Universidad Central “Marta Abreu” de Las Villas

Facultad de Ingeniería Eléctrica

Centro de Estudios de Electrónica y Tecnologías de la Información. (CEETI)

TRABAJO DE DIPLOMA

Análisis de técnicas de reducción de ruido para el reconocimiento robusto del habla.

Autor: Dayana Ferrer de Armas.

Tutor: Dr. Julián Cárdenas Barreras.

Santa Clara

2014

"Año 56 de la Revolución"

Universidad Central “Marta Abreu” de Las Villas

Facultad de Ingeniería Eléctrica

Centro de Estudios de Electrónica y Tecnologías de la Información. (CEETI)

TRABAJO DE DIPLOMA

Análisis de técnicas de reducción de ruido para el reconocimiento robusto del habla.

Autor: Dayana Ferrer de Armas.

ferrero@uclv.edu.cu

Tutor: Dr. Julián Cárdenas Barreras.

julian@uclv.edu.cu

Santa Clara

"Año 56 de la Revolución"

Hago constar que el presente trabajo de diploma fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de estudios de la especialidad de Ingeniería Biomédica, autorizando a que el mismo sea utilizado por la Institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos, ni publicados sin autorización de la Universidad.

_____

Firma del Autor

Los abajo firmantes certificamos que el presente trabajo ha sido realizado según acuerdo de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del Tutor

Firma del Jefe de Departamento donde se defiende el trabajo

_____

Firma del Responsable de Información Científico-Técnica

PENSAMIENTO

"Lo fundamental es que seamos capaces de hacer cada día algo que perfeccione lo que hicimos el día anterior. Tenemos que ir sobre nuestros errores, marchar sobre ellos, analizarlos y que no se repitan."

Ernesto Che Guevara

DEDICATORIA

- *A mis padres quienes han estado a mi lado en cada momento y con su amor infinito, confianza, preocupación y entrega total, me han llevado a darle este, su mayor regalo. Los quiero.*
- *A mi novio Damián que ha sido mi fiel compañero en los buenos y malos momentos, dándome aliento para salir adelante. Te amo.*
- *A la memoria de mi abuelo papa, para que esté donde esté se sienta orgulloso de mí.*

AGRADECIMIENTOS

- *A mis padres por haberme apoyado siempre durante el transcurso de la carrera.*
- *A mi amigo y novio Damián, por confiar en mí y darme fuerzas para seguir adelante cuando pensaba que no podía.*
- *A mi tutor por su ayuda incondicional en todos los momentos que he necesitado, así como por su interés en transmitirme sus conocimientos y motivarme cada día a la superación profesional.*
- *A mis padrinos por demostrarme que puedo contar con ellos.*
- *A Lisandra, Maidelis y Merlin que hicieron muy amena mi estancia en la UCLV el último curso.*
- *A todas las personas que de una forma u otra contribuyeron a la realización de este trabajo.*

A todos…

MUCHAS GRACIAS.

TAREA TÉCNICA

- Estudio de algoritmos de reducción de ruido empleados en señales de voz.
- Estudio de métodos que permiten evaluar el desempeño de los algoritmos de reducción de ruido.
- Implementación en Matlab de los algoritmos elegidos.
- Hacer uso de bases de datos de voz internacional que permitan realizar los experimentos de filtrado.
- Diseño de los experimentos que permitan extraer conclusiones respecto al desempeño real de los diferentes métodos de reducción de ruido propuestos.
- Realización de los experimentos.
- Confección del informe final.

Firma del Autor

Firma del Tutor

RESUMEN

El reconocimiento robusto del habla es un área de investigación activa en la actualidad. La existencia de dispositivos móviles que pueden ser manejados en cualquier entorno hace del reconocimiento del habla una habilidad deseable y útil. No obstante los efectos nocivos del ruido, en la mayor parte de los casos, con características altamente no estacionarias, atentan contra el buen desempeño de los sistemas de reconocimiento, convirtiéndolos, en muchos casos en herramientas prácticamente no utilizadas.

La reducción del ruido en estos adversos entornos constituye un reto de investigación y desarrollo. La presente tesis pretende comparar el desempeño de varias técnicas de reducción de ruido que puedan ser empleadas como pre-procesamiento de la señal de habla a fin de incrementar las tasas de reconocimientos de los sistemas mencionados. Los algoritmos a elegir deben caracterizarse no solo por su buen desempeño ante diferente relaciones señal a ruido (SNR), sino también por su eficiencia computacional, de manera que no constituyan una carga de cómputo significativa a los sistemas que lo empleen.

TABLA DE CONTENIDOS

INTRODUCCIÓN

El habla es el medio de comunicación más natural, cómodo y versátil del que disponemos las personas. No es de extrañar, por lo tanto, el interés hacia el uso de las tecnologías del habla, y en particular el reconocimiento automático de habla (RAH), como método de interacción hombre-máquina. En las recientes décadas, los sistemas de reconocimiento de discurso han mejorado significativamente. No obstante, todavía en los ambientes ruidosos sigue siendo una tarea muy desafiante.

Las tecnologías de reconocimiento del habla han alcanzado un alto grado de madurez, de modo que son un componente básico en el diseño de las interfaces conversacionales de las que están dotadas muchas aplicaciones, incluyendo las desarrolladas sobre nuevos dispositivos como teléfonos móviles y agendas personales.

En este contexto, los sistemas de reconocimiento robustos del habla se han vuelto una habilidad deseable y útil, no obstante los efectos nocivos del ruido, en la mayor parte de los casos, con características altamente no estacionarias, atentan contra el buen desempeño de los sistemas de reconocimiento, convirtiéndolos, en muchos casos en herramientas prácticamente no utilizadas.

Los efectos adversos sobre la señal de voz del entorno acústico y del canal de comunicaciones suponen uno de los inconvenientes más importantes en el ámbito del RAH. A pesar de los importantes avances logrados durante las últimas décadas, los sistemas para reconocimiento automático de habla usados actualmente presentan importantes limitaciones prácticas. Estas limitaciones son las principales responsables de la lenta y aun escasa incorporación en la vida diaria de la tecnología de reconocimiento automático de habla.

Las propiedades que se exigen en la actualidad a estos sistemas (robustez ante ambientes y canales de comunicaciones adversas, adaptabilidad, multimodalidad, capacidad multilingüe, bajo consumo de recursos, etc.) no se ven satisfechas en su totalidad por la tecnología predominante.

Si bien es cierto que los reconocedores de habla actuales proporcionan buenos resultados en tareas y entornos acústicos controlados, su comportamiento se degrada rápidamente en situaciones más realistas, donde la señal de voz está contaminada con diversos tipos de ruido (calle, oficina, reverberación, ruido aditivo, u otras causas).

Por esta razón, las líneas de investigación que constituyen esta tesis se orientan hacia el análisis de varias técnicas para el mejoramiento robusto del habla. Los algoritmos a elegir deben caracterizarse no solo por su buen desempeño ante las diferentes relaciones señal a ruido (SNR), sino también por su eficiencia computacional, de manera que no constituyan una carga de cómputo significativa a los sistemas que lo empleen.

Con el propósito de mejorar las prestaciones en condiciones ruidosas de los sistemas para reconocimiento automático de habla se han propuesto numerosas técnicas, no excluyentes entre sí. En [1]; [2]; [3]; [4]; [5]; entre otros se puede encontrar una amplia revisión de dichas técnicas. Estas han sido desarrolladas para el reconocimiento robusto de habla y han sido aplicadas con diversos grados de éxito en multitud de estudios que cubren un amplio rango de situaciones.

A pesar del esfuerzo investigativo realizado en este ámbito, cabe señalar que el reconocimiento robusto de habla es un problema para el que aún no se han encontrado soluciones completamente satisfactorias. Por lo antes expresado, el problema científico de esta investigación es: ¿Cuáles de los algoritmos utilizables para la reducción de ruido presentan un mejor resultado y un menor costo computacional, para el Reconocimiento Robusto del Habla?

De aquí que el objetivo general de este trabajo es: Comparar el desempeño de varias técnicas de reducción de ruido que puedan ser empleadas como pre-procesamiento de la

señal de habla a fin de incrementar las tasas de reconocimientos de los sistemas mencionados.

De esta forma, se declaran como objetivos específicos los siguientes:

1. Identificar al menos tres algoritmos de reducción de ruido aplicables como métodos de procesamiento de la señal de voz, en sistemas de reconocimiento automático del habla.

2. Realizar el filtrado de señales del habla contaminadas con diferentes tipos de ruido a diferentes SNR.

3. Comparar el desempeño de los diferentes algoritmos empleando medidas objetivas que cuantifiquen la reducción del ruido y la inteligibilidad de la señal procesada.

4. Comparar el desempeño de los algoritmos empleando un sistema de reconocimiento del habla y el costo computacional que ellos añaden al sistema.

Con este proyecto se pretende identificar algoritmos de reducción de ruidos de bajo costo computacional que permitan mejorar las tasas de reconocimientos de sistemas automáticos de reconocimiento del habla; extensible además a reconocimiento automático de locutores.

Los resultados de investigación poseen aplicabilidad práctica y teórica de trascendencia al tema de reconocimiento automático del habla y de locutores. El desarrollo exitoso del trabajo permitirá la consolidación de aplicaciones de tecnología del habla, especialmente en aspectos relacionados con la robustez.

Organización del informe:

El informe de la investigación se estructurará en introducción, capitulario, conclusiones, referencias bibliográficas, bibliografía y anexos.

En el capítulo 1 se presentara la caracterización de los métodos de reducción de ruido, y las metodologías que permiten realizar un estudio comparativo entre ellos. El objetivo de este

capítulo será introducir el estado actual de la reducción de ruido para el reconocimiento robusto del habla, así como establecer la importancia y motivación de este trabajo.

En el capítulo 2 se hará el diseño metodológico de la investigación.

En el capítulo 3 se tendrán los resultados y la validación mediante la comparación con ejemplos prácticos de la efectividad de los métodos elegidos permitiendo seleccionar el método o combinación de métodos más apropiada para su inclusión en sistemas de reconocimiento del habla.

En las conclusiones y recomendaciones se ofrecerán los aspectos más significativos del cumplimiento de los objetivos trazados en el trabajo de diploma. Se mencionaran también las limitaciones del mismo y las direcciones a seguir en futuras investigaciones.

CAPÍTULO 1. MARCO TEÓRICO.

En este capítulo se presenta el marco teórico alrededor del tema del Reconocimiento Robusto del Habla. Primeramente comenzamos con una breve generalización del Reconocimiento Robusto del Habla 1.1. La mención de las principales limitaciones y obstáculos a los que se enfrenta esta tecnología en la actualidad 1.3, así como las posibles vías de mejora planteadas hasta el momento.

1.1 El Reconocimiento Robusto del Habla.

El reconocimiento robusto del habla es un área de investigación activa en la actualidad, debido al creciente interés en la interacción hombre-máquina, dicho interés se ha visto favorecido por los avances realizados en las diversas tecnologías involucradas en el Reconocimiento Automático del Habla (RAH), por el aumento de las capacidades de los terminales de usuario y de las redes de comunicaciones, así como por las exigencias de una sociedad demandante de una mayor cantidad y calidad de servicios.

El proceso de reconocimiento se produce de manera óptima cuando las condiciones de los datos que se evalúan son idénticas a aquellas con las que se entrenó el sistema de reconocimiento. Esto no ocurre casi nunca en el mundo real de las aplicaciones de reconocimiento del habla. Existen muchas fuentes de variabilidad que producen desajustes entre las condiciones de entrenamiento y las de evaluación.

Cuando la variabilidad se produce por cambios en el entorno del hablante, y/o en la posición y características del canal que éste utiliza, las estrategias para combatirlas se denominan estrategias de robustecimiento del reconocedor de habla. El reconocimiento

robusto del habla es por tanto aquel que esta inmunizado o es lo menos vulnerable posible a los cambios de las condiciones del entorno en que se produce la evaluación.

Durante este período de tiempo podemos afirmar que el avance científico-tecnológico en Tecnología del Habla ha sido espectacular. Este avance queda patente al contemplar el nacimiento y la difusión de productos y aplicaciones como programas de dictado y servicios de telefonía. Pero a pesar de lo anterior son todavía muchos e importantes los retos a afrontar para alcanzar la ansiada consolidación de los productos y aplicaciones de la Tecnología del Habla.

1.2 El ruido y sus efectos.

El ruido se define como todo sonido no deseado, que distorsiona la información transmitida por la onda acústica dificultando su correcta percepción. Existen dos tipos fundamentales de distorsiones de la señal de voz: el ruido aditivo y la distorsión de canal.

El ruido aditivo se define como el que se suma a la señal de voz en el dominio del tiempo, y será estacionario si además tiene una densidad de potencia espectral que no varía con el tiempo. Dentro de esta categoría existen ruidos blancos, que son aquellos que tienen un espectro de potencias plano, en contraste con los ruidos aditivos coloreados, cuyo espectro de potencias tienen peculiaridades para ciertas frecuencias. Los aditivos no estacionarios son aquellos cuyas propiedades estadísticas cambian con el tiempo.[6]

A esta última categoría pertenecen las voces espontáneas, efectos de los labios o la respiración, etc. Los efectos del ruido aditivo en la señal son los más difíciles de eliminar, ya que tienen la peculiaridad de transformarla no linealmente en ciertos dominios de análisis. El ruido aditivo se puede considerar el motor de la investigación que hay en marcha en el campo del reconocimiento automático robusto del habla en estos momentos.

La distorsión de canal es el ruido que se mezcla de manera convolucional con la señal de voz en el tiempo. Puede estar provocado por reverberaciones de la señal en el medio de transmisión, por la respuesta en frecuencias del micrófono que se utilice, o por peculiaridades del medio de transmisión. Sus efectos han sido combatidos con éxito ya que son lineales y se evitan procesando linealmente la señal con métodos como el filtrado RASTA, cancelación de ecos, o sustracción del valor medio de los coeficientes MFCC.[7]

1.3 Principales limitaciones a las que se enfrenta RAH.

Con el progreso de las nuevas tecnologías y la introducción de sistemas interactivos, se ha incrementado enormemente la demanda de interfaces para comunicarse con las máquinas, pero lamentablemente los mecanismos para incorporar la información a los reconocedores del habla, han resultado difícil desde los primeros trabajos realizados en la década de los años 50 hasta la actualidad, aunque es cierto que se han logrado avances importantes hasta llegar a los sistemas disponibles hoy en día. Las dificultades técnicas más importantes a las que se enfrenta el reconocimiento automático de habla son las siguientes:

- Variabilidad fonética inter e intra-locutor. La diversidad que se observa en las características fonéticas de las clases acústicas consideradas tiene varios orígenes. Por una parte, se produce como consecuencia de las diferencias fisiológicas y culturales entre los distintos locutores. Por otra parte, puede darse en un mismo locutor en función de su estado físico y anímico, o del contexto conversacional en el que se encuentre. Por último, las características de la pronunciación de un sonido también se ven influidas por el contexto acústico.
- Ambigüedades que dificultan la determinación de la clase acústica correspondiente a un segmento de voz. Las propiedades del aparato fonador humano y del lenguaje empleado hacen que ciertas clases acústicas puedan presentar características fonéticas parecidas, dificultando por tanto su distinción. Así mismo, a menudo resulta complicado establecer fronteras claras entre segmentos de voz correspondientes a distintas clases. Para reducir las consecuencias de estas ambigüedades, en los últimos años se ha impulsado el uso de técnicas discriminativas en la etapa de modelado acústico en principio más adecuadas que los modelos generativos empleados tradicionalmente.
- Efectos asociados al habla espontánea. Las prestaciones del reconocimiento automático de habla dependen en gran medida del estilo empleado en la locución. Así, sucede con frecuencia que en una conversación espontánea, especialmente en un ambiente relajado, se descuida la articulación de los sonidos que se pronuncian, produciéndose recortes, supresiones o fusiones de los mismos. Así mismo, pueden darse otros efectos de difícil tratamiento como toses, carraspeos, vacilaciones,

interrupciones, etc. que discutan de forma notable la tarea del reconocimiento automático de habla.

- Entorno acústico adverso. Los efectos del entorno acústico (ruido, interferencias, reverberaciones, etc.) sobre la señal de voz constituyen uno de los inconvenientes más importantes en el ámbito del reconocimiento automático de habla. Por su importancia, podemos destacar la pérdida de información debida a la naturaleza aleatoria del ruido, y la distorsión de las funciones de distribución de los vectores de parámetros espectrales respecto a los modelos acústicos entrenados en condiciones limpias.[8]

1.4 Estrategias de robustecimiento frente al ruido.

Existen varias clasificaciones de las técnicas clásicas de robustecimiento frente a la distorsión producida por el ruido. Una muy ampliamente aceptada es la cancelación del ruido que considera tres familias de técnicas basándose en la filosofía usada para afrontar los efectos del ruido:

- Técnicas de pre-procesado de la señal para cancelar del ruido antes de parametrizar la señal de voz, con el objetivo de que al parametrizarla sea lo más parecida posible a una señal limpia. Su objetivo es proporcionar a las etapas de parametrización o/y modelado acústico una versión de la señal de entrada lo más limpia posible de ruido. Con este fin se usan diversas técnicas de filtrado lineal óptimo, modelado paramétrico autorregresivo de la señal de voz o de su espectro, enmascaramiento del ruido y sustracción espectral, que eliminan, en la medida de lo posible, el ruido que afecta a la señal de voz de entrada al sistema.
- Técnicas de compensación de características. La cancelación de la distorsión del ruido se hace una vez que la señal está parametrizada. Mediante diferentes operaciones como puede ser el filtrado cepstral paso alto, el uso de modelos del efecto del ruido, etc., se recuperan en la medida de lo posible los parámetros de voz limpia a partir de los de voz ruidosa.
- Técnicas de modificación del clasificador para que la clasificación sea óptima teniendo voz ruidosa. La primera intuición es reentrenar el modelo con datos contaminados. Esta

opción podría ser útil aunque tiene un coste computacional, de capacidad de almacenamiento y tiempo elevado. Sería además necesario disponer de datos suficientes en las condiciones de evaluación, lo que no es posible en la mayor parte de las situaciones. En general es más deseable que el sistema se adapte a entornos cambiantes y no conocidos al entrenar y en un tiempo relativamente pequeño. Los mecanismos de adaptación de modelos permiten la modificación de éstos usando una pequeña porción de datos.[2]

Las técnicas desarrolladas para el reconocimiento robusto de habla, de las cuáles se han mencionado aquí únicamente las más relevantes, han sido aplicadas con diversos grados de éxito en multitud de estudios que cubren un amplio rango de situaciones. No obstante, a pesar del esfuerzo investigativo realizado en este ámbito, cabe señalar que el reconocimiento robusto de habla es un problema para el que aún no se han encontrado soluciones completamente satisfactorias.

1.4.1 Técnicas de cancelación del ruido.

Entre todas las dificultades posibles, los efectos adversos sobre la señal de voz del entorno acústico y del canal de comunicaciones suponen uno de los inconvenientes más importantes en el ámbito del RAH. Los reconocedores de habla actuales proporcionan buenos resultados en entornos acústicos controlados pero este se ve afectado en situaciones más realistas como cuando la señal de voz está afectada con diversos tipos de ruido.

El objetivo de las técnicas de cancelación del ruido en la señal es eliminarlo antes de que la señal sea procesada y sometida al reconocimiento. Se basan en el supuesto de aditividad de la voz y el ruido en el dominio del tiempo por lo cual el espectro de potencias de la señal ruidosa será la suma de los espectros de la voz el ruido. El ruido se considera estacionario, al menos en las tramas en las que se divide la señal para su tratamiento, y es posible estimar su densidad espectral de potencia. Estas técnicas siguen dos pasos:

i. Estimación del espectro del ruido.
ii. Atenuación de dicho espectro de ruido en el espectro de la señal contaminada.

Los distintos algoritmos de atenuación espectral, se basan en diferentes métodos. Una posible clasificación de los mismos es la que los divide en tres grupos: los algoritmos de

sustracción espectral (de potencia o de amplitud), el filtrado de Wiener y el algoritmo geométrico para la sustracción espectral.[1, 9]

1.4.1.1 Técnicas basadas en el pre-procesamiento.

Entre las técnicas empleadas en el pre-procesamiento de la señal de entrada, el filtro de mediana y sus diversas versiones (filtros de mediana ponderados, adaptativos, etc.) destacan por su sencillez y su eficacia para eliminar el ruido impulsivo. No obstante, estas técnicas presentan diversos inconvenientes que limitan su uso práctico en el RAH. En primer lugar, el filtro de mediana afecta a todas las muestras de la señal de entrada, ocasionando una cierta distorsión sobre aquellas no contaminadas por el ruido impulsivo. En segundo lugar, esta solución no resulta adecuada cuando el ruido aparece en forma de ráfagas o afecta a un porcentaje elevado de las muestras de la señal. La razón es que en este caso se requieren ventanas de filtrado de gran tamaño, lo que acentúa la degradación de los segmentos no contaminados de la señal.

Una aproximación muy interesante para la eliminación del ruido impulsivo presente en la señal de entrada consiste en la detección de las muestras contaminadas y su sustitución por una estimación apropiada de los valores originales de la señal. La detección de las muestras contaminadas se realiza mediante un análisis de predicción lineal de la señal. Este modelado es apropiado para la voz pero no para el ruido impulsivo, por lo que su residuo de predicción será mayor que el de la señal de interés.

Esta diferencia aumenta aún más aplicando un filtro adaptado al filtro inverso de predicción. Los principales inconvenientes de este método son la complejidad añadida que introduce el sistema de pre procesamiento y la restricción sobre la duración máxima de las ráfagas del ruido impulsivo que impone la etapa de reconstrucción. Entre las técnicas empleadas comúnmente para la extracción de características robustas, podemos destacar la eficacia del procedimiento para la normalización en media y varianza de los coeficientes cepstrales. Una alternativa interesante consiste en la extracción de los coeficientes cepstrales a partir de representaciones de la señal más robustas o en las que resulta más fácil aislar y eliminar el ruido[10].

1.4.1.2 Técnicas basadas en el modelado acústico.

En la etapa de modelado acústico, las técnicas más eficaces para el reconocimiento robusto del habla contaminada con ruido impulsivo se basan bien en el entrenamiento de los modelos ocultos de Markov (HMMs) con voz contaminada, bien en el uso de modelos del ruido para la adaptación de los modelos acústicos entrenados en ausencia de ruido. El entrenamiento de los modelos ocultos de Markov con voz contaminada se emplea frecuentemente como referencia con la que comparar otras técnicas; sin embargo, la utilidad práctica de este procedimiento es escasa ya que requiere un conocimiento a priori muy preciso acerca del entorno acústico en el que operará el sistema (lo que incluye ejemplos reales de los ruidos).

En parte, las técnicas de combinación de modelos como las descritas en [11] presentan este mismo inconveniente. En estos casos, el reconocedor de habla incorpora un nuevo modelo acústico (generalmente un HMM) que modela las características del ruido que contamina la señal de voz. El modelo del ruido se combina con los modelos de las unidades acústicas consideradas, entrenados en ausencia de ruido, para formar modelos de la voz contaminada que se ajusten mejor a los vectores de observaciones obtenidas en el momento de operación.

Los modelos del ruido también pueden emplearse para compensar su efecto sobre la señal de voz o los vectores de parámetros extraídos de esta. En este caso, se realiza una estimación inicial de la pareja de estados acústicos, correspondientes a los modelos de la voz y del ruido, que maximizan la verosimilitud de la observación. Con esta información es posible estimar las funciones de densidad de probabilidad de los espectros de la señal y del ruido, que se usarán a continuación para diseñar el filtro de Wiener o el estimador de mínimo error cuadrático medio encargados del realce de la señal de voz o de sus coeficientes espectrales.

Este procedimiento se emplea en [12] para el reconocimiento robusto de habla o el realce de voz contaminada con diversos tipos de ruido impulsivo real. En comparación con otros métodos similares de realce estadístico, estas técnicas permiten un mejor modelado de los ruidos no estacionarios gracias al uso de los HMMs.

El principal inconveniente de estas técnicas consiste en que, salvo en aplicaciones muy concretas, resulta complicado conocer a priori y con suficiente precisión las características

del entorno de operación del sistema. Si dichas características no varían o lo hacen muy lentamente, una solución adecuada consiste en adaptar los modelos acústicos mediante métodos sencillos como MAP (máximum a posteriori), MLLR (máximum likelihood linear regression), MCELR (mínimum classification error linear regression), etc.

1.4.1.3 Técnicas basadas en la decodificación.

Centrándonos ahora en la etapa de decodificación, en [13] se propone una modificación del algoritmo de Viterbi consistente en prescindir de un determinado número de observaciones en la búsqueda de la secuencia de estados óptima. En este trabajo se asume que ciertas tramas están tan distorsionadas por el ruido impulsivo que su consideración en la etapa de decodificación sólo puede producir confusiones. Por esta razón, se desechan las verosimilitudes de los vectores más degradados. Este método proporciona buenos resultados en una tarea de reconocimiento de habla contaminada con distintas clases de ruido impulsivo típicas en redes de telefonía móvil e IP. Su dificultad principal reside en determinar el número óptimo de tramas que se deben descartar, proceso que se realiza en paralelo con el reconocimiento de la locución.

El trabajo que se presenta en [14, 15] pretende limitar la influencia de las tramas corruptas sobre la búsqueda de la secuencia de estados óptima. El ruido que afecta a la señal de entrada produce ciertos desajustes entre las funciones de distribución de los vectores de características y los modelos acústicos.

Como resultado, las verosimilitudes de las tramas contaminadas con ruido impulsivo son, por lo general, muchos menores a las obtenidas en condiciones ideales. Para remediarlo, este método divide las componentes del vector de características en varios grupos, dependiendo de su sensibilidad frente al ruido. Las expresiones matemáticas de las funciones de densidad de probabilidad en cada estado, dadas por los modelos de mezclas de Gaussianas, se factorizan conforme a esta división, asignándose a cada término un umbral mínimo. Dichos umbrales, cuyos valores dependen del grado de distorsión del vector de entrada, evitan que la evaluación de los modelos acústicos produzca verosimilitudes anómalas.

Las técnicas de (missing features), por su parte, buscan identificar las regiones corruptas del espectrograma de la señal para reducir su influencia en la etapa de decodificación. Una

vez localizadas, existen diversas alternativas como realizar la decodificación sin considerar dichas componentes, reconstruirlas a partir de las regiones adyacentes del espectrograma, etc. La principal dificultad consiste en identificar con suficiente precisión las regiones distorsionadas en presencia de ruido impulsivo no estacionario que aparece de forma intermitente.

El uso de estas técnicas como el modelado de las distribuciones de probabilidad de emisión de las clases acústicas mediante modelos de mezclas de Gaussianas y la estimación del espectro de la señal de voz en la etapa de parametrización mediante la transformada discreta de Fourier (DFT) en las etapas de parametrización y de modelado acústico de los reconocedores de habla se generalizó hace varias décadas, en un contexto tecnológico muy distinto al actual. Así, entre las razones para su adopción primaron su gran versatilidad y su sencillez algorítmica. En la actualidad, el coste computacional de estas etapas no supone el principal problema en el RAH.

1.5 Aspectos computacionales.

El coste computacional del proceso de reconocimiento se puede dividir en dos partes, el coste computacional de la parametrización de señal de voz, y el coste computacional del proceso de decodificación:

- El coste computacional de la parametrización unido al costo de almacenamiento, aumentará con los algoritmos más elaborados como son las parametrizaciones basadas en los modelos auditivos, siendo mínimo para el caso de sustracción espectral de la media por ejemplo.
- En el proceso de reconocimiento, la decodificación es la que tiene un costo computacional que puede ser bastante elevado, especialmente para el reconocimiento de habla continua.

Los costos mencionados se deben ajustar según los requerimientos de tasa de error de la aplicación, y los medios con los que se cuenta para su implementación.[16]

1.6 Tasa de error y precisión del reconocimiento.

La tasa de error de palabra (WER) se define como la capacidad del reconocedor automático de habla de cometer errores de reconocimiento. Para el caso del reconocimiento del habla continuo, el reconocimiento se hace frase a frase, y se definen tres tipos de errores: errores de inserción, errores de eliminaciones y errores de sustitución.[16]

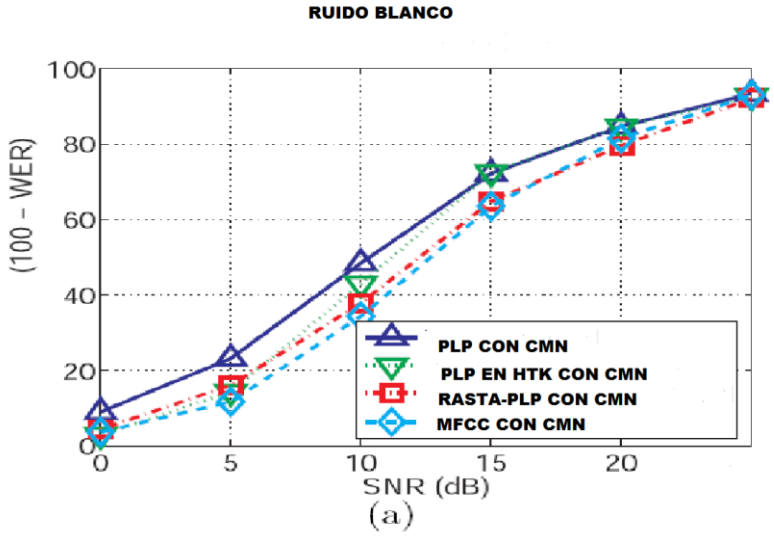
Conseguir que los sistemas de RAH proporcionen unas tasas de reconocimiento aceptables ante cualquier entorno acústico supondría un gran adelanto, ya que las interfaces orales podrían llegar a ser realidades contrastadas y elementos fundamentales en muchas redes de comunicaciones, independientemente del entorno acústico que rodeara al usuario. En la actualidad las tasas de reconocimiento más optimistas están en un orden de magnitud por debajo de las que serían atribuibles al ser humano y el reconocimiento automático de voz continua es un campo de trabajo al que la comunidad científica dedica un esfuerzo importante.

Los sistemas monolocutor y multilocutor consiguen mejores tasas de reconocimiento que los sistemas independientes de locutor, pero requieren un costoso período de aprendizaje o adaptación para cada nuevo locutor. Los sistemas independientes del locutor no tienen este inconveniente porque se entrenan con muchos locutores, pero sus tasas de acierto son menores debido a que la mayoría de las representaciones de la señal de voz son altamente dependientes del locutor.

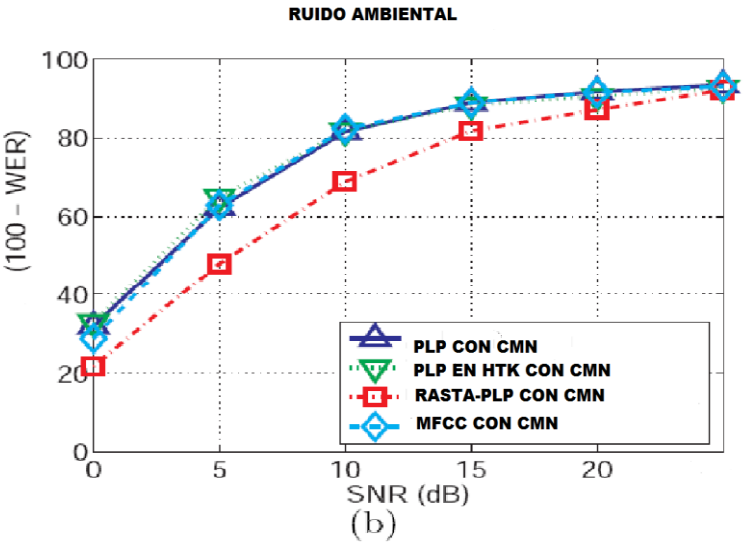
La inmensa mayoría de los sistemas de reconocimiento del habla se diseñan suponiendo que las condiciones ambientales en las que van a funcionar no van a afectar sustancialmente a la señal de voz, lo cual supone una sustancial simplificación del problema general del reconocimiento. Una importante fuente en el incremento en las tasas de error en el reconocimiento la constituyen el entorno y el canal de transmisión: ruidos, interferencias, reverberaciones del entorno, tipo de micrófono, características frecuenciales de una línea de transmisión (caso de haberla), etc.

La parametrización de los coeficientes MFCC (coeficientes cepstrales en escala MEL) y PLP (predicción lineal perceptual) han sido estudiadas en [17] y en [18] con el propósito de desarrollar y evaluar un procedimiento general y sencillo desde el punto de vista algorítmico para el reconocimiento robusto del habla en presencia de distintos tipos de ruido.

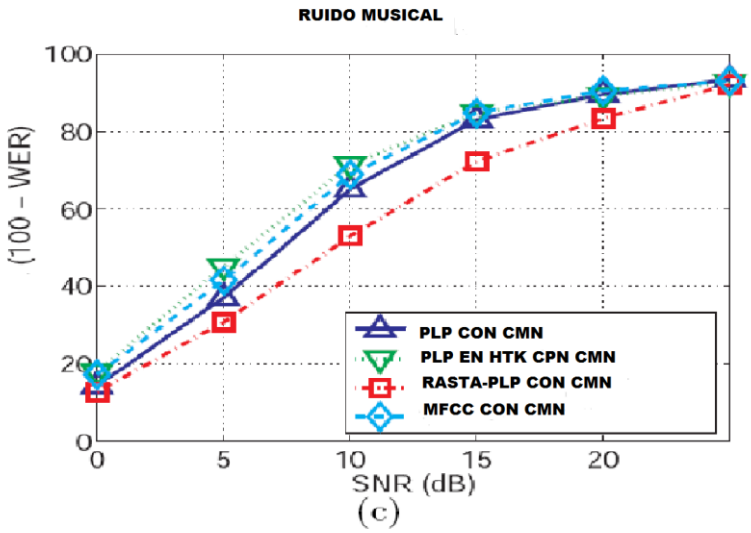
A continuación se muestran unas figuras, en las cuales se observan comparaciones entre las parametrizaciones de los coeficientes MFCC y PLP usando entornos contaminados con diferentes tipos de ruido.



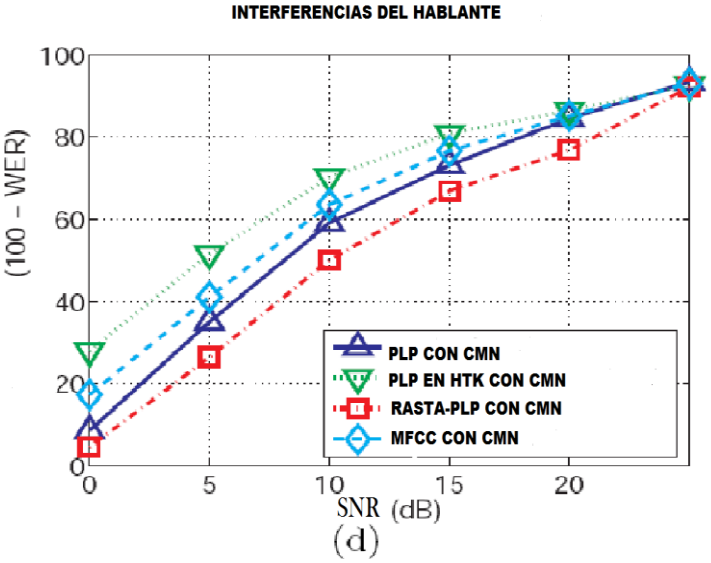
(a)



(b)


RUIDO MUSICAL
(100 – WER)
SNR (dB)
PLP CON CMN
PLP EN HTK CPN CMN
RASTA-PLP CON CMN
MFCC CON CMN

(c)


INTERFERENCIAS DEL HABLANTE
(100 – WER)
SNR (dB)
PLP CON CMN
PLP EN HTK CON CMN
RASTA-PLP CON CMN
MFCC CON CMN

(d)

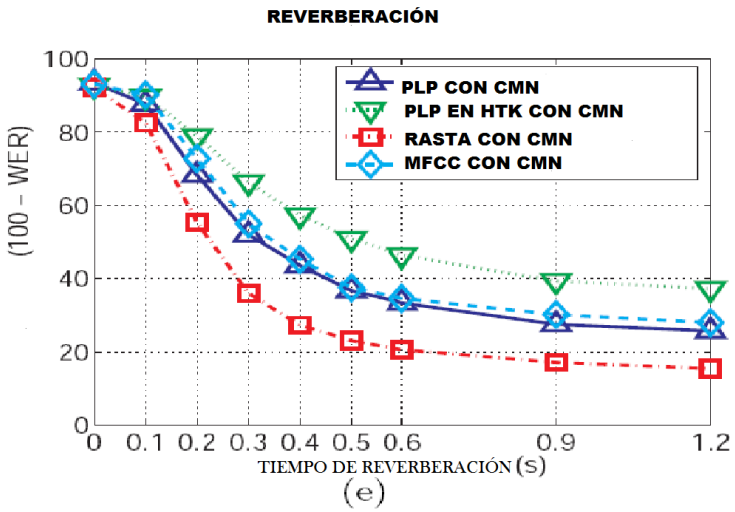


Figura 1.1: Comparaciones entre las parametrizaciones de los coeficientes MFCC y PLP usando entornos contaminados con: (a) Ruido blanco, (b) Ruido ambiental, (c) Ruido musical, (d) Interferencias del hablante, y (e) Reverberaciones.

1.7 Modelado acústico: HMMs.

Las propiedades que se exigen en la actualidad a los sistemas del RAH no se ven satisfechas en su totalidad por la tecnología predominante, basada en el uso de modelos ocultos de Markov (HMMs) para el modelado temporal y modelos de mezclas de Gaussianas (GMMs) para el modelado acústico.

Los sistemas basados en modelos ocultos de Markov (HMMs)[19] constituyen actualmente tecnología de vanguardia en el ámbito del reconocimiento automático de habla. La principal fortaleza de estos sistemas reside precisamente en los HMMs, que permiten un tratamiento mucho más adecuado de la dinámica temporal de la señal de voz, siendo este uno de los problemas principales en el reconocimiento de habla. Las HMMs son autómatas de estados finitos gobernados por un conjunto de probabilidades de transición. Cada estado del HMM tiene asociada una determinada distribución de probabilidad de emisión, generalmente modelada mediante una mezcla gaussiana (GMM).

Tradicionalmente se han empleado los modelos ocultos de Markov por su versatilidad y fácil implementación práctica, si bien en los últimos años han comenzado a estudiarse diversas técnicas con una mayor capacidad expresiva o que generalizan a los HMMs como

los modelos gráficos. Hasta la fecha este método es el que proporciona mejores resultados y el más utilizado.

1.8 Verificación del reconocimiento.

En la mayoría de aplicaciones de reconocimiento automático del habla es necesario disponer de un mecanismo que nos permita verificar las hipótesis generadas por el sistema de reconocimiento. La verificación del reconocimiento asigna una medida de confianza a las hipótesis generadas por el reconocedor de forma que nos permita detectar la presencia de errores de reconocimiento (inserciones, eliminaciones y sustituciones). Las técnicas más utilizadas en la actualidad asignan una medida de confianza a las palabras reconocidas por el sistema de reconocimiento mediante el cálculo de una tasa de probabilidades y son aceptadas o rechazadas comparando la medida de confianza con un umbral de decisión.

La tasa de probabilidades está definida como la relación entre la probabilidad de un modelo oculto de Markov que modela el espacio de reconocimiento correctos con respecto a la probabilidad de un modelo oculto de Markov que modela el espacio de falsas alarmas[20, 21]. Mediante un proceso de entrenamiento de tipo discriminativo se aprende sobre una base de datos de entrenamiento las distribuciones del espacio de reconocimiento correctos y de falsas alarmas. La inclusión de la información de falsas alarmas en el proceso de reconocimiento permite además de minimizar el número de falsas alarmas el aumentar las tasas de reconocimiento [20].

Otros tipos de aproximación al problema de verificación se encaminan hacia la utilización de ciertos parámetros del proceso de reconocimiento (duración, número estados activos, etc.) como información para definir una medida de confianza [22, 23]. Cualquiera que sea el método utilizado, la finalidad es dar una medida robusta sobre la confianza del reconocimiento tanto a nivel acústico como de lenguaje.

CAPÍTULO 2. MATERIALES Y MÉTODOS.

Este trabajo se ha realizado con el propósito de realizar un estudio comparativo de las técnicas de reducción de ruido seleccionadas para el reconocimiento automático del habla, en presencia de diferentes tipos de ruidos y distintos valores de SNR.

En el presente capítulo se describirán los algoritmos que serán objeto de análisis y comparación, conjuntamente al diseño del experimento que evalúa criterios de carácter objetivo, el cual permitirá determinar la efectividad de cada uno de ellos.

2.1 Algoritmos propuestos.

Los algoritmos que se proponen evaluar basan su funcionamiento en la estimación del ruido a partir de una única señal de habla. Cada uno de ellos determina segmentos sin actividad de voz, y extrapola el comportamiento del ruido a los segmentos con presencia de voz para reducir su impacto negativo. En este proceso, deben considerarse las posibilidades de cada método para reducir el ruido, comúnmente no estacionario, y mantener la información del habla inalterada.

La inteligibilidad se mide por medio del porcentaje de palabras reconocidas correctamente por sujetos normoyentes, y se identifican las confusiones más frecuentes ocurridas con cada algoritmo mediante matrices de confusión de consonantes [1]. La calidad del habla obtenida con cada técnica se evalúa en forma subjetiva de acuerdo a la calificación de los oyentes con respecto a un grupo de factores intervinientes en su percepción. También se efectúa la evaluación de calidad mediante un grupo de medidas objetivas seleccionadas al efecto.

2.1.1 Sustracción Espectral Multi-Banda (MBSS).

Un enfoque intuitivo para la supresión de ruido es substraer una estimación del mismo al habla contaminada. Las técnicas de sustracción espectral realizan esta tarea sobre la representación de la señal ruidosa en el dominio de Fourier [24]. Apoyándose en la mayor importancia de la magnitud en la percepción del habla, el enfoque consiste en estimar la magnitud del espectro de la señal limpia, y asignarle la fase de la señal contaminada. En cada tramo de análisis, la estimación de la magnitud del habla limpia se obtiene restando una estimación del espectro de magnitud del ruido al espectro de magnitud de la señal ruidosa. Debido a que no se substrae el espectro real del ruido sino sólo una versión estimada, es necesario introducir una rectificación para asegurar que los valores de magnitud obtenidos sean positivos.

Las técnicas de sustracción espectral aplican una atenuación dependiente de la relación señal-ruido (SNR) del tramo analizado. Sin embargo, la rectificación introducida y las fluctuaciones de las características reales del ruido con respecto al valor estimado, originan la aparición de una distorsión conocida como ruido musical [24]. Este ruido se presenta como picos espectrales aislados que aparecen aleatoriamente sobre frecuencias cambiantes entre los distintos tramos de análisis, tomando características altamente no estacionarias.

El ruido musical tiene efectos negativos en la inteligibilidad y su aparición se debe a las incorrecciones en el cálculo del espectro de ruido de tiempo corto, lo que provoca diferencias espectrales en el habla resultante [1].

Existen ventajas al utilizar métodos no lineales para el procedimiento de sustracción espectral encontradas en la literatura [19], este enfoque se debe a la variación de la relación señal-ruido (SNR) presente en el espectro del habla. A diferencia del ruido blanco gaussiano, que tiene un espectro plano, el espectro del ruido de mundo real no es plano. Por lo tanto, la señal de ruido no afecta uniformemente el espectro del habla, es decir, algunas frecuencias son afectadas más que otras. En el ruido multi-locutor, por ejemplo, las frecuencias bajas donde reside la mayor parte de la energía de habla son afectadas más que las altas frecuencias. Por lo tanto, se impone calcular un factor apropiado que restará sólo la cantidad necesaria de ruido al espectro del habla. En la figura (2.1) se puede apreciar el diagrama en bloque del método de sustracción espectral multi-banda.

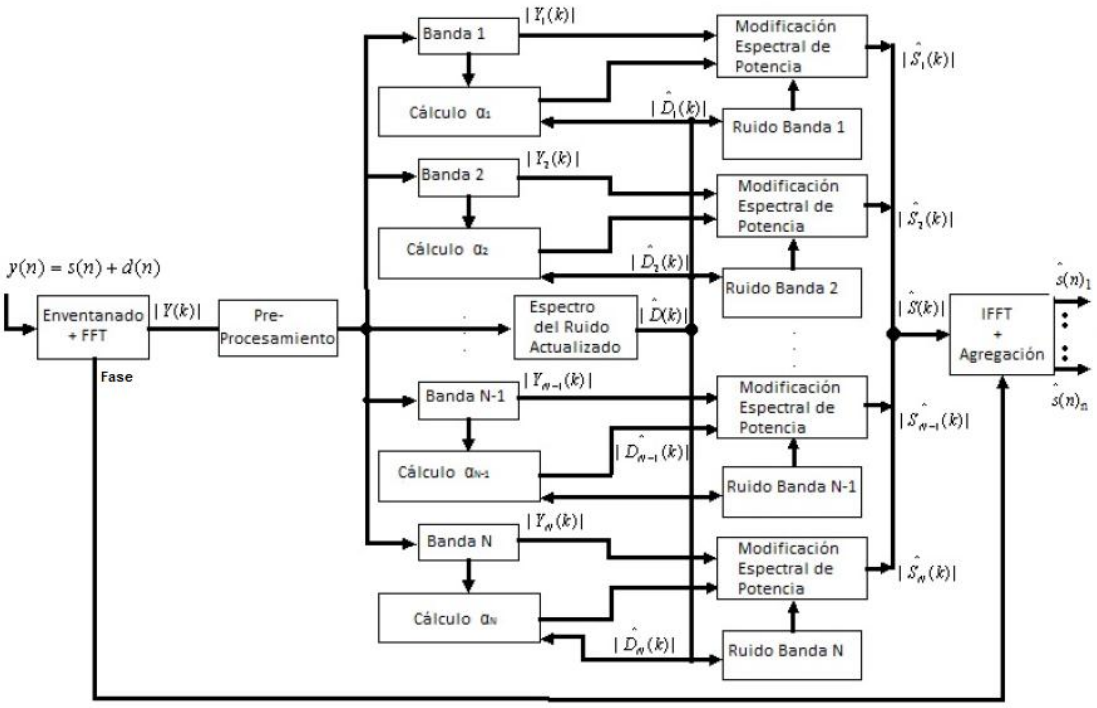


Figura 2.1: Diagrama en bloques del método de sustracción espectral multi-banda (MBSS).

Asumiendo que el ruido aditivo sea estacionario y no correlacionado con la señal de habla limpia, el resultado de la señal de habla contaminado por ruido puede ser expresado como:

$$y(n) = s(n) + d(n) \tag{2.1}$$

Donde y(n) es la señal de habla contaminada, s(n) señal limpia y d(n) la estimación de ruido. Entonces la potencia espectral del habla contaminada se puede expresar como:

$$|Y_{(k)}|^2 \approx |S_{(k)}|^2 + |D_{(k)}|^2 \tag{2.2}$$

Donde $|S_{(k)}|^2$ y $|D_{(k)}|^2$son las potencias espectrales del habla limpia y el ruido respectivamente.

El ruido es estimado en los primeros segmentos de señal y luego es actualizado en los momentos de silencio. El cálculo aproximado de espectro de habla, la Ecuación (2.2) puede ser rescrita como:

$$|Y_{(k)}|^2 \approx |S_{(k)}|^2 + \alpha|D_{(k)}|^2 \tag{2.3}$$

Donde α es un factor de sobre-sustracción en función con la relación señal-ruido (SNR) ecuación (2.5) y (2.6). La implementación asume que el efecto del ruido sobre el espectro del habla es uniforme, y el factor de sobre-sustracción α sustrae una sobre-estimación del ruido en el espectro completo. Este no es el caso de los ruidos presentes en el mundo real, como por ejemplo, el ruido de un carro, de una cafetería, etc. Para tomar en cuenta el efecto del ruido coloreado sobre el espectro de la señal del habla de forma diferente para varias frecuencias, se propone el enfoque de sustracción espectral multi-banda. El espectro del habla es dividido en N bandas no solapadas y la sustracción espectral se realiza independientemente en cada banda. En este caso, la estimación del espectro de la señal limpia se expresa como:

$$\left|\hat{S}_{(k)}\right|^2 \approx \left|Yi_{(k)}\right|^2 + \alpha i \delta i \left|D_{(k)}\right|^2 \qquad bi \le k \le ei \tag{2.4}$$

Donde *bi* y *ei* son el principio y final respectivamente de las tramas en la i-enésima banda de frecuencia, *αi* es el factor de sobre-sustracción en la i-enésima banda y *δi* factor de retoque que puede ser individualmente escogido para cada banda, logrando así personalizar las propiedades de sustracción del ruido. El factor de sobre-sustracción *αi* específico para cada banda se calcula en función de la SNR de la i-enésima banda de frecuencia (SNRi) mediante la ecuación siguiente:

$$SNRi(dB) = 10\log_{10}\left(\frac{\left|Yi_{(k)}\right|^2}{\left|\hat{D}i_{(k)}\right|^2}\right) \tag{2.5}$$

Usando la SNRi se calcula entonces αi mediante:

$$\alpha i = \begin{cases} 5 & SNR < -5 \\ 4 - \frac{3}{20} * (SNR) & -5 \le SNRi \le 20 \\ -100 * (SNRi) & SNRi > 20 \end{cases} \tag{2.6}$$

El uso de *αi* provee de un grado de control sobre los niveles de sustracción de ruido en cada banda y, consecuentemente, el uso de múltiples bandas de frecuencia y el uso del peso *δi* provee una grado adicional de control para cada banda.

Los posibles valores negativos en el espectro realzado según la ecuación (2.4) se reducen empleando:

$$\left|\hat{S}i_{(k)}\right|^2 = \begin{cases} \left|\hat{S}_{(k)}\right|^2 & \left|S_{(k)}\right|^2 > 0 \\ \beta\left|Yi_{(k)}\right|^2 & resto \end{cases} \quad (2.7)$$

Donde el parámetro de reducción β es igual a 0.002.

2.1.1.1 Implementación.

La señal del habla contaminada es enventanada usando una ventana de Hamming de 20ms con un solapamiento de 10ms entre tramas. Luego se calcula la transformada de Fourier para cada trama. El espectro del ruido es inicialmente estimado sobre las 10 primeras tramas de la señal a través de un promedio espectral ponderado que se toma sobre el paso de las tramas y se calcula como:

$$\left|\hat{Y}j_{(k)}\right|^2 = \sum_{I=-M}^{M} W_I * Y_{j-I}(k) \quad (2.8)$$

Donde j es el índice de la trama. El número de tramas, M, es limitado a 2 para prevenir corrimiento espectral. Los factores de peso en cada trama son calculados empíricamente y se muestran a continuación: W= [0.09, 0.25, 0.32 0.25, 0.09]. Posteriormente el espectro del ruido es actualizado para los instantes de silencio, pero esta vez cambia los factores de peso, siendo estos:

W= [0.001, 0.01, 0.02, 0.01, 0.001].

Los valores de δi para la ecuación (2.4) son determinados empíricamente también y se muestran a continuación:

$$\delta i = \begin{cases} 1 & fi < 1kHz \\ 2.5 & 1kHz \leq fi \leq \frac{Fs}{2} - 2kHz \\ 1.5 & fi > \frac{Fs}{2} - 2kHz \end{cases} \quad (2.9)$$

Donde es fi la frecuencia central de la banda i-enésima y *Fs* es la frecuencia de muestreo. La motivación de usar pequeños valores de δi para las bajas frecuencias, es minimizar la distorsión del habla producto a que es allí donde la mayor parte de la energía del habla está presente.

Finalmente, el realce de la señal es obtenido para cada banda mediante la transformada inversa de Fourier usando la fase del espectro original ruidoso. Luego, métodos de

solapamiento y agregación son usados para obtener en cada banda la respectiva señal limpia.

2.1.2 Filtrado de Wiener.

El filtrado de Wiener es una técnica relacionada con las estrategias de filtrado óptimo que intenta minimizar el error cuadrático medio entre la señal limpia y la estimada. Este filtro es el mejor estimador lineal que cumple con esta tarea, y es también óptimo entre los estimadores no lineales cuando los procesos involucrados responden a modelos gaussianos [25].

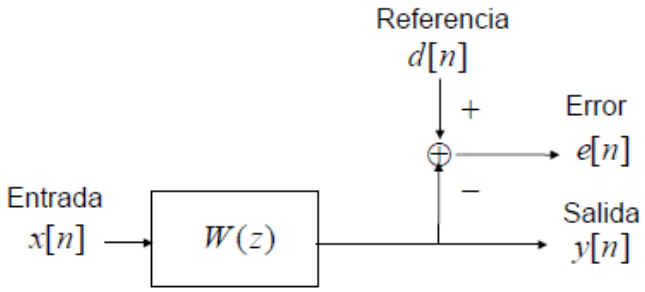


Figura 2.2: Filtro Wiener.

El filtro de Wiener es una de las herramientas fundamentales del tratamiento estadístico de señales (predicción, modelado, identificación, igualación, etc); para derivar este filtro necesitamos únicamente conocer estadísticos de segundo orden: autocorrelación de la entrada y correlación cruzada entre la entrada y la salida deseada. El filtro extrae la parte de la entrada que tiene correlación lineal con la salida deseada: si la salida esta incorrelada con la entrada, el filtro de Wiener es nulo y la señal de error resultante esta incorrelada con la entrada del filtro.

En el dominio espectral, y teniendo en cuenta que la señal de voz sólo puede considerarse localmente estacionaria, el filtro de Wiener puede expresarse de la siguiente manera [25]:

$$H_w(k,n) = \frac{\hat{\Gamma}_s(k,n)}{\hat{\Gamma}_s(k,n)+\hat{\Gamma}_N(k,n)} \qquad (2.10)$$

donde $\hat{\Gamma}_s(k,n)$ y $\hat{\Gamma}_N(k,n$ representan estimadores de corta duración de la densidad espectral de potencia de la señal limpia y del ruido, respectivamente. Usualmente se toma:

$$\hat{\Gamma}_s(k,n = |S(k,n)|^2 \quad (2.11)$$

$$\hat{\Gamma}_N(k,n) = |N(k,n)|^2 \quad (2.12)$$

Dado que no se conoce la densidad espectral de potencia de la señal de voz limpia, en la práctica, el filtro debe reducirse a alguna aproximación de la fórmula anterior. Una alternativa simple es utilizar estimadores similares a los utilizados en sustracción espectral. En este caso, ambas técnicas quedan vinculadas por una relación de cuadrados. Otro enfoque consiste en estimar el espectro de la señal limpia de forma iterativa, modelando la voz como la respuesta de un sistema autorregresivo [26]. Un tercer enfoque consiste en reformular la ecuación (2.10) en términos de la SNR, y emplear un estimador de la SNR, como los usados en las reglas [27] A pesar de su sencillez, estos estimadores han mostrado tener muy buenas propiedades.

La implementación de filtrado óptimo escogida para este trabajo hace uso de un algoritmo de detección de pausas [28], este algoritmo primeramente estima la potencia de la envolvente de la señal *E(p)* como la sumatoria al cuadrado de las componentes espectrales de tiempo corto para señal a la entrada dividida en tramas *p* de longitud 20ms con solapamiento de un 50% enventanadas utilizando una ventana de Hamming :

$$E(p) = \sum_k |X(p,wk)|^2 \quad (2.13)$$

aquí *X(p,wk)* denota las componentes espectrales de la señal ruidosa a la entrada en la frecuencia *wk*, para la trama *p*. Posteriormente se calculan las potencias de las envolventes de *X(p,wk)* para cuando esta es filtrada paso-bajo y paso-alto.

$$E_{LP}(p) = \sum_l |X(p,wl)|^2 \quad (2.14)$$

$$E_{HP}(p) = \sum_m |X(p,wm)|^2 \quad (2.15)$$

donde l y m representan todas las componentes espectrales por debajo y por encima de la frecuencia de corte respectivamente.

La diferencia entre el filtrado de Wiener y la sustracción espectral de potencias es que el filtrado de Wiener usa los valores esperados y la sustracción espectral usa los valores instantáneos. Son similares en la práctica aunque las filosofías subyacentes difieren.

2.1.2.1 Implementación.

El método implementado para la reducción de ruido después de detectar los momentos de silencio o pausas en el habla, hace uso del filtrado de Wiener para eliminar el ruido presente en la señal del habla. En caso de que se detecte que la trama que se está analizando es silencio, es decir ruido, simplemente se realiza el filtrado de Wiener y se calcula la transformada inversa de Fourier para señal S(p,K), y posteriormente utilizando medidas de solapamiento y agregación finalmente se obtiene la señal realzada. En caso de que no sea silencio la trama actual, es decir, que se esté en presencia de actividad vocal, se deberá estimar el ruido. La estimación de este ruido se define como:

$$|N''(p,k)| - |N(r,k)|/pc * \sum_{r'=1}^{r'=r-1} |N(r',k)| \qquad (2.16)$$

Donde pc es la fracción declarada anteriormente y que ahora es utilizada como peso, |N(r,k)| representa las tramas de silencio detectadas anteriormente. Luego se realiza el filtrado de Wiener y se calcula la transformada inversa de Fourier para señal S(p,K), para posteriormente, utilizando medidas de solapamiento y agregación, obtener finalmente la señal realzada. En la figura 2.3 se muestra una representación del diagrama en bloques del algoritmo de reducción de ruido utilizando el filtro de Wiener y el detector de pausas del habla.

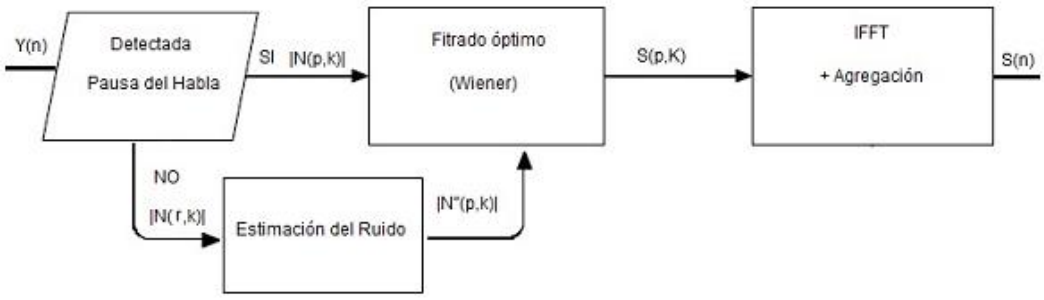


Figura 2.3.Diagrama en bloque del algoritmo de reducción de ruido utilizando Filtrado de Wiener y Detección de pausas en el habla.

2.1.3 Algoritmo Geométrico de sustracción espectral.

El enfoque de sustracción espectral basado en principios geométricos [9] aborda dos defectos muy importantes de la sustracción espectral: el ruido musical y las suposiciones de existencia de términos cruzados. El enfoque se basa en la representación del espectro del habla con ruido en el plano complejo como la suma del vector espectral ruidoso y el vector

espectral señal limpia. Representar el espectro del habla ruidoso geométricamente en el plano complejo provee de una perspicacia al método, la cual no resulta común en los demás enfoques de sustracción espectral.

En el enfoque de sustracción espectral utilizando algoritmos geométricos (AG_SE), al igual que en el método de MBSS después de proponer la ecuación (2.1), se calcula la transformada de Fourier de tiempo corto a la señal del habla con ruido (véase ecuación (2.17)), donde X(wk) y D(wk) son el espectro del habla limpia y del ruido respectivamente para $\mathrm{wk} = 2\pi\mathrm{k}/\mathrm{N}$ y $\mathrm{k} = 0,1,2 \dots \dots \dots \mathrm{N} - 1$, donde N es la longitud en muestras de la trama.

$$\mathrm{Y(wk)} = \mathrm{X(wk)} + \mathrm{D(wk)} \qquad (2.17)$$

Luego de multiplicar Y(wk) en la ecuación anterior por su conjugada $\mathrm{Y}'(\mathrm{wk})$ y aproximando los términos resultantes que no pueden ser obtenidos directamente, se puede calcular la estimación del espectro del habla limpia [29].

En la ecuación (2.17) se muestra como Y(wk) para la frecuencia wk es obtenida mediante la sumatoria de dos valores complejos espectrales a la frecuencia wk. Sin embargo, Y(wk) puede ser representada en el plano complejo como la suma de dos números complejos, X(wk) y D(wk). En la Figura 2.4 se muestra la representación de Y(wk) como el vector adición de X(wk) y D(wk).

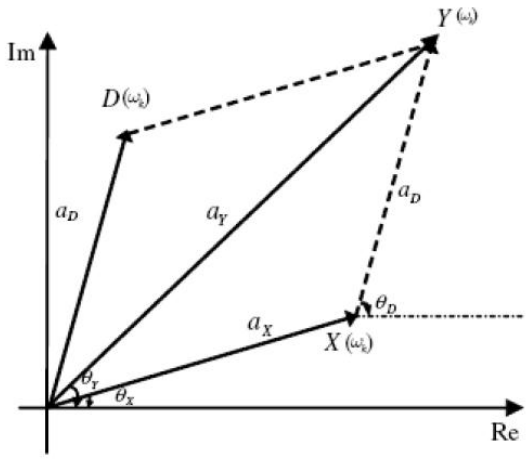


Figura 2.4. Representación del espectro del habla contaminada Y(wk) en el plano complejo como la suma del espectro de la señal limpia X(wk) y el espectro del ruido D(wk).

H(wk) es comúnmente usada como función de ganancia del algoritmo de sustracción espectral de potencia y es obtenida en [29], después de ser asumido que los términos cruzados son ceros o equivalentemente, que la diferencia de fase $[\theta_x(k) - \theta_D(k)]$ es igual $a \pm \pi/2$. Teniendo esto en cuenta, primeramente se escribe $y(n) = x(n) + d(n)$ en forma polar:

$$a_Y e^{j\theta_Y} = a_X e^{j\theta_X} + a_D e^{j\theta^{j\theta_D}} \qquad (2.18)$$

Donde $\{a_Y, a_X, a_D\}$ son las magnitudes y $\{\theta_Y, \theta_X, \theta_D\}$ son las fases del espectro del habla ruidoso, habla limpia y ruido respectivamente.

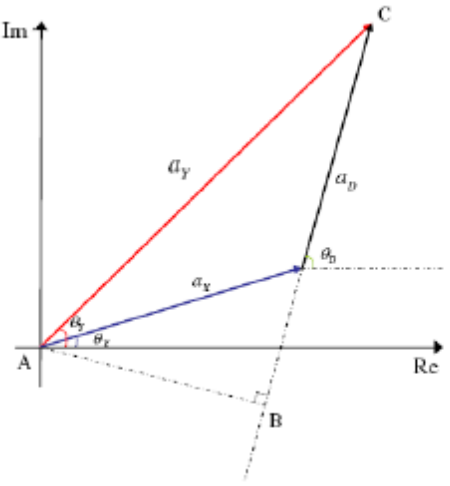


Figura 2.5. Triángulo que muestra la relación entre las fases de habla con ruido, ruido y habla limpia espectral.

Para el triángulo mostrado en la Figura 2.5, usando la ley de los senos se puede obtener una nueva función de ganancia:

$$H_{GA} = \frac{a_X}{a_Y} = \sqrt{\frac{1-C_{YD}^2}{1-C_{XD}^2}} \qquad (2.19)$$

La función (2.19) depende del cálculo de las diferencias de fase entre el ruido y la señal de ruido. Uno de las posibilidades para obtener y utilizar las diferencias de fases es a través de los principios trigonométricos. Dicho lo anterior se puede calcular C_{YD} y C_{XD}.

$C_{YD} = \frac{\gamma+1-\xi}{2\sqrt{\xi}}$ (2.20)

$$C_{XD} = \frac{\gamma - 1 - \xi}{2\sqrt{\xi}} \qquad (2.21)$$

Donde las variables γ y ξ son definidas como:

$$\gamma \triangleq \frac{a_{Y^2}}{a_{D^2}} \qquad (2.22)$$

$$\xi \triangleq \frac{a_{X^2}}{a_{D^2}} \qquad (2.23)$$

Note que los términos γ y ξ son versiones instantáneas de la SNR a posteriori y a priori respectivamente. Luego, mediante la sustitución de ecuaciones se obtiene la siguiente función de supresión en función de los términos γ y ξ:

$$H_{GA}(\xi, \gamma) = \sqrt{\frac{1 - \frac{(\gamma + 1 - \xi)^2}{4\gamma}}{1 - \frac{(\gamma - 1 - \xi)^2}{4\xi}}} \qquad (2.24)$$

2.1.3.1 Implementación.

La ganancia obtenida en la ecuación (2.24) es ideal, en la práctica esta necesita ser estimada por la observación del ruido. La implementación de la función de ganancia requiere estimaciones de γ y ξ. De acuerdo a las ecuaciones (2.22) y (2.23), γ y ξ son valores instantáneos y de corto plazo. Para calcular ξ, se propone usar información espectral actual tanto como información espectral pasada. Más específicamente, se puede utilizar la magnitud mejorada de espectro obtenido en la trama anterior y aproximar ξ como:

$$\xi_I(\lambda, k) = \frac{a_{Y^2(\lambda-1,k)}}{a_{D^2(\lambda-1,k)}} \qquad (2.25)$$

Donde $\xi_I(\lambda, k)$ indica la estimación de ξ en la trama λ y muestra k, el subíndice I indica la medición instantánea. La estimación anterior de los valores instantáneos de ξ solo utiliza información espectral pasada. También se puede utilizar la relación existente entre los valores verdaderos de γ y ξ para obtener un estimado de ξ basado en la información espectral válida para la trama presente. Combinando las dos estimaciones de ξ se puede entonces obtener información espectral pasada y presente mediante la ecuación siguiente:

$$\xi(\lambda, k) = \alpha \left[\frac{a_Y(\lambda-1,k)}{a_D(\lambda-1,k)}\right] + (1 - \alpha) * (\sqrt{r(\lambda, k)} - 1)^2 \qquad (2.26)$$

Donde α es la constante de suavizado, y $a_D(\lambda - 1, k)$ es un estimado de la magnitud espectral del ruido. La ecuación (2.26) es un promedio ponderado de las mediciones

instantáneas de la SNR pasada y presente y la constante de suavizado controla el peso de la información pasada y futura.

Para $r(\lambda, k)$, se usa la siguiente estimación instantánea:

$$r(\lambda, k) = (\frac{a_Y(\lambda,k)}{a_D(\lambda,k)})^2 \qquad (2.27)$$

Donde $a_D(\lambda, k)$ es una estimación del espectro de ruido obtenido usando un algoritmo de estimación de ruido. Se considera el suavizado y el limitado de los valores de para reducir las fluctuaciones rápidas relacionadas con el cálculo anterior de y también limitar la sobre-supresión producto a valores grandes de $r(\lambda, k)$. Entontes se calcula el suavizado de como:

$$Y_{GA}(\lambda, k) = \beta * Y_{GA}(\lambda - 1, k) + (1 - \beta) * min[\gamma_I(\lambda, k), 20] \qquad (2.28)$$

Donde $Y_{GA}(\lambda, k)$ es la estimación del suavizado de γ, $\gamma_I(\lambda, k)$ y β es una constante de suavizado. El operador min se usa para limitar el valor de $\gamma_I(\lambda, k)$ al máximo de $13db = 10\log_{10}(20)$ y evitar así la sobre-atenuación de la señal.

Las estimaciones anteriores de γ y ξ (es decir $Y_{GA}(\lambda, k)$ y $\xi(\lambda, k)$) son usadas para aproximar la función de ganancia (2.22). En principio, la función de transferencia obtenida en (2.22) está basada en los valores instantáneos de γ y ξ., en la práctica, como siempre, los valores verdaderos de γ y ξ pueden variar drásticamente de trama en trama, lo cual resulta extremadamente desafiante para estimar esos valores con alto grado de exactitud y confiabilidad.

Por último se decide limitar $H_{GA}(\xi, \gamma)$ a ser siempre menor o igual que 1 por el hecho de que para $H_{GA}(\xi, \gamma) > 1$, el espectro de magnitud del error, puede ser grande.

A modo de resumen y para aumentar la comprensión del AG_SE se muestra su diagrama en bloques.

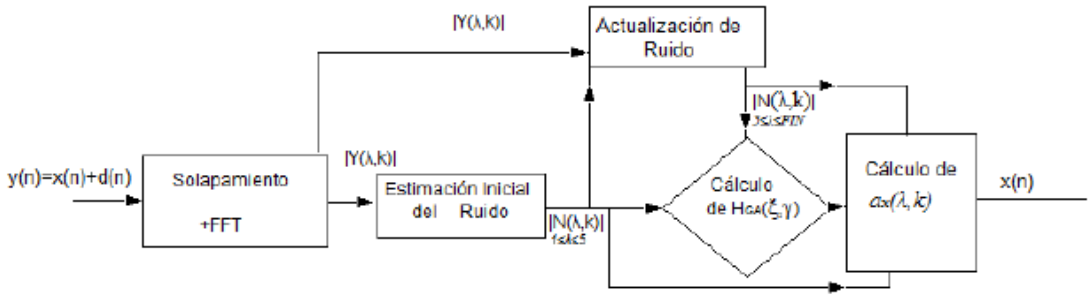


Figura 11. Diagrama en bloque del AG_SE.

2.2 Sistema utilizado.

La herramienta escogida para la implementación del presente proyecto ha sido el programa MATLAB (abreviatura de *MATrix LABoratory*, "laboratorio de matrices" en su versión 7.4). MATLAB es un lenguaje de programación de alto nivel basado en matrices y vectores que gracias a su gran potencia de cálculo y a su agradable entorno, que incorpora la posibilidad de una visualización gráfica de los resultados, hacen que sea la herramienta ideal para el procesamiento de señales [30, 31].

Este software, además de hacer cálculos matemáticos, facilita la implementación de entornos gráficos que hacen posible la interactividad del usuario con los datos que maneja el programa. Por sus características, este software permite realizar cálculos con vectores y matrices de grandes dimensiones, lo cual resulta bastante adecuado para manejar el gran volumen de datos con el que debemos trabajar.

2.3 Características de las bases de datos utilizadas.

Para obtener resultados válidos y repetibles en los experimentos que se proponen realizar para el reconocimiento automático del habla, se seleccionaron grabaciones de audio provenientes de bases de datos internacionales de dominio público. De estas bases de datos se eligieron un subconjunto de alocuciones provenientes de un escenario limpio y en presencia de información ruidosa proveniente de escenarios clásicos, como restaurantes, mercados, estaciones de transporte, automóviles, etc. Esta selección permite analizar el comportamiento del reconocedor automático del habla en situaciones reales. Se utilizaron alocuciones provenientes de bases de datos en inglés y español, grabadas en contextos de mínimo ruido.

El material de habla utilizado para las pruebas de evaluación de los algoritmos fueron las bases de datos TiDigits y Noizeus. A continuación se explicarán las características principales de algunas de ellas.

2.3.1 TIDigits.

La base de datos TIDigits fue elaborada con el propósito de diseñar y evaluar los algoritmos para facilitar el reconocimiento.

Contiene grabaciones de alocuciones en inglés en un entorno limpio, esta posee 12546 archivos de audio de 4 tipos de voces: hombre, mujer, niño y niña, en total son 326 portavoces (111 hombres, 114 mujeres, 50 niños y 51 niñas) cada uno de ellos pronuncia 77 sucesiones de dígitos. Cada grupo del portavoz se divide en condición de prueba y condición de entrenamiento, con una frecuencia de muestreo de 20kHz.

2.3.2 NOIZEUS.

La base de datos Noizeus [32] al igual que TIDgits fue desarrollada para facilitar la comparación de algoritmos con vistas a la mejoría del reconocimiento automático del habla.

Noizeus contiene 30 frases a distintos valores de SNRs. Estos ruidos fueron tomados de la base de datos AURORA, de la base de datos de IEEE para incluir todos los fonemas en el idioma inglés americano, producidas por 3 hombres y 3 mujeres, contaminadas con 8 tipos de ruidos reales a diferentes incluyen el ruido de tren, de automóvil, restaurante, aeropuerto, entre otros. Las frecuencias de muestreo utilizadas se encuentran en un intervalo de 25 kHz y 8 kHz.

2.4 Método para la evaluación de los algoritmos.

Es evidente que para comparar los resultados obtenidos con diferentes métodos de filtrado es necesario disponer de un método de evaluación que permita determinar la calidad de cada método a evaluar. Al igual que se hace en gran parte de la bibliografía consultada, se utilizó un criterio estrictamente visual. Para ello se desarrollaron un conjunto de funciones que muestran ambas señales y sus correspondientes espectrogramas.

Se realizarán experimentos controlados en los cuales señales de habla limpia serán contaminadas con ruidos reales. Específicamente se eligieron dos tipos de ruido: Ruido de

Múltiples Parlantes (Babble) y Ruido de Automóvil (Car), a múltiples SNRs: 0dB, 5dB y 10db en correspondencia con estudios realizados por otros autores [29]. Las señales ruidosas han sido tomadas de la base de datos NOIZEUS.

Para determinar los algoritmos más recomendados a emplear en la propuesta de reconocimiento automático del habla, se sugiere el empleo de sistemas de reconocimiento del habla, como es el caso de Sphinx3 [33], el cual posee un buen desempeño ante las voces libres de ruido y es de acceso público.

Los sistemas de reconocimiento automático del habla con arquitectura integrada suelen constar de dos módulos básicos: el módulo de parametrización y el módulo de clasificación o reconocedor propiamente dicho. El primero se encarga de la extracción de una serie de parámetros o rasgos acústicos que son una representación compacta de la señal de voz. El segundo compara dichos parámetros acústicos con los modelos acústicos de cada sonido y decide la palabra o frase reconocida con mayor probabilidad. Ambos módulos son susceptibles de ser modificados para aumentar la robustez del sistema completo a las diferentes distorsiones antes mencionadas.

La selección de rasgos acústicos apropiados es quizás la tarea más importante en el plan de un sistema de reconocimiento de discurso robusto, como él directamente afecta la actuación del sistema. Estos rasgos deben seleccionarse con los criterios siguientes:

- Ellos deben contener la información máxima necesaria para el reconocimiento del discurso.
- Ellos deben ser insensibles a las características del portavoz, la manera de hablar, el ruido del fondo, entre otros.
- Nosotros debemos poder estimarlos con precisión y fiablemente.
- Nosotros debemos poder estimarlos a través de su eficacia computacional en el procedimiento.
- Ellos deben tener un significado físico (preferentemente consistente con el proceso de percepción auditorio del humano).

Obviamente, es muy difícil de seleccionar un juego de rasgos acústicos de que satisfacen todos estos requisitos.

La mayoría de las técnicas propuestas para mejorar las prestaciones de los reconocedores de habla en entornos ruidosos están orientadas a minimizar los efectos de los desajustes que se producen entre las condiciones acústicas en las se entrenan los modelos y las condiciones reales en las que opera el sistema. En consecuencia, las soluciones más habituales consisten en el desarrollo de parametrizaciones robustas frente al ruido, el entrenamiento de los modelos acústicos con voz contaminada o su adaptación al ruido ambiental.

Las parametrizaciones robustas son, por tanto, el conjunto de técnicas que se aplican sobre el módulo de parametrización del sistema de RAH, para conseguir que las prestaciones del sistema no se degraden en presencia de diversos tipos de distorsiones.

En este trabajo se usa una parametrización convencional basada en 12 coeficientes cepstrales MFCC (Mel-frequency cepstral coeficient), los coeficientes cepstrales se normalizan fichero a fichero, lo que mejora las prestaciones de los sistemas en ambientes ruidosos, donde las condiciones de entrenamiento y prueba no coinciden [18].

La evaluación de los algoritmos se realizará tomando como base el desempeño del sistema RAH ante señales limpias, ante señales contaminadas con los ruidos seleccionados a diferentes SNR y, ante señales realzadas con cada uno de los algoritmos. El desempeño del sistema RAH de las señales realzadas que más semeje al desempeño ante señales limpias será la métrica principal que permitirá la elección adecuada. Es posible además que no exista un método superior al resto, en cuyo caso se realizarán los análisis correspondientes para determinar en qué escenarios cada método tiene mayores perspectivas de éxito.

CAPÍTULO 3. RESULTADOS Y DISCUSIÓN.

El objetivo de este capítulo es comparar el desempeño y la efectividad de los algoritmos de reducción de ruido para el reconocimiento robusto del habla.

3.1 Desempeño de los algoritmos de reducción de ruido.

Se realizó un experimento que contempla el desempeño del algoritmo geométrico de sustracción espectral (AG_SE), el algoritmo de sustracción espectral multi-banda (MBSS) y el Filtrado de Wiener utilizando detección de pausas en el habla (FW), frente a ruidos típicos y simulados. Las señales fueron contaminadas artificialmente con dos tipos de ruido: murmullo (babble) y ruido presente en autos (car). Se tomaron en consideración los siguientes niveles de SNR: 0dB, 5dB y 10dB.

Con el fin de apreciar el desempeño de estos algoritmos de reducción de ruido, se ofrecen gráficas de los resultados obtenidos para cada tipo de ruido con los que se contaminaron las señales analizadas. En las Figuras 3.1, 3.2 y 3.3 muestran las formas de onda de la primera señal utilizada para la comparación de los algoritmos: señal limpia en la parte superior (SNR = ∞), la señal contaminada con ruido car y una SNR de 0dB en el centro y la señal filtrada por el método de reducción de ruido basado en MBSS, AG_SE y FW respectivamente (Ver Anexo I para observar las formas de ondas de la señal contaminada con ruido Babble). Se eligió una SNR = 0dB, porque es el caso más crítico para el que se evaluaron los algoritmos.

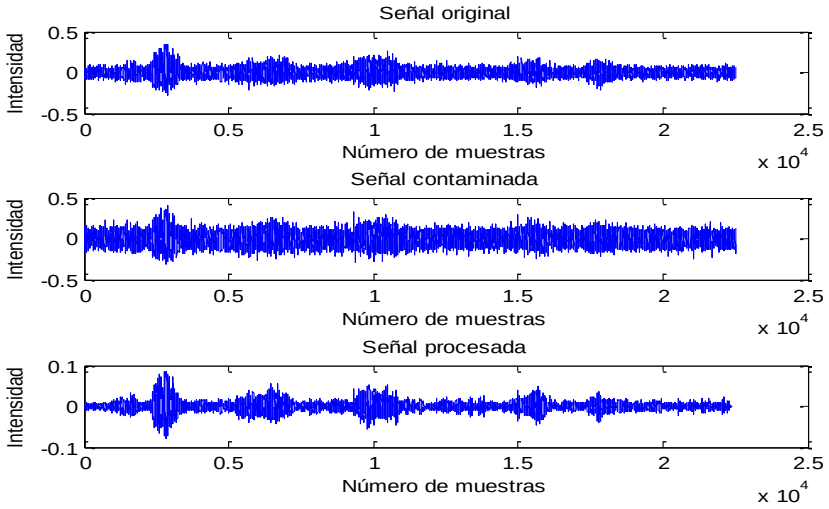


Figura 3.1. Formas de onda de la señal limpia, contaminada (Ruido Car y SNR=0dB) y filtrada por el algoritmo basado en MBSS.

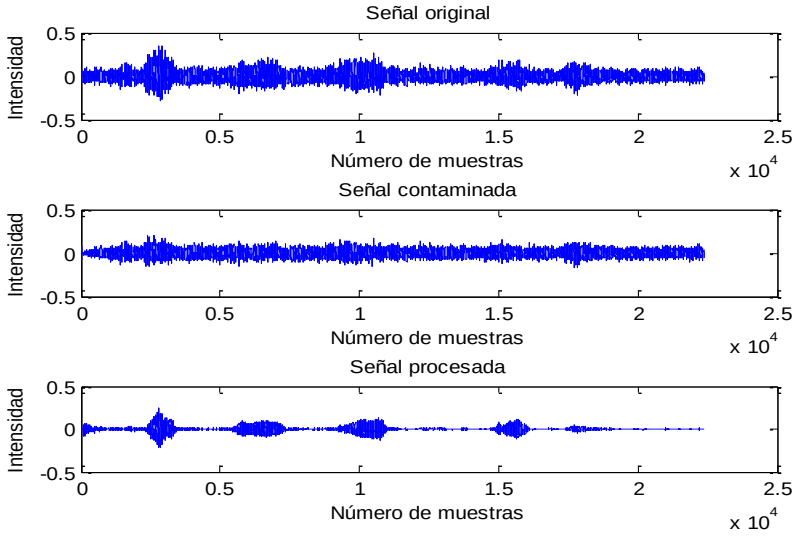


Figura 3.2. Formas de onda de la señal limpia, contaminada (Ruido Car y SNR=0dB) y filtrada por el algoritmo basado en AG_SE.

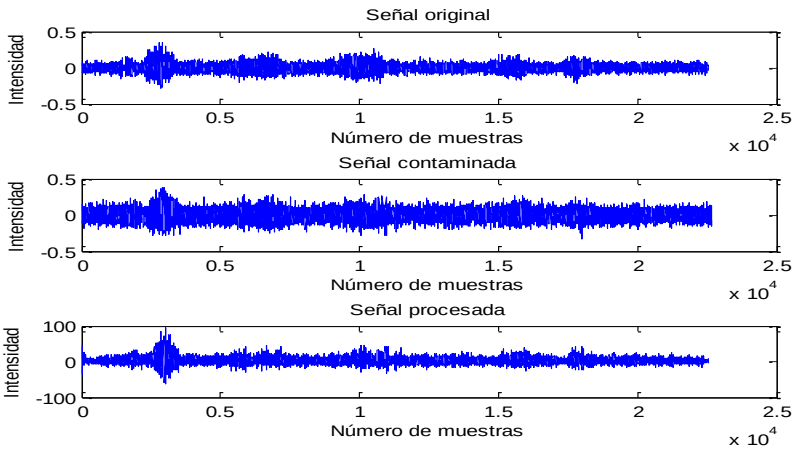


Figura 3.3. Formas de onda de la señal limpia, contaminada (Ruido Car y SNR=0dB) y filtrada por el algoritmo basado en FW.

Con el objetivo de poder definir mejor las pérdidas por procesamiento de los métodos de reducción de ruido, las Figuras de la 3.4 a la 3.10 muestran espectrogramas que expresan el comportamiento de estos algoritmos ante las SNR extremas (SNR=0dB y SNR=10dB) utilizadas en el experimento con ruido car. (Véase Anexo II para contaminación con ruido Babble).

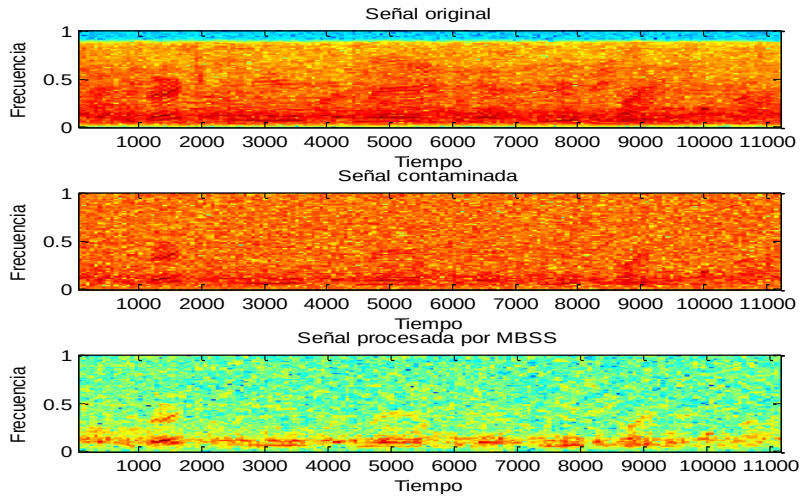


Figura 3.4. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=0dB) filtrada por el algoritmo de MBSS.

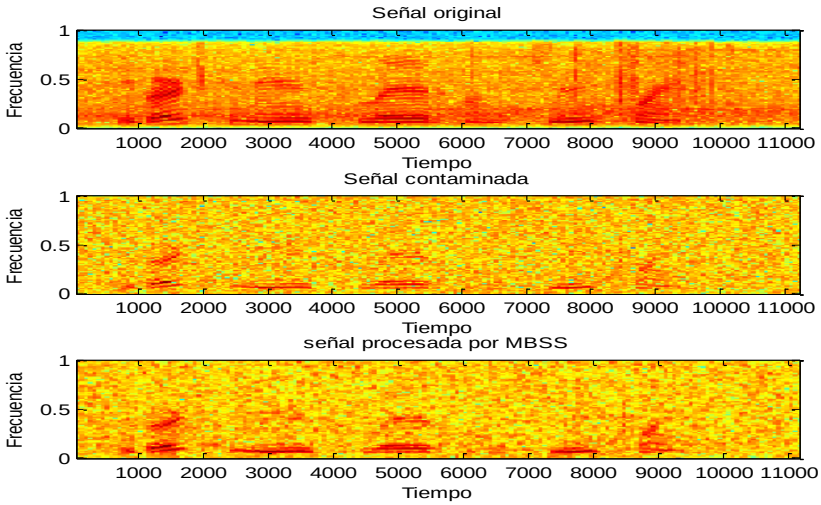


Figura 3.5. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=10dB) filtrada por el algoritmo de MBSS.

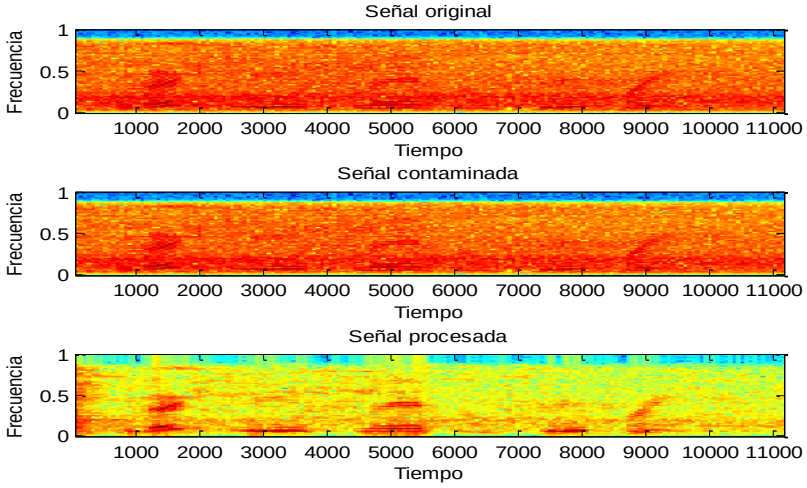


Figura 3.6. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=0dB) filtrada por el algoritmo de AG_SE.

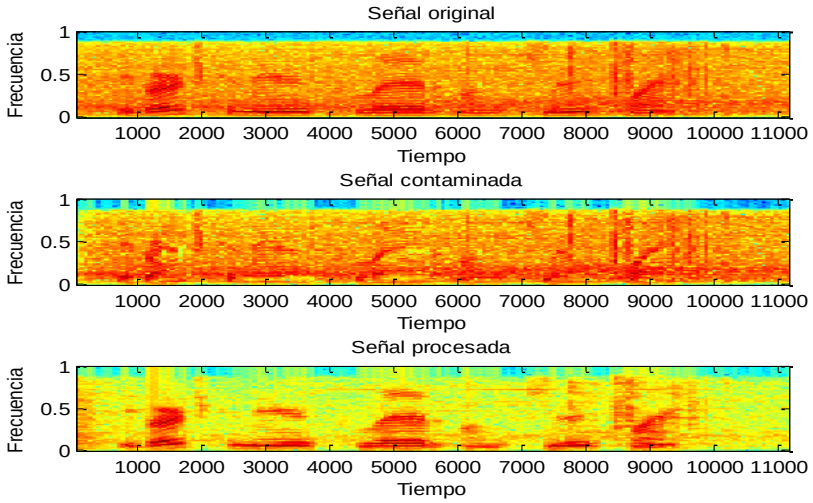


Figura 3.7. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=10dB) filtrada por el algoritmo de AG_SE.

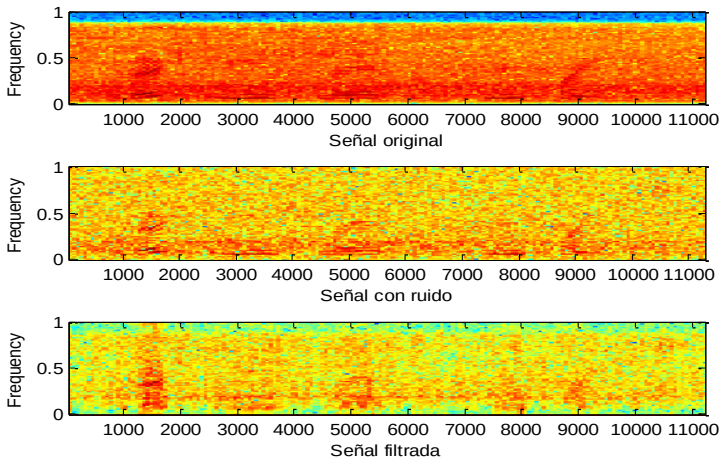


Figura 3.8. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=0dB) filtrada por el algoritmo de FW.

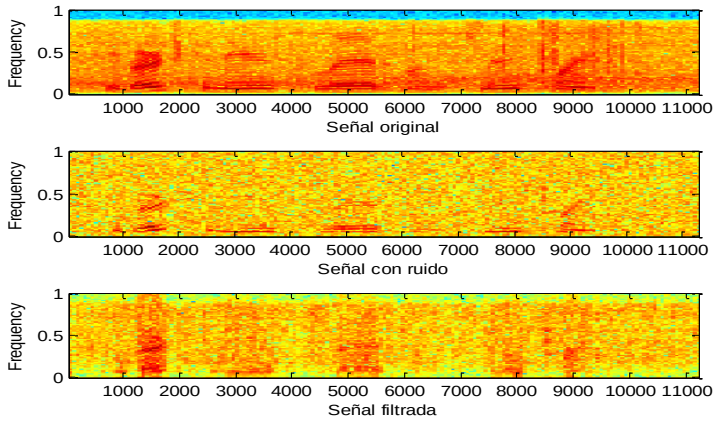


Figura 3.9. Espectrogramas de la señal limpia, señal contaminada (Ruido Car y SNR=10dB) filtrada por el algoritmo de FW.

De los espectrogramas presentados se observa que de los tres algoritmos el AG_SE es el que mejor preserva la información del espectro, incluso al procesar señales con SNR = 0dB (Figura 3.6), en comparación con los otros dos algoritmos. El peor desempeño para los ejemplos graficados se percibe con el filtrado de Wiener (Figuras 3.8 y 3.9), en los que la información espectral está significativamente deteriorada, indicando que no sólo elimina el ruido sino además, información de voz potencialmente importante. Para las demás Figuras los espectrogramas conservan gran parte de la información de la estructura de formantes presentes en la señal de habla.

3.2 Evaluación de los algoritmos de reducción de ruido utilizados.

Se llevaron a cabo varios experimentos con el fin de evaluar los algoritmos de reducción de ruido seleccionados y el sistema de reconocimiento utilizado, después de haber filtrado las señales del habla.

El trabajo de investigación que se presenta en este capítulo se centra en un escenario de operación especialmente adverso para el reconocimiento automático de habla:

Condición de entrenamiento	Condición de prueba
1. Modelos de entrenamiento con la señal del habla limpia.	• Archivos de prueba con la señal limpia.
2. Modelos de entrenamiento con la	• Archivos de prueba con la señal

señal del habla limpia.	contaminada.
3. Modelos de entrenamiento con la señal del habla limpia.	• Archivos de prueba con el filtrado aplicado.

La Tabla 3.1 presenta los resultados del reconocimiento previo a la aplicación de los algoritmos de reducción de ruido. Se ilustran los distintos tipos de error en que incurre el sistema automático de reconocimiento del habla: sustituciones, inserciones y eliminaciones ante tres relaciones señal a ruido (∞ dB, 10dB y 0dB), y dos tipos de ruidos (Babble y Car). Conjuntamente ofrece también las tasas de error de palabra (WER) obtenidas para cada uno de los entornos de prueba. Estos resultados se considerarán de referencia para identificar los efectos de la aplicación de los algoritmos de reducción de ruido. El anexo III presenta los resultados para SNR = 5dB.

Las Figuras 3.10 y 3.11 presentan estos resultados de manera gráfica. La figura 3.10 muestra cómo a 10 dB y para ambos tipos de ruidos el número de sustituciones, inserciones y eliminaciones crecen significativamente pero de manera proporcional a su distribución en el reconocimiento de señales limpias, con un predominio significativo del número de inserciones. Sin embargo, a 0dB el WER se ve más afectado por el número de eliminaciones ante ruidos tipo Car y por el número de sustituciones ante ruidos Babble. Estos resultados pueden explicarse por las características de ambos ruidos. El ruido Car tiene un comportamiento de ruido blanco, mientras que el ruido babble tiene un comportamiento espectral semejante al de las alocuciones a reconocer.

Tabla 3.1: Resultado del reconocedor automático del habla con archivos de prueba

Escenario de prueba	Ruidos	Sustituciones	Inserciones	Eliminaciones	WER (%)
Señal Limpia (∞ dB)	-	97	100	35	0.81
SNR=10dB	Babble	6019	17168	857	89.01
	Car	7952	16162	1220	93.78
SNR=0dB	Babble	13483	3716	9950	112.50
	Car	5821	471	17169	105.45

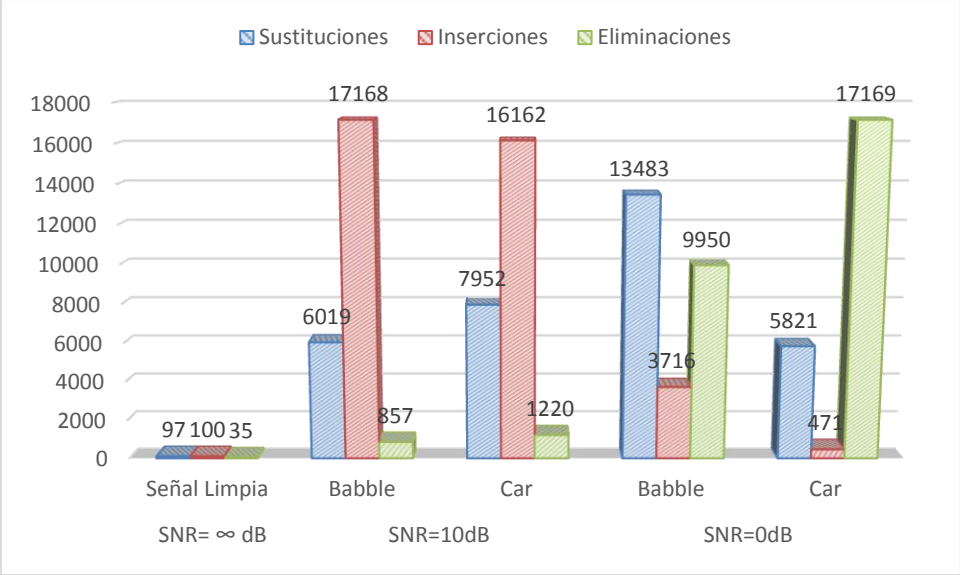


Figura 3.10. Número de inserciones, sustituciones y eliminaciones del SARH ante diferentes SNR y tipos de ruido.

Evidentemente, el desempeño del sistema de reconocimiento se degrada en gran cuantía ante la presencia de ruido si no se emplea ningún mecanismo de compensación del ruido. La tabla 3.1 y la Figura 3.11 muestran cómo se afecta el WER para los diferentes escenarios. De un reconocimiento cercano al 100% de aciertos (WER = 0.81%) ante señales limpias, las tasas de error llegan a ser superiores al 100% a SNR = 0dB y cercanas a esta cifra a SNR = 10dB. Es de esperar que la aplicación de los métodos de reducción de ruido permita reducir estas tasas de error.

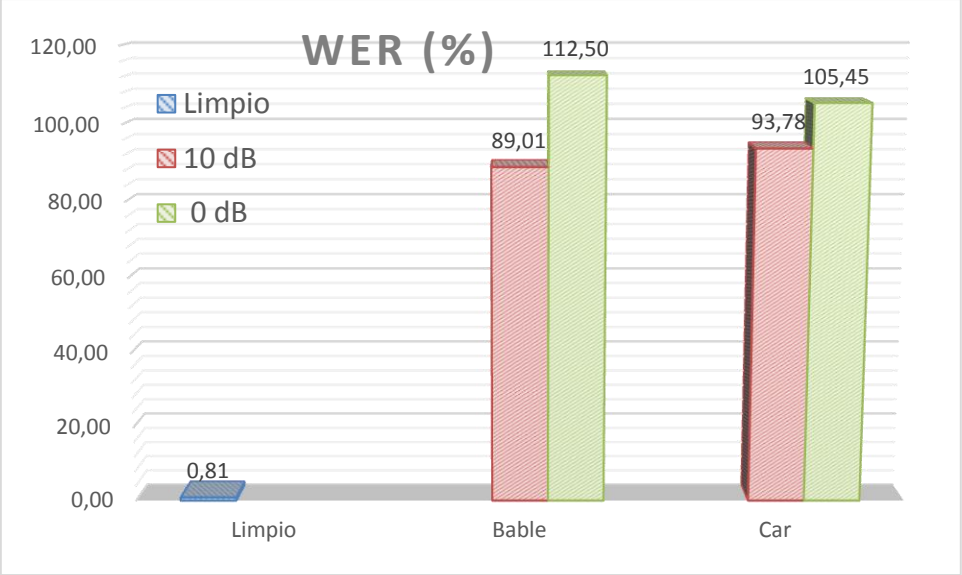


Figura 3.11. Tasa de error de palabras del SARH ante diferentes SNR y tipos de ruido.

En la Tabla 3.2 se presentan las medidas de desempeño empleadas anteriormente luego de aplicar los algoritmos de reducción de ruido bajo estudio. Podemos apreciar que a pesar de las capacidades de las técnicas del filtrado, los resultados, si bien reducen las tasas de error WER en todos los casos, todavía resultan elevadas, sugiriendo que han de tomarse otras alternativas para minimizar los efectos del ruido.

Tabla 3.2: Resultado del reconocedor automático del habla con archivos de prueba con el filtrado aplicado.

	Ruidos	**Sustituciones**	**Inserciones**	**eliminaciones**	**WER (%)**
AG_SE SNR=0dB	Babble	9292	7433	3502	74.88
	Car	7478	5833	3851	63.53
MBSS SNR=0dB	Babble	7865	4828	15811	92.52
	Car	11399	11842	1968	93.32
FW SNR=0dB	Babble	12085	2552	9121	87.95
	Car	2131	10558	10552	86.03
AG_SE SNR=10dB	Babble	3342	295	6003	35.69
	Car	2665	21	7747	38.62
MBSS SNR=10dB	Babble	5837	9999	1671	64.81
	Car	5502	10387	1236	63.39
FW SNR=10dB	Babble	8416	4710	3765	62.53
	Car	7031	3227	391	52.45

En correspondencia con los resultados presentados en el epígrafe 3.1, donde se inspeccionó visualmente, y a través de un ejemplo, el comportamiento de los diferentes algoritmos; los resultados sistema de reconocimiento del habla favorecen también al algoritmo AG_SE. Observando la Figura 3.12 resulta evidente que el método AG_SE reduce grandemente el número de inserciones en comparación con los otros métodos, también son menores el número de sustituciones, aunque el número de eliminaciones es superior al resto de los métodos. Estos resultados inducen a pensar que el algoritmo elimina bien el ruido pero a su vez elimina información útil importante que conduce al elevado número de eliminaciones.

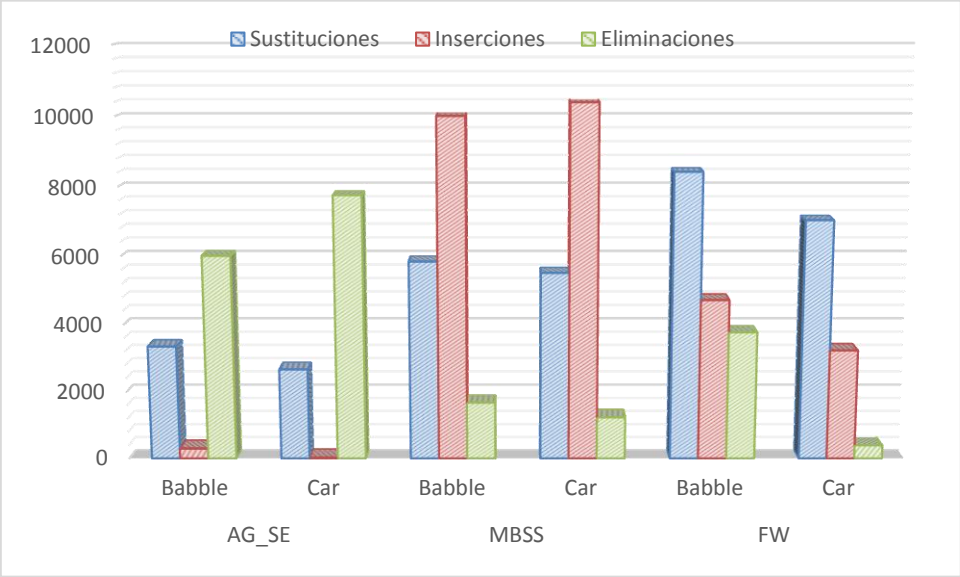


Figura 3.12 Número de Sustituciones, Inserciones y Eliminaciones luego de aplicar los algoritmos de filtrado a señales con SNR = 10 dB

A SNR=0dB, el comportamiento de los métodos se afecta, aunque sigue siendo AG_SE, el de mejor desempeño y consistencia independientemente del tipo de ruido (véase Figura 3.13). En comparación, los otros métodos exhiben comportamientos diferentes en dependencia de si el ruido es de características de ruido blanco (Car) o coloreado (Babble).

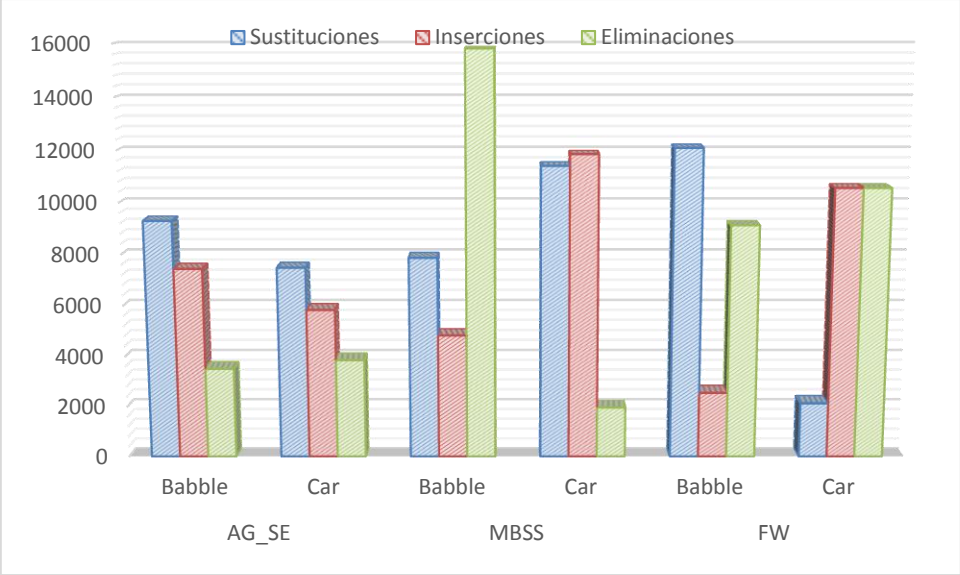


Figura 3.13 Número de Sustituciones, Inserciones y Eliminaciones luego de aplicar los algoritmos de filtrado a señales con SNR = 0dB

Las tasas de error que logra el SARH ante señales filtradas, no permiten concluir que la sola aplicación del algoritmo de reducción de ruido pueda conducir a un sistema de reconocimiento confiable, pues en todos los casos, las cifras WER superan el 60%, lo que implica menos de un 40% de aciertos (Véase Figura 3.14)

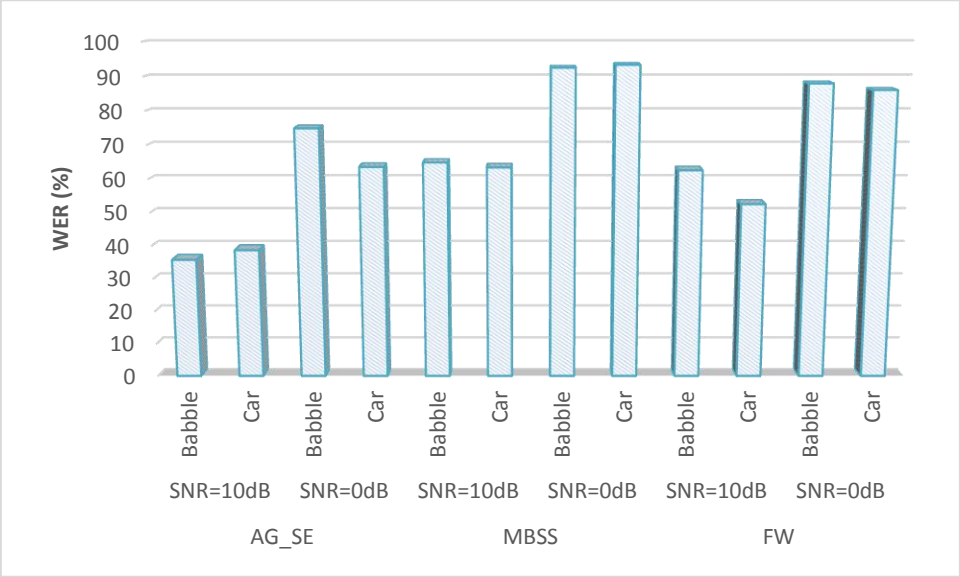


Figura 3.14. Tasa de error de palabras del SARH ante diferentes SNR y tipos de ruido luego de filtrar las señales con los algoritmos de reducción de ruido.

Los resultados obtenidos permiten afirmar que el empleo de estos algoritmos, aunque insuficientes desde el punto de vista del reconocimiento, sí reducen en un porciento elevado los errores de clasificación. Es recomendable para trabajos futuros, realizar experimentos en los que se combine el filtrado en el dominio de la señal, con otras técnicas aplicadas en etapas subsiguientes del SARH, con vistas a lograr otros incrementos de la calidad del sistema.

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

En la presente tesis hemos estudiado la problemática del reconocimiento automático del habla en entornos realistas caracterizados por la presencia de distorsiones de distinta naturaleza que afectan negativamente a la tasa de reconocimiento del sistema. Las conclusiones que hemos extraído al concluir este proyecto son las que describimos a continuación.

- En los anteriores capítulos se han mostrado los fundamentos teóricos de las decisiones tomadas para el desarrollo de esta tesis y cómo éstas se han llevado a cabo en la implementación, sin entrar exhaustivamente en detalles del código proporcionado por los autores de los mismos. Además, se ha comprobado el correcto funcionamiento del reconocedor realizando experimentos con las base de datos Noizeus y TiDigits.
- Se realizó un experimento para analizar el desempeño del Algoritmo Geométrico de Sustracción Espectral (GA_SE), el Algoritmo de Sustracción Espectral Multi-Banda (MBSS) y el Filtrado de Wiener (FW), frente a ruidos típicos y simulados y ante diferentes niveles de SNR y se concluyó que el de mejor desempeño fue Algoritmo Geométrico para la Sustracción Espectral (GA_SE).
- Por los resultados presentados podemos afirmar que el empleo de los algoritmos de reducción de ruido AG_SE, MBSS y FW, aunque no se obtuvo la tasa de error de palabra (WER) esperada, sí reducen un porciento elevado de clasificación.

RECOMENDACIONES

- Realizar experimentos en los que se combine el filtrado en el dominio de la señal, con otras técnicas aplicadas en etapas subsiguientes del SARH, con vistas a lograr otros incrementos de la calidad del sistema.
- Utilizar otros escenarios de entrenamiento, para evaluar la robustez del sistema de reconocimiento del habla.

- Tomarse otras alternativas para minimizar los efectos del ruido y de esta forma reducir las tasas de error WER, ya que en este proyecto a pesar de que las técnicas de filtrado utilizadas reducen las tasas de error WER de forma significativa, aun estas resultan elevadas.

REFERENCIAS BIBLIOGRÁFICAS

1. Loizou., P.C., *Speech enhancement. Theory and practice.* 2007.
2. Ascensión Gallardo Antolín, J.M.G., Rubén San Segundo Hernández, Javiern Ferreiros López y José Manuel Pardo Muñoz., *Técnicas de robustez frente al ruido para sistemas de reconocimiento de habla en teléfonos móviles y PDAs.*
3. Pericas, F.J.H., *Técnicas de procesado y represetación de la señal de voz para el reconocimiento del habla en ambientes ruidosos.*, in *Departamento de Teoría de la señal y comunicaciones.*1993, Universidad Politécnica de Cataluña
4. Rodríguez., J.L.O., *Técnicas de robustez frente al ruido para sistemas de reconocimiento de habla en teléfonos móviles y PDAs.* 2006.
5. J. C. Segura, M.C.B., Ángel de la torre and A. J. Rubio., *Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR.* 2002.
6. Gallardo., A., *Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo.* 2002.
7. Ascensión Gallardo Antolín, J.M.G., Rubén San Segundo Hernández,Javier Ferreiros López y José Manuel Pardo Muñoz., *Técnicas de robustez frente al ruido para sistemas de reconocimiento de habla en teléfonos móviles y PDAs.* 2004.
8. Solera., R., *Máquinas de vectores soporte para reconocimiento robusto de habla.* 2011.
9. Yang Lu, P.C.L., *A geometric approach to spectral subtraction.* 2008.
10. Marquina., A.Á., *Algoritmos de extracción de características.*
11. Gales, M.J. and S.J. Young, *A fast and flexible implementation of parallel model combination.*, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*1995, ICASSP 1995: Detroit, Michigan, Estados Unidos.
12. Nilsson, M., M. Dahl, and I. Claesson, *HMM-based speech enhancement applied in non-stationary noise using cepstral features and log-normal approximation.* 2003.
13. Siu., M. and Y.C. Chan, *Robust Speech Recognition Against Short-Time Noise.*, in *Proceedings of the 7th International Conference on Spoken Language Processing*2002, ICASLP 2002: Denver, Colorado, Estados Unidos.

14. Ding, P., *Soft Decision Strategy and Adaptive Compensation for Robust Speech Recognition Against Impulsive Noise.*, in *Proceedings of the 9th European Conference on Speech Communication and Technology*.2005, Interspeech 2005: Lisboa, Portugal.

15. Ding., P., *Flooring the Observation Probability for Robust ASR in Impulsive Noise.*, in *Proceedings of the 8th European Conference on Speech Communication and Technology*.2003, EUROSPEECH 2003: Ginebra, Suiza.

16. Martínez., D.L.G., *Ecualización de histogramas en el procesado robusto de la voz.*, in *Departamento de teoría de la señal, telemática y telecomunicaciones.*2007, Universidad de Granada.

17. Tuomas Virtanen, R.S.a.B.R., *Techniques for noise robustness in automatic speech recognition.* 2013.

18. Rueda., L., *Mejoras en reconocimiento del habla basadas en mejoras en la parametrización de la voz.* 2011.

19. Boudy., P.L.a.J., *Experiments whitn a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars.* 1992.

20. Lleida., E. and R.C.Rose, *Efficient decoding and training prodcedures for utterance verification in continuous speech recognition.* 1996.

21. Sukkar, R.A. and C.H. Lee, *Vocabulary independent discriminative utterance verification for nonkeywords rejection in subword based speech recognition.*, in *IEEE*nov. 1996.

22. S.Cox and R.C.Rose, *Confidence measures for the Switchboard database.*, in *ICSLP 1996*1996: Philadelphia.

23. T.Schaff and T.Kemp, *Confidence measures for spontaneous speech recognition.*, in *ICASSP 1997*1997: Munich.

24. Loizou, S.D.K.a.P.C., *A multi-band spectral subtraction method for enhancing speech corrupted colored noise.* 2002.

25. D. R. Tomassi, L.A., C.E. Martínez, D. H. Milone, M.E. Torres y H. L. Rufiner., *Evaluación de técnicas clásicas de reducción de ruido en señales de voz.* 2005.

26. A.O., J.S.L.a., *All-pole modeling of degraded speech.* 1978.

27. I., D.M.E.a.C., *Relaxed statistical model for speech enhancement and a priori SNR estimation.* 2005.

28. Marzinzik., M., *Noise reduction schemes for digital hearing aids and their use for the hearing impaired.* 2000.

29. Jiménez., R., *Desarrollo de una prótesis auditiva digital.* 2010.

30. Arrieta, C.G., *Procedimiento de diseño de algoritmos en Matlab para el análisis de filtros digitales no adaptativos aplicados en el procesamiento de señales con perturbaciones.* 2009.

31. Gené., O.B., *Reductor de ruido mediante resta espectral en entorno Matlab.* 2006.

32. Loizou., P., *http://www.utdallas.edu/~loizou/speech/noizeus*, 2007.

33. Pavón., E.P.P., *Construcción de un reconocedor de voz utilizando Sphinx y el corpus DIMEx100*. 206.

ANEXOS

Con el objetivo de mostrar el funcionamiento de los algoritmos y el diseño en general se presenta gráficamente en estos Anexos el comportamiento de estos frente a otro ruido (babble), las señales utilizadas fueron las mismas.

Anexo I.

Forma de onda de la primera señal contaminada con ruido Babble (SNR= 0dB, ya que es la situación más crítica para la que se trabajó), filtrada con las diferentes técnicas de reducción de ruido utilizadas.

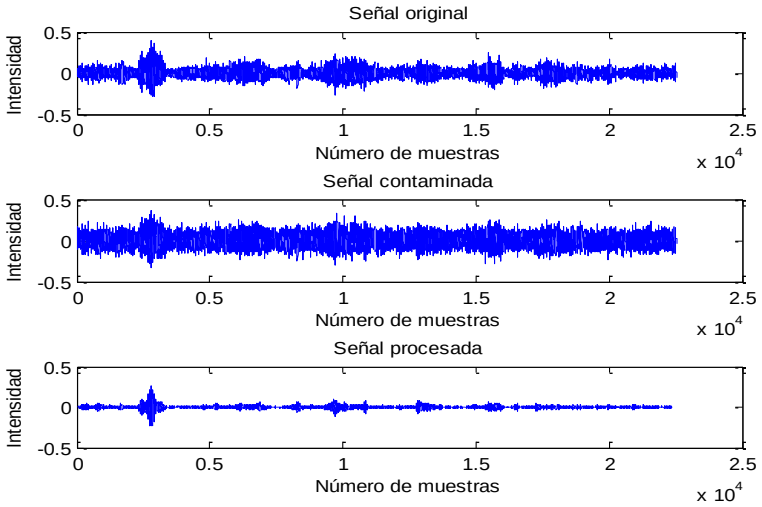


Figura 4.1: Formas de onda de la señal limpia, contaminada (Ruido Babble y SNR=0dB) y filtrada por el algoritmo basado en MBSS.

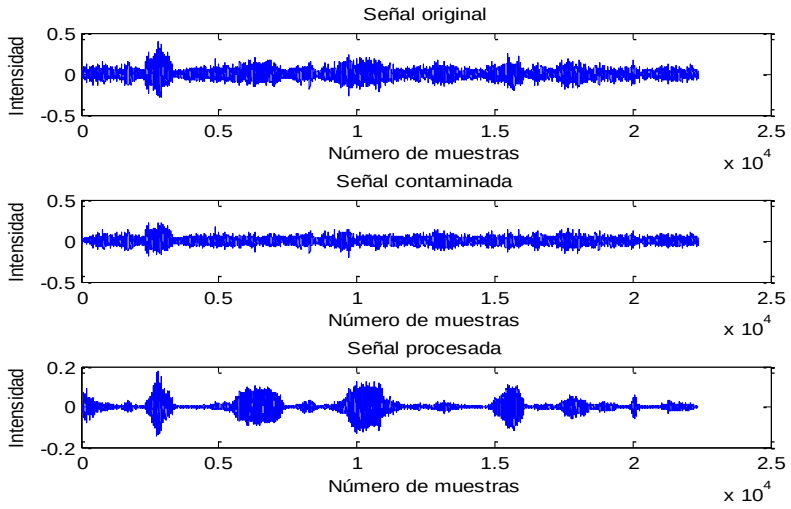


Figura 4.2: Formas de onda de la señal limpia, contaminada (Ruido Babble y SNR=0dB) y filtrada por el algoritmo basado en GA_SE.

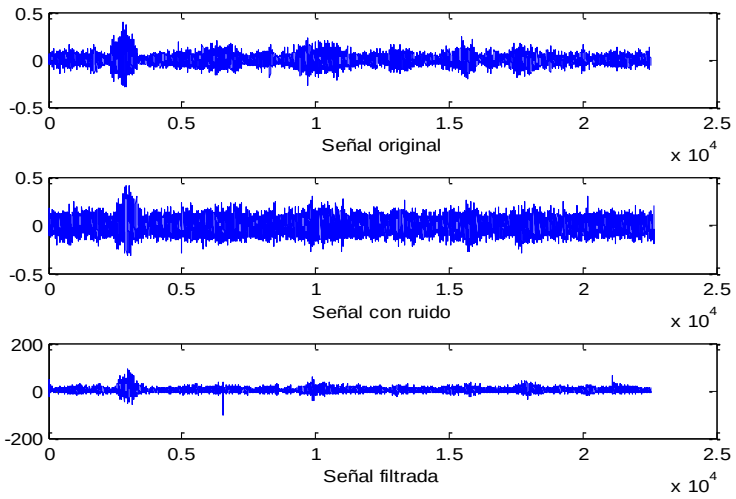


Figura 4.3: Formas de onda de la señal limpia, contaminada (Ruido Babble y SNR=0dB) y filtrada por el algoritmo basado en FW.

Anexo II.

Espectrogramas de la primera señal contaminada con ruido Babble (SNR= 0dB y SNR=10dB), filtrada con las diferentes técnicas de reducción de ruido utilizadas.

Algoritmo Geométrico de Sustracción Espectral (GA_SE)

A continuación se muestra el funcionamiento de AG_SE frente a ruido Babble a diferentes SNR [0 10] dB.

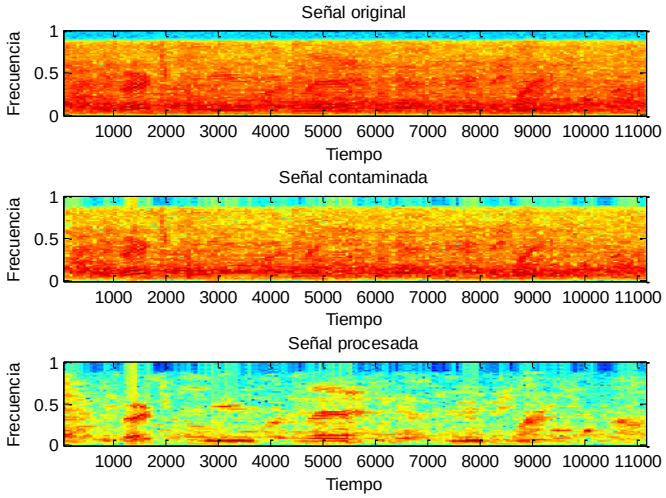


Figura 4.4: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=0dB) filtrada por el algoritmo de AG_SE.

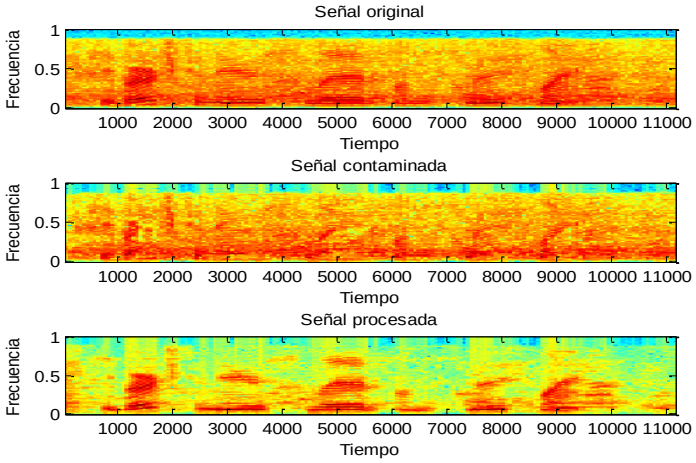


Figura 4.5: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=10dB) filtrada por el algoritmo de AG_SE.

Algoritmo de sustracción espectral multi-banda (MBSS)

A continuación se muestra el funcionamiento de MBSS frente a ruido Babble a diferentes SNR [0 10] dB.

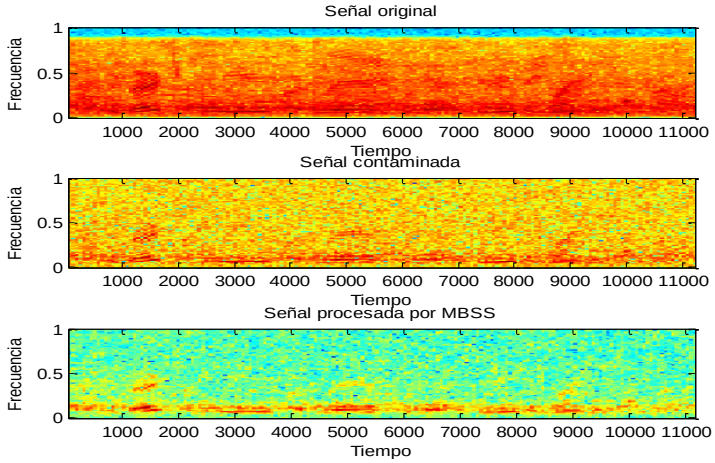


Figura 4.6: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=0dB) filtrada por el algoritmo de MBSS.

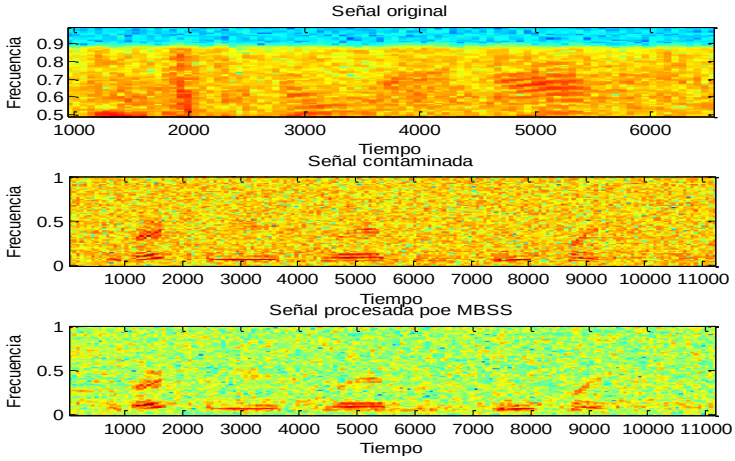


Figura 4.7: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=10dB) filtrada por el algoritmo de MBSS.

Algoritmo de Filtrado de Wiener (FW)

A continuación se muestra el funcionamiento de FW frente a ruido Babble a diferentes SNR [0 10] dB.

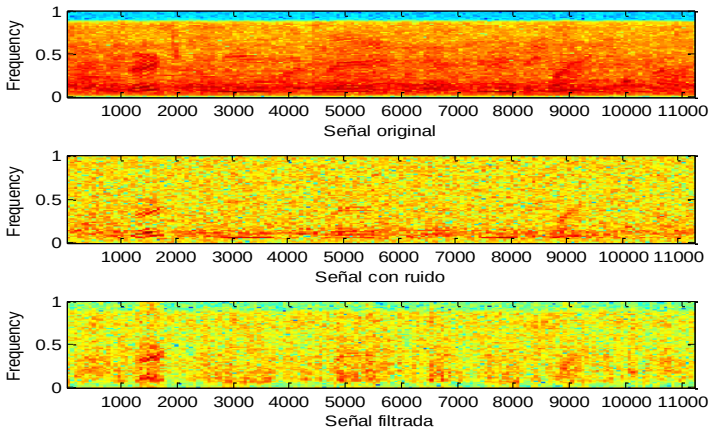


Figura 4.8: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=0dB) filtrada por el algoritmo de FW.

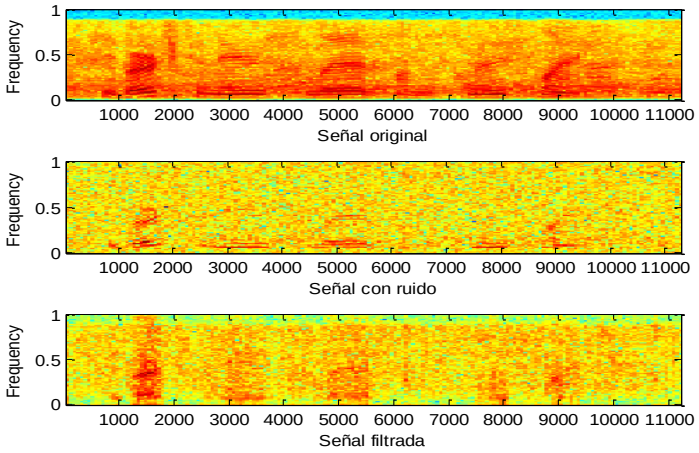


Figura 4.9: Espectrogramas de la señal limpia, señal contaminada (Ruido babble y SNR=10dB) filtrada por el algoritmo de FW.

Anexo III.

Las tablas 4.1 y 4.2 nos muestran la cantidad de inserciones, sustituciones y eliminaciones que produjo este procesamiento, además del porciento de la tasa de error de palabra (WER) obtenido para cada uno de los entornos de prueba y para un SNR=5dB.

Tabla 4.1: Resultado del reconocedor automático del habla con archivos de prueba con la señal contaminada.

Escenario de prueba	Ruidos	Sustituciones	Inserciones	Eliminaciones	WER (%)
Ruidoso SNR=5dB	Babble	2398	131	22552	92.69
	Car	13939	944	4702	102.04

Tabla 4.2: Resultado del reconocedor automático del habla con archivos de prueba con el filtrado aplicado.

	Ruidos	Sustituciones	Inserciones	Eliminaciones	WER (%)
GA_SE SNR=5dB	Babble	5204	8131	1957	56.61
	Car	4089	6325	1843	45.37
MBSS SNR=5dB	Babble	11430	10243	2884	90.90
	Car	16054	7959	4472	90.45
FW SNR=5dB	Babble	10974	4083	6175	78.60
	Car	10036	3380	6342	73.14