En este capítulo se explica en detalle el funcionamiento de los dos algoritmos que se utilizan en este proyecto para preprocesar la señal vocal como vía para mejorar el comportamiento del reconocedor de voz. En primer lugar se realiza un estudio del algoritmo de filtrado morfológico propuesto por Hory en [1] concentrándonos en las peculiaridades que serán objeto de optimización en el siguiente capítulo, mostrando al final los logros conseguidos hasta el momento en el contexto del reconocimiento de voz. Posteriormente se pasa a explicar en detalle la técnica de substracción espectral con estimación de ruido basada en quantiles, y se realizará un experimento cuyos resultados servirán de referencia para la mejora del filtrado morfológico en el siguiente capítulo y que motivarán el estudio de la posible combinación de ambas técnicas que se realiza en el Capítulo 5. Esto se detalla en el análisis comparativo de ambas técnicas que se realiza finalmente y donde se definen los objetivos que se perseguirán en los siguientes capítulos

Capítulo 3

FILTRADO MORFOLÓGICO Y SUBSTRACCIÓN ESPECTRAL QBNE

3.1 ALGORITMO DE SEGMENTACIÓN de HORY

El algoritmo de segmentación se basa en las diferentes propiedades estadísticas que muestran los píxeles del espectrograma en las zonas donde sólo el ruido esta presente y las zonas donde tenemos mezcla de señal determinista y ruido. Por ello, en primer lugar realizamos un estudio de las características estadísticas de un espectrograma de una señal inmersa en WGN. A continuación se detallará paso a paso el funcionamiento del algoritmo iterativo que segmenta las zonas de señal determinista. En todos los apartados se explicarán los cálculos y procedimientos utilizados en base a consideraciones prácticas y en relación con los objetivos de este proyecto. En [1] se puede encontrar una explicación en profundidad del algoritmo y de la teoría estadística involucrada.

3.1.1 Análisis estadístico del espectrograma

Los coeficientes Tiempo-Frecuencia del espectrograma, $S_x[n,k]$, son considerados como píxeles de un imagen con unas características estadísticas determinadas dependiendo del área al que pertenezcan (señal + ruido o solo ruido).

Para analizar las características estadísticas de las distintas zonas del espectrograma, el primer paso es asociar a cada píxel una célula formada por los 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 espectrograma. Las dimensiones de esta célula se toman como 3 x 3, es decir, a cada píxel se le asocian los 8 píxeles que tiene alrededor.

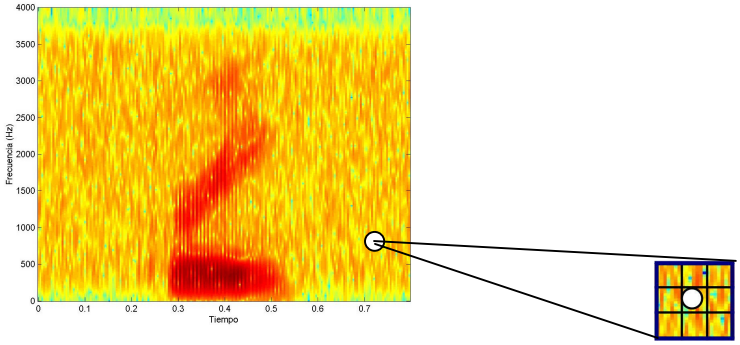


Figura 3.1 Asociación de célula a cada píxel del espectrograma

Para cada una de las células que hemos derivado (tantas como píxeles tenemos en el espectrograma) calculamos dos características estadísticas:

- Media local (*Característica 1*):

$$F_1 = \frac{1}{N}\sum_{C} S_x[n,k] \tag{3.1}$$

- Desviación estándar local (*Característica 2*):

$$F_2 = \left[\frac{1}{N}\sum_{C}\left(S_x[n,k] - F_1\right)^2\right]^{1/2} \tag{3.2}$$

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. Una vez aquí, tenemos cada píxel del espectrograma asociado con un par (F_1, F_2). El conjunto de todos estos pares forman lo que llamamos el *Espacio Característico*. *Característica 1* indica si un píxel está situado en una región de proximidad a píxeles de alta o de baja energía, mientras que *Característica 2* indica si la región es de alta o de baja variación de energía.

En la representación gráfica de la *Característica 1* frente a la *Característica 2* podemos observar la formación de racimos de píxeles pertenecientes a zonas con características estadísticas similares, por lo que 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.

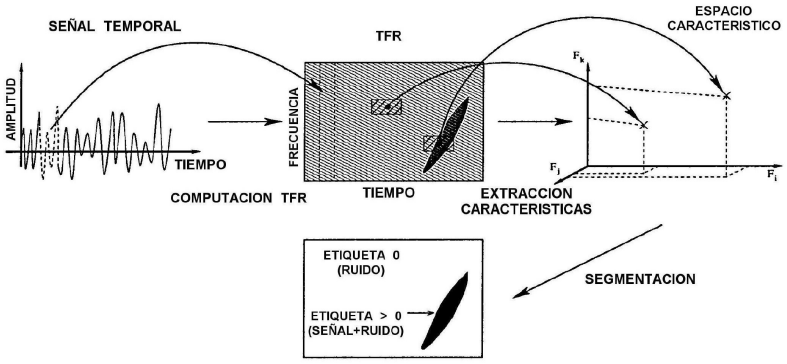


Figura 3.2 Visión general del sistema

3.1.2 Modelo estadístico teórico

Para determinar qué posiciones del *Espacio Característico* se corresponden con píxeles pertenecientes a zonas de ruido y qué localizaciones son las correspondientes a píxeles con más o menos probabilidad de contener señal determinista se desarrolla un modelo estadístico teórico del espectrograma [1] . A través de este modelo se derivan las siguientes expresiones teóricas para la media local y la desviación estándar local:

$$E\{F_1\} = \hat{\sigma}^2 (1 + pr) \tag{3.3}$$

$$E\{F_2\} = \hat{\sigma}^2 \left[\frac{\sum_{i,j=0}^{4} a_{ij} p^i r^j}{\sum_{i,j=0}^{2} b_{ij} p^i r^j} \right]^{\frac{1}{2}} \tag{3.4}$$

Donde a_{ij} y b_{ij} son coeficientes que solo dependen del tamaño de la célula. Asumiendo ergodicicidad , esto es, los parámetros estadísticos de una muestra o secuencia de un proceso son representativos de las del proceso en conjunto, los valores teóricos $E\{F_1\}$ y $E\{F_2\}$ se corresponderán con los estimadores imparciales F_1 y F_2 computados a partir de los valores reales del espectrograma.

Las fórmulas dependen de los dos siguientes parámetros:

- $p = P/N$: 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. En el siguiente apartado se describe la forma en la que se estima la potencia de ruido presente en el espectrograma.

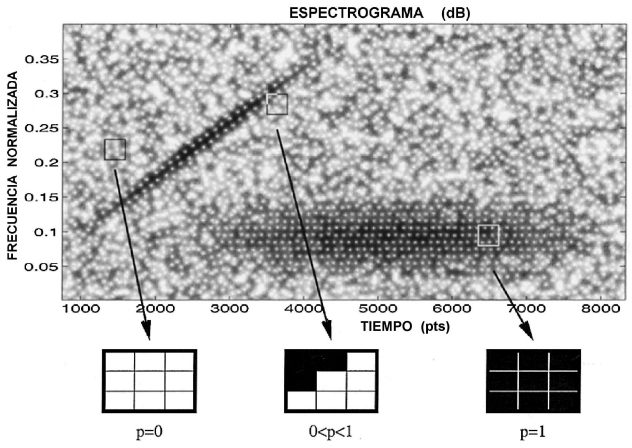


Figura 3.3 Valores del parámetros p para células en una zona sin señal determinista (a), una zona mixta (b) y una zona de señal determinista (c)

Calculando $E\{F_1\}$ y $E\{F_2\}$ para el rango de posibles valores de p y r derivamos lo que llamamos el *Grid teórico*, que se trata de una red de puntos que se superponen al *Espacio Característico* para indicar las zonas con mayores probabilidades de tener señal determinista **para cada p y r determinado.** Si utilizamos 15 puntos para recorrer los posibles rangos de valores tanto de p como de r tendremos 15x15 posibilidades que se traducirán en 225 diferentes localizaciones indicadas por el Grid teórico.

p = [0,1/15, 2/15, , , , , , , , , ,14/15, 1]

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

maxvalue= máximo valor posible del píxel en el espectrograma

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* = maxvalue). 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 (puntos cercanos a localizaciones de p y s con valores altos). A estos elementos se les denominarán semillas.

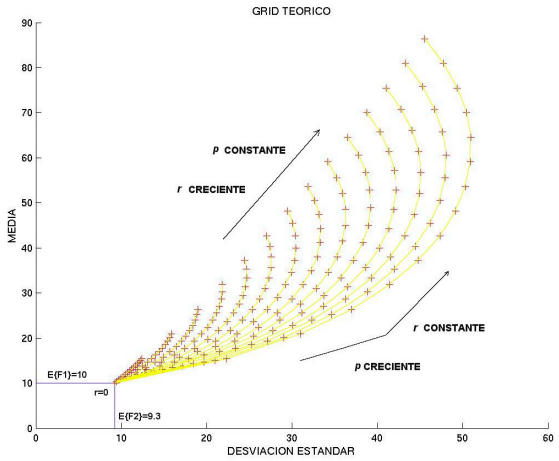


Figura 3.4 Grid Teórico $\left(E\{F_1\}, E\{F_2\}\right)$ computado con una varianza de ruido $\sigma^2 = 10$ y una célula de $N = 3x3$ puntos. El punto $\left(E\{F_1\} = 10, E\{F_2\} = 9.3\right)$ es el valor de ruido esperado. Los 15 valores del los parámetros *r* y *p* son regularmente espaciados entre, respectivamente, [0,6] y [0,1] (+)]

3.1.3 Estimación de la PDF de ruido WGN

La estimación de la PDF de la *Característica 1* pertenecientes a células en zonas de ruido tiene tres objetivos:

- Estimar la potencia de ruido σ^2 necesaria para calcular el Grid teórico
- Definir el Área de trabajo y la Zona de confianza de ruido
- Determinar la condición de parada de las iteraciones

Se demuestra [1] que la media local (*Característica 1*) de los coeficientes de un espectrograma WGN siguen una distribución gamma cuyos parámetros de forma y escala μ y ν pueden ser estimados mediante MLNE (teoría de estimación de máxima verosimilitud):

$$u_1 = \frac{1}{N}\sum_{n=1}^{N}\ln(X_n) \qquad u_2 = \frac{1}{N}\sum_{n=1}^{N}X_n$$

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

Siendo X_n la *Característica 1* para todos los píxeles del espectrograma desde n=1 hasta n=N.

3.1.3.1 Estimación de la potencia de ruido necesaria en el cálculo del Grid teórico

Se demuestra [1] que la potencia de ruido WGN presente en el espectrograma puede ser estimada de forma eficiente como:

$$\sigma^2 = \boldsymbol{u}_2 = \frac{\hat{\mu}}{\hat{\nu}}$$

3.1.3.2 Definición del Área de trabajo y la Región de confianza de ruido

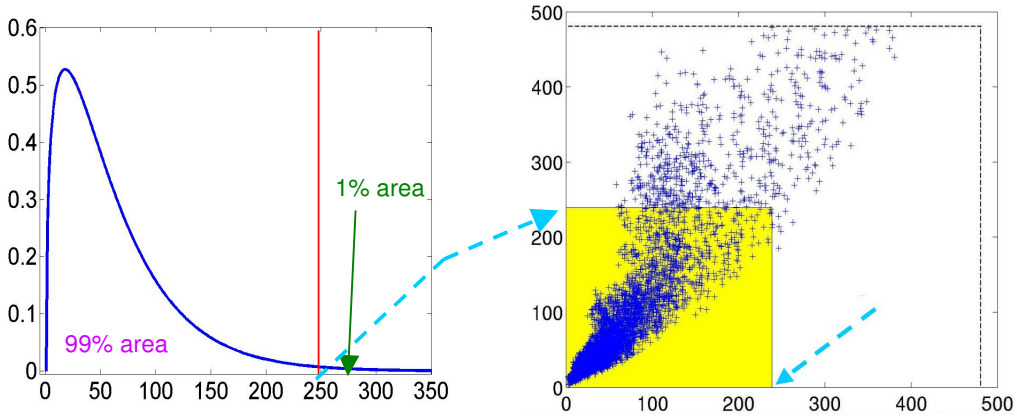
A través de la PDF estimada determinamos un límite para la *Característica 1* por debajo del cual la mayor parte de los píxeles serán de ruido, por lo que no los tendremos en cuenta a la hora de

buscar semillas. Asumiendo una probabilidad de error de detección P_e (típicamente 0.01), se deriva el límite de la distribución l a partir de la distribución de manera que se cumpla:

$$P_e = Prob\{F_1 \geq l\}$$

Las dos zonas en las que queda dividido el *Espacio Característico* son:

1. **Región de Confianza de ruido**: formada por el espacio $[0,l]x[0,l]$
2. **Área de Trabajo**: la Región perteneciente a: $F_1 > l$ y $F_2 > l$



a b

Figura 3.5 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*).

3.1.3.3 Determinación de la condición de parada de las iteraciones

Iteración tras iteración, más señal determinista va siendo extraída y las estimaciones de la PDF de ruido se va haciendo más precisa ya que se basan en una mayor proporción de píxeles ruidosos. Cuando esta estimación llega a un cierto grado de convergencia se da por concluido el proceso de segmentación. Este proceso se explica en detalle en el apartado 3.1.4.4.

3.1.4 Algoritmo iterativo de segmentación

El diagrama de flujo principal del algoritmo se muestra a continuación:

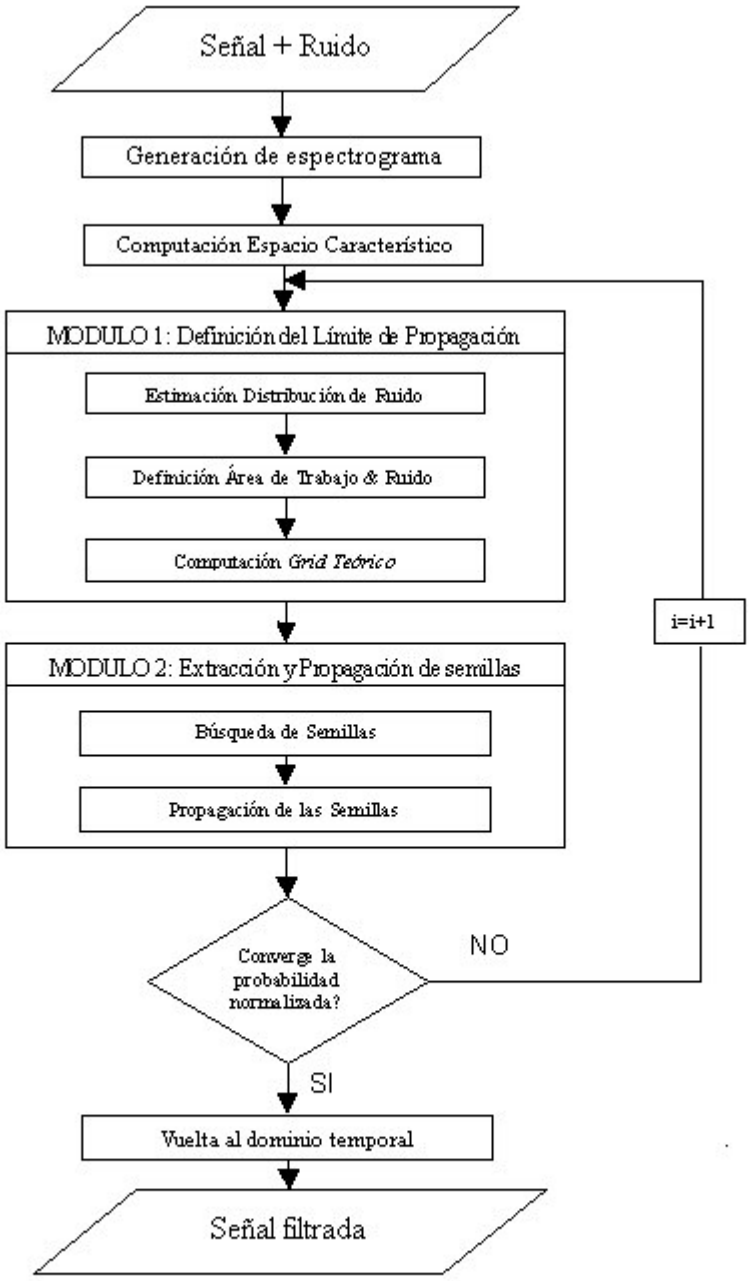


Figura 3.6 Diagrama de flujo del algoritmo iterativo

3.1.4.1 Generación de espectrograma y computación del Espacio Característico

El espectrograma se genera tal y como se describe en el apartado 2.1.3 con los siguientes parámetros:

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

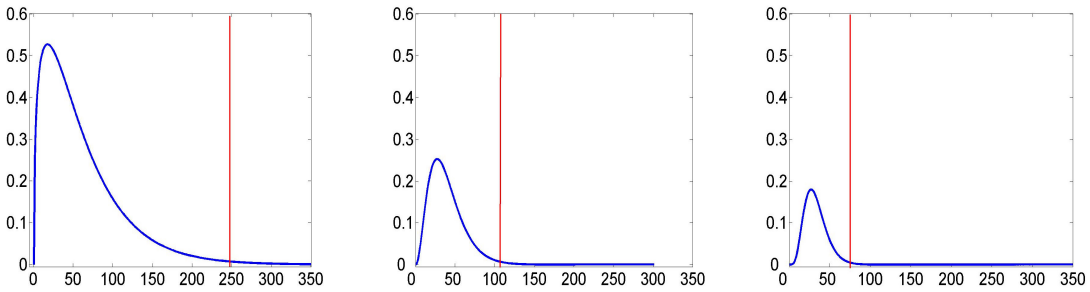
Para conseguir mejores resultados en la segmentación se lleva a cabo un filtrado en el dominio cepstrum que elimina las variaciones debidas a la frecuencia fundamental, dando como resultado un espectrograma mucho más suavizado.

Una vez obtenido el espectrograma suavizado se calcula para cada píxel la media local y la desviación estándar local tal como se describe en el apartado 3.1.1 derivando así el Espacio Característico.

3.1.4.2 MODULO 1

A continuación comienza el ciclo iterativo del algoritmo, empezando cada iteración con la estimación de la PDF de ruido, la definición del límite de propagación y la computación del Grid teórico.

En cada iteración del algoritmo una parte de señal determinista es extraída por lo que cada vez la estimación de la PDF de ruido es más precisa. En este proceso el límite de propagación se va haciendo más pequeño en cada iteración, de forma que la Región de confianza de ruido será más pequeña y exacta y el Área de trabajo mayor.



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

Figura 3.7 Ilustración del límite decreciente al incrementar el número de iteraciones.

El Grid teórico también va cambiando en cada iteración ya que así lo hacen los parámetros en los que se basa: maxvalue y σ^2, ajustándose de esta forma al nuevo Espacio Característico después de cada extracción. Aunque el Espacio Característico solo es computado una vez al comienzo del proceso, iteración tras iteración vamos eliminando de él los puntos ligados a los píxeles extraídos en la anterior iteración.

3.1.4.3 MODULO 2

Una vez obtenida toda la información necesaria en los pasos anteriores pasamos a la búsqueda de semillas y su posterior propagación. Para cada punto del Grid teórico dentro de el Área de trabajo trazamos un círculo de radio:

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

Los puntos del Espacio Característico que se encuentren dentro de este círculo se convertirán en semillas que procederemos a propagar. El primer círculo será alrededor del punto del Grid teórico correspondiente a p y r máximas (mayor probabilidad de contener señal determinista). Después, para cada valor de p (recorriendo los valores de máximo a mínimo) se buscan semillas en los puntos para cada valor de r (recorriendo los valores de máximo a mínimo).

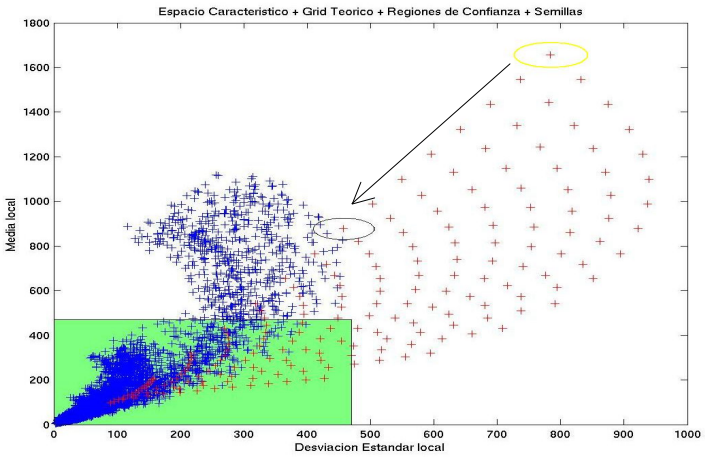


Figura 3.8 Ilustración del proceso de selección de las semillas comenzando por el punto más alto del Grid. Nótese que el radio d los círculos es distinto al depender estos de los valores de *p* y *r* en el Grid

Cada vez que se evalúa uno de los puntos del Grid teórico, se procede a la propagación de las semillas encontradas antes de evaluar el siguiente punto.

La propagación se lleva a cabo mediante un algoritmo de región creciente basándose tanto en el Espacio Característico como en el espectrograma. 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. Si estos píxeles se corresponden con puntos dentro del Área de trabajo pasan a su vez a ser semillas, procediéndose a su extracción del espectrograma, y a evaluar de nuevo sus 8 píxeles vecinos de forma análoga a la descrita. 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.

3.1.4.4 Convergencia de la PDF de ruido

Como ya se ha comentado anteriormente, las estimaciones de la PDF de ruido se van haciendo más precisas en cada iteración y el límite de propagación derivado se va haciendo cada vez más pequeño y exacto.

Llega un momento en que casi toda la región determinista ha sido extraída por lo que la estimación de la PDF no cambia apreciablemente y el límite de propagación se mantiene constante. Este hecho será el indicativo de que el algoritmo debe finalizar por lo que la condición de parada se basa en la diferencia entre el límite de propagación computado en una iteración y en la siguiente:

$$\frac{\left|\text{último límite - límite previo}\right|}{\text{último límite}} \cdot 100\% < 1\%$$

Si el cambio es menor del 1% (valor típico) se considera que toda la señal determinista ha sido extraída y se finaliza el proceso de segmentación. Si por el contrario el porcentaje es mayor se procedería a otra iteración del algoritmo.

3.1.5 Aplicación al contexto del reconocimiento vocal

El algoritmo que lleva a cabo el filtrado morfológico fue propuesto por Hory en [1] demostrando la capacidad de discriminar en un espectrograma las zonas que contienen solo ruido WGN y las que contienen además señal determinista. En [3] se lleva a cabo una implementación independiente del algoritmo corroborando los resultados obtenidos por Hory y como extensión se aplica el algoritmo por primera vez a señales vocales como pre-procesamiento para mejorar el comportamiento del reconocimiento automático de voz.

Se utiliza la base de datos Aurora 2 en su típica configuración (8440 ficheros de entrenamiento limpios y 7 grupos de 1001 ficheros de test) pero con ruido WGN como se explica en el apartado 2.5.1 El algoritmo de segmentación se aplica tanto a los ficheros de entrenamiento como

a los de test; y antes de volver al dominio temporal la energía de las áreas del espectrograma consideradas deterministas se ponen a un nivel alto constante y las consideradas como ruido a cero, de forma que se elimina toda información de variaciones en las amplitudes espectrales a la vez que se elimina todo tipo de ruido. Los resultados de exactitud en reconocimiento vocal con esos conjuntos de entrenamiento y test se muestran en la figura 3.9 frente al resultado base obtenido con los mismos conjuntos de datos pero sin ningún tipo de procesamiento.

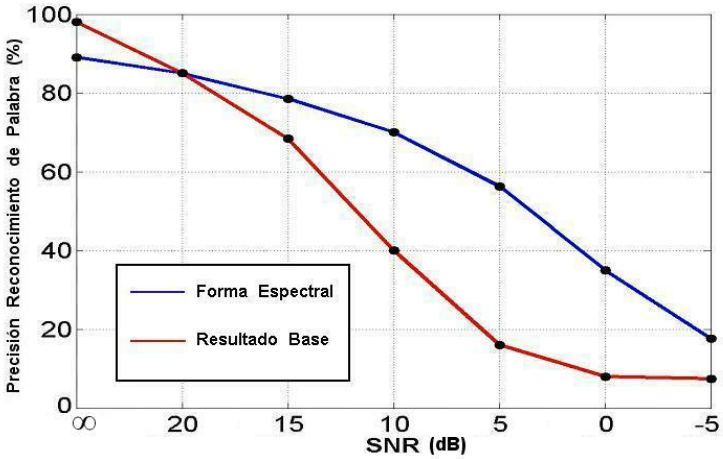


Figura 3.9 Comparación de la Precisión en el Reconocimiento entre el Resultado Base y la Forma Espectral obtenida tras la segmentación.

Observamos que para el caso limpio los resultados para la Forma Espectral son sorprendentemente buenos (89%), teniendo en cuenta que toda la información de variaciones espectrales ha sido eliminada, poniendo de manifiesto la importancia en reconocimiento vocal de las formas espectrales. Además observamos una mayor robustez frente al ruido, cayendo solo hasta el 56% a 5dB frente al 16% de la señal sin filtrado morfológico, lo cual implica una mejora del 40%.

3.2 SUBSTRACCIÓN ESPECTRAL QBNE

La Substracción Espectral 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. La potencia del ruido se estima como una media de valores del espectrograma en los periodos donde solo el ruido esta presente, y posteriormente se resta 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.

3.2.1 Notación y Asunciones

Se asume que la señal de ruido observada es una realización de un proceso estacionario en sentido amplio con densidad espectral de potencia $N(w)$. La estimación de $N(w)$,llamada $\hat{N}(w,t)$, se calcula a partir de los valores de la densidad espectral de potencia de la señal original desde el comienzo de la señal hasta el instante t . Asimismo, se asume que la señal vocal limpia en cada trama t es una realización de otro proceso estacionario en sentido amplio con densidad espectral de potencia $S(w,t)$. Como se supone que la señal vocal y el ruido son aditivas e independientes, la densidad espectral de potencia de la señal ruidosa observada es:

$$X(w,t) = S(w,t) + N(w)$$

La densidad espectral de potencia se estima como el cuadrado de la magnitud de los coeficientes de la transformada de Fourier de la señal en cada ventana t. La densidad espectral de potencia de la señal limpia puede ser estimada como:

$$\hat{S}(w,t) = X(w,t) - \hat{N}(w,t)$$

3.2.2 Estimación del 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 (o intevalos

intermedios donde la señal vocal este austente) y hace un promedio tal que: $\hat{N}(w,t) = E[X(w,t)]$ a lo largo de un intervalo t que cumpla con la condición. 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). Además el método depende de que realmente existan suficientes periodos de no actividad vocal en la señal.

La estimación de ruido llevada a cabo en este proyecto es propuesta por Stahl en [11] y se denomina estimación de ruido basada en quantiles (QBNE: Quantile Based Noise Estimation). La idea consiste en estimar la energía del ruido en cada banda frecuencial mediante quantiles temporales de la densidad espectral de potencia. Martín [8] sentó las bases de este enfoque proponiendo un algoritmo basado en estadísticas mínimas. Como los mínimos son sensibles a excepciones que se produzcan, etc. en este método utiliza un quantil distinto del mínimo.

La experiencia demuestra que incluso en las secciones de actividad vocal no todas las bandas de frecuencia están permanentemente ocupadas con señal vocal. De hecho, la energía en cada banda de frecuencia esta a nivel de ruido durante un porcentaje considerable del tiempo. Los límites entre las zonas de ruido y actividad vocal son implícitamente detectados para cada banda haciendo una estimación de la potencia de ruido que se va actualizando a lo largo de toda la señal. De esta manera no es necesario un detector de actividad vocal y la estimación de ruido puede reaccionar a cambios en el mismo durante intervalos de actividad vocal.

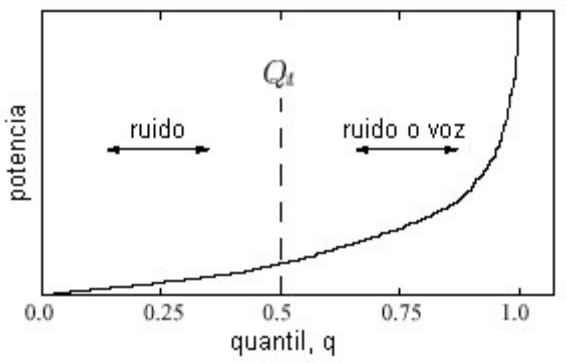


Figure 3.10 Ilustración de los quantiles de una señal inmersa en ruido. Los valores por debajo de Q_t se asume que proporcionan una estimación fiable de los valores instantáneos de ruido

Para cada frecuencia w a lo largo de un periodo T, la potencia a esa frecuencia en cada trama se coloca en una pila FIFO que se ordena numéricamente, tomando como estimación del ruido el valor que se encuentra en la posición central de dicha pila. La estimación de ruido QBNE, $\hat{N}(w,t_0)$ a la frecuencia w y en el instante t_0 se define como:

$$\hat{N}(w,t_0) = N_{\frac{n}{2}+1}(w) \text{ asumiendo que n es impar}$$

donde $N_q(w)$ es un buffer ordenado numéricamente de longitud n que contiene los valores de $X(w,t)$ donde $t-\frac{T}{2} < t < t+\frac{T}{2}$. El proceso es continuo y nuevos valores instantáneos van remplazando los más antiguos en el buffer, a medida que vamos moviendo la ventana temporal de longitud T.

3.2.3 Eliminación del ruido de la señal vocal

Una vez hecha la estimación de ruido se procede a la substracción espectral propiamente dicha. Como hemos comentado anteriormente la densidad espectral de potencia de la señal limpia puede ser estimada como $\hat{S}(w,t) = X(w,t) - \hat{N}(w,t)$ pero resultados experimentales revelan que es mejor llevar a cabo una variación de esta esquema basada en dos parámetros: α y β

$$Y(w,t) = X(w,t) - \alpha\hat{N}(w,t)$$

$$\hat{S}(w,t) = \begin{cases} Y(w,t), \text{ si } Y(w,t) > \beta X(w,t) \\ \beta X(w,t) \text{ en otro caso} \end{cases}$$

El parámetro α sirve para hacer una sobre-estimación de la potencia de ruido y β controla el nivel base que dejamos en el espectrograma. El espectro resultante tras la substracción consiste en una serie de picos y valles repartidos aleatoriamente, por lo que se realiza una sobre-substracción que trata de reducir dicha variabilidad en el espectro.

El valor de α es función de una estimación de la relación señal a ruido:

$$\alpha = \alpha_0 - \frac{3}{20} SNR \qquad \alpha_0 \in (3,6)$$

Este parámetro tiene que ser elegido cuidadosamente para prevenir por una parte el ruido remanente pero no producir demasiada distorsión en la señal.

La introducción de un nivel base hace que el espectro no baje del valor $\beta X(w,t)$, 'llenando' de esta forma los valles profundos que rodean a los picos (del espectrograma mejorado). Esto reduce la variabilidad y con ello el ruido remanente. Para niveles de ruido altos β debería estar en el rango (0.1 , 0.01). Para niveles de ruido menores β debe estar por debajo de 0.01.

La magnitud resultante de esta operación se combina con la fase original de la señal degradada para regenerar una nueva señal temporal con una mejor relación señal a ruido, y por consiguiente, de mejor calidad.

3.2.4 Experimentos y Resultados

En [7] se lleva a cabo una evaluación de este método llevando a cabo el siguiente experimento de reconocimiento automático de voz con muestras inmersas en ruido producido en un coche.

Se utiliza la base de datos Aurora 2 en su típica configuración (8440 ficheros de entrenamiento limpios y 7 grupos de 1001 ficheros de test) Los ficheros de test contaminados por ruido de coche se proporcionan como parte del paquete Aurora 2 y han sido generados mediante adición de ruido real de coche a muestras vocales limpias mediante el software y las

recomendaciones de la ITU. El reconocedor es entrenado con las 8440 señales de voz limpias y no tratadas con ningún algoritmo.

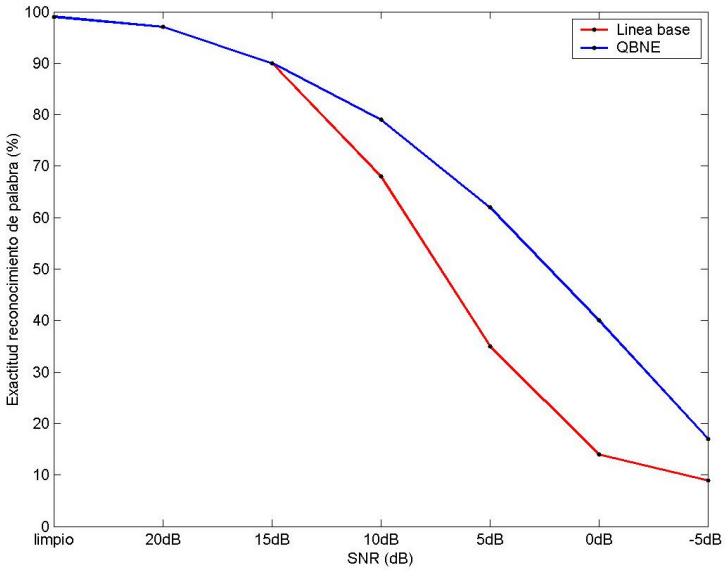


Figura 3.11 Comparación del Reconocimiento de palabra frente a SNR para la línea base sin tratar en rojo; y con el tratamiento de substracción espectral en azul (ruido de coche)

La línea base (en rojo) muestra los resultados en exactitud del reconocimiento para las señales vocales inmersas en ruido de coche sin ningún tipo de tratamiento y la curva QBNE (en azul) muestra los resultados que se obtienen para las señales tratadas con el algoritmo de substracción espectral. Para las relaciones señales a ruido altas no existe ninguna mejora apreciable sobre la línea base. Para todas las SNR por debajo de 15dB se observa una mejora notable en la exactitud del reconocimiento por lo que se demuestra la validez de esta técnica para mejorar la señal vocal bajo condiciones adversas de ruido de coche.

El algoritmo de filtrado morfológico de Hory [1] esta diseñado para trabajar segmentando señales deterministas inmersas en ruido WGN. Los resultados obtenidos en [3] en reconocimiento automático de voz se hacen como es lógico bajo condiciones de este mismo ruido; por lo que para poder llevar a cabo un estudio comparativo de resultados y poder juzgar objetivamente la efectividad del filtrado morfológico se lleva a cabo el mismo experimento que en [7] pero con señales de voz inmersas en WGN (los ficheros de test inmersos en WGN se obtienen de la misma

forma que los contaminados por ruido de coche como se explica en el apartado 2.5.1). La obtención de ambas curvas, como en el experimento de [7] se obtienen con un reconocedor entrenado con señales limpias sin ningún tipo de tratamiento.

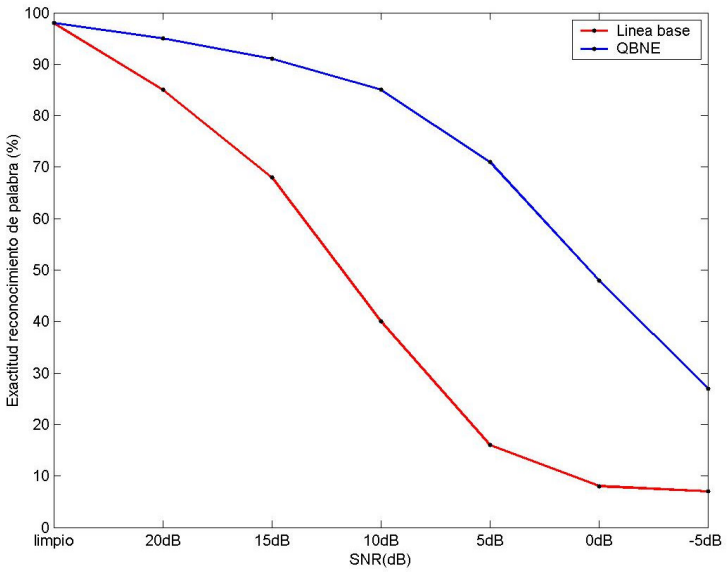


Figura 3.12 Comparación del Reconocimiento de palabra frente a SNR para la línea base sin tratar en rojo; y con el tratamiento de substracción espectral en azul (wgn)

El espectrograma es derivado según los parámetros que se detallan en el apartado 3.1.4.1 y la ventana que se toma para hacer la estimación de ruido tiene una longitud de 128 muestras que equivale a un intervalo temporal de 1 segundo aproximadamente. La línea base muestra los resultados en exactitud del reconocimiento para las señales vocales inmersas en WGN sin ningún tipo de tratamiento y la curva QBNE muestra los resultados que se obtienen para las señales tratadas con el algoritmo de substracción espectral. Observamos una mejora bastante más acusada que la obtenida para ruido de coche sobre su correspondiente línea base, teniendo mejoras notables para todas las SNR desde 20dB a –5dB. El punto donde se obtiene máxima mejora es para 5dB donde la exactitud en el reconocimiento de la señal preprocesada es de un 55%.

Es necesario destacar que los parámetros α y β de la substracción espectral (sobre-estimación y nivel base) son dependientes de la relación señal a ruido y han sido optimizados en cada una de las diferentes SNR en el experimento para obtener el mejor resultado en exactitud en reconocimiento de palabra posible con esta técnica. Esto significa que para alcanzar estos resultados

en un caso real tendríamos que tener algún tipo de información sobre la SNR de la señal ya que los parámetros del algoritmo dependen de ella. Esto no es posible en la mayoría de los casos; por lo tanto las cifras de este experimento hay que considerarlos como umbrales máximos teóricos que la substracción espectral puede conseguir preprocesando señales inmersas en WGN. Los parámetros utilizados para cada SNR se muestran a continuación.

SNR		
20dB	1.8125	0.0125
15dB	1.75	0.01875
10dB	3.25	0.01875
5dB	3.625	0.01875
0dB	5.25	0.0125
-5dB	6.3125	0.0125

Tabla 3.1 Valores óptimos de α y β para cada relación señal a ruido

Vemos como a medida que el nivel de ruido va aumentando es mejor llevar a cabo una sobreestimación del ruido más acusada, mientras que el parámetros que controla el nivel base no tiene tanta influencia y se mantiene más o menos en los mismos valores para todas las SNR.

3.3 ANÁLISIS COMPARATIVO Y OBJETIVOS

El algoritmo presentado en [1] se muestra como una técnica efectiva para identificar las zonas de señal determinista en un espectrograma contaminado por WGN y poder realizar así una segmentación que elimina el ruido de las zonas que el algoritmo considera ausentes de señal. La extensión realizada en [3] muestra el potencial que el filtrado morfológico tiene en el contexto de

las señales vocales y el reconocimiento automático de voz, obteniendo un comportamiento aceptable incluso en condiciones de señal a ruido negativas. A pesar de los prometedores resultados obtenidos con el filtrado morfológico, el comportamiento del reconocedor esta todavía lejos de los niveles de exactitud obtenidos mediante otras técnicas de mejora de la señal vocal como es el tratamiento con substracción espectral. En la figura 3.13 se muestra una comparación entre los resultados obtenidos mediante la substracción espectral con estimación de ruido basada en quantiles, y los resultados obtenidos en [3]

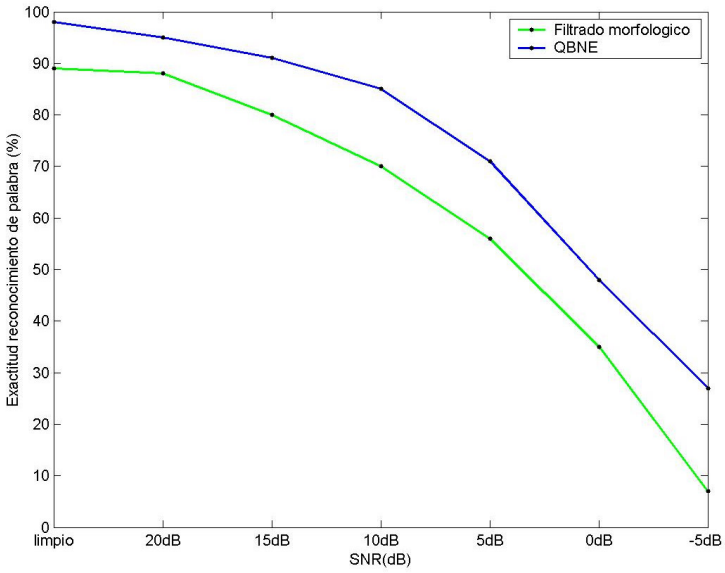


Figura 3.13 Comparación del Reconocimiento de palabra frente a SNR para el tratamiento mediante substracción espectral en azul y el filtrado morfológico en verde (wgn)

La diferencia existente muestra que el algoritmo de filtrado morfológico, aún teniendo buen comportamiento esta por debajo de otras técnicas bien establecidas como la substracción espectral. Pero el filtrado morfológico debe ser adaptado para trabajar con señales vocales ya que el diseño de [1] se basa en la segmentación de señales sintéticas (chirps) cuyas características se alejan mucho de la voz. De esta forma, se hace necesaria la toma de una serie de medidas en el algoritmo para solucionar las insuficiencias de éste en el campo de las señales vocales. El presente proyecto lleva a cabo un estudio detallado del proceso llevando a cabo ajustes en todos los estadios del algoritmo para dar lugar a unos resultados óptimos en el área del reconocimiento vocal. Este estudio abarca los siguientes puntos que serán tratados en el Capítulo 4:

- Computación del Grid teórico
- Inclusión de las variaciones de amplitud espectral
- Estimaciones de la PDF de ruido y los límites de propagación
- Condición de parada del algoritmo
- Condiciones de Entrenamiento / Test

Además, animados por los resultados obtenidos con el tratamiento mediante substracción espectral se realiza un estudio de la posible combinación de ambas técnicas. Se proponen diversas configuraciones para aprovechar las variaciones de amplitud espectral 'mejoradas' que ofrece la substracción espectral, determinando la mejor configuración a usar en dicha combinación a través de la realización de una serie de experimentos que se detallarán en el Capítulo 5.

Por otra parte llama la atención el elevado requerimiento computacional que tiene el filtrado morfológico en comparación con la Substracción Espectral. En el ámbito de la telefonía móvil por ejemplo, hay un claro requerimiento de algoritmos eficientes que no necesiten muchos recursos, ya que típicamente deben trabajar en tiempo real. Incluso cuando tales algoritmos de mejora de la señal vocal sean se ejecuten de manera remota (por ejemplo en un servidor de la red) el elevado tráfico que este tipo de redes soporta hace que el requerimiento de eficiencia sea igualmente importante. Por ello se proponen diversas maneras de reducir la carga computacional, planteando modificaciones del algoritmo que aumenten su eficiencia pero sin perder competitividad. En primer lugar se estudia la manera de eliminar la necesidad del proceso iterativo e ir a las condiciones finales del algoritmo mediante una serie de estimaciones y haciendo uso de datos experimentales que indican las condiciones óptimas para los distintos parámetros implicados en el proceso. A continuación se propone un algoritmo que elimina la necesidad de la propagación de semillas basándose en el estado en el que queda el Espacio Característico tras dicho proceso y por último se estudia una manera mucho más eficiente de evitar las iteraciones realizando una estimación de las condiciones finales a través del mismo método que sigue el algoritmo original pero de una forma más eficaz. Las prestaciones de todos estos acercamientos y la combinación entre ellos serán evaluadas en el Capítulo 6 a través de experimentos que midan tanto la reducción del coste computacional como la competitividad en resultados de reconocimiento automático de palabra.

Las herramientas utilizadas en el desarrollo de este proyecto han sido:

- Ordenadores, DELL GX-240, Pentium IV. 2 GHz
- Base de datos Aurora 2.0
- MatLab TM v6.1.
- Software HTK de reconocimiento automático de voz
- Sistema operativo Linux por la estabilidad requerida en los experimentos