En este capítulo se realiza una introducción sobre todos los fundamentos de tratamiento de señal de voz, especialmente en lo referente al contexto de la mejora de voz para el reconocimiento, necesarios para la correcta interpretación del resto del proyecto. Así en primer lugar se exponen las bases del análisis de la señal vocal tanto en tiempo como en frecuencia, introduciéndose el espectrograma que será la representación que utilizarán todos los algoritmos propuestos es el presente trabajo. Tras realizar una introducción sobre el ruido que corrompe las señales vocales y los objetivos que persigue la eliminación de dicho ruido, se pasa a explicar más en detalle las bases del objetivo perseguido en este proyecto que es el reconocimiento automático. Posteriormente se pasa a explicar la base de datos utilizada para realizar los diferentes experimentos que se llevan a cabo y el reconocedor concreto utilizado. Finalmente se hace una breve introducción sobre los dos algoritmos de mejora de la señal vocal que se estudian en este proyecto y que se explicarán al detalle en el siguiente Capítulo.

Capítulo 2

FUNDAMENTOS DEL PROCRESAMIENTO DE VOZ Y RECONOCIMIENTO AUTOMÁTICO

2.1 LA SEÑAL VOCAL

La voz es una onda de presión que se transmite, como cualquier otro sonido, haciendo vibrar las moléculas de aire. En su generación y recepción por el ser humano existen varios procesos y órganos implicados. La comprensión de estos mecanismos se hace imprescindible para poder tratar esa onda como una señal, y llevar a cabo procesos tales como el reconocimiento de voz, síntesis, compresión, mejora de calidad, etc...

En un esquema general de una comunicación vocal existen tres procesos:

- *proceso lingüístico*: el cerebro del emisor tiene una idea y decide qué se va a decir; y el del receptor decide qué entiende.
- *proceso fisiológico*: en emisión los músculos se combinan de una cierta forma y en una cierta secuencia para generar la onda de presión; y en recepción los órganos son excitados de una cierta forma y secuencia por la onda de presión
- *proceso acústico*: la onda se propaga por el medio que normalmente es el aire

2.1.1 Producción de la voz

En el proceso de generación de la voz, el aire sale de los pulmones hacia la laringe en donde, dependiendo de si el sonido a generar es sonoro o sordo, se produce una vibración de las cuerdas vocales (aportando una oscilación a la frecuencia fundamental) o no. Este flujo de aire pasa entonces a la faringe, la cavidad bucal y posiblemente la nasal donde, dependiendo de la posición de labios, lengua, dientes, etc... el sonido es modulado de una u otra forma para dar lugar a las distintas vocales y consonantes. El resultado es una señal que puede ser descrita como una mezcla de ondas de diferente frecuencia con diferentes amplitudes. El proceso de generación de dicha onda puede ser modelado como un sistema con dos tipos diferentes de excitación que pasan por un filtro que añade las resonancias del tracto vocal (formantes). Estas resonancias son los picos de la envolvente del espectro de la señal de voz y son muy importantes para reconocer o sintetizar voz (la posición de estas resonancias es el discriminante entre los distintos fonos)

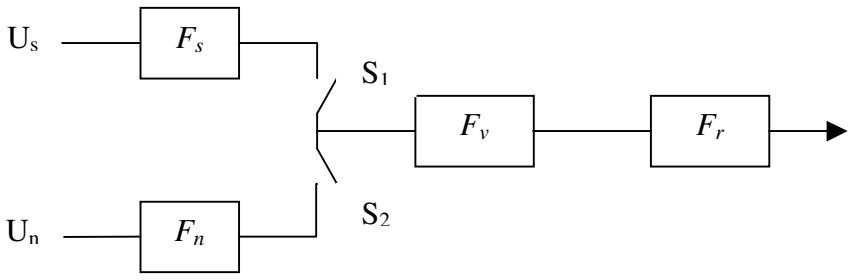

Figura 2.1 Modelo de producción de la voz

La Figura 2.1 muestra un modelo simplificado del Tracto Vocal. En ella, U_s es la excitación sonora, U_n la excitación sorda (fuente de ruido); F_s y F_n son filtros de forma. Los interruptores S_1 y S_2 permiten la selección de una u otra fuente de excitación. F_v es el Tracto Vocal, y F_r se encarga de la impedancia de radiación.

2.1.2 Características de la señal vocal

Si la voz es grabada a través de un micrófono y la señal resultante se muestrea a 8KHz para un posterior tratamiento digital el resultado es una onda quasi-estacionaria en intervalos cortos de tiempo, que se corresponden con el tiempo empleado para pronunciar un determinado fono. Estos intervalos son de unos 10-30 ms que se corresponden con 80-240 muestras de la señal digital.

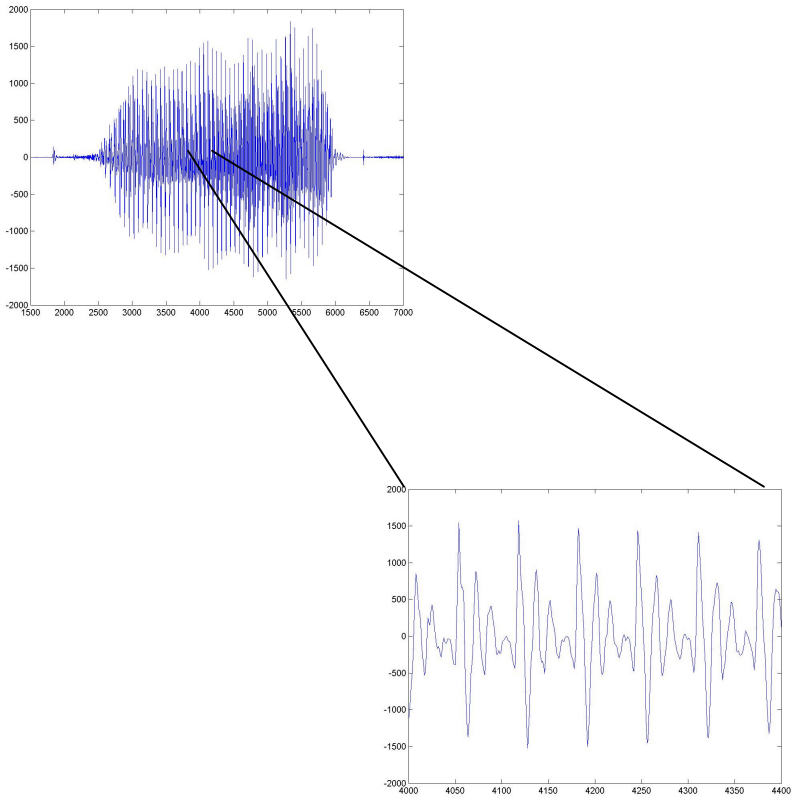

Figura 2.2 Señal vocal en tiempo y zoom mostrando la propiedad estacionaria local

Por ello, para realizar cualquier estudio de la señal vocal es necesario un análisis localizado por tramas; que deberán ser lo suficientemente cortas para abarcar situaciones estacionarias y lo suficientemente largas para hacer un análisis fiable así como minimizar el número de tramas a analizar.

2.1.3 Espectrograma

El espectrograma es una de las representaciones más utilizadas en el análisis de las señales vocales, y es la que se utiliza en este proyecto en los diferentes algoritmos propuestos. Se trata de una representación Tiempo-Frecuencia en la que se puede ver la evolución temporal del contenido armónico de la señal. El espectrograma de una señal en cada instante n se deriva de la siguiente manera:

$$S_n(e^{jw}) = \sum_{m=-\infty}^{\infty} s_n[m] \cdot e^{-jwm}$$

donde

$$s_n[m] = s[m] \cdot w[n-m]$$

Como se ha comentado anteriormente, cualquier análisis de la señal vocal debe ser un análisis localizado por tramas y el espectrograma no es una excepción. Para derivar esta representación se toma una ventana de la señal temporal $s_n[m]$ y se calcula la DFT $S_n(e^{jw})$, obteniendo el espectro de ese trozo de voz. A continuación se toma la magnitud de dicho espectro y se representa de manera vertical utilizando colores (o niveles de gris) para los valores de dicha magnitud. La ventana temporal la vamos deslizando y vamos obteniendo para cada posición temporal el contenido armónico de la señal, formando columna tras columna la representación que llamamos espectrograma. La figura 2.3 muestra gráficamente dicho proceso.

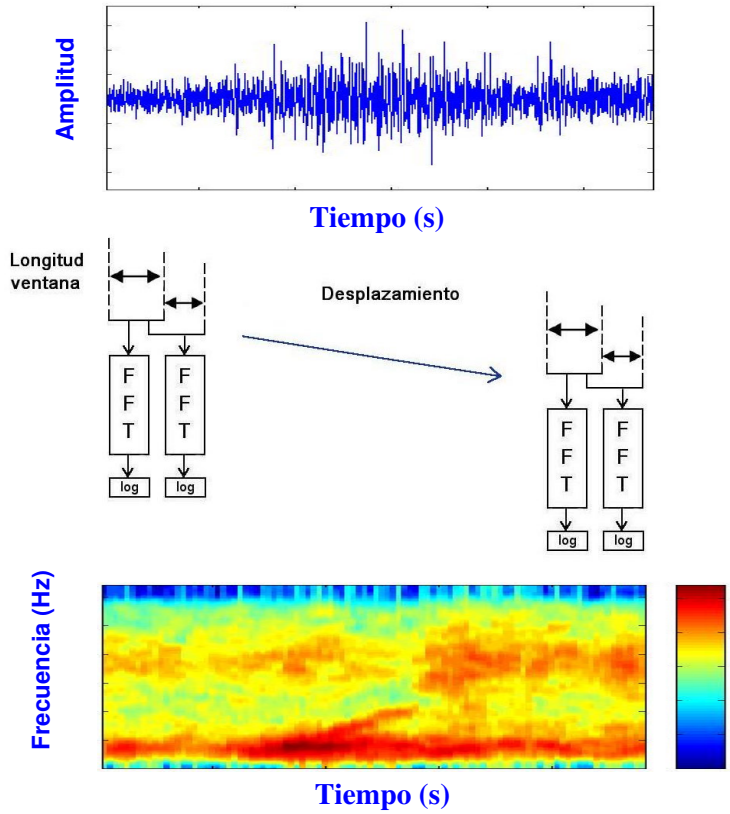

Figura 2.3 Proceso de generación del espectrograma

Existen varios parámetros implicados en la obtención del espectrograma que, dependiendo de los valores que tomen, darán unas características determinadas a la representación. Los más importantes son la forma en la que enventanamos la señal temporal, ya que dará lugar una determinada resolución en tiempo y frecuencia y como consecuencia a diferentes tipos de espectrograma. Con una longitud de ventana corta en tiempo tendremos buena resolución temporal (se detectarán cambios rápidos) pero no distinguiremos lóbulos que sean estrechos en el contenido espectral (Espectrograma de Banda Ancha). En el caso contrario, es decir utilizando una ventana larga en tiempo, obtendremos mucha precisión en el contenido espectral, pero estaremos evaluando comportamientos temporales medios (Espectrograma de Banda Estrecha). Además también afectará en la resolución frecuencial el tipo de ventana temporal utilizada: Rectangular, Hamming, Hanning, etc. Los otros parámetros implicados son el número de puntos utilizados para calcular la DFT y el solapamiento temporal entre las ventanas, que creará un representación más o menos redundante para tener una visión más o menos suavizada de los cambios que se producen a lo largo del eje temporal.

El resultado es una representación donde se puede distinguir con suma facilidad cómo va evolucionando el contenido espectral de una señal vocal, y donde se pueden diferenciar claramente (en condiciones de señal a ruido razonables) los patrones espectrales pertenecientes a la señal vocal. Mediante sencillas mediaciones, además, podemos identificar por ejemplo la frecuencia fundamental de la señal vocal o los formantes de los distintos fonos que se van pronunciando.

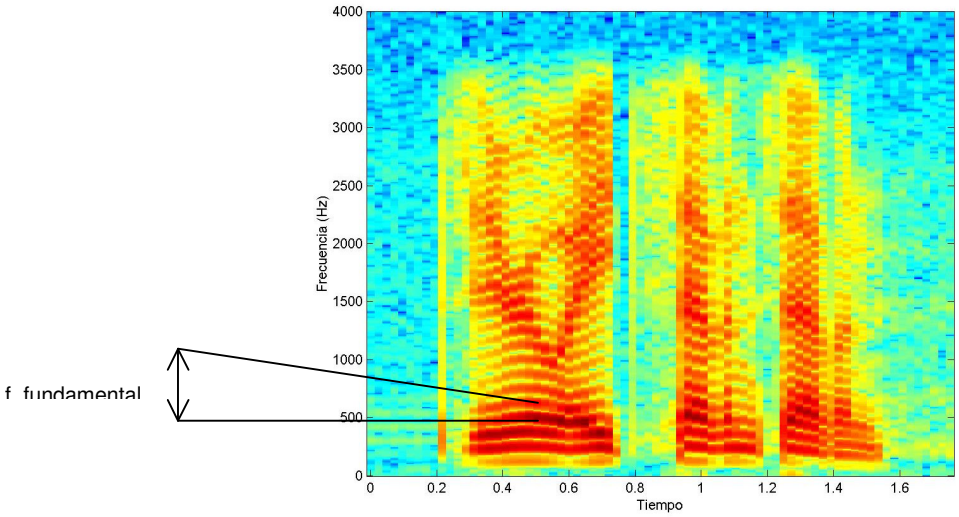

Figura 2.4 Espectrograma de banda estrecha

Como se ha dicho anteriormente, la frecuencia fundamental es la velocidad de oscilación de las cuerdas vocales en los sonidos sonoros. Esta oscilación va variando ligeramente a lo largo del discurso para dar la entonación a las palabras, pero en media tiene una frecuencia de unos 125Hz en el hombre y de unos 250Hz en la mujer. Esta periodicidad se muestra en el espectrograma de banda estrecha como un conjunto de bandas horizontales separadas exactamente la distancia equivalente a la frecuencia en cuestión, como se puede comprobar en la figura 2.4

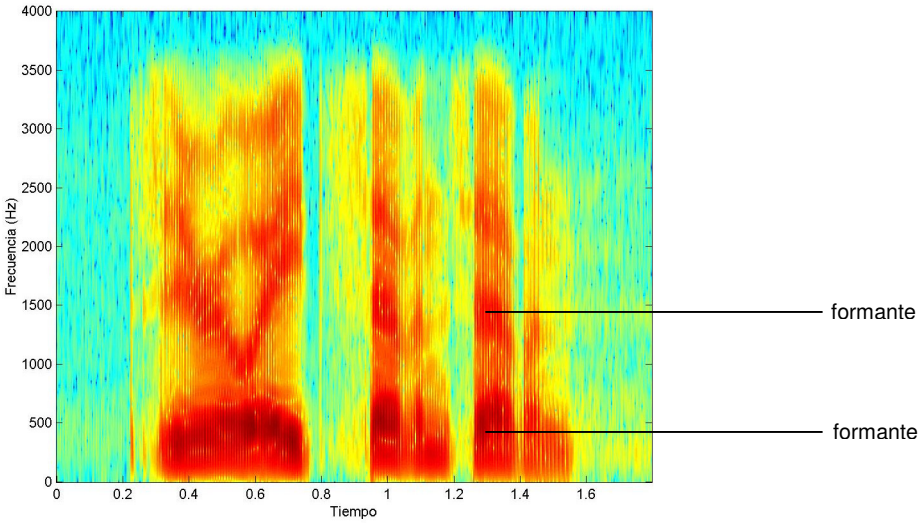

Figura 2.5 Espectrograma de banda ancha

Para distinguir los formantes es más útil el espectrograma de banda ancha, ya que la menor resolución en frecuencia hace que no sea posible capturar la frecuencia fundamental, sino sólo lóbulos frecuenciales más anchos correspondientes a las resonancias del tracto vocal. En el espectrograma aparecen como bandas anchas horizontales de energía que van cambiando a lo largo del eje temporal dependiendo del fono que se este pronunciando en ese momento concreto. De esta forma, viendo la posición de, principalmente, los cuatro primeros formantes se puede deducir el fono al que corresponde; y por ello esa es la información que se extrae en los reconocedores de voz para identificar los sonidos y con ello las palabras pronunciadas.

El Espectrograma es completamente reversible, ya que esta basado en una herramienta reversible: la DFT . Cada columna del espectrograma puede combinarse con la fase (que debió ser almacenada en el proceso de generación) y volver al dominio temporal mediante la IDFT, formando así una trama de señal temporal. Llevando a cabo este proceso columna tras columna y tratando las tramas temporales para tener en cuenta el solapamiento que se llevó a cabo en la generación, podemos regenerar de forma exacta la señal temporal original. Esta propiedad de reversibilidad es imprescindible para los algoritmos propuestos en este proyecto ya que se basan en un tratamiento de la imagen espectrograma para después volver a un señal temporal mejor que la señal original. El proceso ideal que persiguen los diferentes algoritmos se muestra en la figura 2.6. La señal temporal degradada se pasa al dominio frecuencial (a) para ser procesada por un algoritmo

que extraiga sólo los patrones espectrales pertenecientes a la señal vocal (b) ; y una vez aquí volver al dominio temporal (c) dando como resultado una señal limpia.

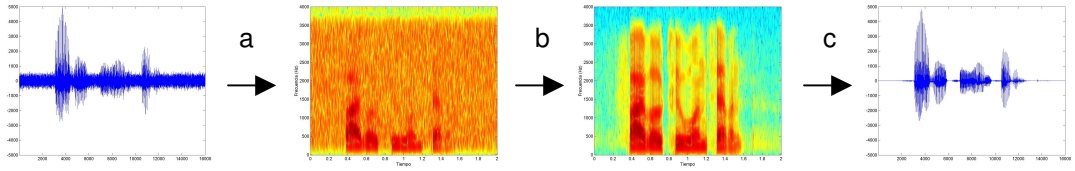

Figura 2.6 Proceso ideal de extracción de señal a través del espectrograma

2.2 RUIDO

El ruido afecta a la señal vocal de diferentes maneras dependiendo de la naturaleza de éste y de los procesos implicados en el registro de ambos: señal vocal y ruido. Así, el tipo de ruido es un factor decisivo a la hora de elegir un determinado sistema de mejora de la señal vocal, haciendo necesaria una caracterización y modelado de la señal de ruido para, por una parte diseñar adecuadamente el sistema de mejora y por otra parte evaluar cómo se comporta dicho sistema ante ruidos de diferentes características. Los diferentes tipos de ruido tienen diferentes propiedades estadísticas, espectrales o espaciales. Las posibles características pueden resumirse como muestra la tabla 2.1

Propiedad	Tipos
Estructura	Continuo / Impulsivo / Periódico
Tipo de interacción	Aditivo / Multiplicativo / Convolucional
Comportamiento temporal	Estacionario / No estacionario
Rango frecuencial	Banda ancha / Banda estrecha
Dependencia con la señal	Correlado / Incorrelado
Propiedades estadísticas	Dependiente / Independiente
Propiedades espaciales	Coherente / Incoherente

Tabla 2.1 Clasificación del ruido basada en varias propiedades

Basándose en la naturaleza y propiedades de las fuentes de ruido, éste puede clasificarse de las siguientes maneras:

1. *Ruido Gaussiano blanco (WGN)*: se trata del ruido más genérico, compuesto por muestras aleatorias cuyas amplitudes se distribuyen como una campana de gauss. Es un ruido aditivo, plano en frecuencia, estacionario e incorrelado.
2. *Ruido de fondo*: se trata de ruido aditivo, que normalmente no esta correlado con la señal y que esta presente en numerosos escenarios. Como ejemplos podemos citar: ruido de coche, oficinas, calles, ventiladores, Hoth, entorno industrial, acondicionadores, helicóptero, etc. De estos ruidos, el llamado Hoth (ruido blanco filtrado) es estacionario, el de calles y fábricas tiene características más dinámicas. El ruido en fábricas y en helicópteros tiene componentes fuertemente periódicas; y el ruido de ventiladores y el de coche son ruidos reales ejemplo de no estacionariedad, teniendo características variables con el tiempo.
3. *Otros hablantes*: es ruido aditivo, compuesto por uno o más hablantes 'competidores'. Suele llamarse 'babble' y es el ruido que podemos encontrar en una cafetería por ejemplo. Tiene características y rango frecuencial muy similares a la señal de interés por lo que es difícil de tratar.
4. *Ruido impulsivo*: Se trata de ruidos de corta duración como el sonido producido por un objeto al caer o al cerrar violentamente una puerta.
5. *Ruido no aditivo*: Puede ser debido a no linealidades de los micrófonos o a distorsión de canal (señales vocales en líneas de transmisión)
6. *Ruido no aditivo debido al estrés del hablante*: se denomina efecto Lombard y es el efecto que tiene sobre el hablante un entorno ruidoso. En presencia de ruido el hablante tiene tendencia a incrementar el esfuerzo vocal haciendo que la voz tenga diferentes características espectrales.
7. *Ruido correlado con la señal*: reverberaciones y ecos
8. *Ruido convolucional*: resultado de una convolución de la señal vocal en el dominio temporal. Por ejemplo cambios en la señal vocal debido a cambios en al acústica del entorno o micrófonos.
9. *Ruido multiplicativo*: se trata de una distorsión producida por el desvanecimiento en los canales de comunicación.

En general, es más difícil tratar con ruido no estacionario, donde no existe conocimiento a priori sobre sus características. Debido a que este ruido varía con el tiempo, los métodos que tratan de estimar el ruido en intervalos iniciales asumiendo no existencia de señal vocal, son del todo inapropiados. Por otra parte, los tipos de ruido que tienen características similares a la señal vocal en tiempo o frecuencia son los más difíciles de eliminar o atenuar. El ruido 'babble' por ejemplo tiene características muy similares y es particularmente difícil para un algoritmo aislar la señal vocal de este ruido.

2.3 APLICACIONES Y OBJETIVOS DE LA MEJORA DE VOZ

Las señales vocales, cuando son grabadas para su tratamiento, transmisión, etc. siempre son afectadas por distintos tipos de ruidos dependiendo del entorno y del proceso de grabación. El problema de mejorar el comportamiento de los sistemas de comunicación vocal en entornos ruidosos ha sido un desafío para los investigadores desde su creación. Los algoritmos que tratan de reducir el ruido presente y mejorar por tanto la señal vocal tienen aplicación en multitud de entornos entre los que se incluyen [14]:

1. Sistemas de telefonía móvil, que sufren tanto de todo tipo de ruidos de fondo como de ruido de canal.
2. Telefonía de manos libres bajo ruido presente en el interior de un coche, etc.
3. Comunicaciones tierra-aire donde la voz de los pilotos se ve degradada por el ruido presente en la cabina y de los motores.
4. Comunicaciones de larga distancia a través de canales de radio ruidosos
5. Sistemas de tele-conferencia donde una fuente de ruido puede ser recogida y transmitida a los receptores
6. Sistemas en los implantes para mejorar la audición en entornos ruidosos (aulas, cafeterías, etc...)

Además, los esfuerzos por conseguir un mejor calidad y/o inteligibilidad de las señales vocales pueden servir para mejorar el comportamiento de otras aplicaciones en el contexto del procesamiento de voz como pueden ser la codificación, compresión o el reconocimiento automático. Así, los sistemas de mejora de la señal vocal tienen en resumen tres principales objetivos:

1. Mejorar la calidad y la inteligibilidad de las señales vocales degradadas por ruido de fondo, reduciendo así la posible fatiga del receptor.
2. Hacer los sistemas de codificación robustos ante el ruido
3. Hacer mejorar el comportamiento de los sistemas de reconocimiento vocal

2.4 RECONOCEDORES AUTOMÁTICOS DE VOZ

La investigación en reconocimiento automático del habla (ASR: Automatic Speech Recognition) es relativamente reciente, siendo los años 40 los testigos de los primeros intentos en este campo. El primer sistema comercial fue presentado en 1973 y sólo podía reconocer 100 palabras diferentes sin posibilidad de aprendizaje posterior. La investigación y desarrollo en este campo ha recorrido un largo camino desde entonces, gracias en gran medida al gran incremento de la capacidad computacional de los ordenadores de hoy día. Pero el sistema auditivo humano todavía es muy superior al mejor de los sistemas de reconocimiento automático disponible actualmente.

Los sistemas de reconocimiento automático del habla han sido desarrollados debido a las ventajas que ofrecen:

- Dan una clara ventaja fisiológica a la hora de comunicarse con cualquier tipo de máquina, ya que la comunicación se lleva a cabo en las condiciones naturales para los humanos.
- Las manos y visión pueden estar ocupadas haciendo otras tareas mientras nos comunicamos con la computadora.

En la práctica, para alcanzar resultados en el reconocimiento de palabra aceptables se introducen ciertas limitaciones como son la posibilidad de uso de un vocabulario y una gramática limitadas, y que no más de un usuario pueda utilizar el sistema al mismo tiempo.

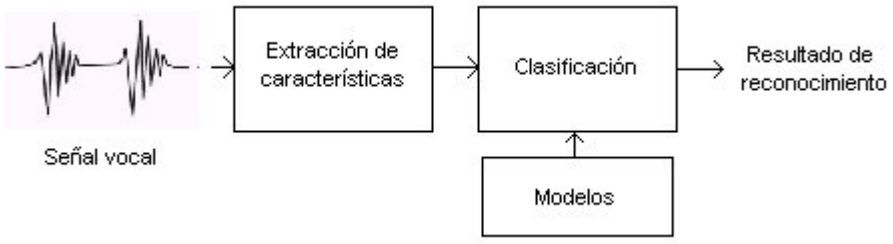

Figura 2.7 Típico sistema de reconocimiento vocal

2.4.1 Fases en la creación de un Reconocedor de voz

En todo sistema de reconocimiento vocal tenemos estas dos etapas claramente diferenciadas:

- *Entrenamiento*: En esta fase se toma un conjunto de muestras vocales etiquetadas, se extraen un conjunto de características de cada muestra vocal y con ellas y las etiquetas se entrena el reconocedor creando una serie de modelos. Estos modelos servirán para que después el reconocedor sea capaz de relacionar muestras de voz similares con las mismas etiquetas que informan de qué palabra o fonema concreto se trata.

- *Test* : Es la fase en la que se usa el reconocedor para determinar qué palabra se esta pronunciando. Para cada muestra de voz se extraen un conjunto de características que se comparan con los modelos creados en el entrenamiento de una u otra forma según el algoritmo que utilice el reconocedor. El modelo más 'cercano' identifica el fonema o palabra que se esta pronunciando a la entrada.

2.4.2 Extracción de características

El primer paso en el reconocimiento es la extracción de características. La señal de voz es segmentada en tramas y para cada trama se calcula un vector de características que consiste en una serie de parámetros que representan las características de la señal vocal. Este proceso se lleva a cabo por las siguientes razones:

- Reducción de la cantidad de datos
- Si se eligen adecuadamente, los parámetros pueden de alguna forma suprimir las características diferenciadoras de dialectos, cierto ruido de fondo y de canal.
- Se pueden enfatizar ciertas características que son útiles a la hora de separar palabras y fonemas.

2.4.3 Clasificación

La clasificación se puede llevar a cabo desde dos puntos de vista diferentes. El primero de ellos y más simple es la comparación con una secuencia de referencia. Aquí, cada palabra o fonema en el vocabulario tiene una secuencia de referencia con la que se compara el vector de características.

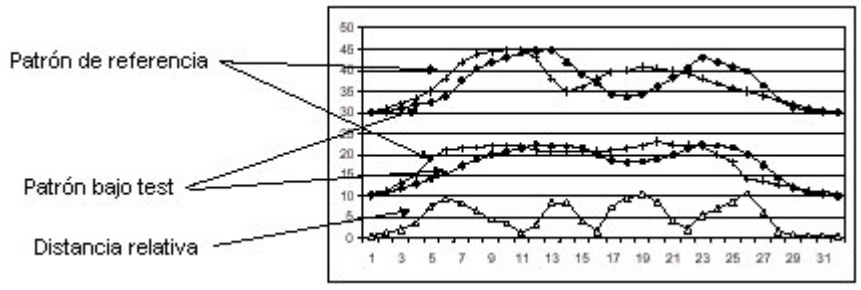

Figura 2.8 Comparación entre el vector de características bajo estudio y el patrón de referencia

El patrón de referencia que mejor se ajuste al vector de características será el que se seleccione. Si el que mejor se ajusta no es lo suficientemente bueno, el reconocedor puede decidir rechazarlo.

El segundo enfoque y más usado se basa en modelos estadísticos. Cada fonema o palabra tiene su propio modelo estadístico y se intentan ajustar a los modelos de referencia creados anteriormente. El método estadístico más utilizado es el de los Modelos Ocultos de Markov (HMM Hidden Markov Models)

2.4.4 Modelos ocultos de Markov

Este método tiene una serie de ventajas:

- Tienen una estructura matemática muy simple
- Los modelos pueden ser adaptados automáticamente por los datos de entrenamiento
- La estructura puede describir secuencias de patrones muy complicadas

En este método la señal de voz se considera un proceso de Markov. Esto significa que la muestra de voz grabada se ve como el resultado de un proceso de generación que consiste en una secuencia de estados. El estado actual sólo depende del estado previo, es decir, el proceso no tiene memoria. La probabilidad de cambiar de estado se denota como probabilidad de transición (ver figura 2.9) Un inconveniente es que la secuencia de estados no puede ser observada directamente, sino que solo pueden hacerse observaciones que tengan relación estadística con los modelos de estados. La probabilidad de hacer una cierta observación en un estado específico se llama probabilidad de observación. Entonces, la probabilidad de que un cierto modelo haya generado la cierta secuencia observada puede ser calculada. Un HMM es definido de forma exacta mediante estos tres parámetros:

p_0 = vector de distribución de las probabilidades de estados iniciales, con elementos: $p_0(j) = P(S_1 = j)$

A = Matriz de transición de estados, con elementos: $a_{ij} = P(S_t = j / S_{t-1} = i)$

$B =$ Funciones de densidad de probabilidad de las observaciones condicionadas: $b_j(x) = f_{X_t / S_t}(x / j)$

Donde S es el estado, t el índice temporal, i y j los índices de los estados y x la secuencia de salida observable.

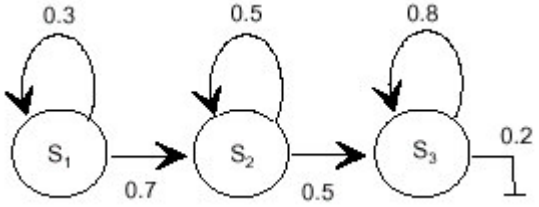

Figura 2.9 Un proceso de Markov simple. Los círculos son estados y las flechas indican transiciones entre los estados con la probabilidad asociada

El entrenamiento de los HMM se hace normalmente con un método recursivo llamado re-estimación de Baum-Welch. Las probabilidades de observación se describen normalmente mediante funciones de distribución gaussianas. Es común usar una suma de varias gaussianas ponderadas, con diferentes medias y diferentes varianzas (ver figura 2.10). Todo esto es utilizado por el paquete Aurora que se describe en el apartado 2.5.

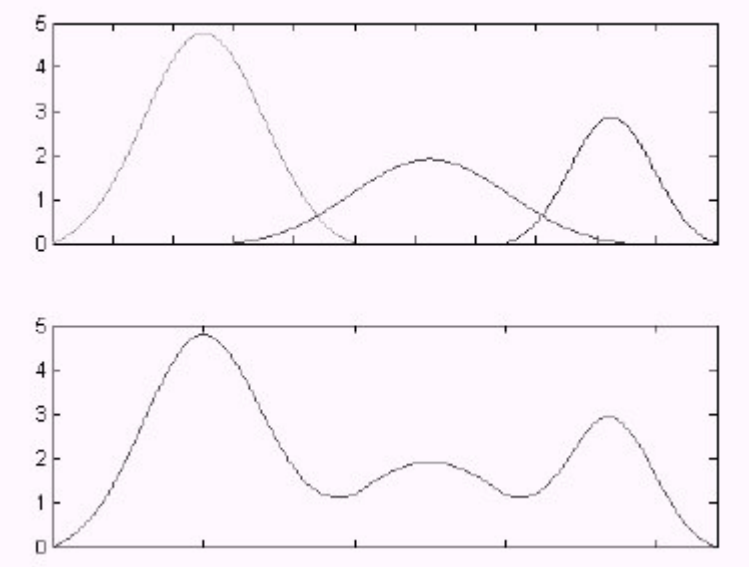

Figura 2.10 Tres funciones gaussianas ponderadas para formar una función de densidad de probabilidad

2.4.5 Dificultades en el reconocimiento

El reconocimiento del habla por parte de los ordenadores es algo totalmente establecido. Existen multitud de sistemas disponibles en el mercado con un altísimo porcentaje de exactitud en el reconocimiento vocal (alrededor del 97%). Pero estos porcentajes son alcanzados en entornos controlados, debiendo el hablante estar en un ambiente sin ruido de fondo y con micrófonos cercanos a los labios. Los profesionales que deben dictar grandes cantidades de texto tales como médicos o abogados deben soportar estas restricciones para aprovechar el considerable ahorro de tiempo y esfuerzo que estos sistemas ofrecen. Sin embargo, un sistema con buen comportamiento en entornos más genéricos, como pueden ser señales vocales grabadas en entornos ruidosos, o por un micrófono de sobremesa como el utilizado en las conferencias, discursos, etc... esta todavía por desarrollar.

Entre los problemas que degradan el comportamiento de los reconocedores y que no permiten su uso en condiciones genéricas podemos citar por ejemplo las variaciones de las características vocales de un hablante a otro, la frecuente ambigüedad de las variables acústicas que no siempre pueden ser directamente mapeadas a variables fonéticas, las variaciones de la voz individual que depende de estados de animo, etc; pero sobre todo debemos citar los ruidos e interferencias que afectan a la señal vocal en distintos ambientes.

El oído humano es fuertemente robusto en toda clase de condiciones de ruido o interferencias. La capacidad del cerebro de procesar varias señales acústicas al mismo tiempo y de tratar grupos de frecuencias más que frecuencias individuales hace que se pueda distinguir una conversación en condiciones de señal a ruido realmente bajas. El mejor de los sistemas de reconocimiento vocal esta todavía muy lejos del oído y cerebro humano, ya que su comportamiento se ve seriamente afectado por la presencia de ruido interferente. De esta manera, el ruido de fondo es hoy día uno de los principales problemas en el campo de los reconocedores de voz, y el creciente número de aplicaciones basadas en reconocimiento vocal hacen que este problema tenga cada vez más importancia.

En este contexto las consecuencias del ruido de fondo son:

- Contaminación directa de las características espectrales en las cuales se basan los reconocedores de voz: estimaciones de parámetros incorrecta, dificultad en la identificación del comienzo y final de las palabras, etc.

- Cambios en la forma de hablar de las personas inmersas en un ambiente ruidoso (efecto Lombard) En presencia de ruido el hablante tiene tendencia a incrementar el esfuerzo vocal haciendo que la voz tenga diferentes características espectrales.

2.5 AURORA 2

Aurora 2.0 es una base de datos de señales de voz limpias y contaminadas con distintos tipos de ruido para evaluar el comportamiento de sistemas de reconocimiento vocal en condiciones ruidosas. La fuente de las señales vocales es la base de datos TIDigits [19] que consiste en secuencias de dígitos pronunciados por hablantes de inglés-americano. La base de datos ha sido 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.

El paquete Aurora también incluye un reconocedor de voz completo para la generación de resultados de referencia y para evaluar el comportamiento de algoritmos de pre-procesamiento de la señal vocal como es el caso de este proyecto. El objetivo perseguido en la creación de esta base de datos usada internacionalmente es 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.

2.5.1 Adición de ruido

La base de datos TIDigits contiene secuencias de hasta 7 dígitos grabadas en un entorno sin ruido. En Aurora 2 las señales son sub-muestreadas de los originales 20KHz a 8KHz y filtradas con un paso de baja ideal de ancho de banda 4 KHz (banda telefónica) para dar como resultado lo que es considerado en Aurora 2 como los datos 'limpios'.

Estos datos limpios son contaminados artificialmente con diferentes tipos de ruido y a diferentes relaciones de señal a ruido. El proceso de adición de ruido es llevado a cabo siguiendo las recomendaciones de la ITU [22][23][24], y utilizando el software proporcionado por este organismo. La contaminación original en la base de datos Aurora 2 se lleva a cabo con ruido grabado en 8 localizaciones diferentes (metro, babble, coche, museo, restaurante calle, aeropuerto y estación de tren) y a 6 diferentes niveles (20dB, 15dB, 10dB, 5dB, 0dB y –5dB).

En la realización de este proyecto se tomaron los datos 'limpios' originales de Aurora2 para generar un nuevo conjunto de datos contaminados por WGN. La contaminación se llevó a cabo siguiendo las mismas recomendaciones y software proporcionados por la ITU y a los mismos niveles de señal a ruido.

2.5.2 Definición de conjuntos de entrenamiento y de test

Aurora 2 proporciona dos conjuntos diferentes de datos de entrenamiento:

- Entrenamiento limpio
- Entrenamiento multicondición, con datos limpios y contaminados

El entrenamiento limpio tiene la ventaja de que se modela la voz sin ningún tipo de distorsión, representando toda la información disponible en una señal limpia. Es un entrenamiento

genérico, completamente independiente del tipo de ruido al que vaya a estar sometida la señal vocal en la utilización del reconocedor.

En el entrenamiento multicondición el reconocedor se entrena con señales vocales contaminadas con distintos tipos de ruido, de forma que estas distorsiones son tenidas en cuenta en los modelos y el reconocedor da mejores resultados cuando es probado en condiciones de ruido similares a las del entrenamiento.

Ambos conjuntos de entrenamiento constan de 8440 ficheros vocales seleccionados de la sección de entrenamiento de la base de datos TIDigits que contiene grabaciones de 55 hombres y 55 mujeres.

El conjunto de datos de test consta de 1001 ficheros a 6 niveles diferentes de señal a ruido (20,15,10,5,0 y –5dB). Además el caso limpio se toma como la séptima condición por lo que en total existen para cada tipo de ruido 7007 archivos de test. Los ficheros limpios originales se tomaron de la sección de test de TIDigits que contiene grabaciones de 52 hombres y 52 mujeres.

2.5.3 Reconocedor del paquete Aurora 2

El extractor de características del reconocedor incluido en Aurora 2 extrae 13 coeficientes MFCC (Mel Frequency Cepstral Coefficients) [19][21] de cada trama de 25ms en las que divide el fichero vocal. Estos coeficientes se refieren al dominio de la frecuencia y se espacian según la escala Mel (logaritmica para representar el patrón de percepción del oído humano).

Los sistemas de entrenamiento y reconocimiento usan el paquete software HTK que crea modelos de cada dígito basados en los HMMs. El entrenamiento del reconocedor se lleva a cabo en varias etapas, como es usual, utilizando el algoritmo de re-estimación de Baum-Welch basado en funciones de probabilidad gaussianas. Una descripción completa del reconocedor puede encontrarse en [18] y [21].

2.6 METODOS DE MEJORA DE LA SEÑAL VOCAL

El problema de reconocer señales vocales fuertemente distorsionadas es un gran desafío hoy día en el área del reconocimiento automático de voz. Mucho trabajo ha sido llevado a cabo en este campo dando como resultado multitud de algoritmos y diferentes enfoques a la hora de intentar mejorar la señal vocal para alcanzar mejores resultados. Existen muchas maneras de clasificar los métodos de mejora de la señal vocal, basándose en el criterio usado o la aplicación del sistema [20].

Dominio	Posibles estrategias
Número de canales	Uno / dos / múltiples
Dominio del proceso	Tiempo / Frecuencia
Tipo de algoritmo	Adaptativo / No adaptativo

Tabla 2.2 Clasificación de métodos de mejora de la señal vocal

Típicamente, los métodos se dividen en técnicas que utilizan un solo canal y técnicas multicanal. Los sistemas que utilizan un solo micrófono para captar la señal son los más comunes en el mundo real (comunicaciones móviles, implantes auditivos, etc.). Estos sistemas son fáciles de construir y de menos coste que los sistemas multicanal, pero es más difícil llevar a cabo la mejora de señal, especialmente en condiciones de ruido no estacionario ya que no disponemos de una referencia del ruido presente en cada momento (como ocurre en los sistemas multicanal). A pesar de todo será la técnica en la que se concentra este proyecto, ya que es el escenario más común donde se aplican los algoritmos de mejora de la señal vocal.

Normalmente es difícil para un algoritmo ser capaz de tratar todo tipo de ruidos. Por lo tanto, un algoritmo se basa en una serie de asunciones y limitaciones que son típicamente dependientes de la aplicación concreta y del entorno. La mayoría de los métodos de mejora trabajan en el dominio de la frecuencia tratando de estimar la potencia de ruido para substraerla de la señal. En caso de estimación perfecta esto resolvería el problema, pero la estimación dista mucho de ser ideal, siendo más o menos difícil dependiendo del ruido que estemos tratando. A continuación introducimos los dos enfoques que se han llevado a cabo en este proyecto.

2.6.1 Substracción Espectral

La Substracción Espectral es el más popular y obvio método de reducir el efecto del ruido de fondo en las señales vocales. La teoría en la que se basa es la asunción de que la señal vocal y el ruido no están correlados y por lo tanto la densidad espectral de potencia de la señal vocal ruidosa es la suma de las densidades espectrales de potencia de la señal vocal limpia y el ruido. Esta técnica estima la potencia del ruido, normalmente mediante la media temporal de los valores de la densidad espectral de potencia de la señal vocal degradada durante los periodos donde solo el ruido esta presente, y resta esta estimación a la densidad espectral de potencia de dicha señal ruidosa. La magnitud resultante y la fase de la señal ruidosa se combinan para regenerar la señal temporal, obteniendo finalmente una mejor relación señal a ruido.

La estimación clásica del ruido [6] toma los valores de la densidad espectral de potencia de la señal ruidosa durante el intervalo precedente al comienzo de la señal vocal y hace un promedio. Este acercamiento a la estimación del ruido ha sido modificado por muchos otros investigadores para intentar solventar los problemas que conlleva, como son la necesidad de un detector de actividad vocal fiable (para estimar la potencia de ruido en periodos de ausencia de señal vocal) y la asunción de que el ruido es estacionario y constante (condiciones ideales que raramente se dan en el mundo real).

Existen multitud de algoritmos basados en la estimación continua de la potencia de ruido que tratan de evitar de diversas formas los problemas del acercamiento clásico. Martín [8] propone un algoritmo basado en estadísticos mínimos por el medio del cual el ruido es estimado de los valores mínimos de la densidad espectral de potencia de la señal degradada tras ser suavizada

mediante un filtro. Arslan [10] propone un estimador de ruido adaptativo por el que la estimación de ruido se hace de forma que aumenta lentamente durante los periodos de actividad vocal pero que cae rápidamente hacia los valores instantáneos de ruido durante los periodos de ausencia de señal vocal. Hirsch-Ehrlicher [12] proponen 2 métodos: uno estima el ruido de una media ponderada de valores anteriores en la densidad espectral de potencia de la señal degradada; en el otro se utiliza un histograma de la energía por bandas. Stahl [11] extiende el anterior enfoque del histograma a una acercamiento basado en quantiles (QBNE Quantile Based Noise Estimation). Un enfoque similar es el de Ealey [13] donde el enfoque de seguimiento de harmónicos estima el ruido instantáneo de las regiones vecinas que no contienen señal vocal. Esto se basa en la idea de que incluso en los periodos de actividad no todas las frecuencias están permanentemente ocupadas con señal vocal, y que durante una considerable cantidad de tiempo la energía en cada banda esta realmente al nivel del ruido. Los resultados obtenidos por Evans y Mason [7] muestran que la estimación basada en quantiles da muy buenos resultados, alcanzando un 35% de mejora en reconocimiento vocal sobre la señal original para ruido producido en el interior de un coche. En el apartado 3.2 se desarrolla la teoría en profundidad de este tipo de estimación junto con el resto de los detalles referentes a la Substracción Espectral.

2.6.2 Filtrado morfológico

Si observamos un espectrograma de una señal vocal inmersa en ruido, comprobamos que para el ojo humano, las regiones y estructuras pertenecientes a la señal vocal son relativamente fáciles de identificar bajo condiciones de ruido razonables como se puede observar en la figura 2.11

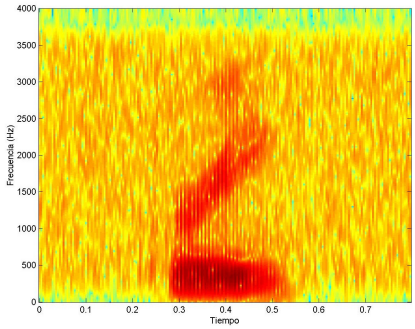

Figura 2.11 Espectrograma de señal vocal inmersa en ruido WGN

Si se quiere automatizar esta detección de formas, una técnica ampliamente conocida en procesamiento de imágenes es el filtrado morfológico. Entre los propósitos para los que se utiliza dicha técnica tenemos la mejora de estructuras, identificación de objetos, suavizado de formas, segmentación, etc... Un ejemplo típico de procesamiento vía filtrado morfológico es el reconocimiento óptico de caracteres, donde una imagen binaria procedente de un escáner es preprocesada para eliminar ruido y marcar los objetos (letras), como mejora para el posterior reconocimiento. El filtrado morfológico se utiliza en este proyecto como herramienta para el pre-procesamiento de espectrogramas de señales de voz inmersas en ruido con el objetivo de separar dichas 'imágenes', y con ello dichas señales.

2.6.2.1 Descripción de algoritmo propuesto por Hory [1]

Conceptualmente, el algoritmo considera el espectrograma como una imagen convencional formada por píxeles. El proceso de segmentación comienza asociando a cada píxel un conjunto de píxeles vecinos, formando sub-imágenes denominadas células que son derivadas para todos los píxeles de la imagen. A continuación se calculan para cada célula: la media (*Característica 1*) y la desviación típica (*Característica 2*) de los valores de los píxeles dentro de la célula. De esta forma tenemos asociado cada píxel en la imagen con dichas características estadísticas locales, formando con ello un dominio paralelo llamado *Espacio Característico,* donde se analiza a qué tipo de señal pertenece cada píxel y donde se aplican los criterios de segmentación. Cada píxel esta asociado a una célula, y aquellas células que están en zonas ocupadas predominantemente por ruido tienen diferentes estadísticas que aquellas en zonas de señal vocal más ruido, ocupando por ello un lugar diferente en el Espacio Característico y haciendo posible su distinción.

La distinción entre las distintas zonas viene controlada por una estimación de la función de densidad de probabilidad de la Característica 1 para píxeles de ruido, que se corresponde con una distribución Gamma como es demostrado en [1]. Los parámetros de esta distribución Gamma son estimados a partir de las estadísticas de la imagen a través de la teoría de estimación de máxima verosimilitud. Esta distribución hace posible definir un límite l que divide el Espacio Característico

en dos zonas: el *Área de trabajo* y la *Región de Confianza de Ruido* asumiendo una probabilidad de error P_e en la estimación de la distribución tal que Pe= Prob{Característica 1}> l . La primera Área de Trabajo resultante propone píxeles candidatos para la selección semillas de un posterior proceso de propagación, mientras que la otra zona contiene píxeles que probablemente se correspondan con píxeles pertenecientes a zonas de ruido en el Espectrograma. A continuación se calcula un *Grid teórico* con parámetros asociados a los contenidos de señal que pueden albergar las células del espectrograma, y se superpone sobre el Espacio Característico para servir como guía en la selección de semillas. Los puntos del Espacio Característico más cercanos a ciertos puntos del Grid Teórico y pertenecientes al Área de Trabajo tendrán más probabilidad de corresponderse con píxeles del Espectrograma pertenecientes a señal vocal, y por ello serán los primeros seleccionados como semillas para la propagación.

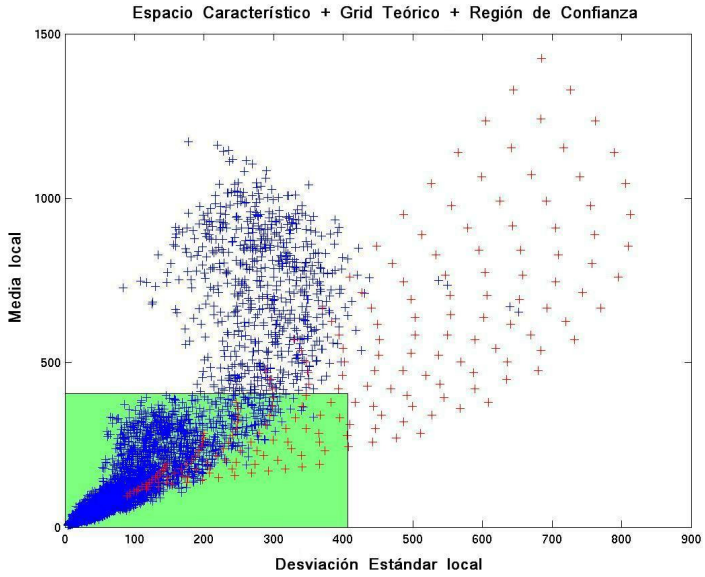

Figura 2.12 Espacio Característico, Región de Confianza de Ruido y Grid Teórico

Las semillas seleccionadas en el Espacio Característico son mapeadas a sus posiciones en el Espectrograma y se comienza la propagación a los píxeles vecinos. Si un píxel vecino se corresponde con un punto en el Espacio Característico dentro del Área de Trabajo se convertirá a su vez en semilla y el mismo proceso se llevará acabo sobre sus nuevos píxeles vecinos. Todos los píxeles del Espectrograma que se han convertido en semillas se asumen correspondientes a zonas donde es dominante la señal vocal y como consecuencia extraídos. Este proceso se repite hasta que

los vecinos de todas las semillas seleccionadas en ese momento caen fuera del Área de Trabajo, lo que dará por concluida una iteración del proceso global.

Para llevar a cabo la siguiente iteración se vuelve a estimar la función de densidad de probabilidad del ruido, se determina un nuevo límite l y se computa una nuevo Grid teórico; todo ello a partir de los puntos correspondientes a los píxeles que no fueron extraídos en la iteración anterior. Debido a las propiedades de la *Característica 1* el nuevo límite siempre será inferior, por lo que tendremos un nuevo grupo de candidatos en el Espacio Característico sobre los que trabajar. Iteración tras iteración, la estimación de la densidad de probabilidad del ruido se hace más y más precisa ya que se basa en regiones que contienen cada vez menos señal vocal. El proceso de segmentación finalmente concluye cuando la estimación converge; es decir, cuando el límite l computado se hace prácticamente constante, indicando que la estimación de la PDF no varía ya que toda el área de señal determinista ha sido extraída.