2.4.3 Vuelta al dominio temporal

Para volver al dominio temporal a partir del espectrograma, aplicamos la DFT inversa a cada columna del espectrograma. El resultado es después multiplicado por la correspondiente fase de cada columna, guardada previamente en el proceso de transformación de Forma de onda - Espectrograma. Obtenemos finalmente la señal filtrada morfológicamente en el dominio temporal.

2.5 ESPECTROGRAMAS DE SEÑALES DE VOZ. PROPIEDADES DE LA VOZ EN EL ESPECTROGRAMA

Una vez descrito el proceso de obtención del espectrograma, pasamos a describir los espectrogramas de señales de voz. Vemos en primer lugar en la Figura 2.7 la imagen de cinco palabras consecutivas separadas por pausas, sin contaminación alguna de ruido.

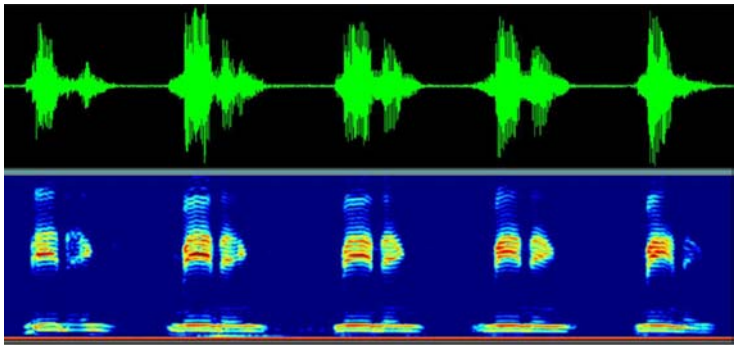


Figura 2.7 : Espectrograma de 5 palabras separadas por pausas.

Las principales características que podemos medir en los espectrogramas de voz son:

1. *Duración de palabra*. Es el tiempo de pronunciación de una palabra.
2. *Rango frecuencial*. El rango de frecuencia audible por los humanos se extiende desde los 20 Hz a los 20 kHz. Los sonidos emitidos por los seres humanos contienen energía en todas las frecuencias del rango audible, aunque la mayor información fonética se encuentra concentrada por debajo de los 8 kHz, siendo éste el principal rango usado por defecto en los espectrogramas.
3. *Periodicidades*. Existen dos estructuras claramente diferenciables:
    a) Tono. Son las frecuencias fundamentales de oscilación generadas en la glotis (vibración de las cuerdas vocales). La repetitiva secuencia de apertura y cierre de las cuerdas vocales ocurre en el hombre unas 125 veces por segundo, y en la mujer aproximadamente el doble (250 Hz),

elevando la sensación del tono. En el espectrograma se reflejan en líneas paralelas horizontales separadas una cierta distancia.

b) Resonancias del tracto vocal o formant. Técnicamente representan un conjunto de harmónicos adyacentes generados en alguna zona del tracto vocal. En el espectrograma aparecen como bandas anchas de energía incluyendo varias líneas producidas por el tono, y lógicamente distintas formas del tracto vocal producen distintos patrones de formants. Estos patrones pueden diferenciarse bastante entre distintas personas, pero un entrenamiento en la lectura de espectrogramas permite reconocer patrones que son independientes de algunas frecuencias, y por tanto identificar algunos fonemas con un alto grado de fiabilidad.

Estas estructuras harmónicas y periodicidades dan potencial para la reducción de datos. De hecho, los parámetros LPC consiguen una buena tasa de compresión. Sin embargo, la resonancia existente en el tono no suministra apenas información en el espectrograma de cara al reconocimiento de voz, y para el tratamiento de la imagen digital suele realizarse previamente un suave filtrado paso de baja para eliminarlas y realzar los formants. La imagen típica de un espectrograma de voz es por tanto la de la Figura 2.8.

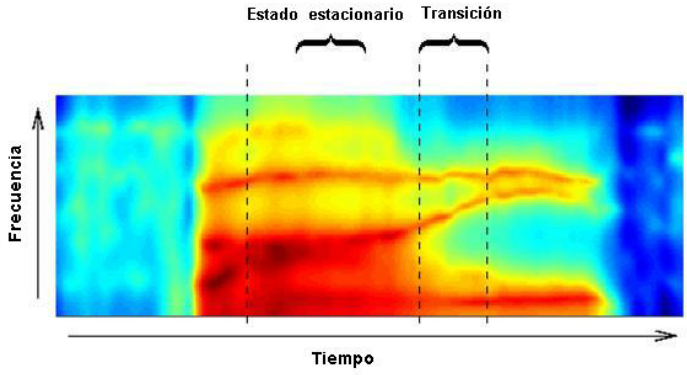


Figura 2.8 : Estados de la voz en el espectrograma.

Esta imagen muestra el espectrograma de una única palabra. Un siguiente acercamiento muestra que podemos distinguir principalmente en ella dos estados: el *estacionario*, y el de *transición,* al pasar de un fonema a otro.

El análisis de dichos estados es en realidad bastante complicado debido a la monumental cantidad de posibilidades existentes gracias al enorme vocabulario que forman nuestras lenguas, la gran cantidad de fonemas y todas sus posibles combinaciones, los distintos idiomas hablados, entonaciones, tipos de voz....

A pesar de todo, el espectrograma sigue siendo una transformación extremadamente útil de cara al tratamiento de la señal de voz. El estudio de las características espectrales es básico para la comprensión y el correcto procesamiento de dichas señales. El espectrograma es por tanto la representación en tres dimensiones por excelencia de las señales de voz.

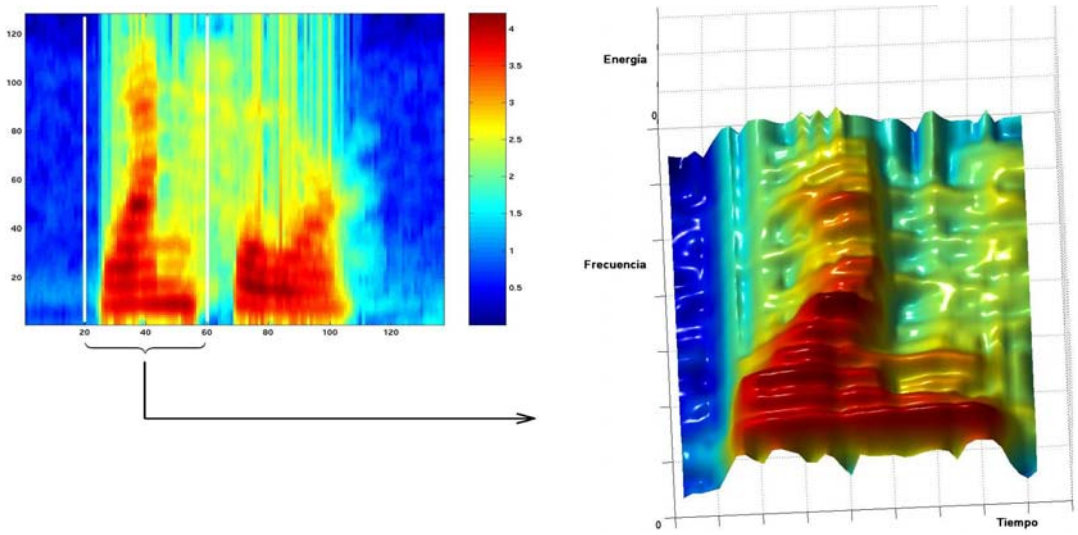


Figura 2.9 : Representación en tres dimensiones de la imagen Espectrograma.

Una vez que se ha conseguido pasar del dominio temporal de una señal de voz a su representación mediante una imagen digital, comienza el tratamiento de dicha composición. El tratamiento del espectrograma mediante el procesamiento digital de imágenes es muy amplio y ofrece innumerables posibilidades. El presente algoritmo se basa en la extracción de las zonas del espectrograma que contienen señal determinista. Dicha indicación en la imagen de cuáles son las regiones deseadas se conoce comúnmente como segmentación de una imagen.

2.6 SEGMENTACIÓN DE IMÁGENES

2.6.1 Introducción

La segmentación consiste en la división de una imagen en los objetos que lo constituyen, y es el primer paso en el análisis de la imagen. Este proceso tiene una gran cantidad de aplicaciones: militares, inspecciones industriales, medidas de control, imágenes médicas… Uno de los mayores problemas de la segmentación reside en que la subdivisión depende completamente de la aplicación, lo que contribuye a la dificultad de la automatización del algoritmo. Pueden distinguirse dos tipos, en función de las:

- Similitudes
- Discontinuidades

2.6.2 Segmentación basada en Similitudes

Dentro de este grupo, existen tres métodos principales:

a) Umbralización
b) Región creciente
c) División y fusión de regiones

2.6.2.1 Umbralización

Para las imágenes en blanco y negro, la técnica consiste en el cálculo del histograma de la imagen y la posterior umbralización de éste, ya sea de forma bimodal o multimodal. El umbral puede ser de distinto tipo en función de si sólo depende de la imagen original (umbral global), si sólo depende de las propiedades locales de la imagen (umbral local), o si depende además de los casos anteriores de cada píxel en concreto (umbral dinámico).

Si se trata de imágenes a color puede procederse de manera análoga pero con histogramas de una, o de varias dimensiones.

La principal desventaja de los métodos de umbralización es que no tienen en cuenta la información espacial, al fijarse únicamente en el valor de los píxeles de la imagen.

2.6.2.2 Región Creciente

Este procedimiento consiste en la selección de unos píxeles iniciales llamados semillas en base a unas condiciones. Estas semillas irán propagándose en función de unas determinadas propiedades, haciendo que las regiones segmentadas vayan creciendo. La dificultad de este algoritmo reside en la correcta elección de las semillas y en la selección de las propiedades de inclusión en una región de un píxel adyacente a una semilla, ya que ambas características dependen fuertemente de la aplicación deseada. Normalmente para imágenes a color la inclusión se hace en base a una distancia.

Este tipo de segmentación es el usado en el algoritmo automático propuesto en el presente Proyecto, y es explicado al detalle en los sucesivos capítulos.

2.6.2.3 División y fusión de regiones

Esta técnica no parte de semillas ni de cálculos de histogramas. Directamente divide la imagen en regiones adyacentes, y el método consiste en la sucesiva comparación de dichas regiones para fusionarlas o por el contrario dejarlas separadas. Esta comparación suele realizarse en base a una estructura de datos.

2.6.3 Segmentación basada en discontinuidades

La manera más usual de detectar discontinuidades en una imagen es utilizando una máscara. La suma del producto entre los coeficientes de la máscara y los píxeles de la imagen indicará las regiones a destacar.

Este grupo consta de la detección basada en tres tipos de discontinuidades:

a) Detección de puntos
b) Detección de líneas
c) Detección de bordes

2.6.3.1 Detección de puntos

El procedimiento consiste en la aplicación de una máscara "paso de alta", y comparar posteriormente si el resultado es mayor o menor que un cierto umbral.

2.6.3.2 Detección de líneas

En este caso la máscara debe ser la adecuada para destacar líneas horizontales, verticales u oblicuas. Se procede de igual manera que en el caso anterior.

2.6.3.3 Detección de bordes

Esta detección suele ir implementada en base a un operador de derivación. Distinguimos de nuevo tres tipos:

- *Operador Gradiente*: el gradiente da la dirección de máxima variación de la imagen *f* en *(x,y)*. Suele aplicarse de la siguiente manera:

$$|\nabla f| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}$$

  Existen dos tipos de detectores: direccionales (enfatizan una dirección en particular del borde) y no direccionales (no lo hacen).

  Su aplicación en imágenes a color es realizada aplicando el operador a cada componente o aplicando un gradiente vectorial.

- *Operador Laplaciano*: En este caso el operador consiste en

$$\Delta f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

  Su problema es que detecta bordes dobles. Además, es muy sensible al ruido.

- *Operador de Canny*: un filtro detector de borde debe realizar una buena detección, una buena localización y ser inmune a objetos interferentes. Buscando maximizar la SNR, Canny definió las cualidades que debería cumplir el filtro [15].

CAPÍTULO 3
VISIÓN GENERAL DEL SISTEMA

Una vez introducidas las generalidades del Procesamiento de Voz e Imágenes, pasamos a describir la globalidad del sistema implementado. En este capítulo se explican además de las principales cualidades del sistema, sus especificaciones y una breve descripción del proceso del filtrado morfológico. Asimismo, se realiza una aproximación a la Teoría de la Estimación de Ruido de Máxima Probabilidad desarrollada en [2], elemento básico en el algoritmo. Por último, se detallan los requerimientos y las características del sistema utilizado en el desarrollo y utilización del algoritmo.

3.1 VISIÓN GENERAL DEL SISTEMA

La Figura 3.1 introduce la estructura principal del sistema. La entrada es una señal en el dominio temporal, ya sea sintética o de voz real. Esta señal se encuentra contaminada por ruido, y es transformada tal y como se ha detallado en al capítulo anterior en la imagen espectrograma. A esta representación Tiempo-Frecuencia, se le aplica el Filtrado Morfológico, produciendo su segmentación, esto es, extrayendo las regiones consideradas como señal determinista o voz. Finalmente, el nuevo espectrograma segmentado es transformado al dominio temporal inicial para obtener la señal mejorada.

Figura 3.1 : Diagrama del sistema.

Los tipos de señales de entrada han sido descritos brevemente en las Secciones 2.4.2.1 y 2.4.2.2. Las transformaciones Onda Temporal - Espectrograma y viceversa es detallada en la Sección 2.4, mientras que el Filtrado Morfológico corresponde al Capítulo 4.

3.2 ESPECIFICACIONES DEL SISTEMA

Los requerimientos del sistema pueden ser resumidos de la siguiente manera. El sistema debe tener la capacidad de:

1. Transformación de la señal de audio a espectrograma y viceversa.
2. Estimación de la potencia del ruido que corrompe la señal.
3. Extracción de las regiones que contienen la señal determinista en el caso de las señales sintéticas, o extracción de las regiones que contienen la voz en los espectrogramas de voz.

3.3 CARACTERÍSTICAS DEL SISTEMA

3.3.1 Software

El algoritmo es implementado en Matlab por las siguientes razones:

- Es una eficiente herramienta numérica computacional basada en matrices y vectores.
- Disponibilidad de funciones básicas de cálculo y transformación de señales.
- Facilidad para mostrar la información gráficamente.
- Matlab es ejecutado bajo el Sistema Operativo Linux porque:
    - Es más estable que el sistema Operativo Microsoft para las largas horas de simulación requeridas en los experimentos realizados.

3.3.2 Hardware

- Este es un Proyecto de software. El único hardware involucrado es el de un ordenador personal Intel Pentium 4 con microprocesador a 2 GHz.

3.4 BREVE DESCRIPCIÓN DEL PROCESO DEL FILTRADO MORFOLÓGICO

El algoritmo usado en el proceso del filtrado morfológico es descrito con todo detalle en el Capítulo 4. Aun así, a continuación se muestra una breve descripción del procedimiento para mostrar la relevancia de la teoría introducida en la Sección 3.5.1.

En primer lugar, consideramos que un espectrograma es una imagen formada por elementos individuales llamados píxeles. Estos píxeles describen la señal de entrada al sistema, la cual es una señal sintética o de voz contaminada por ruido. Algunos grupos conexos de estos píxeles contienen por tanto señal determinista más ruido, siendo denominados patrones espectrales. Otros en cambio contienen solamente ruido. El objetivo final del algoritmo es identificar y separar estos dos tipos de regiones en el espectrograma mediante la segmentación.

La segmentación comienza identificando los píxeles más cualificados, llamados semillas. Son los píxeles con mayores expectativas de contener señal determinista entre el resto de píxeles (los criterios de selección de semillas son argumentados en la Sección 4.2.3.1). Estas semillas contaminan a los píxeles vecinos en el espectrograma a través del filtrado morfológico usando un algoritmo de región creciente, convirtiendo por tanto a los píxeles contaminados en nuevas semillas. Las áreas formadas por semillas y píxeles contaminados van creciendo a lo largo de la ejecución del algoritmo, iteración a iteración, hasta que todos los patrones espectrales son eventualmente identificados. Obviamente las semillas necesitan saber a qué vecinos contaminar y a cuáles no, así como cuándo parar la contaminación. De lo contrario, no se identificarían bien todas las regiones deseadas.

Un gran beneficio señalado en [1] y [2] es que la extracción de la señal determinista es automática, en el sentido que no hace falta ningún conocimiento previo sobre la señal de entrada no-estacionaria. Por 'ningún conocimiento previo' entendemos que no son necesarios conocer datos tales como la potencia del ruido que corrompe la señal. También significa que no se conocen de antemano el período de tiempo o el rango de frecuencias de la señal determinista y/o del ruido. Ninguno de estos datos es necesario para la detección explícita de la señal. De hecho, tras la segmentación del espectrograma, los anteriores parámetros se obtienen a partir del espectrograma segmentado, así como una estimación de la potencia del ruido bastante precisa. El algoritmo propuesto es por tanto ciego al tipo de señal de entrada, ya sea sintética o de voz, y corrompida por un nivel de ruido bajo o alto. El proceso de segmentación se detiene sólo cuando cree que toda la señal determinista ha sido extraída y que el resto de las regiones no extraídas están formadas únicamente por ruido, habiendo obtenido una última estimación de ruido por aproximaciones considerablemente precisa.

El algoritmo realiza estas tareas "inteligentes" gracias a un proceso incluido en la Teoría de la Estimación de Ruido de Máxima Probabilidad (MLNE: Maximum Likelihood Noise Estimation). A continuación se muestra un resumen de dicha Teoría, la cual es desarrollada con detalle en [2].

3.5 TEORÍA INVOLUCRADA EN EL PROCESO DE SEGMENTACIÓN AUTOMÁTICA

3.5.1 Maximum Likelihood Noise Estimation (MLNE)

3.5.1.1 Teoría Básica

Los coeficientes del espectrograma tienen distintas propiedades estadísticas en función de su procedencia. De esta forma, se demuestra que los coeficientes del espectrograma formado sólo por ruido Gausiano blanco siguen una distribución central χ^2 mientras que los componentes de señales deterministas poseen una distribución χ^2 no central. Por tanto, la tarea de diferenciar coeficientes de uno u otro tipo se convierte en la de detectar coeficientes de una u otra distribución.

A continuación se escogen dos características estadísticas: la media local y la desviación-estándar de los coeficientes del espectrograma. La media de un conjunto de variables χ^2 centrales es una χ^2 central con un número conocido de grados de libertad. Si las variables están correladas, como normalmente ocurre en el espectrograma, el número de

grados de libertad no es un número entero y la media local no puede ser considerada una variable χ^2. En este caso es una distribución γ de parámetros desconocidos.

La tarea de segmentación consiste en identificar los coeficientes de ruido ajustando los histogramas de la media a los de una distribución γ. Es por tanto necesaria una estimación rápida y fiable de los parámetros de dicha distribución γ. MLNE muestra el camino para una buena estimación de estos parámetros.

3.5.1.2 Derivación de los estimadores MLE para los parámetros de la distribución genérica Gamma [2]

Dado un conjunto de *N* variables aleatorias independientes e idénticamente distribuidas, $\mathrm{X} = \{\mathrm{X}_n\}_{n=1}^{N}$ con una distribución gamma $\gamma(\mu,\nu)$, las Funciones de Densidad de Probabilidad (PDF: Probability Density Function) de X_n son:

$$f_\gamma(x_n;\mu,\nu) = \frac{\nu^\mu}{\Gamma(\mu)} x_n^{\mu-1} e^{-\nu x_n} \tag{3.1}$$

donde $\Gamma(.)$ es la función gamma completa y (μ,ν) son respectivamente los desconocidos parámetros de escala y de forma.

La función probabilidad de *N* muestras $\boldsymbol{x} = \{x_n\}_{n=1}^{N}$ de $\boldsymbol{X}$ es:

$$L_\gamma(x;\mu,\nu) = \prod_{n=1}^{N} f_\gamma(x_n;\mu,\nu) = \left(\frac{\nu^\mu}{\Gamma(\mu)}\right)^N e^{N((\mu-1)u_1 - \nu u_2)} \tag{3.2}$$

donde $\boldsymbol{u} = (u_1, u_2)$ es la muestra de la estadística bidimensional

$$U = (U_1, U_2) = \frac{1}{N}\left(\sum_{n=1}^{N} \ln(X_n), \sum_{n=1}^{N} X_n\right) \tag{3.3}$$

La estadística $\boldsymbol{U}$ es suficiente para conocer el parámetro bidimensional $\theta = (\mu,\nu)$ y los estimadores $\hat{\mu}$ y $\hat{\nu}$ de μ y ν satisfacen el sistema de ecuaciones:

$$\begin{cases} \Psi(\hat{\mu}) - \dfrac{1}{\hat{\mu}} - \ln(\hat{\nu}) = u_1 \\ \dfrac{\hat{\mu}}{\hat{\nu}} = u_2 \end{cases} \tag{3.4}$$

donde $\psi(\hat{\mu}) = \dfrac{d\ln(\Gamma(\hat{\mu}+1))}{d\hat{\mu}}$ es la llamada función *psi*.

Este sistema de ecuaciones no admite una solución analítica debido a las funciones *psi* y logaritmo. Una solución numérica es costosa en tiempo computacional, y nosotros estamos interesados en una obtención rápida de las estimaciones de μ y ν. Se propone por tanto una aproximación analítica a las soluciones del sistema (3.4).

La función *psi* puede ser expandida en series:

$$\Psi(\hat{\mu}) = \ln(\hat{\mu}) + \frac{1}{2\hat{\mu}} - \frac{1}{12\hat{\mu}^2} + o(\frac{1}{\hat{\mu}^4}) \tag{3.5}$$

Aproximar $\Psi(\hat{\mu})$ en la ecuación (3.4) por $o(\frac{1}{\hat{\mu}^2})$ conduce a la aproximación de los estimadores:

$$\hat{\mu} = \frac{3+\sqrt{9+12(\ln(\boldsymbol{u}_2)-\boldsymbol{u}_1)}}{12(\ln(\boldsymbol{u}_2)-\boldsymbol{u}_1)} \tag{3.6}$$

$$\hat{\nu} = \frac{3+\sqrt{9+12(\ln(\boldsymbol{u}_2)-\boldsymbol{u}_1)}}{12\boldsymbol{u}_1(\ln(\boldsymbol{u}_2)-\boldsymbol{u}_1)} \tag{3.7}$$

Nótese que la concavidad de la función logaritmo asegura que $\ln(\boldsymbol{u}_2)-\boldsymbol{u}_1$ es positivo para cualquier muestra de *N* elementos de $\boldsymbol{x}$. Las estimaciones $\hat{\mu}$ y $\hat{\nu}$ son por tanto reales y positivas.

Tal y como se demuestra en [2], los estimadores son obtenidos de forma precisa, siendo el error relativo cometido en la aproximación menor a un 1%. Además, como muestran las ecuaciones (3.6) y (3.7), los estimadores dependen de las estadísticas del espectrograma $\boldsymbol{u}_1$ y $\boldsymbol{u}_2$. Estas estadísticas cambian en cada iteración del algoritmo, dado que en cada una de ellas se realiza extracción de señal determinista o de voz. Esto conduce a una versión cada vez más precisa de la distribución estimada del ruido. Por tanto, a medida que se realicen más iteraciones en el programa, la estimación de la potencia de ruido se irá haciendo cada vez más precisa. La discusión sobre la convergencia de esta estimación queda realizada en la Sección 4.4.3.4.2

La aplicación de los estimadores (3.6) y (3.7) en la PDF de los coeficientes de un espectrograma formado por ruido WGN conduce a los siguientes resultados. Recordemos en primer lugar que el espectrograma discreto de una secuencia discreta $s[n]$ es el módulo al cuadrado de su Transformada de Fourier computada con una ventana de ponderación *f* formada por N_f puntos:

$$S_x[n,k] = \frac{1}{N_f}\left|\sum_{m=1}^{M} f[i-n]x[i]e^{-i2\pi(ik/N_f)}\right|^2 \quad \textbf{(3.8)}$$

Cuando la señal de entrada $s[n]$ es un proceso Gausiano de media cero y varianza σ^2, los coeficientes del espectrograma $S[n,k]$ calculados con una ventana rectangular tienen una PDF $\left(\sigma^2/2\right)\chi_1^2$ si k=0, N_f/2 y $\left(\sigma^2/2\right)\chi_2^2$ en otro caso [4]. Llamemos a F_1 la media de M coeficientes WGN del espectrograma. Si los coeficientes del espectrograma son muestras de M, por tanto variables $\left(\sigma^2/2\right)\chi_2^2$, F_1 es una muestra de una variable $\left(\sigma^2/2M\right)\chi_{2M}^2$. La distribución $\alpha\chi_d^2$ es un caso especial de la distribución $\gamma(\mu,\nu)$ con $\mu = \frac{\boldsymbol{d}}{2}$ y $\nu = \frac{1}{2\alpha}$ donde d es un entero positivo [5]. A pesar de todo, existen diversas fuentes de correlación en los coeficientes del espectrograma: el uso de una ventana incrementa la correlación a lo largo del eje de la frecuencia [6], y el uso del zero-padding y el solapamiento de la ventana entre los segmentos temporales, incrementa la correlación a lo largo del eje temporal [1]. La suma de χ^2 correladas, induce un suavizamiento de la distribución, y el parámetro d pasa a tomar valores reales, con $d \le 2M$. Debemos por tanto considerar la media F_1 como una distribución $\gamma(\mu,\nu)$ con parámetros desconocidos $\mu = \frac{\boldsymbol{d}}{2}$ y $\nu = \frac{\boldsymbol{d}}{2\alpha^2}$, siendo d real. Nótese que $\mu/\nu = \sigma^2$.

Por tanto la estadística $\boldsymbol{U}$ de la ecuación (3.3) es suficiente para la estimación de los parámetros (μ,ν). La teoría general de las estadísticas suficientes prueba que la estadística $\boldsymbol{U}_2$ es un eficiente estimador de la función $\frac{\mu}{\nu}$. La muestra de $\boldsymbol{U}_2$ es $\boldsymbol{u}_2 = \hat{\mu}/\hat{\nu}$ (ver ecuación (3.4)), y por tanto, en el caso de los coeficientes de ruido del espectrograma, $\boldsymbol{u}_2 = \sigma^2$. La estadística $\boldsymbol{U}_2$ es por consiguiente un estimador eficiente de la potencia de ruido σ^2. La estimación de $\hat{\sigma}^2$ viene dada por:

$$\sigma^2 = \boldsymbol{u}_2 = \frac{\hat{\mu}}{\hat{\nu}} \quad \textbf{(3.9)}$$

donde $\hat{\mu}$ y $\hat{\nu}$ son los estimadores de la ecuación (3.6).

3.5.1.3 Contribución al Algoritmo de Segmentación.

La teoría del MLNE es usada para estimar de forma fiable los parámetros de la distribución gamma, i.e. la estimación de la función de distribución de ruido. Los datos observados en este caso son la media local y la desviación estándar de las sub-imágenes[2] del espectrograma. A partir de ellos, se realiza la estimación de la distribución de ruido, necesaria para la determinación de un límite en la selección y propagación de semillas. La incidencia en el algoritmo de segmentación de este límite es explicada con claridad en la Sección 4.4.3.2.

Tras la primera iteración del algoritmo, donde se ha realizado la primera extracción de señal, vuelve a realizarse una nueva estimación de la distribución de ruido, obteniendo por tanto un nuevo límite que guíe la continuación de la extracción de la señal en la nueva iteración. Claramente, a medida que se extraen más grupos de regiones de señal determinista, va quedando en el espectrograma una mayor proporción de píxeles de ruido. Por este motivo la estimación del ruido es cada vez más precisa. Este proceso se repite siempre al final de cada iteración, tras la última extracción de señal, obteniendo para la siguiente iteración una estimación más precisa de la distribución del ruido. Cuando esta estimación alcanza un determinado nivel de exactitud, el proceso de segmentación se detiene.

De forma breve, la distribución estimada actúa como guía de toda la segmentación y como condición de parada del algoritmo.

[2] A cada píxel se le asocia un grupo de píxeles vecinos, llamando a todo el conjunto célula. La media local y la desviación estándar local son la media y la desviación estándar de la célula (sub-imagen). Cada píxel queda por tanto representado por su media local. En la Sección 4.4.3.1.1 se explica este proceso.

CAPÍTULO 4
FILTRADO MORFOLÓGICO

El presente capítulo describe todo el proceso de filtrado morfológico realizado sobre la señal de entrada al sistema. Al haber introducido la conversión de onda temporal a espectrograma en la Sección 2.4.2, este capítulo sólo se centrará en las operaciones de filtrado morfológico para la consecución de la segmentación. Primeramente se explica el método en el que está basado dicho proceso. A continuación se plantea su interpretación estadística, seguida de las propiedades estadísticas de las llamadas *Características* seleccionadas del espectrograma.

Posteriormente se ilustra el diagrama de flujo de la función principal del algoritmo (Figura 4.5), detallando a continuación cada uno de los módulos que lo componen y explicando la teoría en la que están basados. Paralelamente, se irán mostrando ilustraciones del proceso. Por último, se exponen dos ejemplos completos de segmentación paso a paso de dos tipos de señales distintas inmersas en ruido Gausiano blanco. En primer lugar con señales sintéticas chirp, y posteriormente con señales de voz.

4.1 MÉTODO

Como ya se ha comentado anteriormente, un espectrograma es considerado como una imagen formada por píxeles, donde cada uno de ellos está caracterizado por un nivel de energía llamado coeficiente Tiempo-Frecuencia. Segmentar un espectrograma consiste en decidir si un coeficiente es un componente de una región determinista o si pertenece a una región de ruido. Desde el punto de vista visual, se puede afirmar qué píxeles forman la región de señal determinista (alta energía) y cuáles la región de ruido (baja energía). A pesar de todo, una segmentación precisa no puede realizarse únicamente basándose en el nivel de energía individual de cada píxel debido a dos razones:

1. Desigualdad Heisenberg-Gabor: esta teoría prueba que la energía contenida en un espectrograma en un instante *n* y frecuencia *k* recae sobre una vecindad de ese punto del espectrograma. Por tanto, el nivel de energía de un píxel no puede describir completamente el contenido de señal en el instante *n* y en la frecuencia *k*.
2. El ruido que corrompe la señal afecta aleatoriamente a los coeficientes del espectrograma. Por tanto, los vecinos de un píxel deben ser tenidos en cuenta para obtener una media de la potencia del ruido.

Se propone por tanto la asociación de otras características distintas al nivel de energía de cada localización. Para tener en consideración el principio de indeterminación, definimos dos *Características* que envuelven a grupos de coeficientes del espectrograma. Estas características son estadísticas de los coeficientes del espectrograma.

Primeramente, es necesaria la asociación de cada píxel a una célula formada por coeficientes vecinos de ese píxel. El tamaño *N* de la célula debe ser pequeño en comparación con el número total de puntos del espectrograma para poder asegurar así una descripción local del espacio Tiempo-Frecuencia. Una justificación más detallada sobre la elección del tamaño de la célula queda mostrada en la Sección 4.4.1

Segmentar el espectrograma requiere la asociación de una etiqueta a cada localización de éste, denominándola ruido, o señal más ruido. Como este procedimiento de segmentación es dificultado tanto por la desigualdad Heisenberg-Gabor como por la corrupción del ruido en el nivel de energía, se propone la transformación del nivel de energía del espectrograma. A continuación se calculan dos conjuntos de estadísticas sobre cada célula. En un nuevo espacio, referido en adelante como *Espacio Característico*, cada localización Tiempo-Frecuencia del espectrograma es posicionado con respecto a las *Características* estadísticas extraídas de cada célula formada por vecinos de un píxel, tal y como se muestra en la Figura

4.1. De esta forma, en el *Espacio Característico* se forman racimos formados por células con similares propiedades estadísticas. A continuación, se procesa un algoritmo de región creciente que segmenta asociando una etiqueta común a puntos conectados en el espectrograma que posean las mismas propiedades estadísticas en el *Espacio Característico*. Se demuestra que esta técnica es estable con respecto al ruido, y por tanto puede ser aplicada a un gran rango de señales sin especificaciones. Una de las grandes ventajas es que suministra una caracterización de los patrones espectrales sin conocimiento previo alguno sobre su situación u orientación en el espectrograma, basándose por tanto únicamente en las variaciones de la magnitud. De esta manera, no es necesario el ajuste de parámetros de tiempo, como ocurre en la mayoría de los procedimientos existentes, los cuales requieren una estimación del tiempo de inicio del patrón. La elección de las características es tal que facilitan parámetros descriptivos de las estructuras del espectrograma.

Estas *Características*, como combinación de variables aleatorias, son otras variables aleatorias. En el *Espacio Característico*, aparecen como racimos cuya localización y dispersión son, respectivamente, medidos por su media y varianza. A continuación mostramos un estudio teórico de estas propiedades estadísticas que permiten prever la posición de los patrones espectrales en el *Espacio Característico*.

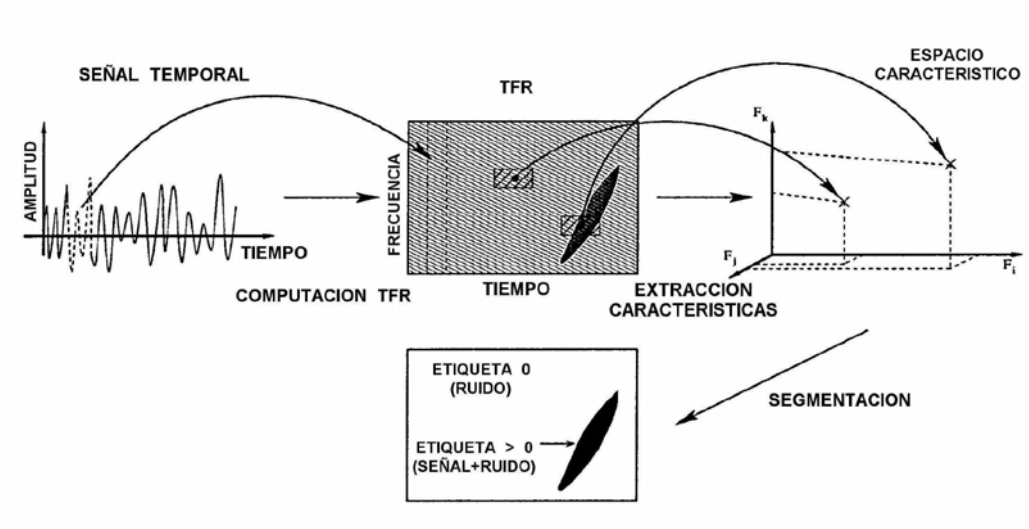


Figura 4.1 : Visión general del método.

4.2 INTERPRETACIÓN ESTADÍSTICA

En esta Sección se deriva la Función de Densidad de Probabilidad de los coeficientes del espectrograma de una secuencia determinista contaminada por ruido WGN. Posteriormente se propone un modelo estadístico local del espectrograma.

4.2.1 Propiedades estadísticas del espectrograma

Consideremos la señal $x[m]$, la cual es suma de una determinada secuencia discreta $d[m]$ de M muestras, y de un proceso Gausiano blanco $b[m]$ de media cero y varianza σ^2.

$$x[m] = d[m] + b[m], \text{ con } 1 \le m \le M \quad \textbf{(4.1)}$$

El espectrograma discreto $S_x[n,k]$ en el instante de tiempo *n* y frecuencia *k* de $x[m]$ es el periodograma de la señal "weighted" por una ventana $f[m]$ de $M_f \le M$ muestras

$$S_x[n,k] = \frac{1}{M_f}\left|\sum_{m=1}^{M} f[m-n]x[m]e^{-i2\pi(mk/M_f)}\right|^2 \quad \textbf{(4.2)}$$

Cuando $f[m]$ es una ventana rectangular, es decir, vale 1 para $1 \le m \le M_f$ y cero en otro caso, los coeficientes del espectrograma del ruido Gausiano blanco $b[m]$ tienen una distribución central χ^2 con dos grados de libertad y parámetro de proporcionalidad $\sigma^2/2$

$$S_b[n,k] \sim \frac{\sigma^2}{2}\chi_2^2, \text{ si } k \ne 0 \text{ y } k \ne \frac{M_f}{2} \quad \textbf{(4.3)}$$

si M_f tiende al infinito. Es bien conocido que la PDF $f_{\alpha\chi_D^2}(x)$ es una distribución gamma $\gamma(D/2, 1/2\alpha)$

$$f_{\alpha\chi_D^2}(x) = \frac{1}{(2\alpha)^{D/2}\,\Gamma\left(\frac{D}{2}\right)} x^{D/2-1}\exp\left(-\frac{x}{2\alpha}\right) \quad \textbf{(4.4)}$$

donde $\alpha = \sigma^2/2$ es el parámetro de proporcionalidad, y $D = 2$ es el número de grados de libertad.

La secuencia $x[m]$ de (4.1) es un conjunto de variables Gausianas independientes $d[m]$ de media distinta de cero y varianza σ^2. Se concluye que $S_x[n,k]$ es una distribución χ^2 no-central con dos grados de libertad, parámetros no centrales $S_d[n,k]$ y parámetro de proporcionalidad $\sigma^2/2$.

$$S_x[n,k] \sim \frac{\sigma^2}{2}\chi^2_{2,S_d[n,k]}, \text{ si } k \ne 0 \text{ y } k \ne \frac{M_f}{2} \quad \textbf{(4.5)}$$

Ver el Apéndice A para las expresiones de distribuciones χ^2 no centrales. La distribución χ^2 central de (4.3) es un caso especial de χ^2 no central con un parámetro no central nulo.

Nótese que un pequeño número de coeficientes del espectrograma puede tener diferentes distribuciones debido al uso de distintas ventanas en el tiempo. No las tenemos en cuenta en el modelo de distribución. El comportamiento de las variables aleatorias queda

totalmente caracterizado por sus momentos. En el Apéndice A se propone una expresión general del momento *q* sobre una distribución χ^2 no-central. En particular, la media y la varianza esperada de $S_x[n,k]$ son derivadas de :

$$E\{S_x[n,k]\} = \sigma^2 + S_d[n,k] \quad \textbf{(4.6)}$$

$$Var\{S_x[n,k]\} = \sigma^4 + 2\sigma^2 S_d[n,k] \quad \textbf{(4.7)}$$

Estos valores se incrementan linealmente con el parámetro no-central $S_d[n,k]$. Este parámetro describe el contenido de la señal determinista $d[m]$ únicamente en el instante *n* y frecuencia *k*. Si $d[m]$ es un proceso no-estacionario, entonces la no-estacionariedad a ser analizada está contenida en los momentos anteriores por el parámetro no-central $S_d[n,k]$. El conjunto de parámetros no-centrales es por tanto una firma de la no-estacionariedad.

4.2.2 Modelo Estadístico Local

Cada célula $C(n,k)$ es un conjunto de N coeficientes del espectrograma con PDFs χ^2 con parámetros no estacionarios $S_d[n,k]$ y varianza σ^2. Por tanto, la variable padre $S_C[n,k]$ asociada a $C(n,k)$ tiene una PDF definida por $N+1$ parámetros desconocidos. Para definir características tales como las estadísticas de esta variable padre, debe definirse un modelo de la célula para reducir el número de parámetros desconocidos.

Consideremos una célula[3] *C*, donde *P* de los *N* puntos contienen energía de la señal determinista. Cada uno de los *P* coeficientes asociados es una distribución χ^2 no-central con parámetros no-centrales $S_d[n,k]$. Como el tamaño de la célula *C* es pequeño en comparación al tamaño del espectrograma (lo cual garantiza un acercamiento local), la contribución de energía de la señal determinista puede ser considerada como una variación lenta sobre los *P* coeficientes. El parámetro no-central *P* puede entonces ser aproximado por el mismo parámetro *S*, el cual es la media de $S_d[n,k]$ de toda la célula

$$S = \frac{1}{P}\sum_C S_d[n,k] \quad \textbf{(4.8)}$$

Los restantes $N-P$ coeficientes son distribuciones χ^2 centrales. Por tanto, cada coeficiente $S_x[n,k]$ de *C* es una muestra de la variable aleatoria $\left(\sigma^2/2\right)\chi^2_{2,S}$ con probabilidad $p = P/N$ y una muestra de la variable aleatoria $\left(\sigma^2/2\right)\chi^2_2$ con probabilidad $1-p$ (ver

[3] Los índices (n, k) son omitidos cuando tratemos una única célula sin confusión.

Figura 4.2). Gracias a la fórmula de la probabilidad total, cada coeficiente de *C* puede ser considerado como una muestra de la variable padre S_C, cuya PDF $f(x;p,S)$ es una mezcla de PDFs de χ^2:

$$f(x;p,S) = pf_{(\sigma^2/2)\chi^2_{2,S}}(x) + (1-p)f_{(\sigma^2/2)\chi^2_2}(x) \tag{4.9}$$

donde $f_X(x)$ es la PDF de una variable aleatoria X. La ecuación (4.9) es el modelo estadístico propuesto para aplicar a la célula. Bajo la condición (4.8), la PDF de la variable

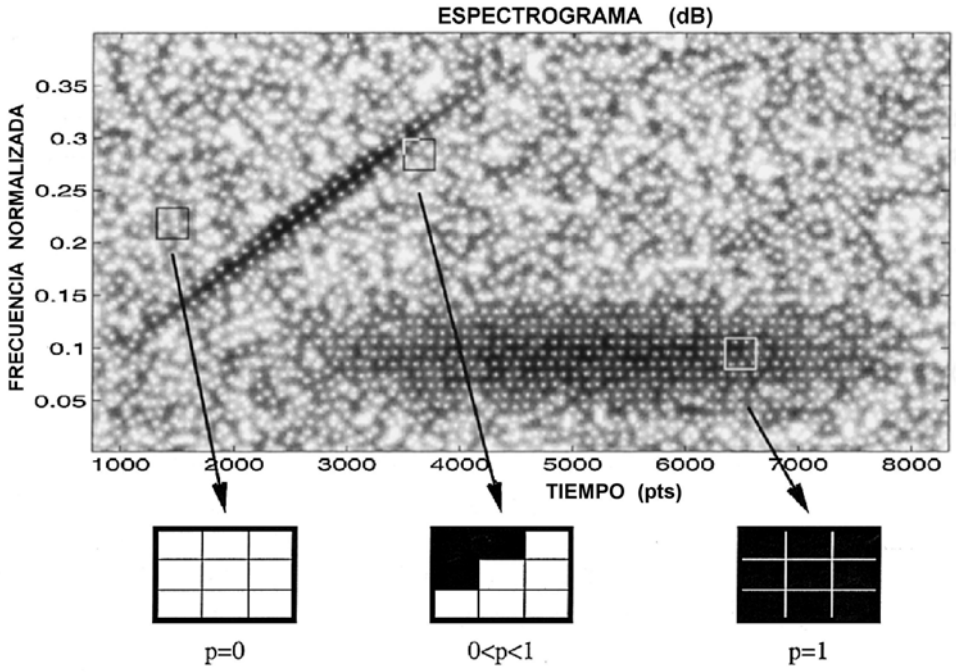


Figura 4.2 : Descripción de la célula. Los cuadrados blancos son coeficientes del espectrograma χ^2 centrales, mientras que los negros son coeficientes del espectrograma χ^2 no-centrales.

padre depende solamente de los tres parámetros desconocidos p, σ^2 y *S*. La propiedad de linealidad de la Transformada de Fourier lleva a escribir la primera función característica $\phi(u;p,S)$ de la variable padre como

$$\phi(u;p,S) = p\phi_{(\sigma^2/2)\chi^2_{2,S}}(u) + (1-p)\phi_{(\sigma^2/2)\chi^2_2}(u) \tag{4.10}$$

El momento q $\mu_q = E\left\{S_C^q\right\}$ de la variable padre es derivado de (4.10):

$$\mu_q = (-i)^q \frac{d^q\phi(u;p,S)}{du^q} = p\mu_q^d + (1-p)\mu_q^c \tag{4.11}$$

Las expresiones de μ_q^d y μ_q^c, momentos q de las distribuciones χ^2 central y no central son derivadas de la ecuación A.10 del Apéndice A, con $\alpha = \sigma^2/2$, $D = 2$ y $\Theta = S$ para μ_q^d, y $\Theta = 0$ para μ_q^c. Tenemos por tanto

$$\mu_q = q!(\sigma^2)^q \left[1 + p \sum_{n=1}^{q} \binom{n}{q} \frac{1}{n!} r^n \right] \quad \textbf{(4.12)}$$

donde $r = S/\sigma^2$ es la relación señal a ruido local calculada sobre la célula. Nótese que esta interpretación es habitual en Teoría de la señal, donde el parámetro no-central es asimilado a una relación señal a ruido (SNR).

Finalmente, cada célula es definida sobre el espectrograma y descrita por su variable padre. Los coeficientes del espectrograma de la célula son muestras de esta variable aleatoria. Sus momentos, los cuales son expresados en (4.12), dependen de dos parámetros:

- $p = P/N$: proporción de los coeficientes del espectrograma que poseen señal determinista.
- $r = S/\sigma^2$: SNR de la célula.

La variación simultánea de estos parámetros sobre todo el espectrograma es relacionado con las variaciones de $S_d[n,k]$ a lo largo de las dimensiones *n* y *k*. Además, caracterizan las variaciones de la magnitud de los componentes de señal. Buscamos extraer las características cuyo comportamiento estadístico, el cual es relacionado con las variaciones de *p* y *r*, suministre una discriminación obvia de puntos que pertenezcan a patrones espectrales de diferentes variaciones de magnitud. Para limitar los efectos del ruido en la lectura del espectrograma, seleccionamos como primera característica el valor esperado de la variable padre

$$E\{S_C\} = \mu_1 \quad \textbf{(4.13)}$$

Esta característica es relevante para la caracterización de las regiones en el espectrograma de poca densidad de variación de energía. A continuación, se propone la combinación de este proceso con el de la extracción de una segunda característica: la desviación estándar de la célula

$$SD\{S_C\} = \sqrt{Var\{S_C\}} = \sqrt{\mu_2 - \mu_1^2} \quad \textbf{(4.14)}$$

Esta característica determina las variaciones de densidad de alta energía del espectrograma.

4.3 PROPIEDADES ESTADÍSTICAS DE LAS CARACTERÍSTICAS

En la sección previa, las C*aracterísticas* son definidas como el valor esperado y la desviación-estándar de las variables padres asociadas a cada célula local en el espectrograma. En esta Sección, se proponen estimadores de estas características. Asimismo, también se dan las expresiones de sus dos primeros momentos. Estas propiedades estadísticas son necesarias

para describir los racimos que se obtienen en el *Espacio Característico*. Además, se hace una breve discusión de la influencia de la ventana de tiempo $f[m]$ sobre la distribución de los coeficientes del espectrograma, dado que la teoría derivada en la sección previa asume independencia de estos coeficientes.

4.3.1 Media Local

Asumiendo ergodicidad, esto es, los parámetros estadísticos de una muestra o secuencia de un proceso son representativos de las del proceso en conjunto, la característica de (4.13) es estimada extrayendo la media empírica de la célula

$$F_1 = \frac{1}{N}\sum_C S_x[n,k] \tag{4.15}$$

Tal y como se demuestra en el Apéndice B, (4.15) es un estimador imparcial de $E\{S_C\}$. Cuando la célula contiene energía de la componente determinista, la distribución de F_1 es la convolución repetida *N* veces de la distribución con un parámetro no central no nulo *S* si los coeficientes del espectrograma son independientes. No daremos la expresión analítica de esta distribución, pero en el Apéndice B se derivan expresiones del primer y segundo momento de F_1, las cuales son, respectivamente, polinomios de primer y segundo orden de *r* y *p*. Considerando el caso de células que contienen sólo coeficientes de ruido del espectrograma ($r = 0$ y $p = 0$), las ecuaciones (B.4) y (B.5) del Apéndice B toman la forma:

$$E\{F_1\} = \sigma^2 \tag{4.16}$$

$$Var\{F_1\} = \frac{\sigma^4}{N} \tag{4.17}$$

En este caso de sólo ruido, F_1 es una variable aleatoria $\left(\sigma^2/2N\right)\chi^2_{2N}$ tal y como se demuestra en [1].

4.3.2 Correlación de los coeficientes

En muchas situaciones, la representación Tiempo-Frecuencia presenta redundancia de información que significa correlación entre los coeficientes de la representación. Comencemos considerando el espectrograma de un proceso Gausiano blanco. Sus coeficientes a lo largo del eje del tiempo están correlados si la ventana temporal se solapa. A lo largo del eje de la frecuencia están asintóticamente incorrelados, pero el uso de una ventana de ponderación también introduce correlación a lo largo del eje de la frecuencia. Tal y como

se demuestra en [1], el solapamiento de la ventana temporal, así como el uso de una ventana Hanning, inducen a un suavizamiento de la PDF de F_1.

Tal y como veremos en la siguiente sección, el procedimiento de segmentación está controlado por la PDF de la característica F_1 extraída de las células que contienen sólo energía de ruido. La teórica PDF extraída en la Sección 4.2 no es válida en presencia de correlación, siendo ésta la situación la mayoría de las ocasiones. Por tanto, debe ser considerado el efecto de la correlación en la verdadera PDF. Supongamos que las características F_1 son $\left(\sigma^2/D\right)\chi_D^2$ con valores σ^2 y D inespecificados. Tal y como se ha demostrado en la Sección 3.5.1.2, los MLNE $\hat{\sigma}^2$ y $\hat{\boldsymbol{D}}$ de una distribución $\left(\sigma^2/D\right)\chi_D^2$ son aproximados de forma precisa por

$$\hat{\sigma}^2 = M \tag{4.18}$$

$$\hat{\boldsymbol{D}} = \frac{3+\sqrt{9+12(\ln(\boldsymbol{M})-\boldsymbol{M}_l)}}{6(\ln(\boldsymbol{M})-\boldsymbol{M}_l)} \tag{4.19}$$

con $M = 1/N_S \sum_{N_S} F_1$, $M_l = 1/N_S \sum_{N_S} \ln(F_1)$ y N_S el número de coeficientes del espectrograma. Las estadísticas M y M_l son suficientes para el número de grados de libertad del parámetro no-central. La varianza del proceso Gausiano $b[m]$ es entonces eficientemente estimada por $\hat{\sigma}^2$ con un bajo coste computacional, tal y como se demuestra en [1].

La conclusión es que sea cual sea la forma de computación del espectrograma, la potencia de ruido puede ser estimada de forma precisa considerando la medida F_1 calculada sobre toda la célula de ruido como una variable $\alpha\chi_D^2$ con $\alpha = \sigma^2/D$.

En presencia de componentes deterministas, el histograma de F_1 es una mezcla de PDFs χ^2 centrales y no centrales. Un esquema de clasificación binario basado en un algoritmo de maximización de la esperanza, por ejemplo, debería ser usado para identificar datos de la mezcla. En lugar de este procedimiento, se propone añadir al proceso una segunda característica F_2.

4.3.3 Desviación estándar Local

Asumiendo ergodicidad, la característica de (4.14) es extraída estimando la desviación estándar F_2 de la célula

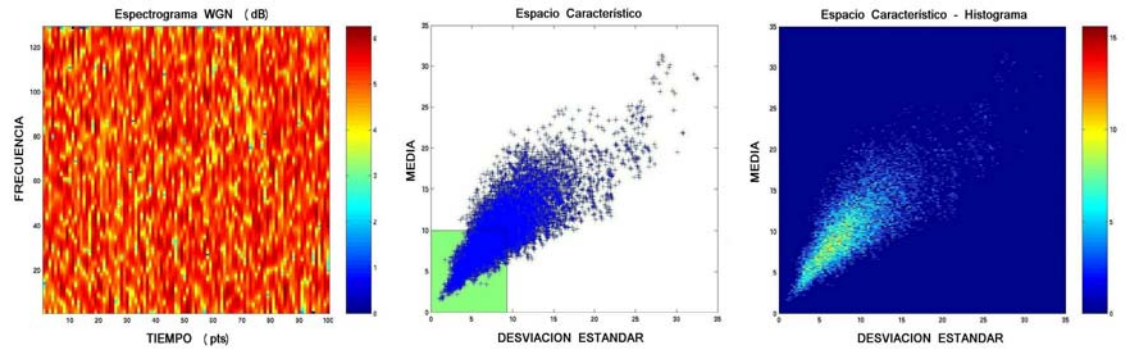
$$F_2 = \left[\frac{1}{N} \sum_{C} \left(S_x[n,k] - F_1 \right)^2 \right]^{1/2} \quad \textbf{(4.20)}$$

El primer y segundo momento de F_2 son proporcionados en el Apéndice B por (B.9) y (B.10) respectivamente. El caso de $p = r = 0$ es el caso de una célula de ruido. Las últimas expresiones toman por tanto la forma

$$E\{F_2\} = \left[1 - \frac{3}{N} + \frac{3}{2N^2} \right]^{1/2} \sigma^2 \quad \textbf{(4.21)}$$

$$Var\{F_2\} = \frac{4N-3}{2N^2} \sigma^4 \quad \textbf{(4.22)}$$

Las ecuaciones (4.16) y (4.21) muestran que los puntos de ruido forman un racimo en el *Espacio Característico* localizado en torno a los valores $\left(E\{F_1\} = \sigma^2, E\{F_2\} = \left[1 - (3/N) + (3/2N^2) \right]^{1/2} \sigma^2 \right)$. La Figura 4.3 ilustra este resultado en el caso de un proceso Gausiano de media cero y varianza $\sigma^2 = 10$ con $N = 21$. Tal y como se esperaba por (4.16) y (4.21), el racimo está localizado en torno a $\left(E\{F_1\} = 10, E\{F_2\} = 9.3 \right)$. La Figura 4.3 (a) es el espectrograma de la señal. Las Figura 4.3 (b) y (c) muestran el *Espacio Característico* normal y su función de densidad respectivamente.



(a)	(b)	(c)

Figura 4.3 : (a) Espectrograma de un proceso Gausiano blanco de media cero y varianza $\sigma^2 = 10$. (b) *Espacio Característico* asociado $\left(F_1, F_2 \right)$. (c) Histograma (función de densidad) del *Espacio Característico*. Tamaño de la célula usado: 3x7.

Consideremos ahora de nuevo el caso de una componente determinista inmersa en ruido Gausiano blanco. Los parámetros *r* y *p* se incrementan cuando la célula se desliza desde puntos ruidosos a patrones de componentes deterministas. La evolución de $E\{F_1\}$ y $E\{F_2\}$ es descrita en términos de *r* y *p* por (B.4) y (B.10). Una simulada red de curvas

$\left(E\{F_1\}, E\{F_2\}\right)$ parametrizada por *r* y *p* es mostrada en la Figura 4.4 para ilustrar esta evolución. La firma del patrón espectral determinístico en el *Espacio Característico* es un racimo curvado que se extiende desde la región ruidosa al área de *r* y *p* distintas de cero, dependiendo de la magnitud máxima del patrón, y del tamaño del patrón respecto al tamaño de la célula. La forma del racimo depende de las variaciones simultáneas de *r* y *p*, describiendo las variaciones de magnitud del patrón espectral. Consideremos por ejemplo un patrón espectral con una SNR local constante, el cual es un caso extremo de patrón espectral de borde afilado. Cuando la célula se desliza a lo largo del patrón, la proporción *p* crece primero y posteriormente decrece, mientras que la SNR local, *r*, es constante. Su representación en el *Espacio Característico* es un racimo siguiendo una línea *r* (línea amarilla).

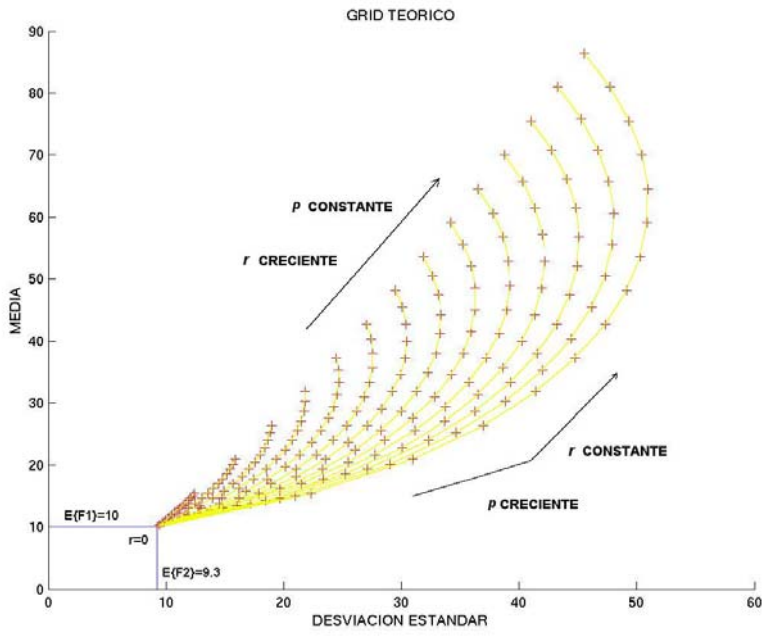


Figura 4.4 : Grid Teórico $\left(E\{F_1\}, E\{F_2\}\right)$ computado con una varianza de ruido $\sigma^2 = 10$ y una célula de $N = 7x3$ puntos. El punto $\left(E\{F_1\} = 10, E\{F_2\} = 9.3\right)$ es el valor de ruido esperado. Los 15 valores del los parámetros *r* y *p* son regularmente espaciados entre, respectivamente, [0,6] y [0,1] (+).

El acercamiento local permite simplificar el modelo estadístico de la célula. Las propiedades derivadas de las C*aracterísticas* dependen de los dos parámetros caracterizados *r* y *p*. Esto provee una descripción del *Espacio Característico* que debe ser usada tanto en la segmentación del espectrograma como en la descripción de las variaciones de magnitud de los patrones espectrales extraídos. Una mayor descripción de la obtención del *Grid Teórico* es realizada en la Sección 4.4.3.2.3.

4.4 ALGORITMO DE SEGMENTACIÓN EN EL ESPACIO CARACTERÍSTICO

Antes de explicar el algoritmo de segmentación, es necesaria la argumentación de la elección del tamaño de la célula. Posteriormente se expondrá el diagrama de flujo de la función principal, para continuar explicando detalladamente cada uno de los módulos que la componen. Por último, se realiza una demostración del algoritmo paso a paso mediante ilustraciones de cada tipo de señal estudiado.

4.4.1 Tamaño de la célula

La manera en que las *Características* se agregan al *Espacio Característico* depende del tamaño de la célula. Por un lado, una célula pequeña en comparación al tamaño del espectrograma asegura un acercamiento local. Por otro lado, la dispersión de los racimos del *Espacio Característico* decrece a medida que el tamaño de la célula aumenta, ya que F_1 y F_2 son estimadores consistentes de los momentos. Esto induce a una separación creciente de los datos en el *Espacio Característico*. Una caracterización local a través de células extensas requiere grandes cantidades de solapamiento y de zero-padding. La consecuencia del solapamiento y del zero-padding es el suavizamiento de los datos debido al aumento de la correlación de los coeficientes del espectrograma. Definimos el tamaño de la célula en función de la importancia de la correlación de su punto central, el cual depende de la configuración del espectrograma: tamaño y forma de la ventana, solapamiento, y zero-padding. Esto nos lleva por tanto a un compromiso entre el tamaño de la célula y la dispersión de las *Características* debido a la correlación. Esta elección también permite la caracterización de cada punto por su región de influencia.

Una justificación más detallada sobre el cálculo del tamaño de la célula a partir de estos parámetros es realizada en [1]. No obstante, los detalles sobre esta obtención no se encuentran entre los objetivos primordiales de estudio de este Proyecto, escogiendo finalmente como tamaño de la célula los valores más usados con esta técnica: células de tamaño 3x3, 3x5 o 3x7. Posteriores pruebas del algoritmo con otros tamaños justifican éstas elecciones como acertadas.

4.4.2 Diagrama Principal

A continuación mostramos el Diagrama de Flujo de la Función Principal del algoritmo:

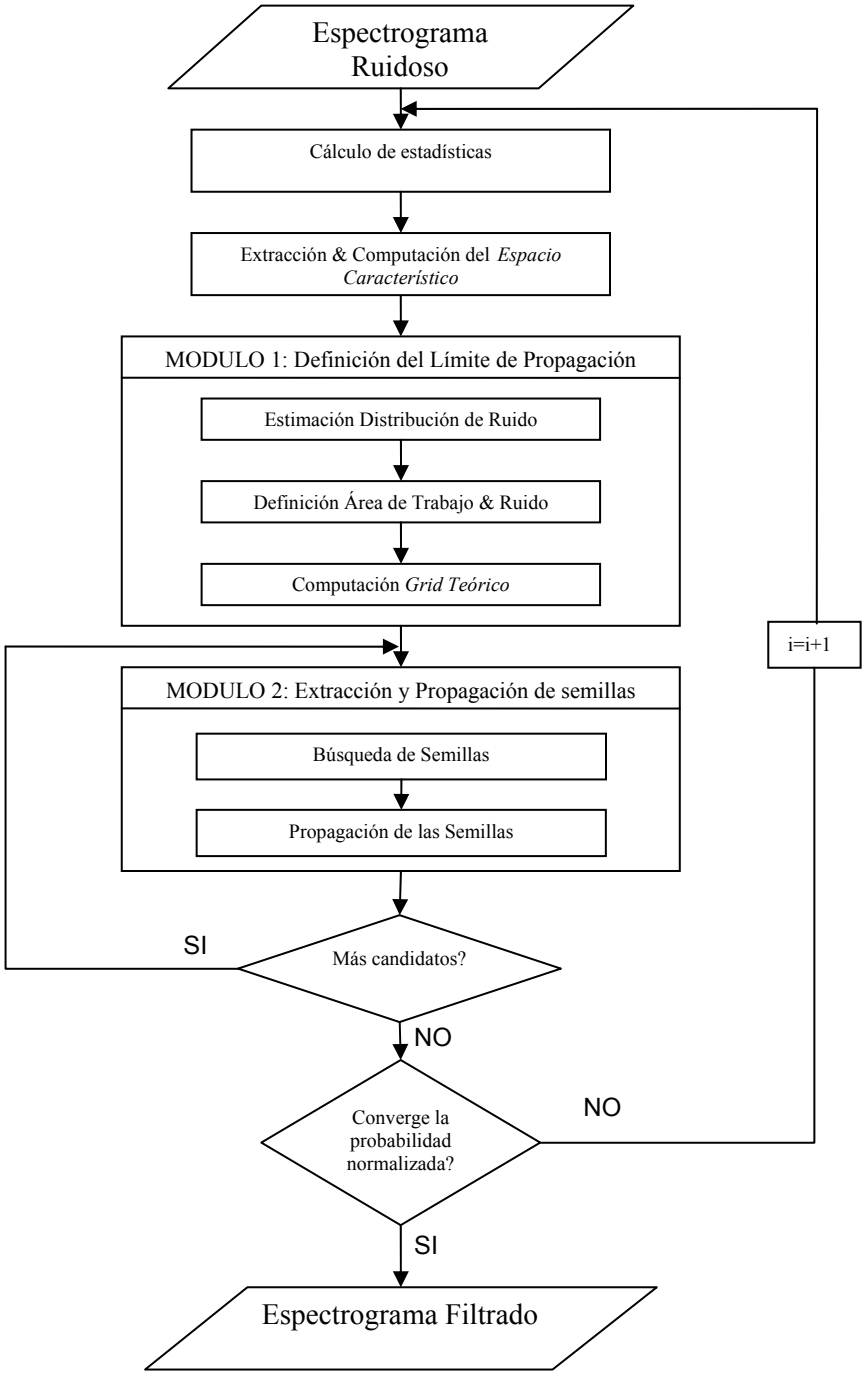


Figura 4.5 : Diagrama de Flujo de la Ffunción Principal.

4.4.3 Procedimientos

Descripción de cada uno de los módulos del algoritmo.

4.4.3.1 Extracción de las Características Locales y Computación del *Espacio Característico*

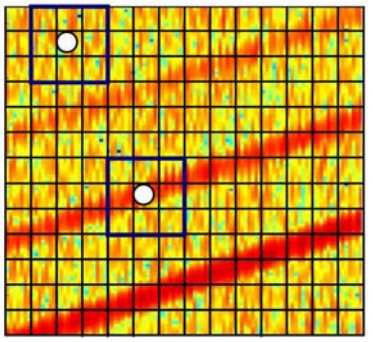
4.4.3.1.1 Extracción de las Características Locales

Dos características estadísticas locales son determinadas a partir de cada célula para caracterizar a cada píxel. Este píxel estará representado a lo largo de todo el algoritmo por dos cualidades de la célula que gobierna:

- ***Característica 1 = Media local*** = media de todos los coeficientes Tiempo-Frecuencia de los píxeles de la célula.

- ***Característica 2 = Desviación Estándar Local*** = desviación estándar de los coeficientes Tiempo-Frecuencia de los píxeles de la célula

La extracción de estos parámetros estadísticos es ilustrada en la Figura 4.6. El círculo blanco representa la variable padre (píxel central de la célula) y el cuadro azul es la célula.



$$\text{Característica 1}, F_1 = \frac{1}{N}\sum_{C} S_x[n,k]$$

$$\text{Característica 2}, F_2 = \left[\frac{1}{N}\sum_{C}(S_x[n,k] - F_1)^2\right]^{\frac{1}{2}}$$

Donde N es el número de píxeles de la célula C, y $S_x[n,k]$ es el coeficiente Tiempo-Frecuencia (i.e. energía) de un píxel en el instante n y frecuencia k.

Figura 4.6 : Ilustración de la extracción de las características locales.

Por tanto, cada píxel deja de ser caracterizado por su valor energético, y pasa a ser estudiado por las dos propiedades estadísticas de la célula que gobierna: *Característica 1* y *Característica 2*. Los píxeles pertenecientes al margen externo del espectrograma no son considerados y se dejan inalterados. El efecto de esto es inapreciable debido a que la cantidad de píxeles no estimados es despreciable frente a todos los que forman el espectrograma. Además, hay que tener en cuenta que por su situación en el espectrograma, estos píxeles son los menos importantes.

4.4.3.1.2 Computación del Espacio Característico

El llamado *Espacio Característico* es construido a partir de *Característica 1* y *Característica 2*. Estas forman las coordenadas del nuevo *Espacio Característico*, colocando *Característica 1* en el eje vertical y *Característica 2* en el horizontal. Por tanto, cada píxel en el espectrograma se corresponde con un elemento en el *Espacio Característico* a través de la célula que preside. En otras palabras, podemos imaginar los píxeles del espectrograma reposicionados con respecto a los valores de *Característica 1* y *Característica 2* tal y como muestra la Figura 4.7:

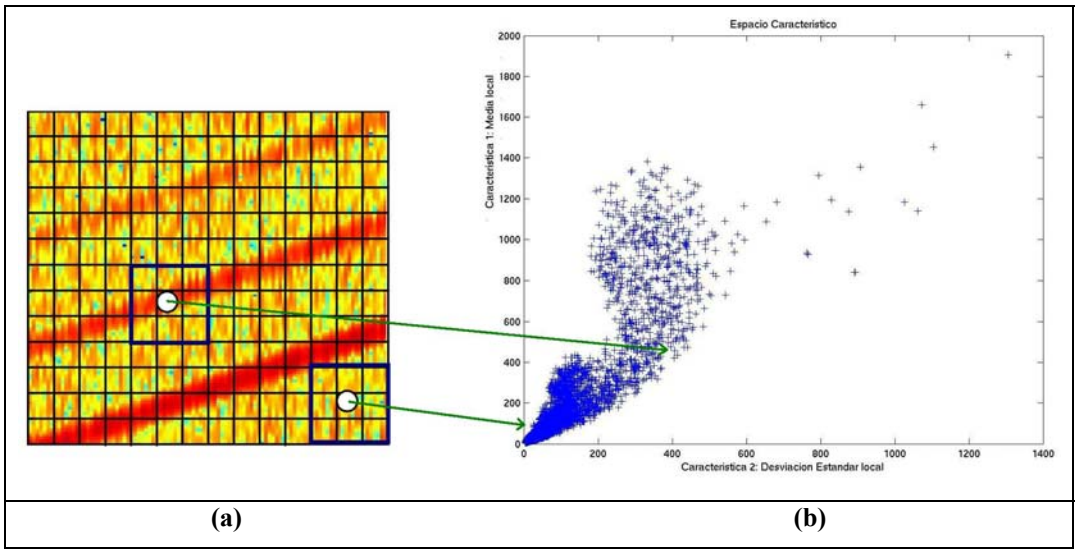


Figura 4.7 : Ilustración de la computación del *Espacio Característico* de una señal sintética formada por chirps y corrompida por ruido Gausiano blanco. (b) *Espacio Característico* correspondiente del espectrograma *(a)*.

Los beneficios de tener este *Espacio Característico* son obvios. En primer lugar, podemos analizar dos criterios a la vez. *Característica 2* indica si un píxel está situado en una región de proximidad a píxeles de alta o de baja energía, mientras que *Característica 1* indica si la región es de alta o de baja variación de energía. La localización correspondiente de cada píxel en el *Espacio Característico* determinará a lo largo del algoritmo si ese píxel pertenece a una región de señal determinista o a una región de ruido. Para tomar esta decisión es necesaria la determinación previa de varios parámetros, tales como unas estimaciones de la distribución de ruido, la cual conllevará a la definición de dos áreas de trabajo en el *Espacio Característico*. Asimismo, será necesaria la superposición de un *Grid Teórico*, el cual guiará el proceso de segmentación mediante la técnica de región creciente. La obtención de estos parámetros es realizada en el siguiente apartado.

4.4.3.2 Módulo 1: Definición del límite de Propagación

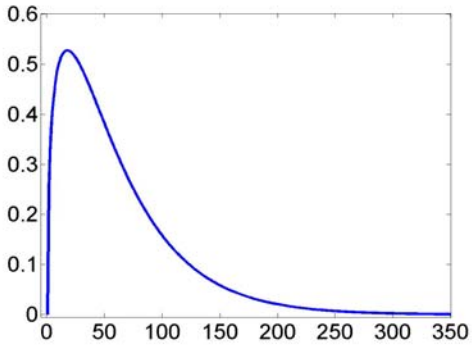
En éste módulo se realizan principalmente dos tareas. La primera, es la determinación de un límite de propagación de las semillas a partir de una estimación de la distribución de ruido. Este límite definirá dos áreas complementarias de trabajo en el *Espacio Característico*. La segunda, es el cálculo del llamado *Grid Teórico*, necesario para la extracción y propagación de las semillas. Este módulo queda incluido en el bucle principal del programa. Por tanto, en cada iteración de este bucle (llamado en adelante simplemente iteración), se realizarán los cálculos de nuevo, obteniendo valores de estimación de ruido cada vez más precisos. El módulo está formado por tres pasos:

1. Estimación de la Distribución de Ruido
2. Definición del Área de Trabajo y de la Región de Confianza de ruido
3. Computación del *Grid Teórico*

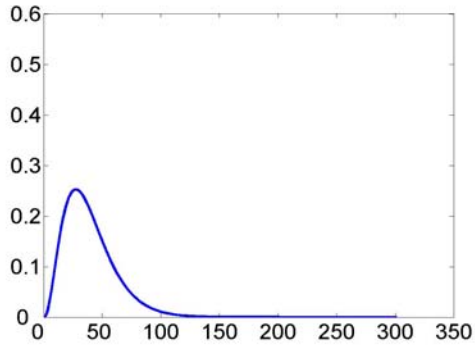
4.4.3.2.1 Estimación de la Distribución de Ruido

La función de densidad de Probabilidad del ruido de los píxeles es estimada usando los estimadores MLNE descritos en la Sección 3.4.1.2. Primero, $\boldsymbol{U}_1$ y $\boldsymbol{U}_2$ (ecuación 3.3) son determinados para poder calcular los parámetros $\hat{\boldsymbol{u}}$ y $\hat{\boldsymbol{v}}$ (ecuaciones 3.6 y 3.7). Estas estimaciones son los parámetros de forma y escala de la función de densidad de probabilidad PDF del ruido. Con estos dos parámetros estimados, la función de distribución gamma queda determinada.

Obviamente, para la primera iteración, $\boldsymbol{U}_1$ y $\boldsymbol{U}_2$ son computados a partir de todos los píxeles del espectrograma al no haberse segmentado aún píxel alguno. Hay que remarcar que para las siguientes iteraciones estos parámetros son sólo computados a partir de los restantes píxeles, es decir, de los píxeles que no han sido segmentados en las anteriores iteraciones. La Figura 4.8 muestra dos ejemplos de funciones de distribución de ruido estimadas, para la segunda y cuarta iteración. Nótese que el rango del eje horizontal decrece en cada iteración. Este dato es fundamental para el resto del algoritmo, quedando aclarado en el siguiente apartado.



(a) Iteración 2: *u*=1.391; *v*=0.023



(a) Iteración 4: *u*=3.323; *v*=0.085

Figura 4.8 : Funciones de distribución de ruido estimadas en la segunda y cuarta iteración respectivamente. El rango del eje x decrece a medida que las iteraciones aumentan.

4.4.3.2.2 Definición del Área de Trabajo y de la Región de Confianza de ruido

La distribución estimada del ruido obtenida en el apartado anterior sirve fundamentalmente para la obtención del llamado límite de propagación. Asumiendo una probabilidad de error de detección P_e, derivamos el límite de la distribución $l^{[i]}$ [4] a partir de la distribución $\left(\hat{\sigma}^{2[i]} \Big/ \hat{D}^{2[i]} \right) \chi^2_{\hat{D}^{2[i]}}$ de tal forma que:

$$P_e = Prob\left\{ F_1 \geq l^{[i]} \right\} \tag{4.23}$$

Un valor típico usado en [1] para el algoritmo es $P_e = 0.01$. La forma de determinar el límite dada la distribución es realizada a través de un algoritmo de búsqueda dicotómico, el cual demuestra ser eficiente desde el punto de vista del coste computacional.

A continuación identificamos dos áreas de trabajo en el *Espacio Característico*:

1. **Región de Confianza de ruido**: formada por el espacio $[0, l^{[i]}]x[0, l^{[i]}]$
2. **Área de Trabajo**: formada por todo el *Espacio Característico*, menos la Región de Confianza. Matemáticamente, la Región perteneciente a: $F_1 > l^{[i]}$ y $F_2 > l^{[i]}$

Estas dos áreas dividen el *Espacio Característico* en dos regiones complementarias. Los píxeles cuya representación (a través de la célula que gobiernan) en el *Espacio*

[4] Donde el superíndice $[i]$ se refiere a la iteración *i*

Característico caiga en el Área de Trabajo, serán los candidatos para la selección y propagación de las semillas[5]; aquellos que en cambio caigan en el Área de Confianza de ruido es debido a que se estima que corresponden a regiones formadas únicamente por ruido en el espectrograma.

La Figura 4.9 muestra en primer lugar la manera de determinar el límite a partir de la distribución estimada de ruido (a). Una $P_e = 0.01$ es equivalente a dejar a la izquierda del límite el 99% del área de la PDF. En (b) se muestra cómo el límite determina las dos regiones en el *Espacio Característico*.

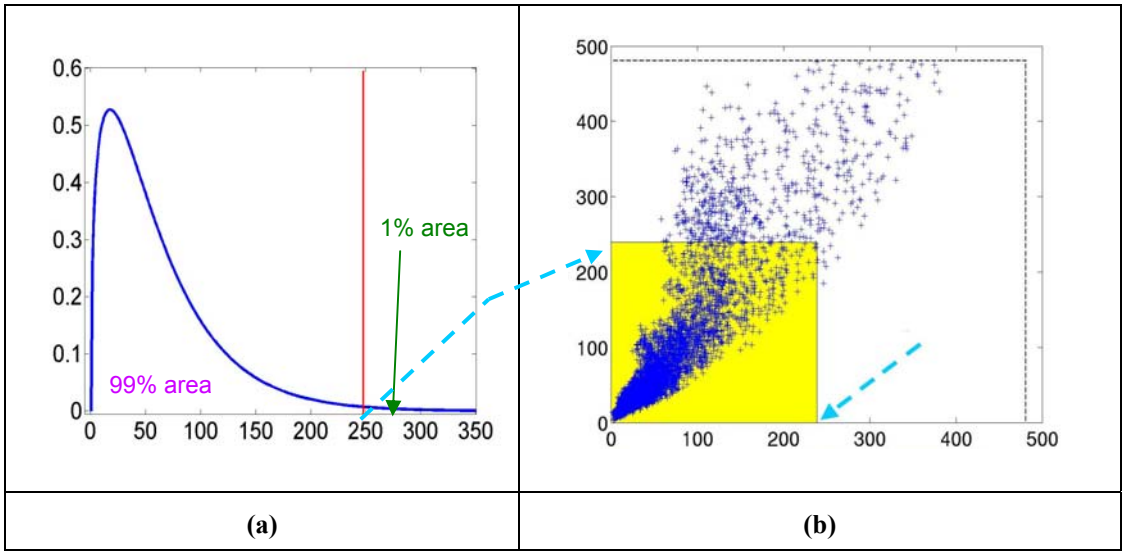
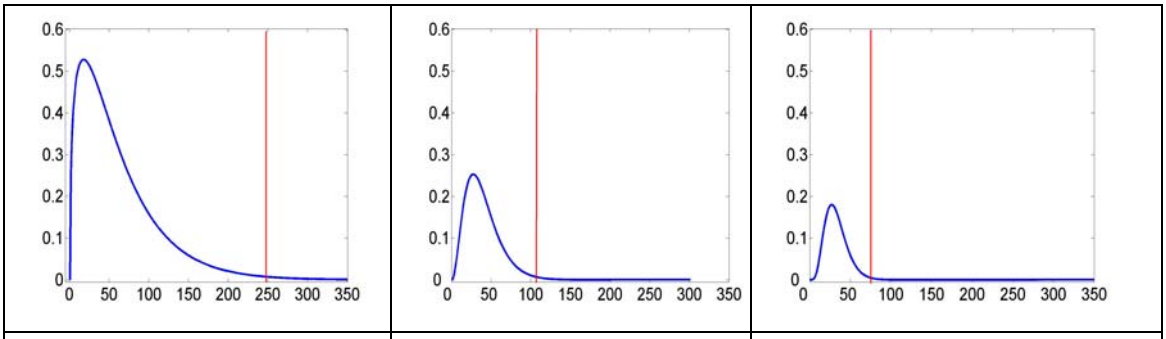


Figura 4.9 : En (a) se muestra la obtención del límite de propagación. Para Pe=0.01, el área a dejar a la izquierda del límite es el 99%. Este límite es llevado al *Espacio Característico* (b), donde se determinan la Región de Confianza (zona amarilla), y el Área de Trabajo (resto del *Espacio Característico*).

Evidentemente, estas áreas son sólo validas en la iteración en que han sido calculados los parámetros de estimación de ruido. Por tanto, en cada iteración, al ir obteniendo nuevas estimaciones de ruido a partir de los píxeles aún no extraídos, se irá determinando un nuevo límite, dando lugar a una nueva Región de Confianza de ruido y Área de Trabajo. Como en cada iteración la estimación de ruido es más precisa, el límite obtenido en cada iteración será también más preciso, y de valor menor al obtenido en la iteración anterior. Esto se debe a que los valores que toman los píxeles tomados para el cálculo de la distribución son de menor valor energético, al haber extraído ya los de mayor valor en las anteriores iteraciones. La Figura 4.10 muestra un ejemplo de la distribución de ruido para distintas iteraciones, llevando por tanto a un límite cada vez menor, y consecuentemente a una Región de Confianza cada vez más pequeña.

[5] píxeles con opciones más altas de contener señal determinista, los cuales son seleccionados para contaminar a sus vecinos en el espectrograma. Ver Sección 4.4.3.3 para más detalles.



(a) Iteración 2	(b) Iteración 4	(c) Iteración 6
l **=246.98**	***l*** **=108.31**	***l*****=77.39**

Figura 4.10 : Ilustración del límite decreciente al incrementar el número de iteraciones.

Como resultado, la Región de Confianza será cada vez más pequeña, y el Área de Trabajo será cada vez más grande a medida que las iteraciones avancen.

4.4.3.2.3 Computación del Grid Teórico

A continuación introducimos cómo calcular el llamado *Grid Teórico* $\left(E\{F_1\}, E\{F_2\}\right)$, para que al superponerlo en el *Espacio Característico* pueda guiar el proceso de segmentación.

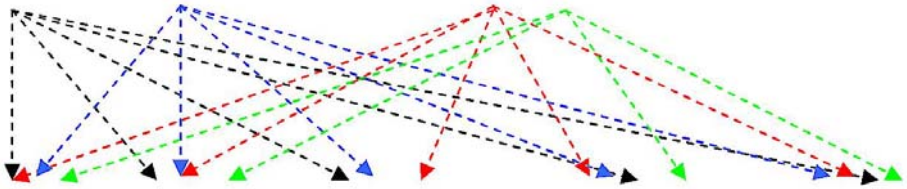
Recordamos primero los parámetros ***p*** y ***r***, descritos en la Sección 4.2.2:

- $p = P/N$: proporción de los coeficientes del espectrograma que poseen señal determinista en una célula.
- $r = S/\sigma^2$: SNR de la célula.

Donde P es el número de píxeles que contienen señal en una célula, N el número de píxeles de la célula, S el valor total de la señal determinista en una célula y σ^2 la potencia de ruido.

Posteriormente, se computa el *Grid Teórico* a partir de las ecuaciones 4.24 y 4.25. usando como σ^2 la estimación de la potencia de ruido $\hat{\sigma}^{2[i]}$. Los parámetros p y r son regularmente espaciados entre, respectivamente, [0,1] y [0,max$\{S_x[n,k]/\hat{\sigma}^{2[i]}\}$]. El valor de espaciamiento es elegido por nosotros, y dará la resolución de Grid. Un valor intermedio igual a 15, tal y como es usado en [1], será el aceptado.

Si el espaciamiento = 15, entonces 15*15 => Grid de 225 puntos:
p=[0,1/15, 2/15, , , , , , , , , ,14/15, 1]



r = [0, maxvalue*1/15, , , , , , , , , , , maxvalue*14/15, maxvalue]

Figura 4.11 : Espaciado regular de *p* y *r* y su posterior entrelazado para el uso en las ecuaciones 4.24 y 4.25.

Al estudiar las definiciones de *p* y *r*, se observa que los límites de los rangos seleccionados corresponden a los valores mínimos y máximos posibles de cada parámetro. La variable *maxvalue* corresponde al máximo valor posible del píxel en el espectrograma. El *Grid Teórico* es computado espaciando regularmente *p* y *r* a lo largo de sus rangos y entrelazando posteriormente sus valores, tal y como muestra la Figura 4.11. Todas estas posibilidades se introducen en las ecuaciones 4.24 y 4.25 para calcular el *Grid Teórico*.

Estas ecuaciones describen los primeros momentos de las C*aracterísticas* F_1 y F_2. Una mayor descripción de su obtención es mostrada en el Apéndice B. La elección de valores de *p* y *r* a lo largo de todo su rango, más la posible combinación entre ellas, da como resultado un Grid cuyos puntos representan las zonas con mayores probabilidades de tener señal determinista **para cada p y r determinado.** Por tanto, el *Grid Teórico* nos muestra los puntos en el *Espacio Característico* que debemos observar en la búsqueda de los primeros elementos que contienen señal. A estos elementos se les denominarán semillas.

$$E\{F_1\} = \hat{\sigma}^2 (1 + pr) \tag{4.24}$$

$$E\{F_2\} = \hat{\sigma}^2 \left[\frac{\sum_{i,j=0}^{4} a_{ij} p^i r^j}{\sum_{i,j=0}^{2} b_{ij} p^i r^j} \right]^{\frac{1}{2}} \tag{4.25}$$

Los coeficientes *a* y *b* dependen únicamente del tamaño de la célula *N*. Sus definiciones quedan descritas de nuevo en el Apéndice B. La Figura 4.12 muestra el *Grid Teórico* obtenido en la primera iteración de la señal representada en la Figura 4.7.

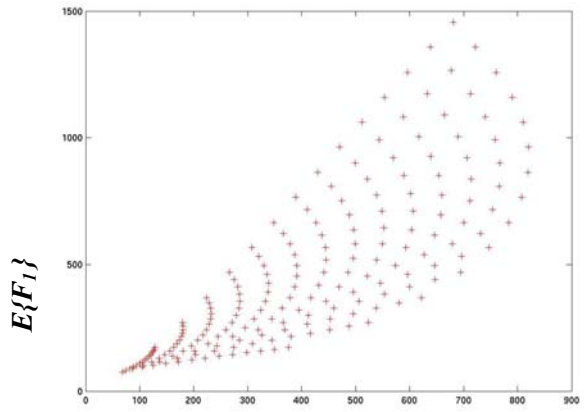
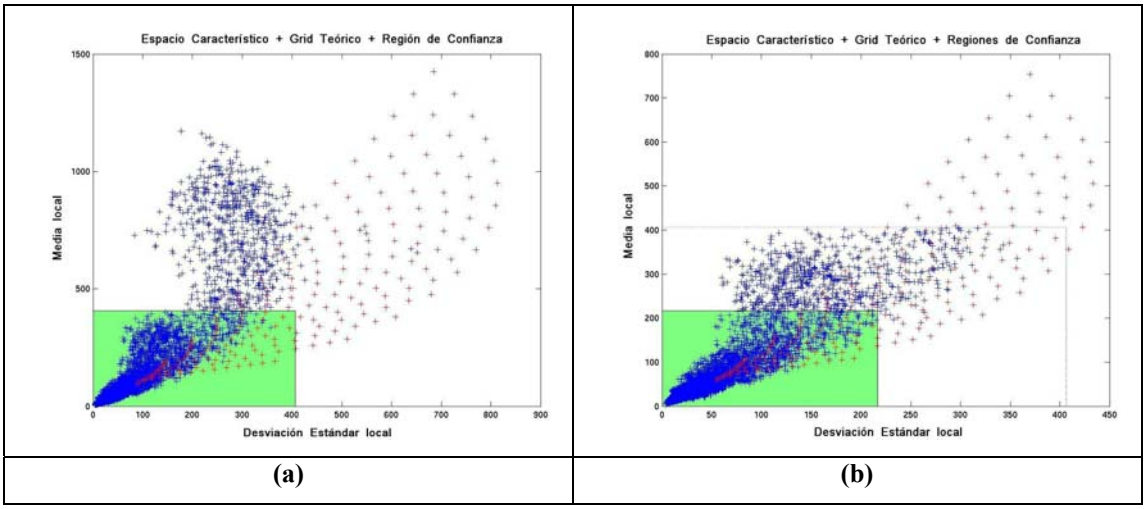


Figura 4.12 : Ilustración de las características $E\{F_2\}$ Teórico. El ejemplo aquí mostrado es el computado a partir de la señal de la figura 4.7 para la primera iteración.

Cada punto del *Grid Teórico* tiene por tanto su significado desde el punto de vista de estimaciones de propiedades de la célula que determinaría ese punto. De esta manera, el punto con mayores p y r representa el punto estimado donde hay mayor proporción de los coeficientes del espectrograma que poseen señal determinista en una célula (p = 1), y a la vez, la mayor SNR posible de la célula ($r = \max\{ S_x[n,k] / \hat{\sigma}^{2[i]} \}$). Evidentemente este punto determinará el comienzo de la búsqueda de semillas. El proceso de segmentación es explicado en el siguiente apartado.

El *Grid Teórico* es entonces superpuesto al *Espacio Característico* tal y como se muestra en la Figura 4.13 (a). No hay que olvidar que en cada iteración los parámetros y estimaciones son recalculados de nuevo a partir del espectrograma resultante de las anteriores iteraciones, el cual ha sufrido la extracción de parte de la señal determinista. Esto quiere decir que *maxvalue* decrece en cada nueva iteración, resultando un *Grid Teórico* más pequeño, tal y como se muestra en la Figura 4.13 (b), donde se contempla el *Grid Teórico* de la segunda iteración. Veremos más adelante que en cada iteración el *Grid Teórico* se hace más denso, y se inclina unos cuantos grados más en el sentido contrario a las agujas del reloj. Este comportamiento permite ir ajustándose cada vez mejor a los sucesivos *Espacios Característicos.*



(a) **(b)**

Figura 4.13 : Espacios Característicos, Grids Teóricos, Áreas de Trabajos y Regiones de Confianza, computados para (a) la primera iteración y (b) segunda iteración. Nótese que el Grid Teórico en *(b)* es de mayor resolución que el de *(a)*.

Resumiendo, el objetivo del *Grid Teórico* es el de identificar los elementos del *Espacio Característico* (por tanto, píxeles del espectrograma) de mayor *p* y *r*. Su mayor contribución es el guiado en la selección de semillas, discutido en el apartado 4.4.3.3.1. Es importante destacar que la composición entre el *Espacio Característico* y el *Grid Teórico* habilita cuatro criterios independientes a tener en cuenta: *p*, *r*, *Característica 1* y *Característica 2*. La combinación de todos ellos asegura que las mejores semillas son seleccionadas.

4.4.3.3 Módulo 2: Extracción y Propagación de semillas

En este módulo se realiza la inspección para la selección de las semillas y su posterior extracción. Una vez extraídas, las semillas se propagarán mediante la técnica de región creciente.

4.4.3.3.1 Selección de las semillas

Una vez conocidas las propiedades de los puntos del *Grid Teórico*, es lógico deducir la secuencia de selección de semillas. Primeramente, es necesaria la asociación de una zona de proximidad a cada punto del *Grid Teórico*. No hay que olvidar que este Grid está calculado a partir de estimaciones, y que en la realidad, los puntos del *Espacio Característico* se encontrarán alrededor de los del *Grid Teórico*.

Por tanto, a cada punto del *Grid Teórico* se le asocia un círculo con centro en ese punto y radio definido por $SD\{F_1(p,r)\}/2$. La deducción del segundo momento de la *Característica 1* es realizada en el Apéndice B. La determinación del círculo queda de esta manera determinada por el parámetro de variación de la *Característica 1*, suceso bastante lógico. La expresión final del radio del círculo es:

$$radio = \frac{\sigma^4}{2N}\left(1 + 2pr + p\left(r - p\right)r^2\right) \quad \textbf{(4.26)}$$

El mecanismo consiste básicamente en la búsqueda de elementos del *Espacio Característico* dentro del círculo asociado a cada punto del *Grid Teórico*. Primero se escoge el punto del Grid con mayor *p* y *r*. Si hubiera elementos dentro de su círculo asociado éstos pasarían a ser semillas, y se procedería a su propagación (proceso explicado en la Sección 4.4.3.3.2). Si no hubiera elemento alguno dentro del círculo, se pasaría a buscarlo en el círculo asociado al punto del *Grid Teórico* con mismo valor máximo *p*, pero inmediatamente menor *r*. Este proceso se repite hasta que se encuentre alguna semilla. Hay que añadir una condición extra: el punto del *Grid Teórico* que está siendo analizado a través de su círculo debe pertenecer al Área de Trabajo. Esta condición es básica, puesto que de lo contrario se pasaría a extraer semillas de la Región de Confianza, zona en la que los elementos son considerados ruidosos, y se acabarían extrayendo erróneamente zonas ruidosas en lugar de señal determinista.

Cuando ya se han analizados todos los puntos del Grid correspondientes al valor máximo *p* y que pertenezcan al Área de Trabajo, se procede a disminuir este parámetro, y a estudiar los nuevos puntos del Grid para todos los valores de *r*, comenzando de nuevo por su valor máximo posible. Como siempre, hay que añadir la condición de que los puntos del Grid deben pertenecer al Área de Trabajo. Este proceso es iterativo, disminuyendo el valor *p*, y volviendo a inspeccionar los nuevos puntos para los distintos valores de *r*.

Este proceso es ejecutado hasta que se encuentre alguna semilla, o se finalice la inspección de todas las zonas de los puntos del Grid en el Área de Trabajo sin encontrar semilla alguna.

La Figura 4.14 muestra el proceso de selección de las primeras semillas en la primera iteración, las cuales son llevados de nuevo al espectrograma, donde comienza su propagación.

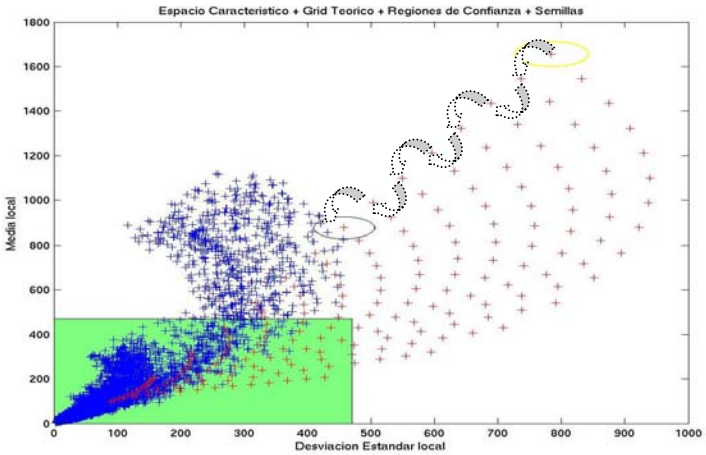


Figura 4.14 : Ilustración del proceso de selección de las semillas comenzando por el punto más alto del Grid. Nótese que el radio de los círculos es distinto al depender éstos de los valores de *p* y *r* en el Grid.

La Figura 4.15 muestra una zona aumentada de la Figura 4.14. Es el ejemplo de extracción de las primeras semillas en la primera iteración, las cuales se mapean para encontrar su posición en el espectrograma, donde comienza su propagación. Podemos observar que las primeras semillas corresponden al chirp con mayor SNR.

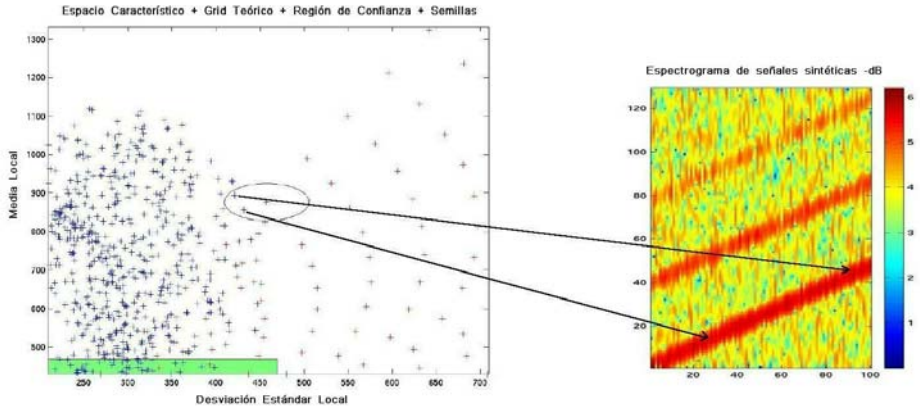


Figura 4.15 : Primer grupo de semillas seleccionado y su mapeo en el espectrograma.

4.4.3.3.2 Propagación de las semillas

Una vez extraídas las semillas encontradas mediante el proceso ya definido, se procede a su propagación mediante la técnica de región creciente. Dicha expansión es realizada bajo condiciones simultáneas del espectrograma y del *Espacio Característico*.

Cada semilla encontrada en el *Espacio Característico* es mapeada en el espectrograma. El píxel correspondiente a una semilla es extraído del espectrograma, formando ya parte de la señal determinista extraída. A continuación, se examinan los 8 píxeles vecinos que rodean a la semilla, convirtiéndose en candidatos a nuevas semillas. Estos píxeles son mapeados de vuelta al *Espacio Característico*, y se procede a evaluar su posición. La condición para que estos candidatos pasen a ser semillas es que su posición en el *Espacio Característico* caiga en el Área de Trabajo. Si esto ocurre, los candidatos pasan a ser denominados semillas, procediéndose a su extracción del espectrograma, y a evaluar de nuevo sus 8 píxeles vecinos de forma análoga a la descrita. Hay que subrayar que la posición adyacente de píxeles en el espectrograma no implica necesariamente la proximidad de sus representaciones en el *Espacio Característico*, ya que éste únicamente representa propiedades estadísticas de las células que gobierna cada píxel. Lógicamente, los píxeles que ya han sido seleccionados como semillas, no vuelven a ser considerados como candidatos para la propagación de semillas. De lo contrario, obtendríamos un algoritmo con un bucle infinito.

Este proceso se repite hasta que no haya más vecinos de los píxeles extraídos que cumplan las condiciones de contaminación: posición en el *Espacio Característico* en el Área de Trabajo, y no haber sido contaminado anteriormente.

La Figura 4.16 detalla aún más este proceso mediante un diagrama de flujo. La correspondiente explicación de cada paso queda mostrada en la Tabla 4.1.

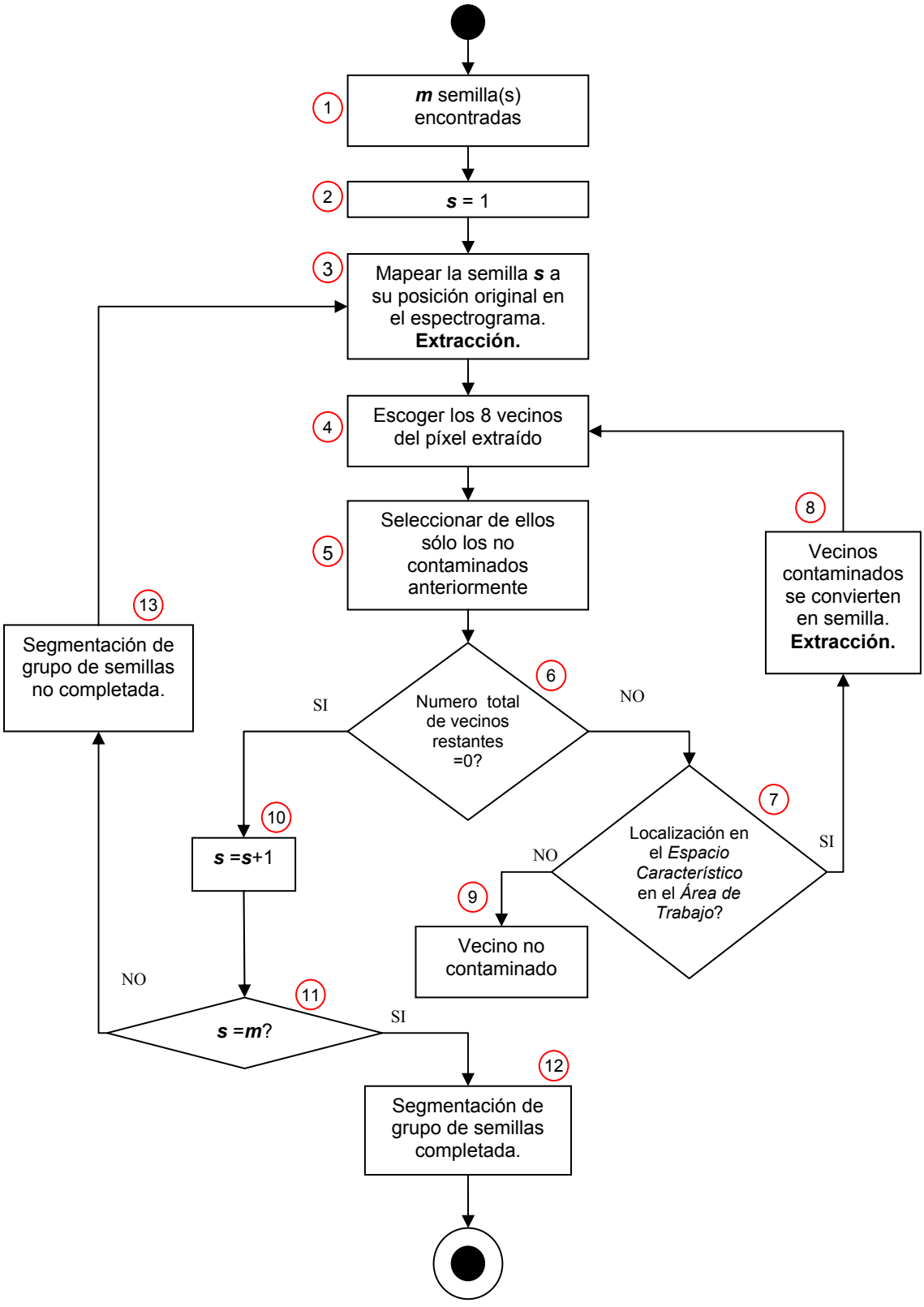


Figura 4.16 : Diagrama de flujo del proceso de propagación de las semillas.

Paso	Proceso
1	Identificar el número de semillas ***m*** encontradas
2	Introducir un contador ***s*** e inicializarlo a 1. Es usado en los Pasos 10 y 11.
3	Mapear las semillas a su localización original en el espectrograma. Extraer esa semilla.
4	Seleccionar los 8 vecinos que rodean a la semilla en el espectrograma.
5	De los 8 píxeles, seleccionar sólo aquellos que no han sido contaminados anteriormente.
6	Si el número de vecinos restantes es cero, significa que esa semilla no puede propagarse en ninguna dirección. Por tanto hay que considerar la semilla como muerta y el algoritmo va al Paso 10 para comprobar si quedan semillas por propagarse. Por otro lado, si quedan vecinos no contaminados anteriormente, el algoritmo continúa en el Paso 7 para comprobar las propiedades de esos vecinos.
7	Para los restantes vecinos, comprobar si su correspondiente localización en el *Espacio Característico* cae en al Área de Trabajo.
8	Si algún vecino tiene una localización en el *Espacio Característico* en la actual Área de Trabajo, hay que contaminarlo, i.e. convertirlo en semilla. Esta nueva semilla tiene que ir al Paso 4.
9	Si un vecino tiene su correspondiente localización en el *Espacio Característico* en la Región de Confianza de ruido, el vecino no es contaminado.
10	Cuando una semilla muere, incrementar el contador ***s*** en 1 para registrar cuántas semillas se han propagado todo lo que podían.
11	Comprobar si ***s*** es iguala ***m***.
12	Si lo es, significa que todas las semillas han muerto. Por tanto, la propagación del grupo de ***m*** semillas queda completada.
13	De lo contrario, significa que quedan semillas por propagarse. Ir al Paso 3 a continuar con la siguiente semilla.

Tabla 4.1 : Pasos del proceso de propagación de las semillas.

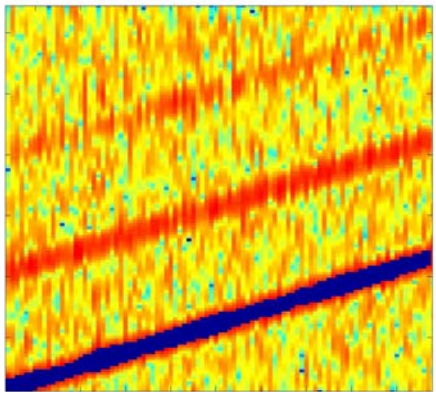


Figura 4.17 : Resultado de propagación de las semillas seleccionadas en la Figura 4.15.

La Figura 4.17 muestra el resultado de la propagación del primer grupo de semillas seleccionado. La región azul (energía cero) es el área contaminada por este primer grupo. Estas regiones tienen energía cero porque son las áreas extraídas, las cuales contienen señal determinista. Las siguientes iteraciones se realizarán calculando los parámetros a partir de este nuevo espectrograma. Los dos chirps restantes aún no han sido detectados porque sus píxeles no son vecinos de las actuales semillas, pertenecientes por tanto sólo al primer chirp.

4.4.3.4 Tests

A continuación se presentan los dos exámenes que se realizan tras las operaciones de los dos primeros módulos. Tal y como se muestra en el Diagrama de flujo principal en la Figura 4.5, cada uno de estos tests dará lugar a un bucle en el sistema. El primer test, el cual provocará un bucle en torno al Módulo 2, irá destinado a completar toda la extracción de la señal determinista dentro de una iteración. El segundo, el cual provocará un bucle al comienzo del algoritmo, es el bucle principal del sistema. Cada iteración de este bucle representa una extracción más de señal, un nuevo acercamiento al límite que es capaz de conseguir este algoritmo.

4.4.3.4.1 Test de comprobación de candidatos

Cuando la propagación del primer grupo de semillas ha terminado, la selección de semillas propuesta en la Sección 4.4.3.3.1 continúa, pero en esta ocasión sin los elementos del *Espacio Característico* que ya han sido extraídos anteriormente. Es necesaria la repetición de todos los cálculos realizados en el Módulo 2. De esta manera, se pasa al siguiente punto del *Grid Teórico*, para evaluar si en su círculo asociado se encuentra alguna semilla. Si hay alguna, de nuevo se procede a su extracción y propagación. Encontrar nuevas semillas normalmente significa encontrar un nuevo patrón espectral. Si no la hay, es el momento de pasar a buscar en el siguiente punto, y así sucesivamente hasta que hayan sido examinados todos los puntos del *Grid Teórico* que, como ya se ha destacado antes, no se encuentren en la Región de Confianza. Por tanto, el bucle ocasionado por este test tiene el fin de garantizar que se establezca una búsqueda de semillas en todos los puntos del *Grid Teórico* que pertenezcan al Área de Trabajo. Este proceso es ilustrado mediante el diagrama de flujo de la Figura 4.18:

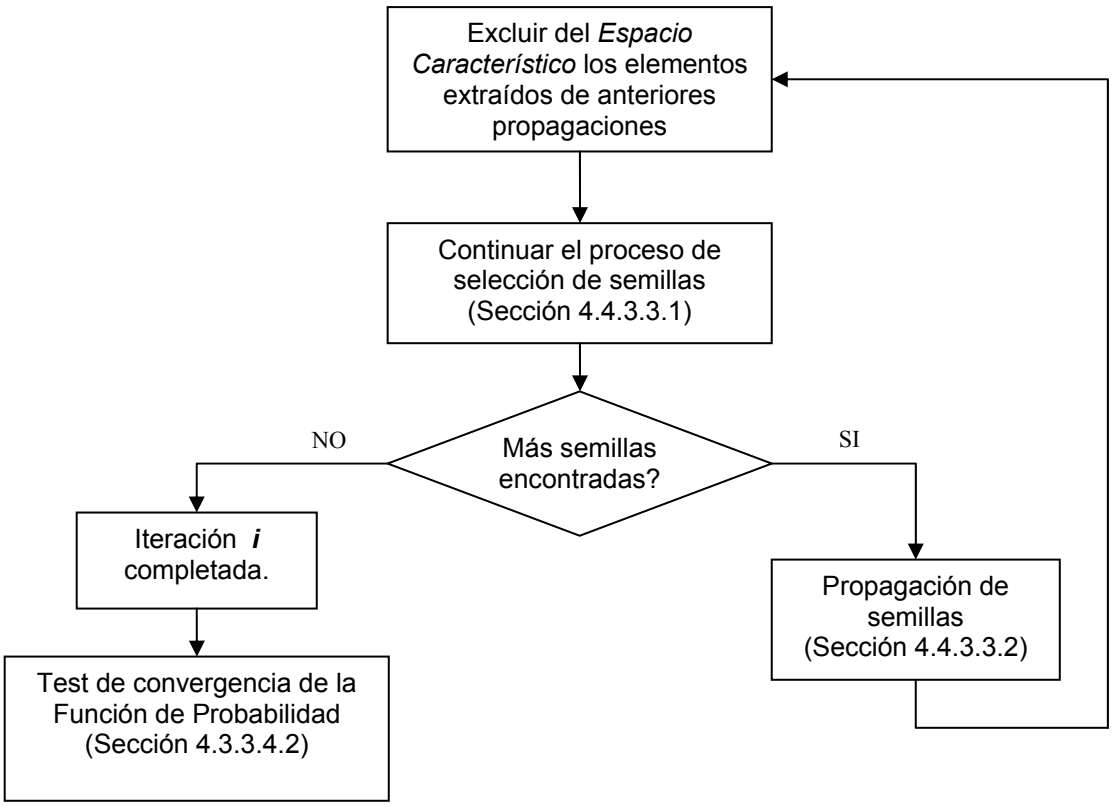


Figura 4.18 : Diagrama de flujo de los tests.

4.4.3.4.2 Test de convergencia de la Función de Probabilidad

Si no se encuentran más semillas, significa que todos los candidatos de la actual Área de Trabajo han sido examinados, y la presente iteración queda completa. A continuación se realiza el llamado Test de Convergencia.

Este test es llevado a cabo al final de cada iteración para decidir si hay necesidad de una nueva iteración o no. Esta es la parte que conduce al automatismo del sistema. Sólo se producirá una nueva iteración si el sistema considera que aún quedan píxeles pertenecientes a regiones de señal determinista que no han sido extraídos. La decisión está basada en la función de convergencia de probabilidad mencionada en la Sección 3.5.1.1.

Así, todos los parámetros determinados desde el principio del algoritmo son recalculados a partir del espectrograma sin considerar los píxeles ya extraídos. Como ya se ha comentado anteriormente, el hecho de haber extraído señal determinista deja una mayor proporción de píxeles ruidosos en el espectrograma, y consecuentemente esto dará lugar a unas nuevas estimaciones de ruido más precisas. Por tanto, para cada nueva iteración se calcula una nueva estimación de la potencia de ruido, una nueva PDF, un nuevo *Espacio Característico*, unas nuevas Área de Trabajo y Región de Confianza y un nuevo *Grid Teórico*. Tal y como ya se demostró en la Figura 4.8, el límite computado decrece en las sucesivas

iteraciones gracias a la extracción de las regiones de señal determinista. Es por tanto razonable afirmar que si después de una particular iteración casi toda la región de señal determinista ha sido extraída, el límite computado será casi constante. La diferencia entre un límite y el consecutivo se hace cada vez más pequeña, confirmando la Convergencia predicha en la Teoría. Vemos un ejemplo en la Figura 4.19:

Iteración	Límite computado
1	485.65
2	217.60
3	132.72
4	104.70
5	92.06
6	80.68
7	79.81

Tabla 4.2 : Límite computado para cada iteración.

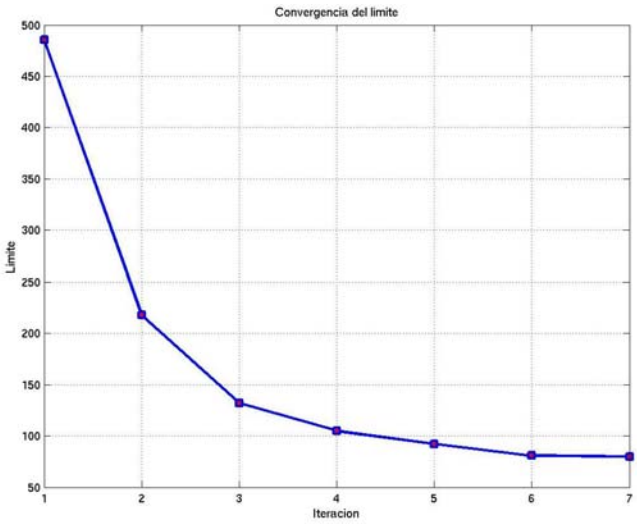


Figura 4.19 : Demostración de la convergencia del límite.

A continuación se define una estimación de error en tanto por ciento P_l , para que la segmentación pare automáticamente cuando la diferencia relativa entre el nuevo límite y el anterior se haga menor que el error P_l . Matemáticamente, al final de cada iteración se calcula:

$$\frac{\left|\text{último límite - límite previo}\right|}{\text{último límite}} \cdot 100\%$$

y se compara con P_l . Si el resultado es mayor que este error, significa que el límite ha variado sustancialmente porque aún queda señal determinista por extraer. Por tanto, hay que continuar con el algoritmo. Si por el contrario el resultado es menor, significa que el nuevo límite obtenido es casi igual que el obtenido en la anterior iteración, y que ya no queda prácticamente señal determinista por extraer. Es la condición de parada de nuestro algoritmo de segmentación.

El porcentaje de error es introducido de forma experimental en torno a un 2%. De esta forma, si nos fijamos en el ejemplo de la Tabla 4.2, el algoritmo pararía después de la séptima iteración ya que:

$$\frac{\left|79.81 - 80.68\right|}{79.81} \cdot 100\% = 1.09\% < 2\%$$

El test de Convergencia es mostrado en el Diagrama de la Figura 4.20:

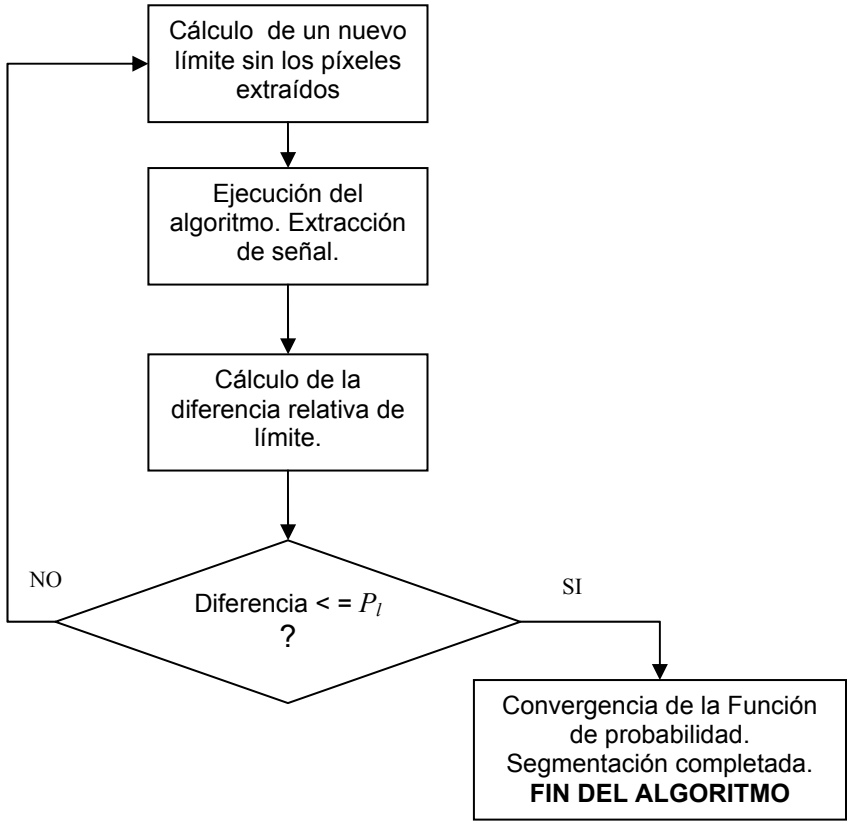


Figura 4.20 : Diagrama de flujo del Test de Convergencia de la Probabilidad.

4.4.4 Ilustración del proceso Paso a Paso

Los siguientes apartados ilustran el proceso de segmentación paso a paso. En primer lugar, se realizará una demostración de segmentación de señales sintéticas chirp, de forma muy análoga a la efectuada en el principal artículo de referencia [1]. Posteriormente, entramos en el campo de la investigación al aplicar nuestro algoritmo de segmentación a señales de voz corrompidas por ruido Gausiano blanco.

4.4.4.1 Ejemplo de señales sintéticas: Chirps

A continuación mostramos las ilustraciones pertenecientes a un proceso completo de segmentación de señales sintéticas. La señal de entrada está formada por 3 chirps inmersos en ruido Gausiano blanco. La Figura 4.21 muestra en la parte superior (a) el espectrograma y el *Espacio Característico* asociado de 3 chirps limpios, sin contaminación alguna de ruido. En la parte inferior (b) se muestra las representaciones análogas de los mismos 3 chirps pero inmersos en ruido WGN cuya potencia es la necesaria para que la SNR de los chirps sea de 5

dB (chirp inferior), 0 dB (intermedio) y -5 dB (chirp superior). Los dibujos de la Figura 4.21(b) corresponden a la iteración número cero, al ser los primeros cálculos de la señal, y no haber extraído aún señal determinista alguna.

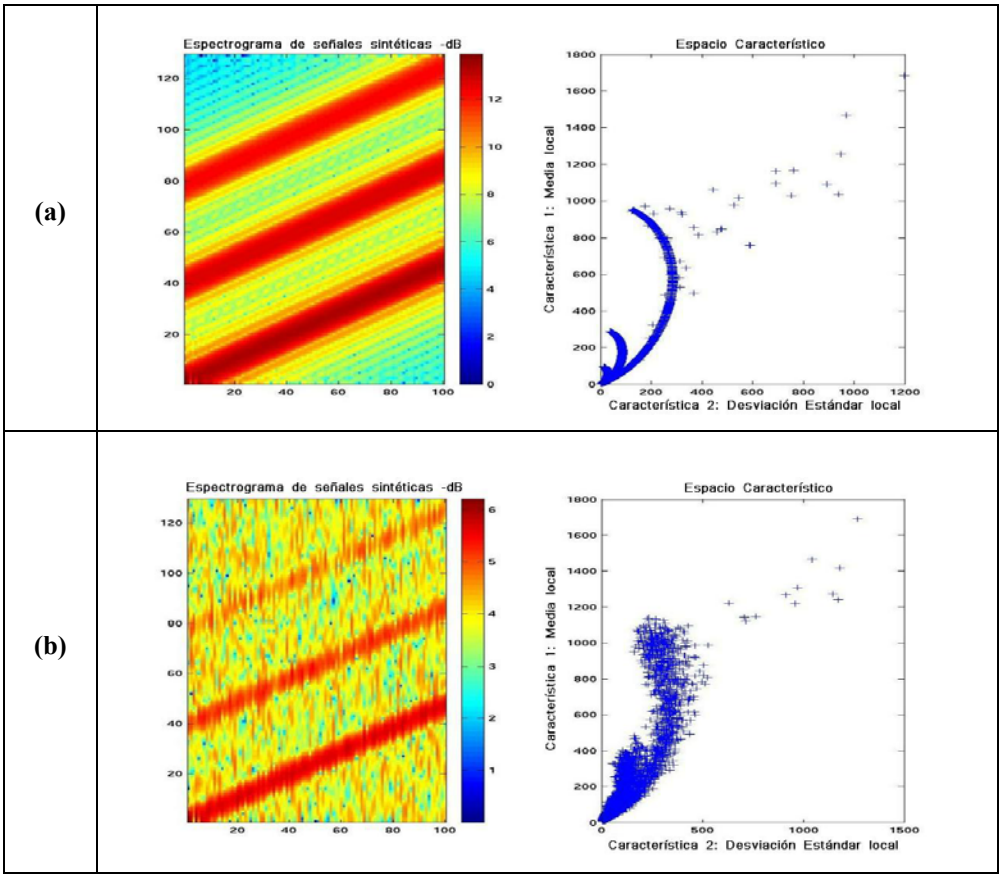


Figura 4.21 : (a) Espectrograma y *Espacio Característico* de chirps limpios. (b) Representación de los mismos chirps pero inmersos en ruido WGN.

Puede apreciarse cómo la introducción del ruido afecta al espectrograma y al *Espacio Característico*, distorsionando las líneas y bordes bien definidos en las condiciones ideales. Como curiosidad, cabe destacar la forma curva y tan precisa de cada chirp en el *Espacio Característico*. Si observamos en el espectrograma de los chirps limpios cualquier parte central de un chirp, la media de las células es alta, pero su desviación estándar (variación entre los coeficientes) es baja. Estos puntos corresponden en el *Espacio Característico* al extremo superior de la curva de cada chirp. Si comenzamos a desplazarnos de nuevo en el espectrograma hacia el borde del chirp, la media comienza a disminuir, y la desviación estándar a aumentar, ya que dentro de la misma célula hay valores bajos y altos al encontrarnos en la zona margen. La correspondencia es la zona intermedia de la curva azul. Por último, a medida que nos alejamos del chirp tanto la media como la desviación estándar tienden a cero, conduciendo al extremo inferior de la curva.

Los cambios del *Espacio Característico* y del espectrograma para cada iteración son mostrados en la Figura 4.22. Nótese que la señal determinista es extraída poco a poco del espectrograma hasta que sólo se abandonan píxeles de ruido.

Como a mayor número de iteraciones se extrae cada vez más señal determinista, pueden observarse los siguientes cambios con cada nueva iteración:

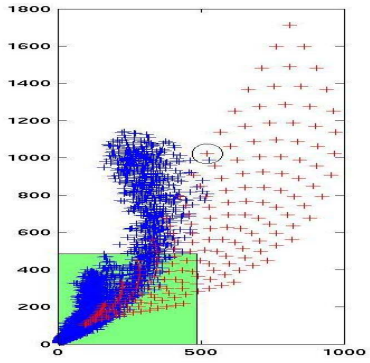
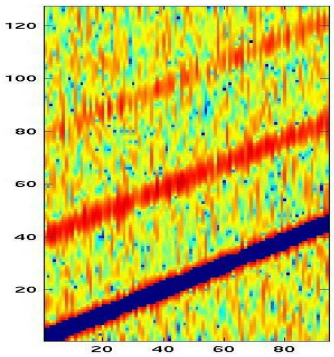
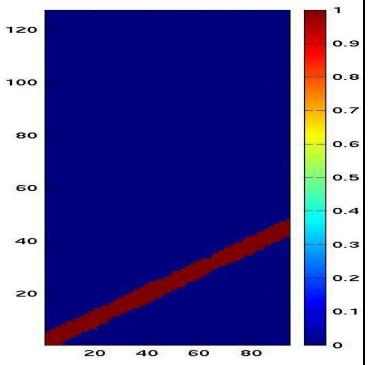
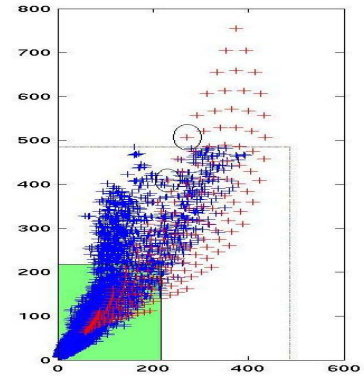
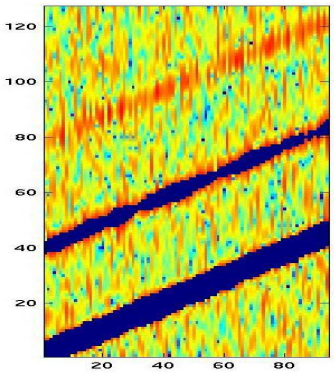
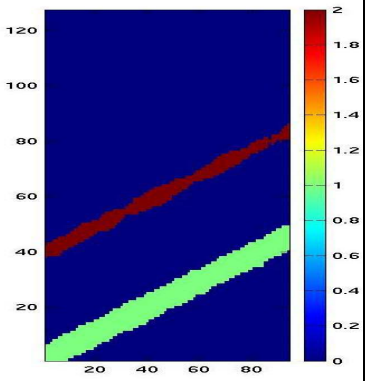
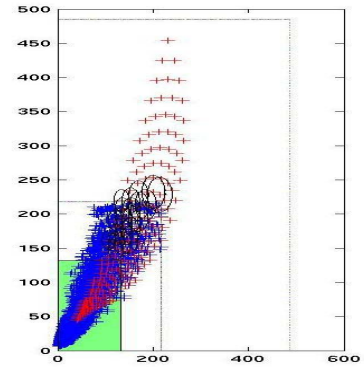
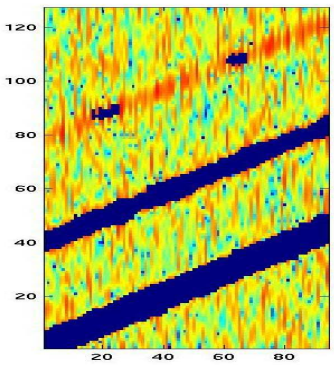
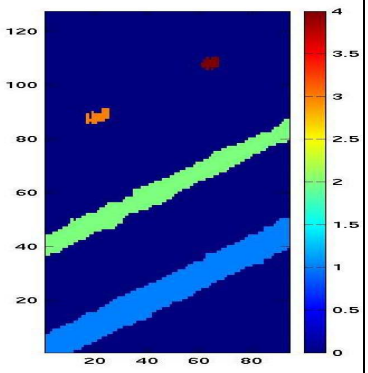
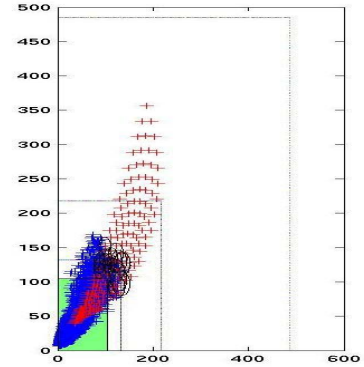
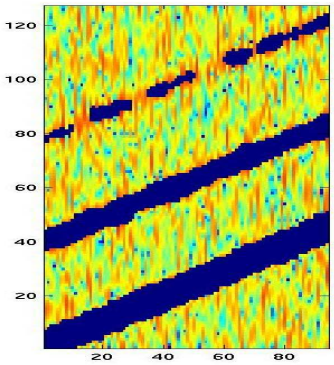
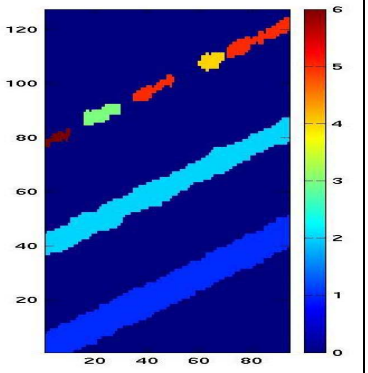
- El número de elementos del *Espacio Característico* disminuye.
- El nuevo límite computado en cada iteración siempre será inferior al calculado anteriormente debido al decrecimiento en *Característica 1*.
- Consecuentemente, hay una mayor Área de Trabajo, y por tanto una menor Región de Confianza de ruido.
- El nuevo *Grid Teórico* será de mayor resolución debido al decrecimiento de *maxvalue*, habilitando que las semillas sigan siendo encontradas eficientemente a partir del decrecimiento del tamaño del *Espacio Característico*.

Mostramos el proceso de segmentación paso a paso, para describir lo sucedido en cada iteración posteriormente. Cada fila representa una iteración.

La primera columna es la ilustración del *Espacio Característico* para esa iteración. En él, se encuentran superpuestos el *Espacio Característico* (cruces azules), el Grid Teórico (cruces rojas), la Región de Confianza (área verde) y el Área de Trabajo (región blanca), así como las semillas encontradas que producen la extracción de parte de la señal (son los elementos del *Espacio Característico* que caen dentro del círculo asociado a un punto del Grid Teórico).

La segunda columna es el espectrograma resultante de la extracción de señal realizada hasta ese momento. Así, entre una fila y la siguiente puede observarse qué cantidad de extracción ha sido realizada.

La tercera columna representa las regiones que han sido segmentadas hasta ese momento. Estas áreas van etiquetadas con un número en función del orden de detección, y de las posteriores fusiones con otras regiones.

	Espacio Característico	Espectrograma restante	Segmentación etiquetada
1			
2			
3			
4			

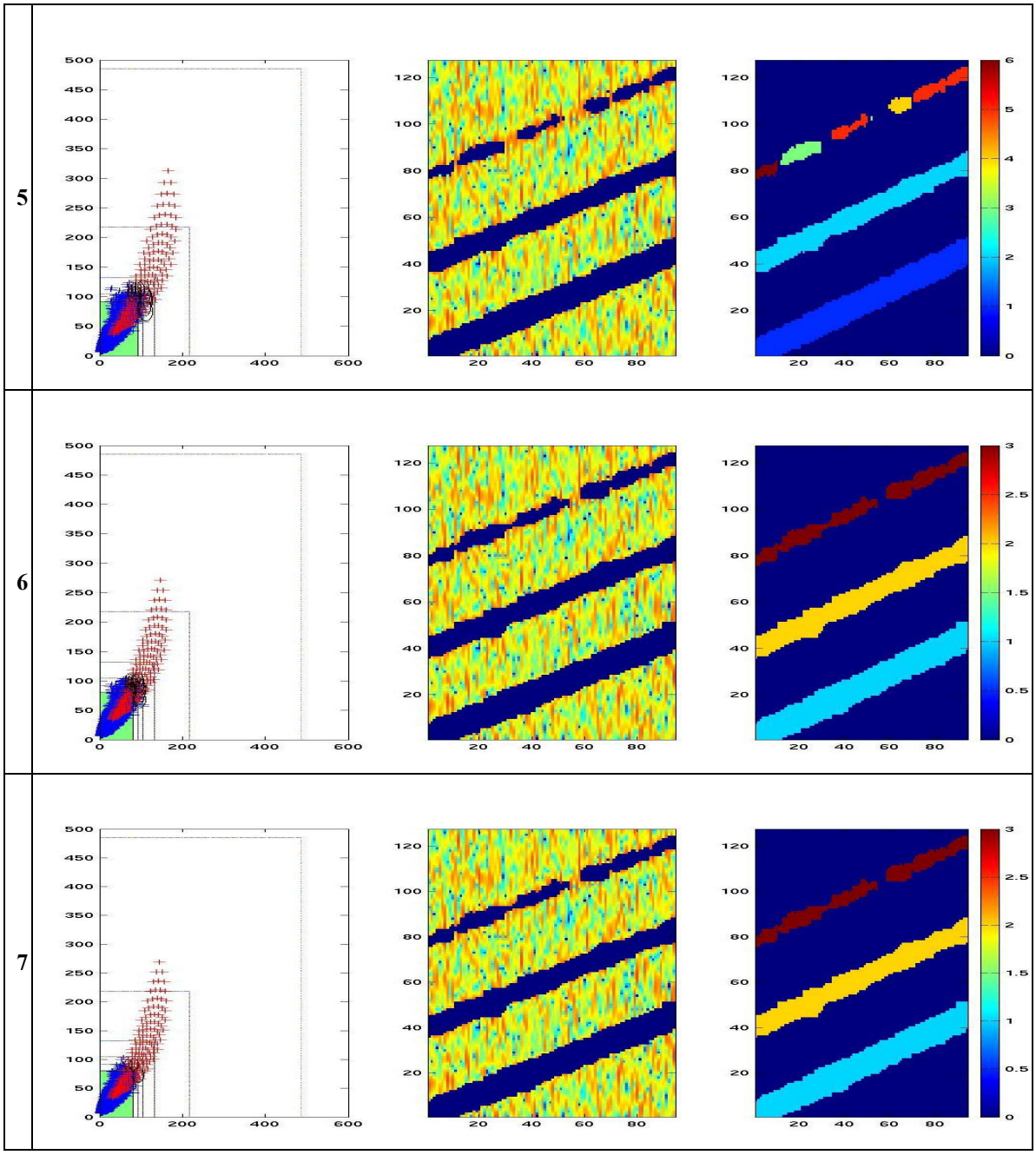


Figura 4.22 : Ilustración de los cambios iteración a iteración en el proceso de segmentación de tres chirps.

ITERACIÓN 1: Se detecta el chirp de mayor SNR mediante una semilla. Se extrae la zona de máxima energía de ese chirp, etiquetándola con *1*.

ITERACIÓN 2: Se encuentran dos semillas. Una de ellas corresponde a la fuerza a la detección del segundo chirp, etiquetándolo en adelante como *2*. La otra semilla puede corresponder también a la detección del chirp *2*, o bien a la continuación de la extracción del chirp número *1*. En cualquier caso es apreciable la expansión de la región detectada del chirp *1*, cuyo motivo puede deberse también a la propagación de las viejas semillas de la primera iteración en la nueva Área de Trabajo. De cualquier manera, el resultado de esta iteración es

la extracción de la mayor zona de energía del chirp *2*, cuya SNR es de 0 dB, y de la expansión en la extracción del chirp *1*, cuya SNR es de 5 dB.

ITERACIÓN 3: Multitud de semillas encontradas. La mayoría de ellas corresponden a pequeñas expansiones de los chirps *1* y *2*. Sólamente dos semillas han conseguido detectar pequeñas regiones del tercer chirp, cuya SNR es de -5 dB. Estas dos regiones, al estar inconexas, son denominadas en principio *3* y *4*.

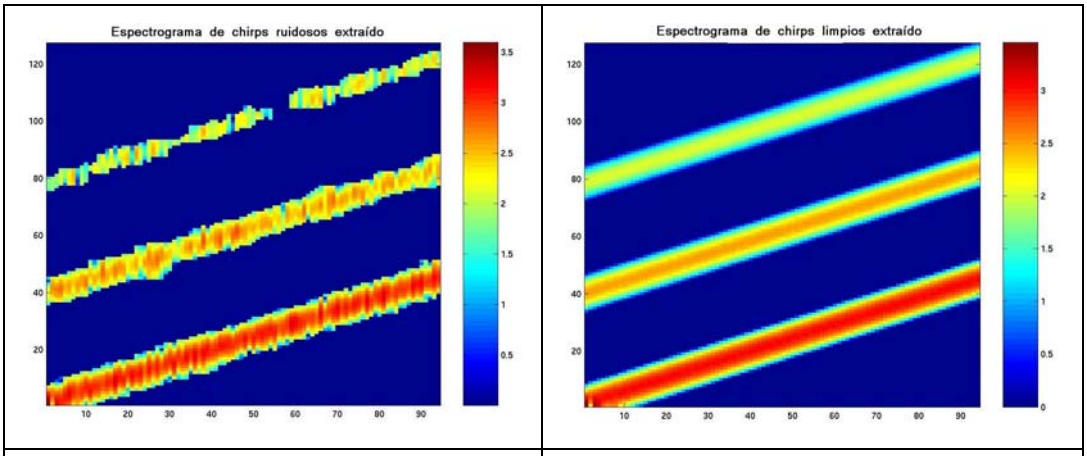
ITERACIÓN 4: De nuevo multitud de semillas encontradas. En este caso, corresponden tanto a la expansión de los dos chirps ya detectados completamente, como a nuevas pequeñas áreas pertenecientes al tercer chirp. Al igual que antes las sub-regiones del tercer chirp tienen distintas etiquetas.

ITERACIÓN 5: Más semillas encontradas para la nueva Área de Trabajo. En este caso ya no se encuentran más regiones nuevas, sino que se expanden fundamentalmente las ya encontradas del tercer chirp.

ITERACIÓN 6: Las nuevas semillas encontradas producen una expansión que hace que la mayoría de las sub-regiones detectadas del tercer chirp se fundan formando una o dos regiones. El algoritmo pasa a denominarlas con la etiqueta *3*.

ITERACIÓN 7: En esta última iteración, apenas hay variación respecto a la anterior, debido a la convergencia del límite ya explicado. El nuevo límite obtenido es casi igual ya que se sigue estimando la misma potencia de ruido. En este caso, al no variar prácticamente nada, sólo se encuentran 3 semillas más, las cuales apenas se expanden. El algoritmo considera de la forma detallada anteriormente que ya ha extraído toda la información posible, y no procesa ninguna iteración más. Es la condición de salida del algoritmo.

Claramente se ve a lo largo del proceso que antes de la búsqueda de nuevas semillas hay que realizar la propagación en el último y nuevo *Espacio Característico* computado de las semillas de la anterior iteración. Además, de no ser así, nunca podrían ser extraídos puntos del *Espacio Característico* que no cayeran dentro del Grid, y se demuestra mediante las anteriores ilustraciones la imposibilidad de este suceso gracias a la propagación. Mientras nuevas semillas normalmente implican la detección de nuevos patrones espectrales de voz, la propagación de viejas semillas puede ser interpretada como la continuación de la segmentación de las regiones previamente detectadas en anteriores iteraciones, y que no pudieron ser completadas. La razón es que algunos vecinos de semillas de previas iteraciones no eran candidatos para la propagación en ese tiempo, pero lo son posteriormente para una mayor Área de Trabajo. Esto puede ser resaltado en el espectrograma para la primera y segunda iteración en la Figura 4.22.



(a) **(b)**

Figura 4.23 : Espectrograma final de las regiones extraídas. (a) Chirps contaminados por ruido. (b) Chirps limpios (caso ideal).

La Figura 4.23 (a) muestra los píxeles del espectrograma que han sido extraídos en el caso de los chirps contaminados por ruido. Puede establecerse la comparación con la extracción de señal de los chirps limpios, en la Figura 4.23 (b). Estos píxeles son los representativos de la señal determinista formada por los tres chirps. Podemos apreciar que el chirp con mayor SNR (5 dB) es detectado perfectamente. El segundo chirp, con una SNR de 0 dB también es detectado con bastante precisión, aunque obviamente no tanta como el primer chirp. Por último, el tercer chirp, con una SNR bastante baja (-5 dB) sigue siendo detectado, aunque claramente se observa una menor precisión y área segmentada. Este último chirp es el peor extraído por razones obvias: su baja SNR. Un estudio en profundidad sobre la precisión del algoritmo en función de la SNR es realizado en el apartado 5.1.2.

En busca de una mayor comprensión del algoritmo, se introdujeron dos nuevas Figuras. En primer lugar, se estudió de nuevo el histograma o función de densidad del *Espacio Característico*. Queda mostrado en la Figura 4.24. Puede apreciarse claramente cómo esta Transformación es muy buena de cara a la segmentación, porque distribuye en el *Espacio Característico* regiones muy claras en función de las propiedades estadísticas. Puede observarse que casi todo el ruido cae en la esquina inferior izquierda, mientras que los chirps parten de esa esquina situándose por otras regiones del espectrograma. Obviamente, el *Espacio Característico* restante tras la última iteración coincide con la zona de mayor densidad, al tratarse la mayor parte de los píxeles del espectrograma de ruido.

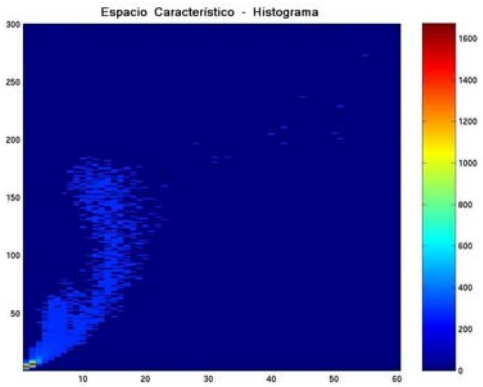
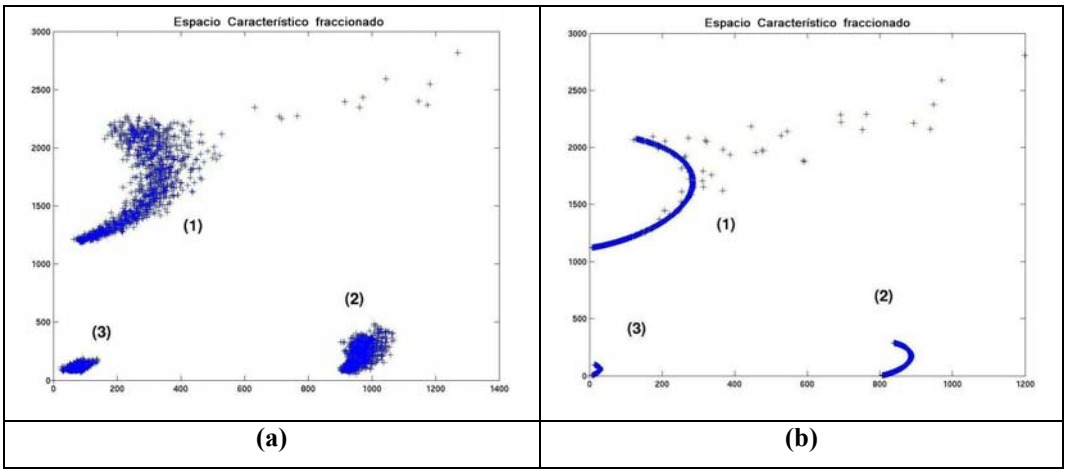


Figura 4.24 : Histograma del *Espacio Característico* inicial.

La Figura 4.25 muestra otra perspectiva del *Espacio Característico*. Una vez finalizada la extracción de la señal, queda la curiosidad de observar cómo es la representación de cada uno de los chirps por separado en el *Espacio Característico*. Tal y como ya se explicó anteriormente, cada uno tiene una forma curva. Evidentemente la observación en el *Espacio Característico* sin ruido es clara, no siéndolo así en el ruidoso. Tras segmentar y etiquetar cada uno de los chirps, se procedió a la separación de cada una de las detecciones, observando de nuevo la forma genérica curva, distorsionada por el ruido. La Figura 4.25 (a) muestra la separación del caso ruidoso, en comparación con la Figura 4.25 (b), caso ideal sin ruido.



(a) **(b)**

Figura 4.25 : *Espacio Característico* fraccionado.

4.4.4.2 Ejemplo de señales de voz

El siguiente paso consiste en mostrar el proceso de segmentación de una señal de voz inmersa en ruido WGN. Una descripción detallada de la procedencia de este tipo de señales es realizada en el apartado 5.2. Por el momento, cabe destacar que el siguiente ejemplo mostrado paso a paso corresponde a una señal de voz cuya SNR es de 5 dB. Para el tratamiento de estas señales se deshabilitó la herramienta de asignación de etiquetas, ya que en este caso lo

importante es extraer la mayor cantidad de señal de voz posible, independientemente del orden o de la numeración.

La Figura 4.26 (a) muestra el espectrograma de la señal de voz limpia, junto al *Espacio Característico* asociado. La Figura 4.26 (b) contempla la misma señal de voz inmersa en ruido WGN de tal forma que la SNR es de 5 dB. Puede apreciarse la considerable corrupción del ruido en todo el espectrograma.

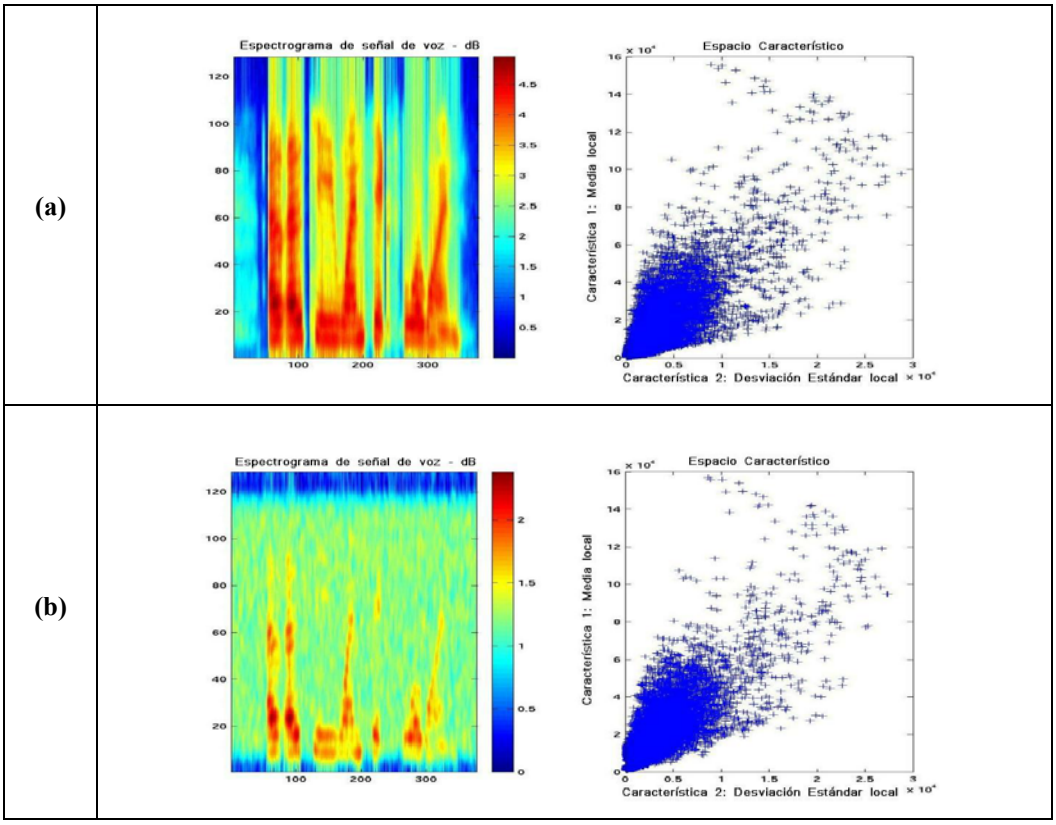
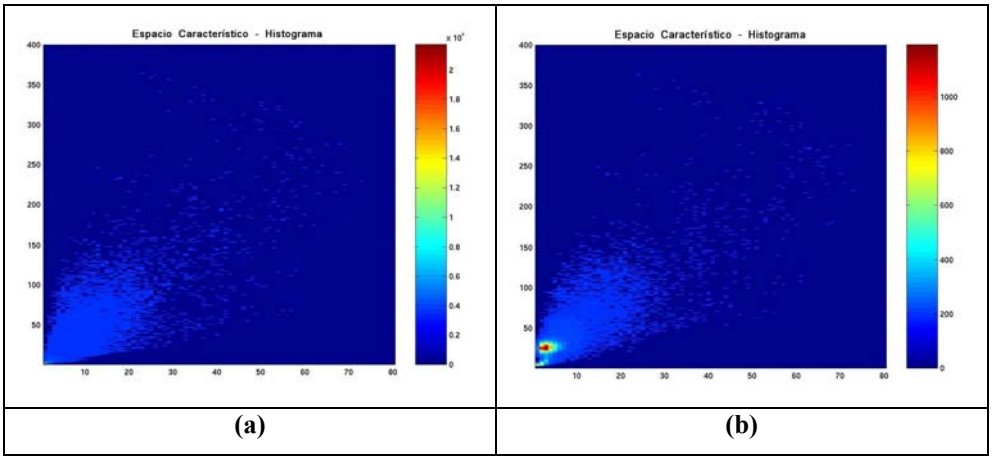


Figura 4.26 : (a) Espectrograma y *Espacio Característico* de señal de voz limpia. (b) Representación de la misma señal de voz pero inmersa en ruido WGN.

Cabe apreciar que ahora la forma de los racimos en el *Espacio Característico* es distinta en el sentido de que contienen una mayor aleatoriedad. Anteriormente podíamos dar una explicación de la forma obtenida en el caso de los chirps, pero ahora se completa prácticamente todo el *Espacio Característico* de forma casual.

Un estudio de la densidad de estos *Espacios Característicos* aclara la situación aproximada de la mayoría de los píxeles de ruido. La Figura 4.27 (a) muestra que en el caso limpio hay una distribución por igual de los píxeles a lo largo del Espacio. Por el contrario, la Figura 4.27 (b) muestra el caso ruidoso en el que la mayoría de los píxeles, los cuales corresponden al ruido, se sitúan en torno a un determinado punto (obviamente en función de la potencia del ruido).

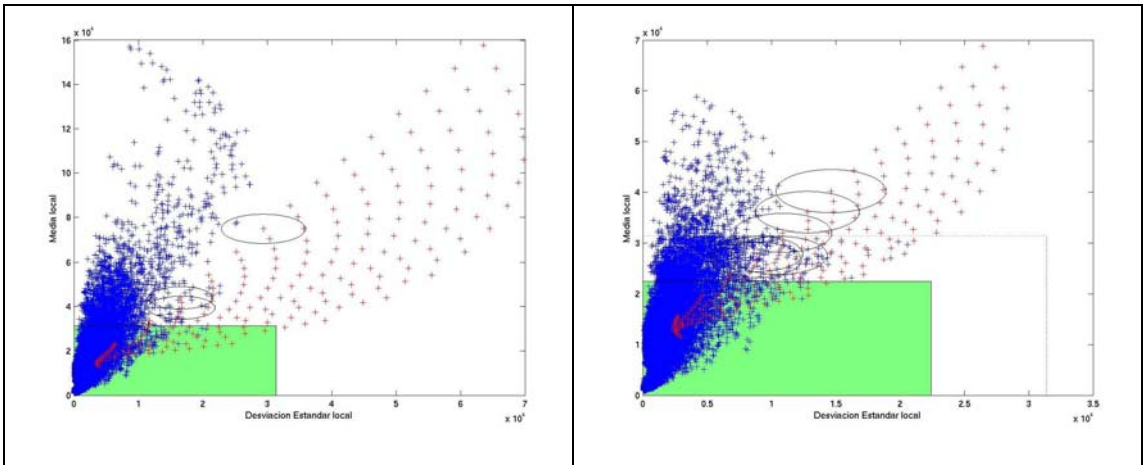


(a) (b)

Figura 4.27 : Función de densidad de los *Espacios Característicos*.

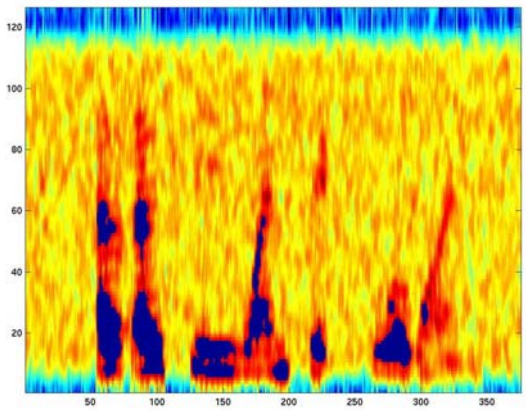
Otro punto de vista del objetivo final de la segmentación es claramente dejar en la Región de Confianza todos esos píxeles que se vislumbran como ruidosos, extrayendo por tanto el resto.

Los cambios del *Espacio Característico* y del espectrograma para cada iteración son mostrados en las Figura 4.28 y Figura 4.29. En la primera de ellas, se muestra el estudio completo realizado en los *Espacios Característicos* de las dos primeras iteraciones: *Grids Teóricos*, semillas seleccionadas, Región de Confianza y Área de trabajo. En la siguiente Figura se muestran las sucesivas extracciones de señal para cada iteración. De nuevo, las zonas azules en el espectrograma indican las regiones de señal que han sido extraídas. Pueden observarse en cada iteración los mismos cambios que los comentados para el caso de los tres chirps.

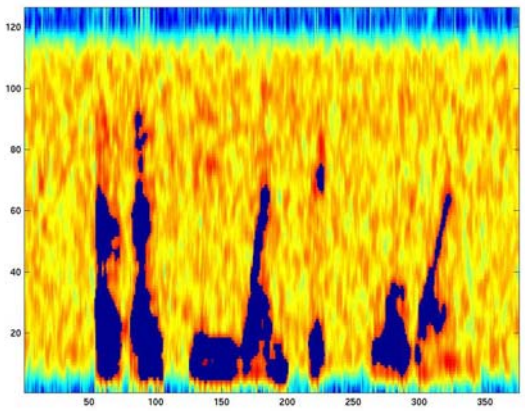


(a) (b)

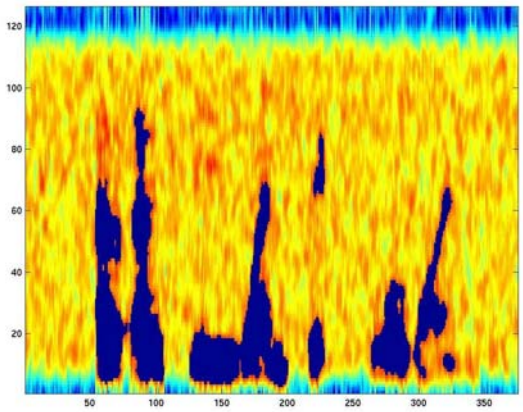
Figura 4.28 : Espacios Característicos, Grids Teóricos, semillas, Región de Confianza y Área de Trabajo de las Iteraciones 1 y 2.



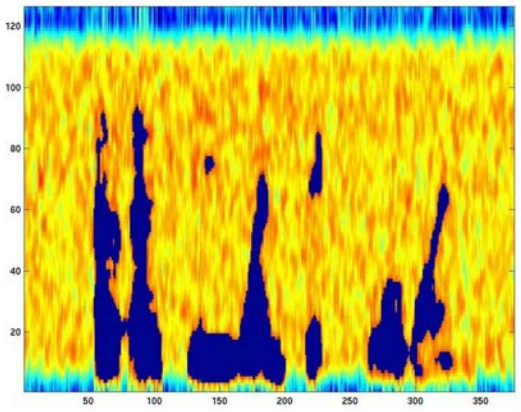
(a) Segmentación tras iteración 1



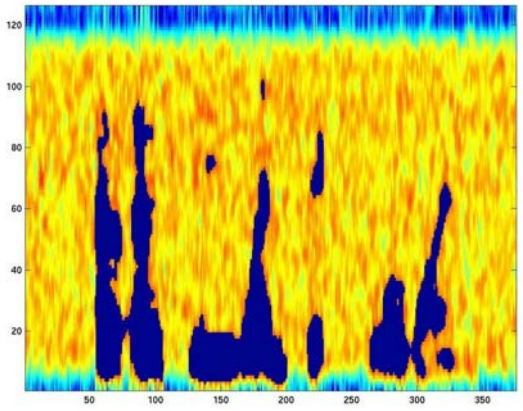
(b) Segmentación tras iteración 2



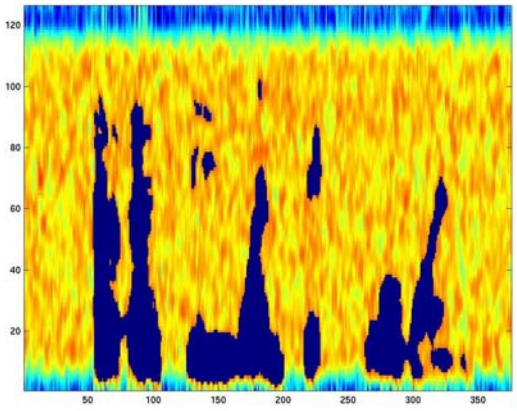
(c) Segmentación tras iteración 3



(d) Segmentación tras iteración 4



(e) Segmentación tras iteración 5



(f) Segmentación tras iteración 6

Figura 4.29 : Extracción de señal realizada en cada iteración.

Puede apreciarse cómo a medida que avanzan las iteraciones se consigue haber extraído una mayor cantidad de señal. Finalmente, tras la última iteración, se ha conseguido

claramente la segmentación de prácticamente toda la región de voz detectable para este caso de SNR.

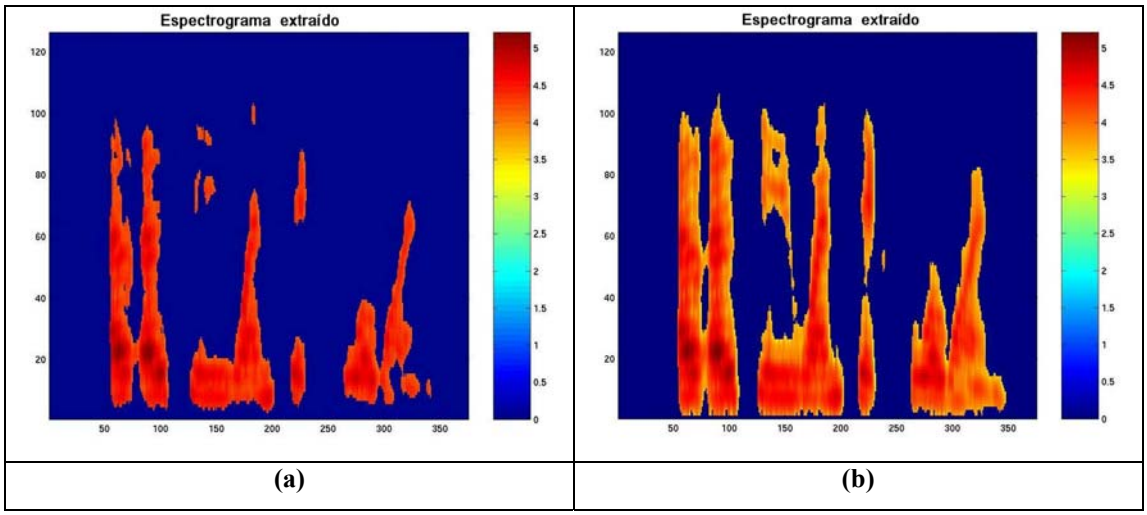


Figura 4.30 : Espectrograma final de las regiones extraídas. (a) Voz contaminada por ruido WGN. (b) Voz limpia (caso ideal).

La Figura 4.30 muestra los píxeles finalmente extraídos. En (a) vemos la extracción de la señal para el último ejemplo d SNR = 5 dB, en comparación con la segmentación realizada sobre la señal limpia, cuyo resultado podemos apreciar en (b). Se observa la esperada disminución de área segmentada respecto al caso ideal.

Para la segmentación de voz en estas condiciones también se realizó un estudio detallado en términos de precisión de segmentación de señal de voz en función de la SNR. La descripción y los resultados quedan explicados en el siguiente capítulo.

CAPÍTULO 5

EXPERIMENTOS, RESULTADOS Y DISCUSIONES

En este capítulo se describen todas las pruebas y experimentos realizados con el algoritmo de segmentación. Distinguimos dos grupos, en función del tipo de señal a examinar:

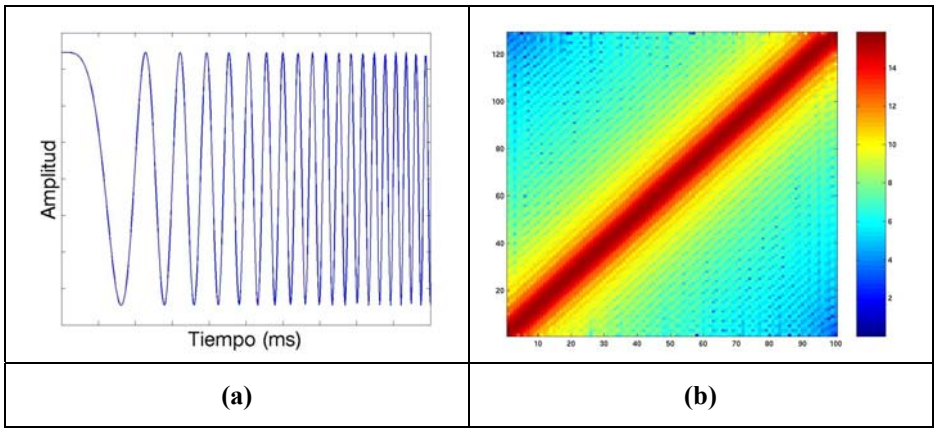
- *Resultados y discusión de los experimentos sobre señales sintéticas de chirps.* Tras mostrar los resultados de la segmentación para distintos chirps, se realiza un estudio para determinar tanto la precisión de la segmentación en función de la SNR, como la precisión de la estimación de la potencia de ruido.
- *Resultados y discusión de los experimentos sobre muestras de señales de voz de la Base de Datos Aurora 2*. De nuevo, tras mostrar los resultados para distintas señales de entrada, se realiza un estudio para establecer la importancia de la determinación de la forma espectral de cara a un Reconocimiento de Voz Automático. Fueron necesarias largas horas de simulaciones para obtener una medida fiable. En este caso, además, se estudió el funcionamiento del algoritmo para distintos tipos de ruidos reales.

5.1 RESULTADOS Y DISCUSIÓN DE LOS EXPERIMENTOS SOBRE SEÑALES SINTÉTICAS DE CHIRPS

5.1.1 Muestra de Resultados

Probamos el algoritmo con señales sintéticas basadas en chirps inmersos en distintos niveles de ruido Gausiano blanco. El rango de SNR estudiado es de 20 dB a -5 dB. Recordamos que la señal chirp $x(t)$ es únicamente una señal sinusoidal cuya frecuencia aumenta linealmente con el tiempo y a la que se le añade la señal de ruido Gausiano blanco $n(t)$, resultando la suma de ambas la señal de entrada al sistema $y(t)$:

$$y(t) = x(t) + n(t) = A \cdot \cos(w(t) \cdot t) + n(t) = A \cdot \cos((a + b \cdot t) \cdot t) + n(t) \quad \textbf{(5.1)}$$



(a) **(b)**

Figura 5.1 : (a) Forma de onda del chirp; (b) Correspondiente espectrograma.

La Figura 5.1 muestra en primer lugar (a) la forma de onda de un chirp limpio (sin contaminación alguna de ruido) luego la SNR = ∞. Seguidamente (b) su correspondiente espectrograma. La Figura 5.2 muestra la actuación del algoritmo sobre los chirps en función de la SNR. Los niveles representados son los más característicos: 10 dB, 5 dB, 0 dB, y – 5 dB. Las representaciones son la forma de onda del chirp contaminado, su espectrograma original, y el espectrograma segmentado. En ellas podemos observar en primer lugar cómo es la evolución de la influencia del ruido en la señal, tanto en su forma de onda temporal, como en el espectrograma. Y por supuesto, también podemos percibir la degradación del resultado de la segmentación a medida que la SNR empeora. Antes de realizar unas medidas de la precisión podemos comentar dicha evolución:

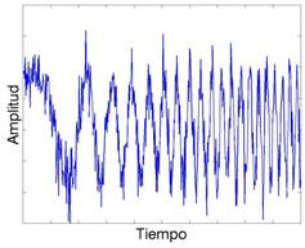
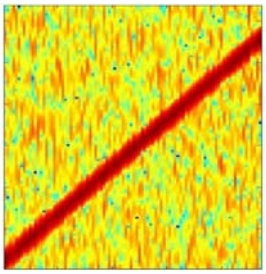
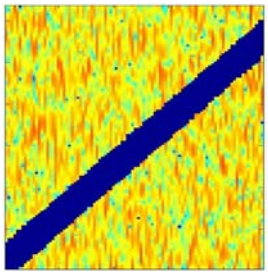
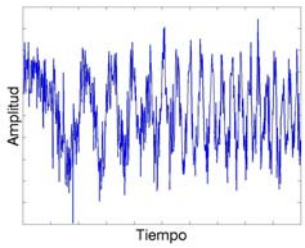
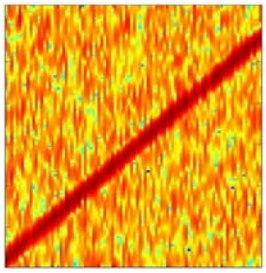
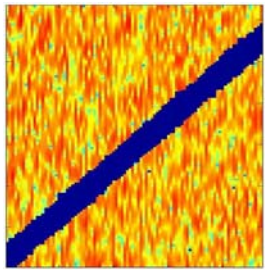
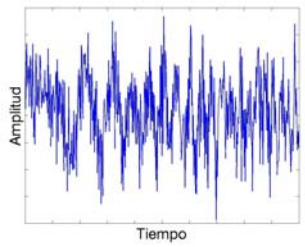
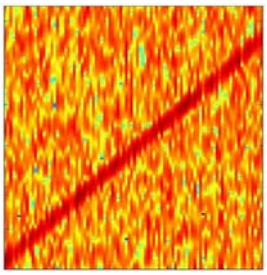
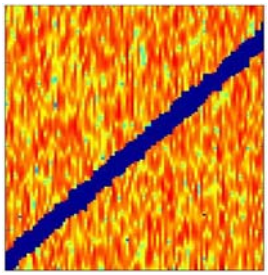
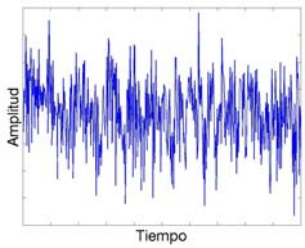
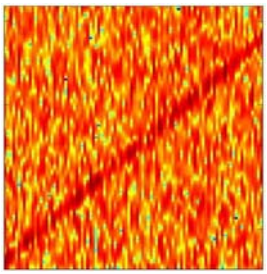
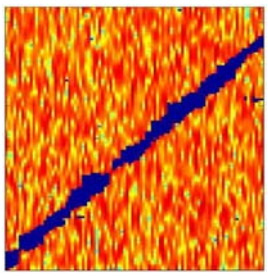
SNR (dB)	Forma de Onda	Espectrograma original	Espectrograma segmentado
10 dB			
5 dB			
0 dB			
-5 dB			

Figura 5.2 : Resultado de la segmentación de señales sintéticas chirps con SNR 10 dB, 5 dB, 0 dB y -5 dB. Remarcar que el espectrograma segmentado es el espectrograma con su señal determinista extraída (regiones azules).

Tal y como muestra la Figura 5.2, el área extraída del chirp original decrece a medida que la SNR disminuye. Para una SNR de 10 dB, la segmentación es prácticamente perfecta, ya que incluso los bordes del área segmentada están nítidamente definidos de forma lineal. Para una SNR de 5 dB, la segmentación es muy buena, de nuevo detectando el chirp y con los bordes también linealmente definidos aunque algo peor que en el anterior caso. Para una SNR de 0 dB puede observarse una correcta segmentación pero comienza a apreciarse un mayor deterioro en la definición de los bordes. Finalmente, para la peor condición SNR estudiada (-5

dB), se observa una detección del chirp, aunque de manera imperfecta. La línea de chirp detectada no llega a ser continua del todo, y es bastante más estrecha que la de los anteriores casos. De hecho, puede incluso observarse cómo a ese nivel el algoritmo comienza a confundir pequeñas regiones de ruido como parte de señal determinista. A pesar de todo, un altísimo porcentaje de la señal segmentada pertenece al chirp, demostrando que el algoritmo es capaz de detectar regiones de señal incluso a SNR negativas, algo que la mayoría de los algoritmos basados en la SNR no consiguen. Estos sistemas tienen un muy buen funcionamiento a altas SNR, pero simplemente confunden regiones de ruido con regiones de señal determinista cuando la SNR se aproxima a 0 dB, limitando bastante su rango de funcionamiento. Los siguientes experimentos fueron llevados a cabo para comprobar con mayor precisión la robustez del algoritmo bajo distintas SNRs.

5.1.2 Evaluación de la robustez del sistema bajo distintas SNRs

5.1.2.1 Experimento 1

Objetivo

Evaluar la robustez del algoritmo en términos de área segmentada. Obtención de una sola imagen que muestre la evolución de la segmentación del algoritmo: relación entre la SNR y el área segmentada.

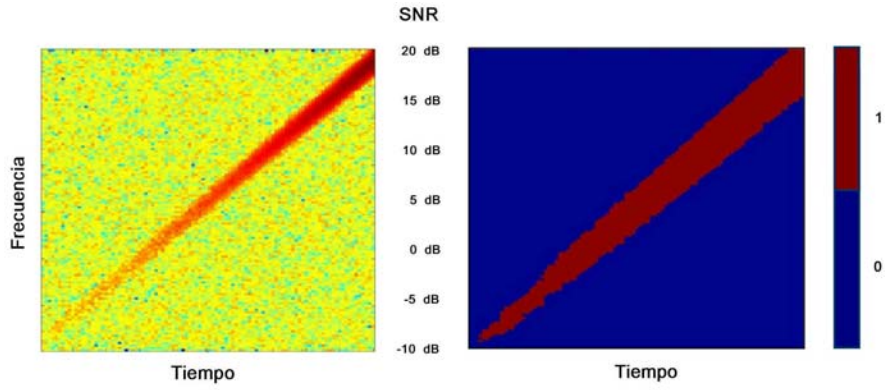
Procedimiento

Producimos una señal chirp cuya amplitud es logarítmicamente creciente a lo largo del eje temporal en un entorno de ruido WGN constante:

$$y(t) = x(t) + n(t) = A(t) \cdot \cos(w(t) \cdot t) + n(t) = A(a + b \cdot 10^t) \cdot \cos((c + d \cdot t) \cdot t) + n(t) \quad \textbf{(5.2)}$$

Esta señal presenta una SNR en aumento lineal en decibelios a lo largo del eje de la frecuencia (rango desde -10 dB a 20 dB). Tras la aplicación de esta señal a nuestro algoritmo, representamos la señal extraída con un valor unitario, imponiendo la señal considerada como ruido con valor cero.

Resultado



(a)	(b)

Figura 5.3 : Actuación del sistema en función de la SNR.

Discusión

La Figura 5.3 muestra claramente que el área segmentada se reduce en proporción al decrecimiento de la SNR. Se comprueba que el algoritmo es robusto al poder detectar señal incluso cuando ésta puede verse con dificultad en el espectrograma debido a la baja SNR, tal y como puede apreciarse en la esquina inferior izquierda de la Figura 5.3 (a). La segmentación de este espectrograma, Figura 5.3 (b), muestra que la señal es prácticamente perdida en torno a un valor de SNR realmente bajo: -9 dB. Este experimento es sobre todo un tanteo visual. El siguiente experimento obtiene resultados estadísticos en función de la SNR.

5.1.2.2 Experimento 2

Objetivo

Evaluar la robustez del algoritmo usando una función coste objetivo en términos de área correctamente segmentada. Analizar matemáticamente mediante esta función coste el porcentaje de área correctamente segmentada en función de la SNR.

Procedimiento

En primer lugar aplicamos el algoritmo de segmentación a una señal chirp limpia. Su segmentación fue la base para las siguientes pruebas, al considerar el área segmentada en este experimento como la correcta e ideal para la segmentación de los chirps con distintas SNRs. Seguidamente, se realizó la segmentación de 100 chirps con distintos niveles de SNR, y se comparó el área segmentada en cada caso respecto a la ideal. La precisión de la segmentación obtenida fue calculada por tanto:

$$\text{Seg.Prec.}|_{\text{SNR}=x\text{dB}} = \frac{\sum_{n,k} \text{Seg}(n,k)|_{x\text{dB}}}{\sum_{n,k} \text{Seg}(n,k)|_{\text{limpio}}} * 100\% \quad \textbf{(5.3)}$$

donde la precisión a x dB viene dada por el porcentaje de área segmentada sobre el área ideal. En la ecuación 5.3, *n* y *k* son las dimensiones de la imagen del espectrograma. A partir de esta ecuación, se define el error asociado:

$$\text{Seg.Err.}|_{\text{SNR}=x\text{dB}} = 100\% - \text{Seg.Prec.}|_{\text{SNR}=x\text{dB}} \quad \textbf{(5.4)}$$

donde el error de segmentación a x dB es definido como el 100% menos la precisión de la segmentación a x dB. Finalmente, representamos la precisión de la segmentación frente a la SNR para evaluar la dependencia del algoritmo con la SNR. El resultado es mostrado en la Figura 5.4.

Resultado

El porcentaje de área segmentada es representado en función de la SNR a partir de los 100 resultados obtenidos. En la Figura 5.4 se muestra asimismo la aproximación de cuarto orden del resultado experimental.

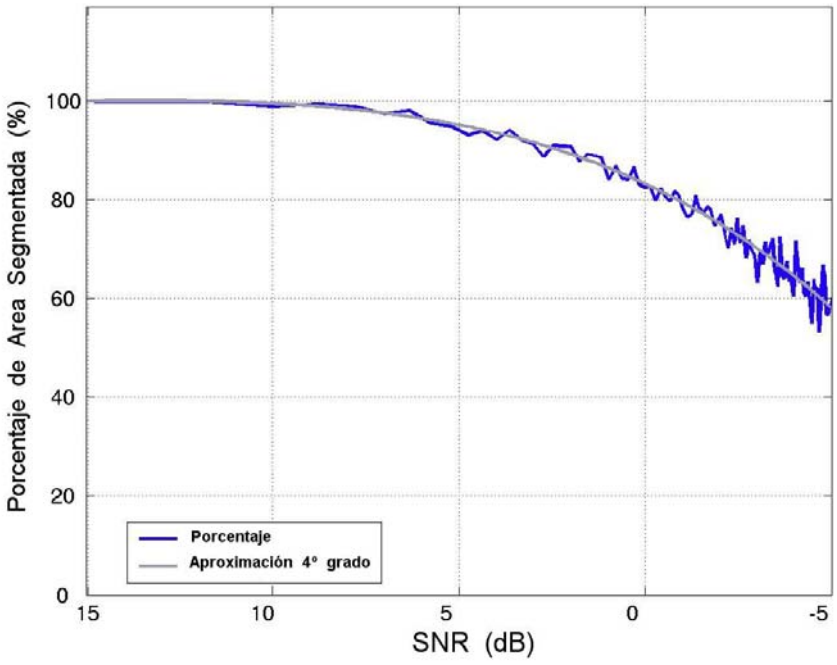


Figura 5.4 : Precisión de la segmentación frente a la SNR.

Discusión

La Figura 5.4 muestra que la precisión de la segmentación es perfecta por encima de los 8 dB. A partir de ese valor, la precisión decae gradualmente, aunque el algoritmo sigue

detectando un más que aceptable 82.8% a 0 dB e incluso un 60% a -5 dB. Queda demostrado por tanto que el algoritmo es robusto incluso bajo condiciones de señal a ruido muy pobres.

5.1.3 Evaluación de la precisión de la estimación de la potencia de ruido contaminante

Objetivo

Evaluar la precisión de la estimación de la potencia del ruido contaminante, i.e. validar la ecuación 3.9.

Procedimiento

El proceso de segmentación es aplicado a las 100 señales de chirps de la base de datos usada en el experimento de la Sección 5.1.2.2. La potencia del ruido de contaminación es estimada al final de cada segmentación usando la ecuación 3.9. Por último, comparamos la potencia de ruido estimada con la potencia de ruido real añadida, para poder así evaluar la precisión de la estimación. El resultado de las 100 comparaciones es mostrado gráficamente en la Figura 5.5.

Resultado

Los puntos azules representan la potencia de ruido estimada a una SNR de *x* dB. La línea roja es la potencia real añadida, es decir, el resultado que debería haberse obtenido si la estimación de la potencia del ruido fuera del 100%.

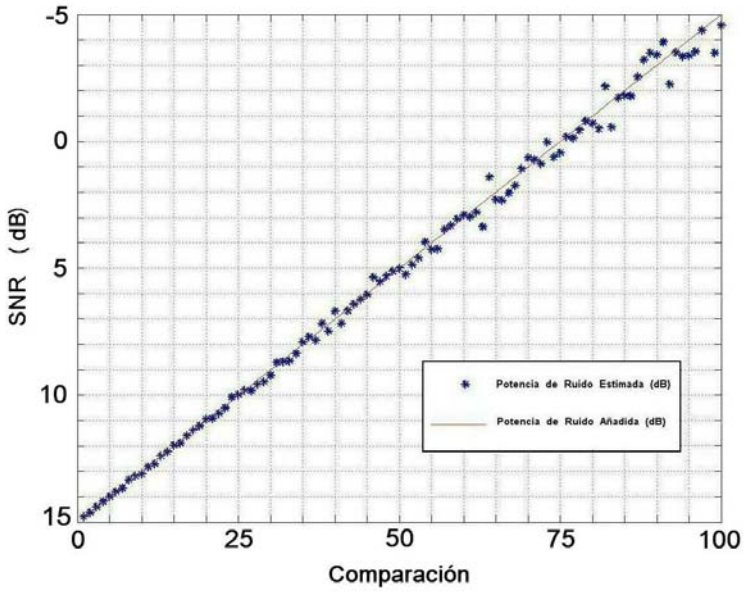


Figura 5.5 : Estimación de la Potencia del ruido contaminante frente a la Potencia real del ruido añadido.

Discusión

Tal y como se muestra en la Figura 5.5, los puntos azules quedan bastante cercanos de la línea roja. En concreto, vemos que la estimación es muy precisa para una alta SNR, y bastante aproximada para bajas SNR. Estos resultados son parte de la explicación del comportamiento observado en la Sección 5.1.2.1. De esta forma, se corresponde la segmentación muy precisa por encima de los 8 dB con la alta precisión de estimación, así como el decaimiento gradual de la segmentación en concordancia con la caída de precisión de estimación de la potencia de ruido para valores de SNR cada vez más bajos.

5.2 RESULTADOS Y DISCUSIÓN DE EXPERIMENTOS EN MUESTRAS DE VOZ

El último objetivo del desarrollo del algoritmo es su aplicación a espectrogramas de voz para la mejora del reconocimiento de voz. Esta sección evalúa la conveniencia del algoritmo para este propósito según qué condiciones.

5.2.1 Evaluaciones del potencial del algoritmo en el filtrado de espectrogramas de voz

Procedemos a evaluar visualmente el resultado de la aplicación del algoritmo de segmentación a señales de voz. Distinguiremos dos grupos de experimentos. En primer lugar, examinaremos un grupo de señales de voz inmersas en ruido WGN. Posteriormente, evaluaremos las imágenes de señales de voz inmersas en ruidos reales de situaciones cotidianas.

5.2.1.1 Experimento 1

Objetivo

El primer paso es la evaluación del algoritmo ante señales de voz corrompidas por ruido WGN. Obviamente estas señales siguen siendo desde el punto de vista teórico unas señales sintéticas, al ser combinación de una señal real de voz con una señal ideal de ruido Gausiano blanco. En la realidad nunca se presentan señales de este tipo, pero constituyen el paso intermedio entre la segmentación de señales sintéticas puras, y la segmentación de señales reales.

Procedimiento

En primer lugar hay que describir la fuente de señales de voz usada. Los experimentos han sido realizados en un subgrupo de la base de datos TIDigits [8]. Esta base de datos es

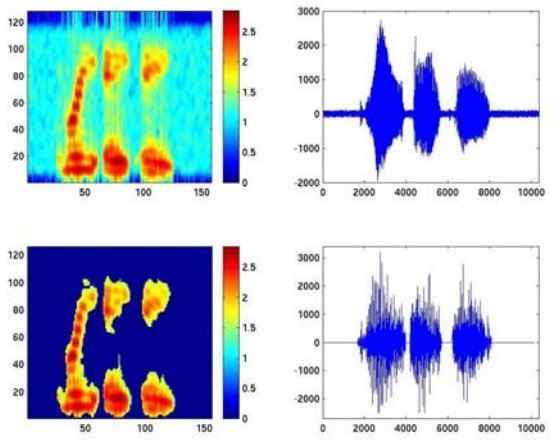
AURORA 2, obtenida a través de la Agencia de Distribución de los recursos de Evaluación Lingüísticos (ELDA: Evaluations and Language resources Distribution Agency). Dicha Agencia tiene el acuerdo de distribución de la Asociación Europea de los Recursos del Lenguaje (ELRA: European Language Resources Association), la cual es una Organización Europea sin ánimo de lucro establecida en Luxemburgo en Febrero de 1995. Dentro de sus diferentes Proyectos, se encuentra AURORA, una enorme base de datos formada por miles de ficheros de voz ante distintas características como son: tipo de ruido, potencia de ruido, idioma, distorsión, tipo de vocabulario… El objetivo perseguido es el de la estandarización de los resultados de los diferentes algoritmos de extracción de señal contaminada por ruido, al evaluar todos ellos los mismos ficheros de voz.

Dentro de las diferentes versiones de este Proyecto, hemos utilizado la base de datos AURORA 2, la cual permite el uso controlado de adición de ruido según las normas de filtrado de la ITU. Estos miles de ficheros fueron procesados para añadirles ruido WGN a distintos niveles de SNR [3].

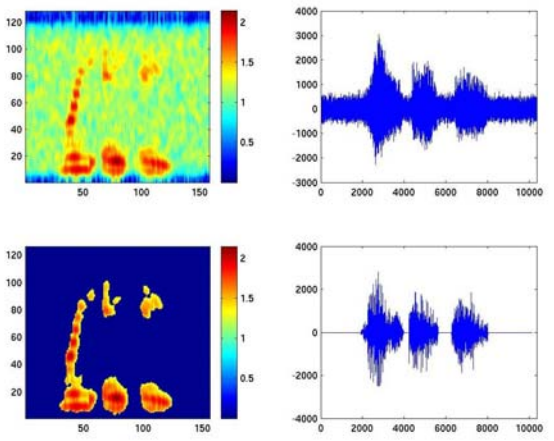
Una vez obtenida esta base de datos, la procesamos con nuestro algoritmo. Los resultados que se muestran a continuación pertenecen a la segmentación de siempre la misma señal de voz pero contaminada por distintos niveles de ruido.

Resultado

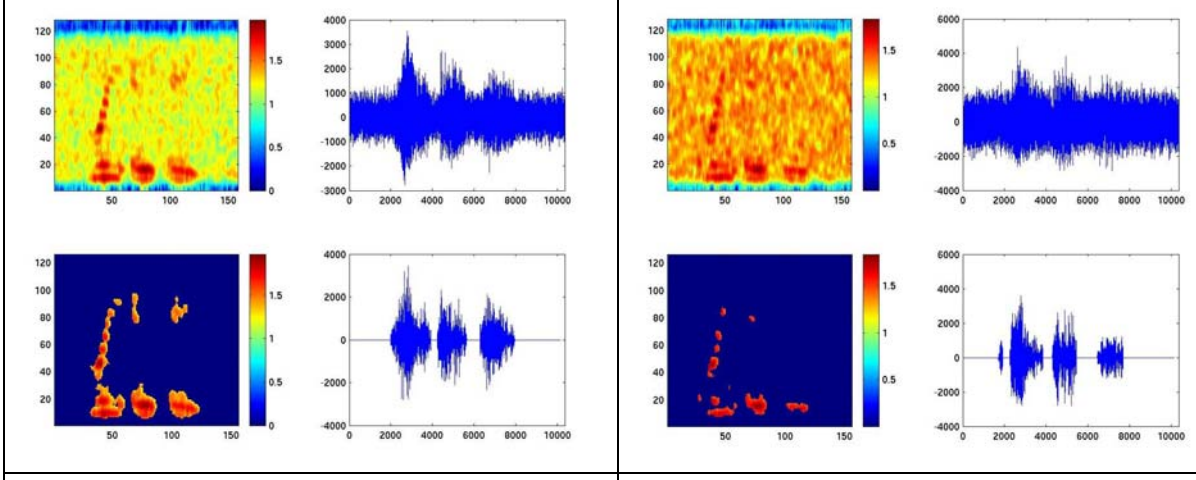
En cada cuadrícula de la Figura 5.6 se representan en la parte superior el espectrograma de la señal y su forma de onda en el tiempo. La parte inferior contiene el espectrograma segmentado y su transformación al dominio temporal de nuevo.



(a) SNR = 20 dB.



(b) SNR = 5 dB.



(c) SNR = 0 dB.	(d) SNR = -5 dB.

Figura 5.6 : Resultados de segmentación de imágenes de voz inmersas en ruido WGN

Discusión

La Figura 5.6 muestra ejemplos de contaminación de ruido WGN a señales de voz. En la Figura 5.6 (a) se contempla que la mayor parte de la región de voz es detectada. Este ejemplo es claro al tener una SNR de 20 dB, y el mismo comportamiento es observado hasta una SNR de aproximadamente 8 dB. A partir de ahí, se observa una menor región de detección. Esta pérdida de detección es progresiva, tal y como se muestra en la Figura 5.6 (b) y (c) con una SNR de 5 y 0 dB respectivamente. Por último, la Figura 5.6 (d) evalúa el resultado para el peor caso de SNR estudiado: -5 dB, donde aún pueden ser detectadas las regiones más energéticas de voz.

Este resultado es bastante esperanzador, pues incluso en las peores condiciones de SNR se detectan sustanciales regiones de voz.

5.2.1.2 Experimento 2

Objetivo

Observar la conveniencia y efectividad del algoritmo en el filtrado de espectrogramas de voz contaminados por ruido de diferentes situaciones reales, tales como el ruido existente dentro de un restaurante o dentro de un coche en movimiento.

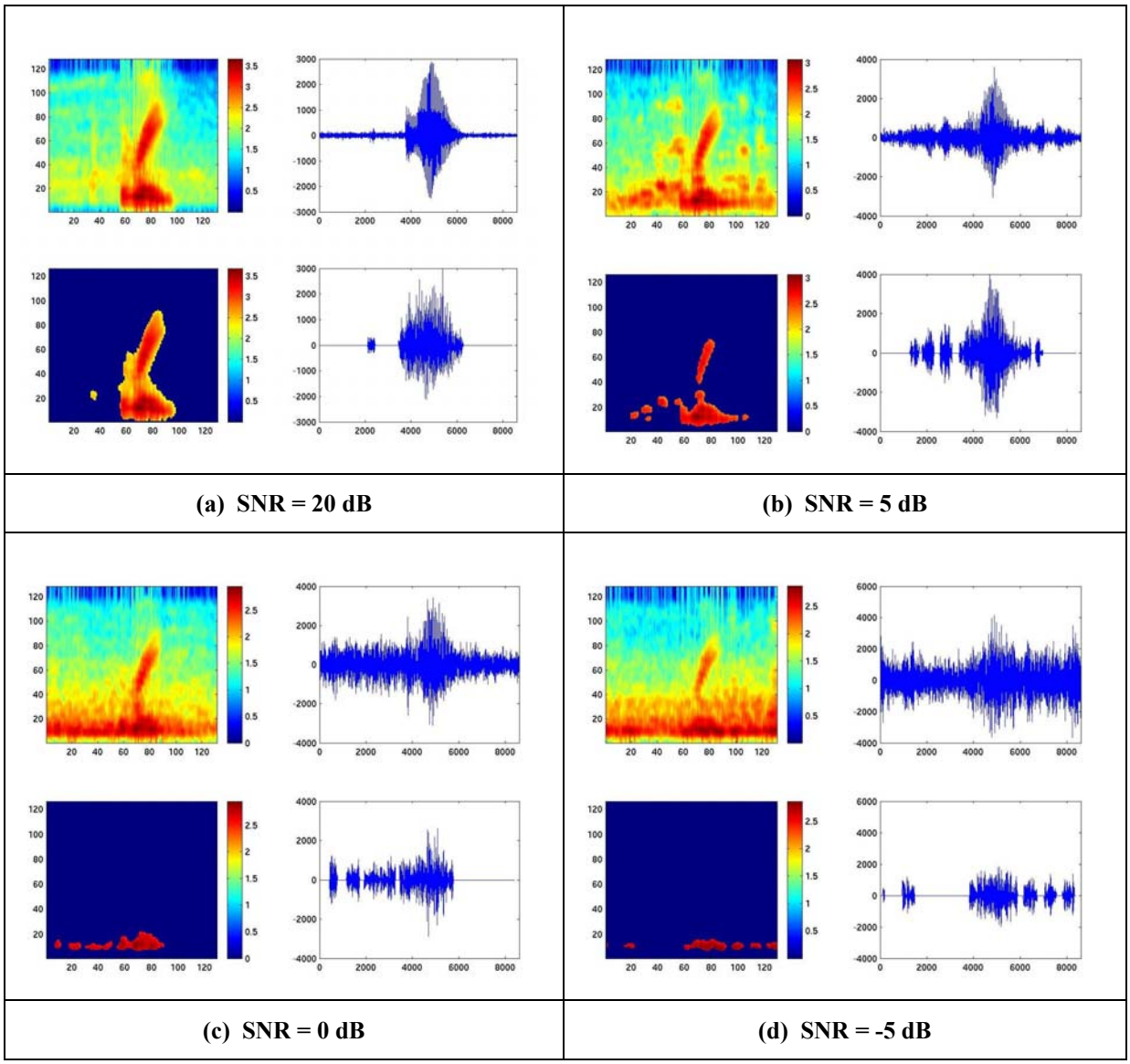
Procedimiento

El algoritmo es aplicado a los espectrogramas de voz obtenidos de los ficheros de la base de datos AURORA 2. La diferencia en este caso es que el ruido contaminante de las señales es real y viene incluido en la base de datos, no siendo necesaria manipulación alguna.

Al igual que en el experimento anterior, los espectrogramas filtrados son de nuevo transformados al dominio temporal como señal mejorada.

Resultado

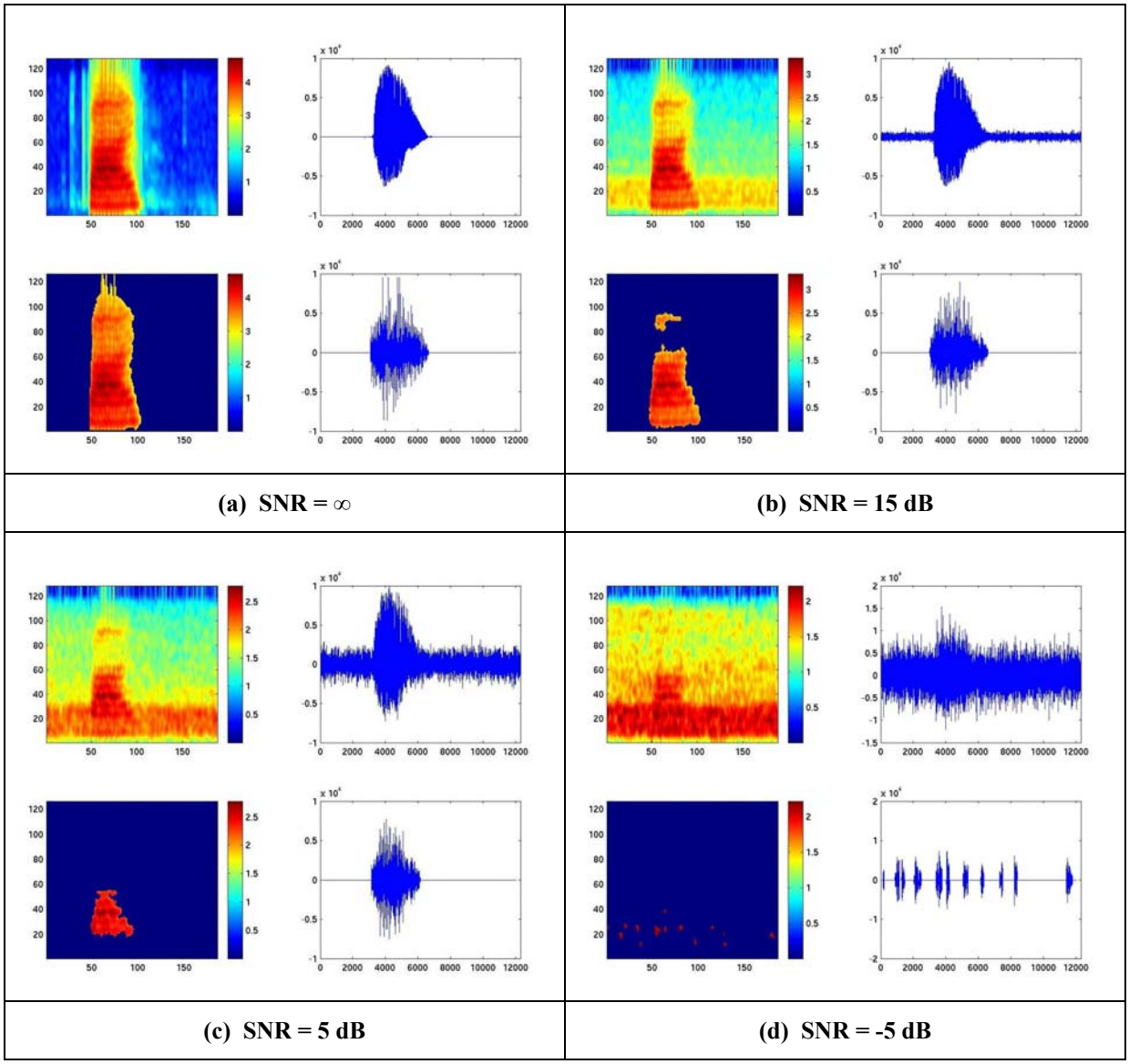
Las Figuras 5.7 y 5.8 muestran algunos ejemplos de segmentación sobre muestras de espectrogramas de voz obtenidas en condiciones reales de ruido en el interior de un restaurante y en el interior de un coche en movimiento respectivamente. Cada figura muestra la segmentación de la misma señal bajo cuatro potencias de ruido diferentes, pudiendo comparar así la región segmentada a distintos niveles de SNR. Hay que señalar que en ningún momento se está procediendo a la eliminación del ruido que cae dentro de la señal extraída, sino que únicamente se delimitan las regiones de voz en el espectrograma para cada caso.



(a) SNR = 20 dB	(b) SNR = 5 dB
(c) SNR = 0 dB	(d) SNR = -5 dB

Figura 5.7 : Resultados de segmentación de señales de voz contaminadas por el ruido existente en el interior de un restaurante.

El resultado de la Figura 5.7 muestra que el algoritmo trabaja correctamente como filtradora de la onda de voz en el tiempo para unas SNRs elevadas. A partir de 5 dB el algoritmo deja de segmentar efectivamente, al evaluar incorrectamente la potencia de ruido para todo el espectrograma. La Figura 5.7 (d) muestra la escasa e incorrecta detección de señal para el peor caso de -5 dB.



(a) SNR = ∞	(b) SNR = 15 dB
(c) SNR = 5 dB	(d) SNR = -5 dB

Figura 5.8 : Resultados de la segmentación de señales de voz contaminadas por el ruido existente en el interior de un coche en movimiento.

De nuevo, el resultado de la segmentación queda mostrado en la Figura 5.8, donde puede analizarse la evolución de los resultados del algoritmo, similares a los obtenidos en la Figura anterior. Una de las diferencias es que para este tipo de ruido, la detección y extracción

es bastante considerable hasta los 5 dB. En el caso de SNR de -5 dB la segmentación falla de nuevo completamente.

Discusión

Somos conscientes de que la actuación del algoritmo en espectrogramas reales de voz no es tan buena como en la de señales sintéticas basadas en chirps. La razón es que el ruido posee distintas características. El ruido WGN tiene característica espectral plana, afectando por igual a todas las regiones del espectrograma. Tal y como acabamos de comprobar, los ruidos de situaciones reales presentan características muy distintas. Por ejemplo, el ruido de restaurante tiende a corromper más la señal a bajas frecuencias, y menos a altas. Aun así, de vez en cuando aparecen zonas a altas frecuencias, probablemente procedentes de otras conversaciones de voz ajenas, tal y como se aprecia en la Figura 5.8 (b). En el caso del ruido en el interior de un coche en movimiento, la evolución de la señal es más clara. Este tipo de ruido está muy concentrado en torno a una banda de baja frecuencia perfectamente definida, procediendo casi con toda seguridad del ruido constante del motor del coche. El algoritmo segmenta buena parte de la señal de voz hasta el momento en el que la potencia del ruido supera a la de la señal, momento en el que el algoritmo considera prácticamente todo el espectrograma como ruido.

A pesar de todo, hay que señalar que aunque el resultado ante ruidos como los del restaurante o el interior de un coche no sean tan buenos como en caso del ruido WGN, sigue teniendo un comportamiento aceptable a SNRs más lógicas, tal y como muestran las Figuras 5.7 (b) y 5.8 (c), En ellas, un buen porcentaje de región de voz es detectado a 5 dB. El algoritmo tiene por tanto el potencial de ser optimizado para una mejor actuación. Aunque queda comentado en el Capítulo 8, una mejora evidente del algoritmo es su aplicación a distintas regiones del espectrograma por separado, extrayendo señal por zonas y evaluando así las distintas potencias del ruido de cada región de la imagen.

5.2.2 Evaluación de la Importancia de la forma espectral en la actuación del ASR

El algoritmo es capaz de señalar cuáles son las regiones de señal en un espectrograma ruidoso. De forma implícita, elimina todo el ruido que cae fuera de la región extraída, pero no realiza en momento alguno eliminación del ruido que cae dentro de la región de voz. La evaluación de la forma espectral es el siguiente paso:

Objetivo

Tasar el valor de la determinación de la forma espectral. Para ello, únicamente habrá que tener en cuenta la forma de la región extraída, y eliminar la información de amplitud. De esta forma podremos evaluar la contribución de la forma de los patrones espectrales de la voz en la actuación del ASR.

Procedimiento

En la Sección 5.2.1 hemos descrito la procedencia de nuestra base de datos. Para este experimento hay que distinguir dos grupos de datos. En primer lugar, se escogieron 1001 señales limpias, y se procedió a contaminarlas con ruido WGN de distinta potencia. Las SNRs resultantes de estos ficheros manipulados son: 20, 15, 10, 5, 0 y -5 dB. Este constituye nuestro primer grupo de estudio, formado de esta manera por 7 subconjuntos de 1001 ficheros cada uno: 1 por las señales limpias, y 6 por cada tipo de SNR. Remarcamos que los 6 últimos grupos están compuestos por las 1001 señales de voz del primer grupo, pero afectadas por distintas potencias de ruido.

A continuación se utilizó nuestro algoritmo para segmentar cada uno de estos grupos de distintas SNR. La variación introducida en el algoritmo fue la de eliminar toda información de amplitud a la salida del espectrograma, imponiendo los valores constantes '0' a las regiones no extraídas, y una amplitud constante 'A' a las regiones extraídas. Cada uno de estos espectrogramas es transformado de nuevo al dominio temporal, obteniendo por tanto una forma de onda cuya única información es la relativa a la forma espectral. La Figura 5.9 muestra una salida de esta variación del algoritmo. Este constituye nuestro segundo grupo de estudio, obteniendo exactamente el mismo número de señales que las existentes en el primer grupo. Este grupo ha extraído la información de la forma espectral de las anteriores señales.

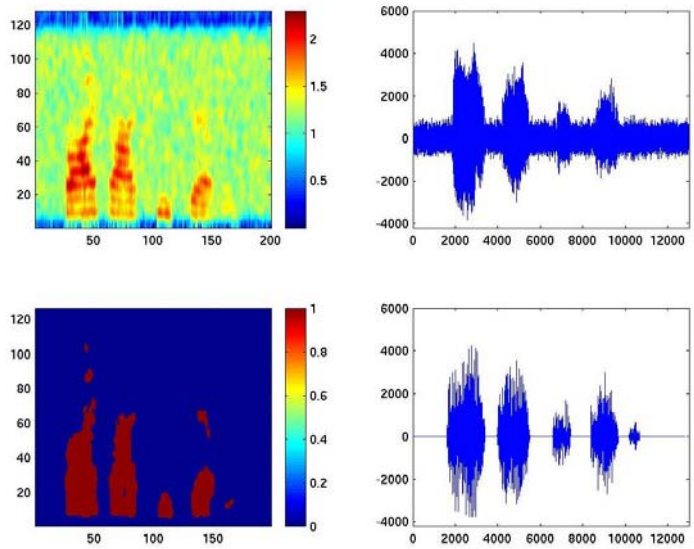


Figura 5.9 : Salida del algoritmo eliminando toda información de amplitud.

El siguiente paso consistió en utilizar un algoritmo de reconocimiento de palabra automático (ASR: Automatic Speech Recognition). En primer lugar se realizó su entrenamiento, utilizando para ello 8440 ficheros de voz limpios. Una vez que el ASR estaba entrenado, se le pasó cada uno de los subgrupos pertenecientes al primer grupo. Es decir, conjuntos de 1001 ficheros con distintas SNRs sin manipulación alguna, sin ningún tipo de filtrado o segmentación. De esta manera obtuvimos una primera medida de precisión de reconocimiento de palabra para cada caso de SNR formando nuestro Resultado Base (Figura 5.10).

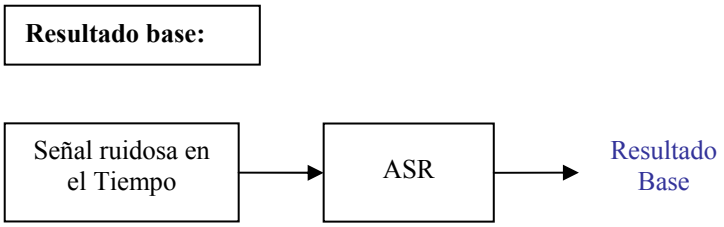


Figura 5.10 : Obtención del Resultado Base.

A continuación, se introdujeron en el ASR cada uno de los subconjuntos del segundo grupo de estudio. La Figura 5.11 muestra el proceso completo de la obtención de los datos de la Forma Espectral.

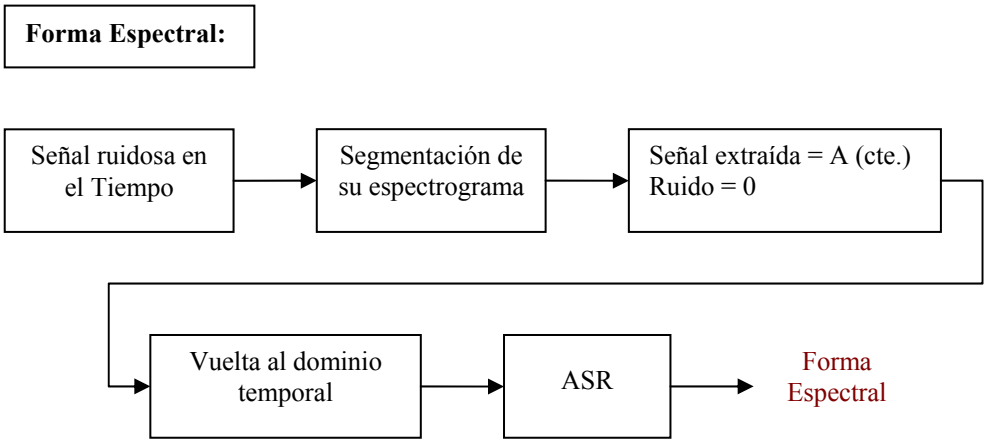


Figura 5.11 : Obtención de la Forma Espectral.

El ASR dio sus resultados de precisión para cada caso. La comparación del Resultado Base y de la Forma Espectral es mostrada en la Figura 5.12.

Resultado

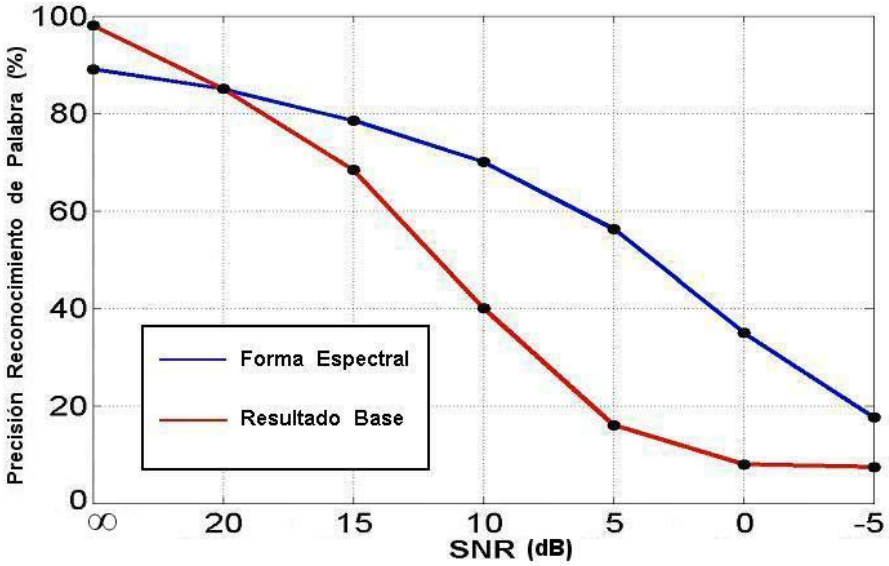


Figura 5.12 : Comparación de la Precisión en el Reconocimiento de Palabra del ASR entre el Resultado Base y la Forma Espectral obtenida tras la segmentación.

Discusión

El grupo de estudio morfológicamente filtrado da unos resultados asombrosos. Teniendo en cuenta que no hay información de amplitud, se obtiene nada menos que un 89% de precisión de reconocimiento de palabra en el caso limpio. A partir de una SNR de 20 dB se obtienen mejores resultados por la sola detección de la Forma Espectral que por el Resultado Base. A su lado, una precisión de reconocimiento de palabra del 56% es conseguida a 5 dB, lo cual da una mejora del 40% comparado con el Resultado Base, formada por un 16%. Incluso a 0 dB, se obtiene una mejora del 25%. Como conclusión, se muestra que la forma espectral por si sola sin información alguna de amplitud de energía lleva a unos resultados de precisión de reconocimiento de palabra muy buenos incluso en malas condiciones de ruido. El resultado de este experimento es arrollador en el sentido de que demuestra la gran relevancia de la detección de la forma espectral, e indirectamente, la utilidad del algoritmo para la mejora en el reconocimiento de voz.

CAPÍTULO 6
CONCLUSIONES

El principal objetivo de este Proyecto consistía en la implementación independiente del algoritmo de segmentación automático propuesto en [1,2]. Este primer paso fue completado exitosamente, obteniendo resultados y figuras muy similares a las publicadas.

El siguiente paso radicaba en una ampliación del estudio de las características de este algoritmo, en términos por ejemplo de SNR. De esta manera, hemos podido comprobar que la técnica implementada consigue resultados sorprendentemente buenos al filtrar el ruido WGN que corrompe la señal sintética, asociando una robustez de una precisión del 60% en la detección de SNR a -5 dB.

Por último, se decidió la aplicación de este algoritmo en el campo de la detección de voz. Se creó una amplia base de datos de ficheros de voz contaminados por ruido WGN a partir de la Base de datos Europea AURORA 2. Los extensos experimentos llevados a cabo dieron como conclusión el gran potencial de la aplicación de la técnica de cara al reconocimiento automático de palabras. Una muestra es la mejora del 40% obtenida a 5 dB sobre el resultado base logrado por el ASR. Una de las mejores características del algoritmo es consecuentemente su buen comportamiento tanto en señales sintéticas como en señales de voz a bajas SNR, algo difícil de conseguir. A la vez, se ha demostrado la grandísima importancia de la localización de la forma espectral. Pensamos que la segmentación de las regiones de voz puede constituir el primer y crítico paso en el proceso de la detección de voz automática.

Como conclusión podemos afirmar que la realización de este Proyecto constituye un buen acercamiento en las tareas de estimación de ruido, segmentación de regiones de señal sintéticas y de voz, y evaluación de la importancia de la forma espectral de cara al reconocimiento automático de palabras. A pesar de las muy buenas y conocidas capacidades del filtrado morfológico para el procesamiento de imágenes, pocos habían pensado en las posibilidades de su aplicación para el filtrado de voz con ruido. Aunque es un trabajo bastante preliminar en las señales de voz, este Proyecto no deja de ser definitivamente un buen comienzo para la continuación de la investigación en este área. Al mismo tiempo, contribuye a las limitadas fuentes de referencia disponibles para este campo específico.

CAPÍTULO 7
LÍNEAS FUTURAS

A pesar del éxito de este Proyecto, deben llevarse a cabo continuaciones y posteriores desarrollos experimentales para optimizar el algoritmo y continuar explorando el potencial del filtrado morfológico.

Una vez demostrada la gran eficiencia del algoritmo para la detección de señales sintéticas, el principal objetivo debe ser el de mejorar las prestaciones del algoritmo en la tarea de detección de voz. Ya ha quedado señalada la enorme importancia de la detección de la forma espectral para el reconocimiento automático de palabras, luego el tipo de algoritmo aquí presentado podría representar el primer paso esencial en toda detección de voz. Los siguientes pasos serían los relacionados con la aplicación de algunas de las muy diversas técnicas para añadir la información de amplitud a las regiones segmentadas eliminando el ruido contaminante en esa región. Un ejemplo podría ser la utilización de la substracción espectral.

Por tanto, la mejora del primer paso en la detección resulta fundamental. Sólo con los resultados obtenidos ya se vislumbra que la efectividad del proceso final de reconocimiento queda muy condicionada por la robustez del algoritmo de segmentación. Aunque los resultados obtenidos con el algoritmo son muy buenos, siempre podrán ser mejorados. En este sentido, las claves para una optimización recaen en la mejora de la precisión de la estimación del ruido para bajas SNR, y en la inclusión de alguna técnica para perfeccionar a posteriori el resultado de la segmentación. Así, en base a algún patrón o característica conocida de la voz,

podrían detectarse y corregirse regiones mal segmentadas, y viceversa: optimizar y ampliar zonas reconocidas como voz.

Junto a esta mejora comentada del algoritmo cabe también destacar otros entornos en los que su inclusión o ampliación puede ser muy interesante. Como ya se ha demostrado, el algoritmo obtiene muy buenos resultados en la detección de patrones espectrales inmersos en ruido WGN, pero no tanto ante la contaminación de otros tipos de ruido. El ruido WGN se caracteriza por tener una representación espectral uniforme, y el algoritmo esta preparado para realizar correctas estimaciones de este tipo de ruido, pero los ruidos reales presentan una dependencia de la frecuencia. Típicamente, y en entornos ruidosos como son los ejemplos de restaurantes, trenes o coches, la característica espectral del ruido es decreciente a lo largo de la frecuencia. Es esta función dependiente de la frecuencia la que hace que el algoritmo no funcione bien en estas condiciones, ya que la estimación de la potencia de ruido sería por tanto sólo correcta en un pequeño rango de ruido plano, e incorrecta en el resto.

En vista de este análisis, la principal sugerencia que se propone es el estudio de la aplicación de este algoritmo para distintas sub-regiones horizontales del espectrograma. En cada una de ellas se realizaría una distinta y precisa estimación de ruido, segmentando cada sub-región de la forma más adecuada. Otra opción es el diseño de un estimador de ruido función de la frecuencia, y la correspondiente aplicación al algoritmo de segmentación. Claramente, el nivel matemático y computacional asociado es visiblemente más elevado.

APÉNDICE A

MOMENTOS DE UNA DISTRIBUCIÓN χ^2 NO CENTRAL

Consideremos un conjunto de *D* variables Gausianas independientes X_d de media θ_d y varianza α. La variable aleatoria $X = \sum_{d=1}^{D} X_d^2$ es, por definición, una variable aleatoria $\alpha\chi^2_{D,\theta}$ con parámetro no central $\Theta = \sum_{d=1}^{D} \theta_d^2$ distinto de cero. Su PDF es de la forma

$$f_{\alpha\chi^2_{D,\theta}}(x) = \frac{1}{2\alpha}\left(\frac{x}{\theta}\right)^{(D-2)/4} e^{-(\theta+x/2\alpha)} I_{D/2-1}\left(\frac{\sqrt{x\theta}}{\alpha}\right) \qquad \textbf{(A.1)}$$

donde $I_c(x)$, es la función modificada de Bessel de orden *c* de primera clase, y viene definida por

$$I_c(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{c+2k}}{k!\Gamma(c+k+1)}, \; con\; x \geq 0 \qquad \textbf{(A.2)}$$

siendo $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$ la función Gamma.

Podemos encontrar en [10] la expresión general del momento *q* de la variable aleatoria $R = \sqrt{\sum_{d=1}^{D} X_d^2}$, la cual tiene, por definición, una ley de distribución Rice

$$E\left\{R^q\right\} = (2\alpha)^{q/2} e^{-\Theta/2\alpha} \frac{\Gamma\left(\frac{(D+q)}{2}\right)}{\Gamma\left(\frac{D}{2}\right)} F_1^1\left(\frac{D+q}{2}, \frac{D}{2}; \frac{\Theta}{2\alpha}\right) \quad \textbf{(A.3)}$$

donde $F_1^1(a,b;x)$ es la función confluente hipergeométrica definida por

$$F_1^1(a,b;x) = \sum_{n=0}^{+\infty} \frac{\Gamma(a+n)\Gamma(b)x^n}{\Gamma(a)\Gamma(b+n)n!} \quad \textbf{(A.4)}$$

La distribución $\alpha\chi_{D,\theta}^2$ no-central es una generalización de la distribución de Rice ya que $X = R^2$ sigue una distribución $\alpha\chi_{D,\theta}^2$ no-central. Sustituyendo (A.4) en (A.3) con $\Gamma(n) = (n-1)!$ y asumiendo $r = \theta/2\alpha$

$$\mu_q = E\left\{R^{2q}\right\} = (2\alpha)^q e^{-r} \sum_{n=0}^{+\infty} \frac{(q+n+\frac{D}{2}-1)!}{(n+\frac{D}{2}-1)!} \frac{r^n}{n!} \quad \textbf{(A.5)}$$

El primer cociente de la suma puede ser expresado como

$$\frac{(q+n+\frac{D}{2}-1)!}{(n+\frac{D}{2}-1)!} r^n = r^{1-D/2} \frac{d^q}{dr^q}\left[r^{q+n+D/2-1}\right] \quad \textbf{(A.6)}$$

lo que conduce a

$$\mu_q = (2\alpha)^q e^{-r} r^{1-D/2} \frac{d^q}{dr^q}\left[r^{q+D/2-1} \sum_{n=0}^{+\infty} \frac{1}{n!} r^n\right] \quad \textbf{(A.7)}$$

La suma bajo la expansión en derivadas de la serie de Taylor de la función exponencial en torno a cero:

$$\mu_q = (2\alpha)^q e^{-r} r^{1-D/2} \frac{d^q}{dr^q}\left[r^{q+D/2-1} e^r\right] \quad \textbf{(A.8)}$$

Finalmente, después de haber expresado la derivada de orden *q*

$$\frac{d^q}{dr^q}\left[r^{q+D/2-1} e^r\right] = \sum_{n=0}^{q} \binom{n}{q} \frac{(q+\frac{D}{2}-1)!}{(n+\frac{D}{2}-1)!} r^{n+D/2-1} \quad \textbf{(A.9)}$$

el momento μ_q de una variable $\alpha\chi_{D,\theta}^2$ viene dado por

$$\mu_q = (2\alpha)^q \sum_{n=0}^{q} \binom{n}{q} \frac{(q+\frac{D}{2}-1)!}{(n+\frac{D}{2}-1)!} \left(\frac{\Theta}{2\alpha}\right)^n \quad \textbf{(A.10)}$$

Esta expresión incluye también el caso en el que $\Theta = 0$ (distribución χ^2 central). La demostración es la misma que la presentada en [10] sustituyendo la distribución de Rice por una distribución de Rayleigh.

APÉNDICE B

MOMENTOS DE LAS *CARACTERÍSTICAS*

En este Apéndice se deriva la expresión del primer y segundo momento de las Características F_1 y F_2. Estas expresiones son necesarias para la descripción del *Espacio Característico*.

Consideremos un conjunto de *N* variables aleatorias independientes e idénticamente distribuidas $(X_1, X_2, ..., X_N)$ con muestras $(x_1, x_2, ..., x_N)$ y el momento q , μ_q, de su variable padre X

$$\mu_q = E\{X^q\} = E\{X_1^q\} = E\{X_2^q\} = ... = E\{X_N^q\} \qquad \textbf{(B.1)}$$

Las estadísticas $m_q = 1/N \sum_{n=1}^{N} x_n^q$ son estimadores imparciales de los momentos relativos μ_q

$$E\{m_q\} = \mu_q \qquad \textbf{(B.2)}$$

con varianza

$$Var\{m_q\} = \frac{(\mu_{2q} - \mu_q^2)}{N} \qquad \textbf{(B.3)}$$

Las anteriores fórmulas llevan a las expresiones del primer y segundo momento de la primera Característica $F_1 = m_1$ de una célula dados los momentos μ_1 y μ_2 de su variable padre

$$E\{F_1\} = \sigma^2(1+pr) \tag{B.4}$$

$$Var\{F_1\} = \frac{\sigma^4}{N}\left(1+2pr+p(1-p)r^2\right) \tag{B.5}$$

La derivación del primer y segundo momento de la segunda Característica F_2 no es directa al ser F_2 la raíz cuadrada del segundo momento empírico de la media $m_2^{'} = 1/N\sum_{n=1}^{N}(x_n - m_1)^2$

La estadística $m_2^{'}$ es un estimador asintóticamente imparcial de la varianza $\mu_2^{'}$

$$E\{m_2^{'}\} = \frac{N-1}{N}\mu_2^{'} \tag{B.6}$$

donde $\mu_q^{'}$ es el momento q de la media de la variable padre $\mu_q^{'} = E\left\{\left(X - E\{X\}\right)^q\right\}$. Su varianza viene dada por

$$Var\{m_2^{'}\} = \frac{N-1}{N^3}\left[(N-1)\mu_4^{'} - (N-3)\mu_2^{'2}\right] \tag{B.7}$$

Como $Var\{m_2^{'}\}$ varía proporcionalmente a $1/N$ y la raíz cuadrada de la derivada existe a lo largo de R_+^*, podemos asumir la siguiente aproximación [16]:

$$Var\left\{\sqrt{m_2^{'}}\right\} \cong \frac{Var\{m_2^{'}\}}{4E\{m_2^{'}\}} \tag{B.8}$$

Sustituir (B.6) y (B.7) en la aproximación anterior, conduce a la expresión de la varianza de la Característica $F_2 = \sqrt{m_2^{'}}$

$$Var\{F_2\} \cong \frac{1}{4N^2}\left[(N-1)\frac{\mu_4^{'}}{\mu_2^{'}} - (N-3)\mu_2^{'}\right] \tag{B.9}$$

Finalmente podemos expresar el valor esperado de F_2, $E\{F_2\} = \left[E\{F_2^2\} - Var\{F_2\}\right]^{1/2}$ sustituyendo la $\mu_q^{'}$ obtenida en (4.12)

$$E\{F_2\} = \sigma^2\left[\frac{\sum_{i,j=0}^{4} a_{ij}p^i r^j}{\sum_{i,j=0}^{2} b_{ij}p^i r^j}\right]^{\frac{1}{2}} \tag{B.10}$$

donde los a_{ij} y b_{ij} distintos de cero son

$$a_{00} = 4N^2 - 12N + 6;\ a_{11} = 16N^2 - 148N + 24$$

$$a_{12} = 8N^2 - 48N + 36;\ a_{22} = 8N^2 + 24N - 36$$

$$a_{13} = -12N + 12\,;\; a_{23} = 16N^2 + 12N - 36$$

$$a_{33} = -16N^2 + 24$$

$$a_{14} = -N + 1\,;\; a_{24} = 4N^2 + N - 7$$

$$a_{34} = -8N^2 + 12\,;\; a_{44} = 4N^2 - 6$$

$$b_{00} = 4N^2\,;\; b_{11} = 8N^2\,;\; b_{12} = 4N^2\,;\; b_{22} = -4N^2\,.$$

Las $\mu_q^{'}$s no dependen de *N*. Las varianzas (B.5) y (B.9) varían entonces como $1/N$ y tienden a cero considerando que *N* tiende a infinito. Las Características F_1 y F_2 son por tanto estimadores consistentes de la media y la desviación estándar de la célula.

APÉNDICE C

"Morphological Filtering of Speech Spectrograms in the Context of Additive Noise"

El siguiente artículo de investigación constituye la Referencia [3], y es el documento aceptado para su presentación y posterior publicación en la conferencia europea que se celebrará en Génova en Septiembre del presente año en el llamado Eurospeech 2003.

La aceptación de esta publicación constituye el primer resultado inmediato de la realización del presente proyecto fin de carrera.

Morphological Filtering of Speech Spectrograms in the Context of Additive Noise

Francisco Romero Rodriguez[1], *Wei M. Liu*[2], *Nicholas W. D. Evans*[2], *John S. D. Mason*[2]

[1]Escuela Superior de Ingenieros, Seville, Spain
franciscororo@hotmail.com
[2]School of Engineering, University of Wales Swansea, UK
w_ming99@yahoo.com, {n.w.d.evans, j.s.d.mason}@swansea.ac.uk

Abstract

A recent approach to signal segmentation in additive noise [1, 2] uses features of small spectrogram sub-units accrued over the full spectrogram. The original work considered chirp signals in additive white Gaussian noise. This paper extends this work first by considering similar signals at different signal-to-noise ratios and then in the context of speech recognition. For the chirp case, a cost function based on spectrogram area is introduced and this indicates that the segmentation process is robust down to and below 0 dB SNR. For the speech experiments the objectives are again to assess the segmentation capabilities of the process. White Gaussian noise is added to clean speech and the segmentation process applied. The cost function now is automatic speech recognition (ASR) accuracy. After segmentation speech areas are set to one constant level and non-speech areas are set to a lower constant level, thereby assessing the segmentation process and the importance of spectral shape in ASR. For the ASR experiments the TIDigits database is used in a standard AURORA 2 configuration, under mis-matched test and training conditions. With 5 dB SNR for the test set only (clean training) a word accuracy of 56% is achieved. This compares with 16% when the same noisy test data is applied directly to the ASR system without segmentation. Thus the segmentation approach shows that spectral shapes alone (without normal spectral amplitude variations) leads to perhaps surprisingly good ASR results in noisy conditions. The next stage is to include amplitude information along with appropriate noise compensation.

1. Introduction

The task of separating speech from noise has proven to be a particularly challenging one over a number of years. Whether the speech is destined for an automatic recognition system or for a person the normal goal of the task is essentially the same and can be summarised as that of extracting a representation of the speech signal that leads to improved recognition. In this context the early work of Boll [3] is generally acknowledged as the forerunner of experimental investigations of many variants under the general heading of spectral subtraction. All involve deriving noise estimates which are then subtracted from the corrupted signal. Typically, these estimates relate to short-term discrete Fourier transform frequency bins and in the early procedures they were derived from non-speech intervals. However, more recently procedures have been examined which derive noise estimates continuously, during speech and non-speech intervals. These include the quantile-based approach of Stahl *et al* [4] and extensions which utilise both local time and local frequency bins [5]. Another recent approach which uses local time and frequency bins to derive noise estimates is the harmonic tunnelling of [6]. Clearly these latter approaches have the potential for deriving better noise estimates since they make use not only of the complete time course but also can provide estimates from within the same instantaneous window that is subjected to compensation, a feature particularly beneficial for distinctly non-stationary noise.

This paper considers another approach which possesses these benefits in that it attempts to separate a signal from noise. However here the process is one of segmentation, classifying regions of a short term spectrogram as either signal or non-signal (noise). The process is based on the statistical properties of the short term spectrogram and morphological filtering. Regions in the spectrogram are identified as either noise or signal and then the signal regions are grown via morphological processing. Very little work has been published on the morphological filtering of speech. The work of Hansen [7] considers morphologically based feature enhancement in the context of noisy speech and Lombard, but little else has followed. The motivation for the work presented here stems from the recent work of Hory *et al* in signal segmentation [1, 2].

The remainder of the paper is structured as follows. Section 2 contains a description of the morphological filtering approach as proposed by Hory *et al* [1, 2] including supporting evidence independently obtained for chirp-based signals extracted from additive white Gaussian (AWG) noise. The basic work is first extended by considering segmentation performance for different signal-to-

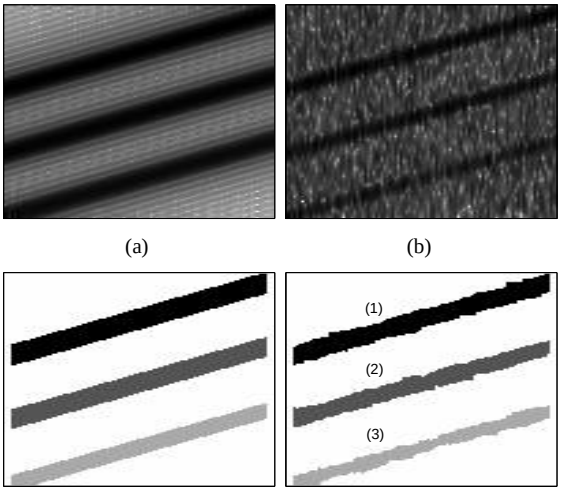


Figure 1: Spectrograms of chirp test signal: (a) without and (b) with AWG noise added at +5 dB (higher chirp), 0 dB (middle chirp), -5 dB (lower chirp), and corresponding segmentation results (c) and (d). Horizontal time axis (0-0.5 seconds), vertical frequency axis (0-4 kHz).

noise ratios, again for the same synthesised chirp signal (Section 3). Then, in Section 4, some initial speech recognition results are presented using the TIDigits database [8], followed by some observations and suggestions for further work.

2. Background to Signal Segmentation

The procedures considered here to segment signals from noise follow closely those of Hory *et al* [1, 2]. Consequently, since the details and background are covered extensively in [1, 2], only the concepts and outlines of the procedures are presented. In conceptual terms it is convenient to consider the spectrogram as an image. Then the segmentation process is based on the statistics of features derived from sub-images and accrued over the full image. The assumption is that the statistics associated with noise differ from those associated with the signal. The sub-image features recommended in [1, 2] are the mean and standard deviation of pixel values (ie power spectra) of each sub-image. Figure 1 (a) and (b) show spectrograms of clean and noisy chirp test signals respectively (as used in [1, 2]). The corresponding spectrograms after segmentation are shown in Figure 1 (c) and (d).

The segmentation procedure begins by computing the local features across the spectrogram and then locating seeds for morphological growth of signal regions. A grid is obtained from estimations of the SNR of the sub-images. The grid is superimposed onto the feature space and determines the seed selection. Examples of seeds are illustrated in Figure 2 (a), which shows a plot of the two features, mean against standard deviation for the 3-chirp

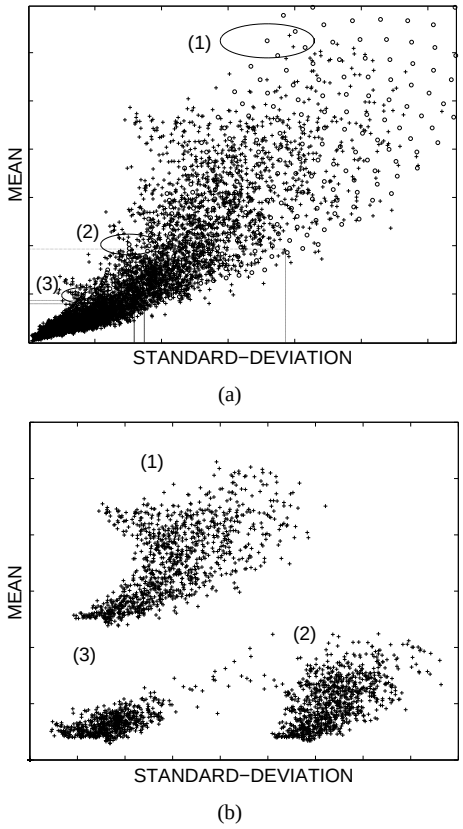


Figure 2: An illustration of seeds (a) for the test signals in Figure 1 (b) and (d), and the resulting segmented feature space (b) showing three distinct regions, one for each chirp.

examples shown in Figure 1 (b) and (d). Seeds are taken, one by one, starting with the highest values of the grid (region 1 in Figure 2 (a)) until reaching a region deemed to be noise. Each seed is grown in the spectrogram to provide signal regions. Subsequently a new grid and a new estimated noise-region are computed from the un-segmented elements. The process is repeated until the normalised maximum likelihood calculated to estimate the noise converges [1, 2]. The segmented feature space is given in Figure 2 (b).

3. Signal to Noise Ratio

The results presented in Figures 1 and 2 corroborate those reported in [1, 2]. In this section the procedures are applied for different levels of signal-to-noise ratio in an attempt to assess the robustness of the process. An objective cost function is defined in the form of segmentation areas. The procedure is applied first to a clean signal leading to the baseline segmentation area. Then the cost function comes from the integration of the image difference signal:

$$\text{Seg. Acc.}|_{\text{SNR}=x\text{dB}} = \frac{\sum_{n,k} \text{Seg}(n,k)|_{x\text{dB}}}{\sum_{n,k} \text{Seg}(n,k)|_{\text{clean}}},$$

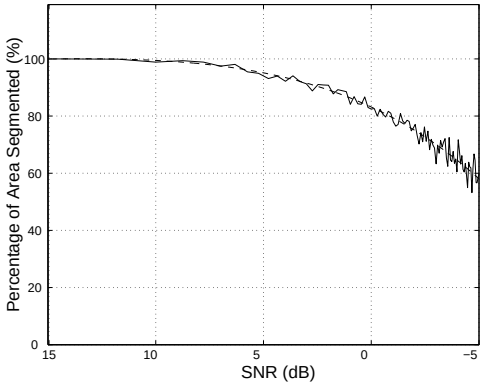


Figure 3: Dependence on SNR: area based segmentation error against SNR for chirp signal.

where the segmentation accuracy at x dB is given by the ratio of segmented areas at x dB and the original clean area, n and k are the dimensions of the spectrogram image. Figure 3 shows values of this cost function for chirp signals with different SNR, decreasing from +15 dB to -5 dB. The associated segmentation error shows a marked increase when the SNR reaches -5 dB, but is reassuringly level for higher SNRs. Example time waveforms, spectrograms and results of segmentation are given in Figure 4. At -5 dB (Figures (b), (d) and (f)) some of the signal area has been lost and the chirp is broken at one point early along the time course; also, there are some small areas wrongly classed as signal mainly at the lower end of the frequency range. The corresponding segmentation cost function in Figure 3 indicates a 40% signal area error for the same conditions.

4. Speech Experiments

The objective of the experiments reported here are twofold. The first is to assess the contribution of speech shapes or structures in the spectrogram to ASR performance and second, to extend the assessment of SNR dependence reported in Section 3, to the speech context.

The experiments are performed on a subset of the TIDigits database [8]. A set of 8440 utterances are used for training and a set of 1001 utterances for testing. The experimental setup is the same as that of the AURORA 2 database [9] except that here AWG noise is added to the test set only at six different noise levels (20, 15, 10, 5, 0, and -5 dB).

In order to investigate the contribution of speech shapes or structures, the segmentation is applied to speech spectrograms as described in Section 2. Following segmentation, the energy at all areas of the spectrogram deemed to be speech are set to one common ceiling value. Similarly, all areas of the spectrogram deemed to be noise are set to one common floor value. An illustration of the results is given in Figure 5 where the spectrogram of an original utterance from the TIDigits database is given in

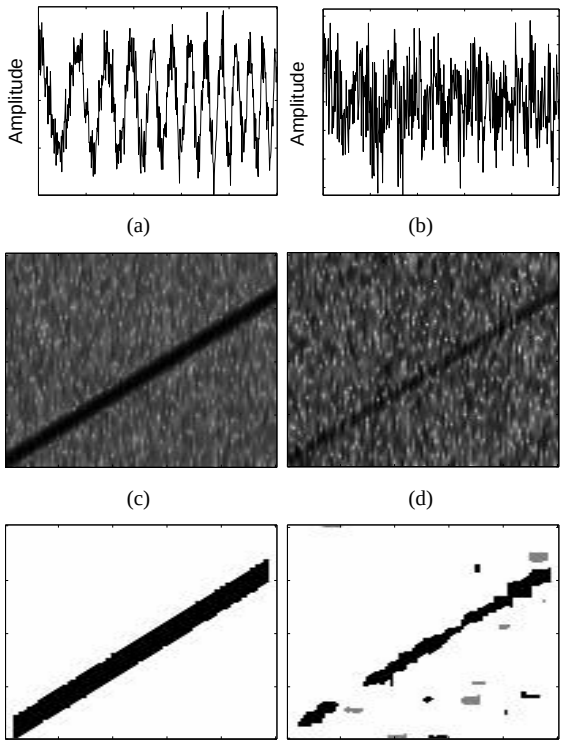


Figure 4: Dependence on SNR: time waveforms at (a) 10 dB and at (b) -5 dB, corresponding spectrograms (c) and (d), and results of segmentation (e) and (f). Horizontal time axis (0-0.5 s), vertical frequency axis (0-4 kHz).

(a), the spectrogram of the same utterance corrupted by AWG noise at 5 dB is given in (b) and the result of segmentation given in (c).

Figure 6 shows ASR results in terms of word accuracy plotted against SNR levels of mismatched noise conditions for the untreated data set and the morphologically filtered set (morph set). In the clean case for both test and training the morph set gives perhaps surprisingly good results at 89%, given that all spectral amplitude information has been replaced by a single level (cf 98% for untreated set). Furthermore the robustness of the shape segmentation is illustrated by the more gradual degradation in the case of the morph set, falling to just 56% at 5 dB (cf 16% for untreated). The next stage is to replace some of the amplitude variation to the segmented speech signal areas.

5. Conclusions

In this paper experimental results relating to a recently proposed process for segmenting signals in the domain of spectrograms [1, 2] are described. The original work considers chirp test signals and experiments reported here use similar signals thereby independently corroborating the original findings. These first results are extended by then

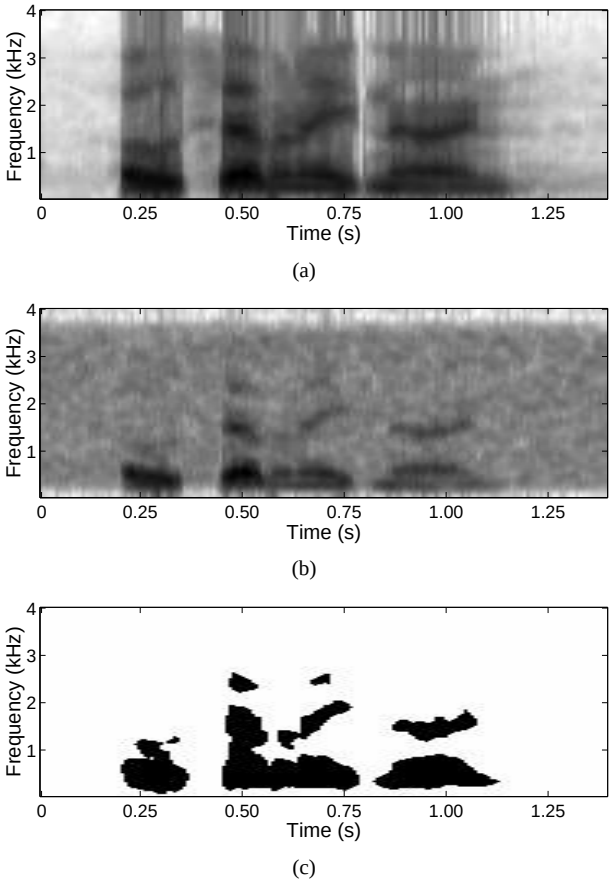


Figure 5: Three spectrograms for (a) original clean speech utterance from the TIDigits database, (b) the same utterance degraded by AWG noise at 5 dB and (c) the result of signal segmentation.

examining segmentation performance at different SNRs. It is shown through a spectrogram area cost function that the segmentation process is robust down to 0 dB and below.

The experimental work is then extended to the speech domain. Areas deemed to be dominated by speech are segmented and set to a constant high amplitude. All other areas are set to a constant low floor level. Clearly this process preserves only the spectral shape of the speech. In so doing it looses local amplitude information of the speech but in doing so it also removes all noise (accept that which has corrupted the spectral shapes of the speech).

For the ASR experiments the TIDigits database is used in a standard AURORA 2 configuration. Under mismatched test and training conditions of 5 dB SNR for the test set only (clean training) a word accuracy of 56% is achieved. This compares with 16% when the same noisy test data is applied directly to the ASR system. In conclusion the approach to signal segmentation applied here to speech, shows that spectral shapes alone (without normal spectral amplitude variations) leads to perhaps surprisingly good ASR results in noisy conditions and provides good noise suppression. The next stage is to include spectral amplitude information along with appropriate noise compensation in the spectrogram regions occupied by speech.

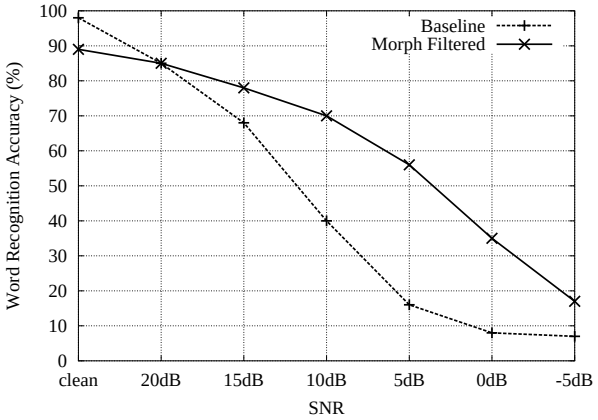


Figure 6: Word recognition accuracy against SNR for the baseline and morphologically filtered sets. The recognisers are trained on clean data and tested against data with noise added at the indicated level.

6. References

[1] Hory, C., Martin, N. and Chehikian, A., "Spectrogram Segmentation by means of Statistical Features for Non-stationary Signal Interpretation", IEEE Trans. on Signal Processing, 50:2915–2925, 2002.

[2] Hory, C. and Martin, N., "Maximum Likelihood Noise Estimation for Spectrogram Segmentation Control", Proc. ICASSP, Vol. 2, 2002, 1581–1584.

[3] Boll, S. F., "Suppression of Acoustic Noise in Speech using Spectral Subtraction", IEEE Trans. on Acoustics Speech and Signal Processing, 27(2):113–120, 1979.

[4] Stahl, V., Fischer, A. and Bippus, R., "Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering", Proc. ICASSP, Vol. 3, 2000, 1875–1878.

[5] Evans, N. W. D. and Mason, J. S., "Time-Frequency Quantile-Based Noise Estimation", Proc. EUSIPCO, Vol. 1, 2002, 539–542.

[6] Ealey, D., Kelleher, H. and Pearce, D., "Harmonic Tunnelling: Tracking Non-stationary Noises During Speech", Proc. Eurospeech, Vol. 1, 2001, 437–450.

[7] Hansen, J. H. L., "Morphological Constrained Feature Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect", IEEE Trans. on Speech and Audio Processing, 2(4):598-614, 1994.

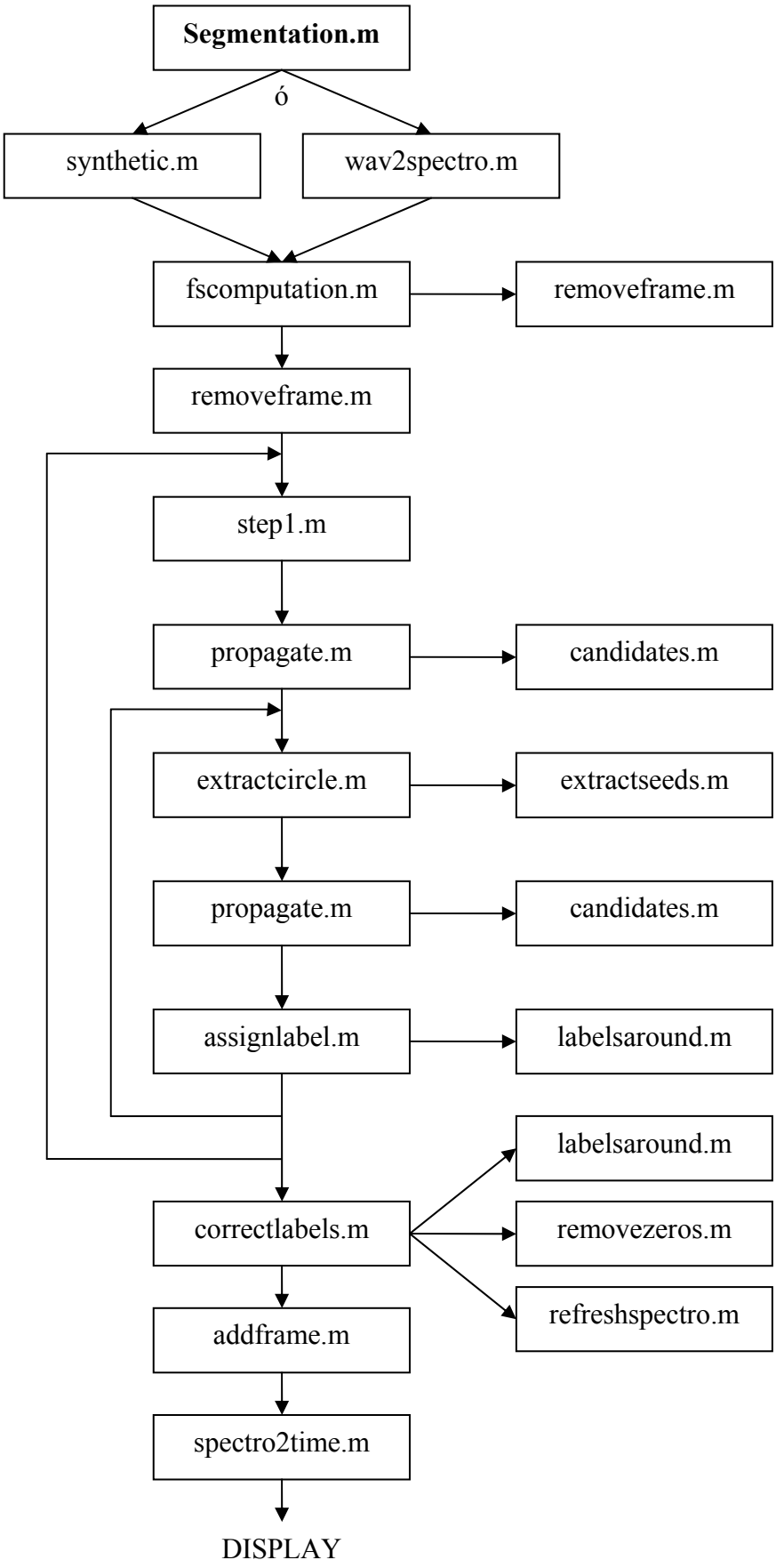
[8] Leonard, R. G., "A database for Speaker Independent Digit Recognition", Proc. ICASSP, Vol. 3, 1984, 42.11–14.

[9] Hirsch, H. G. and Pearce, D., "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions", ISCA ITRW ASR2000 'Automatic Speech Recognition: Challenges for the next Millenium', 2000.

APÉNDICE D

DIAGRAMA DE BLOQUES DEL ALGORITMO DE SEGMENTACIÓN DE ESPECTROGRAMAS

A continuación se muestra el diagrama de bloques de la función principal del algoritmo de segmentación de espectrogramas, el cual fue realizado en MATLAB. Las siguientes funciones son por tanto las más representativas del proceso:


Segmentation.m
ó
synthetic.m
wav2spectro.m
fscomputation.m
removeframe.m
removeframe.m
step1.m
propagate.m
candidates.m
extractcircle.m
extractseeds.m
propagate.m
candidates.m
assignlabel.m
labelsaround.m
labelsaround.m
correctlabels.m
removezeros.m
refreshspectro.m
addframe.m
spectro2time.m
DISPLAY

REFERENCIAS

[1] C. Hory, N. Martin and A. Chehikian, “Spectrogram Segmentation by Means of Statistical Features for Non-stationary Signal Interpretation”, *IEEE Trans. on Signal Processing*, 50:2915-2925, 2002.

[2] C. Hory and N. Martin, “Maximum Likelihood Noise Estimation for Spectrogram Segmentation Control”, in *Proc. ICASSP*, Orlando, FL, Vol. 2, 2002, pp. 1581-1584.

[3] F. Romero, W. Liu, N. Evans, J. Mason, “Morphological Filtering of Speech Spectrograms in the Context of Additive Noise”. Aceptado para Eurospeech 2003.

[4] L. H. Koopsmans, *The spectral analysis of time series*, Academic press, 1974.

[5] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, vol. II, Wiley and sons, second edition, 1995.

[6] P.D. Welch, “A direct digital method of power spectrum estimation”, *IBM J. res. And dev.*, vol 5, no. 2, pp. 141-156, April 1961.

[7] Boll, S. F., “Suppression of Acoustic Noise in Speech using Spectral Subtraction”, *IEEE Trans. on Acoustics Speech and Signal Processing*, 27(2):13-120,1979.

[8] Leonard, R. G., “A database for Speaker Independent Digit Recognition”, Proc. ICASSP, Vol. 3, 1984, 42.11-14.

[9] Hirsch, H. G. and Pearce, D., “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions”, ISCA ITRW ASR2000 ‘Automatic Speech Recognition: Challenges for the next Millenium’, 2000.

[10] J. G. Proakis, *Digital Communications*. New York: Mc Graw Hill, 1995.

[11] L. Rabiner, B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[12] A. Webb, *Statistical Pattern Recognition*. New York: Arnold, 1999.

[13] T. Parsons, *Voice and Speech Processing*, McGraw Hill, 1987.

[14] M. Cooke, S. Beet, M. Crawford, *Visual Representations of Speech Signal*, Wiley, 1993.

[15] A.k. Jain, "Fundamentals of digital image processing". Prentice Hall, 1989.

[16] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*. London, U.K.:Charles Griffin, 1963, vol.2.

[17] R. von Mises, *Mathematical Theory of Probability and Statistics*. New York: Academic, 1964.

[18] B. Leprettre and N. Martin, "Extraction of pertinent subsets from time-frequency representations for detection and recognition purposes," *Signal Process*., vol. 82, no. 2, pp. 229-238, Feb. 2002.

[19] Evans, N. W. D. and Mason, J. S., "Time-Frequency Quantile-Based Noise Estimation", Proc. *EUSIPCO*, Vol. 1, 2002, 539-542.

[20] C. Hory, N. Martin, A. Chehikian, and L. E. Solberg, "Time-frequency space characterization based on statistical criterions," in Proc. *EUSIPCO*, Tampere, Finland, Sept. 4-8, 2000, pp. 214-217.

[21] C-H. Lee, F. K. Soong and K. K. Paliwal. *Automatic Speech and Speaker Recognition: Advanced Topics.* Kluwer Academic Publishers, 1997.

[22] Eric Keller, *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State-of-the-Art and Future Challenges*. John Wiley & Sons, 1994.