El algoritmo de filtrado morfológico presentado en [1] y aplicado al contexto del reconocimiento vocal por primera vez en [3] dista mucho de estar diseñado para llevar a cabo dicha tarea. Por ello, tras detectar una serie de insuficiencias y posibles formas de mejorar el comportamiento en este contexto se llevan a cabo una serie de modificaciones que resultan en un algoritmo optimizado para la mejora de la señal vocal para el reconocimiento automático. En primer lugar se mejora la forma de encontrar los puntos a propagar en al algoritmo de región creciente, mediante un reajuste de la estructura que sirve como guía para su búsqueda. A continuación se expone la forma en la que se deben incluir las variaciones de amplitud espectrales, obviadas en el acercamiento de [3], y se comprueba mediante experimentos la utilidad de dicha inclusión. Por último se realiza un proceso de optimización de las áreas segmentadas, tanto en las muestras de entrenamiento como en las de test, mediante un ajuste de los parámetros del algoritmo que hace mejorar de manera notable los resultados de exactitud en reconocimiento.

Capítulo 4

OPTMIZACIÓN DEL FILTRADO MORFOLÓFICO PARA TRATAR SEÑALES VOCALES

4.1 ALINEAMIENTO DEL GRID TEÓRICO

El Grid teórico es la estructura de puntos que se superpone al Espacio Característico para asistir en la búsqueda de semillas que han de ser propagadas. De esta manera, la correcta posición de este conjunto de puntos calculados a partir de características estadísticas teóricas, determinará la precisión con que identificamos los píxeles con más o menos probabilidad de contener señal determinista y con ello los resultados de la segmentación.

4.1.1 Descripción del problema

El Grid teórico se deriva de manera teórica pero en base a parámetros que son estimados del espectrograma; de forma que la precisión en la estimación de estos parámetros condicionará la corrección de la posición. Estos parámetros estimados del espectrograma ruidoso son los siguientes:

- σ^2: potencia de ruido que corrompe la señal determinista. Este parámetro se determina a partir de la estimación de la distribución gamma que sigue la Característica 1 en un espectrograma de un proceso WGN. La estimación de dicha distribución en la primera iteración se basa en todos los píxeles del espectrograma entrando a formar parte los píxeles con características deterministas que disminuyen la precisión de la estimación y con ello se desvirtúa la posición del Grid teórico.

- r: relación señal a ruido de la célula. La expresión de este parámetro es $r = S/\sigma^2$ donde S el valor total de la señal determinista en una célula y σ^2 la potencia de ruido. Teóricamente los valores que se le dan a r deben cubrir todas las SNR locales que existan en el espectrograma pero de manera de forma aproximada estos valores se estiman con un rango que va desde 0 hasta el máximo valor para los píxeles del espectrograma ruidoso. Estos valores obviamente distan de ser los reales para las distintas SNR en el espectrograma por lo que el Grid teórico no se localizará en la posición teórica exacta.

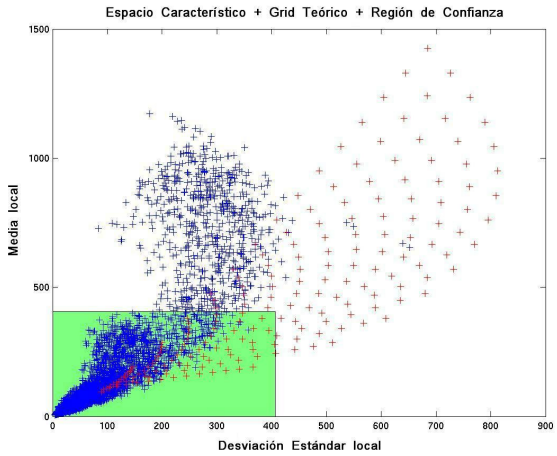


Figura 4.1 Desajuste de la posición del Grid teórico respecto del Espacio Característico

Esto hace que la superposición entre los puntos del Espacio Característico y el Grid teórico no sea completamente perfecta como se puede comprobar en la figura 4.1, lo que hace que píxeles con altas probabilidades de contener señal determinista debido a su posición en el Espacio Característico no sean identificados como tales y el proceso de segmentación pierda efectividad.

La figura 4.1 muestra la distribución del Espacio Característico para un espectrograma de chirps inmersos en WGN, que es en lo que se centran los estudios de [1] y [3]. En el caso de señales vocales el problema se agrava aún más, ya que para mejorar las prestaciones del proceso de segmentación es necesario realizar un suavizado de la imagen espectrograma de tal manera que se eliminen las variaciones debidas a la frecuencia fundamental como se describe en el apartado 4.2.1 Esto hace que cambien ligeramente las características estadísticas del espectrograma haciendo que la distribución de los puntos del Espacio Característico se desplace ligeramente hacia valores menores de la Característica 2. El resultado es un desalineamiento crítico con el Grid teórico que puede desembocar en un fallo en la ejecución del algoritmo al no encontrar ninguna semilla que propagar como ocurre en la situación que se muestra en la figura 4.2

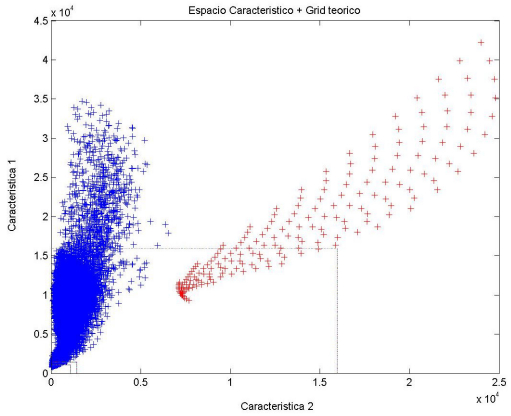


Figura 4.2 Espacio Característico y Grid teórico de señal vocal inmersa en WGN a –5dB

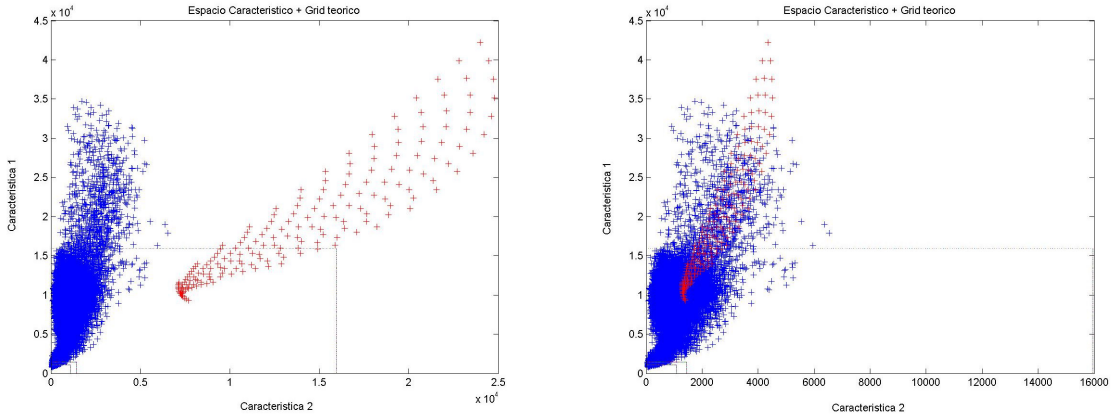
4.1.2 Ajuste de la posición

Una vez descrito el problema vemos que es necesario tomar algún tipo de medida para evitar la pérdida de efectividad del algoritmo, sobre todo al tratar con señales vocales. Lo primero que se puede pensar es en una mejor manera de estimar los parámetros en los que se basa la derivación del Grid teórico mediante una modificación del modelo estadístico en el que se basan que tenga en consideración el suavizado que se realiza en los espectrogramas de señales vocales.

Este enfoque es bastante complejo y además sufriría de la pobre estimación inicial de la potencia de ruido y la SNR de la célula por lo que tendría además que contar con reajustes posteriores basados en la experiencia.

4.1.2.1 Ajuste utilizado en [3]

Un enfoque más eficiente es reajustar la posición del Grid teórico a posteriori teniendo en cuenta la posición en la que se encuentran los puntos del Espacio Característico. Este enfoque es el que se lleva a cabo en [3] cuando se lleva a cabo el experimento con señales vocales, donde queda demostrada la efectividad en cuanto a resultados de segmentación. El ajuste que se lleva a cabo es tan simple como multiplicar por un coeficiente (entre 0 y 1) los valores de $E\{F_2\}$, lo que hace que el Grid teórico se redistribuya hacia valores menores de esta característica consiguiendo un mejor solape con el Espacio Característico. El resultado de dicho ajuste se puede comprobar en la figura 4.3.



a b

Figura 4.3 Espacio Característico y Grid teórico (a) antes y (b) después del ajuste realizado en [3]

Observamos que el solapamiento conseguido es realmente bueno y así lo demuestran también los resultados de la segmentación. Pero el problema de este ajuste es que el coeficiente por el que debemos multiplicar $E\{F_2\}$ es distinto dependiendo de la SNR de la señal de la que se deriva el espectrograma. En los experimentos de [3] se utiliza un coeficiente diferente para cada una de las 7 SNR por lo que el algoritmo deja de ser completamente independiente de la señal de entrada y es necesario un conocimiento a priori del ruido que esta contaminando la señal. Para evitar esta dependencia se desarrolla en este proyecto el alineamiento que describimos a continuación.

4.1.2.2 Ajuste eficiente

El ajuste que se utiliza en este proyecto es completamente independiente de la SNR de la señal ya que simplemente hace coincidir la posición que ocupan Espacio Característico y Grid teórico recalculando el Grid teórico en base a la localización del Espacio Característico. Se observa que tanto la inclinación del Grid como la posición relativa es errónea por lo que el ajuste consistirá en una rotación del Grid y a continuación un desplazamiento hacia la posición adecuada.

En primer lugar se lleva a cabo la rotación del Grid teórico para conseguir el mismo ángulo de inclinación que tiene la distribución de puntos del Espacio Característico. Calculando las rectas de regresión que se ajustan al Espacio Característico y al Grid teórico podemos determinar el ángulo de rotación necesario.

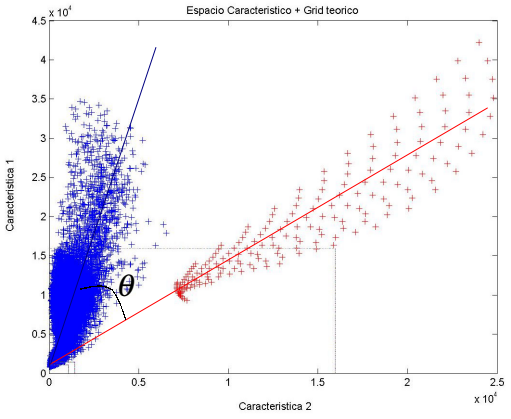


Figura 4.4 Determinación del ángulo de rotación del Grid teórico

Con este ángulo calculamos una matriz de rotación que multiplicamos por la matriz que forman los puntos del Grid teórico para conseguir el cambio de inclinación deseado.

$$[E\{F_1\}, E\{F_2\}] = [E\{F_1\}, E\{F_2\}] \cdot \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

La mayoría de los puntos del Espacio Característico se corresponden con píxeles de ruido y se encuentran concentrados en un punto tal y como muestra la figura 4.5. Este punto es el que debe ocupar el punto del Grid teórico con p=s=0 ya que es el que indica dónde se localizan los puntos de ruido.

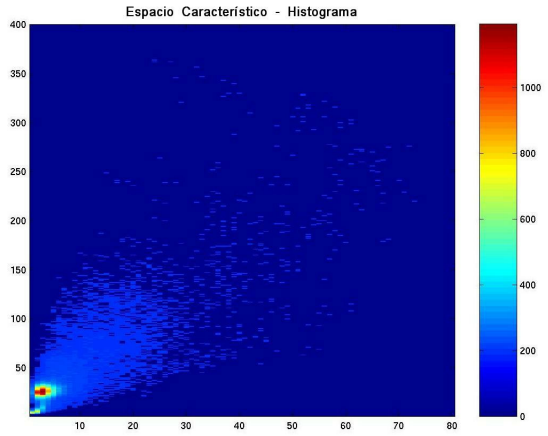


Figura 4.5 Histograma del Espacio Característico de espectrograma de voz inmersa en ruido mostrando la densidad de puntos en cada localización.

Como es donde existe más concentración de puntos, dicha localización quedará determinada calculando el centro de masas de la Característica 1 y Característica 2. A continuación se determinan las distancias de dichos centros de masas al primer punto de $E\{F_1\}$ y $E\{F_2\}$ (p=s=0) lo que nos mostrará la cantidad que debemos restar a cada punto del Grid teórico para conseguir el desplazamiento horizontal y vertical deseado.

En la figura 4.6 podemos ver el espectrograma de una señal limpia y en la figura 4.7 el estudio de esa misma señal inmersa en WGN a 10dB. La primera fila de la tabla muestra los Espacios Característicos y los Grid teóricos con y sin realizar el ajuste. La segunda fila muestra los resultados de la segmentación para ambos casos. Como se puede observar el ajuste hace que el Grid

teórico se solape perfectamente con los puntos del Espacio Característico de forma que las semillas se encontrarán de manera más rápida y eficiente. Esto implicará por una parte un menor número de iteraciones ya que en la primera de ellas se extrae un número mucho mayor de señal lo que contribuye a mejores estimaciones de los parámetros de ruido y mejor estimación de los límites de propagación, lo que acorta el proceso considerablemente. Por otra que se encuentra un mayor número de puntos pertenecientes a señal determinista por lo que los resultados de la segmentación mejoran, como se puede comprobar en la figura 4.7 (d)

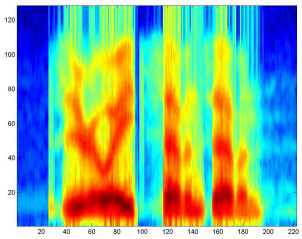


Figura 4.6 Espectrograma de señal en condiciones limpias

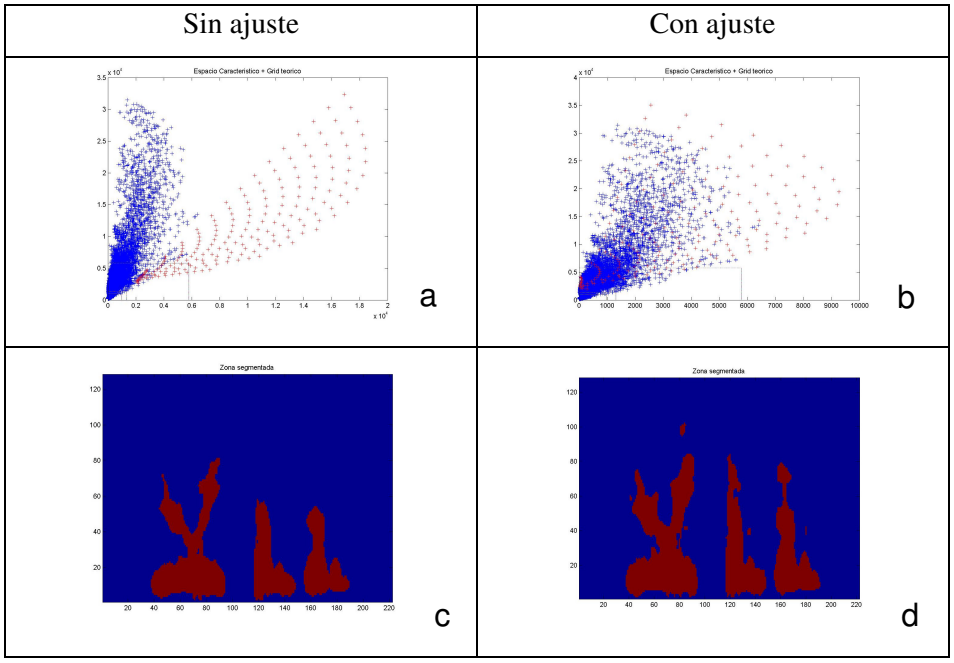


Figura 4.7 (a) Espacio Característico y Grid teórico para espectrograma de señal de figura 4.6 inmersa en WGN a 10dB sin ajuste del Grid y (b) con ajuste. (c) y (d) muestran los resultados del proceso de segmentación para ambos casos.

4.2 RECUPERACIÓN DE AMPLITUDES ESPECTRALES

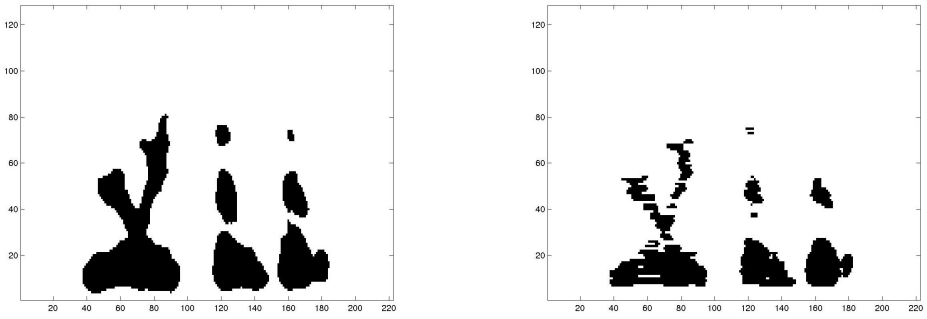
En los experimentos llevados a cabo en [3] la onda temporal que se envía al reconocedor, tanto en el periodo de entrenamiento como en el de test, es generada a partir de un espectrograma muy simplificado en el que no existen variaciones de las amplitudes espectrales. Las zonas del espectrograma donde el algoritmo de segmentación indica que existe señal vocal se ponen a un valor alto (el máximo para el rango de representación de dicho valor) y a las zonas donde el algoritmo considera que solo existe ruido se les da el valor cero. El espectrograma resultante sólo tienen entonces dos posibles valores de amplitud, preservándose de esta forma sólo la información de las formas espectrales de la señal vocal. Haciendo esto se pierde toda información sobre los valores de magnitud del espectro y en principio es perjudicial para un posterior proceso de reconocimiento de voz. Pero las variaciones de magnitud están contaminadas por el ruido de manera que, si no se incluyen, se elimina también la influencia del mismo (aparte de la anterior corrupción de las formas espectrales) y se tiene un algoritmo más robusto. De esta forma, en [3] se lleva a cabo un entrenamiento del reconocedor con señales temporales obtenidas a partir de formas espectrales exclusivamente y posteriormente se comprueba la respuesta del mismo ante ese mismo tipo de señales, donde las formas espectrales se han obtenido a partir de espectrogramas

contaminados con diferentes niveles de ruido. Como se muestra en el apartado 3.1.5 lo resultados obtenidos son realmente buenos poniendo de manifiesto la importancia de las formas espectrales en el contexto del reconocimiento de voz, y dando como resultado un algoritmo que hace el reconocedor mucho más robusto frente al ruido.

A pesar de las notables mejoras en el reconocimiento de voz frente a la línea base (sin tratamiento de ningún tipo) las ondas temporales generadas con ese tipo de espectrogramas son del todo antinaturales, de manera que para un oído humano son completamente ininteligibles. Así, las variaciones en las amplitudes espectrales son críticas para un correcto reconocimiento por parte del oído humano de manera que el siguiente paso en el desarrollo del filtrado morfológico aplicado a señales vocales es incluir estas variaciones de alguna manera para comprobar el comportamiento del reconocedor de voz y poder determinar de manera objetiva la importancia relativa de las formas espectrales. Como se ha indicado anteriormente, las variaciones espectrales que tenemos disponibles están afectadas por el ruido que corrompe la señal de forma que, tras comprobar el comportamiento del reconocedor ante estas amplitudes degradadas, en el capítulo 5 se llevará a cabo un estudio de la posible compensación del ruido en las áreas de señal vocal mediante substracción espectral.

4.2.1 Obtención de amplitudes espectrales

El espectrograma que se computa al comienzo del proceso como base para llevar a cabo todo el filtrado morfológico se genera eliminando cierta información que no es necesaria para el proceso de segmentación como se comenta en el apartado 3.1.4.1. El espectrograma así calculado tiene unas variaciones de amplitud mucho más suaves que hacen que el algoritmo trabaje de forma mucho más eficiente de forma que se lleva a cabo una mejor detección de las áreas que contienen señal vocal . Como ejemplo gráfico se muestra en la figura 4.8 los resultados de la segmentación del mismo espectrograma generado con y sin el proceso de suavizado, comprobándose como la información de la frecuencia fundamental presente en el espectrograma sin tratar hace que se degrade el comportamiento de la segmentación, no detectando de manera adecuada las formas espectrales y produciéndose discontinuidades debido a las bandas tan marcadas que produce la frecuencia fundamental.



a b

Figura 4.8 Resultados de segmentación de espectrograma suavizado (a) y sin suavizar (b)

Así, el espectrograma tiene buenas características para el proceso de segmentación, pero la información de las variaciones espectrales está degradada por el suavizado. Las variaciones espectrales que deberían ser utilizadas tienen que conservar toda la información disponible en la onda temporal original; por ello, el nuevo algoritmo que incluye las variaciones de las amplitudes espectrales genera otro espectrograma sin ningún tipo de tratamiento adicional que se comprueba capaz de reconstruir la señal temporal original de manera exacta. Ese nuevo espectrograma es el que sirve de base para el proceso de inclusión de las amplitudes mediante aplicación de una máscara que se describe en el siguiente apartado.

4.2.2 Inclusión mediante aplicación de máscara

El proceso de segmentación original nos da como resultado una matriz de las mismas dimensiones que el espectrograma en donde se indican con valor ‘1’ las zonas que el algoritmo considera que contienen señal vocal y con valor ‘0’ las zonas consideradas como ruido. Esta matriz servirá entonces como una máscara que se multiplica píxel a píxel por el nuevo espectrograma generado sin ningún tipo de tratamiento adicional. El resultado obtenido es un espectrograma con los valores originales de variaciones espectrales en las zonas que contienen señal vocal (resultado de multiplicar por ‘1’) y silencio en las zonas consideradas como ruido (resultado de la multiplicación por ‘0’). A continuación se muestra gráficamente el proceso:

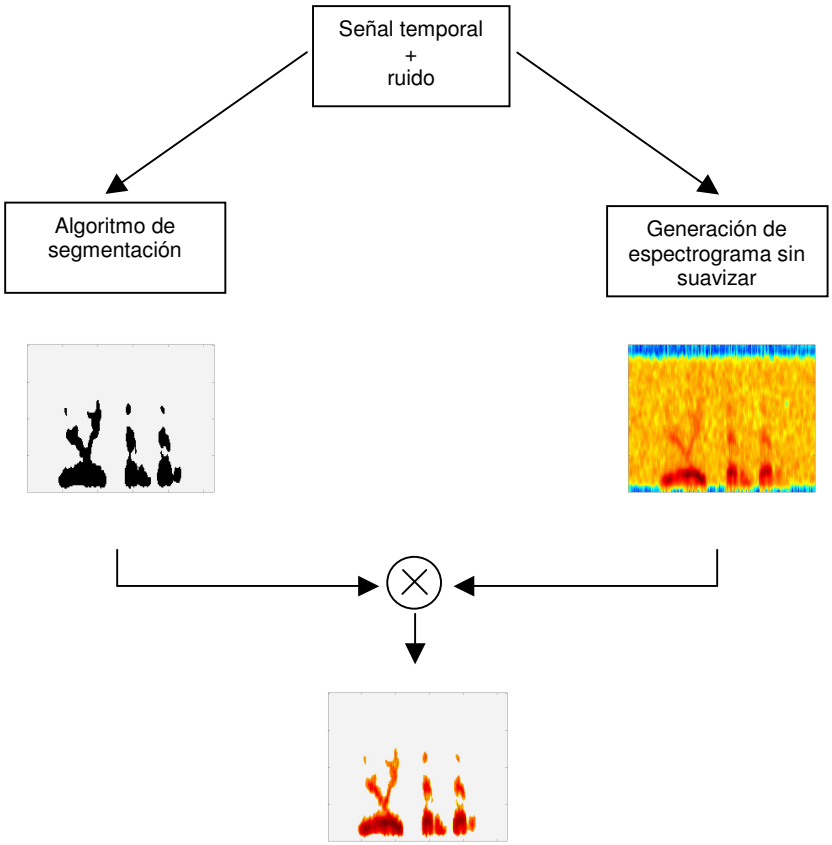


Figura 4.9 Proceso de aplicación de la máscara para recuperar las variaciones de amplitud espectrales

Finalmente, a partir de este espectrograma en el que se ha llevado a cabo una 'limpieza' de las zonas donde sólo existe ruido, se regenera la onda temporal que se enviará al reconocedor de voz.

Este procedimiento da pie al tratamiento de la máscara resultado del proceso de segmentación para dar mejores características a la señal temporal resultante. En principio la máscara tiene simplemente los valores '1' y '0' para las zonas consideradas como voz y como ruido respectivamente. El nivel alto no importa demasiado ya que al regenerar la onda temporal se lleva a cabo un proceso de normalización que da lugar siempre a la misma potencia; pero el nivel base es susceptible de optimización ya que recomendaciones sobre el uso de reconocedores de voz y resultados experimentales muestran que no es bueno utilizar señales cuyos espectrogramas tienen

valores nulos, sino que siempre es bueno tener un cierto valor base. Esto viene apoyado por el hecho de que un espectrograma de una señal limpia nunca llega a valores nulos incluso en las zonas de silencio, por lo que ceros en el espectrograma significan anti-naturalidad y alejamiento de las características ideales que queremos alcanzar. Por ello, en lugar de utilizar '0' como nivel base de la máscara se utilizará un valor pequeño que al multiplicar por el ruido llevará a cabo una fuerte atenuación pero sin crear valores nulos. Por otra parte, los cambios en los valores de amplitud en un espectrograma nunca son muy fuertes, sino que siguen curvas suaves, por lo que también se realiza un suavizado en la transición del nivel alto al nivel base de la máscara con el objetivo de amortiguar las transiciones y evitar posibles cambios bruscos en el espectrograma resultante. La optimización del nivel base y el proceso de suavizado de la máscara se describen en los apartados siguientes.

4.2.3 Optimización del nivel base

El nivel base que se utilice en la máscara determinará el nivel de limpieza de ruido ya que los valores pertenecientes a las zonas de ruido serán más o menos atenuados al multiplicarse por dicho número. Como se ha comentado anteriormente no es bueno realizar una limpieza total dejando los valores del espectrograma a cero, sino que ha de utilizarse un valor que atenúe de manera optima el ruido. En este apartado se muestran los resultados de una serie de experimentos llevados a cabo para ver en qué medida afecta la elección de dicho nivel base y para llevar a cabo la elección que optimice el comportamiento del reconocedor de voz.

Los experimentos realizados se llevan a cabo con utilizando la base de datos Aurora 2 utilizando la configuración habitual pero reduciendo el número de archivos de entrenamiento a 1000 y los de cada categoría de test a 125. Esta agilización del proceso dará como resultado un empeoramiento de los valores de exactitud del reconocimiento debido a la disminución de los archivos de entrenamiento, pero no afectará al objetivo del mismo ya que lo que nos interesan son valores relativos que determinen el nivel base a utilizar. Se realizan 8 experimentos diferentes con 8 niveles base cubriendo un rango que va desde 0 a 0.5. En cada experimento tanto los archivos limpios de entrenamiento como los 7 grupos de archivos de test (limpio, 20dB, 15dB 10dB, 5dB, 0dB y –5dB) se tratan con el filtrado morfológico utilizando el nivel base correspondiente. Los resultados que obtenemos de exactitud en el reconocimiento de voz se muestran en la Tabla 4.1:

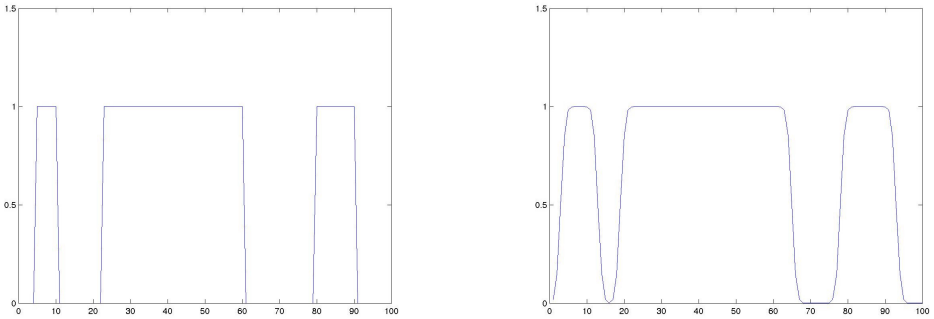
N. base	Limpio	20dB	15dB	10dB	5dB	0dB	-5dB
0	79.17	68.86	56.42	50.12	33.27	15.32	7.57
0.0001	79.59	69.04	56.42	50.92	35.55	15.37	7.57
0.001	79.59	69.04	56.42	50.92	35.55	15.37	7.57
0.01	79.59	69.04	56.42	50.92	35.55	15.37	7.57
0.05	79.59	69.04	56.42	50.92	35.55	15.37	7.57
0.1	79.59	69.04	56.42	50.92	35.55	15.37	7.57
0.25	79.92	70.34	58.03	48.39	28.90	11.70	7.11
0.5	80.50	70.34	57.34	46.10	20.18	10.55	6.72

Tabla 4.1 Optimización del nivel base de la máscara

Como se puede comprobar por los resultados obtenidos, el nivel base no tiene una gran importancia salvo en circunstancias extremas. En primer lugar comprobamos que los resultados son en general ligeramente mejores cuando se utiliza un cierto nivel en lugar de dejar la zona de ruido a cero, pero dichos resultados no cambian en el rango que va de 0.0001 a 0.1. Si se utilizan niveles base más elevados como 0.25 o 0.5 se observa una ligera mejora para las relaciones señal a ruido altas. Esto se debe a que las zonas segmentadas nunca son perfectas y siempre hay cierta parte que contiene señal que no es detectada por el algoritmo (en la sección 4.3 se optimizan los parámetros del algoritmo para minimizar este problema). Cuando esas zonas de señal no son fuertemente atenuadas utilizando un nivel base alto, su contribución hace que el comportamiento del reconocedor mejore. Pero cuando tenemos un nivel de ruido considerable, la no atenuación del mismo de una manera apropiada hace que el comportamiento decaiga considerablemente, por lo que estos niveles base no son aconsejables. De esta forma se puede elegir para el nivel base cualquier valor dentro del intervalo 0.1 a 0.0001, eligiendo en el desarrollo de este proyecto el valor 0.001 para todos los experimentos posteriores.

4.2.4 Suavizado de los bordes de la máscara

En un espectrograma de una señal limpia los cambios que se producen en la magnitud son siempre suaves, creciendo o decreciendo progresivamente para ir creando los diferentes formantes. El proceso de aplicación de la máscara con sólo dos niveles hace que los píxeles contiguos en el espectrograma que coincidan con los bordes de la máscara sean multiplicados por un valor 1000 veces diferente. Dependiendo de la zona donde el algoritmo haya establecido el límite de lo que considera señal, la transición originada en los valores resultantes será más o menos abrupta, lo que se deben evitar a toda costa ya que este tipo de cambios no existen en los espectrogramas de señales vocales. Para evitar este efecto en la aplicación de la máscara se realiza un suavizado de los bordes de manera que sigan un patrón de decrecimiento gaussiano.



a b

Figura 4.10 Perfil de la máscara antes (a) y después del suavizado (b)

Así, los valores de las amplitudes espectrales se multiplicarán por '1' en la zona que el algoritmo clasifica como conteniendo señal, pero en la zona inmediatamente contigua no se pasará directamente al nivel base, sino que se irá atenuando progresivamente de manera que se elimina toda transición brusca de valores hasta llegar al nivel base. Además, de esta forma las zonas que todavía contienen señal que se encuentran alrededor del área segmentada y que no son detectadas por el algoritmo no serán completamente atenuadas al nivel de ruido, por lo que podrán contribuir con su información a mejorar el comportamiento del reconocedor de voz.

A continuación observamos en la figura 4.11 una máscara real resultado del proceso de segmentación de un espectrograma de voz inmersa en WGN antes y después del suavizado, y la

manera en la que afecta a la onda temporal. Vemos que en el caso sin suavizar, el anormal salto en los valores de magnitud espectral hace que las muestras temporales se comporten de manera brusca en esas zonas. En el caso suavizado se evita todo tipo de comportamiento anormal acercándonos más al perfil ideal de la señal de voz limpia.

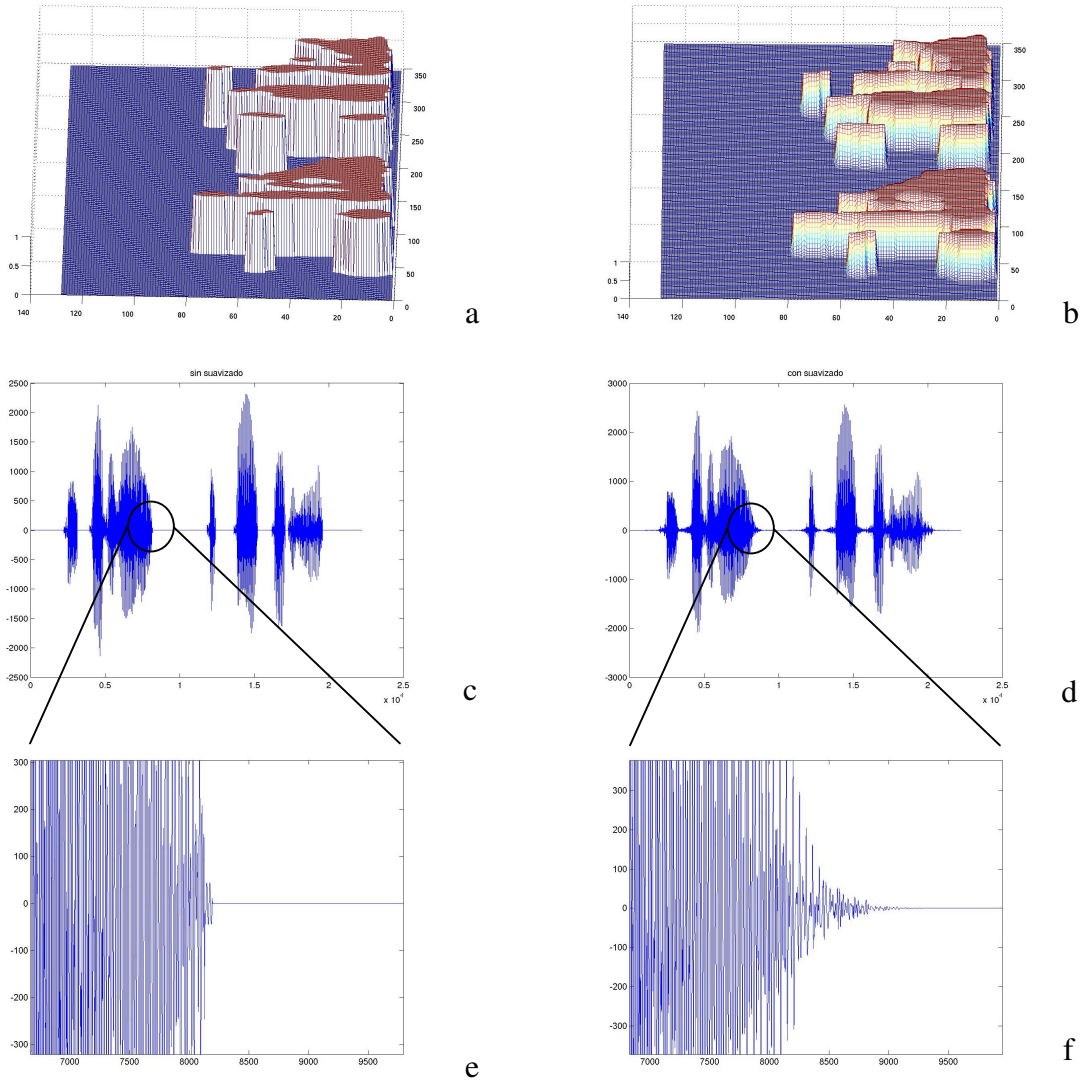
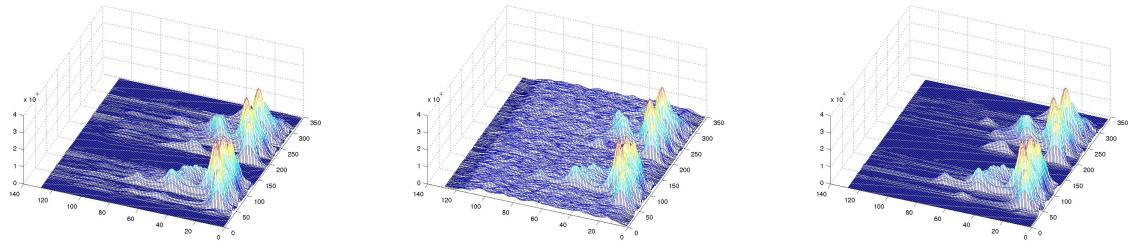


Figura 4.11 Máscara sin suavizado (a) y tras el proceso de suavizado (b). Ondas temporales correspondientes tras la multiplicación (c) y (d) y zoom en discontinuidad debido a transición de ambas ondas (e) y (f).

4.2.5 Experimentos y resultados

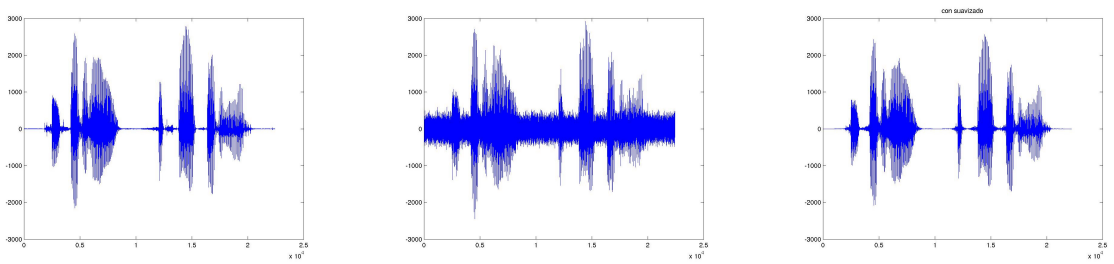
El proceso completo hasta el momento consiste entonces en la segmentación del espectrograma de voz suavizado para obtener una matriz que indica las zonas que se considera que contienen señal vocal. A continuación se crea una máscara a partir de la anterior matriz donde se ajusta un nivel base (‘0.001’) para las zonas consideradas de ruido y se crea un patrón de transición suave entre ese nivel base y el nivel alto (‘1’) que se le da a las áreas de señal. Una vez aquí se genera un espectrograma que sea completamente reversible y se multiplica píxel a píxel por la máscara creando un nuevo espectrograma donde las zonas ruidosas están fuertemente atenuadas y las zonas de señal permanecen inalteradas. En la figura 4.12 se pueden observar los espectrogramas en tres dimensiones de una señal limpia en (a), la misma señal inmersa en ruido WGN a 5 dB en (b) y el resultado de todo el proceso descrito en (c). Como puede apreciarse, las zonas de señal en (c) guardan completa relación con el espectrograma original mientras que el área alrededor donde sólo existe ruido tiene un nivel notablemente menor que esa misma área en el espectrograma (b).



a b c

Figura 4.12 Vista tridimensional de espectrograma original limpio (a), espectrograma de señal a 5dB (b) y resultado del proceso de segmentación con la inclusión de lasa variaciones espectrales (c)

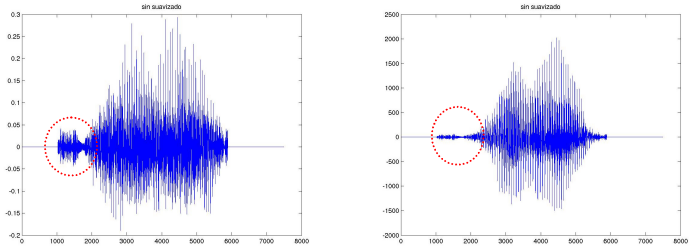
La figura 4.13 muestra las ondas temporales (a), (b) y (c) correspondientes a los tres espectrogramas que se muestran en la figura 4.12. En ellas se observa cómo el proceso limpia el ruido de todas las zonas donde no existe señal vocal y se puede comprobar además la similitud entre las zonas de voz de la señal limpia original (a) y la recuperada (c) a partir de la señal degradada de (b).



a b c

Figura 4.13 Señales temporales correspondientes a los espectrogramas de la figura 4.12

Una ventaja adicional que se consigue al incluir las variaciones espectrales frente a la configuración de [3], donde sólo se utiliza la forma espectral, es que las zonas erróneamente clasificadas como señal vocal, siendo zonas de ruido, no afectan de una manera tan negativa. En [3] se le da a toda la zona clasificada la misma amplitud, de forma que las zonas de ruido erróneamente segmentadas tendrán exactamente el mismo valor que las zonas de señal vocal y su contribución afectará decisivamente al comportamiento del reconocedor de voz. En el nuevo esquema, si existen zonas erróneamente segmentadas, su importancia decrece en gran medida cuando se tienen en cuenta sus valores de amplitud, que estarán a nivel de ruido. En la figura 4.14 se puede observar el resultado de volver al dominio temporal del mismo espectrograma sin incluir los valores de amplitud (a) e incluyéndolos (b). Se puede observar como la zona erróneamente clasificada como señal vocal justo antes del comienzo real pierde la importancia que tiene en (a) cuando se incluyen las variaciones de las magnitudes espectrales (b).



a b

Figura 4.14 Ondas temporales con una zona erróneamente segmentada para el esquema sin y con amplitudes espectrales en (a) y (b) respectivamente

Para obtener una medida objetiva de la importancia de la inclusión de las variaciones de amplitud espectral se realiza el mismo experimento que el llevado a cabo en [3] (al que se hace

referencia en el apartado 3.1.5 y al principio de este capítulo) pero extendiendo el proceso como se ha descrito en esta sección. El experimento utiliza la base de datos Aurora 2 y la configuración experimental típica descrita en la sección 2.5 tal y como se hace en [3]. El entrenamiento se lleva a cabo con 8440 archivos de señal limpia tratados con el filtrado morfológico y el test se realiza con 7 conjuntos de 1001 archivos contaminados con diferentes niveles de ruido y tras ser procesados por el mismo algoritmo de filtrado. Los resultados comparativos entre la línea base en azul, el perfil obtenido en [3] con sólo la forma espectral en verde y los resultados obtenidos en este experimento en rojo se muestran en la siguiente figura.

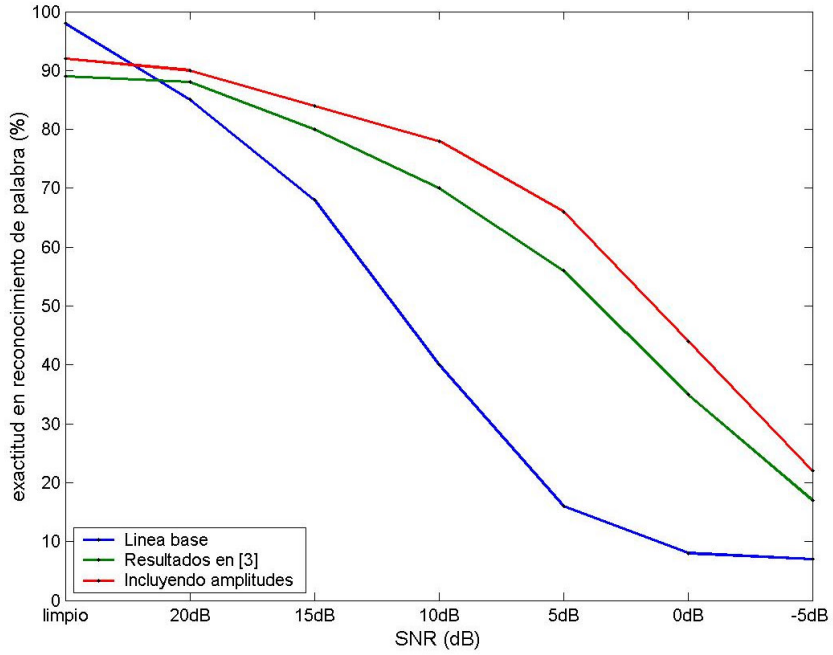


Figura 4.15 Resultados de exactitud en el reconocimiento para la línea base en azul, morfológicamente procesado utilizando solo la forma espectral en verde y con las amplitudes espectrales en rojo.

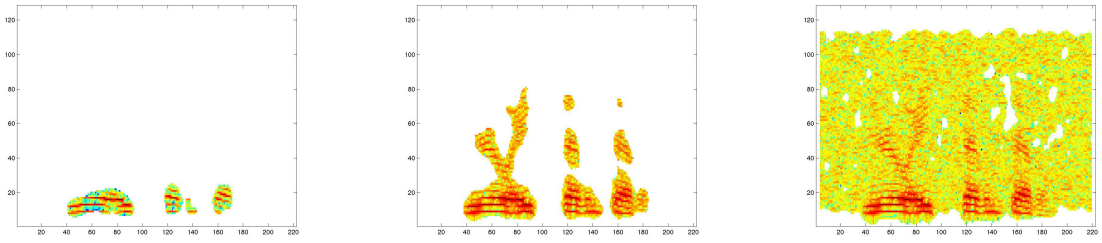
La inclusión de la información de magnitudes espectrales hace que los resultados mejoren para todas las relaciones señal a ruido a pesar de que éstas están degradadas por el ruido. En el caso limpio vemos como el comportamiento se acerca más a la línea base aunque no se llega al mismo resultado. Esto pone de manifiesto la no completa exactitud de las formas espectrales detectadas por el algoritmo, ya que si se segmentara de manera ideal el resultado para el caso limpio y la línea base deberían teóricamente coincidir. Así se motiva la optimización, que se lleva a cabo en el siguiente

apartado, de los parámetros del algoritmo tanto en entrenamiento como en test para conseguir una optimización de las formas espectrales detectadas. Como se ha dicho la inclusión de las amplitudes a través del proceso descrito en este apartado consigue mejorar el comportamiento para todas las SNR llegando a mejoras sobre [3] de hasta el 10% para 5dB y aumentando el rendimiento incluso cuando las amplitudes están fuertemente degradadas en el caso de –5dB en el que se pasa de una exactitud del 17% en [3] al 22 %. El siguiente paso en el tratamiento de las magnitudes espectrales es tratar de compensar el ruido presente en las zonas de señal vocal mediante substracción espectral ya que hasta ahora las magnitudes espectrales utilizadas son las ruidosas que tenemos disponibles en el espectrograma de la señal degradada, sin ningún tipo de tratamiento. El estudio en profundidad de la combinación del Filtrado morfológico y la Substracción espectral se lleva a cabo en el Capítulo 5.

4.3 OPTIMIZACIÓN DE ÁREAS SEGMENTADAS

Como se ha venido comentando a lo largo de las secciones anteriores una detección correcta de las áreas que contienen señal es decisiva para, en el proceso del filtrado morfológico, atenuar sólo las zonas que están completamente dominadas por el ruido y que no aportan ninguna información útil para el reconocedor de voz. En la sección anterior vemos como, en el caso limpio, el resultado del reconocedor cuando procesa señales filtradas todavía esta bastante por debajo de la línea base. Esto nos pone en conocimiento de que las áreas segmentadas por el algoritmo son demasiado pequeñas (al menos para el caso limpio) de forma que se clasifican como ruido zonas que contienen información sobre la señal vocal que mejoran el comportamiento del reconocedor. De esta forma se hace necesaria una optimización del algoritmo que se traduzca en una óptima determinación de las áreas que contienen señal vocal, de manera que no se pierdan altas frecuencias que sean decisivas para la discriminación entre fonos, se detecten de manera correcta las consonantes que se corresponden con los bordes de las áreas segmentadas o se haga una correcta transición de los formantes de un fono a otro, etc. Fundamentalmente, son dos los parámetros que

gobiernan la cantidad de área segmentada dentro del algoritmo; por una parte esta el parámetro que controla la condición de parada de las iteraciones y por otra el que ajusta el umbral que discrimina entre las zonas consideradas ruido o señal vocal. En el siguiente apartado se hace un análisis en detalle de ambos parámetros y de la influencia relativa en la determinación del área segmentada; posteriormente se lleva a cabo el proceso de optimización de los mismos tanto en el caso de las muestras de entrenamiento como en el de las de test.



a b c

Figura 4.16 Modificación de las áreas segmentadas al variar los parámetros del algoritmo

4.3.1 Parámetros que determinan el área segmentada

Como podemos observar en las figuras siguientes, cambios en los parámetros del algoritmo hacen variar de manera radical el área que es clasificada como señal vocal de manera que una elección inadecuada de estos parámetros hace que el algoritmo pierda su efectividad por completo. Como se puede observar, en la figura 4.16 (a) el área segmentada es demasiado pequeña dejando sin clasificar zonas de señal vocal que pueden contener información decisiva para un buen comportamiento del reconocedor de voz y en la figura 4.16 (c) el algoritmo empieza a clasificar zonas de ruido como señal vocal de manera masiva de manera que no limpia de manera adecuada la señal ruidosa y como consecuencia no mejora los resultados de reconocimiento vocal frente a la señal base. A continuación se describen los dos parámetros implicados y su sensibilidad relativa a la hora de la determinación del área segmentada.

4.3.1.1 Condición de parada de las iteraciones

Como se explica en el apartado 3.1.4.4 la condición de parada de las iteraciones se basa en la diferencia entre el límite de propagación computado en una iteración y en la siguiente. De manera que, cuando la estimación de la PDF de ruido tiene una precisión adecuada, el límite de propagación computado no cambia apreciablemente y el algoritmo se detiene considerando que ya no existe más señal determinista que extraer. Teóricamente el parámetro que mide la diferencia entre los límites, llamado *diff*, será en principio el que marque cuando se va a detener el algoritmo y, como consecuencia, el número de iteraciones del proceso y la mayor o menor área segmentada.

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

En el caso de la aplicación del algoritmo a espectrogramas de señales vocales, los límites de propagación van cambiando de manera notable a lo largo de las iteraciones hasta que llega un momento en que se estabilizan de manera relativamente brusca. Tras esa estabilización, la PDF de ruido estimada no cambia de manera apreciable ya que ha alcanzado un gran nivel de exactitud, y la diferencia entre los límites de propagación computados es ínfima. De esta manera cambios en el parámetro *diff*, cuyo valor recomendado en [1] es del 1%, no afectan realmente al número de iteraciones ni al área segmentada salvo casos extremos no aconsejables.

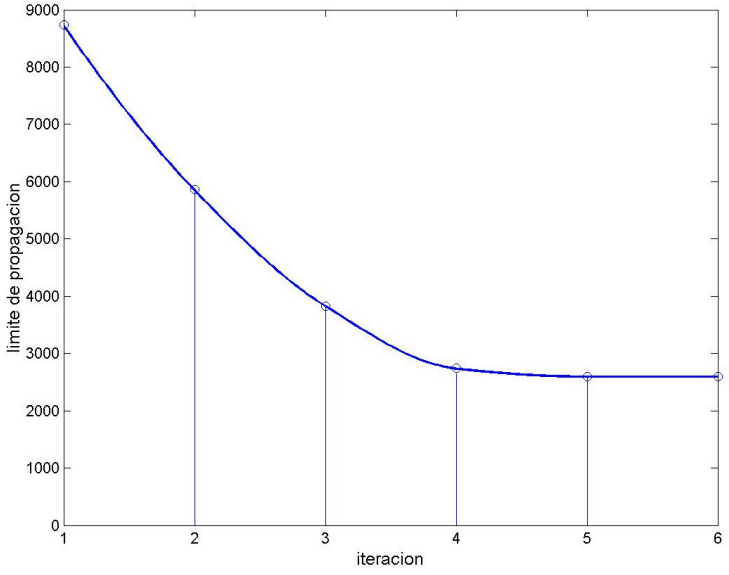


Figura 4.17 Variación del límite de propagación a lo largo de las iteraciones

Si por ejemplo se ajusta al 20% el algoritmo se detendrá antes efectivamente, pero la PDF de ruido no tendrá la suficiente exactitud por lo que el límite final computado no se corresponderá con el valor real de umbral de ruido existente en el espectrograma, determinando un área segmentada errónea. Por el contrario, si se ajusta a valores del orden del 0.001% el algoritmo llevará a cabo más iteraciones, pero sin utilidad alguna ya que la PDF de ruido no cambia y como consecuencia tampoco el límite computado, sin afectar por lo tanto al área segmentada.

El parámetro *diff* lo que controla realmente es que se haya alcanzado la precisión adecuada en la estimación de la PDF de ruido, y esto ocurre en el caso de señales vocales para un amplio rango de valores de dicho parámetro. El 1% aconsejado por Hory en [1] es completamente válido para la función que tiene en el algoritmo por lo que será el que se utilice al aplicarlo en este proyecto a las señales vocales.

4.3.1.2 Probabilidad de error

El proceso del filtrado morfológico se basa en la estimación de la PDF de ruido que obtiene en cada iteración del algoritmo. Cuando el proceso realiza esta estimación a partir de la parte del espectrograma que quede sin segmentar en ese momento concreto, se asume que se esta cometiendo cierto error en dicha estimación ya que los datos en los que se basa no son exclusivamente de ruido, sino que siempre tenemos cierta parte de señal determinista que corrompe la estimación. Además, en la señal vocal también existen zonas de baja energía que tienen características estadísticas similares a un proceso WGN, de forma que no se puede asegurar que la estimación de la PDF de ruido esta determinando con toda exactitud cómo se distribuyen los píxeles de ruido. Contando con un error que se ajusta mediante el parámetro P_e, el algoritmo determina entonces un límite en la PDF a partir del cual confía en que todos los puntos bajo el mismo serán de ruido y todos los puntos por encima contendrán señal determinista. Este límite se determina encontrando el punto en la PDF a partir del cual el área bajo la curva es un tanto por ciento concreto del área total; es decir, se establece una probabilidad de error en la estimación. El tanto por ciento aconsejado por Hory cuando aplica el algoritmo a señales sintéticas en [1] es del 1%, lo que revela la confianza de Hory en la precisión de las estimaciones para este tipo de señales. La figura 4.18 muestra gráficamente lo que acabamos de comentar.

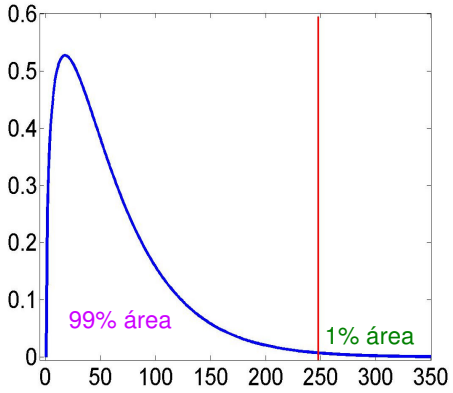
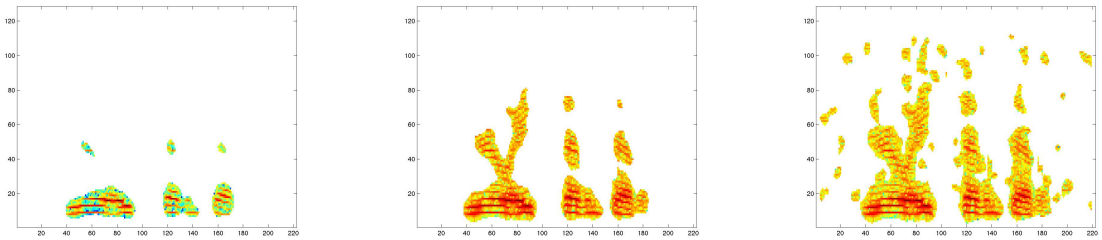


Figura 4.18 Determinación del límite de propagación a partir de una probabilidad de error en la PDF de ruido.

De esta manera el tanto por ciento que se elija para la probabilidad de error determinará el límite de propagación y con ello el Área de trabajo y la Región de confianza de ruido. A lo largo de las iteraciones, cuando la estimación de la PDF de ruido no goza de gran exactitud, el valor para P_e afecta acelerando el proceso cuanto mayor es el valor asignado a este parámetro ya que se extrae una mayor cantidad de señal en cada iteración. Pero es en la última iteración cuando P_e afecta de una manera decisiva, estableciendo la última Área de trabajo y dando la posibilidad de extender la propagación a más o menos píxeles del espectrograma y con ello realizar una segmentación de más o menos área. La figura 4.19 muestra como dependiendo de la P_e elegida el área segmentada cambia en gran medida.



P_e = 1% P_e = 10% P_e = 20%

Figura 4.19 Variación del área segmentada en función de la probabilidad de error usada en la PDF de ruido.

De esta forma, P_e será el parámetro que deberá optimizarse para llevar a cabo una segmentación óptima que extraiga toda el área que aporte información al proceso de reconocimiento. Por lo tanto, el proceso de optimización que llevaremos a cabo a continuación tiene como objetivo determinar cual es la mejor P_e desde el punto de vista del comportamiento del reconocedor de voz. Esto incluye determinar cual es la óptima P_e a utilizar con las muestras limpias del entrenamiento; es decir, con qué cantidad de información es mejor entrenar al reconocedor, teniendo en cuenta que posteriormente va a ser puesto a prueba con señales segmentadas que pierden más o menos información dependiendo de la SNR. Así, se ajustará el algoritmo para se segmente, tanto en entrenamiento como en test, el área en los espectrogramas qué de lugar a una máxima precisión en el reconocimiento automático de palabra.

4.3.2 Optimización de los parámetros

Utilizar una P_e óptima se traduce en llevar a cabo una óptima distinción entre el Área de trabajo y la Región de confianza de ruido; es decir, en ajustar el mejor límite de propagación tanto para las muestras de entrenamiento como para cada uno de los distintos casos de SNR bajo test. De esta manera, en el proceso de optimización los límites de propagación serán ajustados de manera manual de forma que se cubran todos los rangos y posibilidades existentes, tanto para el entrenamiento como para el test. Así podremos identificar cuál es la mejor combinación de estos límites de propagación, que dan lugar a distintas áreas segmentadas, desde el punto de vista del comportamiento del reconocedor de voz. Ajustando de manera manual los límites tenemos un control más directo sobre las áreas segmentadas y el algoritmo que utilizamos sólo tiene que ejecutar una iteración ya que estamos definiendo las condiciones finales del algoritmo, por lo que el proceso de optimización es mucho más rápido y eficiente a llevar a cabo pruebas cambiando el parámetro Pe. Además este proceso de identificación de las condiciones finales óptimas, que ajustamos manualmente, dará pie a una modificación del algoritmo para reducir el coste computacional que se explicará con detalle en el apartado 6.1

Una vez identificados los límites óptimos para tratar con cada uno de los casos de SNR se realizará un estudio del algoritmo para ajustar el parámetro Pe, de tal manera que el proceso original

determine de manera automática los mismos límites óptimos que hemos establecido empíricamente. En el caso de las muestras de entrenamiento se revelará un valor óptimo claro para el parámetro Pe, mientras que en las muestras de test este parámetro será ligeramente diferente dependiendo de la SNR concreta, por lo que al final se determinará un valor para P_e óptimo en media que de el mejor comportamiento del reconocedor de voz para todo el rango de posibles valores de SNR.

En principio, el ajuste manual de los límites de propagación se llevaría a cabo determinando un valor concreto para la Característica 1 que daría lugar a un Área de trabajo y una Región de confianza de ruido determinadas. Esto sería la opción acertada siempre y cuando, dentro de una mismo grupo de SNR, la potencia de ruido para todos los archivos de test fuera la misma de forma que ese límite localizaría de manera perfecta el punto bajo el cual sólo tenemos puntos correspondientes a píxeles de ruido, dando como resultado la mejor área posible para ese conjunto de muestras. Sin embargo, dentro de los conjuntos de muestras con la misma SNR la potencia de ruido es diferente ya que el software de la ITU añade más o menos ruido a la muestra dependiendo de la potencia relativa que tiene la señal vocal para dar lugar a una relación señal a ruido determinada. Como consecuencia cada archivo tendrá una potencia de ruido ligeramente diferente dependiendo de la potencia de la señal vocal y el límite óptimo para la Característica 1 será diferente para cada uno de los archivos. Ajustar un límite diferente para cada uno de los archivos es completamente inviable y además no tiene ningún sentido para nuestro proceso de optimización global, así que lo que se hace es ajustar el límite de propagación en relación a la potencia de la señal, tal y como hace el software de adición de ruido. Para cada muestra, después de computar el Espacio Característico se determina el máximo valor para la Característica 1 y se divide por un determinado *'ratio'* para dar lugar a unos límites de propagación ligeramente distintos según la potencia de señal presente (traducida en un valor mayor o menor del máximo para la Característica1).

$$l = \frac{máximo(Característica1)}{ratio}$$

Así, el proceso de optimización consistirá primeramente en determinar el ratio adecuado para cada conjunto de SNR y posteriormente en relacionar ese parámetro con un valor de la P_e adecuado para dar lugar a los mismos límites de propagación y como consecuencia a segmentar una misma cantidad de área en el espectrograma. A continuación pasamos a describir una serie de experimentos llevados a cabo para identificar los ratios adecuados tanto para entrenamiento como

para test y para relacionarlos posteriormente con el parámetro P_e del algoritmo original y obtener los resultados de comportamiento del reconocedor de voz tras optimizar todos los parámetros.

4.3.2.1 Experimento 1: Mejores entrenamientos y límites generales en test

El experimento que se lleva a cabo en primer lugar busca determinar el límite óptimo para utilizar con las muestras de entrenamiento, así como obtener los límites aproximados que se han de utilizar con las muestras de test para cada una de las SNR. Posteriormente se llevarán a cabo experimentos para determinar con mayor exactitud los límites de test, pero utilizando ya las conclusiones sobre el entrenamiento que obtenemos en este apartado.

El entrenamiento es la parte más decisiva a la hora de utilizar un reconocedor de voz. Llevar a cabo un entrenamiento suficientemente extenso y de calidad llevará a resultados muy buenos en la posterior etapa de test, mientras que una reducción en la cantidad de muestras de entrenamiento o la utilización de muestras que no son las apropiadas hace decaer el comportamiento de manera notable.

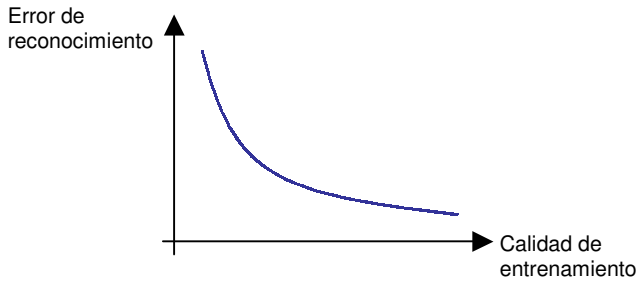


Figura 4.20 Error de reconocimiento frente a calidad del entrenamiento

En este entorno, utilizar un entrenamiento de calidad suele traducirse en entrenar el reconocedor con muestras de voz en las mismas condiciones en las que se va a poner bajo prueba posteriormente, de tal manera que los modelos que se crean en la etapa de entrenamiento reflejarán de manera más precisa los modelos que se crearán durante el test llevando a un mayor número de encuentros y como consecuencia una mayor exactitud en el reconocimiento. De esta manera, si sabemos por ejemplo el tipo de ruido que vamos a tener presente en test, o se conocen

características de algún tipo que van a tener las señales vocales, como en ciertos entornos industriales, cabinas de aviones, etc... puede llevarse a cabo un entrenamiento con ese tipo de señales vocales concretas grabadas en el mismo entorno, consiguiendo así resultados óptimos en la posterior etapa de reconocimiento. Se puede consultar a continuación un ejemplo de resultados de este enfoque donde se prueba una algoritmo de reducción de ruido con entrenamiento limpio y entrenamiento en múltiples condiciones, que consiste en entrenar el reconocedor con muestras de voz contaminadas con los mismos posibles diferentes niveles de ruido con los que luego se hace el test de comportamiento [20]:

SNR (dB)	Entrenamiento limpio	Entrenamiento multicondición
Limpio	99.05	98.63
20	94.45	97.78
15	85.02	97.02
10	63.85	94.68
5	35.76	87.00
0	15.38	57.58
-5	9.21	23.60
Media 0-20	58.89	86.81

Tabla 4.2 Comparación entre entrenamiento limpio y multicondición

Se comprueba que el entrenamiento multicondición es muy útil y da buenos resultados pero en entornos generales como en las comunicaciones móviles, donde la señal vocal esta sometida a todo tipo de ruidos de fondo de diferentes características, realizar un entrenamiento multicondición resulta completamente inviable. De todas formas los resultados mostrados ponen de manifiesto la importancia que tiene entrenar el reconocedor con las muestras de voz adecuadas, ya que esto aumentará notablemente la precisión en el reconocimiento.

En nuestro caso el reconocedor se entrena con muestras limpias pero tras ser filtradas morfológicamente para asemejar los modelos del reconocedor a los que va a encontrar en la etapa

de test. De esta manera, es fundamental encontrar los límites de propagación que hemos de utilizar con las muestras de entrenamiento ya que el comportamiento en la etapa de test será radicalmente diferente si entrenamos con unas áreas más o menos extensas. Por ello, se diseña el siguiente experimento con el que obtendremos el mejor tratamiento que deben recibir las muestras de entrenamiento para llevar a cabo una óptimo reconocimiento de señales filtradas morfológicamente.

4.3.2.1.1 Conjuntos de entrenamiento y de test

Para determinar el mejor conjunto de entrenamiento se pondrán bajo prueba 10 conjuntos de muestras diferentes que cubrirán todo el rango posible de áreas segmentadas. Para ello se definen 10 ratios diferentes que se corresponden con 10 tantos por ciento distintos de área segmentada en la imagen espectrograma de las muestras de entrenamiento. A continuación mostramos los ratios que utilizamos en cada caso y el tanto por ciento medio de área segmentada que se obtiene tras la segmentación:

Ratio	∞	5000	3200	1000	500	200	100	50	20	10
% área segmentada	100	90	80	70	60	50	40	30	20	10

Tabla 4.3 Ratios utilizados en los conjuntos de entrenamiento y % segmentado

Como puede comprobarse, los ratios elegidos siguen una curva exponencial para realizar las segmentaciones a intervalos regulares del 10%. Esto se debe a la característica más o menos exponencial (una función gamma) que tiene la distribución de puntos en el Espacio Característico.

Para determinar el mejor conjunto de entrenamiento tenemos que llevar a cabo la fase de test para ver cual es el que tiene mejor comportamiento. De esta forma se preparan, para cada SNR, 12 conjuntos de muestras segmentadas utilizando 12 ratios diferentes cubriendo todo el espectro de posibles segmentaciones, desde una segmentación de todo el espectrograma completo (ratio=∞) hasta una segmentación mínima (ratio=2). Así podremos determinar la mejor pareja de límites de propagación entrenamiento-test para cada SNR.

4.3.2.1.2 Experimento y resultados

Los experimentos realizados se llevan a cabo con utilizando la base de datos Aurora 2 utilizando la configuración habitual pero reduciendo el número de archivos de entrenamiento en cada categoría a 1000 y los de test a 125. Esta agilización del proceso dará como resultado un empeoramiento de los valores de exactitud del reconocimiento debido a la disminución de los archivos de entrenamiento, pero no afectará al objetivo del mismo ya que lo que nos interesan son valores relativos obtenidos con los distintos límites de propagación. En el experimento siguiente, donde se optimizarán de forma fina los límites de los conjuntos de test, se utilizará la configuración típica de 8440/1000 de manera que servirá para corroborar los resultados obtenidos aquí con los conjuntos reducidos. Esta optimización de límites de entrenamiento / test se lleva a cabo para cada una de las típicas SNR con las que solemos trabajar: limpio, 20sB, 15dB, 10dB, 5dB, 0dB y –5dB. A continuación mostramos los resultados completos obtenidos para el caso de 10dB, cada fila muestra es uno de los entrenamientos bajo test y cada columna muestra los resultados de precisión de reconocimiento para cada uno de los ratios utilizados en las muestras de test.

	2	**4**	**8**	**12**	**16**	**24**	**32**	**48**	**64**	**96**	**128**	
10%	21.56	34.86	42.89	42.66	31.45	27.02	23.47	19.04	18.58	19.04	19.04	18.58
20%	22.12	34.03	43.07	44.16	35.13	28.71	22.94	19.79	20.87	19.43	20.87	20.87
30%	18.81	34.63	43.69	45.77	34.71	28.33	23.12	20.77	20.77	19.14	20.24	20.77
40%	22.14	40.21	52.51	55.13	48.27	44.14	40.65	34.28	33.76	32.91	33.76	32.44
50%	19.43	39.78	52.19	53.07	44.23	40.38	32.14	31.73	31.73	30.29	31.23	30.78
60%	17.29	33.27	37.13	42.85	40.01	37.04	30.43	27.79	27.79	27.68	27.73	27.82
70%	14.85	30.73	36.24	36.24	34.29	32.43	27.14	25.79	25.31	25.25	25.25	25.79
80%	15.34	24.12	30.71	29.25	28.41	26.33	24.39	24.01	24.28	24.17	24.00	24.28
90%	13.97	17.89	24.12	24.00	23.89	23.72	23.49	23.12	23.57	23.24	23.41	23.57
100%	14.20	16.34	18.43	21.12	21.79	22.40	22.84	23.73	23.49	23.97	24.07	23.49

Tabla 4.4 Resultados de optimización entrenamiento / test para 10dB

Gráficamente:

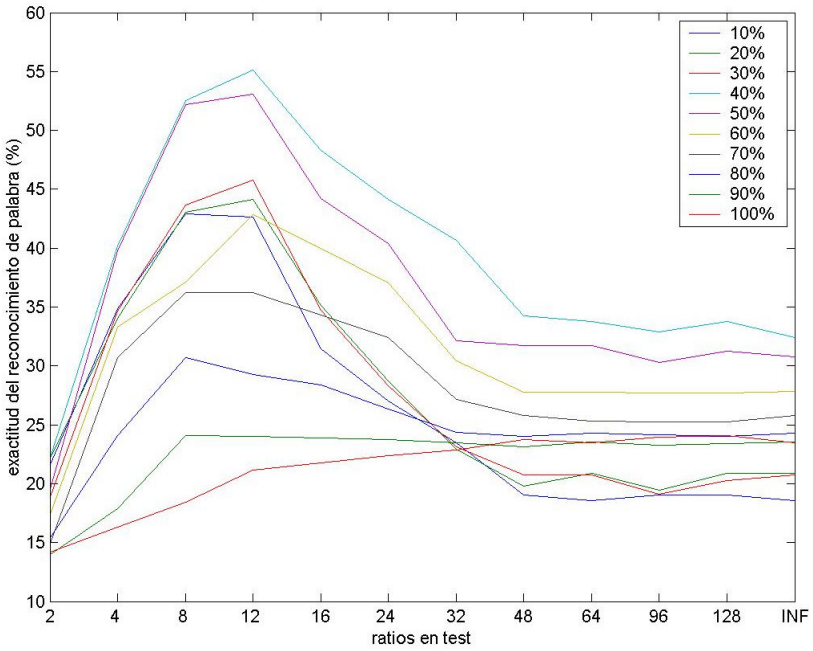


Figura 4.21 Resultados de optimización entrenamiento / test para 10dB

4.3.2.1.3 Conclusiones

En primer lugar observamos que el entrenamiento segmentando el 40% del área, que se corresponde con utilizar un ratio=100, es el mejor independientemente del ratio que se utilice en las muestras de test ya que la curva que traza esta en todo momento por encima de las demás. A continuación los siguientes mejores entrenamientos son los que se hacen con muestras segmentadas un 50% y un 30%, aunque este último esta en ciertas zonas por debajo de otros entrenamientos pero lejos del lugar donde se alcanza el pico de reconocimiento. Este pico de reconocimiento siempre ocupa la zona de los ratios entre 8 y 12 para las muestras de test, que debe ser el umbral a partir del

cual los puntos en el Espacio Característico se corresponden con píxeles de zonas dominadas por ruido. De esta forma, comprobamos que los mejores resultados siempre se obtienen segmentando sólo las zonas donde la señal vocal es dominante. Hacia la derecha de la curva comienzan a segmentarse zonas de ruido y el comportamiento del reconocedor decae para todos los entrenamientos; y hacia la izquierda de las curvas se segmenta un área demasiado pequeña de manera que se comienza a perder información útil haciendo que el comportamiento del reconocedor empeore.

Para las demás SNR bajo estudio las curvas tienen características similares salvo que el pico en el reconocimiento, como es lógico, esta centrado en otros ratios. Una conclusión muy importante es que los mejores entrenamientos siempre son los del 40%, 50% y 30% independientemente de la SNR de las muestras de test, lo que es muy bueno desde el punto de vista de la utilización del algoritmo de manera genérica, ya que en un entorno real se usaría un entrenamiento del reconocedor que resulta óptimo para todas las SNR bajo estudio. A continuación se llevará a cabo una optimización fina de los ratios para cada una de las SNR en las zonas donde tenemos los correspondientes picos en el reconocimiento:

limpio	20dB	15dB	10dB	5dB	0dB	-5dB
96	32	16-24	8-12	4-8	4	2

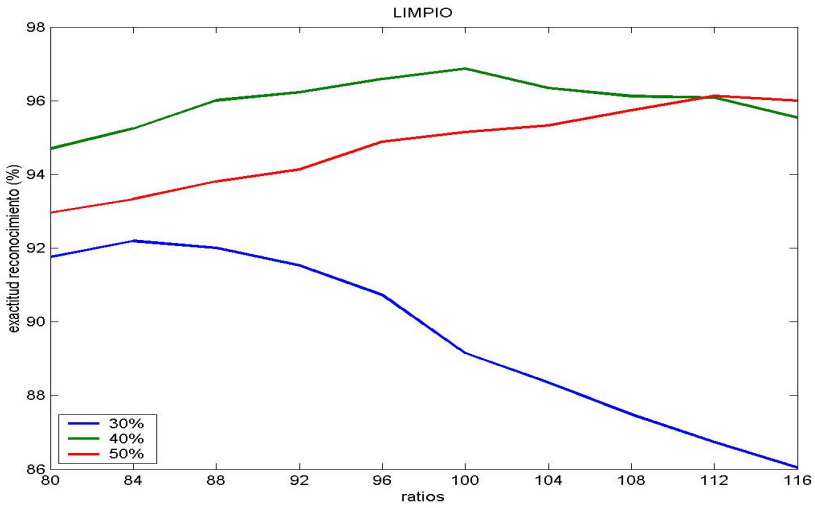
Tabla 4.5 Ratios donde se encuentra el pico de reconocimiento para cada una de las SNR

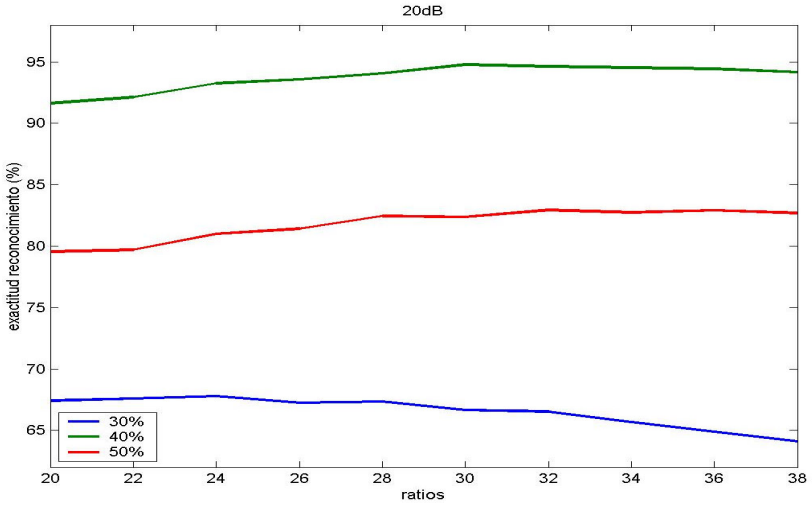
4.3.2.2 Experimento 2: Optimización fina de ratios en test

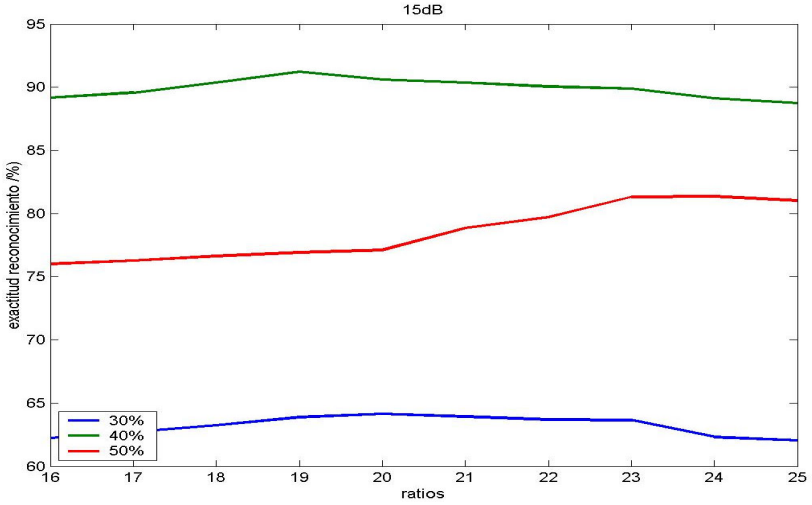
El experimento anterior nos da los valores aproximados de los ratios que han de utilizarse en cada SNR para obtener resultados óptimos en la exactitud del reconocimiento de voz. De esta forma, el siguiente paso es hacer un estudio más en detalle en cada pico para determinar con mayor exactitud el ratio adecuado para cada una de las SNR bajo estudio, ya que los valores de los ratios en el experimento anterior se diseñaron para cubrir todo el espectro posible de valores. Con ello obtendremos una mayor precisión en el proceso de ajuste de los parámetros del algoritmo y llegaremos a los máximos resultados para el reconocimiento vocal que pueden alcanzarse mediante este método de ajuste manual.

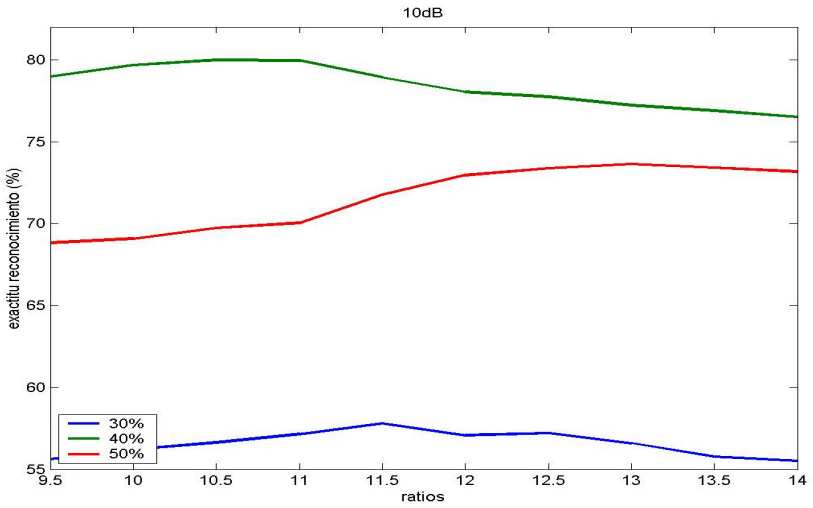
Para cada una de las SNR, se eligen 10 nuevos valores para los ratios en las zonas donde se alcanzan los picos correspondientes. La separación entre dichos valores será distinta para cada una de las SNR ya que, dependiendo de la potencia de ruido presente en la muestra, el límite de propagación debe ser ajustado en una zona más o menos cercana al valor máximo de la Característica 1, por lo que existe distinta sensibilidad de cambio en el límite de propagación a cambios relativos de los ratios. Por ejemplo en el caso limpio la distancia entre los ratios se ajusta a 4; suponiendo que el máximo valor en el espectrograma sea 40000 cambios de los ratios de 4 unidades alrededor de 96 dan como resultado cambios en el límite de propagación de unas 16 unidades (40000/96 – 40000/100 ~ 16). Por el contrario, para -5dB la distancia se ajusta a 0.1 ya que cambios de esa cantidad alrededor de 2 dan lugar a cambios en el límite de propagación de aproximadamente 900 unidades (40000/2 – 40000/2.1 ~ 900). Estos cambios tan diferentes en los límites de propagación se explican por la distribución exponencial que tienen los puntos del Espacio Característico. Alrededor de 400, donde se juega con los límites para el caso limpio la densidad es bastante alta, por lo que pequeños cambios del límite de propagación de traducen en grandes variaciones del área segmentada; mientras que alrededor de 20000, para el caso de –5dB, la densidad de puntos es mucho menor por lo que los cambios en los límites de propagación deben ser mayores para apreciar diferencias en el área segmentada.

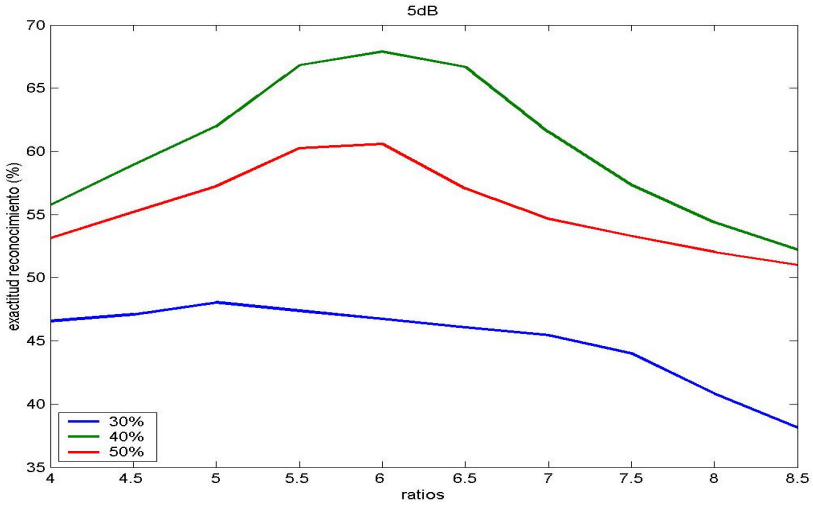
Para cada una de las SNR se obtendrán los resultados de exactitud en el reconocimiento vocal con el reconocedor entrenado con los 3 mejores conjuntos de muestras de entrenamiento determinadas en el experimento anterior; pero extendiendo el número de muestras de entrenamiento a 8440 en lugar de 1000, lo que mejorará de manera notable los valores de exactitud en el reconocimiento obtenidos en este apartado. El número de muestras de test también será aumentado de 125 a 1001 para cumplir con la configuración estándar del paquete Aurora 2; por lo que tendremos para cada una de las SNR 10 grupos de 1001 muestras, cada grupo filtrado morfológicamente utilizando un límite de propagación distinto. A continuación presentamos gráficamente los resultados de esta optimización para cada una de las SNR.

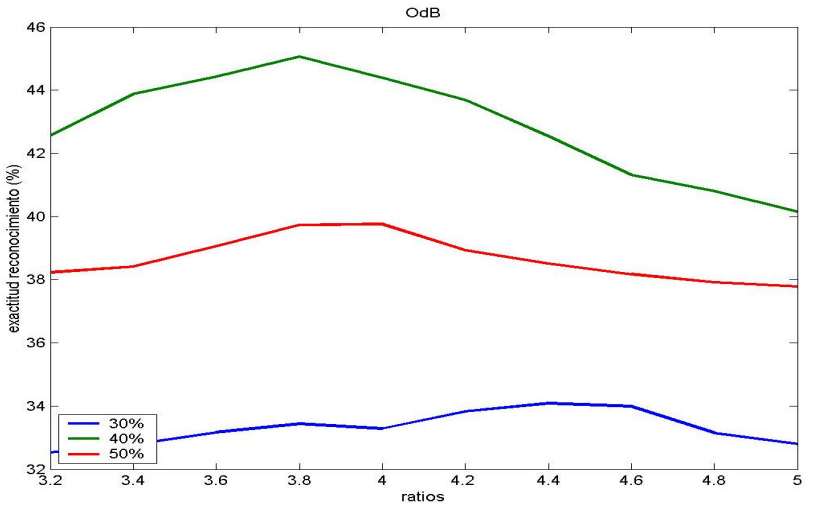

LIMPIO
exactitud reconocimiento (%)
ratios
30%
40%
50%

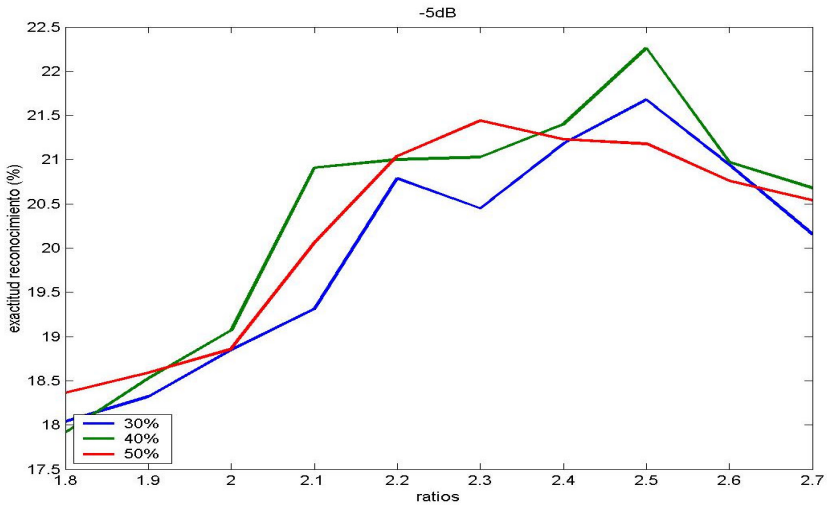

20dB
exactitud reconocimiento (%)
ratios
30%
40%
50%


15dB
exactitud reconocimiento (%)
ratios
30%
40%
50%


10dB
exactitu reconocimiento (%)
ratios
30%
40%
50%


5dB
exactitud reconocimiento (%)
ratios
30%
40%
50%


OdB
exactitud reconocimiento (%)
ratios
30%
40%
50%



Figuras 4.22 Optimización fina de los ratios para las 7 SNR bajo estudio. Resultados de reconocimiento vocal con tres entrenamientos distintos: segmentando el 30%, 40% y 50%.

En principio observamos que, al aumentar las muestras de entrenamiento y test para cumplir con el estándar Aurora2, se sigue cumpliendo que el entrenamiento del 40% es el que da mejores resultados de exactitud en el reconocimiento para todas las SNR. De esta forma quedan determinadas las áreas correctas a utilizar desde el punto de vista de la información necesaria en los modelos para reconocer muestras tratadas con filtrado morfológico.

En cada una de las SNR tenemos un máximo para un ratio determinado, que es el que nos indica el límite de propagación óptimo a utilizar. Este máximo no se da exactamente para el mismo ratio en los tres entrenamientos con los que estamos probando el reconocedor, aunque si que se alcanza en el rango que hemos elegido para hacer el estudio. De esta forma, comprobamos que los máximos para los demás entrenamientos son, en general, notablemente menores que el alcanzado para el entrenamiento del 40%. A continuación mostramos una tabla resumen de la exactitud de reconocimiento alcanzada para cada SNR y el ratio con el que se ha obtenido.

SNR	Limpio	20dB	15dB	10dB	5dB	0dB	-5dB
Exactitud (%)	96.87	94.78	91.22	79.88	67.88	45.06	22.26
ratio	100	30	19	10.5	6	3.8	2.5

Tabla 4.6 Resumen de resultados de optimización de ratios para cada SNR

La tabla 4.6 muestra los ratios óptimos que han de utilizarse en cada una de las SNR para alcanzar máxima exactitud en el reconocimiento de voz. Estos resultados se alcanzan para un reconocedor entrenado con muestras segmentadas utilizando ratio = 100 (da lugar a una segmentación del 40%). Se observa cómo, en el caso limpio, el óptimo se alcanza también para ratio = 100, lo que apoya la idea de que el entrenamiento ha de hacerse siempre en las condiciones que vaya a encontrase en la etapa de test ya que los habrá una mayor proporción de encuentros debido a la similitud de los modelos. Se comprueba que, tras el proceso de optimización, obtenemos una mejora para todas las SNR respecto a los resultados obtenidos en la sección 4.2. El resultado para el caso limpio se acerca mucho más al obtenido con la línea base (98%), por lo que vemos que las áreas resultantes de utilizar los nuevos ratios preservan más información útil para el reconocedor de voz.

De todas formas estos resultados han sido obtenidos mediante un algoritmo simplificado que ajusta de forma manual los límites respecto al valor máximo de la Característica 1 en el Espacio Característico, con el objetivo de determinar los límites de propagación óptimos y con ello poder ajustar el parámetro P_e del algoritmo original. De esta manera, el siguiente experimento consistirá en identificar la P_e que da como resultado unos ratios lo más parecidos a los encontrados aquí experimentalmente. Una vez obtenida la P_e a utilizar tanto en las muestras de entrenamiento como en las de test se llevará a cabo el experimento con el algoritmo original optimizado, que lleva a cabo una mejor selección del límite de propagación ya que se basa en la potencia de ruido existente en cada muestra particular y de una manera mucho más precisa e individualizada.

A pesar de todo, la modificación del algoritmo que hemos realizado en este apartado nos da pie a pensar en una forma de reducir considerablemente el coste computacional manteniendo resultados competitivos. El algoritmo utilizado sólo necesita una iteración ya que manualmente se ajustan las condiciones finales para el límite de propagación, de manera que se elimina todo el proceso de búsqueda a través de estimaciones cada vez más precisas de la PDF de ruido. Estos límites se ajustan en función de la SNR de la muestra por lo que es necesario un conocimiento a priori sobre el ruido que esta degradando la señal. Esto hace que el algoritmo no sea independiente, pero si se realiza de alguna forma una estimación de dicha SNR se podría automáticamente relacionar con los límites óptimos determinados en este apartado y tener un algoritmo completamente autónomo y mucho más rápido que el original. Esta idea será desarrollada en detalle en el Capítulo 6 junto con otras formas de reducción del coste computacional.

4.3.2.3 Experimento 3: Ajuste de P_e en el algoritmo original

Una vez tenemos los ratios óptimos a utilizar tanto para las muestras de entrenamiento como para las de test tenemos que buscar una forma de relacionarlos con el parámetro P_e del algoritmos original. La forma de hacerlo es ejecutar el algoritmo original y registrar el límite de propagación donde el proceso se detiene, de tal manera que podemos calcular el ratio relativo al máximo valor para la Característica 1. El algoritmo original se ajustará con distintos valores de la Probabilidad de error de manera que el caso en que el conjunto de ratios para todas las SNR sea más cercano al conjunto de ratios obtenidos experimentalmente en el apartado anterior nos indicará el valor del parámetro P_e a utilizar en el algoritmo optimizado.

El algoritmo original utiliza un método para definir el límite de propagación completamente diferente al utilizado en el proceso de optimización por lo que el ratio que se compute para cada muestra no será en absoluto el mismo. De esta forma la elección del parámetro P_e se hará en base a consideraciones medias; es decir, para cada caso de Probabilidad de error y SNR concretas se registrarán los ratios que resultan para un conjunto de muestras y se obtendrá la media de todos ellos para compararla con el ratio obtenido manualmente. Así, el parámetro P_e finalmente elegido dará como resultado una segmentación de áreas cuya extensión, en media, ha sido determinada como óptima.

En primer lugar procedemos con el caso de las muestras de entrenamiento. Se procede a la segmentación de 1000 muestras de entrenamiento registrando el ratio relativo entre el máximo de la Característica 1 y el límite de propagación final, calculándose a continuación la media de todos estos valores para compararlos con el óptimo de referencia determinado en el apartado anterior. En la tabla siguiente se muestran los resultados para los 5 valores diferentes del parámetro P_e que vamos que son puestos bajo estudio.

Pe (%)	**12%**	**10%**	**8%**	**5%**	**1%**
Media de ratios	105.7	**96.6**	85.1	58.6	19.5

Tabla 4.7 Media de ratios de parada para muestras de entrenamiento

El óptimo de referencia obtenido en el apartado anterior es ratio = 100 por lo que el parámetro P_e que utilizaremos para las muestras de entrenamiento será el 10% ya que en media es el que da ratios de parada es más cercanos. Es de interés destacar la gran diferencia existente con la media calculada para el caso del 1% que es el que propone Hory en [1], lo que reafirma la idea de la necesidad de un reajuste de los parámetros del algoritmo para trabajar en el contexto de la mejora de voz para el reconocimiento automático. A continuación mostramos un gráfico de los ratios computados para cada uno de los 1000 archivos y la media de todos ellos para el caso P_e = 10%. La variabilidad que se aprecia en los ratios resulta en una selección más exclusiva del límite de propagación que la realizada mediante el ajuste manual, llevando a un ajuste particular para cada archivo que resultará en un comportamiento mejorado como se comprobará en el apartado siguiente.

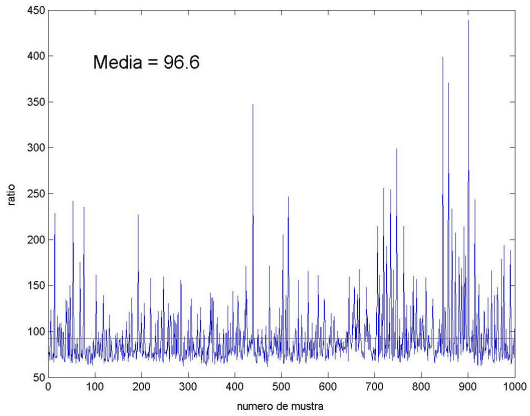


Figura 4.23 Ratios en los que se detiene el algoritmo con Pe=10% en las muestras de entrenamiento

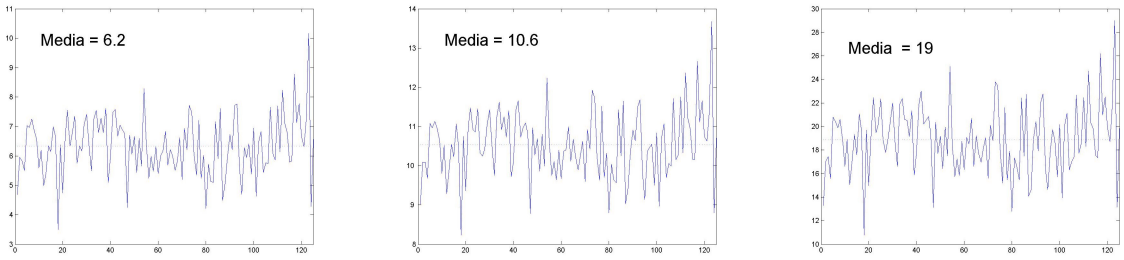
En el caso de las muestras de test se procede de la misma forma pero obteniendo las medias de grupos de 125 archivos para cada una de las SNR. Los experimentos se realizan para condiciones medias de SNR: 5dB, 10dB y 15dB dentro del rango típico bajo estudio que propone el estándar Aurora 2 lo que nos da suficiente información para identificar cual es la P_e óptima que ha de utilizarse en el algoritmo original para el caso de segmentar muestras de test. A continuación mostramos los resultados obtenidos para las 5 P_e en cuestión.

Pe (%)	12%	10%	8%	5%	1%
Media de ratios 5dB	6.6	**6.2**	5.8	5.5	5.1
Media de ratios 10dB	11.8	**10.6**	10.0	8.8	6.7
Media de ratios 15dB	20.3	**19**	18	16.1	13.4

Tabla 4.8 Media de ratios de parada para muestras de test en 3 diferentes condiciones de SNR

El conjunto de medias más cercano a los óptimos de referencia (6 para 5dB, 10.5 para 10dB y 19 para 15dB) son indudablemente los obtenidos con P_e = 10%, de manera que una vez más se reafirma la teoría de utilizar en entrenamiento y en test el mismo tipo de muestras vocales (en este caso un mismo tratamiento de eliminación de ruido) para obtener resultados óptimos en la exactitud del reconocimiento de voz.

Al igual que en el caso de las muestras de entrenamiento vemos en la figura siguiente la variabilidad en los ratios que se obtiene con el algoritmo original. El patrón de variabilidad que resulta es el mismo independientemente de la SNR en cuestión o de la P_e que se utilice pero, dependiendo del caso, este patrón se centra alrededor de un valor determinado. De esta manera, lo que estamos realizando en este capítulo puede verse como un ajuste del offset de este patrón de variación para conseguir resultados óptimos, lo que apoya el método seguido de determinación de la P_e óptima en base a criterios de proximidad a referencias bajo consideraciones medias.



a b c

Figura 4.24 Patrones de variación de los ratios del algoritmo original ajustado con Pe = 10% para 5dB (a), 10dB (b) y 15dB (c).

4.3.2.4 Experimento 4: Algoritmo original optimizado

Los experimentos anteriores han servido para señalar que los parámetros que debemos utilizar en el algoritmo, tanto en las muestras de entrenamiento como en las de test son P_e = 10% y *diff* = 1%. Ajustando esos parámetros las áreas segmentadas en los espectrogramas son las óptimas, de manera que preservan la máxima información disponible que es útil para el reconocimiento de voz mientras que atenúan todo el ruido alrededor de éstas. Para comprobar la efectividad del algoritmo optimizado se lleva a cabo el experimento que describimos a continuación, siguiendo las directrices, como hasta ahora del protocolo de entrenamiento y pruebas recomendado en [18].

En primer lugar se toman los 8440 ficheros de entrenamiento y se filtran morfológicamente con el algoritmo ajustado según los resultados de la optimización. Estos ficheros sirven para entrenar el reconocedor obteniendo unos modelos que serán óptimos para reconocer muestras de voz tratadas con el filtrado morfológico. A continuación se aplica el filtrado morfológico con el mismo algoritmo a los 7 conjuntos de 1001 ficheros de test contaminados con diferentes potencias de ruido y se envían al reconocedor para comprobar la exactitud en el reconocimiento. Como comprobación de que hemos ajustado los parámetros óptimos del algoritmo se realiza el mismo experimento pero para los dos valores de P_e que se encuentran inmediatamente por encima y por debajo de los valores que tenemos bajo estudio. Así, la tabla 4.9 muestra el comportamiento del reconocedor para los 3 valores de Pe: 12%, 10% y 8%.

	Limpio	**20dB**	**15dB**	**10dB**	**5dB**	**0dB**	**-5dB**
Pe = 8%	93.61	92.78	91.94	87.30	76.73	54.29	28.14
Pe = 10%	95.33	94.84	93.34	88.58	78.11	57.32	29.54
Pe = 12%	94.47	94.20	92.72	87.10	73.44	49.12	26.34

Tabla 4.9 Resultados de exactitud en reconocimiento para el algoritmo ajustado con Pe = 8%, 10% y 12%

Los resultados indican que efectivamente el 10% es el mejor valor para P_e en el algoritmo original para todos los valores de SNR. Los valores obtenidos para el caso limpio son ligeramente más bajos que los obtenidos ajustando el límite manualmente, pero para el resto de SNR la mejoría es notable, llegando a una diferencia del 12% para 0dB. Esto pone de manifiesto la efectividad del algoritmo original ajustando el límite de propagación de manera particular, dependiendo de las

características de cada archivo concreto. A continuación mostramos una figura comparativa de los resultados obtenidos en el apartado 4.2 antes de optimizar los parámetros que controlan la cantidad de área segmentada y los resultados tras la optimización de dichos parámetros.

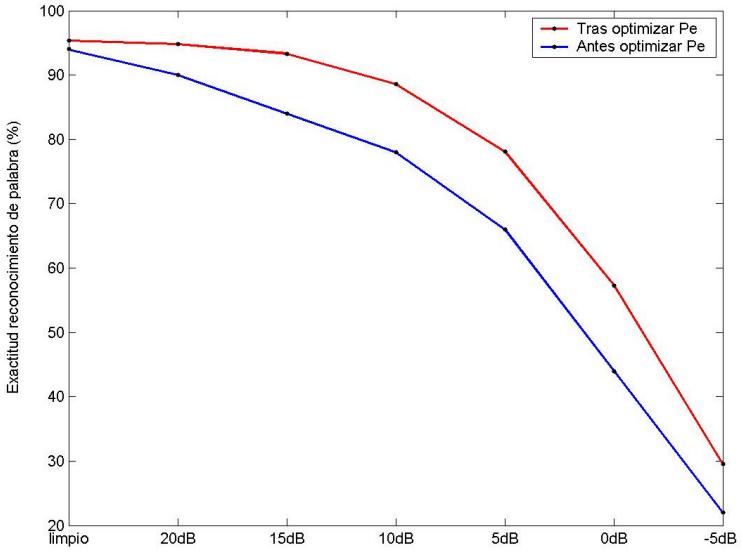


Figura 4.25 Comparación de comportamiento del reconocedor antes y después de optimizar el parámetro Pe en entrenamiento y test

La optimización del parámetro P_e hace que los resultados mejoren de manera notable para todas las SNR por debajo de 20dB de forma que se consigue un algoritmo mejorado mucho más robusto frente al ruido que con los antiguos parámetros. La curva que muestra el comportamiento tras la optimización de P_e es en media algo más de un 8% mejor, obteniéndose la máxima diferencia para 0dB, donde el porcentaje de mejora es del 13%.

4.4 Conclusiones

En este capítulo se han llevado a cabo una serie de mejoras en el algoritmo orientadas a optimizar el comportamiento cuando se aplica en el entorno de la mejora de las señales vocales para el reconocimiento automático. En principio se estudió el proceso de manera detallada para encontrar las deficiencias y las posibles mejoras al aplicar el algoritmo a espectrogramas de voz inmersas en ruido WGN. Como consecuencia de ello se detectó en primer lugar una deficiencia en la manera de

encontrar los puntos candidatos a una propagación y se aplicó un método para solucionar este problema de forma efectiva en el apartado 4.1. A continuación, en el apartado 4.2, estudiamos la forma de incluir las variaciones de las amplitudes espectrales, ya que en [3] los experimentos se llevaron a cabo con espectrogramas en los que la parte clasificada como voz tenía el valor '1' y la parte clasificada como ruido el valor '0'. De esta manera se aplica el procedimiento de multiplicación por una máscara que permite ajustar el nivel de atenuación del ruido circundante y realizar una transición suave entre de los valores de amplitud en el espectrograma, vitales para dar lugar a una onda temporal sin cambios abruptos que restan naturalidad. Por último se detecta la necesidad de optimizar las áreas segmentadas tanto en las muestras del entrenamiento del reconocedor como en las muestras de test. Esto motiva un estudio de los parámetros del algoritmo que gobiernan la cantidad de área segmentada y un proceso de optimización de los mismos que se lleva a cabo a lo largo del apartado 4.3. La figura 4.26 muestra los valores obtenidos para la exactitud del reconocimiento de voz obtenidos tras el proceso completo de optimización frente a la línea base, los resultados obtenidos en [3] y el comportamiento que se obtiene con el tratamiento mediante substracción espectral (QBNE).

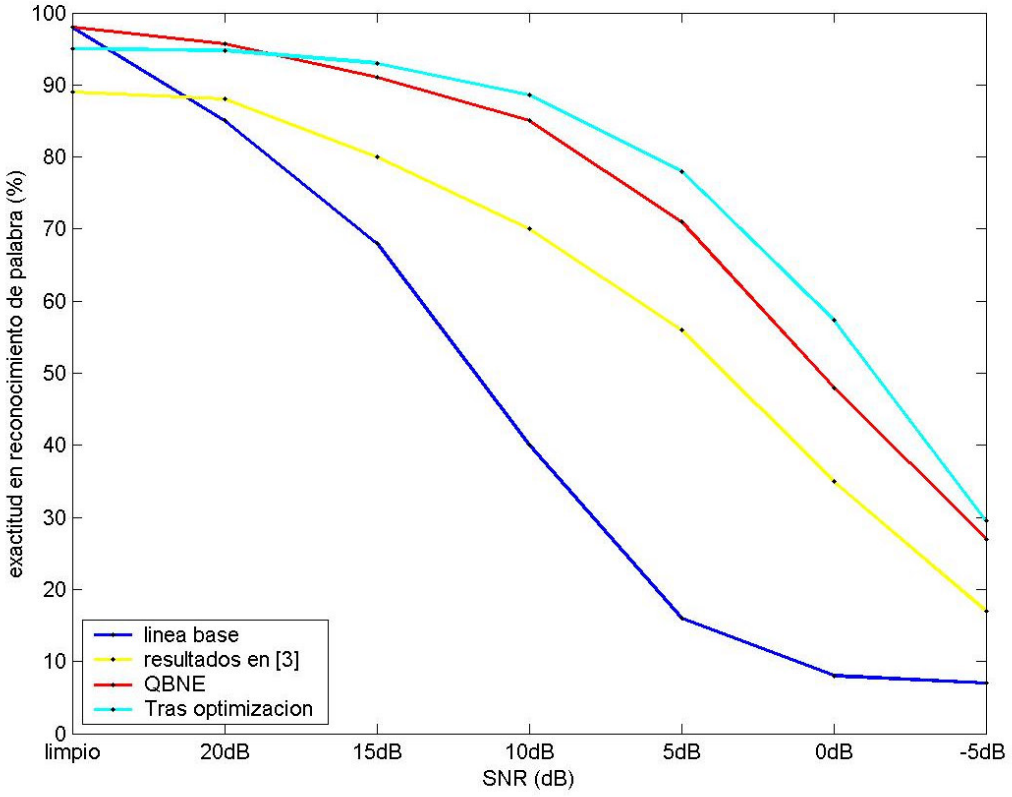


Figura 4.26 Comparación de resultados en exactitud del reconocimiento de palabra entre la línea base en azul, los resultados obtenidos en [3] en amarillo, utilizando substracción espectral en rojo y el filtrado morfológico optimizado en celeste.

Tras todo el proceso de adaptación a trabajar con señales vocales observamos una mejora realmente notable frente a los resultados obtenidos en [3], de manera que se incrementa la exactitud del reconocimiento de palabra para todas las SNR, con una mejora media del 14.6%. Si sumamos esta mejora a la que conseguían por si solos los resultados de [3] sobre la línea base, el filtrado morfológico optimizado obtiene comportamiento que en media es un 30% superior a la señal sin tratamiento, lo que suponen resultados realmente competitivos en el entorno de la mejora de voz para el reconocimiento automático. Apoyando este hecho vemos como el filtrado morfológico optimizado obtiene resultados del mismo orden o superior que el obtenido mediante el tratamiento con substracción espectral basada en quantiles que es una técnica bien establecida y reconocida en este ámbito. De esta manera, el proceso de optimización del algoritmo para tratar señales vocales se considera un éxito, dando como resultado un algoritmo competitivo en el mundo del pre-procesamiento de señales vocales para el reconocimiento automático. A partir de aquí, los siguientes pasos consisten en: por una parte intentar mejorar aún más el comportamiento mediante una compensación del ruido presente en las zonas segmentadas, a través de una combinación con la substracción espectral llevada a cabo en el Capítulo 5; y por otra tratar de reducir el tiempo de procesamiento del algoritmo, tratado en el Capítulo 6.