UNIVERSIDAD DE GRANADA

NUEVOS AVANCES EN DETECCIÓN DE ACTIVIDAD DE VOZ MEDIANTE HOS Y ESTRATEGIAS DE OPTIMIZACIÓN

TESIS DOCTORAL

Juan Manuel Górriz Sáez

2006

Departamento de Arquitectura y Tecnología de Computadores

Editor: Editorial de la Universidad de Granada
Autor: juan Manuel Gorriz Saez
D.L.: Gr. 1116 - 2006
ISBN: 84-338-3863-6

UNIVERSIDAD DE GRANADA

NUEVOS AVANCES EN DETECCIÓN DE ACTIVIDAD DE VOZ MEDIANTE ESTADÍSTICA DE ALTO ORDEN Y ESTRATEGIAS DE OPTIMIZACIÓN

Memoria presentada por

Juan Manuel Górriz Sáez

Para optar al grado de

DOCTOR POR LA UNIVERSIDAD DE GRANADA

Fdo. Juan Manuel Górriz Sáez

D. Carlos García Puntonet, Profesor Titular de Universidad del Departamento de Arquitectura y Tecnología de Computadores y **D. Javier Ramírez Pérez de Inestrosa**, Profesor Titular de Universidad del Departamento de Teoría de la Señal, Telemática y Comunicaciones.

CERTIFICAN

Que la memoria titulada: **"Nuevos Avances en Detección de Actividad de Voz mediante Estadísticos de Alto Orden y Estrategias de Optimización"** ha sido realizada por **D. Juan Manuel Górriz Sáez** bajo nuestra dirección en el Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada para optar al grado de Doctor por la Universidad de Granada.

Granada, a 5 de Mayo de 2006

Fdo. Carlos G. Puntonet
Director de la Tesis

Fdo. Javier Ramiréz Pérez de Inestrosa
Director de la Tesis

A mi creadora

AGRADECIMIENTOS

De bien nacido es ser agradecido...

Me vuelven a faltar palabras para agradecer el apoyo brindado por mi amigo Carlos a todo este trabajo, en los buenos y malos momentos.

Quisiera también dar las gracias a Javier, por hacerme redescubrir el concepto de amistad y además por su trabajo, codo con codo, en esta "nuestra tesis".

A mis compañeros del Departamento de Teoría de la Señal, Telemática y Comunicaciones por su buen recibimiento a mi llegada, haciéndome sentir uno más. En especial a Ángel, por su compañía, a José Carlos, por sus ideas...

A mis "MDPs" (Paco, Juan, Manolo y un largo etc...) que siempre han tenido que soportar mis conversaciones sobre temas científicos a altas horas de la madrugada. A mis amigos Paco Montes de Oca, José Castro, Paco Hernández y todos aquellos que se pierden en mi memoria.

A la suerte o a la ausencia de ella, que ha sido fundamental (en un 3 %) para que me ponga a "currelar" en esta innombrable.

A mi familia: mi padre (D. Enrique), mi madre (mi creadora) y mis hermanos/as (Espe, Manme, Fanny, Quique, Toto, María, Eva y Javi).

A la Srta. Pilar Benavente, ha sido un revulsivo para hacer esta Tesis, sin ella, esto no habría llegado a buen puerto [. . .](una vez más, bis).

Juan Manuel Górriz Sáez

Persona "humana y democrática" de talante amistoso

Granada, 5 de Mayo de 2006

RESUMEN

En este trabajo se presentan nuevos avances en el campo de la detección de actividad de voz (VAD, del inglés "Voice Activity Detection") para su aplicación a Reconocimiento robusto del Habla en entornos ruidosos. Los nuevos detectores de actividad de voz (VADs) se basan en distintas metodologías: i) Tests estadísticos basados en promedios biespectrales sobre la rejilla en el dominio bi-frecuencia; ii) Tests estadísticos basados en el cociente de probabilidad (LRT) de magnitudes biespectrales integradas; iii) Análisis cluster para modelado del espacio de ruido y formulación de una regla de decisión basada en divergencia cluster; y iv) Máquinas de vectores soporte (SVMs) aplicadas a las clases de SNRs en subbandas de energía. El rendimiento de los VADs propuestos es superior en tasa de acierto de detección para una falsa alarma dada, cuando los comparamos con los VADs estándar, como los de ITU-T G.729, ETSI GSM AMR y ETSI AFE, y con los recientemente publicados, usando las bases de datos más representativas de ETSI como son AURORA2&3, y al formar parte de un sistema automático de reconocimiento (ASR), mejoran sensiblemente la tasa de reconocimiento de palabra en entornos ruidosos.

ABSTRACT

In this work we present new advances in Voice Activity Detection (VAD) and their application to robust Speech Recognition in noisy environments. The new set of VADs are based in several methodologies: i) Statistical Tests based on bispectrum averages over the grid in the bi-frequency domain; ii) Statistical Tests based on the likelihood ratio (LRT) on integrated bispectrum magnitude; iii) Cluster Analysis for noise subspace modeling and the formulation of a cluster divergence based decision rule; and iv) Support Vector Machines (SVMs) applied to classify energy subband SNRS. The proposed VADs present higher performance in detection hit rate, for a given false alarm, than standard VADs such as ITU-T G.729, ETSI GSM AMR and ETSI AFE, and a representative set of recently reported VAD algorithms. An exhaustive analysis is conducted on the ETSI Aurora2 and Aurora3 databases in order to assess the performance for distributed speech recognition (DSR) in noisy environments.

PRÓLOGO

Dentro de las tecnologías del habla, la investigación en el campo de la interacción oral hombre-máquina es una de las áreas de investigación más populares. En este campo se incluyen los sistemas de reconocimiento automático del habla (ASR: Automatic Speech Recognition) que se van introduciendo paulatinamente en productos comerciales. Uno de los problemas más importantes que tienen este tipo de aplicaciones es su sensibilidad al ruido presente en el entorno de operación. Los sistemas ASR se caracterizan mediante modelos estadísticos de probabilidad entrenados previamente con objeto de representar adecuadamente las propiedades de la señal de voz. En estas condiciones, los sistemas de reconocimiento proporcionan buenos resultados en condiciones de laboratorio; sin embargo, en un entorno más realista, la señal de voz se encontrará afectada por ruido y los modelos entrenados no corresponderán al de las condiciones de operación del sistema. De esta manera, la precisión de los sistemas de reconocimiento que operan en entornos ruidosos se degrada notablemente siendo ésta aún mayor cuando disminuye la relación señal-ruido (SNR: Signal-to-Noise Ratio). El reconocimiento robusto de voz trata con el mantenimiento de la tasa de reconocimiento incluso cuando la calidad de la señal de voz de entrada se ve degradada por el entorno. El estudio de nuevas técnicas de reconocimiento robusto ha sido un campo de investigación de gran interés en los últimos años. En este trabajo se estudia el problema de la detección de voz en ruido, un problema sin resolver que tiene

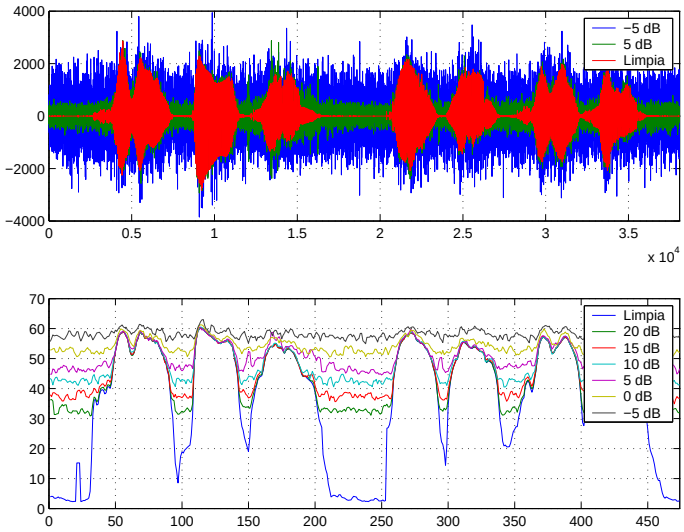


Fig. 0.1: El problema de la Detección de Actividad de Voz (VAD) desde el punto de vista de la energía.

una gran influencia en la aplicación de los algoritmos de supresión de ruido y en la efectividad de los sistemas de reconocimiento. Las prestaciones del algoritmo de detección se evalúan de forma directa por medio de la determinación de las tasas de clasificación de segmentos cortos de voz y silencio y, de forma indirecta, por medio de la tasa de reconocimiento. Los experimentos de reconocimiento que se consideran son el reconocimiento de dígitos conectados mediante modelos ocultos de Markov (HMM: Hidden Markov Models) utilizando una representación basada en los coeficientes cepstrales en escala Mel (MFCC: Mel-frequency cepstral coefficient). Los algoritmos evaluados consiguen elevar la precisión del reconocedor de forma significativa.

En particular en este trabajo se van a presentar nuevas propuestas para VAD usando distintos conceptos como magnitudes biespectrales, clustering, tests de cociente de probabilidades, etc.

La estadística de alto orden se ha usado para diversas aplicaciones dado su enorme potencialidad para la caracterización de variables aleatorias [Brillinger, 1975]. Como herramienta fundamental, en el análisis estadístico de procesos aleatorios, cabe destacar el uso de momentos y cumulantes, que son promedios estadísticos de funciones de variables aleatorias. Estas variables pueden caracterizarse así como sus propiedades más interesantes mediante sus momentos y cumulantes; por ejemplo si una variable aleatoria sigue una estadística gaussiana, su función de distribución queda totalmente caracterizada por sus momentos y cumulantes de orden uno y dos [Górriz, 2003]. En general deberemos recurrir a estadísticos de orden superior a dos para el propósito de caracterización de variables aleatorias.

Por otra parte, la resolución del problema de VAD puede afrontarse mediante clasificadores óptimos bayesianos para la distinción entre los dos estados presentes en el canal, silencio y voz (digamos H_0 y H_1). Éstos están basados en el cociente de probabilidades del patrón de entrada dada la hipótesis alternativa y nula ($H_0 \& H_1$) y obtienen buenos resultados para el conjunto empírico empleado para la clasificación aunque carecen del poder de generalización que otorgan otras estrategias basadas en Teoría Estadística de Vapnik-Chervonenkis [Vapnik, 1995a] para detección de Actividad de Voz [Ramírez et al., 2006b].

Finalmente la técnica de "clustering" (agrupamiento en castellano) ha sido aplicada a una gran cantidad de campos como por ejemplo, máquinas de aprendizaje, biología, psiquiatría, geología, geografía,etc. de manera satisfactoria [Jain and Dubes, 1988; Fisher, 1987]. El objetivo de esta segmentacion

de datos es agrupar un conjunto de objetos en subconjuntos o "clusters" de tal forma que dentro de cada subconjunto los objetos estén más relacionados (en cierto sentido) que los objetos asignados a otros subconjuntos. El análisis mediante clustering también se usa para formar estadísticos descriptivos que aseguren que un conjunto de objetivos, así agrupados, tienen propiedades estadísticas, en esencia, distintas. Esa será la filosofía de los detectores basados en clusters [Górriz et al., 2006b].

El trabajo se ha organizado de la siguiente manera:

- En los Capítulos 1 y 2 se introduce el entorno "Tecnologías del Habla" dentro del cual está enmarcado el presente trabajo. Además se repasa el estado del arte en detección de actividad de voz, presentando los VADs más significativos desarrollados para las distintas aplicaciones que sirven de motivación para el desarrollo de este trabajo.

- En el Capítulo 3 se presentan los primeros detectores de actividad de voz propuestos usando estadísticos de alto orden (HOS, del inglés "higher order statistics"), por ejemplo la magnitud biespectral. Mas concretamente usamos tests de probabilidades generalizados basados en promedios biespectrales en la rejilla fina. Los resultados para distintas configuraciones y métodos han sido publicados en [Górriz et al., 2006c; Górriz et al., 2005e; Górriz et al., 2005b; Górriz et al., 2005d; Górriz et al., 2005c; Górriz et al., 2005a] (Véase Apéndice). De estos trabajos se obtuvieron resultados prometedores para seguir investigando en este campo y adaptar el problema de detección de actividad de voz a este tipo de test genéricos desarrollados por [Tugnait, 1994; Subba-Rao, 1982], que usaban costosos promedios biespectrales. El resultado de esta adaptación (promedio en una única dimensión espectral y uso de tests óptimos de cociente de probabilidades (LRT, del inglés

"likelihood ratio tests") sobre ventanas de múltiple observación como clasificadores de Bayes) se presenta en el siguiente punto.

- En el siguiente Capítulo se presentan los principales logros alcanzados en detección de actividad de voz como un compromiso entre eficiencia computacional y rendimiento en la tasa de acierto. Son los VADs basados en tests de cociente de probabilidades (LRT) sobre promedios de la magnitud biespectral (biespectro integrado) que adolecen de una alta tasa de acierto de los segmentos de voz sin necesidad de incluir filtros de reducción de ruido que preceden al VAD. Estos trabajos han sido publicados recientemente en [Górriz et al., 2005e; Ramírez et al., 2006a; Górriz et al., 2006e; Górriz et al., 2006a] (véase Apéndice), siendo óptimos para un conjunto empírico de realizaciones muestrales, o lo que es lo mismo aquellos VADs que consideran exclusivamente el riesgo empírico [Vapnik, 1995a].

- La técnica "clustering" junto con otras estrategias de optimización pueden ser usadas para formular reglas de decisión robustas para VAD. En el Capítulo 5 presentamos la principal aportación de la aplicación del conocido "soft computing" al campo de procesado de señales para detección. La principal ventaja de estos detectores es la eficiencia computacional en su implementación, que los hace idóneos para aplicaciones en tiempo real, además de la mejora sustancial que producen cuando los comparamos con otros VADs recientemente publicados y los estándares para transmisión discontinua y reconocimiento de voz. Algunos trabajos relacionados con el Capítulo 5 han sido recientemente publicados en [Górriz et al., 2006b; Górriz et al., 2006a; Górriz et al., 2006f; Culebras et al., 2006; Górriz et al., 2006d].

- En el Capítulo 6 se aplica la teoría basada en SVM al problema de VAD, obteniendo un conjunto de detectores óptimos como clasificadores que minimizan el riesgo actual. La principal ventaja de estos VADs es su capacidad de generalización superior (teóricamente hablando) a todos los VADs propuestos hasta la fecha. Estos detectores presentan un gran sustento teórico [Vapnik, 1995a], siendo su mayor inconveniente la aproximación de aprendizaje "por lotes" que representan. Sin duda una estrategia a seguir en el futuro será su reformulación como algoritmos adaptativos de aprendizaje con el mismo poder de generalización en entornos estacionarios [Górriz, 2003]. Los trabajos publicados en este contexto han sido varios: [Ramírez et al., 2006b; Yélamos et al., 2006].

- Finalmente en los últimos capítulos se discuten conjuntamente los resultados experimentales mostrados en este trabajo y se presentan las conclusiones y líneas futuras del mismo.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ALGORITMOS

1. INTRODUCCIÓN

Una tarea común a muchas aplicaciones de las tecnologías del habla es la detección de los periodos de voz y silencio en una señal dada. La codificación de voz, la supresión de ruido acústico o el reconocimiento automático del habla son ejemplos de estas aplicaciones siendo el propósito del detector bien distinto en cada una de ellas. Por ejemplo, el detector de actividad de voz (VAD) empleado en codificación de voz permite reducir la velocidad de transmisión y realizar un mejor aprovechamiento del ancho de banda disponible. En otra clase de aplicaciones en las que el entorno acústico degrada las prestaciones del sistema, el objetivo del detector será la caracterización del ruido para así compensar su efecto. En este contexto, los sistemas robustos de reconocimiento emplean técnicas de detección de actividad oral para reducir la tasa de error del reconocedor que se degrada rápidamente por la influencia del ruido.

La efectividad del VAD es de gran importancia para la mayoría de las aplicaciones dentro del campo de las tecnologías del habla. Su misión no es tan trivial como pudiera parecer en un principio y la dificultad aumenta considerablemente cuando la señal se encuentra mezclada con ruido de carácter no estacionario. En este trabajo de investigación se describen nuevas técnicas robustas para detección de actividad de voz que tienen como principal objetivo mejorar las prestaciones de los sistemas de reconocimiento de voz que operan en entornos ruidosos. La primera propuesta se basa en la de-

terminación de la divergencia entre la voz y el silencio utilizando un test estadístico basado en el biespectro integrado promediado por bloques. Este algoritmo, además se caracteriza respecto a otros publicados recientemente en la forma en que se formula la regla de decisión. En vez de fundamentar la decisión en un único segmento de señal, lo que se prefiere es utilizar el biespectro integrado con una ventana de datos de entrada más amplia. Esta aproximación reduce la tasa de error del detector permitiendo así una mejor discriminación de los segmentos de voz y de silencio. La segunda técnica se basa en una generalización de un test estadístico óptimo de múltiples observaciones biespectrales que demuestra una robustez creciente con el número de observaciones siendo su complejidad computacional reducida.

La evaluación de los VADs propuestos se basa en la utilización de las bases de datos empleadas por el grupo de trabajo AURORA de ETSI en la fase de desarrollo del estándar ETSI ES 202 050 para reconocimiento de voz distribuido (DSR: "*Distributed Speech Recognition*") [ETSI, 2002]. En primer lugar, se evalúa el rendimiento del VAD por medio de las tasas de aciertos de los segmentos de voz y de silencio. Con posterioridad se analiza la influencia del VAD en un sistema robusto de reconocimiento de voz que emplea el VAD para filtrado de ruido y "*frame-dropping*", una técnica utilizada habitualmente en reconocimiento robusto que excluye los segmentos de silencio del proceso de reconocimiento y que reduce significativamente el número de errores de inserción. Este estudio experimental se completa con una comparativa con los algoritmos de detección de actividad oral incluidos en los estándares de codificación ITU G.729 y ETSI AMR [ETSI, 1999] para transmisión discontinua, así como el estándar ETSI AFE ("*Advanced front-end*")para DSR [ETSI, 2002], y cuatro de los VADs más representativos de la bibliografía.

Este trabajo ha sido desarrollado en el marco de los proyectos de investigación:

- "SR3-VoIP: Sistemas robustos de reconocimiento y reconstrucción de voz sobre IP" (Ref.: TEC2004-03829/TCM).
- "HIWIRE: Human Inputs that Work in Real Environments" (IST Contract No. 507943, Sexto Programa Marco).
- "SESIBONN" Separación de Señales y Sistemas Adaptativos en Biomedicina, Comunicaciones, Imágenes y Predicción (Ref TEC2004-06096).

Se trata de un trabajo de investigación realizado por el doctorando y basado en los artículos que han sido recientemente aceptados para su publicación en las revistas más relevantes dentro del campo: *IEEE Signal Processing Letters*, *IEE Electronic Letters*, *Journal of Acoustical Society of America* ó *Lecture Notes in Computer Science* como parte de números especiales en procesado de señal, tal y como se refleja en la documentación adjunta en los apéndices.

1.1. El problema de la Detección de Actividad de Voz (VAD)

Un problema importante en numerosas aplicaciones de procesado de voz es la determinación de presencia de periodos de voz en una señal dada. Esta tarea se puede identificar como un problema de *test* estadístico de hipótesis y su objetivo es determinar a qué categoría o clase pertenece la señal. La decisión se suele basar en un vector de observación, frecuentemente conocido como vector de características, que sirve como entrada a un bloque o regla de decisión que asigna a cada vector una de las dos clases. La tarea de clasificación no es tan trivial como pudiera parecer puesto que el nivel creciente de ruido presente en el entorno acústico de operación degrada la efectividad del clasificador y origina numerosos errores de detección.

Existen numerosas situaciones en las aplicaciones emergentes de las tecnologías del habla (por ejemplo, comunicaciones móviles, reconocimiento robusto del habla o dispositivos de ayuda a la audición) que requieren un esquema de reducción de ruido en combinación con un detector de actividad (VAD) preciso [Bouquin-Jeannes and Faucon, 1994; Bouquin-Jeannes and Faucon, 1995b]. En las últimas décadas numerosos investigadores han estudiado diferentes estrategias para detectar la presencia de voz en ruido y la influencia de la precisión del VAD sobre sistemas de procesado de voz [Freeman et al., 1989; ITU, 1996; Sohn et al., 1999; ETSI, 1999; Marzinzik and Kollmeier, 2002; Sangwan et al., 2002; Karray and Martin, 2003]. La mayoría de los autores que tratan el tema de la reducción de ruido utilizan el término "detección de pausas" en el contexto del problema de estimación de la estadística del ruido. El algoritmo de detección de los segmentos de silencio es una pieza sensible y clave para la mayoría de los algoritmos existentes de reducción de ruido mediante un único micrófono. Existen algoritmos de supresión bien conocidos [Berouti et al., 1979; Boll, 1979; Ephraim and Malah, 1984],

tales como filtrado de Wiener o sustracción espectral, que se utilizan frecuentemente en reconocimiento robusto del habla, y para los cuales el VAD resulta crítico para alcanzar altos niveles de efectividad. Estas técnicas estiman el espectro de ruido durante los periodos de silencio para compensar el efecto degenerativo que éste causa a la señal de voz. Del mismo modo, el VAD resulta aún más crítico en entornos de ruido no estacionario puesto que se hace necesario actualizar las propiedades constantemente variantes del ruido y un error de detección tiene una mayor influencia sobre el rendimiento del sistema. Con el objetivo de eliminar el problema de la importancia del VAD en los sistemas de reducción de ruido, Martin [Martin, 1993] propuso un algoritmo que actualizaba el ruido de manera continua. Estas técnicas son más rápidas en cuanto se refiere a la estimación o seguimiento de las propiedades del ruido; sin embargo, capturan la energía de la señal con lo que se degrada la calidad de la señal de voz compensada. Por tanto, resulta claramente mejor utilizar un VAD para la mayoría de los sistemas de supresión de ruido y aplicaciones.

1.2. Estado actual

Los detectores de actividad oral se utilizan en numerosas áreas y aplicaciones de procesado de voz. Recientemente se han propuesto varios algoritmos de detección para diferentes aplicaciones incluyendo servicios de comunicación móvil [Freeman et al., 1989], trasmisión de voz en tiempo real a través de Internet [Sangwan et al., 2002] y reducción de ruido en dispositivos de ayuda a la audición [Itoh and Mizushima, 1997]. El interés de los investigadores se ha centrado en el desarrollo de algoritmos robustos con especial atención al estudio de características robustas de la señal de voz y la derivación de reglas de decisión precisas. Sohn [Sohn and Su, 1998] presentó

un algoritmo basado en la adaptación del espectro de ruido con decisión suave sobre la presencia de voz en una señal ruidosa. La regla de decisión consistía en una razón de probabilidades en el que se asume un modelo para la señal y un conocimiento *a priori* del ruido. Una versión mejorada del VAD original [Sohn et al., 1999] se consiguió incorporando a éste, un esquema de suavizado de la decisión ("*hang-over*") que consideraba las observaciones previas de un proceso de Markov de primer orden para modelar los segmentos de voz. El algoritmo mejoraba o al menos era comparable al estándar de la UIT G.729 [ITU, 1996] en términos de las probabilidades de detección correcta de los segmentos de voz y de las falsas alarmas. A partir de entonces, otros investigadores han introducido mejoras sobre el algoritmo de Sohn siendo éste una de las referencias más utilizadas en el campo para evaluar nuevos algoritmos. Cho [Cho et al., 2001a; Cho et al., 2001c; Cho et al., 2001b] presentó varias mejoras sobre la propuesta de Sohn [Sohn and Su, 1998; Sohn et al., 1999]. En [Cho et al., 2001a; Cho et al., 2001b] se propuso un test suavizado de razón de probabilidades para reducir los errores de detección demostrándose superior a G.729 y comparable al estándar de ETSI AMR (opción 2) [ETSI, 1999]. Así mismo, Cho propuso un esquema de adaptación de ruido con decisión mixta que obtenía mejores resultados que la adaptación con decisión suave propuesta por Sohn [Sohn and Su, 1998].

Recientemente, ETSI ha desarrollado un nuevo estándar de reducción de ruido y de extracción de características robustas para sistemas de reconocimiento de voz distribuidos. El estándar así llamado AFE ("*Advanced Front-End*") [ETSI, 2002] incorpora dos VADs. El primero de ellos se basa en la energía y tiene como objetivo estimar el espectro del ruido durante los periodos de silencio para reducción de ruido mediante filtrado de Wiener. El segundo de ellos se basa en una medida de la SNR en frecuencia y se utiliza

para eliminar de la entrada del reconocedor los segmentos de señal que son detectados como silencio. Esta técnica, conocida mediante el término "*frame-dropping*", se utiliza frecuentemente en reconocimiento de voz con el fin de reducir el número de errores de inserción y mejorar el rendimiento de estos sistemas en condiciones adversas de ruido.

Quizá la aplicación que mayor interés suscita se presenta en el campo de la codificación para telefonía móvil digital. El VAD permite realizar la compresión del silencio en los sistemas modernos de telecomunicaciones móviles reduciendo la velocidad de transmisión mediante el modo de transmisión discontinua (DTX). En muchos sistemas prácticos, tales como el sistema global de comunicaciones móviles GSM, se emplea la detección del silencio y la transmisión de ruido de comodidad a la audición para mejorar la eficiencia en la codificación de la señal y, por tanto, la utilización del canal de comunicación. En 1996, la Unión Internacional de Telecomunicaciones (UIT) adoptó un algoritmo de codificación de voz con calidad telefónica conocido como G.729 que operaba en combinación con un VAD (anexo G.729B) [ITU, 1996] en el modo de transmisión discontinua. La recomendación G.729B utiliza un vector de características formado por el espectro de predicción lineal (LP), la energía en la banda completa, la energía en la banda de 0 a 1 kHz y la tasa de cruces por cero (ZCR: zero-crossing rate). El estándar se desarrolló con la colaboración de investigadores de France Telecom, la Universidad de Sherbrook, NTT, y AT&T Bell Labs. La efectividad del VAD se evaluó mediante pruebas subjetivas de calidad de la señal de voz y la reducción de la velocidad de transmisión [Benyassine et al., 1997]. Igualmente, se utilizaron otros tests objetivos evaluando la correcta identificación de periodos sordos, sonoros, silencios y transiciones.

Otro estándar para transmisión discontinua es el codificador AMR de

ETSI [ETSI, 1999] desarrollado para el sistema GSM. El estándar tiene dos variantes para el VAD. En la opción 1, la señal se descompone mediante un banco de filtros calculándose el nivel de señal en cada banda. La decisión se formula en términos de una medida de la SNR junto con: información sobre la periodicidad de la señal ("*pitch*"), la salida de un detector de tonos y un módulo de detección de correlación en la señal. La opción 2 es una versión mejorada del VAD original y utiliza parámetros del codificador para formular la decisión siendo más robusto que G.729 y AMR1. Estos tres detectores se utilizan frecuentemente en la bibliografía como referencia para comparar nuevos algoritmos de detección. Marzinzik [Marzinzik and Kollmeier, 2002] propuso un VAD para estimación del espectro de ruido basado en el seguimiento de la dinámica de la envolvente espectral. El algoritmo se comparaba con el estándar G.729 por medio de las curvas características de operación (ROC) demostrando una reducción de las falsas alarmas del silencio junto con una mejora de la eficacia en la detección del silencio para un conjunto significativo de ruidos y condiciones. Beritelli [Beritelli et al., 1998] propuso un VAD basado en lógica difusa con un bloque de identificación de patrones que consistía en seis reglas difusas. La evaluación de este VAD se realizó utilizando medidas objetivas, psico-acústicas y subjetivas siendo los estándares G.729 y AMR utilizados como referencia [Beritelli et al., 2002]. Nemer [Nemer et al., 2001] presentó un algoritmo robusto basado en la estadística de alto orden en el dominio LPC. Sus prestaciones se compararon con el estándar G.729 en diferentes condiciones de ruido cuantificándose en términos de las probabilidades de clasificación.

1.3. Motivaciones

Dentro de las tecnologías del habla, la investigación en el campo de la interacción oral hombre-máquina es una de las áreas de investigación más populares. En este campo se incluyen los sistemas de reconocimiento automático del habla (ASR: Automatic Speech Recognition) que se van introduciendo paulatinamente en nuevos productos comerciales. Uno de los problemas más importantes que tienen este tipo de aplicaciones es su sensibilidad al ruido presente en el entorno de operación. Los sistemas ASR se caracterizan mediante modelos estadísticos de probabilidad entrenados previamente con objeto de representar adecuadamente las propiedades de la señal de voz. De este modo, los sistemas de reconocimiento proporcionan buenos resultados en condiciones de laboratorio; sin embargo, en un entorno más realista, la señal de voz se encontrará afectada por ruido y los modelos entrenados no corresponderán a las condiciones de operación del sistema. De esta manera, la precisión de los sistemas de reconocimiento que operan en entornos ruidosos se degrada notablemente siendo la degradación más importante cuando disminuye la SNR.

El objetivo del reconocimiento robusto es el mantenimiento de la tasa de reconocimiento incluso cuando la calidad de la señal de voz de entrada se vea seriamente afectada por el entorno acústico. El estudio de nuevas técnicas ha sido un campo de investigación de gran interés en los últimos años. En este trabajo se estudia el problema de la detección de voz en ruido, un problema sin resolver que tiene una gran influencia en la aplicación de los algoritmos de supresión de ruido y en la efectividad de los sistemas de reconocimiento.

La selección de un vector de características adecuado y una regla de decisión robusta para detección de actividad oral en una señal ruidosa afecta profundamente al rendimiento del VAD. La mayoría de los algoritmos son

efectivos en numerosas aplicaciones pero con frecuencia cometen errores de detección principalmente por causa de la pérdida de poder discriminativo de la regla de decisión en condiciones extremas de SNR [ITU, 1996; ETSI, 1999]. Por ejemplo, un detector basado en el nivel de energía puede operar eficientemente en condiciones de alta SNR pero fallará significativamente cuando la SNR disminuya. Se han propuesto numerosos algoritmos para paliar estos inconvenientes por medio de la definición de reglas de decisión más robustas. Las diferentes aproximaciones utilizadas incluyen aquellas basadas en umbrales de energías [Woo et al., 2000], detección de la frecuencia fundamental ("*pitch*") [Chengalvarayan, 1999], análisis espectral [Marzinzik and Kollmeier, 2002], tasas de cruces por cero [ITU, 1996], medidas de periodicidad [Tucker, 1992], o combinaciones de diferentes características de la señal [ITU, 1996; ETSI, 1999; Tanyer and Özer, 2000].

En este trabajo se estudian nuevas alternativas para mejorar la robustez de la detección en entornos adversos. La idea se basa en la utilización de información de largo periodo de la señal de voz y la magnitud biespectral como elemento discriminador voz/silencio. Concretamente, se desarrolla un método de detección de actividad de voz que emplea un test óptimo basado en cociente de probabilidades (LRT) de magnitud biespectral promediada en un bloque de datos y la generalización de este test a múltiples observaciones (MO-LRT) con una implementación eficiente.

Por otro lado, la detección incorrecta de segmentos de voz y silencio es una importante fuente de error en sistemas de reconocimiento automático del habla. Existen dos claras razones para ello:

i) La mayoría de los algoritmos de realce emplean un VAD para estimar la estadística del ruido. Por tanto, la efectividad de los algoritmos de compensación de ruido depende de la precisión del VAD.

ii) La técnica conocida como "*frame-dropping*" se utiliza frecuentemente en reconocimiento de voz para reducir el número de errores de inserción. Puesto que esta técnica se basa en la salida del VAD, los segmentos de voz incorrectamente etiquetados como silencio introducen errores irrecuperables de borrado; del mismo modo, los segmentos de silencio incorrectamente etiquetados como voz aumentan el número de errores de inserción.

En este trabajo de investigación, las prestaciones del algoritmo de detección se evaluarán de forma directa por medio de la determinación de las tasas de clasificación de segmentos cortos de voz y silencio y, de forma indirecta, por medio de la tasa de reconocimiento. Los experimentos de reconocimiento que se consideran son el reconocimiento de dígitos conectados mediante modelos ocultos de Markov (HMM: Hidden Markov Models) utilizando una representación basada en los coeficientes cepstrales en escala Mel (MFCC: Mel-frequency cepstral coefficient). Los algoritmos evaluados consiguen elevar la precisión del reconocedor de forma significativa.

2. FUNDAMENTOS DE LA DETECCIÓN DE ACTIVIDAD DE VOZ

La mayoría de los VADs pueden esquematizarse mediante el diagrama de bloques mostrado en la figura 2.1. En general, los VADs procesan la señal de entrada en segmentos (frames) cortos de tiempo de entre 20−40 ms y a continuación se extrae un conjunto de características de los frames pre-procesados. La decisión del VAD se realiza en base a la información suministrada por estas características y finalmente, esta decisión se suaviza en el tiempo, para caracterizar la naturaleza estacionaria en cortos periodos de tiempo del silencio y de la voz. La decisión final del VAD puede ser suave con un cierto valor de significación ó una decisión binaria abrupta. El tipo de aplicación determina que tipo de decisiones es preferible. El problema de diseño de cualquier VAD se caracteriza por la selección de las características, cómo se estima la estadística del ruido y la selección del método de clasificación. En el presente capítulo discutimos los métodos clásicos para selección de características y de la regla decisión

---

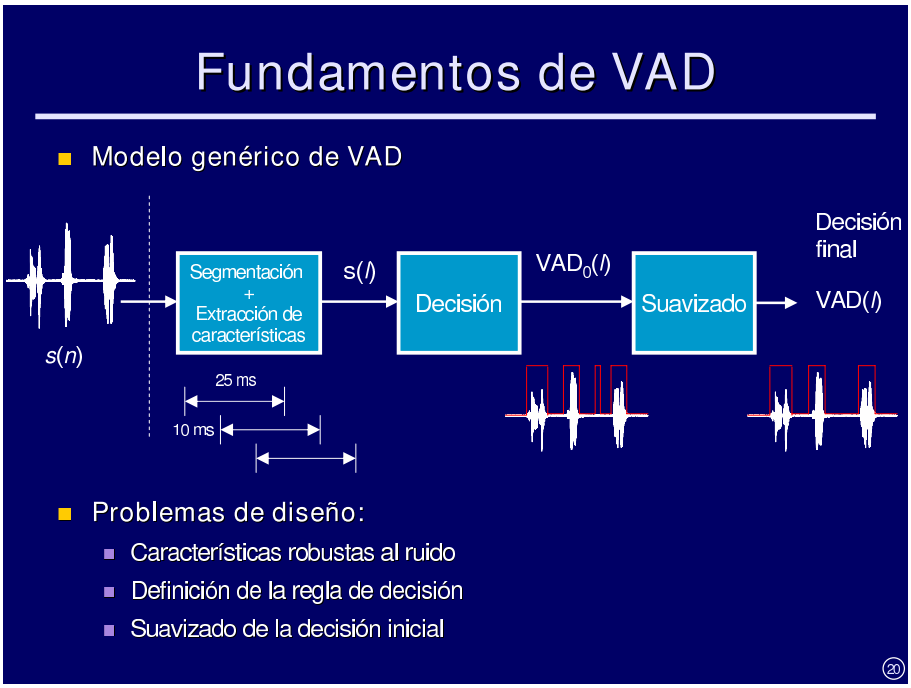


Fig. 2.1: Diagrama de bloques de un VAD

2.1. Hipótesis de partida

Basándonos en las propiedades y en los modelos de producción de la señal voz, se pueden realizar distintas asunciones para discriminar los periodos de voz y silencio. Dado que no se conoce ninguna información a priori sobre el ruido, se asume que presenta ciertas propiedades. Las primera de ellas y más usual, es la asunción de que el ruido es aditivo, lo que significa que la energía en los periodos de voz consistirá en la energía del ruido ambiente más la de la señal limpia de voz, por lo tanto la energía del periodo de voz será mayor. Además, se asume que las propiedades estadísticas de la voz varían, mientras que las del ruido permanecen constantes en un periodo de tiempo más largo. Algunas de las hipótesis más usuales realizadas en la mayoría de los algoritmos detección de actividad de voz se listan a continuación:

- El ruido de ambiente es aditivo a la señal de voz.
- El segmento de la señal de voz tiene un valor de energía superior que el segmento de ruido de ambiente.
- La voz es estacionaria en periodos cortos de tiempo, por ejemplo $T < 40$ ms.
- La voz es no estacionaria sobre periodos más largos, por ejemplo $T >$ 0,5s.
- El ruido del ambiente es estacionario para periodos mucho más largos, por ejemplo $T > 2$s.
- La voz tiene más componentes periódicas que el ruido.
- El espectro de la voz está más "organizado" que el espectro de ruido.

Usando estas hipótesis de partida, se pueden diseñar fácilmente algoritmos de detección de actividad de voz (VAD) para discriminar periodos de voz y silencio. Dadas las condiciones restrictivas impuestas al ruido ambiente, existirán siempre situaciones en las que no se cumplan todas las hipótesis realizadas, por lo que el VAD debería utilizar un subconjunto de ellas que lo hagan lo suficientemente robusto. De hecho, en la mayoría de los VADs, los estadísticos del ruido suelen ser actualizados para proporcionar una regla de clasificación más robusta.

2.2. Extracción de Características

La extracción de características intenta presentar el contenido de la señal de voz de manera compacta, de tal forma que la información propia de la señal se preserva. Antes de la extracción de características, la señal puede

limitarse a un rango de frecuencias determinado, pudiendo filtrar las componentes bajas de frecuencia. Mas aún, el rango dinámico de la señal puede comprimirse con un filtro de pre-énfasis, amplificando las frecuencias más altas. El filtro de pre-énfasis aplana el espectro, lo cual es deseable dado que, de otra forma, los formantes de baja frecuencia que contienen más energía, serían modelados con preferencia sobre los de alta frecuencia. La señal de entrada $s(n)$ puede así ser filtrada con este filtro para obtener una señal pre-enfatizada $\hat{s}(n)$ como sigue:

$$\hat{s}(n) = s(n) - \alpha s(n-1) \tag{2.1}$$

donde α tiene un valor próximo a la unidad (por ejemplo 0.95) [Furui, 2001]. Usualmente las características se extraen mediante segmentos solapados, estando por tanto correlacionados y suavizando el cambio espectral de frame a frame. Un frame es un intervalo temporal corto de señal obtenido mediante la multiplicación de la señal por una función ventana, $w(n)$. La función de ventana determina la porción de la señal de voz que es procesada haciendo nula la señal fuera de la misma. Las ventanas más comunes son la rectangular $w_R(n)$ y la de Hamming $w_H(n)$ definidas como:

$$w_R(n) = \left\{ \begin{array}{ll} 1, & \text{cuando} \quad n = [0, \ldots, N-1] \\ 0 & \text{otro caso} \end{array} \right\} \tag{2.2}$$

$$w_H(n) = \left\{ \begin{array}{ll} 0{,}54 - 0{,}46 cos\left(\frac{2\pi n}{N-1}\right), & \text{cuando} \quad n = [0, \ldots, N-1] \\ 0 & \text{otro caso} \end{array} \right\} \tag{2.3}$$

donde N es la longitud del segmento. La longitud de la ventana N se elige atendiendo a consideraciones prácticas por lo que varía según la aplicación. Por ejemplo, para el cálculo de energías en un tiempo breve, una N pequeña provocaría que la energía variara muy rápido de frame a frame. Al contrario

el aumento del tamaño de ventana N provocaría un promedio sobre un intervalo temporal muy grande, por lo tanto la energía que modela un intervalo temporal pequeño no reflejaría adecuadamente las propiedades dinámicas de la señal. En la práctica, se suele elegir una N en torno a 120-240 muestras a 8000 Hz de frecuencia de muestreo (por ejemplo el segmento dura entre 15-30 ms). La longitud del frame y el solapamiento entre frames definen la tasa de frames. La tasa de frames describe la frecuencia a la que los frames son extraídos de la señal, por lo que define el intervalo temporal entre decisiones del VAD.

Se han propuesto varias características y combinaciones de ellas para distintos algoritmos VAD [Tanyer and Özer, 2000]. Una de la más usuales, es la características que representa las variaciones de niveles de energías [Woo et al., 2000] ó la diferencia entre los espectros de ruido y voz [Marzinzik and Kollmeier, 2002]. Sin embargo, una característica podría no ser suficiente para suministrar información relevante de la señal, habiéndose propuesto una combinación de ellas para formar los conocidos como vectores de características. La elección correcta de características puede aportar grandes ventajas [Liu and Motoda, 1998], como el requerimiento de un menor número de datos para entrenamiento, aumento de la precisión en la clasificación o reducción de la dimensión del espacio de características. Las características más conocidas hasta la fecha se presentan en las siguientes secciones.

2.2.1. Energía

El uso del promedio de energía en periodos cortos en la forma más sencilla de clasificar la señal en segmentos de voz y silencio (ruido), dado que la señal presenta, a priori, una mayor energía que el ruido cuando la señal limpia está

presente. El promedio temporal de energía puede escribirse como:

$$E = \frac{1}{N} \sum_{n=0}^{N-1} x^2(n) \tag{2.4}$$

donde $x(n)$ denota la señal discreta temporal considerada. Para evitar los problemas que acarrea esta característica en ambientes de baja SNR se suele trabajar en subbandas de energía, las cuales describen la distribución de la energía de la señal en el dominio de la frecuencia con una determinada resolución [Sohn et al., 1999].

Este tipo de detectores pueden proporcionar un buen rendimiento en clasificación cuando la energía de los periodos de voz es suficientemente mayor que la del ruido ambiente [Woo et al., 2000], sin embargo, cuando la energía del ruido ambiente es también elevada, la característica "energía de la señal" proporcionará un rendimiento muy bajo. En particular, las sonidos sordos de la voz , caracterizados por un nivel bajo de energía, serán incorrectamente clasificados. Sin embargo la mayoría de los VADs usan en esencia esta magnitud (transformaciones no lineales) para la tarea de la detección de actividad de voz.

2.2.2. Tasa de cruces por cero

La tasa de cruces por cero (ZCR, del inglés "Zero Crossing Rate") es una medida sencilla del contenido de frecuencia de la señal de entrada. El valor de la ZCR se obtiene calculando el número de veces que la secuencia cambia de signo, por lo que viene dada por:

$$ZCR = \frac{1}{N} \sum_{n=0}^{N-1} sign(x(n)) - sign(x(n-1)) \tag{2.5}$$

donde "sign" representa la función signo.

En ambientes con alta SNR, el valor de la ZCR aumenta considerablemente en los periodos de voz debido a sus formantes de alta frecuencia, y el valor de la ZCR es más bajo para los segmentos de ruido [Rabiner and Sambur, 1975]. En cambio en ambientes ruidosos, la ZCR del ruido se asume considerablemente mayor que la ZCR de la señal de voz. Además esta primera es muy sensible al tipo de ruido, encontrándose el valor de la ZCR de los sonidos sordos muy próximo al valor de la ZCR del ruido. A pesar de esto la ZCR se usa ampliamente en los algoritmos VAD, dado que suministra una estimación sencilla de la distribución de frecuencias de la señal.

2.2.3. Periodicidad

La periodicidad es una técnica que remarca las componentes sonoras de la señal de entrada, pudiéndose definir de varias formas. El objetivo esencial de los algoritmos basados en periodicidad es encontrar esta periodicidad, comparando la similitud de la señal original con una versión desplazada de la misma. Si la distancia de desplazamiento es igual a la frecuencia de la señal, las dos señales comparadas presentarán una gran similitud (correlación). Una de las técnicas más usadas es localizar el máximo de la función de auto-correlación. La función de auto-correlación en el instante k se define como:

$$r_x(k) = \sum_{n=0}^{N-k-1} x(n)x(n+k); \text{para} \quad k = 0, \ldots, N-1 \tag{2.6}$$

Como puede verse, para el instante $k = 0$ esta característica es equivalente a la energía definida en la ecuación 2.4. Desafortunadamente esta función puede presentar falsos picos, lo que complicaría la búsqueda de la característica periodicidad, debido a cambios rápidos en las frecuencias de los formantes [Lee et al., 2003]. Otra característica basada en periodicidad fue la propuesta por [Tucker, 1992], el estimador de periodicidad de míni-

mos cuadrados (LSPE, del inglés "Least Square Periodicity Estimator"). El objetivo del LSPE es encontrar el periodo de "pitch", que minimice el error cuadrático medio (MSE) entre la señal y la señal periódica reconstruida. La característica de periodicidad puede reducir significativamente las falsas alarmas debidas a impulsos y ruidos blancos, aumentando la precisión en detección cuando lo comparamos con la característica de auto-correlación [Tucker, 1992]. Sin embargo, también es sensible a cualquier otra señal periódica presente (por ejemplo un ruido periódico en el ambiente).

2.2.4. Entropía

La Entropía es una medida cuantitativa de la aleatoriedad de una fuente de información. La entropía de Shannon mide la longitud promedio de palabra binaria por símbolo bajo codificación óptima y para una fuente de información S. Dada una variable aleatoria discreta $X(\omega)$ generada por una fuente de información S, se define la incertidumbre o entropía de la misma como el valor esperado de su *Información* $\mathbf{I}(\omega)$:

$$\mathbf{H}(X) = \mathbf{E}\{\mathbf{I}(\omega)\} = -\sum_{\omega \in \Omega} P(\omega) log\{P(\omega)\} = -\sum_{x} p(x) log\{p(x)\} \qquad (2.7)$$

donde ω denota el evento que genera la variable aleatoria y $p(x)$ es la probabilidad de la realización x de la variable aleatoria $X(\omega)$. La aplicación de estos conceptos a la magnitud espectral normalizada de la señal de entrada, hace de la entropía un serio candidato para la formulación de VADs robustos para distintos niveles de ruido aunque muy sensibles a ruidos musicales [Huang and Yang, 2000].

2.2.5. Coeficientes de Predicción Lineal

La codificación de predicción lineal (LPC del inglés "Linear Prediction Coding") es uno de los métodos de análisis de la voz más utilizados. Se ha aplicado en muchas campos del procesado de voz digital, como por ejemplo el diseño de filtros, análisis espectral e identificación de sistemas [Markel and Gray, 1976]. En este método se asume que la señal de entrada se deja modelar como un proceso Auto-Regresivo como sigue:

$$x(t) = \sum_{k=1}^{p} a_p(k)x(t-k) + e(t) \tag{2.8}$$

donde $e(t)$ es una variable estadística decorrelada con media nula y varianza σ^2. El objetivo de LPC es encontrar este conjunto de coeficientes que mejor representen la señal de entrada en un "cierto sentido", este es el de minimizar el error cuadrático medio. Una vez encontrados el conjunto de parámetros $\{a_p(k)\}, k = 1, \dots p$ a partir del principio de ortogonalidad que satisface la solución óptima [Markel and Gray, 1976], el VAD se formula en términos de la medida de distancia espectral entre modelo y señal de entrada [Nemer et al., 2001].

2.2.6. Coeficientes Cepstrales

El cepstrum de una señal x se define como la transformada de Fourier inversa de la magnitud logarítmica del espectro. El cepstrum $\mathcal{C}$, puede ser calculado como:

$$\mathcal{C} = \mathcal{F}^{-1}\{\ln \mathcal{F}(x)\} \tag{2.9}$$

donde $\mathcal{F}$ denota el operador transformada de Fourier. Existen dos aproximaciones ampliamente usadas para el cálculo de los coeficientes cepstrales;

la primera es el cálculo directo de los coeficientes cepstrales de frecuencia mel (MFCC) basado en el espectro de potencia que se deriva del análisis de Fourier; la segunda aproximación (indirecta) está basada en el análisis LPC: el cepstrum LPC puede obtenerse como una DFT de los coeficientes LPC.

Actualmente, el análisis cepstral es probablemente la técnica de extracción de características más usada. Éste proporciona un buen modelo de las variaciones de la voz. Esta variaciones no son exhibidas por las señales de ruido por lo que se podría usar esta propiedad para la formulación de algoritmos VAD robustos. Un de los VADs propuestos basados en cepstrum es el presentado en [Haigh and Mason, 1993], el cual proporciona buenos resultados sin ninguna medida explícita de la señal o del nivel del ruido (como en la sección anterior se aplica un modelo a la voz).

2.3. Realización de la Decisión

La decisión del VAD (véase figura 2.1) se lleva a cabo en el módulo de decisión del algoritmo VAD. La decisión puede ser binaria o suave con un nivel de significación, dependiendo tal elección de la aplicación donde el VAD se esté empleando. Usualmente se prefiere una decisión binaria, pero por ejemplo, en los sistemas de reconocimiento de voz se suelen emplean ambos tipos. La decisión del VAD se ha realizado hasta la fecha respecto a diversos criterios que pasamos a resumir a continuación:

2.3.1. Regla de decisión mediante Modelos Estadísticos

Los modelos estadísticos proporcionan una metodología flexible de trabajo para los algoritmos VAD. En esencia éstos modelos utilizan estadísticos de decisión, que miden la distancia entre un modelo y el vector de características de observación.

Sea $\mathcal{O}$ un modelo estadístico y $P(\mathbf{x}^l|\mathcal{O})$ la probabilidad de observación del vector de características $\mathbf{x}^{(l)}$. Con la suposición de que los vectores de características $x^{(l)}, \quad l = \{0, \ldots, N-1\}$ sean estadísticamente independientes, la probabilidad de la secuencia de observaciones $\mathbf{X} = (\mathbf{x}^0, \ldots, \mathbf{x}^{N-1})$ puede calcularse como un productorio de probabilidades de observación dado por:

$$P(\mathbf{X}|\mathcal{O}) = \prod_{l=0}^{N-1} P(\mathbf{x}^{(l)}|\mathcal{O}) \tag{2.10}$$

La probabilidad de la secuencia se suele presentar como un logaritmo de probabilidad para preservar la precisión numérica:

$$\mathcal{L}(\mathbf{X}|\mathcal{O}) = \sum_{l=0}^{N-1} P(\mathbf{x}^{(l)}|\mathcal{O}) \tag{2.11}$$

Asumiendo que las observaciones de voz y de ruido son generadas por diferentes distribuciones, se puede calcular la probabilidad de que la observación fuera producida por la voz o el ruido. La decisión de clasificación puede hacerse mediante el principio de máximo a posteriori (MAP del inglés "Maximum a Posterior Principle") asignando la observación que maximiza la probabilidad a posteriori [Kay, 1993] al modelo. El principio MAP puede formularse como:

$$\hat{\mathcal{O}} = arg \max_{\mathcal{O}_i} P(\mathcal{O}_i|\mathbf{X}) \tag{2.12}$$

donde $P(\mathcal{O}_i|\mathbf{X})$ es la probabilidad a posteriori para la clase $\mathcal{O}_i$, es decir, la probabilidad de la clase $\mathcal{O}_i$ después de que la observación $\mathbf{X}$, se haya producido. Generalmente $P(\mathcal{O}_i|\mathbf{X})$ es desconocida, pero usando la fórmula de Bayes puede reescribirse como sigue:

$$P(\mathcal{O}_i|\mathbf{X}) = \frac{P(\mathcal{O}_i)P(\mathbf{X}|\mathcal{O}_i)}{P(\mathbf{X})} \tag{2.13}$$

donde $P(\mathcal{O}_i)$ es la probabilidad de la clase $\mathcal{O}_i$, y el numerador $P(\mathbf{X})$ es la función de densidad de probabilidad incondicional para la secuencia $\mathbf{X}$, que

no depende de la clase $\mathcal{O}_i$. Asumiendo que todas las clases son igualmente probables, esto es, las probabilidades $P(\mathcal{O}_i)$ son las mismas para todas clases, la maximización de la ecuación 2.12 puede escribirse:

$$\hat{\mathcal{O}} = arg \max_{\mathcal{O}_i} \mathcal{L}(\mathbf{X}|\mathcal{O}_i) \tag{2.14}$$

Usando este principio y un modelo para las probabilidades condicionadas de la secuencia de observación (por ejemplo un modelo gaussiano), se han desarrollado un conjunto de detectores de actividad de voz de una gran precisión en la clasificación de los segmentos de voz y de silencio [Sohn et al., 1999].

2.3.2. Regla de decisión mediante distancias

Tradicionalmente los algoritmos VAD asumen que los estadísticos del ruido ambiente son estacionarios sobre un periodo mayor que los de la voz. Esto hace posible la estimación de los parámetros para el ruido, conformando así un "modelo". Los parámetros del ruido pueden ser estimados y actualizados durante los periodos inactivos del canal con la ayuda de la decisión propia del VAD. El uso de "distancias" entre el vector de características de observación y el modelo puede ser utilizado como regla de decisión.

Se han usado funciones distancia en distintos algoritmos para clasificación de patrones de manera satisfactoria [Rabiner and Juang, 1993], midiendo la desemejanza entre dos vectores de características, usualmente calculados a partir de la señal de entrada y de unidades almacenadas (modelos ó plantillas de referencia). Los sistemas que usan funciones distancia para su regla de decisión, adolecen de problemas de clasificación en ambientes en los que se modifican, con el tiempo, las condiciones del ruido ambiente, como por ejemplo el ruido de murmullo (en inglés "bable"). Las funciones distancia en

los algoritmos VAD suelen basarse en distintas características como energía, energías en sub-bandas, tasa de cruces por cero o espectro de potencia.

Dados dos vectores de características, una función distancia cumple las siguientes propiedades:

- Definida Positiva : $0 \leq d(\mathbf{x}, \mathbf{y}) < \infty, \quad \forall \mathbf{x}, \mathbf{y}$ y $d(\mathbf{x}, \mathbf{y}) = 0$ entonces $\mathbf{x} = \mathbf{y}$
- Simetría : $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y}$.
- Desigualdad triangular : $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z}$.

Una de las distancias más usadas es la Euclídea [1], la cual será empleada en el Capítulo 5 para la formulación de una regla de decisión basada en la característica energía en sub-bandas.

2.3.3. Regla de decisión mediante heurísticas

Últimamente se han desarrollado distintas estrategias para VAD basadas en diversas heurísticas con mayor o menor éxito. Caben destacar los detectores de actividad de voz basados en lógica difusa [Beritelli et al., 2002] (clasificación basada en reglas difusas que caracterizan los segmentos de voz y/o silencio); en redes neuronales artificiales basadas en funciones de base radial (RBF) y un entrenamiento basado en mínimos cuadrados medios (LMS) [Kim and Park, 2000]; su generalización **NO HEURÍSTICA** a máquinas de vectores soporte (SVMs, véase el Capítulo 6), en la que se minimiza en el proceso de entrenamiento el error actual para la obtención de los parámetros de la red neuronal (vectores soporte, SVs); ó *incluso* Algoritmos Genéticos (GAs), como una búsqueda óptima en un espacio de vectores de características (la energía del segmento, la energía diferencia, la energía diferencia en

[1] $d(\mathbf{x}, \mathbf{y}) \equiv \sqrt{\sum_{k=1}^{N} (x(k) - y(k))^2}$

banda baja, la tasa de cruce por cero diferencia y la distorsión espectral (al igual que el estándar G.729B), en un proceso de entrenamiento inicial que usa señales limpias y ruidosas [Estévez et al., 2000].

2.4. Post-Procesado de la Decisión (Suavizado)

La decisiones binarias se realizan usualmente segmento a segmento. Sin embargo esta decisión basada en segmentos no tiene en cuenta la naturaleza continua de los eventos "voz" y "silencio". Por lo tanto, cualquier suavizado a la decisión puede ser aplicado para prevenir recortes de voz ó segmentos de ruido con características similares a la voz (picos ó "spikes"), que son en realidad producidos por un ruido impulsivo. Una versión sencilla de suavizado para eliminar recortes, consiste en extender la decisión en la detección de voz [Kim and Park, 2000] a los frames futuros, lo que produce un aumento en la tasa de acierto de segmentos de voz y una disminución en la de silencio. Este esquema es aplicado a periodos de voz suficientemente largos para evitar prolongar los segmentos "spikes".

Por último las características extraídas en fases anteriores, pueden reutilizarse en el post-procesado para producir reglas de decisión de suavizado aún más complejas, por ejemplo se podría incorporar el conocimiento a priori del modelo de lenguaje para la obtención óptima de los finales de palabra.

3. TESTS ESTADÍSTICOS APLICADOS A COEFICIENTES BIESPECTRALES PARA VAD

En esta sección se desarrolla un algoritmo de detección de actividad de voz eficiente y robusto para la mejora de la tasa de reconocimiento de voz en ambientes ruidos. La aproximación consiste en el uso de tests estadísticos basados en la determinación del biespectro de cumulantes de tercer orden de ruido y voz, y en el uso de un filtro precediendo al VAD siguiendo la filosofía de [Ramírez et al., 2005a] y que en la siguiente sección comentamos. Este algoritmo difiere de otros muchos en la forma en que la decisión queda formulada, fundamentada en el uso de tests estadísticos aplicados sobre una ventana de múltiple observación (MO) que contiene coeficientes biespectrales promediados de la señal de voz. La motivación fundamental del uso de HOS es la habilidad que presentan estos promedio para detectar señales no gaussianas en entornos gaussianos [Tugnait, 1994] y en definitiva caracterizar la complejidad espectral de unas señales frente a otras.

Mejoras significativas en la discriminación de voz/silencio demuestran la efectividad del VAD propuesto. Se muestra que la aplicación de estos

tests de detección conduce a una separación óptima de las distribuciones de ruido y voz, por lo tanto permitiendo una mayor discriminación y un mejor compromiso entre complejidad y rendimiento. El análisis experimental llevado a cabo sobre la base de datos Aurora 3 suministra una evaluación rigurosa del VAD junto con la comparación exhaustiva con respecto a los VADs estandarizados tales como ITU G.729, GSM AMR para transmisión discontinua y ETSI AFE para reconocimiento distribuido de voz (DSR), y otros VAD presentados recientemente.

3.1. Introducción

Un conjunto representativo de métodos de detección de actividad de voz formula la regla de decisión segmento a segmento utilizando valores instantáneos de la divergencia entre la voz y el ruido [Sohn and Su, 1998; Woo et al., 2000]. Sin embargo, experimentos preliminares demuestran que se puede mejorar la robustez del VAD empleando información de largo periodo en la formulación de la regla de decisión [Ramírez et al., 2004b; Ramírez et al., 2004a; Górriz et al., 2005e]. Un método interesante es el algoritmo propuesto por Li [Li et al., 2002], que se basa en el detector de contornos óptimo propuesto en primer lugar por Canny [Canny, 1986], y utiliza filtros FIR óptimos para detección de contornos. Sin embargo, la utilización de métodos alternativos para detección de actividad de voz tales como el uso de filtros no lineales se encuentra todavía como un campo de investigación sin estudiar. En particular, los filtros de estadística ordenada (OSFs: order statistics filters) se emplean en numerosas aplicaciones; entre ellas, la detección de contornos en imágenes [Hwang and Haddad, 1994], habiéndose estudiado el diseño de una clase de OSFs, conocidos como filtros L (L-filters) [Öten and de Figueiredo, 2003]. El diseño de filtros L óptimos no es una tarea fácil y, normalmente, se utilizan métodos aproximados. Un primer método es el uso de filtros quasi-rango [Restrepo et al., 1994] definidos como la diferencia entre filtros de tipo *rank-order*. Aunque estas técnicas se desarrollaron principalmente para procesamiento de imágenes, algunos autores han estudiado su aplicación a la discriminación robusta entre voz y silencio. Cox [Cox and Tinothy, 1980] estudió un esquema estadístico paramétrico de detección utilizando filtros de tipo *rank-order*. La técnica se basaba en la descomposición de la señal por medio de un banco de filtros de cuatro canales en el que se realizaba la operación de ordenación en el dominio del tiempo sobre periodos

de 15 ms y muestras de los cuatro canales. Aunque el algoritmo funcionaba bien y era robusto al ruido, este sufría algunos inconvenientes que tienen que ser estudiados. En primer lugar, es computacionalmente ineficiente puesto que requiere ordenar conjuntos de datos de 400 muestras cada 15 ms. En segundo lugar, el algoritmo de decisión resulta adecuado únicamente para ruido blanco y fallará cuando se consideren otros tipos de ruido de naturaleza paso baja como por ejemplo, el ruido de coche. Estos estudios preliminares motivaron continuar avanzando en el campo explorando áreas directamente relacionadas y su aplicación a la detección de actividad de voz en entornos ruidosos. El algoritmo propuesto en [Ramírez et al., 2005a] utiliza un bloque previo de reducción de ruido y OSFs en subbandas para formular la regla de decisión.

3.2. Etapa de previa de filtrado

Los filtros no lineales entre los que se incluyen los OSF [Öten and de Figueiredo, 2003] y los filtros L, han demostrado ser más efectivos y robustos que los filtros lineales en ciertas aplicaciones [Arce et al., 1986; Ko and Lee, 1991; Pitas and Pitas, 1993]. Como ejemplo, los filtros basados en estadística ordenada se han empleado con éxito en la restauración de señales e imágenes afectadas por ruido aditivo.

El OSF más común es el filtro de mediana que resulta fácil de implementar y exhibe un buen comportamiento en la eliminación de ruido de carácter impulsivo. La salida de un OSF de orden L se define sobre el conjunto de datos $\{x(l-N), ..., x(l), ..., x(l+N)\}$ mediante:

$$y(l) = \sum_{i=1}^{L} a_i x_{(i)} \tag{3.1}$$

donde $L = 2N+1$ and $x_{(i)}$ representa al conjunto de datos anterior reordenado en orden creciente, es decir:

$$\{x_{(1)} \leq x_{(2)} \leq ... \leq x_{(L)}\} \tag{3.2}$$

Nótese que $x_{(1)}$ es el mínimo, $x_{(L)}$, el máximo, y $x_{(N+1)}$, la mediana. Los pesos a_i definen el OSF. Por ejemplo, el filtro de mediana es un caso especial de filtro L con coeficientes $a_i = 1$ para $i = N + 1$ y $a_i = 0$ en cualquier otro caso.

En este Capítulo se estudia la utilización de OSFs [Ramírez et al., 2005a] con VADs basados en tests de coeficientes biespectrales. La aproximación utilizada para el filtro de reducción de ruido se define sobre las energías logarítmicas de la señal en subbandas. En primer lugar se realiza un filtrado de la señal para reducción de ruido y la decisión se realiza sobre la señal filtrada, lo cual añade un retardo adicional al VAD.

3.2.1. Etapa de reducción de ruido

La entrada $x(n)$ se descompone en segmentos de voz de 25 ms con un desplazamiento de 10 ms. Sea $X(m, l)$ la magnitud espectral del segmento l en la banda m ($m = 0, 1, ..., N_{FFT} - 1$). El diseño del bloque de reducción de ruido se basa en la teoría de filtrado óptimo de Wiener siendo la atenuación dependiente de la SNR de la señal procesada. La figura 3.1 muestra un diagrama de bloques de esta etapa de reducción de ruido. Nótese que la salida del VAD se formula sobre la señal filtrada con objeto de mejorar la precisión del VAD en condiciones de ruido. El bloque de reducción de ruido consta de cuatro etapas[Ramírez et al., 2005a]:

i) **Suavizado espectral**. El espectro de potencia se promedia sobre dos segmentos consecutivos y tres bandas espectrales adyacentes.

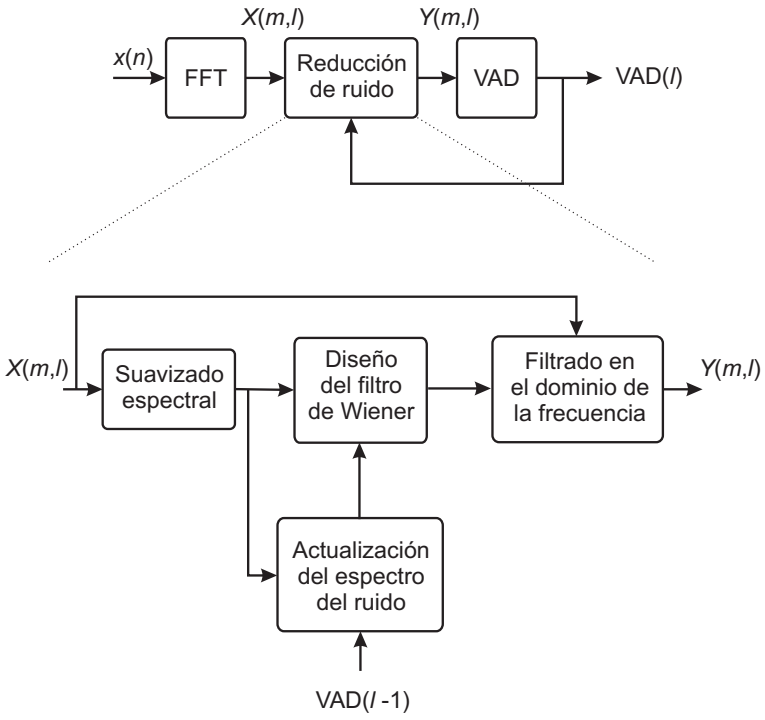


Fig. 3.1: Diagrama de bloques del VAD propuesto.

ii) **Estimación del espectro de ruido**. El espectro del ruido $N_e(m,l)$ se estima por medio de un filtro IIR de primer orden sobre el espectro suavizado $X_s(m,l)$:

$$N_e(m,l) = \lambda N_e(m,l-1) + (1-\lambda)X_s(m,l) \tag{3.3}$$

siendo $\lambda = 0{,}99$ y $m = 0, 1, \ldots, N_{FFT}/2$.

iii) **Diseño del filtro de Wiener (WF)**. En primer lugar se estima el espectro de la señal limpia $S(m,l)$ por medio de una sustracción espectral suavizada:

$$S(m,l) = \gamma S'(m,l-1) + (1-\gamma)\,\text{máx}(X_s(m,l) - N_e(m,l), 0) \tag{3.4}$$

donde $\gamma = 0.98$. A continuación se diseña el WF $H(m,l)$ mediante:

$$H(m,l) = \frac{\eta(m,l)}{1+\eta(m,l)} \tag{3.5}$$

con

$$\eta(m,l) = \text{máx}\left[\frac{S(m,l)}{N_e(m,l)}, \eta_{\text{mín}}\right] \tag{3.6}$$

seleccionando $\eta_{\text{mín}}$ para que el filtro H tenga una atenuación máxima de 20 dB. Nótese que $S'(m,l)$ representa el espectro de la señal filtrada asumiéndose que toma un valor nulo al principio del proceso con objeto de inicializar la iteración. Esta señal se obtiene mediante:

$$S'(m,l) = H(m,l)X(m,l) \tag{3.7}$$

El filtro $H(m,l)$ se suaviza antes de obtener la señal de salida con el fin de eliminar cambios bruscos entre frecuencias próximas que suelen generar ruido musical de carácter no estacionario. De este modo, se consigue reducir la varianza del ruido residual y mejorar la robustez del VAD. El suavizado se realiza truncando la respuesta impulsiva del filtro a 17 coeficientes empleando una ventana de Hanning.

iv) **Filtrado en el dominio de la frecuencia**. La señal de entrada se filtra en el dominio de la frecuencia utilizando el filtro suavizado H_s a fin de obtener una señal de salida ($Y(m,l) = H_s(m,l)X(m,l)$) de mejor SNR .

3.2.2. Test Binario GLRT para VAD

Tras haber realizado la reducción de ruido de la señal de entrada, vamos a desarrollar un test binario basado en GRLT para VAD. Para ello comenzamos denotando a $\{x(t)\}$ las medidas discretas de señal en el sensor. Consideremos el conjunto de variables estocásticas $y_k(t), \quad k = 0, \pm 1, \ldots, \pm M$ obtenidas usando un desplazamiento en la señal de entrada de la forma:

$$y_k(t) = x(t+k) \tag{3.8}$$

donde k es un retraso diferencial (o avance) entre las muestras. Esta definición proporciona un conjunto nuevo de $2M+1$ variables vector $\mathbf{y}_k = [y_k(1), \ldots, y_k(N)]$ mediante selección de N muestras de la señal de entrada. Este conjunto puede ser representado mediante la matriz de Toeplitz asociada:

$$T_{x(t)} = \begin{pmatrix} y_{-M}(1) & \ldots & y_{-M}(N) \\ y_{-M+1}(1) & \ldots & y_{-M+1}(N) \\ \ldots & \ldots & \ldots \\ y_M(1) & \ldots & y_M(N) \end{pmatrix} \tag{3.9}$$

Usando este modelo la detección de voz/silencio puede describirse usando dos hipótesis esenciales (reorganizando índices):

$$\begin{array}{c} H_0 : y_k(t) = n_k(t) \\ H_1 : y_k(t) = s_k(t) + n_k(t) \end{array} \; ; \; \begin{array}{c} t = 1, \ldots, N \\ k = 0, \pm 1, \ldots, \pm M \end{array} \tag{3.10}$$

donde $s_k(t)$ denota una señal de voz no Gaussiana con un retraso k y $n_k(t)$ representa las secuencias de ruido aditivo, respectivamente. Se asume que todos los procesos involucrados en este desarrollo son conjuntamente estacionarios y de media cero. Consideremos la función cumulante de tercer orden definido como $C_{\mathbf{y}_k\mathbf{y}_l} \equiv E[y_0(t)y_k(t)y_l(t)]$, y la transformada discreta de Fourier bidimensional (2D DFT) de $C_{\mathbf{y}_k\mathbf{y}_l}$, esto es la función biespectral:

$$\mathcal{C}_{\mathbf{y}_k\mathbf{y}_l}(\omega_1, \omega_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} C_{\mathbf{y}_k\mathbf{y}_l} \cdot \exp(-j(\omega_1 k + \omega_2 l)) \tag{3.11}$$

Muestreando la ecuación 3.11, el estimador biespectro puede expresarse como:

$$\hat{\mathcal{C}}_{\mathbf{y}_k\mathbf{y}_l}(n, m) = \sum_{k=-M}^{M} \sum_{l=-M}^{M} C_{\mathbf{y}_k\mathbf{y}_l} \cdot w(k, l) \cdot \exp(-j(\omega_n k + \omega_m l)) \tag{3.12}$$

donde $\omega_{n,m} = \frac{2\pi}{M}(n, m)$ con $n, m = -M, \ldots, M$ son las frecuencias discretas, $w(k, l)$ es la función de ventana (para reducir "aliasing" [Nikias and Petropulu, 1993]) y $C_{\mathbf{y}_k\mathbf{y}_l} = \frac{1}{N}\sum_{t=1}^{N} y_0(t)y_k(t)y_l(t) = \frac{1}{N}\mathbf{y}_0 \cdot \mathbf{y}_k \cdot \mathbf{y}_l$. La estimación del biespectro es discutida en profundidad en [Brillinger and Rossenblatt, 1975] y muchos otros, dándose las condiciones necesarias para alcanzar consistencia en la estimación. La estimación es asintóticamente consistente cuando la varianza del estimador es nula, es decir, cuando el número de muestras tiende a infinito.

La decisión de nuestro algoritmo está basada en tests estadísticos como son el test Generalizado de cociente de probabilidades (GLRT) [Subba-Rao, 1982] y el tests distribuido como una χ^2 central bajo H_0 [Hinich, 1982]. Los

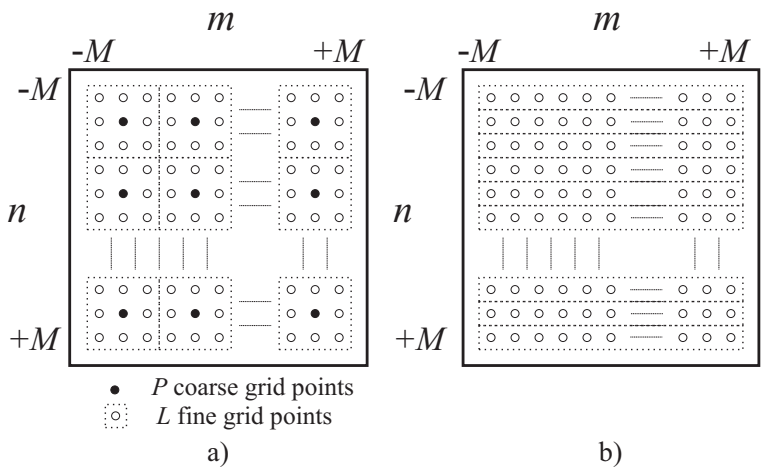


Fig. 3.2: Vecindades usadas en la estimación de los parámetros biespectrales a) Mallas fina y gruesa. Distribuimos uniformemente P puntos con L puntos vecinos. b) Promedio por filas para la estimación del biespectro integrado.

abreviaremos como tests GLRT y χ^2. Los tests están basados en distribuciones asintóticas, donde distintas simulaciones en [Tugnait, 1993] muestran que el test χ^2 requiere un conjunto de datos mayor para alcanzar una distribución asintótica consistente con el modelo teórico.

3.2.3. GLRT

Consideremos el dominio completo en frecuencias biespectrales $0 \leq \omega_{n,m} \leq 2\pi$ y definamos P puntos uniformemente distribuidos en esta malla (m, n), llamada "malla gruesa" como muestra la figura 3.2. Definamos a su vez la "malla fina" que contiene los L puntos más cercanos a las parejas de frecuencias de la malla gruesa. Se cumple en este caso que $2M + 1 = P \cdot L$. Si reordenamos las componentes del conjunto de L estimaciones Biespectrales $\hat{\mathcal{C}}(n_l, m_l)$ donde $l = 1, \ldots, L$, sobre la malla fina alrededor de la pareja bifrecuencia en un vector de dimensión L, β_{ml} donde $m = 1, \ldots, P$ indexa

la malla gruesa [Subba-Rao, 1982] y definimos P-vectores $\phi_i(\beta_{1i}, \ldots, \beta_{Pi})$, $i = 1, \ldots, L$; el test GLRT, con las hipótesis anteriores es:

$$H_0 : \mu = \mu_n \quad against \quad H_1 : \eta \equiv \mu^T \sigma^{-1} \mu > \mu_n^T \sigma_n^{-1} \mu_n \tag{3.13}$$

donde μ y σ son las estimaciones de máxima probabilidad de la media y la covarianza del vector $\mathcal{C} = (\mathcal{C}_{\mathbf{y}_k \mathbf{y}_l}(m_1, n_1) \ldots \mathcal{C}_{\mathbf{y}_k \mathbf{y}_l}(m_P, n_P))$, es decir:

$$\begin{array}{c} \mu = 1/L \sum_{i=1}^{L} \phi_i \\ \sigma = 1/L \sum_{i=1}^{L} (\phi_i - \mu)(\phi_i - \mu)^T \end{array} \tag{3.14}$$

Por lo tanto, la presencia de voz es detectada si:

$$\eta > \eta_n \tag{3.15}$$

donde η_n es un umbral determinado con un nivel de significación, por ejemplo la probabilidad de falsa alarma. Notamos que:

1. Asumimos independencia estadística entre las componentes del biespectro de la señal $s(t)$ y del ruido aditivo $n(t)$ [1], por lo tanto:

$$\mu = \mu_n + \mu_s; \quad \sigma = \sigma_n + \sigma_s \tag{3.16}$$

2. El lado derecho de la hipótesis H_1 debe ser estimado en cada "*frame*" (es desconocido *a priori*). En nuestro algoritmo la aproximación se basa en la información de los frames anteriores clasificados como ruido.

El estadístico considerado η está distribuido como una $F_{2P,2(L-P)}$ central bajo la hipótesis nula. Por lo tanto un test Neyman-Pearson puede ser aplicado con un determinado nivel de significación α.

[1] Esta es una asunción aceptable [Górriz et al., 2005e] dado que los resultados que obtenemos de ella son bastante significativos. Aquí, no asumimos en cambio que las secuencias de ruido $n_k(t)$ $k = 0 \ldots \pm M$ son gaussianas, son modeladas como un sesgo biespectral adaptable. Esta hipótesis será eludida en el Capítulo 4 donde los parámetros de la hipótesis alternativa serán calculados de manera teórica.

3.2.4. Tests χ^2

A continuación consideramos el test estadístico distribuido como χ^2_{2L}[Hinich, 1982] dado por:

$$\eta = \sum_{m,n} 2KN_B^{-1}|\Gamma_{\mathbf{y}_k\mathbf{y}_l}(m,n)|^2 \tag{3.17}$$

donde K es el número de segmentos no solapados, cada uno de tamaño N_B, dada la secuencia de muestras, $\Gamma_{\mathbf{y}_k\mathbf{y}_l}(m,n) = \frac{|\hat{\mathcal{C}}_{\mathbf{y}_k\mathbf{y}_l}(n,m)|}{[S_{\mathbf{y}_0}(m)S_{\mathbf{y}_k}(n)S_{\mathbf{y}_l}(m+n)]^{0,5}}$ la cual está distribuida asintóticamente como una $\chi^2_{2L}(0)$, donde $S_{\mathbf{y}_k}$ representa la potencia espectral de $y_k(t)$ y L denota el número de puntos en el dominio principal. El tests Neyman-Pearson para un nivel de significación (probabilidad de falsa alarma) α se representa como:

$$H_1 \quad if \quad \eta > \eta_\alpha \tag{3.18}$$

donde η_α es determinado usando tablas de la distribución χ^2 central. Note como el denominador de $\Gamma_{\mathbf{y}_k\mathbf{y}_l}(m,n)$ es desconocido *a priori* por lo que debe ser estimado como la función biespectral (es decir, $\hat{\mathcal{C}}_{\mathbf{y}_k\mathbf{y}_l}(n,m)$). Esto requiere un mayor número de muestras como mencionamos al principio de la presente sección.

3.2.5. Una aproximación eficiente: el Biespectro Integrado

Como primera aproximación para detección y destacar el potencial del método propuesto proponemos una decisión basada en el promedio de las componentes del biespectro en una dimensión de frecuencia (véase figura 3.2) en vez de promediar sobre las mallas gruesa y delgada tal y como plantea [Subba-Rao, 1982]. De este modo definimos η como:

$$\eta = \frac{1}{L \cdot P}\sum_{i=1}^{P}\sum_{j=1}^{L}\hat{\mathcal{C}}(i,j) = \frac{1}{L}\sum_{j=1}^{L}\mu(j) \tag{3.19}$$

donde L,P define la malla seleccionada (altas frecuencias con variabilidad notable). Se puede demostrar fácilmente de [Tugnait, 1994] que:

$$S_{sx}(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{C}(\omega,\omega_2)d\omega_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{C}(\omega_1,\omega)d\omega_1 \tag{3.20}$$

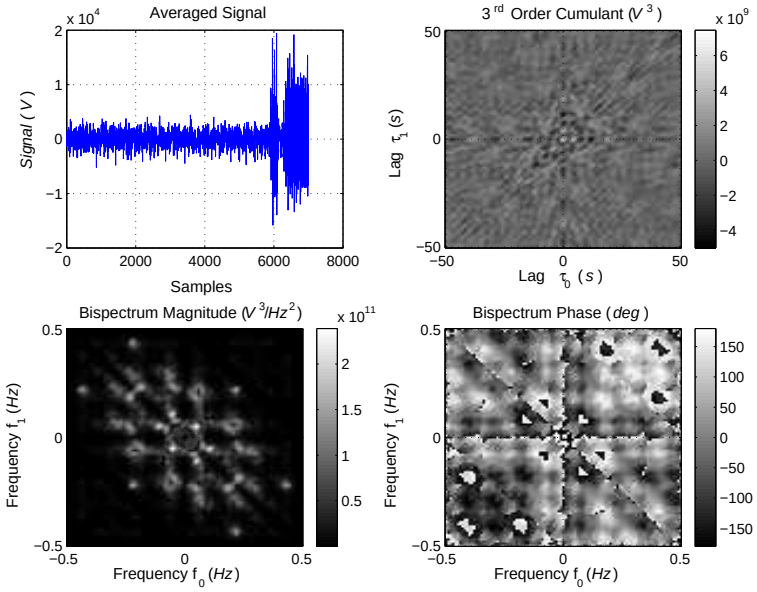
Es decir, el espectro cruzado entre la señal $x(t)$ y su cuadrado $s(t)$ puede interpretarse como un estimador del biespectro integrado de $x(t)$. Este biespectro integrado formará la base para los tests estadísticos usados en este trabajo para detección de señales no gaussianas en ruido y además para el repertorio de VADs desarrollados con LRT, en base a aproximaciones más rigurosas, en el Capítulo 4. La ventaja de esta implementación es el bajo coste computacional a diferencia de los tests basados en el biespectro clásico para detección de actividad de voz usados por primera vez en [Górriz et al., 2005e].

La figura 3.3 muestra las diferencias entre el cumulante y el biespectro en la señal de voz y en el ruido. Puede concluirse claramente que el biespectro de la señal exhibe características discriminativas para una detección voz/silencio.

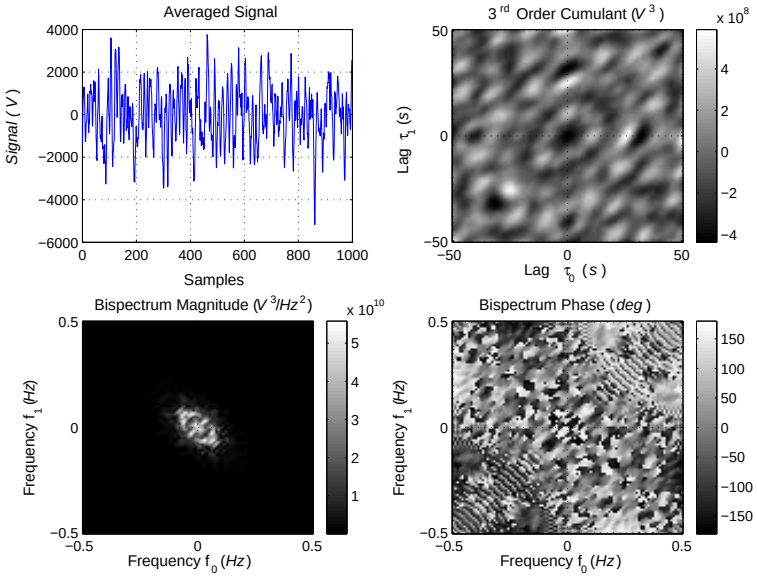
La figura 3.4(a) muestra la transformación de malla fina a gruesa de la ecuación 3.19 cuando se define como un promedio por filas (biespectro integrado) de la representación 2-D del biespectro. La estimación espectral retiene el poder discriminativo y exhibe importantes diferencias entre señales de voz y ruido.

3.2.5.1. Ejemplo ilustrativo

La figura 3.4(b) muestra la operación del VAD propuesto en esta sección sobre una frase de la base de datos Spanish SpeechDat-Car (SDC) [Moreno et al., 2000]. La frase tiene transcripción fonética: [“siete”, “θinko”, “dos”, “uno”, “otSo”, “seis”]. Fig 3.4(b) muestra el valor de η frente al tiempo en esa

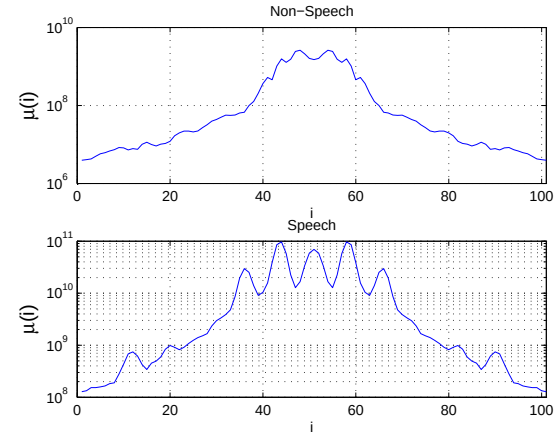


(a)

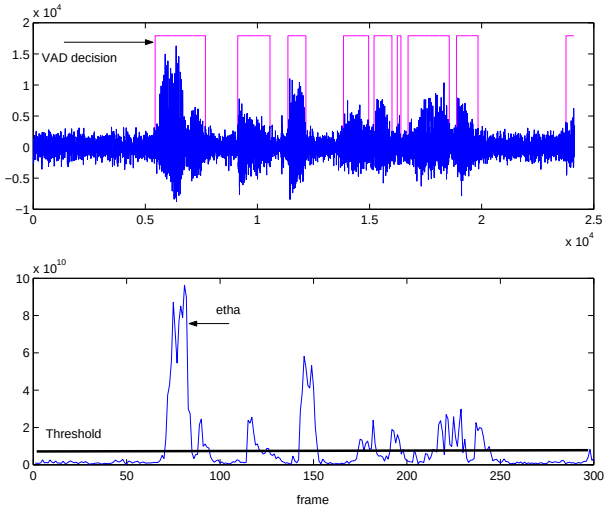


(b)

Fig. 3.3: Distintas características que permiten la detección de actividad de voz: cumulantes de 3^{er} orden, magnitud y fase biespectral sobre $\mathbf{y}_k$. (a) Características de la señal de voz. (b) Características de la señal de ruido.



(a)



(b)

Fig. 3.4: Operación del VAD sobre una frase de la base de datos SDC en español. (a) Biespectro promediado por filas para periodos de voz y silencio. (b) Evaluación de η y decisión del VAD.

frase. Observe como asumiendo un umbral de decisión η_0 ligeramente superior al valor inicial de la magnitud η sobre el primer segmento (ruido), podemos alcanzar una buena decisión para VAD. La figura también muestra el buen comportamiento del método propuesto para detectar sonidos fricativos (como los derivados de la transcripción fonética) incluso cuando usamos una única ventana en estos experimentos preliminares.

Alternativamente, hemos usado en la parte experimental información de retardo largo "long-term information (LTI)" en la decisión del VAD, tal y como se propuso en [Ramírez et al., 2004a], lo cual mejora esencialmente el rendimiento del VAD. Con esta aproximación, la decisión del VAD se formula no solo sobre el segmento actual l, usando $y_k(t)$, $k = 0, \ldots, \pm M$, sino además sobre $2m + 1$ segmentos o "frames" anteriores y futuros, esto es, el conjunto de segmentos $\{l - m, \ldots, l, \ldots, l + m\}$ que incluye múltiples observaciones consecutivas de la señal de entrada $y_k(t+S\cdot j)$, $k = 0, \ldots, \pm M$, $j = 0, \ldots, \pm m$, donde S define el desplazamiento del VAD. De este modo, el VAD realiza una detección anticipada del comienzo y retrasada del fin de las palabras lo cual hace innecesario mecanismos tales como el de periodo de guarda "*hang-over*".

3.2.6. Discusión

El VAD propuesto no resulta sensible al valor del umbral por las siguientes razones: *i*) El uso de un nivel adaptable sólo se encuentra motivado por la mejora obtenida en condiciones de bajo nivel de ruido. El incremento del valor del nivel en estas condiciones ayuda a identificar con mayor precisión los segmentos que contienen sólo ruido sin perjudicar el rendimiento del VAD en condiciones de alto ruido, *ii*) para reducir la sensibilidad del VAD al nivel de ruido se utilizan cotas superior e inferior para el nivel del umbral,

y *iii*) la utilización de este esquema de adaptación ha proporcionado buenos resultados pero no es la clave para el alto nivel de rendimiento del detector propuesto. Del mismo modo, los experimentos que se han realizado sobre distintas bases de datos proporcionaron buenos resultados para todos los ruidos y condiciones de SNR. *iv*) La ventaja del uso de tests estadísticos radica en el hecho de poder trabajar a un determinado nivel de significación, o lo que es lo mismo a una probabilidad de falsa alarma seleccionada por las condiciones óptimas de trabajo en cada aplicación.

El algoritmo propuesto aprovecha el bloque de reducción de ruido para mejorar su robustez frente al ruido ambiental. Es interesante aclarar cómo se ha garantizado la convergencia del bucle de realimentación de la figura 3.1. La solución adoptada ha sido asumir que cada frase de la base de datos contiene un periodo de silencio al principio de la frase para la inicialización del proceso. Si la frase no empezara de esta manera el algoritmo podría fallar al principio en la evaluación del espectro de ruido y la detección posterior podría ser completamente errónea. Sin embargo, la mayoría de los algoritmos de la bibliografía necesitan una estimación de los parámetros del ruido y, normalmente, realizan esta suposición para la inicialización. Un problema común de todos los algoritmos tras la inicialización es la incorrecta estimación de la estadística del ruido cuando el VAD falla. Los resultados presentados en próximas secciones mostrarán la efectividad del algoritmo propuesto que se encuentra libre de problemas de convergencia.

El algoritmo propuesto tiene un retardo de N segmentos correspondiente al tamaño de ventana de múltiple observación. Este hecho puede ser un obstáculo para algunas aplicaciones en tiempo real, pero para otras, como para reconocimiento de voz, no es una limitación y la mejora en las prestaciones justifica su utilización como se demostrará en el resto de este trabajo.

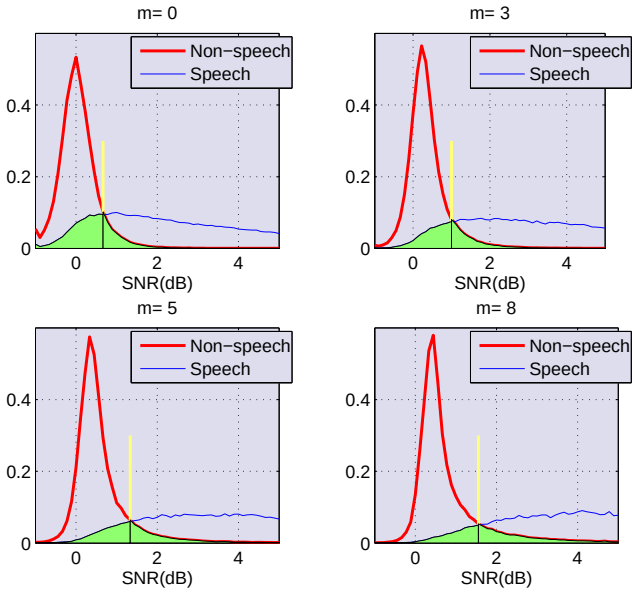


Fig. 3.5: Distribuciones de la voz y del silencio y probabilidades de error para un clasificador óptimo de Bayes con m= 0, 3, 5 y 8.

3.2.7. Análisis del detector

Para clarificar las motivaciones para el algoritmo propuesto se han estudiado las distribuciones de la variable de decisión definida mediante la ecuación 3.13 como función de la longitud de la ventana MO usada m. En el análisis se utilizó etiquetado manual de la base de datos SpeechDat-Car (SDC) española [Moreno et al., 2000]. Esta base de datos consta de grabaciones con dos tipos de micrófonos: un micrófono de proximidad y un micrófono distante, en diferentes condiciones de conducción: a) coche parado y motor encendido, b) tráfico de ciudad, baja velocidad y carretera rugosa y c) alta velocidad sobre buena carretera.

Para este análisis se consideró el entorno más desfavorable (grabaciones realizadas con el micrófono distante a alta velocidad y sobre buena carretera)

con una SNR promedio de aproximadamente 5 dB. Se midió la SNR durante periodos de voz y silencio para diferentes valores de m y se construyeron los correspondientes histogramas y distribuciones de probabilidad. En la figura 3.5 se muestran las distribuciones de probabilidad de los segmentos de voz y silencio para $m = 0$, 3, 5 y 8. Se ve claramente que conforme aumenta la longitud de la ventana de MO se reduce la varianza del ruido y la distribución de la voz se encuentra más desplazada hacia la derecha. Por tanto, el solapamiento entre las dos distribuciones es menor y el error de clasificación se reduce. Como conclusión se obtiene que las distribuciones de la voz y el ruido se pueden discriminar mejor cuando se incrementa la longitud de la ventana haciéndose el VAD más robusto al ruido del entorno.

La reducción del solapamiento entre las distribuciones permite elevar la discriminación entre la voz y el silencio. Este hecho se demuestra calculando los errores de clasificación de la voz y el silencio para un clasificador óptimo de Bayes. Nótese que en la figura 3.5 se han representado las áreas que representan la probabilidad de detectar incorrectamente la voz y el silencio, así como el correspondiente umbral de detección que minimiza el error total. Se demuestra así que el error de detección de la voz se reduce cuando se incrementa la longitud de la ventana (m) mientras que esta mejora en la robustez del detector se ve únicamente afectada por un moderado aumento del error en la detección del silencio [Ramírez et al., 2005a]. Esta mejora se obtiene como consecuencia directa de la reducción del solapamiento entre las distribuciones de probabilidad cuando se eleva el valor de m tal y como se muestra en la figura 3.5. Por tanto, incrementar la longitud de la ventana es beneficioso en entornos de alto ruido puesto que el VAD introduce de forma artificial un periodo de guarda ("*hang-over*") que reduce la pérdida de comienzos y finales de palabra.

Cuando se emplea el bloque de reducción de ruido se consigue una mayor reducción del error de detección de la voz oscilando éste entre el 20 % y el 9 % [Ramírez et al., 2005a]. Como conclusión se obtiene que si se considera el bloque de reducción de ruido se reducen los errores de clasificación haciendo el VAD más robusto frente al ruido.

Por tanto, la utilización de tanto el bloque de reducción de ruido como los tests GLRT basados en promedios biespectrales integrados proporciona importantes ventajas en la detección de voz en una señal ruidosa puesto que los errores de clasificación se reducen de forma significativa.

3.2.8. Evaluación y comparación

Se suelen realizar diferentes análisis y experimentos para evaluar el comportamiento de los algoritmos de detección de actividad de voz. El análisis se suele centrar principalmente en la determinación de la probabilidad de error o error de clasificación a diferentes niveles de SNR [Marzinzik and Kollmeier, 2002], y en la evaluación de la influencia del VAD en sistemas de procesamiento de la voz [Bouquin-Jeannes and Faucon, 1995b]. Del mismo modo, se han considerado tests subjetivos para la evaluación de VADs que operan en combinación con codificadores de voz [Benyassine et al., 1997]]. En esta sección se describe el entorno experimental y los tests objetivos utilizados para evaluar el algoritmo propuesto.

3.2.8.1. Análisis de discriminación frente al ruido

En primer lugar, el VAD propuesto se evaluó en términos de su capacidad para discriminar la voz y el silencio en función de la SNR. En este análisis se empleó la base de datos AURORA-2 [Hirsch and Pearce, 2000] desarrollada para la evaluación y el desarrollo del estándar de ETSI para reconocimiento

distribuido de voz [ETSI, 2002]. Esta base de datos incluye la base de datos limpia TIdigits, que contiene frases de hasta siete dígitos conectados leídas por nativos americanos en inglés como fuente, y una selección de ocho ruidos reales que se han mezclado artificialmente a la voz con SNRs de 20, 15, 10, 5, 0 y -5 dB. Estos ruidos se han grabado en diferentes lugares (metro, coche, sala de exposiciones, restaurante, calle, etc) y representan los escenarios más comunes para terminales de telecomunicaciones. Para realizar este análisis se realizó el etiquetado semiautomático de la base de datos limpia como segmentos de voz y de silencio. El rendimiento del detector se evaluó en función de la SNR en términos de las tasas de acierto de los segmentos de silencio (HR0) y de voz (HR1) que se definen como la fracción de todas los segmentos de silencio o de voz que se detectan correctamente como tales, es decir:

$$HR0 = \frac{N_{0,0}}{N_0^{Ref}} \qquad HR1 = \frac{N_{1,1}}{N_1^{Ref}} \tag{3.21}$$

siendo N_0^{Ref} y N_1^{Ref} el número de segmentos de silencio y de voz reales de la base de datos etiquetada, respectivamente, mientras que $N_{0,0}$ y $N_{1,1}$ denotan los segmentos de silencio y de voz clasificados correctamente por el VAD.

Del análisis del VAD se pueden extraer las siguientes conclusiones(Tabla 3.1):

- El VAD propuesto obtiene el mejor compromiso entre los diferentes VADs analizados. Da buenos resultados en la detección del silencio y exhibe una degradación muy lenta en la detección de la voz en condiciones desfavorables de ruido.

- El estándar G.729 sufre una reducida capacidad de detección de la voz cuando aumenta el nivel de ruido mientras que la detección del silencio

es eficiente (85 %) en condiciones limpias y pobre (20 %) en condiciones de elevado ruido.

- El estándar AMR1 tiene un comportamiento extremadamente conservativo con una elevada tasa de acierto de los segmentos de voz en el rango completo de SNRs; sin embargo, su tasa de acierto del silencio es demasiado baja (10 %) lo cual lo hace poco útil en aplicaciones de procesamiento de voz.

- El estándar AMR2 representa una clara mejora frente a G.729 y AMR1, tiene una mejor capacidad de detección del silencio pero sigue sufriendo una rápida degradación en la detección de la voz cuando disminuye la SNR.

- El detector utilizado en el estándar AFE en la etapa de filtrado de Wiener, al estar basado únicamente en la energía, tiene reducida precisión en la detección de la voz con una rápida degradación cuando la SNR disminuye.

- El VAD utilizado en el estándar AFE para "frame-dropping" tiene alta precisión en la detección de los segmentos de voz y baja efectividad en la detección de los silencios. El objetivo de este detector es el de no perder voz lo cual originaría errores de borrado en el sistema reconocimiento.

La tabla 3.1 resume estos resultados y las ventajas que demuestra el algoritmo propuesto en términos de las tasas de acierto promedio para todos los ruidos y condiciones de SNR. Nótese que en esta tabla se incluyen resultados para otros algoritmos de detección de actividad de voz publicados recientemente [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999].

Tab. 3.1: Tasa de acierto promediada para el rango completo de valores de SNRs. Comparación del VAD propuesto con estándares y otros algoritmos publicados recientemente.

	G.729	AMR1	AMR2	AFE (WF)	AFE (FD)
HR0 (%)	31.77	31.31	42.77	57.68	28.74
HR1 (%)	93.00	98.18	93.76	88.72	97.70
	Woo	Li	Marzinzik	Sohn	**GLRT**/χ^2
HR0 (%)	55.40	57.03	52.69	43.66	60.27/41.28
HR1 (%)	88.41	83.65	93.04	94.46	97.40/92.52

Estos resultados demuestran claramente que no existe un VAD óptimo para todas las aplicaciones. Cada VAD se diseña y optimiza para una aplicación concreta. Por tanto, la evaluación del VAD tiene que realizarse considerando el objetivo específico para el que fue diseñado. Con frecuencia los detectores evitan la pérdida de segmentos de voz pero tienen un comportamiento muy poco preciso en la detección de las pausas (por ejemplo, el estándar AMR1). Por esta razón, los dos parámetros tienen que considerarse conjuntamente en la evaluación. A continuación se realiza un análisis de discriminación más preciso basado en las curvas ROC.

3.2.8.2. Curvas ROC

La tabla de discriminación 3.1 muestra el comportamiento del detector una vez fijado el umbral de detección. Sin embargo, dependiendo de la aplicación, en ocasiones será necesario modificar el punto de operación del detector para hacerlo más efectivo en la detección de las pausas o en la detección de

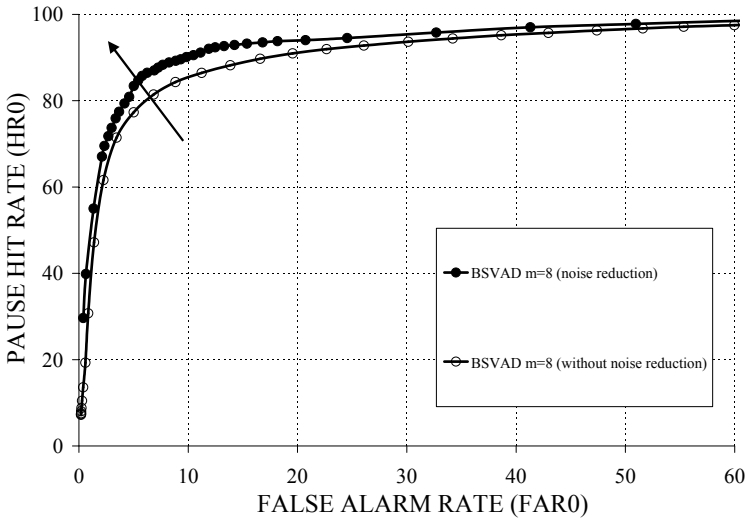


Fig. 3.6: Efectividad del bloque de reducción de ruido (High: alta velocidad, buena carretera, 5 dB de SNR) para una ventana de múltiple observación con $m = 8$.

los periodos de voz. Por esta razón, se introducen las curvas ROC ("Receiver Operating Characteristics") en la evaluación de los algoritmos de detección de actividad de voz. Estas curvas representan la variación de las tasas de error del detector cuando varía el umbral de detección y describen de forma más precisa la tasa de error del VAD.

En este análisis se utiliza la base de datos SpeechDat-Car española [Moreno et al., 2000]. Esta base de datos contiene 4914 grabaciones con micrófonos distante y de proximidad de más de 160 locutores. Las frases se encuentran agrupadas en tres subconjuntos que representan diferentes condiciones de conducción y SNR. El etiquetado semiautomático se realizó sobre las grabaciones realizadas con el micrófono de proximidad que se encuentran menos afectadas por el ruido.

Para construir las curvas ROC se determinaron las tasas de acierto del silencio (HR0) y de la voz (HR1) en cada condición de ruido y se represen-

taron los valores de (HR0, FAR0) estando la tasa de falsas alarmas definida como FAR0= 1-HR1.

3.2.8.3. Eficiencia de la etapa de reducción de ruido

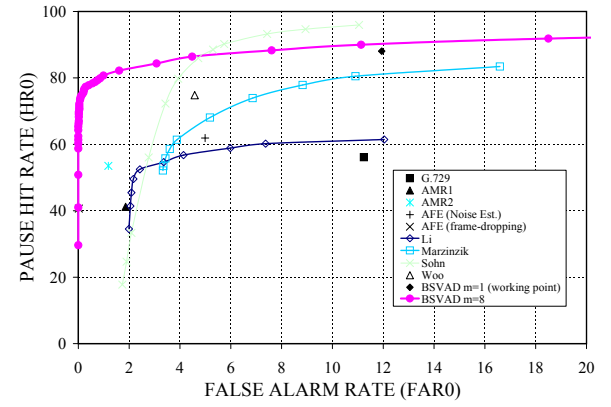
Antes de mostrar resultados comparativos del rendimiento del VAD, vamos ilustrar el efecto del bloque de reducción de ruido previo a la operación del detector. La figura 3.6 muestra la influencia del bloque de reducción de ruido: en primer lugar, no consideramos el bloque de reducción de ruido con objeto de observar con mayor claridad el efecto que tiene este en la detección. De la figura se puede deducir que las curvas ROC se desplazan hacia arriba y hacia la izquierda cuando se usa esa etapa mejorando así la eficacia del detector. Como conclusión, podemos afirmar que cuando se emplea el bloque de reducción de ruido, el VAD demuestra un desplazamiento adicional de la curva tal y como se muestra en la figura 3.6.

3.2.8.4. Comparación con otros algoritmos

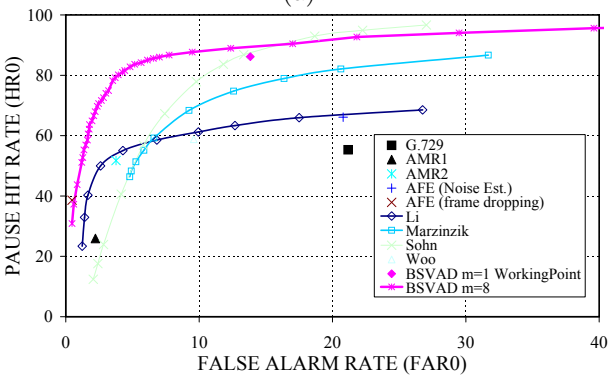
La figura 3.7 muestra las curvas ROC para las tres condiciones de ruido y grabaciones realizadas con el micrófono distante. En las gráficas se muestran los puntos de operación de los estándares G.729 [ITU, 1996], AMR [ETSI, 1999] y AFE [ETSI, 2002], así como las curvas y/o punto de operación de los algoritmos de Li [Li et al., 2002], Marzinzik [Marzinzik and Kollmeier, 2002], Sohn [Sohn et al., 1999] y Woo [Woo et al., 2000]. El detector propuesto proporciona mejores resultados pudiéndose extraer las siguientes conclusiones:

- El punto de operación del estándar G.729 se desplaza hacia la derecha en el espacio ROC cuando disminuye la SNR.

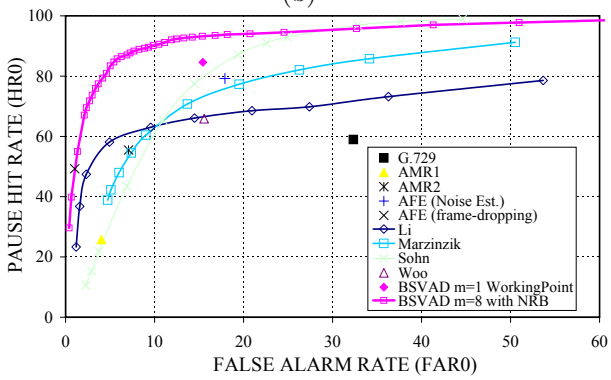
- AMR1 opera en un punto con baja tasa de falsas alarmas pero exhibe



(a)



(b)



(c)

Fig. 3.7: Curvas ROC obtenidas para diferentes subconjuntos de la base de datos SpeechDat-Car española: (a) *Quiet* (coche parado, motor encendido, SNR de 12 dB). (b) *Low* (tráfico de ciudad, baja velocidad, carretera rugosa, SNR de 9 dB). (c) *High* (alta velocidad, buena carretera, SNR de 5 dB).

un valor muy bajo de la tasa de acierto del silencio.

- AMR2 proporciona claras ventajas sobre G.729 y AMR1 consiguiendo una importante reducción de la tasa de falsa alarmas cuando se compara con G.729, y más elevada tasa de acierto del silencio que AMR1.

- El VAD utilizado en el AFE para estimación del ruido en filtrado de Wiener opera en un punto del espacio ROC con buena capacidad de detección del silencio pero con altas falsas alarmas. Sufre una rápida degradación en la precisión cuando las condiciones se hacen más ruidosas.

- El VAD incluido en el AFE para "*frame-dropping*" se ha diseñado únicamente en el estándar para ese propósito y, por tanto, tiene un comportamiento conservativo. Tiene baja tasa de acierto del silencio con baja tasa de falsas alarmas.

- El VAD propuesto opera también en un punto de bajas falsas alarmas pero con muy buenos niveles de precisión en la detección del silencio cuando se compara con el resto de algoritmos.

Por tanto, de entre todos los métodos analizados, el VAD propuesto es el que opera con más bajas falsas alarmas para un valor dado de la tasa de aciertos del silencio y también, con el más alto nivel de precisión en la detección del silencio para una tasa de falsas alarmas dada. Las ventajas son especialmente importantes sobre G.729, que se usa en el estándar para transmisión discontinua, y sobre el algoritmo de Li, que emplea filtros lineales óptimos para detección de los límites de las palabras. El método propuesto también obtiene una mejora considerable sobre el VAD de Marzinzik, que realiza un seguimiento de las envolventes espectrales de la señal, y sobre el

VAD de Sohn, que formula la regla de decisión por medio de un test basado en el cociente de probabilidades. La figura 3.7 muestra también la capacidad del algoritmo de adaptación del umbral utilizado para seleccionar el punto de operación (sobre la curva ROC) más adecuado a cada condición de ruido.

Vale la pena mencionar que los experimentos de evaluación descritos anteriormente proporcionan una primera evaluación de la precisión del VAD. Los experimentos de reconocimiento sobre las bases de datos AURORA son una medida directa de la calidad del VAD en el contexto de la aplicación para la que fue diseñado. Para tal tarea vamos a desarrollar otro conjunto de detectores más específicos para este campo (en el test y en la magnitud biespectral usada) y que no presentan una etapa de reducción de ruido precediéndolos (el gran inconveniente de la anterior aproximación en cuanto a retardo adicional de la etapa de filtrado, y enmascaramiento del rendimiento del detector), dada su capacidad de discriminación que se deriva de la formulación de una regla de decisión suave en cociente de probabilidades, lo cual confiere al algoritmo una gran robustez frente al ruido, como veremos en el siguiente Capítulo.

4. LRT SOBRE VENTANA DE MÚLTIPLE OBSERVACIÓN DE COEFICIENTES BIESPECTRO INTEGRADO

En esta sección presentamos los últimos avances obtenidos en esta línea de trabajo. En concreto se desarrolla una nueva metodología robusta para detección de actividad de voz basada en test estadísticos de cocientes de distribuciones de probabilidad (LRT: "*likelihood-ratio test*") definidos sobre múltiples observaciones (MOLRT: "*multiple observation likelihood-ratio test*") biespectrales de la señal [Górriz et al., 2005e; Ramírez et al., 2005b]. Esta técnica, además de demostrar una inherente robustez frente al ruido, se manifiesta computacionalmente eficiente dada su implementación integrada [Ramírez et al., 2006a]. La idea fundamental se deriva de estudios previos y en la mejora obtenida cuando se emplea información de retardo largo en la formulación de la regla de decisión así como las varianza de los estimadores. El algoritmo de detección se formula desde un punto de vista estadístico siendo necesaria la introducción de un modelo estadístico para la señal y el ruido.

4.1. La Función Biespectral

El biespectro de una señal determinista, continua en el tiempo $x(t)$ se define como [Brillinger and Rosenblatt, 1968; Nikias and Raghuveer, 1987]

$$B(\omega_1, \omega_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C_{3x}(\tau_1, \tau_2) \exp\{-j(\omega_1\tau_1 + \omega_2\tau_2)\} d\tau_1 d\tau_2 \qquad (4.1)$$

donde

$$C_{3x}(\tau_1, \tau_2) = E\{x^*(t)x(t+\tau_1)x(t+\tau_2)\} = \int_{-\infty}^{+\infty} x^*(t)x(t+\tau_1)x(t+\tau_2)dt \quad (4.2)$$

es la función de correlación de tercer orden de $x(t)$, y $\omega = 2 * \pi * f$ con frecuencia normalizada f. Dadas las propiedades de simetría, el biespectro de una señal real queda unívocamente definido por sus valores en la región triangular $0 \leq \omega_2 \leq \omega_1 \leq \omega_1 + \omega_2 \leq \pi$, suponiendo que no se produce solapamiento ("*aliasing*") biespectral.

De manera similar, para señales discretas en el tiempo, el biespectro se define como:

$$B_x(\omega_1, \omega_2) = \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{3x}(i, k) \exp\{-j(\omega_1 i + \omega_2 k)\} \qquad (4.3)$$

donde $C_{3x}(i, k) = E\{x^*(n)x(n+i)x(n+k)\}$ es el cumulante de tercer orden del proceso $x(n)$. Note que, dada la definición anterior, el cumulante de tercer orden puede expresarse como:

$$C_{3x}(i, k) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} B_x(\omega_1, \omega_2) \exp\{j(\omega_1 i + \omega_2 k)\} d\omega_1 d\omega_2 \qquad (4.4)$$

Aunque el biespectro tiene todas las ventajas provenientes del uso de cumulantes y poli-espectro en general, su uso directo tiene dos limitaciones serias:

i) La computación del biespectro en toda la región triangular es costosa, y *ii*) el uso de una plantilla de dos dimensiones para clasificación no es práctica. Para usar el biespectro integrado de manera eficiente, se han propuesto distintos métodos [Tugnait, 1994; Tugnait, 1995] para distintas aplicaciones [Zhang et al., 2001; Liao and Bao, 1998].

4.1.1. Biespectro Integrado

Sea $x(t)$ un proceso aleatorio estacionario de media nula. Si definimos $\tilde{y}(t) = x^2(t) - E\{x^2(t)\}$, la correlación cruzada entre $\tilde{y}(t)$ y $x(t)$ se puede expresar como:

$$r_{\tilde{y}x}(k) = E\{\tilde{y}(t)x(t+k)\} = E\{x^2(t)x(t+k)\} - \overbrace{E\{x^2(t)\}E\{x(t+k)\}}^{0} = C_{3x}(0,k) \tag{4.5}$$

por lo tanto, su espectro cruzado viene dado por:

$$S_{\tilde{y}x}(\omega) = \sum_{-\infty}^{\infty} C_{3x}(0,k)\exp\{-j\omega k\} \tag{4.6}$$

y

$$C_{3x}(0,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\tilde{y}x}(\omega)\exp\{j(\omega_k)\}d\omega \tag{4.7}$$

si comparamos la Eq. 4.4 con la Eq. 4.7, finalmente obtenemos:

$$S_{\tilde{y}x}(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega,\omega_2)d\omega_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_x(\omega_1,\omega)d\omega_1 \tag{4.8}$$

Es fácil demostrar que el biespectro de un proceso gaussiano es nulo, por lo tanto su espectro integrado lo será también. Como conclusión, el biespectro integrado es equivalente al espectro cruzado entre la señal y su cuadrado, y de este modo, queda representado como una función de una única variable de

frecuencia. De manera evidente, su cálculo como espectro cruzado reportará beneficios considerables en cuanto a coste computacional. Mas aún, la varianza de su estimador será del mismo orden que la del estimador de potencia espectral.

4.2. Detección de Actividad de Voz basada en el Biespectro Integrado

En esta sección abordamos el problema de detección de actividad de voz formulándolo en términos de un marco de test clásico de hipótesis binarias:

$$\begin{array}{lll} H_0 & : & x(t) = n(t) \\ H_1 & : & x(t) = s(t) + n(t) \end{array} \tag{4.9}$$

En un test de hipótesis binarias, la regla de decisión que minimiza la probabilidad de error es el clasificador de Bayes. Dado un vector de observación $\hat{\mathbf{y}}$ a clasificar, el problema se reduce a seleccionar la clase (H_0 ó H_1) con la mayor probabilidad condicionada $P(H_i|\hat{\mathbf{y}})$. Usando la regla de Bayes, un test estadístico LRT [Sohn et al., 1999] se define como:

$$L(\hat{\mathbf{y}}) = \frac{p_{\mathbf{y}|H_1}(\hat{\mathbf{y}}|H_1)}{p_{\mathbf{y}|H_0}(\hat{\mathbf{y}}|H_0)} \tag{4.10}$$

y el vector de observación $\hat{\mathbf{y}}$ es clasificado como H_1 si $L(\hat{\mathbf{y}})$ es mayor que $P(H_0)/P(H_1)$, en otro caso se clasifica como H_0. En [Ramírez et al., 2005b] el LRT (que fue primeramente propuesto para VAD en [Sohn et al., 1999], el cual se aplicaba a la potencia espectral), se generaliza y aplica a sucesivas observaciones de la señal ruidosa $\hat{\mathbf{y}}_1$,$\hat{\mathbf{y}}_2$, ..., $\hat{\mathbf{y}}_m$. El así llamado LRT de múltiple observación(MO-LRT) produce mejoras significativas en la robustez del detector cuando el número de observaciones aumenta. Este test implica la

evaluación de la distribuciones condicionales conjuntas de las observaciones bajo H_0 y H_1

$$L_m(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_m) = \frac{p_{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_m|H_1}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_m|H_1)}{p_{\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_m|H_0}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_m|H_0)} \tag{4.11}$$

, expresión fácilmente evaluable si se asume que las observaciones son estadísticamente independientes.

Tomando el biespectro integrado $\{S_{yx}(\omega) : \omega\}$ como vectores rasgo $\hat{\mathbf{y}}$ y asumiendo que son variables gaussianas independientes centradas en presencia y ausencia de voz:

$$\begin{aligned} p(S_{yx}(\omega)|H_0) &= \tfrac{1}{\pi\lambda_0(\omega)} \exp\left[-\tfrac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)}\right] \\ p(S_{yx}(\omega)|H_1) &= \tfrac{1}{\pi\lambda_1(\omega)} \exp\left[-\tfrac{|S_{yx}(\omega)|^2}{\lambda_1(\omega)}\right] \end{aligned} \tag{4.12}$$

la evaluación de los tests de decisión dados en 4.10 y 4.11 solo requieren estimar el biespectro integrado de la señal ruidosa y su varianza. Así, tomando logaritmos en 4.10 y sustituyendo el modelo definido en 4.12 obtenemos:

$$\Phi(\hat{\mathbf{y}}) = \sum_{\omega} \log\left(\tfrac{p(S_{yx}(\omega)|H_1)}{p(S_{yx}(\omega)|H_0)}\right) = \sum_{\omega} \left\{\left(1 - \tfrac{\lambda_0(\omega)}{\lambda_1(\omega)}\right) \tfrac{|S_{yx}(\omega)|^2}{\lambda_0(\omega)} - \log\left(\tfrac{\lambda_1(\omega)}{\lambda_0(\omega)}\right)\right\} \tag{4.13}$$

Finalmente, si definimos las razones de varianza *a priori* y *a posteriori* como:

$$\xi(\omega) = \frac{\lambda_1(\omega)}{\lambda_0(\omega)} - 1 \qquad \gamma(\omega) = \frac{|S_{yx}(\omega)|^2}{\lambda_0} \tag{4.14}$$

la ecuación 4.13 puede expresarse de una forma más compacta:

$$\begin{aligned} \Phi(\hat{\mathbf{y}}) &= \sum_{\omega} \left[\left(1 - \tfrac{1}{1+\xi(\omega)}\right)\gamma(\omega) - \log\left(1 + \xi(\omega)\right)\right] = \\ &= \sum_{\omega} \left[\tfrac{\xi(\omega)\gamma(\omega)}{1+\xi(\omega)} - \log\left(1 + \xi(\omega)\right)\right] \end{aligned} \tag{4.15}$$

La próxima sección aborda los dos temas principales que son necesarios para evaluar el LRT propuesto. Primero, la estimación del biespectro integrado sobre un conjunto de datos finito, y segundo, el cálculo de las varianzas del biespectro integrado $\lambda_0(\omega)$ y $\lambda_1(\omega)$ bajo las hipótesis H_0 y H_1.

4.2.1. Estimación del biespectro integrado

Denotemos por $\hat{S}_{yx}(\omega)$ al estimador consistente de $S_{yx}(\omega)$, donde $y(t) = x^2(t) - E\{x^2(t)\}$. Dado un conjunto finito de datos $x(1), x(2), ..., x(N)$, el biespectro integrado normalmente se estima dividiendo la secuencia de muestras en segmentos o bloques [Brillinger and Rosenblatt, 1968]. Esto es, el conjunto de datos es dividido en K_B segmentos no solapados cada uno de con N_B muestras tal que $N = K_B N_B$. De este modo, el periodograma cruzado del bloque de datos i-ésimo viene dado por

$$\hat{S}_{yx}^{(i)}(\omega) = \frac{1}{N_B} X^{(i)}(\omega) \left[Y^{(i)}(\omega)\right]^* \tag{4.16}$$

donde $X^{(i)}(\omega)$ e $Y^{(i)}(\omega)$ denotan las transformadas discretas de Fourier para el bloque ith. Finalmente, la estimación se obtiene promediando sobre los K_B bloques:

$$\hat{S}_{yx}(\omega) = \frac{1}{K_B} \sum_{i=1}^{K_B} \hat{S}_{yx}^{(i)}(\omega) \tag{4.17}$$

4.2.2. Varianza del biespectro integrado

Las propiedades de este estimador han sido discutidas en profundidad en [Brillinger and Rosenblatt, 1968; Brillinger, 1975]. El test propuesto en las secciones anteriores y el modelo asumido en la ecuación 4.12 están plenamente justificados ya que para valores elevados de N_B, la estimación $S_{yx}^{(i)}(\omega_m)$ es compleja, gaussiana e independiente de $S_{yx}^{(i)}(\omega_n)$ para $m \neq n$ $(m, n =$

$1, 2, ..., N_B/2-1$). Mas aún, su media y varianza para grandes valores de N_B y K_B puede aproximarse [Tugnait, 1994] por:

$$\begin{array}{l} E\left\{\hat{S}_{yx}(\omega)\right\} \approx S_{yx}(\omega) \\ \text{var}\left\{\Re\left[\hat{S}^{(i)}_{yx}(\omega)\right]\right\} \approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) + \Re\left\{S^2_{yx}(\omega)\right\}\right] \\ \text{var}\left\{\Im\left[\hat{S}^{(i)}_{yx}(\omega)\right]\right\} \approx \frac{1}{2K_B}\left[S_{yy}(\omega)S_{xx}(\omega) - \Re\left\{S^2_{yx}(\omega)\right\}\right] \end{array} \tag{4.18}$$

La estimación de $\lambda_0(\omega)$ y $\lambda_1(\omega)$ requiere el cálculo de $S_{xx}(\omega)$ y $S_{yy}(\omega)$ bajo H_0 y H_1:

- Bajo la hipótesis H_0, $x(t) = n(t)$ y $y(t) = x^2(t) - E\{x^2(t)\}$ y por tanto,

$$\begin{array}{l} S_{xx}(\omega) = S_{nn}(\omega) \\ S_{yy}(\omega) = S_{n^2n^2}(\omega) \end{array} \tag{4.19}$$

  La segunda expresión de la ecuación anterior puede escribirse en términos del espectro del ruido y el cuadrado de su varianza como (véase Apéndice B):

$$S_{n^2n^2}(\omega) = 2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4 \tag{4.20}$$

  y, por lo tanto, la varianza del espectro integrado λ_0, puede estimarse evaluando $S_{nn}(\omega)$ y la varianza del ruido:

$$\lambda_0(\omega) = \frac{1}{K_B}\left[2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega)\right] S_{nn}(\omega) \tag{4.21}$$

- Por otro lado, bajo la hipótesis H_1, $x(t) = s(t) + n(t)$, el cálculo de $S_{xx}(\omega)$ y $S_{yy}(\omega)$ requiere cierto desarrollo matemático (véase Apéndice B), tras el cual podemos llegar a la conclusión de que:

$$\begin{array}{l} S_{xx}(\omega) = S_{ss}(\omega) + S_{nn}(\omega) \\ S_{yy}(\omega) = S_{s^2s^2}(\omega) + S_{n^2n^2}(\omega) + 4S_{ss}(w) * S_{nn}(\omega) - 2\pi(\sigma_s^4 + \sigma_n^4)\delta(\omega) \end{array} \tag{4.22}$$

Usando la ecuación 4.20 para $s(t)$ y sustituyéndola en la ecuación 4.22:

$$S_{yy}(\omega) = 2\left[S_{ss}(\omega) * S_{ss}(\omega) + S_{nn}(\omega) * S_{nn}(\omega) + 2\left[S_{ss}(\omega) * S_{nn}(\omega)\right]\right] \tag{4.23}$$

La varianza del bispectro integrado bajo H_1 $\lambda_1(\omega)$, puede estimarse en términos de $S_{ss}(\omega)$ y $S_{nn}(\omega)$ por medio de:

$$\lambda_1(\omega) = \frac{1}{K_B}\left[2S_{ss}(\omega) * S_{ss}(\omega) + 2S_{nn}(\omega) * S_{nn}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega)\right] \left[S_{ss}(\omega) + S_{nn}(\omega)\right] \tag{4.24}$$

Finalmente, tenemos que buscar una manera de estimar el espectro de potencia de la señal limpia y del ruido, $S_{ss}(\omega)$ y $S_{nn}(\omega)$. En este Capítulo, usamos un método basado en filtrado de Wiener y sustracción espectral para la estimación de $S_{ss}(\omega)$ en términos del espectro de potencia de la señal ruidosa $S_{xx}(\omega)$. Durante un periodo corto de inicialización, el espectro de potencia del ruido residual $S_{nn}(\omega)$ es estimado asumiendo un periodo de silencio (ruidoso) en el comienzo de la frase. Note que, $S_{nn}(\omega)$ puede calcularse en términos de la DFT de la señal ruidosa $x(t) = n(t)$. Después del periodo de inicialización, el espectro de potencia de la señal ruidosa $S_{xx}(\omega)$ se calcula para cada segmento o "frame" a través de las ecuaciones 4.16 y 4.17, y $S_{ss}(\omega)$ es obtenido aplicando un bloque de filtrado. El filtrado consiste en la sustracción del anterior espectro suavizado seguido por un filtro de Wiener. Figure 4.1 muestra un diagrama de bloques para la estimación del espectro de potencia de la señal sin ruido $S_{ss}(\omega)$ usando el de la señal ruidosa $S_{xx}(\omega)$. Es necesario clarificar que $S_{nn}(\omega)$ no solo es estimada durante el periodo de inicialización sino que se actualiza durante los segmentos clasificados como silencio basándonos en la decisión del VAD. Resumiendo, el proceso de filtrado se puede resumir en dos etapas:

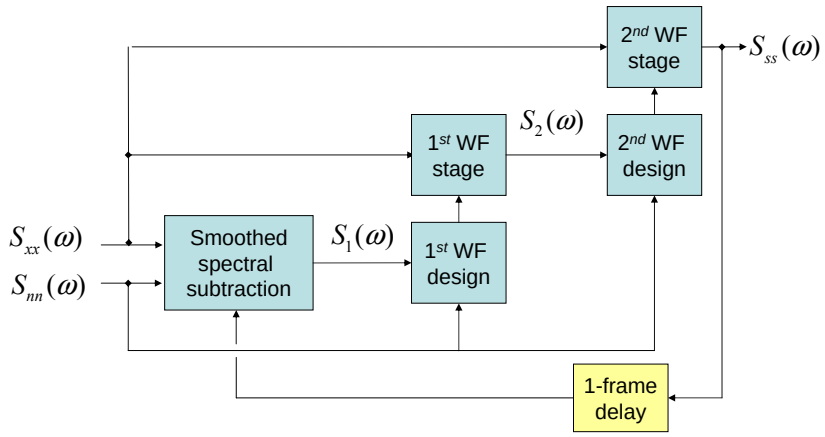


Fig. 4.1: Estimación de $S_{ss}(\omega)$ via sustracción espectral suave y filtrado de Wiener.

1. Sustracción Espectral.

$$S_1(\omega) = L_s S_{ss}(\omega) + (1 - L_s)\,\text{máx}(S_{xx}(\omega) - \alpha S_{nn}(\omega), \beta S_{xx}(\omega)) \quad (4.25)$$

2. Diseño y filtrado de la primera etapa de WF.

$$\begin{aligned} \mu_1(\omega) &= S_1(\omega)/S_{nn}(\omega) \\ W_1(\omega) &= \mu_1(\omega)/(1 + \mu_1(\omega)) \\ S_2(\omega) &= W_1(\omega) S_{xx}(\omega) \end{aligned} \quad (4.26)$$

3. Diseño y filtrado de la segunda etapa de WF.

$$\begin{aligned} \mu_2(\omega) &= S_2(\omega)/S_{nn}(\omega) \\ W_2(\omega) &= \text{máx}(\mu_2(\omega)/(1 + \mu_2(\omega)), \beta) \\ S_{ss}(\omega) &= W_2(\omega) S_{xx}(\omega) \end{aligned} \quad (4.27)$$

donde $L_s = 0{,}99$, $\alpha = 1$ y $\beta = 10^{(-22/10)}$ es seleccionada para asegurar una atenuación máxima de -22dB para el filtro para reducir el ruido de alta varia

que normalmente aparece debido a rápidos cambios a través de celdas de frecuencias adyacentes. El uso de dos etapas de diseño de WF y filtrado es necesario para reducir el ruido musical que obtendríamos con el uso de una única etapa, siendo otra alternativa el uso de una etapa de diseño WF y filtrado y un posterior truncamiento del filtro para suavizar su respuesta (ej. con una ventana de Hanning).

La próxima sección muestra dos aproximaciones diferentes para la estimación del biespectro integrado de la señal de entrada y su varianza así como la formulación del LRT.

4.2.3. Partición en bloques y promediado, VAD BA-IBI

Describimos el primer método de VAD en la presente sección. La señal de entrada $x(n)$ muestreada a 8 kHz se divide en ventanas solapadas de tamaño $N = K_B N_B$ muestras. Un valor típico del tamaño de ventana es 0.2 segundos, idóneo para la obtención de estimaciones precisas del biespectro integrado. El mejor compromiso entre promediado bloques (K_B) y resolución espectral (N_B) será discutido en las siguientes secciones.

La figura 4.2 ilustra la forma en que la señal es procesada y el bloque de datos para el que se está realizando la decisión. Note que, la decisión se realiza para un bloque de datos de T muestras alrededor del punto medio de la ventana de análisis donde T es el desplazamiento de la ventana. Por lo tanto, se usa un conjunto de datos considerable para la estimación del biespectro integrado promediando K_B bloques de datos sucesivos mientras que la decisión se toma para un conjunto menor de datos. Como en la mayoría de los VADs estandarizados [ITU, 1996; ETSI, 1999; ETSI, 2002] el desplazamiento es de 80 muestras por lo que la tasa de frames del VAD es de 100 Hz.

Después de haber estimado el biespectro integrado de la señal limpia

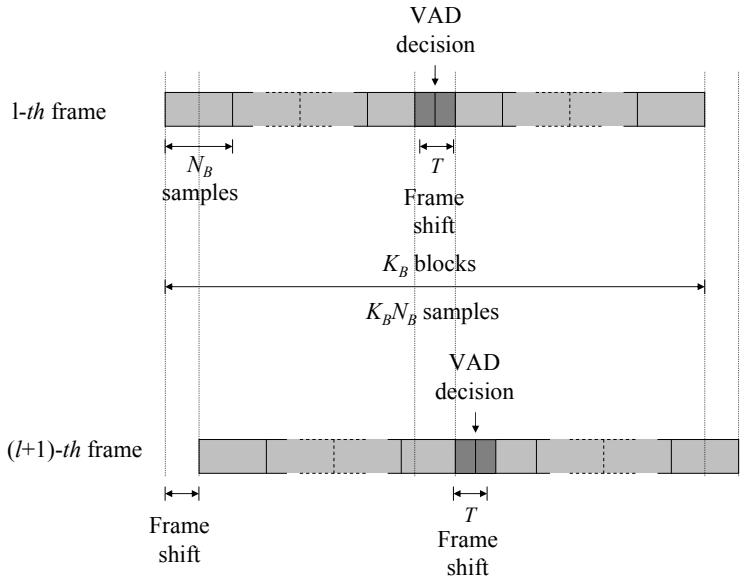


Fig. 4.2: Estimación del biespectro integrado mediante promediado de bloques y decisión del VAD.

$S_{ss}(\omega)$, se han de calcular $\lambda_0(\omega)$ y $\lambda_1(\omega)$ evaluando los operadores de convolución requeridos por las ecuaciones 4.21 y 4.24. Después, los ratios de varianza *a priori* y *a posteriori* definidos en la ecuación 4.14 pueden estimarse y la regla de decisión del VAD se formula comparando el LRT definido en la ecuación 4.15 con un umbral dado η. Si el LRT es mayor que el umbral η, el segmento es clasificado como señal de voz, en otro caso se clasifica como silencio (ruido).

Una vez que se obtiene la decisión del VAD para la ventana que esta siendo procesada, la estimación del espectro de potencia del ruido se actualiza durante los periodos de silencio con el objetivo de capturar la posible no estacionariedad a largo término (suave) de los ambientes ruidosos:

$$S_{nn}(\omega) = L_n S_{nn}(\omega) + (1 - L_n) S_{xx}(\omega) \tag{4.28}$$

donde L_n= 0.98.

4.2.4. Test contextual de cociente de probabilidades, VAD IBI-MO-LRT

La mayoría de los VADs que se usan hoy en día normalmente emplean algoritmos que implementan periodos de guarda "hang-over" basados en modelos empíricos para suavizar la regla de decisión del VAD. Se ha demostrado recientemente [Górriz et al., 2005e; Ramírez et al., 2005b] que, incorporando información de retardo largo a la regla de decisión, se obtienen mejores resultados en la discriminación voz/silencio en ambientes acústicos adversos, por lo tanto se hace innecesario el uso de mecanismos como el de "hang-over" que están basados en un ajuste manual de parámetros y reglas. El VAD previamente propuesto afronta este problema formulando una decisión suave basada en un conjunto de datos extenso. Sin embargo, podemos definir alternativamente un test estadístico óptimo que implica observaciones inde-

pendientes y múltiples de la señal de entrada [Ramírez et al., 2005b] sobre el biespectro integrado de la señal ruidosa usando la ecuación 4.11.

El VAD basado en un test de múltiple observación (MO-LRT) se describe como sigue. Los vectores de observación $\hat{\mathbf{y}}_{l-m}, ..., \hat{\mathbf{y}}_{l-1}, \hat{\mathbf{y}}_l, \hat{\mathbf{y}}_{l+1}, ..., \hat{\mathbf{y}}_{l+m}$ se usan para formular el test definido por la ecuación 4.11 de tal forma que:

$$L_{l,m}(\hat{\mathbf{y}}_{l-m}, ..., \hat{\mathbf{y}}_{l+m}) = \frac{p_{\mathbf{y}_{l-m},...,\mathbf{y}_{l+m}|H_1}(\hat{\mathbf{y}}_{l-m}, ..., \hat{\mathbf{y}}_{l+m}|H_1)}{p_{\mathbf{y}_{l-m},...,\mathbf{y}_{l+m}|H_0}(\hat{\mathbf{y}}_{l-m}, ..., \hat{\mathbf{y}}_{l+m}|H_0)} \tag{4.29}$$

donde l denota el frame que esta siendo clasificado como voz (H_1) ó ruido (H_0). Note que, asumiendo independencia estadística entre los sucesivos vectores de observación, el test logarítmico log-LRT que resulta:

$$\ell_{l,m} = \sum_{k=l-m}^{l+m} \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \tag{4.30}$$

es de manera natural recursivo, y si la función Φ se define como:

$$\Phi(k) = \ln \frac{p_{\mathbf{y}_k|H_1}(\hat{\mathbf{y}}_k|H_1)}{p_{\mathbf{y}_k|H_0}(\hat{\mathbf{y}}_k|H_0)} \tag{4.31}$$

la anterior ecuación puede escribirse como:

$$\ell_{l+1,m} = \ell_{l,m} - \Phi(l-m) + \Phi(l+m+1) \tag{4.32}$$

Ahora, si consideramos el biespectro integrado de la señal ruidosa como un vector rasgo, ecuación 4.31 se escribe como:

$$\Phi(k) = \sum_{\omega} \left[\frac{\xi_k(\omega)\gamma_k(\omega)}{1+\xi_k(\omega)} - \log\left(1 + \xi_k(\omega)\right) \right] \tag{4.33}$$

donde los ratios de varianza *a priori* y *a posteriori* se definen como en la ecuación 4.14 para el *k-ésimo* vector de observación.

Note que, la regla de decisión se formula sobre una ventana deslizante que consiste en $(2m+1)$ vectores de observación alrededor del frame para el

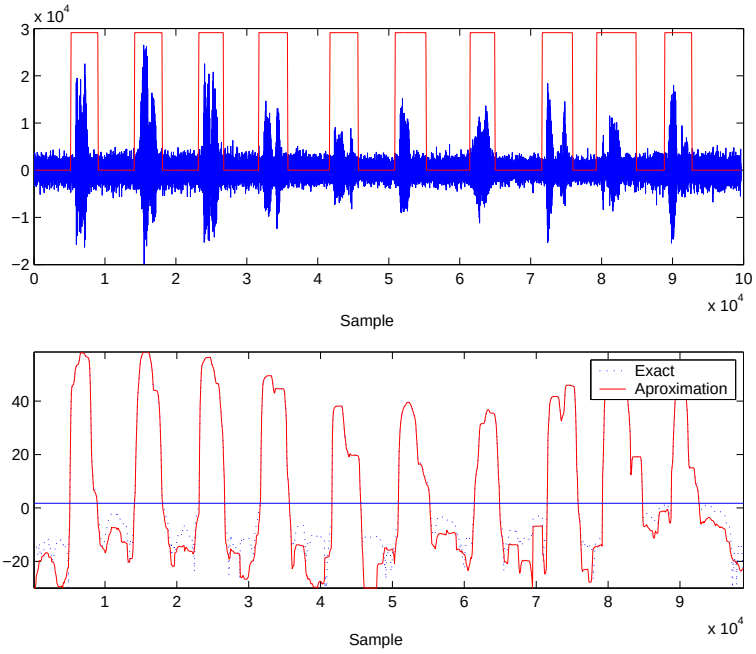


Fig. 4.3: Operación del VAD basado en MO-LRT sobre el biespectro integrado.

cual se realiza la decisión. Este hecho impone un retraso de m-ventanas al algoritmo lo que, para muchas aplicaciones, incluyendo reconocimiento de voz, no es un serio obstáculo para su implementación.

La figura 4.3 muestra un ejemplo de la operación del algoritmo MO-LRT VAD sobre una frase de la base de datos en español Spanish SpeechDat-Car [Moreno et al., 2000]. El uso de este test basado en el espectro integrado de una ventana deslizante de múltiples observaciones reporta beneficios significativos en la detección de voz/silencio. La figura muestra las variables de decisión para el test definidos en la ecuación 4.33 y, alternativamente, para el test con el segundo término de retardo largo eliminado en la ecuación 4.33 cuando usamos un umbral fijo $\eta = 1{,}5$. Note que, esta aproximación reduce la varianza en los periodos de silencio. Para este ejemplo, N_B= 256, y m= 8. Se demuestra que usando una ventana de 8-frames se reduce la variabilidad de la variable de decisión, y por lo tanto la varianza del ruido se mejora, aumentando la discriminación voz/silencio. Por otro lado, la anticipación inherente en la decisión del VAD contribuye a reducir el número de errores por recorte de voz.

4.2.5. Comparación

Es sin duda interesante comparar estos dos métodos propuestos para detección de actividad de voz basados en LRT sobre una única ventana y sobre múltiples observaciones. Ambos métodos exhiben grandes ventajas correspondientes a las propiedades adquiridas por los VADs que emplean información contextual para la formulación de la regla de decisión dado que la variable de decisión se construye sobre un conjunto amplio de datos. El primer método descompone la ventana en K_B bloques de tamaño N_B y el biespectro integrado se calcula promediando K_B bloques de datos para ca-

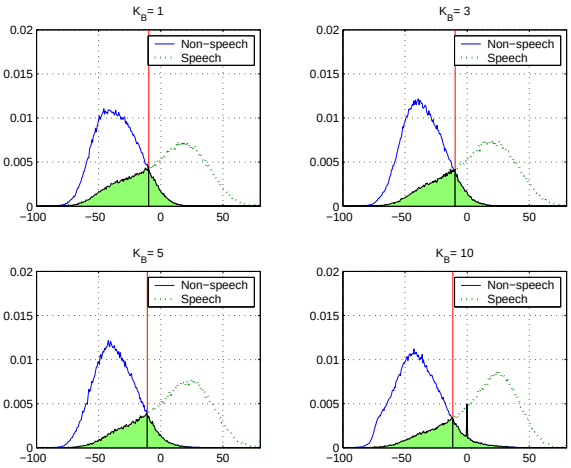
da desplazamiento. Esto puede ser computacionalmente costoso dado que el desplazamiento es usualmente menor que el tamaño de la ventana ($K_B N_B$). El segundo método basado en MO-LRT requiere el cómputo del biespectro integrado de un conjunto pequeño de datos (N_B muestras) después de cada desplazamiento y finalmente, el test se construye de manera dinámica por medio de la ecuación 4.32. Esto es claramente más eficiente en términos de coste computacional. Por otro lado, ambos métodos exhiben una alta precisión en la discriminación entre segmentos de voz/silencio en ambientes ruidosos para configuraciones de retraso equivalentes como se demostrará en las siguientes secciones experimentales.

4.3. Análisis de los métodos propuestos

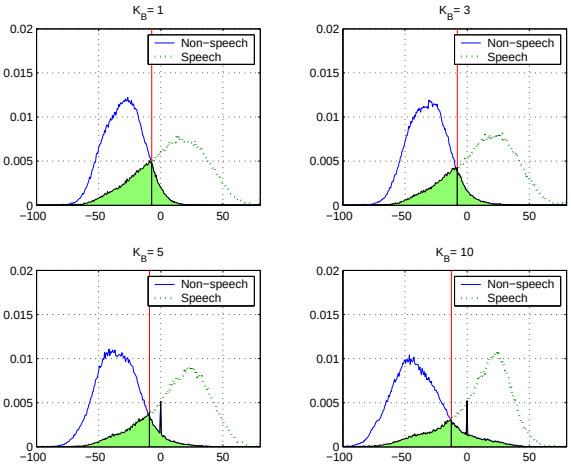
El solapamiento entre las distribuciones de probabilidad de la variable de decisión durante los periodos de voz y silencio representa esencialmente la tasa de error. Para aclarar las motivaciones de los algoritmos que hemos propuesto, las distribuciones de los LRTs definidos por las ecuaciones 4.15 y 4.30 fueron estudiadas como una función de los parámetros de diseño. Las condiciones de ruido más desfavorable (i.e.: alta velocidad, buena carretera y micrófono manos libres, [Moreno et al., 2000]) con una SNR promedio de 5 dB fueron elegidas para los experimentos. Por lo tanto, las variables de decisión definidas en las ecuaciones 4.15 y 4.30 fueron medidas en los periodos de voz y silencio para la base de datos completas, construyéndose de esta forma los histogramas y las distribuciones de probabilidad.

4.3.1. VAD basado en promediado en bloques

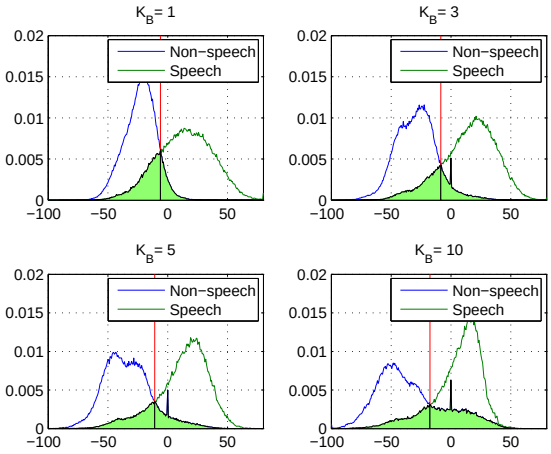
La figura 4.4 muestra las distribuciones de voz y silencio para distintos valores de K_B y N_B. Se ve claramente que las distribuciones de voz y silencio



(a)



(b)



(c)

Fig. 4.4: Distribuciones de la variable de decisión para el VAD basado en promedio por bloques (BA-IBI). (a) $N_B = 64$. (b) $N_B = 128$. (c) $N_B = 256$.

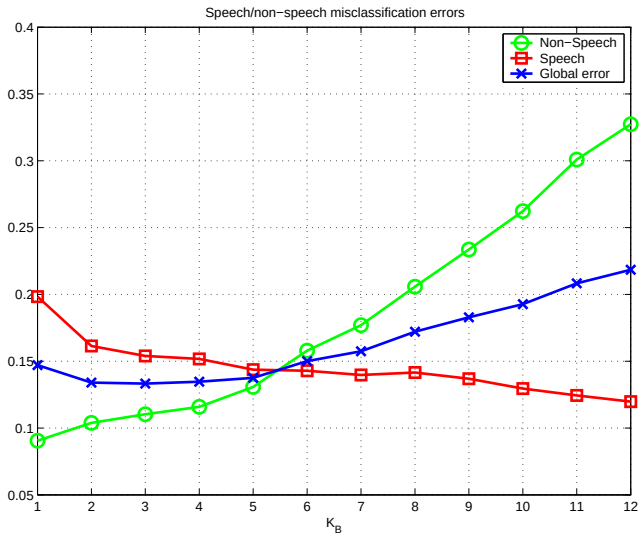


Fig. 4.5: Probabilidad de error como función de K_B para N_B= 256.

se separan mejor cuando se incrementa el número de bloques (K_B). Note como cuando K_B aumenta, la varianza del ruido disminuye y la distribución de voz se desplaza hacia la derecha separándose a su vez de la distribución de ruido. Es decir, las distribuciones de voz y ruido están menos solapadas y consecuentemente se reduce la probabilidad de error del detector.

La reducción del solapamiento provoca mejoras en la discriminación de voz/silencio. Este hecho, antes indicado, puede verse calculando los errores de clasificación de voz y silencio para un clasificador de Bayes óptimo. Note además que la figura 4.4 también muestra las áreas que representan las probabilidades de detectar incorrectamente voz y silencio y el umbral óptimo de decisión. La figura 4.5 muestra los errores de decisión independientes para voz y silencio y la tasa de error global como una función de K_B para $N_B = 256$. El error de detección de voz se reduce claramente cuando incrementamos la

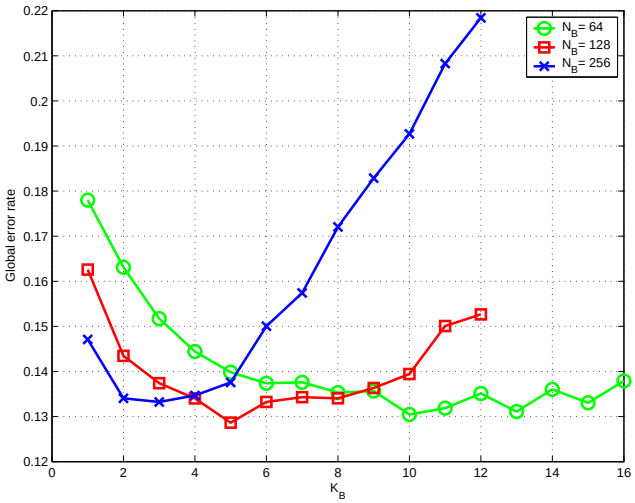


Fig. 4.6: Tasa de error global como una función de K_B para N_B= 64, 128 y 256.

longitud de la ventana ($K_B N_B$) mientras que la robustez sólo se ve afectada con un aumento moderado en el error de detección del silencio. Estas mejoras se alcanzan gracias a la disminución de la región de solapamiento de las distribuciones cuando K_B aumenta como muestra la figura 4.4. Es interesante en este punto notar que los valores óptimos de los parámetros K_B y N_B son elegidos para un tamaño de ventana fijo ($K_B N_B$). Este hecho se muestra en la figura 4.6 donde el mínimo valor de la tasa global de error depende tanto de N_B como de K_B y se obtiene para un tamaño de ventana típico de 80-100 ms.

4.3.2. *VAD MO-LRT usando el Biespectro Integrado*

Unos resultados muy similares se obtienen para el VAD MO-LRT basado en el biespectro integrado de la señal de ruido. La figura 4.7 muestra, las distribuciones de voz y silencio para distintos valores de K_B y $N_B = 256$. Note como las varianzas del ruido decrecen cuando aumentamos m y que la

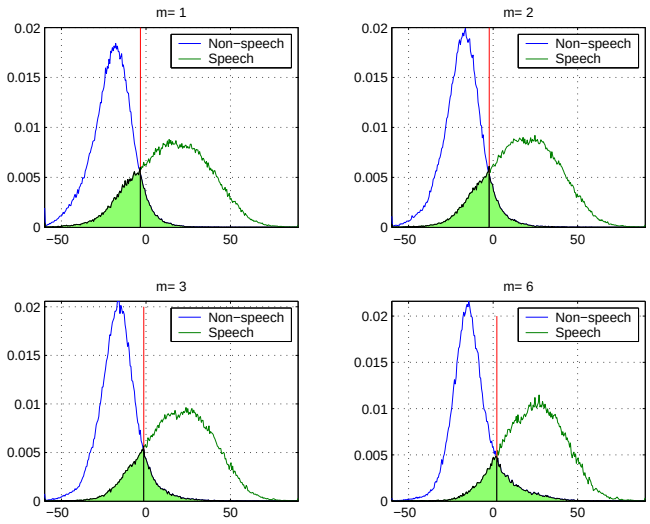


Fig. 4.7: Distribuciones de la variable de decisión para el VAD basado en MO-LRT.

distribución de voz se desplaza a la derecha. La figura 4.8 muestra los errores de decisión independientes para voz y silencio y la tasa de error global como una función de m para $N_B = 256$. El error total se reduce con el aumento de la longitud de ventana (m) y exhibe un valor mínimo fijado un orden. De acuerdo con la figura 4.8, el valor óptimo del orden del VAD es m= 6. Por lo tanto, incrementando la longitud de ventana se obtiene un beneficio evidente en ambientes ruidosos dado que el VAD introduce un periodo de guarda "hang-over" artificial el cual reduce los errores de recorte hacia delante y hacia atrás. Este periodo de guarda es la razón del incremento del error de detección de ruido mostrado en la figura 4.8.

4.4. *Marco Experimental*

Varios autores han realizado diferentes análisis y experimentos para evaluar el comportamiento de los algoritmos de detección de actividad de voz. Este análisis se suele centrar principalmente en la determinación de la proba-

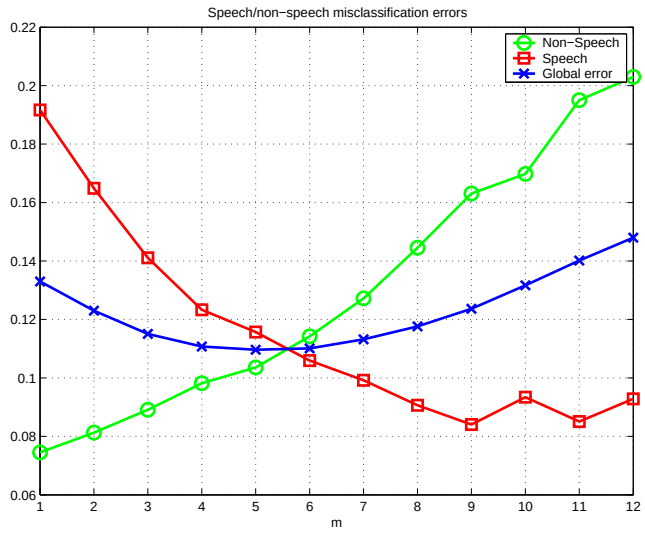


Fig. 4.8: Probabilidad de error como una función de m para N_B= 256.

bilidad de error o error de clasificación a diferentes niveles de SNR [Marzinzik and Kollmeier, 2002], y en la evaluación de la influencia del VAD en sistemas de procesamiento de la voz [Bouquin-Jeannes and Faucon, 1995b]. Del mismo modo, se han considerado tests subjetivos para la evaluación de VADs que operan en combinación con codificadores de voz [Benyassine et al., 1997]. En esta sección, se evalúa el VAD de acuerdo con el objetivo para el cual fue diseñado, es decir, mejorar el rendimiento de los sistemas de reconocimiento que operan en entornos ruidosos. En particular, se determina la influencia del VAD en el rendimiento de un sistema de reconocimiento que considera reducción de ruido espectral y *frame-dropping*.

En esta sección se describe el trabajo experimental y el rendimiento de los tests descritos en las secciones anteriores para evaluar los algoritmos propuestos.

4.4.1. Curvas ROC

Las curvas de discriminación anteriores muestran el comportamiento del detector una vez fijado el umbral de detección. Sin embargo, dependiendo de la aplicación, en ocasiones será necesario modificar el punto de operación del detector para hacerlo más efectivo en la detección de las pausas o en la detección de los periodos de voz. Por esta razón, se introducen las curvas ROC (Receiver Operating Characteristics) en la evaluación de los algoritmos de detección de actividad de voz. Estas curvas representan la variación de las tasas de error del detector cuando varía el umbral de detección y describen de forma más precisa la tasa de error del VAD.

En este análisis se utiliza la base de datos SpeechDat-Car española [Moreno et al., 2000]. Esta base de datos contiene 4914 grabaciones con micrófonos distante y de proximidad de más de 160 locutores. Las frases se encuentran agrupadas en tres subconjuntos que representan diferentes condiciones de conducción y SNR. El etiquetado semiautomático se realizó sobre las grabaciones realizadas con el micrófono de proximidad que se encuentran menos afectadas por el ruido.

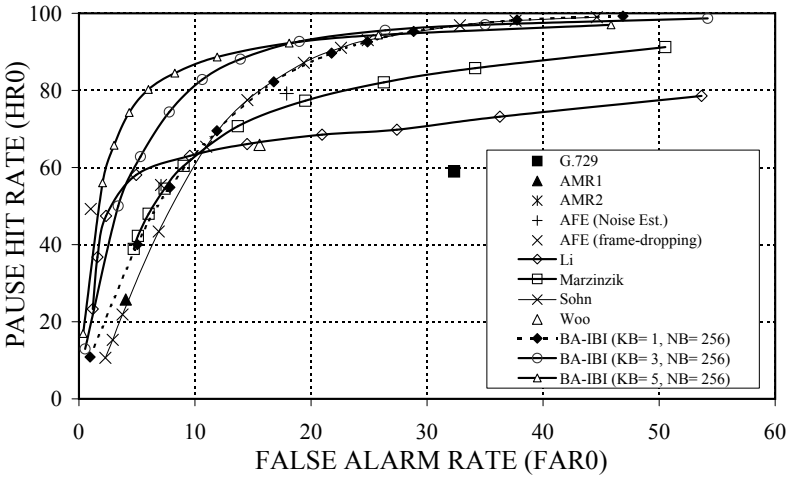
Para construir las curvas ROC se determinaron las tasas de acierto del silencio (HR0) y de la voz (HR1) en cada condición de ruido y se representaron los valores de (HR0, FAR0) estando la tasa de falsas alarmas definida como FAR0= 1-HR1. Para realizar este análisis se realizó el etiquetado semiautomático de la base de datos limpia como segmentos de voz y de silencio. El rendimiento del detector se evaluó en función de la SNR en términos de las tasas de acierto de los segmentos de silencio (HR0) y de voz (HR1) que se definen como la fracción de todas los segmentos de silencio o de voz que se detectan correctamente como tales, es decir:

$$HR0 = \frac{N_{0,0}}{N_0^{Ref}} \qquad HR1 = \frac{N_{1,1}}{N_1^{Ref}} \tag{4.34}$$

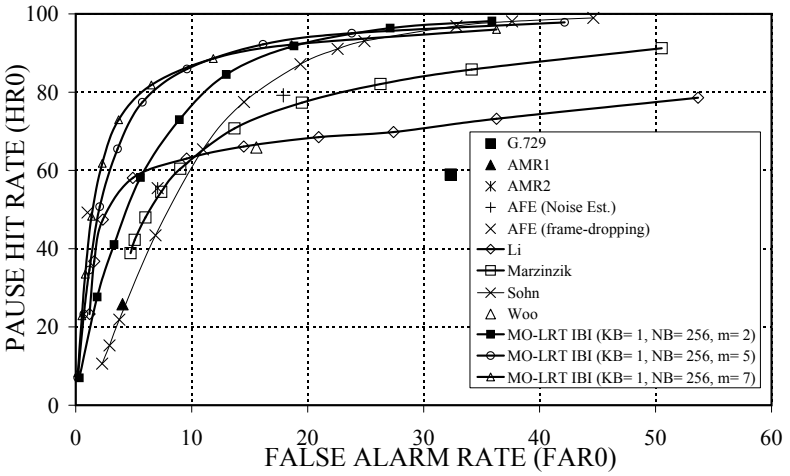
siendo N_0^{Ref} y N_1^{Ref} el número de segmentos de silencio y de voz reales de la base de datos etiquetada, respectivamente, mientras que $N_{0,0}$ y $N_{1,1}$ denotan los segmentos de silencio y de voz clasificados correctamente por el VAD.

La figura 4.9 muestra las curvas ROC de los VADs propuestos y otros algoritmos frecuentemente usados en detección [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999] para grabaciones de micrófono lejano en condiciones de alto ruido. Los puntos de operación de G.729, AMR y AFE VADs también se incluyen. La figura 4.9.a muestra como aumentando el número de bloques (K_B) en el VAD LRT de promedio de bloques biespectrales (BA-IBI) se obtiene un desplazamiento hacia arriba y hacia la izquierda de la curva ROC en el citado espacio. Este resultado es consistente con el análisis mostrado por la figuras 4.4 y 4.6 que predecían una tasa de error mínima para un valor de K_B alrededor de cinco bloques. Resultados similares se obtienen para el eficiente VAD MO-LRT IBI que muestra un desplazamiento de la curva ROC cuando el número de observaciones (m) aumenta como se muestra en la figura 4.9.b. Otra vez, los resultados son consistentes con nuestros experimentos preliminares y los resultados mostrados en las figuras 4.7 y 4.8 que hacen esperar una tasa de error mínima para m próximo a 8 segmentos. Ambos métodos muestran claras mejoras en precisión del detector sobre los VADs estandarizados y sobre un conjunto significativo de los algoritmos VAD más relevantes y recientemente publicados [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999].

En definitiva, de entre todos los VADs examinados, el nuestro aporta las tasas de alarma más bajas para una tasa de aciertos de silencio fija y también, la tasa de aciertos de silencio más elevada para una tasa de alarma



(a)



(b)

Fig. 4.9: Curvas ROC obtenidas para la condición más desfavorable de ruido. (a) LRT VAD biespectral basado en bloques. (b) MO-LRT VAD sobre el biespectro integrado.

dada. La mejora es especialmente importante sobre G.729, el cual es usado en conjunción con un codificador de voz para transmisión discontinua, y sobre el algoritmo de [Li et al., 2002], que está basado en un filtro lineal óptimo para la detección de contornos. El VAD propuesto también mejora el VAD de [Marzinzik and Kollmeier, 2002] que captura la envolvente de potencia espectral, y el VAD de [Sohn et al., 1999], que formula la regla de decisión por medio de un LRT definido sobre el espectro de potencia de la señal de ruido.

Debe destacarse que los experimentos descritos anteriormente aportan una primera medida del rendimiento potencial del VAD. como describimos en anteriores secciones, otra medida de este rendimiento pueden ser el error de recorte [Benyassine et al., 1997]. Estas medidas suministran información relevante acerca del rendimiento del VAD y pueden usarse para la optimización de su operación. Nuestro análisis no distingue entre los segmentos que se clasifican y obtiene las tasa de acierto y de falsa alarma para una primera evaluación del rendimiento del VAD propuesto. Por otro lado, los experimentos de reconocimiento que serán presentados posteriormente usando las bases de datos AURORA serán una medida directa de la calidad del VAD en la aplicación para la que fue diseñado. Los errores de recorte son evaluados indirectamente por el sistema de reconocimiento pues hay una alta probabilidad que ocurra un error por borrado cuando se pierde un periodo de voz después de aplicar frame-dropping.

4.4.2. Experimentos de reconocimiento de voz

Esta sección evalúa los VADs propuestos de acuerdo con el objetivo para el cual fueron diseñados observando la influencia de estos VADs sobre un sistema general de reconocimiento de voz.

El rendimiento de los sistemas ASR que operan sobre redes inalámbricas y entornos ruidosos disminuye rápidamente, siendo una fuente importante de degradación la ineficiente detección de los segmentos de voz y de silencio en una señal ruidosa [Karray and Martin, 2003]. Aunque el análisis de discriminación realizado anteriormente y las curvas ROC describen el comportamiento del VAD, esta sección evalúa su efectividad en el contexto para el que fue diseñado determinando la influencia del VAD en el rendimiento de un sistema de reconocimiento robusto de la voz. Existen dos claras razones para ello. En primer lugar, los algoritmos de supresión de ruido necesitan estimar durante los periodos de silencio los parámetros estadísticos del ruido para tratar de compensarlo y, por tanto, la efectividad del algoritmo depende de la buena estimación del ruido. Por otro lado, la eliminación de las tramas de silencio de la entrada del reconocedor ("*frame-dropping*") reduce el número de inserciones pero puede causar también la pérdida de segmentos de voz clasificados incorrectamente.

El entorno experimental de referencia se enmarca dentro de la iniciativa de ETSI para el desarrollo de un estándar de extracción de características robustas para sistemas de reconocimiento distribuidos. Como referencia (Base) se utiliza la primera propuesta de trabajo del grupo de estandarización AURORA [ETSI, 2000], mientras que el reconocedor se basa en el paquete HTK (Hidden Markov Model Toolkit) [Young et al., 1997]. La tarea consiste en el reconocimiento de digitos conectados que se modelan como modelos ocultos de Markov (HMMs: Hidden Markov Models) de palabra completa con los siguientes parámetros:

1. 16 estados por palabra,

2. modelos simples de izquierda a derecha,

3. mezclas de 3 Gaussianas por estado, y

4. matrices de covarianza diagonal de todos los coeficientes acústicos.

Los modelos de silencio utilizados son de tres estados con una mezcla de 6 Gaussianas por estado. El vector de características para entrenamiento y test del sistema de reconocimiento consta de 39 componentes incluyendo 12 coeficientes cepstrales más la energía logarítmica y las correspondientes derivadas y segundas derivadas. Para la base de datos AURORA-2 se definen dos tipos experimentos: *i*) Entrenamiento con voz limpia (CT: clean training), y *ii*) entrenamiento con voz limpia y voz ruidosa (MCT: multicondition training). Para las bases de datos AURORA-3 (SpeechDat-Car) se definen tres tipos de experimentos con diferente desajuste entre las condiciones de entrenamiento y test del sistema: *i*) desajuste mínimo (WM: well-matched), *ii*) desajuste medio (MM: medium mismatch), y *iii*) alto desajuste (HM: high mismatch). Estas bases de datos contienen grabaciones realizadas con micrófonos de proximidad y distante. En condiciones WM se utilizan ambos micrófonos para entrenamiento y test. En condiciones MM se consideran grabaciones del micrófono distante para entrenamiento y test. En condiciones HM, el entrenamiento se realiza utilizando el micrófono de proximidad en todas las condiciones de conducción mientras que para el reconocimiento o test se emplea el micrófono distante con grabaciones en condiciones de bajo y alto ruido. Finalmente, el rendimiento del sistema se mide en términos de la precisión de palabra (WAcc: *Word accuracy*) definida como:

$$WAcc = \frac{H - I}{N} \times 100\,\% \tag{4.35}$$

siendo H el número de palabras reconocidas correctamente, I, el número de inserciones y N, el número total de palabras.

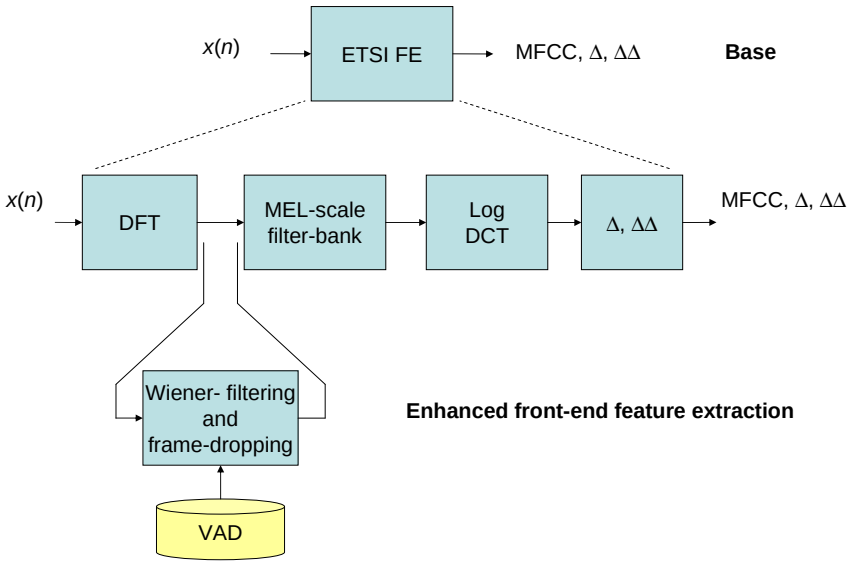


Fig. 4.10: Esquema general de los experimentos de Reconocimiento. Front-end para la extracción de características.

Para analizar con mayor nivel de detalle la influencia de los VADs en sistemas ASR, se ha incorporado un esquema mejorado de extracción de características al algoritmo utilizado como referencia [ETSI, 2000]. Este se basa en la incorporación de filtrado de Wiener (WF) como método de reducción de ruido y "*frame-dropping*" (FD) para reducir el número inserciones del sistema de reconocimiento. El algoritmo de reducción de ruido es similar al que se ha incluido en la propuesta final de estándar [ETSI, 2002] pero con una única etapa de filtrado y escala lineal de frecuencia. No se han considerado otras técnicas presentes en el AFE para reducción del desajuste entre las condiciones de entrenamiento y test del sistema porque no se encuentran afectadas por la decisión del VAD y pueden enmascarar el impacto de este en el rendimiento del sistema.

4.4.2.1. VAD BA-IBI

La tabla 4.1 muestra la precisión del sistema de reconocimiento cuando se emplea el VAD considerado en esta sección para la estimación de ruido en el bloque de filtrado de Wiener y, en su caso, para "*frame-dropping*". Estos resultados corresponden al promedio sobre los tres conjuntos de test de los experimentos de reconocimiento para la base de datos AURORA-2 [Hirsch and Pearce, 2000] y niveles de SNR entre 20 y -5 dB. Nótese que, para los experimentos de reconocimiento basados en los VADs del AFE, se ha utilizado la misma configuración del estándar [ETSI, 2002] que considera dos VADs diferentes, uno de ellos para supresión de ruido y el otro para "*frame-dropping*". Los resultados de la tabla demuestran como el VAD propuesto mejora el rendimiento del sistema de reconocimiento dando una tasa de reconocimiento de palabra más elevada que los estándares G.729 [ITU, 1996], AMR [ETSI, 1999] y AFE [ETSI, 2002] tanto en los experimentos con entrenamiento con voz limpia como con entrenamiento multi-condición. Cuando se compara con algoritmos propuestos recientemente [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999], el VAD propuesto alcanza mejores resultados siendo el que aporta una mayor mejora en WAcc con respecto al sistema de referencia (Base).

La tabla 4.2 muestra la tasa de reconocimiento alcanzada por los diferentes VADs para las bases de datos SpeechDat-Car finlandesa (Finnish) [Nokia, 2000], española (Spanish) [Moreno et al., 2000] y alemana (German) [Instruments, 2001] en las tres condiciones de entrenamiento y test descritas anteriormente. En estos experimentos se ha utilizado el VAD tanto para la estimación del espectro de ruido como para "*frame-dropping*". De nuevo, el VAD propuesto es el que obtiene una tasa de reconocimiento más elevada. Resulta interesante mencionar que las bases de datos SpeechDat-Car tienen,

Tab. 4.1: Precisión promedio de reconocimiento de palabra para la base de datos AURORA 2 y experimentos de entrenamiento con voz limpia y multi-condición. Los resultados son valores medios para todos los ruidos y SNRs entre 20 y -5 dB para WF y FD.

	G.729	AMR1	AMR2	AFE	**BA-IBI**
CT	57.08	65.01	78.48	79.02	**82.78**
MCT	83.55	83.56	87.29	87.55	**89.39**
	Woo	Li	Marzinzik	Sohn	**Base**
CT	76.64	78.63	82.15	79.49	60.49
MCT	85.54	85.58	88.32	88.11	86.47

Tab. 4.2: Precisión promedio de reconocimiento de palabra para las bases de datos SpeechDat-Car.

		Base	Woo	Li	Marz.	Sohn	G729	AMR1	AMR2	AFE	**BA-IBI**
Fin.	WM	92.74	86.81	85.60	93.73	93.84	88.62	94.57	95.52	94.25	95.06
	MM	80.51	66.62	55.63	76.47	80.10	67.99	81.60	79.55	82.42	84.18
	HM	40.53	62.54	58.34	68.37	75.34	65.80	77.14	80.21	56.89	83.28
	Prom.	**71.26**	71.99	66.52	79.52	83.09	74.14	84.44	85.09	77.85	**87.51**
Sp.	WM	92.94	95.35	91.82	94.29	96.07	88.62	94.65	95.67	95.28	96.73
	MM	83.31	89.30	77.45	89.81	91.64	72.84	80.59	90.91	90.23	91.55
	HM	51.55	83.64	78.52	79.43	84.03	65.50	62.41	85.77	77.53	86.76
	Prom.	**75.93**	89.43	82.60	87.84	90.58	75.65	74.33	90.78	87.68	**91.68**
Ger.	WM	91.20	91.59	89.62	91.58	93.23	87.20	90.36	92.79	93.03	94.32
	MM	81.04	80.28	70.87	83.67	83.97	68.52	78.48	83.87	85.43	89.76
	HM	73.17	78.68	78.55	81.27	82.19	72.48	66.23	81.77	83.16	86.39
	Prom.	**81.80**	83.52	79.68	85.51	86.46	76.07	78.36	86.14	87.21	**90.16**
Prom.		**76.33**	81.65	76.27	84.29	86.71	75.29	79.04	87.34	84.25	**89.93**

Tab. 4.3: Tasas de error de palabra para las bases de datos SpeechDat-Car. Resultados obtenidos para el AFE completo y el AFE modificado con el VAD propuesto.

	Finnish	Spanish	German	Danish	Promedio
AFE					
WM(x0.40)	3.96	3.39	4.87	6.02	4.56
MM(x0.35)	19.49	6.21	10.40	22.49	14.65
HM(x0.25)	14.77	9.23	8.70	20.39	13.27
Overall	12.10	5.84	7.76	15.38	**10.27**
AFE + BA-IBI					
WM(x0.40)	3.98	3.15	4.39	6.00	4.38
MM(x0.35)	16.39	6.54	10.37	19.75	13.26
HM(x0.25)	10.89	6.98	8.35	16.90	10.78
Overall	10.05	5.29	7.47	13.54	**9.09**

en general, periodos de silencio más largos entre palabras que la base de datos AURORA-2 y, por tanto, la precisión del VAD se convierte en un factor más importante para mejorar la robustez del sistema de reconocimiento. Este hecho se ve claramente cuando se comparan los resultados obtenidos con el VAD propuesto y el de Marzinzik. La tasa de reconocimiento de ambos es parecida para la base de datos AURORA-2; sin embargo, el algoritmo BA-IBI-LRT obtiene una mejora significativa en la tasa de reconocimiento sobre el VAD de Marzinzik en los experimentos de reconocimiento realizados sobre las bases de datos SpeechDat-Car.

Finalmente, para comparar el método propuesto con los mejores resultados disponibles, se han reemplazado los VADs del estándar AFE [ETSI, 2002] por el VAD propuesto y se han realizado los experimentos de reconocimiento sobre las bases de datos SpeechDat-Car. Los resultados obtenidos se muestran en la tabla 4.3 en términos de la tasa de error de palabra. Se observa una mejora significativa en el rendimiento del sistema de reconocimiento que se mantiene de forma consistente para todas las bases de datos y experimentos. Las mejoras fueron especialmente importantes cuanto mayor es el desajuste entre las condiciones de entrenamiento y test del sistema. Además, la tasa de error se reduce significativamente siendo la mejora relativa en la tasa de error del 18.76 % en HM. Resulta interesante destacar que la mejora obtenida en las prestaciones del sistema se consigue únicamente reemplazando los VADs del AFE por el VAD propuesto y sin introducir nuevos algoritmos de procesado de señal sobre el algoritmo base [ETSI, 2002].

Como conclusión, el rendimiento del VAD tiene un fuerte impacto en los sistemas ASR. Si los periodos de silencio entre palabras son largos y dominan sobre los periodos de voz, los errores de inserción serán una fuente importante de error. Por otro lado, si las pausas son cortas, un VAD conservativo con

una alta tasa de acierto de los segmentos de voz puede ser beneficioso para reducir los errores de borrado ya que los errores de inserción no serán especialmente significativos. El desajuste entre las condiciones de entrenamiento y test también determina la influencia del VAD en la tasa de reconocimiento de los sistemas ASR. Cuando el sistema sufre un alto desajuste entre las condiciones de entrenamiento y test, un VAD preciso tiene un comportamiento más efectivo. Este efecto se encuentra directamente motivado por la eficiencia del algoritmo de reducción de ruido y de la etapa de "*frame-dropping*"

4.4.2.2. VAD IBI-MO-LRT

La figura 4.10 muestra el diagrama de bloques de los experimentos de reconocimiento de voz llevados a cabo para evaluar el VAD propuesto. Como dijimos anteriormente, sobre el sistema base dado en [ETSI, 2000] se ha construido un esquema de extracción de características que incorpora un algoritmo de reducción de ruido y la técnica de "frame-dropping" (FD) para los segmentos de silencio. El algoritmo de reducción de ruido consiste en una etapa filtrado de Wiener (WF) como se describe en el estándar AFE [ETSI, 2002]. Ninguna otra técnica de reducción se ha usado como las presentadas en el estándar AFE dado que no se ven afectadas por el VAD y podrían enmascarar el efecto del VAD en el rendimiento general del sistema.

La tabla 4.4 muestra el rendimiento de reconocimiento alcanzado por distintos VADs que fueron comparados. Estos resultados fueron promediados sobre los tres conjuntos de tests (A, B y C) de los experimentos de reconocimiento de AURORA-2 [Hirsch and Pearce, 2000] y sobre distintas SNRs entre 20 y 0 dBs. Note que, para los experimentos de reconocimiento basados en los VADs AFE, se usó la misma configuración que el estándar [ETSI, 2002], que consiste en considerar distintos VADs para WF y FD.

Tab. 4.4: Tasa de reconocimiento de palabra promedio (%) para AURORA 2 en entrenamiento de condición limpia y múltiple. Los resultados están promediados para todos los ruidos y SNRs que van desde 20 a 0 dB.

	G.729	AMR1	AMR2	AFE	**MO-LRT**
WF	66.19	74.97	83.37	81.57	**84.15**
WF+FD	70.32	74.29	82.89	83.29	**85.71**
	Woo	Li	Marzinzik	Sohn	**Hand-labelled**
WF	83.64	77.43	84.02	83.89	84.69
WF+FD	81.09	82.11	85.23	83.80	86.86

El VAD propuesto "MO-LRT biespectro integrado" obtiene mejoras significativas sobre los estándares VAD G.729, AMR1, AMR2 y AFE en ambas condiciones, limpia y multi-condición de experimentos de entrenamiento/test. Cuando lo comparamos con los algoritmos VAD recientemente publicados, el propuesto proporciona mejores resultados siendo el que más próximo se sitúa del rendimiento ideal obtenido mediante etiquetado manual (aquel que utiliza el etiquetado semiautomático sobre la base de datos limpia para todas las condiciones de SNR).

La tabla 4.5 muestra los resultados de reconocimiento para la base de datos Spanish SpeechDat-Car cuando empleamos sobre el sistema base [ETSI, 2000] WF y FD. De nuevo, el VAD mejora todos los algoritmos usados como referencia, aportando sensibles mejoras en experimentos de reconocimiento de voz. Note que, esta base de datos usada en los experimentos AURORA 3 tiene periodos más largos de silencio que la base de datos AURORA 2,

Tab. 4.5: Tasa de reconocimiento de palabra promedio (%) para las bases de datos Spanish SDC.

	Base	Woo	Li	Marz.	Sohn	**MO-LRT**
WM	92.94	95.35	91.82	94.29	96.07	96.39
MM	83.31	89.30	77.45	89.81	91.64	91.75
HM	51.55	83.64	78.52	79.43	84.03	86.65
Avg.	**75.93**	89.43	82.60	87.84	90.58	**91.60**
	Base	G729	AMR1	AMR2	AFE	**MO-LRT**
WM	92.94	88.62	94.65	95.67	95.28	96.39
MM	83.31	72.84	80.59	90.91	90.23	91.75
HM	51.55	65.50	62.41	85.77	77.53	86.65
Avg.	**75.93**	75.65	74.33	90.78	87.68	**91.60**

y por lo tanto la efectividad del VAD es más relevante para un sistema de reconocimiento. Este hecho puede verse claramente cuando comparamos el rendimiento de nuestro VAD con el de [Marzinzik and Kollmeier, 2002]. La tasa de reconociendo de palabra es similar para ambos para la tarea AURORA 2. sin embargo, sobre la tarea AURORA 3 el propuesto aporta mejores resultados que este anterior [Marzinzik and Kollmeier, 2002].

4.5. Conclusiones

Este trabajo muestra dos esquemas diferentes para la mejora de la robustez de los detectores de actividad de voz y el rendimiento de los sistemas de reconocimiento de voz en entornos ruidos. Ambos métodos están basados en tests estadísticos de cociente de probabilidades definidos sobre el biespectro integrado de la señal que a su vez está definido como un espectro cruzado entre la señal y su cuadrado, mostrando la capacidad de los estadísticos de orden superior para la detección de señales en ruido con otras muchas ventajas adicionales: *i*) su cálculo como espectro cruzado conduce a un ahorro computacional significativo, y *ii*) la varianza del estimador es del mismo orden que la del estimador del espectro de potencia. Los métodos propuestos incorporan información contextual a la regla de decisión, una estrategia que ha proporcionado mejoras significativas en la precisión de los detectores de voz y en distintas aplicaciones en sistemas de reconocimiento. Se distinguen en la manera en que la ventana se construye o, dicho de otro modo, se procesa la señal para obtener estimadores precisos del biespectro integrado y su varianza (para la formulación de reglas de decisión suaves basadas en tests de cociente de probabilidad). La ventana óptima fue determinada analizando el solapamiento entre las distribuciones de la variable de decisión y la tasa de error de un clasificador de Bayes óptimo. El análisis experimental llevado

a cabo sobre la conocida base de datos AURORA ha proporcionado mejoras significativas sobre las técnicas estándares como ITU G.729, AMR1, AMR2 y ESTI AFE VADs, así como respecto a los mejores VADs recientemente publicados. El análisis valoró: *a*) la precisión en detección de los segmentos de voz/silencio mediante el uso de las curvas ROC obteniendo que nuestro VAD proporcionaba mayores tasas de acierto y menores falsas alarmas cuando lo comparábamos a todos los algoritmos de referencia, y *b*) la tasa de reconocimiento cuando se consideraba como parte de un sistema global de reconocimiento mostrando un mayor rendimiento en reconocimiento de voz.

5. CLUSTERING APLICADO AL MODELADO DEL SUBESPACIO DE RUIDO PARA VAD

Este Capítulo muestra un algoritmo eficiente de detección de voz para mejorar el rendimiento de los sistemas de reconocimiento del habla en entornos ruidosos. El método propuesto se basa en una aproximación "clustering" de decisión abrupta que emplea un conjunto de prototipos para caracterizar el ruido del canal. La presencia de voz en el canal se detecta mediante una regla de decisión formulada en términos de la distancia euclídea entre el periodo de análisis y un modelo baso en clusters para el ruido. El algoritmo además aprovecha la información contextual, una estrategia que consiste en considerar no sólo un único segmento sino una vecindad de datos para suavizar la función de decisión y mejorar así la robustez en detección de voz. El esquema que proponemos exhibe un coste computacional reducido haciéndolo adecuado para su uso en aplicaciones en tiempo real, por ejemplo en sistemas de reconocimiento automático de voz (ASR). Un análisis exhaustivo sobre las bases de datos AURORA 2 y AURORA 3 muestra el alto rendimiento del algoritmo cuando lo comparamos con los detectores estandarizados por ITU y ETSI. Los resultados muestran mejoras en la precisión de detección y en la

tasa de reconocimiento de palabra sobre los VADs ITU-T G.729, ETSI GSM AMR estandarizados para transmisión discontinua y ETSI AFE empleado para reconocimiento distribuido del habla (DSR) así como, sobre los VADs más representativos y publicados recientemente.

5.1. Introducción

La discriminación entre voz y silencio puede describirse como una problema de aprendizaje no supervisado. El análisis mediante clusters es una solución apropiada para este problema y se basa en la división del conjunto de datos en grupos, los cuales están relacionados "de cierta forma". A pesar de la simplicidad de los algoritmos de clustering, existe en la actualidad un interés creciente en el uso de estos métodos para reconocimiento de patrones [Anderberg et al., 1973], procesado de imagen [Jain and Flynn, 1996] y la recuperación de información [Rasmussen, 1992; Salton, 1991]. La técnica clustering tiene una rica historia en otras disciplinas [Jain and Dubes, 1988; Fisher, 1987] como máquinas de aprendizaje, biología, psiquiatria, psicología, arqueología, geología, geografía, y marketing. El análisis cluster, también llamado segmentación de datos tiene una amplia variedad de objetivos, todos ellos relacionados con el agrupamiento o segmentación de una colección de objetos en subgrupos ó "clusters" de tal forma que dentro de cada cluster, los elementos se encuentran más relacionados entre ellos que con otros objetos de distintos clusters. También se usa para formar estadísticos descriptivos para discernir si los datos agrupados en distintos subgrupos presentan propiedades esencialmente distintas.

En la Sección 5.2 introducimos la información necesaria acerca del análisis cluster. En la siguiente Sección 5.3 se muestra el proceso de extracción de características mientras que en la Sección 5.4, se describe en profundidad el algoritmo VAD basado en información de retardo largo y C-means (LTCM). En la sección 5.5 discutimos algunos detalles del algoritmo y acabamos el Capítulo, en la Sección 5.6, con una completa descripción experimental comparándolo con un conjunto representativo de métodos VAD valorando su rendimiento como parte de un sistema de reconocimiento.

5.2. Bases del Hard-Clustering

Los algoritmos clustering dividen los datos de entrada en un conjunto determinado de clusters, de tal forma que, los patrones o individuos en cada cluster deberían ser "similares" entre ellos al contrario que los patrones pertenecientes a otros clusters. Dado un conjunto de patrones de entrada $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots \mathbf{x}_N\}$, donde $\mathbf{x}_j = (x_{j1}, \ldots, x_{ji}, \ldots, x_{jK}) \in \mathbb{R}^K$ y cada medida x_{ji} se denomina característica. El Hard Clustering trata de encontrar una partición de tamaño C de $\mathbf{X}$, $P = \{P_1, \ldots, P_C\}$, $C \leq N$, tal que:

- $P_i \neq \emptyset$, $i = 1, \ldots, C$;
- $\cup_{i=1}^{C} P_i = \mathbf{X}$;
- $P_i \cap P_i' = \emptyset$; $i, i' = 1, \ldots, C$ and $i \neq i'$.

La medida de "similitud" se define en términos de una función de criterio. La suma del error cuadrático es uno de los criterios más usados y se define como:

$$J(\mathbf{\Gamma}, \mathbf{M}) = \sum_{i=1}^{C} \sum_{j=1}^{N} \gamma_{ij} ||\mathbf{x}_j - \mathbf{m}_i||^2 \tag{5.1}$$

donde $\mathbf{\Gamma} = \gamma_{ij}$ es una matriz de partición, $\gamma_{ij} = \begin{pmatrix} 1 & if \quad \mathbf{x}_j \in P_i \\ 0 & \text{otro caso} \end{pmatrix}$ con $\sum_{i=1}^{C} \gamma_{ij} = 1, \forall j$, $\mathbf{M} = [\mathbf{m}_1, \ldots, \mathbf{m}_C]$ es la matriz prototipo de cluster o centroide (medias) con $\mathbf{m}_i = 1/N_i \sum_{j=1}^{N} \gamma_{ij} \mathbf{x}_j$ la media muestral para el cluster $|i|$-ésimo y N_i es el número de objetos en el cluster $|i|$-ésimo. La partición óptima, que se obtiene minimizando el anterior criterio, se puede encontrar enumerando todas las posibilidades. Sin embargo, es desaconsejable por el coste computacional habiéndose desarrollado una gran cantidad de heurísticas para esta tarea de optimización.

Alg. 5.1: Pseudocódigo Hard C-medias.

1. Inicializar la partición C aleatoriamente ó basándose en un conocimiento apriori. Calcular la matriz de prototipos culster $\mathbf{M} = \mathbf{m}_1, \ldots, \mathbf{m}_C$
2. Asignar cada objeto de X al cluster más cercano P_i ($\mathbf{x}_j \in P_i$ si $\|\|x_j - m_{i'}\|\| < \|\|x_j - m_i\|\|$ para $j = 1, \ldots, N$, $i \neq i'$, y $i' = 1, \ldots, C$)
3. Recalcular la matriz de prototipos cluster usando la partición actual
4. Repetir pasos 2)-3) hasta que no haya cambio en cada cluster.

El Hard C-medias clustering es la heurística más conocida de las basadas en el cuadrado del error [MacQueen, 1967]. El número de centros cluster (prototipos) C es conocido a priori que son movidos por el algoritmo de C-medias iterativamente para minimizar la varianza total en los clusters. Dado un conjunto inicial de centros, el algoritmo Hard C-medias alterna dos pasos [Hastie et al., 2001]:

- para cada cluster identificamos el subconjunto de puntos de entrenamiento (su cluster) que están más cerca de él que cualquier otro centro;

- las medias de cada característica para los datos en cada cluster son calculadas, y este vector media se convierte en el nuevo centro para ese cluster.

En el algoritmo 5.1 mostramos una descripción más detallada del algoritmo Hard C-medias.

5.3. Extracción de características con información de contexto

Sea $x(n)$ una señal discreta en el tiempo. Denotamos como $y_{n'}$ un segmento que contiene las muestras:

$$y_{n'} = \{x_i^l\} = \{x(i + n' \cdot D)\}; \quad i = 0 \dots L-1 \tag{5.2}$$

donde D es el desplazamiento de la ventana, L es el número de muestras en cada segmento y n' selecciona una determinada ventana de datos. Considere el conjunto de $2 \cdot m+1$ segmentos $\{y_{l-m}, \dots y_l \dots, y_{l+m}\}$ centrados en el frame l, y sea $Y(s, n')$, $n' = l-m, \dots l \dots, l+m$ su transformada discreta de Fourier (DFT):

$$Y_{n'}(\omega_s) \equiv Y(s, n') = \sum_{i=0}^{N_{FFT}-1} x(i + n' \cdot D) \cdot exp\left(-\mathbf{j} \cdot n \cdot \omega_s\right). \tag{5.3}$$

donde $\omega_s = \frac{2\pi \cdot s}{N_{FFT}}$, $0 \leq s \leq N_{FFT} - 1$, N_{FFT} es la resolución de la DFT (si $N_{FFT} > L$ se rellena la DFT con ceros) y $\mathbf{j}$ denota la unidad imaginaria. Las energías promedio para cada frame n'-ésimo, $E(k, n')$, en K subbandas $(k = 1, 2, ..., K)$, son calculadas por medio de:

$$\begin{aligned} E(k, n') &= \left(\frac{K}{N_{FFT}} \sum_{s=s_k}^{s_{k+1}-1} |Y(s, n')|^2 \right) \\ s_k &= \left\lfloor \frac{N_{FFT}}{2K}(k-1) \right\rfloor \;\; k = 1, 2, ..., K \end{aligned} \tag{5.4}$$

donde se ha usado una asignación igualmente espaciada y $\lfloor \cdot \rfloor$ denota la función "floor" (redondeo a la baja). Por lo tanto, la señal energía es promediada sobre K subbandas obteniendo una representación adecuada de la señal de entrada para el VAD [Ramírez et al., 2005a]. El vector de observación en cada frame n' se define como:

$$\mathbf{E}(n') = (E(1, n'), \dots, E(K, n'))^T \in \mathbb{R}^K \tag{5.5}$$

La regla de decisión del VAD se formula sobre una ventana deslizante consiste en $2m+1$ vectores de observación (características) sobre el frame en el que se está tomando la decisión (l), como mostraremos en las siguientes secciones. Esta estrategia, conocida como información de retargo largo "long term information" [Górriz et al., 2005e], proporciona muy buenos resultados en distintas aproximaciones en VAD, sin embargo impone un retraso de m frames o segmentos al algoritmo que, para la mayoría de aplicaciones incluyendo reconocimiento robusto del habla, no es un serio obstáculo de implementación.

En la próxima sección mostramos como aplicamos el **C**-medias para modelar el subespacio de ruido y para encontrar una decisión suave para el detector.

5.4. Hard C Medias para VAD

En el algoritmo VAD LTCM, aplicamos el método de clustering, descrito en la sección 5.2, a un conjunto inicial de frames de silencio para caracterizar el espacio de ruido, es decir $\mathbf{x}_j \equiv \mathbf{E}(n')$. Llamamos este conjunto de clusters ó partición de tamaño C, *prototipos de ruido* dado que, en este trabajo, la palabra cluster se asigna a clases distintas de datos etiquetados, esto es $\mathbf{K}$ se fija a 2 (frames de ruido y de voz). Cada vector de observación ($\mathbf{E}(n')$ a partir de la ecuación 5.5) se etiqueta de manera única, con un entero $j \in \{1, \ldots, N\}$, y se asigna únicamente a un número de prototipos especificado $C < N$, etiquetados por un entero $i \in \{1, \ldots, C\}$. La medida de similitud en términos de vectores de energía que debe ser minimizada esta basada en el cuadrado de la distancia Euclídea:

$$d(\mathbf{E}_j, \mathbf{E}_{j'}) = \sum_{k=1}^{K} (E(k,j) - E(k,j'))^2 = ||\mathbf{E}_j - \mathbf{E}_{j'}||^2 \qquad (5.6)$$

y puede ser definida equivalentemente como:

$$\begin{aligned} J(C) &= \tfrac{1}{2}\sum_{i=1}^{C}\sum_{C(j)=i}\sum_{C(j')=i}||\mathbf{E}_j - \mathbf{E}_{j'}||^2 \\ &= \tfrac{1}{2}\sum_{i=1}^{C}\sum_{C(j)=i}||\mathbf{E}_j - \bar{\mathbf{E}}_i||^2 \end{aligned} \tag{5.7}$$

donde $C(j) = i$ denota un mapa de un conjunto de índices a uno, que asigna la observación j-ésima al i-ésimo prototipo y

$$\begin{aligned} &\bar{\mathbf{E}}_i = (\bar{E}(1,i),\ldots,\bar{E}(K,i))^T = mean(\mathbf{E}_j); \\ &\forall j, \quad C(j) = i, \quad i = 1,\ldots,C \end{aligned} \tag{5.8}$$

es el vector medio asociado con el i-ésimo prototipo ($\mathbf{m}_i$). Por lo tanto, la función de pérdida es minimizada asignando N observaciones a C prototipos de tal forma que dentro de cada prototipo se minimiza la no similitud promedio de las observaciones. Una vez que la convergencia se alcanza, se modelan eficientemente N frames de silencio de dimensión K mediante C vectores prototipo de dimensión K denotados por $\bar{\mathbf{E}}_i^{opt}$, $i = 1,\ldots,C$. En la figura 5.1 observamos como la naturaleza compleja del ruido puede simplificarse (suavizarse) usando esta aproximación cluster. La aproximación cluster acelera la función de decisión de una manera significativa dado que la dimensión de los vectores de características se reduce sustancialmente ($N \rightarrow C$).

5.4.1. Función de decisión suave para VAD

Para clasificar la segunda clase de datos (vectores de energía de los frames de voz) empleamos un esquema algorítmico secuencial básico, relacionado con la cuantización vectorial de Kohonen (LVQ) [Kohonen, 1989], que utiliza un ventana de múltiple observación (MO) centrada en el frame l, como se mostró en la Sección 5.2. Para este objetivo consideremos la misma medida de similitud, un umbral γ y un número máximo de clusters permitidos $\mathbf{K} = 2$.

Sea $\hat{\mathbf{E}}(l)$ el vector de características de decisión que se define sobre la

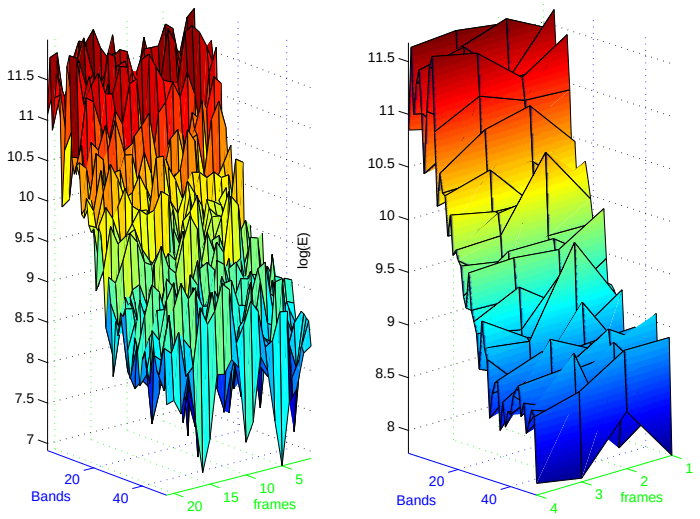


Fig. 5.1: a) 20 Energías logarítmicas de frames de ruido, calculadas usando $N_{FFT} = 256$, y promediadas sobre 50 subbandas. b) La aproximación cluster al conjunto anterior de Energías logarítmicas usando una decisión abrupta de tipo **C**-medias (C=4 prototipos).

ventana MO como sigue:

$$\hat{\mathbf{E}}(l) = \text{máx}\{\mathbf{E}(j)\}, \quad j = l - m, \ldots, l + m \tag{5.9}$$

La selección de este vector de características "envolvente", que describe no sólo un frame instantáneo sino además un conjunto de $(2m{+}1)$ vectores, es útil dado que detecta la presencia de voz en el canal de forma anticipada (transición silencio-voz) y retiene la decisión en alto (voz), suavizando la decisión del VAD (como una estrategia de guarda en un algoritmo basado en "hangover" en la transición voz-silencio [Marzinzik and Kollmeier, 2002; Li et al., 2002]), como se muestra en la figura 5.2. El uso de una ventana de MO impone un retardo de m frames al algoritmo que, para muchas aplicaciones como el reconocimiento de voz no es un inconveniente de implementación.

Finalmente, la presencia de un nuevo cluster (detección de frames de voz) se basa en la siguiente condición:

$$\eta(l) = \log\left(1/K \sum_{k=1}^{K} \frac{\hat{E}(k,l)}{< \bar{\mathbf{E}}_i >}\right) > \gamma \tag{5.10}$$

donde $< \bar{\mathbf{E}}_i >= 1/C \sum_{i=1}^{C} \bar{\mathbf{E}}_i = 1/C \sum_{i=1}^{C} \sum_{j=1}^{N} \gamma_{ij} \mathbf{E}_j$ es el centro del prototipo promediado de ruido y γ es el umbral de decisión.

Una estrategia para adaptar al algoritmo en entornos no estacionarios consiste en actualizar los centros de los prototipos de ruido de acuerdo con la decisión del VAD en los periodos que este detecta silencio (aquellos que no satisfacen la ecuación 5.10). Esta actualización se realiza de manera competitiva (sólo el prototipo de ruido más cercano al vector de características actual se mueve en tal dirección):

$$\begin{aligned} &\bar{\mathbf{E}}_{i'} = arg_{min}\left(||\bar{\mathbf{E}}_i - \hat{\mathbf{E}}(l)||^2\right); \quad i = 1, \ldots C \\ &\Rightarrow \quad \bar{\mathbf{E}}_{i'}^{new} = \alpha \cdot \bar{\mathbf{E}}_{i'}^{old} + (1 - \alpha) \cdot \hat{\mathbf{E}}(l) \end{aligned} \tag{5.11}$$

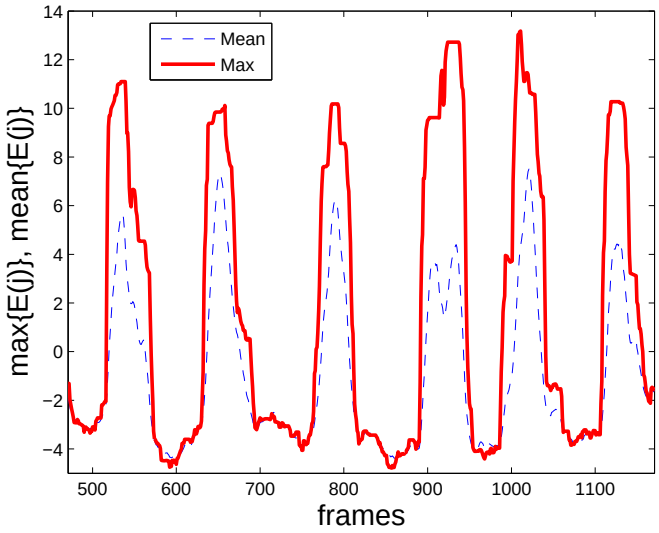


Fig. 5.2: Función de decisión para dos criterios distintos: la envolvente de la energía (equation 5.9) y el promedio de energía.

donde α es un factor de actualización. Su valor es cercano a la unidad para conseguir una regla de decisión suave (por ejemplo seleccionamos en las simulaciones $\alpha = 0{,}99$), para que, los frames de voz incorrectamente clasificados no afecten de manera significativa al modelo de ruido.

5.5. Algunos detalles sobre el algoritmo LTCM

La principal ventaja del algoritmo propuesto es, sin duda, su capacidad de encarar aplicaciones en tiempo real como por ejemplo sistemas DSR. El Hard C-medias descrito en las secciones anteriores se aplica al espacio de ruido una vez (para crear el modelo de prototipos) y después, se actualiza (el modelo) con la ecuación 5.11 durante los frames de silencio, en los cuales la ecuación 5.10 no se satisface. En los experimentos de reconocimiento (Sección 5.6), la selección del umbral está basada en los resultados obtenidos previa-

mente, en los experimentos de detección (puntos de operación para todas las condiciones en las curvas ROC del receptor para todas las condiciones). El punto de operación (umbral seleccionado) debería ser el que proporcione el mejor compromiso entre la tasa de acierto y de falsa alarma, por lo tanto el umbral se adapta dependiendo de la condición de ruido de trabajo (que se puede estimar en los frames de inicio).

El VAD realiza la detección de voz/silecio comparando la decisión no sesgada del LTCM con un umbral adaptativo [Ramírez et al., 2004a], es decir este umbral se adapta a la energía de ruido observada E. Se asume que el sistema trabajará en diferentes condiciones de ruido caracterizados por la energía del ruido de fondo. Los umbrales óptimos (puntos de trabajo) γ_0 y γ_1 pueden derivarse del sistema trabajando en las condiciones más limpias y más ruidosas. Estos umbrales definen una curva de calibración lineal para el VAD que se aplica durante el periodo de inicialización para seleccionar un umbral como una función de la energía del ruido E:

$$\gamma = \left\{ \begin{array}{ll} \gamma_0; & E \leq E_0 \\ \frac{\gamma_0-\gamma_1}{E_0-E_1} + \gamma_0 - \frac{\gamma_0-\gamma_1}{1-E_1/E_2}; & E_0 < E < E_1 \\ \gamma_1; & E \geq E_1 \end{array} \right\} \tag{5.12}$$

donde E_0 y E_1 son las energías del ruido de fondo para las condiciones más limpias y ruidosas que pueden ser determinadas examinando las bases de datos que se usan. De esta forma se asegura con este modelo una discriminación alta de voz/silencio dado que la detección de silencio se mejora en niveles de SNR medio y alto mientras que se mantiene una alta precisión en la detección de segmentos de voz bajo condiciones muy desfavorables de ruido.

El algoritmo descrito hasta ahora se presenta como pseudocódigo en Alg. 5.2. La Figura 5.3 muestra la operación del algoritmo propuesto sobre una frase de la base de datos en español SpeechDat Car [Moreno et al., 2000].

Alg. 5.2: Pseudocódigo del algoritmo LTCM para VAD

1. Inicializar el modelo de ruido:
   - Seleccionar N vectores de características $\{\mathbf{E}_j\}, j = 1, \ldots, N$.
   - Calcular el umbral γ.
2. Apicar clustering C-medias a los vectores de características para extraer los C centros de los prototipos de ruido $\{\bar{\mathbf{E}}_i\}, i = 1, \ldots, C$
3. for l=init to end
   a) Calcular $\hat{\mathbf{E}}(l)$ en la ventana de MO
   b) if $\eta(l) > \gamma$ (ecuación 5.10) then VAD=1
   else Actualizar los centros de los prototipos de ruido $\{\bar{\mathbf{E}}_i\}, i = 1, \ldots, C$ (ecuación 5.11)

Incluye en la figura: *i*) el modelo de energías logarítmicas del ruido (arriba-izqda), *ii*) la aproximación clustering de C-medias (arriba-decha), *iii*) la energía logarítmica del frame actual (frame=3) representada junto con el modelo de ruido (prototipos, $C = 4$) (abajo-izda) y *iv*) la regla de decisión en función del tiempo (segmentos)(abajo-decha).

A continuación se estudia la sensibilidad del método propuesto al número de prototipos utilizados en el modelo. Se encontró de manera experimental que el comportamiento del algoritmo es casi independiente del número prototipos C. La Figura 5.4 muestra que la precisión del algoritmo (tasa de detección de ruido frente a la la tasa de falsa alarma) en la discriminación de voz-silencio no se ve afectada por el número de prototipos siempre que $C \geq 3$, por lo tanto las ventajas de la aproximación clustering es evidente. Note que el objetivo del VAD es trabajar tan cerca posible a la esquina superior izquierda en la curva ROC con una clasificación de segmentos de voz y silencio sin error. La descripción de estas bases de datos para construir las curvas ROC se presentan en la parte experimental (sección 5.6) así como en los Capítulo anteriores dentro de sus partes experimentales.

La Figura 5.5 muestra el efecto de incorporar información contextual a la regla de decisión en la operación del VAD. Un valor elevado de m incrementa la precisión y la robustez del algoritmo (esto es, pequeñas variaciones en γ no afectan al rendimiento del VAD). Esto se justifica por un desplazamiento hacia arriba y hacia la izquierda de la curva ROC lo cual permite trabajar con una tasa de acierto de voz/silencio superior. El efecto del número de bandas utilizadas en el algoritmo se representa en la figura 5.6. De ella se ve como el uso del promedio completo en la energía ($K = 1$) ó los valores de energía de entrada sin promediar ($K = 100$) reduce la efectividad del procedimiento clustering igualando su precisión a otros VADs anteriormente propuestos y

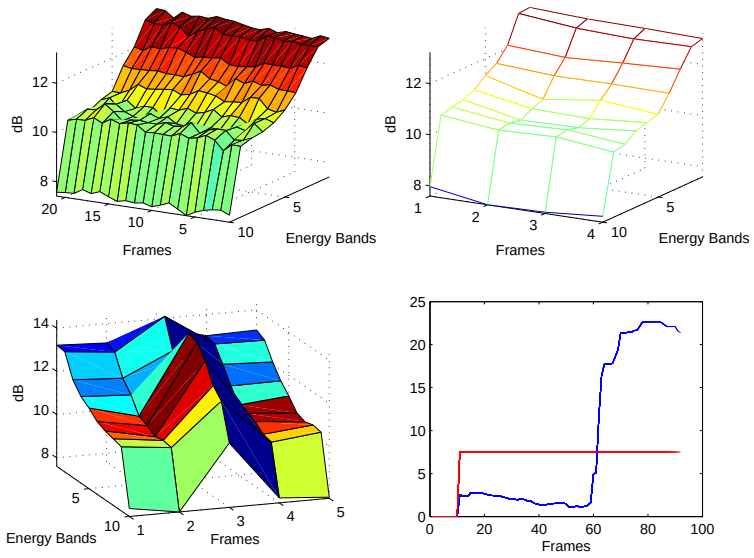


Fig. 5.3: Un paso en el algoritmo. El frame seleccionado es clasificado como frame de voz (VAD=1) como se muestra en la función de decisión a) energías logarítmicas en subbandas para el ruido, b) centros de los prototipos del Hard C-medias, c) comparación entre los prototipos del ruido y la energía-log del frame actual, d) evolución de la función de decisión y umbral.

que veremos en la sección experimental.

Finalmente, la Figura 5.7 muestra la operación del VAD propuesto sobre una frase de la base de datos en español SpeechDat-Car (SDC)[Moreno et al., 2000]. La transcripción fonética es: "tres", "nueve","zero", "siete", "μinko", "dos", "uno", "otSo", "seis","cuatro". También mostramos la función de decisión y el umbral seleccionado en la operación del VAD LTCM para la misma frase.

5.5.1. Distribuciones de la variable de decisión

En esta sección estudiamos las distribuciones de la variable de decisión como una función de la longitud de la ventana de MO (m) para clarificar las motivaciones de proponer esta estrategia. Para ello se utiliza una versión eti-

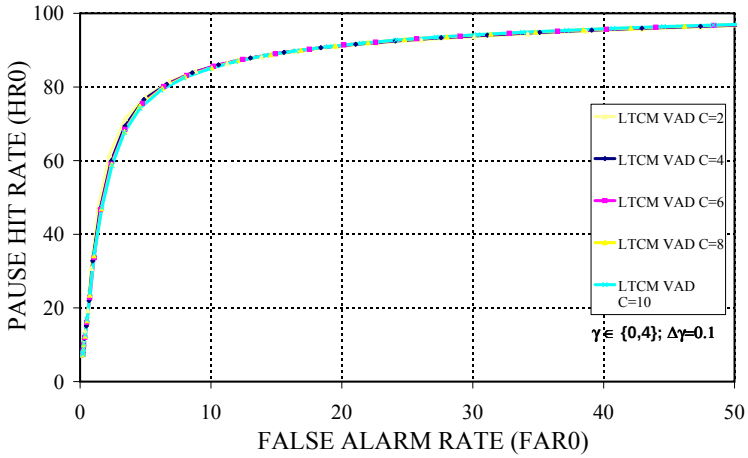


Fig. 5.4: Curvas ROC en condiciones de alto ruido para distinto número de prototipos. La DFT fue calculada con $N_{FFT} = 256$. $K = 10$ subbandas de energías-log fueron usadas para construir los vectores de características y la ventana de MO contenía $2 \cdot m + 1$ frames ($m = 10$).

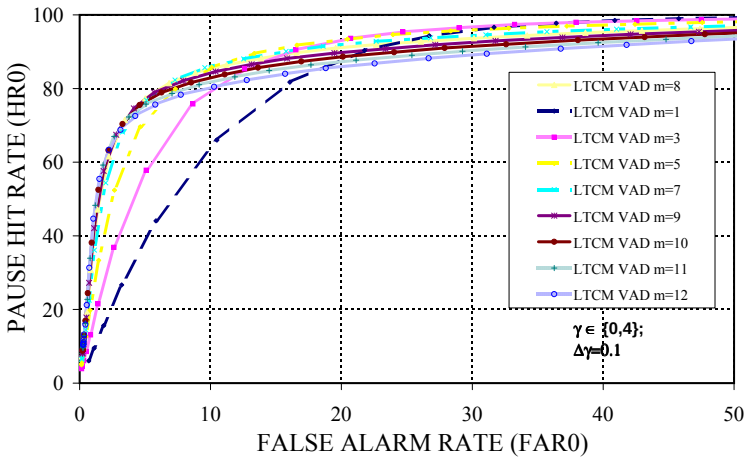


Fig. 5.5: Curvas ROC en condiciones de alto ruido para distinto número de segmentos de la ventana de MO (m). La DFT fue calculada con $N_{FFT} = 256$. $K = 10$ subbandas de energías-log fueron usadas para construir los vectores de características y el número de prototipos de ruido usados fue $C = 8$.

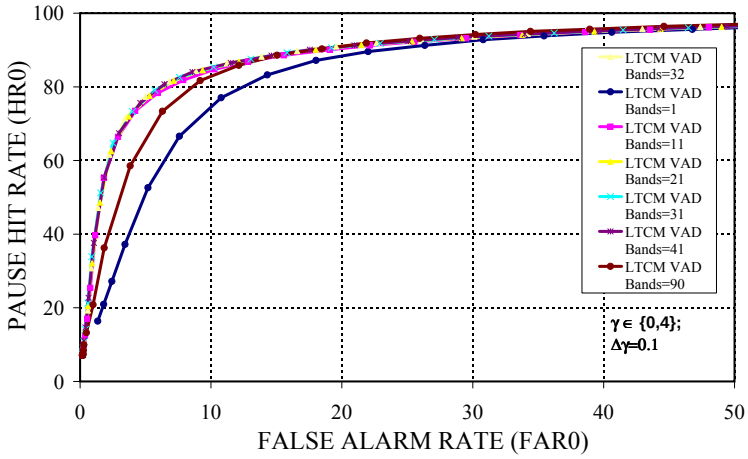


Fig. 5.6: Curvas ROC en condiciones de alto ruido para distinto número de subbandas. La DFT fue calculada con $N_{FFT} = 256$; $C = 10$ prototipos y una ventana de MO de $m = 10$.

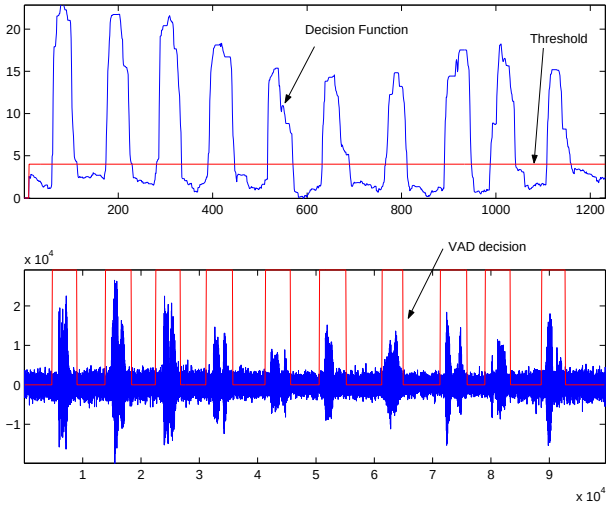


Fig. 5.7: Operación del VAD: Arriba- Función de decisión y umbral frente a los frames. Abajo- Señal de entrada y decisión del VAD como función del tiempo.

quetada de la base de datos en español SpeechDat-Car (SDC) [Moreno et al., 2000] como en capítulos anteriores. Como ya sabemos esta base de datos contiene grabaciones reales con un micrófono manos libres y otro cercano en distintas condiciones de conducción: a) coche parado, motor encendido, b) trafico en ciudad, baja velocidad, carretera rugosa y c) alta velocidad, en autopista. El medio ruidoso más desfavorable (i.e. alta velocidad en autopista) fue el usado con las grabaciones con el micrófono de manos libres para esta sección. De esta forma, medimos la divergencia de orden m entre voz y silencio durante los periodos correspondientes y se construyeron los histogramas y las distribuciones de probabilidad. La señal de entrada muestreada a 8 kHz se descompuso en segmentos de 25 ms solapados con un desplazamiento de 10-ms. La Fig. 5.8 muestra las distribuciones de voz y silencio para m= 0, 2, 5 y 8. Se deduce de ella que estas distribuciones están mejor separadas cuando aumentamos el orden de la ventana de MO. El ruido se confina y exhibe una menor varianza lo que implica una alta tasa de acierto de segmentos de silencio. Este hecho puede corroborarse calculando el error de clasificación de voz y silencio para el clasificador óptimo de Bayes. La figura 5.9 muestra los errores de clasificación como una función de la longitud de la ventana m. El error de clasificación de voz se divide aproximadamente por tres desde el 32 % a el 10 % cuando el orden del VAD se incrementa desde 0 a 8 ventanas. Esto se produce por la separación de las distribuciones que sucede cuando m se incrementa como muestra la Figura. 5.8. Por otro lado, la mejora de la robustez en la detección de voz se ve solamente perjudicado por un aumento moderado del error de detección de silencios. De acuerdo con la figura 5.9, el valor óptimo del orden del VAD es m= 8. Este análisis corrobora el hecho de que usar información contextual [Ramírez et al., 2004a] es beneficioso para el VAD dado que se reducen de manera importante los errores de clasificación

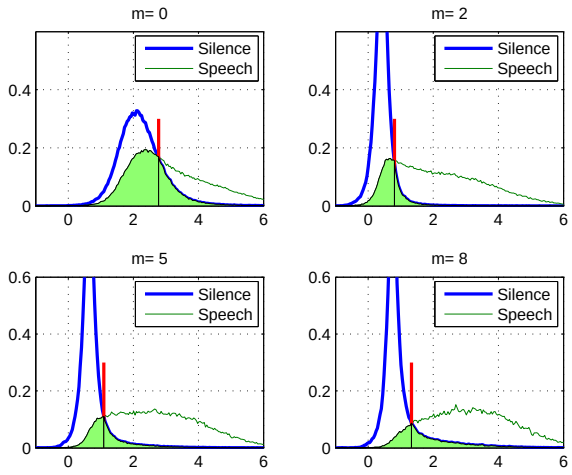


Fig. 5.8: Distribuciones de voz y silencio y las probabilidades de error para el clasificador óptimo de Bayes para m= 0, 2, 5 y 8.

totales.

5.6. *Resultados Experimentales*

Hemos llevado a cabo varios experimentos para valorar el rendimiento del algoritmo propuesto. El análisis se suele centrar en la determinación de las probabilidades de error para distintos escenarios y distintas valores de SNR [Beritelli et al., 2002; Marzinzik and Kollmeier, 2002], y en la influencia de la decisión del VAD en distintos sistemas de procesado de voz [Bouquin-Jeannes and Faucon, 1995a; Karray and Martin, 2003]. En esta sección describimos el trabajo experimental y los tests objetivos del rendimiento realizados para evaluar la eficiencia del VAD.

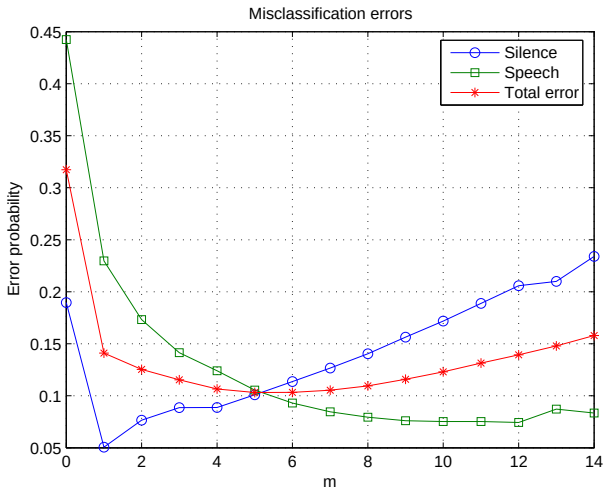


Fig. 5.9: Probabilidad de error como función de m.

5.6.1. *Evaluación bajo distintos ambientes de ruido*

En primer lugar, evaluamos el VAD como discriminador voz/silencio en diferentes escenarios de ruido y a diferentes niveles de SNR. La base de datos AURORA-2 [Hirsch and Pearce, 2000] es la adecuada para llevar a cabo este análisis. Como describimos en otros capítulos anteriores se construye mediante la base de datos limpia TIdigits que consiste en secuencias de hasta siete dígitos grabadas por locutores anglo-americanos como fuentes de voz, y una selección de ocho ruidos reales que se añaden artificialmente a la voz con SNRs de 20dB, 15dB, 10dB, 5dB, 0dB y -5dB. Estas señales ruidosas se han grabado en distintos escenarios (metro, multitudes (babble), coche, salas de exhibición, restaurantes, calles, aeropuertos y estaciones de trenes), que representan los escenarios más probables para los terminales de comunicaciones. En el análisis de discriminación, la base de datos limpia TIdigits se usó como referencia para etiquetar cada frase de ella, segmento a segmento como voz o silencio. El rendimiento de detección se mensuró en términos de la tasa

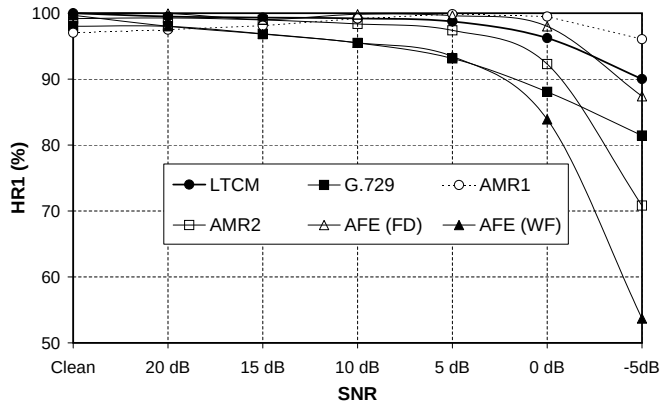


Fig. 5.10: Tasa de aciertos de voz (HR1) de los VADs estándar como una función de la SNR para la base de datos AURORA-2.

de aciertos de silencios (HR0) y la tasa de acierto de voz (HR1) definidos como la fracción de los segmentos de voz o silencio que son correctamente clasificados como voz y silencio, respectivamente:

$$HR1 = \frac{N_{1,1}}{N_1^{ref.}} \qquad HR0 = \frac{N_{0,0}}{N_0^{ref.}} \tag{5.13}$$

donde $N_1^{ref.}$ y $N_0^{ref.}$ son los números de frames de voz y silencio reales en toda la base de datos y $N_{1,1}$ y $N_{0,0}$ son los números de frames de voz y silencio correctamente clasificados, respectivamente.

Las figuras 5.10,5.11,5.12,5.13 establecen los resultados de la comparación para este análisis con los VAD estandarizados como por ejemplo el ITU-T G.729 [ITU, 1996], ETSI AMR [ETSI, 1999] ó ETSI AFE [ETSI, 2002] en términos de la tasa de acierto de silencio (HR0, Fig. 5.12) y de voz (HR1, Fig. 5.10) para la condición limpia y niveles de SNR entre 20 a -5 dB. Note que se representan los resultados para los dos VADs definidos en el estándar

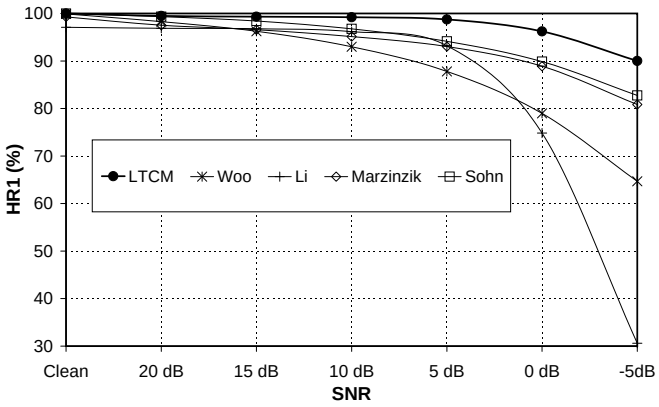


Fig. 5.11: Tasa de aciertos de voz (HR1) de otros VADs como una función de la SNR para la base de datos AURORA-2.

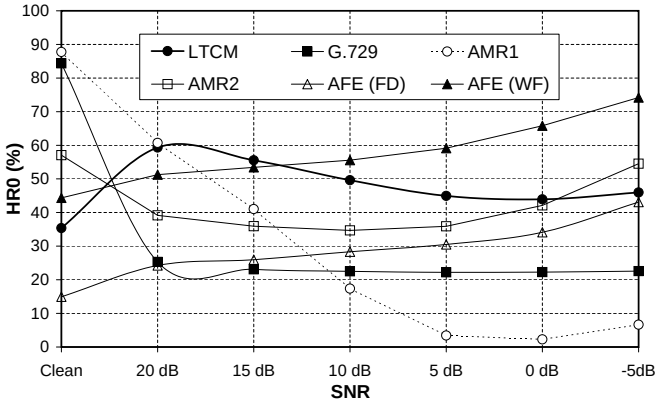


Fig. 5.12: Tasa de aciertos de silencio (HR0) de los VADs estándar como una función para la SNR de la base de datos AURORA-2.

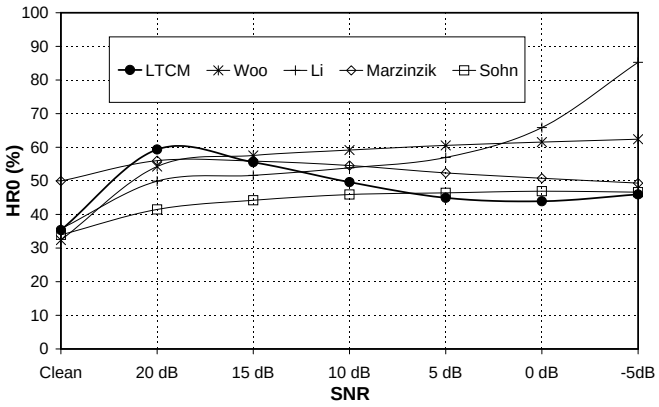


Fig. 5.13: Tasa de aciertos de silencio (HR0) de otros VADs como una función de la SNR para la base de datos AURORA-2.

AFE DSR [ETSI, 2002] para estimación del espectro de ruido en la etapa de filtrado de Wiener (WF) y "frame-dropping" (FD) de segmentos de silencio. Los resultados de estas curvas son los promedios entre todos los tipos de ruido.

Puede derivarse de las figuras 5.12 y 5.10 que: *i*) el VAD ITU-T G.729 presenta una baja precisión en la detección de voz cuando aumenta el nivel de ruido mientras que la detección de silencio es bastante buena en condiciones limpias (85 %) y deficiente (20 %) en condiciones ruidosas, *ii*) ETSI AMR1 presenta un comportamiento extremadamente conservativo con una precisión en detección de voz muy alta para todo el rango de SNR pero muy pobre para la detección de silencios cuando el nivel de ruido aumenta. Aunque AMR1 parece estar bien diseñado para la detección de voz en condiciones desfavorables de ruido, su comportamiento extremadamente conservativo degrada su capacidad de detección de silencio siendo HR0 menos del 10 % por debajo

de 10 dB, haciéndolo inútil en un sistema práctico de procesado de voz, *iii*) ETSI AMR2 conduce a mejoras significativas sobre el VAD G.729 y AMR1 aportando mejor precisión en la detección de silencios aunque sufre una gran degradación en su capacidad para detectar voz en condiciones de niveles altos de ruido, *iv*) El VAD usado en el estándar AFE para la estimación del espectro de ruido en la etapa de filtrado de Wiener está basado en la banda completa de energía. Éste proporciona un rendimiento bajo en la detección de voz con una rápida caída en las tasas de acierto para niveles de SNR pobres. En cambio, el VAD usado en el AFE para frame-dropping alcanza una alta precisión en la detección de voz pero moderada en silencios, y, finalmente *v*) LTCM suministra el mejor compromiso entre todos los VADs analizados. Obtiene un buen comportamiento tanto en la detección de periodos de silencio como de voz exhibiendo una lenta degradación de su rendimiento en condiciones desfavorables (para detección de voz: 90 % a -5 dB).

Las figuras 5.11 y 5.13 comparan el VAD propuesto con un grupo representativo de VADs recientemente publicados [Sohn et al., 1999; Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002]. Merece la pena mencionar que la base de datos AURORA-2 contiene grabaciones con periodos de silencio muy cortos entre dígitos y, consecuentemente, es más importante la correcta clasificación de segmentos de voz que de silencio en este escenario de reconocimiento. Esta es la razón por la que definimos el VAD con una alta tasa de acierto de voz incluso en condiciones altas de ruido. La Tabla 5.1 resume las mejoras que aporta LTCM VAD sobre los distintos métodos VAD en términos de promedios de tasas de aciertos de voz/silencio (sobre todo el rango de SNRs). Como se aprecia, el método desarrollado en esta sección con un 97.57 % de media de HR1 y un 47.81 % de media de HR0 proporciona el mejor compromiso en detección de segmentos de voz y silencios.

Tab. 5.1: Tasas promedio de acierto de voz/silencio para SNRs entre condición limpia y -5 dB. Comparación con: a) VADs estandarizados, y b) otros métodos recientes para VAD.

(a)

	G.729	AMR1	AMR2	AFE (WF)	AFE (FD)	**LTCM**
HR0 (%)	31.77	31.31	42.77	57.68	28.74	**47.81**
HR1 (%)	93.00	98.18	93.76	88.72	97.70	**97.57**

(b)

	Sohn	Woo	Li	Marzinzik	**LTCM**
HR0 (%)	43.66	55.40	57.03	52.69	**47.81**
HR1 (%)	94.46	88.41	83.65	93.04	**97.57**

5.6.2. Curvas ROC

En esta sección aplicamos un test adicional para comparar el rendimiento en la detección de voz por medio de las curvas ROC, una metodología ya usada en el presente trabajo y frecuentemente utilizada en comunicaciones. Ésta se basa, como sabemos, en las probabilidades de acierto y error en detección [Marzinzik and Kollmeier, 2002; Ramírez et al., 2005a; Ramírez et al., 2005b], que describen completamente la tasa de error del VAD. Se usó para este análisis el subconjunto de la base de datos en español AURORA SDC database [Moreno et al., 2000]. Esta base de datos contiene 4914 grabaciones de más de 160 locutores usando un micrófono de manos libres y otro de proximidad. Como en toda la base de datos SDC, los archivos están clasificados en tres condiciones de ruido: condiciones limpia, bajo ruido y alto ruido, que representan diferentes condiciones de conducción y una SNRs promedio de 12dB, 9dB y 5dB, respectivamente. Las grabaciones procedentes del micrófono cercano fueron usadas como referencia en el análisis para etiquetar los

frames de voz y silencio, mientras que las grabaciones del micrófono de manos libres fueron usadas para evaluar los distintos VADs en términos de sus curvas ROC. La tasa de acierto de silencio (HR0) y su falsa alarma (FAR0= 100-HR1) fueron determinadas para cada condición para el VAD LTCM y para los VADs G.729, AMR1, AMR2, y AFE, que fueron usados como referencia. Para el cálculo de la tasa de falsa alarma así como la de acierto, los frames de voz y silencio "reales" fueron determinados como mencionamos anteriormente usando la base de datos etiquetada manualmente en base a las grabaciones del micrófono de proximidad.

La figura 5.14 muestra la tasa de acierto de silencio (HR0) como una función de la falsa de alarma (FAR0= 100-HR1) del VAD LTCM para distintos valores del umbral de decisión y para distintos valores del número de observaciones m. Se muestra como incrementando el número de observaciones (m) proporciona una mejor discriminación de los segmentos de voz y silencio con un desplazamiento hacia arriba y hacia la izquierda de la curva ROC en su espacio. Esto permite al VAD trabajar más cerca del punto "ideal" de trabajo (HR0= 100 %,FAR0= 0 %) donde silencio y voz son perfectamente clasificados sin errores. Estos resultados son consistentes con nuestros experimentos preliminares y con los resultados de las figuras 5.8 y 5.9 que esperaban un una tasa de error mínima para m cerca de 8 frames.

La figura 5.15 muestra la curvas ROC para el VAD LTCM y otros algoritmos VAD de referencia [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999] para grabaciones del micrófono de manos libres en condiciones de nivel alto de ruido. Los puntos de trabajo de los VADs ITU-T G.729, ETSI AMR y ETSI AFE VADs están también representados. Los resultados muestran mejoras en la precisión de detección sobre tanto los VADs estándares como los usados de referencia [Woo et al., 2000;

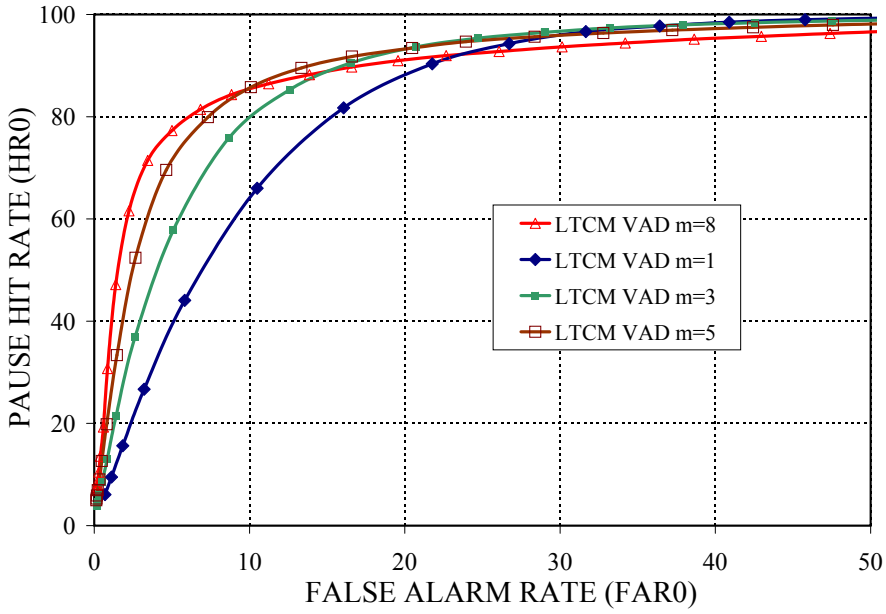


Fig. 5.14: Selección del número de m (Condición High: alta velocidad, buena carretera, SNR promedio de 5 dB).

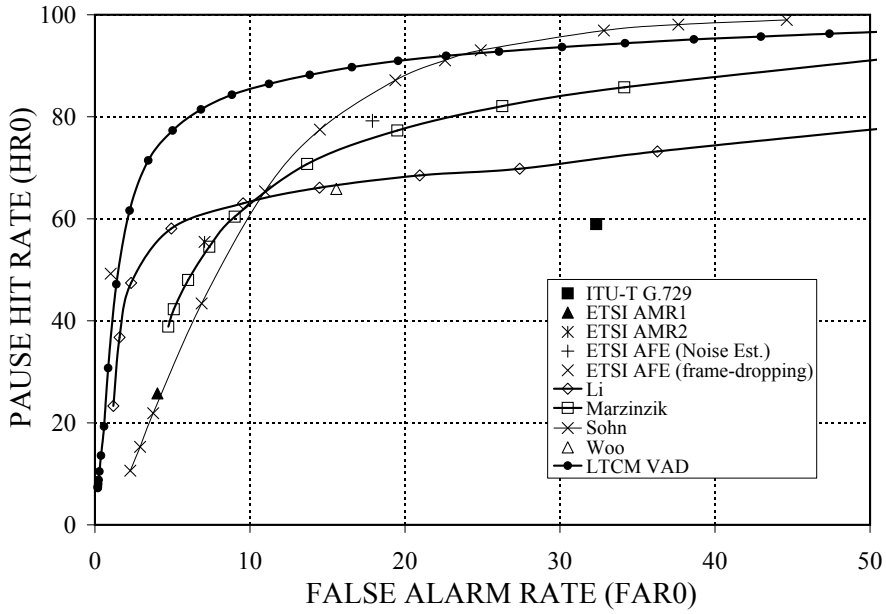


Fig. 5.15: Curvas ROC para la comparación con los VADs estandarizados y otros métodos (condición High: alta velocidad, buena carretera, SNR promedio de 5 dB.

Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999]. De entre todos los VADs examinados, el nuestro proporciona la menor tasa de falsa alarma para una tasa de acierto dado y también, la tasa de acierto más elevada dada una tasa de alarma determinada. Los beneficios son especialmente importantes sobre el VAD ITU-T G.729 [ITU, 1996], el cual es usado usualmente junto con codificadores de voz en transmisión discontinua, y sobre el el algoritmo de [Li et al., 2002], que está basado en un filtro lineal óptimo para la detección de contornos. LTCM también mejora al VAD de Marzinzik [Marzinzik and Kollmeier, 2002] que trata de capturar las envolventes del espectro de potencia, y el de Sohn [Sohn et al., 1999], que formula la regla de decisión por medio de un test estadístico de cociente de probabilidades (LRT) definido sobre el espectro de potencia de la señal ruidosa.

Estos experimentos proporcionan la primera medida del rendimiento del VAD. Otro tipo de tests del rendimiento ha sido propuestos [Benyassine et al., 1997] y están basados en los errores de recorte ("clipping"). Estas medidas suministran información relevante sobre la calidad del VAD y pueden usarse para mejorar u optimizar la operación del mismo. Nuestro análisis no distingue entre los distintos frames que se clasifican y evalúa exclusivamente las tasas de aciertos y fallos de estos. Por otro lado, los experimentos de reconocimiento que serán obtenidos para las bases de datos AURORA serán una medida directa de la calidad del VAD sobre la aplicación para la que fue diseñado. Los errores de recorte son evaluados indirectamente mediante el sistema de reconocimiento dado que hay una alta probabilidad de que suceda el error de borrado cuando parte de la palabra se elimina en la etapa de frame-dropping.

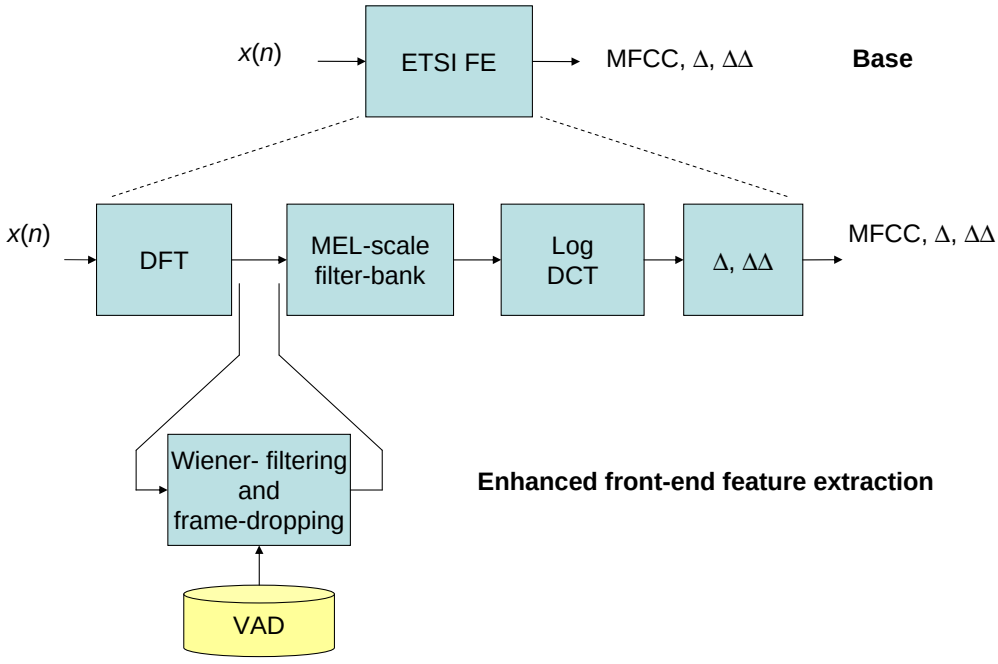


Fig. 5.16: Experimentos de reconocimiento. Extracción de características del Front-end.

5.6.3. Evaluación del VAD en un sistema ASR

Aunque el análisis de discriminación ó el análisis de las curvas ROC presentados en la sección anterior son efectivos para evaluar un algoritmo de discriminación de voz/silencio, vamos a estudiar en esta la influencia del VAD en un sistema de reconocimiento, aplicación principal para la cual se diseñó. Muchos autores han señalado que una buena comparación de los VADs se realiza evaluando su efecto en sistemas de reconocimiento [Woo et al., 2000] dado que una discriminación poco eficiente de frames de voz/silencio es una importante fuente de degradación del sistema completo cuando trabajamos en ambientes ruidosos [Karray and Martin, 2003]. Existen dos claras motivaciones para que esto ocurra: *i*) los parámetros del ruido tales como su espectro son actualizados durante los periodos de silencio; por lo tanto el sistema de realce está fuertemente influenciado por la calidad de la estimación del ruido, y *ii*) la técnica frame-dropping (FD), usualmente utilizada en reconocimiento de voz para reducir el número de errores por inserción causados por el ruido acústico, se basa también en la decisión del VAD por lo que fallos en esta clasificación conducen a una pérdida de voz, lo que a su vez causa pérdidas de segmentos de voz por borrado.

El punto de partida o referencia (Base) es el "front-end" de reconocimiento distribuido de voz (DSR) de [ETSI, 2000] propuesto por el grupo de trabajo ETSI STQ para la evaluación de los algoritmos de extracción robusta de características para DSR en ruido. El sistema de reconocimiento en esta sección está, al igual que en anteriores capítulos, basado en el paquete software HTK (Hidden Markov Model Toolkit) [Young et al., 1997]. La tarea consiste en reconocer dígitos conectados en inglés los cuales están modelados como HMMs (Hidden Markov Models) de palabras completas con los siguientes parámetros: 16 estados por palabra, modelos simples de izquierda a derecha,

mezcla de 3 Gaussianas por estado y sólo las varianzas de todos los coeficientes acústicos (sin la matriz de covarianza completa), mientras que el modelo para silencio consiste en 3 estados con una mezcla de 6 Gaussianas por estado. El vector de 39-parámetros consta de 12 coeficientes cepstrales (sin el coeficiente cepstral de orden cero), las energías logarítmicas del frame más los correspondientes coeficientes derivadas (Δ) y aceleraciones ($\Delta\Delta$).

Se definen dos modos de entrenamiento para los experimentos sobre Aurora2: *i*) entrenamiento usando sólo datos limpios (Clean Training), y *ii*) entrenamiento sobre datos limpios y ruidosos (Multi-Condition Training). Para AURORA-3 SpeechDat-Car, se usan las condiciones well-matched (WM), medium-mismatch (MM) y high-mismatch (HM) (ajuste perfecto, desajuste medio o alto). AURORA3 contiene como hemos dicho varias veces en este trabajo, grabaciones reales de dos micrófonos, uno de manos libres y otro de proximidad. En la condición WM, ambos micrófonos son usados para entrenar y testar. En condición MM, el entrenamiento y el test se lleva acabo con el micrófono de manos libres. En la condición de HM, el entrenamiento se lleva a cabo usando el micrófono de proximidad para todas las condiciones de conducción mientras que el test se realiza mediante el micrófono de manos libres desde bajos niveles de ruido hasta altos. Finalmente, el rendimiento del reconocedor se evalúa en términos de su precisión de reconocimiento de palabra (WAcc) que tiene en cuenta el número de errores de sustitución (S), borrado (D) y de inserción (I):

$$WACC(\,\%) = \frac{N - D - S - I}{N} \times 100\,\% \tag{5.14}$$

donde N es el número total de palabras en la base de datos de test.

Estudiamos en esta sección además la influencia de la decisión del VAD en el rendimiento de distintos esquemas de extracción de características. La

primera aproximación (que se muestra en la figura 5.16) incorpora filtrado de Wiener (WF) al sistema Base ETSI, 2000 como un método de supresión de ruido. El segundo método de extracción de características evaluado usa filtrado de Wiener y frame dropping de segmentos de silencio. El algoritmo implementado sigue el mismo esquema para la etapa de filtrado que el presentado en el front-end avanzado del estándar AFE DSR [ETSI, 2002]. Se usó el mismo esquema de extracción de características para entrenar y testar sin incluir ningún otro tipo de técnica de reducción del desajuste que presenta este estándar (por ejemplo procesado de la forma de onda ó ecualización ciega) dado que éstas no se ven afectadas por la decisión del VAD y podrían enmascarar el impacto del VAD sobre el rendimiento del sistema completo.

Tab. 5.2: Precisión promedio de palabra para la base de datos AURORA-2. (a) Entrenamiento limpio. (b) Entrenamiento Multi-condición.

(a)

	Base	**Base + WF**					**Base + WF + FD**				
		G.729	AMR1	AMR2	AFE	**LTCM**	G.729	AMR1	AMR2	AFE	**LTCM**
Clean	99.03	98.81	98.80	98.81	98.77	98.88	98.41	97.87	98.63	98.78	99.18
20 dB	94.19	87.70	97.09	97.23	97.68	97.46	83.46	96.83	96.72	97.82	98.05
15 dB	85.41	75.23	92.05	94.61	95.19	95.14	71.76	92.03	93.76	95.28	96.10
10 dB	66.19	59.01	74.24	87.50	87.29	88.71	59.05	71.65	86.36	88.67	90.71
5 dB	39.28	40.30	44.29	71.01	66.05	72.48	43.52	40.66	70.97	71.55	75.82
0 dB	17.38	23.43	23.82	41.28	30.31	42.91	27.63	23.88	44.58	41.78	47.01
-5 dB	8.65	13.05	12.09	13.65	4.97	15.34	14.94	14.05	18.87	16.23	19.88
Average	60.49	57.13	66.30	78.33	75.30	**79.34**	57.08	65.01	78.48	79.02	**81.54**

(b)

	Base	**Base + WF**					**Base + WF + FD**				
		G.729	AMR1	AMR2	AFE	**LTCM**	G.729	AMR1	AMR2	AFE	**LTCM**
Clean	98.48	98.16	98.30	98.51	97.86	98.45	97.50	96.67	98.12	98.39	98.78
20 dB	97.39	93.96	97.04	97.86	97.60	97.93	96.05	96.90	97.57	97.98	98.41
15 dB	96.34	89.51	95.18	96.97	96.56	97.06	94.82	95.52	96.58	96.94	97.61
10 dB	93.88	81.69	91.90	94.43	93.98	94.64	91.23	91.76	93.80	93.63	95.39
5 dB	85.70	68.44	80.77	87.27	86.41	87.54	81.14	80.24	85.72	85.32	88.40
0 dB	59.02	42.58	53.29	65.45	64.63	66.23	54.50	53.36	62.81	63.89	66.92
-5 dB	24.47	18.54	23.47	30.31	28.78	31.21	23.73	23.29	27.92	30.80	32.91
Average	86.47	75.24	83.64	88.40	87.84	**88.68**	83.55	83.56	87.29	87.55	**89.35**

La Tabla 5.2 muestra los resultados de reconocimiento sobre la base de datos AURORA-2 como una función de la SNR usando los VADs G.729,

AMR, AFE, y LTCM VAD. Estos resultados fueron promediados sobre los tres conjuntos experientos de de test de AURORA-2. Note que, en particular, para los resultados de reconocimiento para los VADs AFE, hemos usado la misma configuración usada en el estándar [ETSI, 2002] con distintos VADs para WF y FD. Sólo periodos exactos de voz son conservados en la etapa de FD y consecuentemente, todos los frames clasificados por el VAD como silencio son descartados. FD tiene un gran impacto en el entrenamiento de los modelos de silencio dado que provoca que menos frames de silencio estén disponibles para el entrenamiento. Si FD es suficientemente efectiva, muy pocos frames de silencio serán introducidos en el reconocedor en la etapa de test y por lo tanto, habrá poca influencia de los modelos de silencio en el rendimiento del reconocedor de voz. Concluimos señalando que, el VAD propuesto mejora los estándares G.729, AMR1, AMR2 y AFE cuando son usados para WF y también, cuando el VAD es usado para eliminar los frames de silencio. La decisión del VAD es utilizada en la etapa de WF para estimar el espectro de ruido durante los periodos de silencio. En esta aplicación a reducción de ruido, la estimación de la SNR es crítica. Por todo esto, el VAD basado en energía WF AFE sufre una rápida degradación en la detección de voz como se muestra en la figura 5.10, lo que implica un gran número de errores de reconocimiento y el incremento de la tasa de error de palabra, como se muestra en la Tabla 5.2.

Por otro lado, la etapa FD está fuertemente influenciada por el rendimiento del VAD. Un VAD eficiente para reconocimiento robusto de voz necesita un buen compromiso en la precisión en la detección voz/silencio. Cuando este sufre una rápida degradación bajo condiciones de ruido severas, pierde muchos frames de voz y se producen numerosos errores por borrado; si el VAD no identifica correctamente los periodos de silencio causa numerosos errores

de inserción y la correspondiente degradación de FD. El mejor rendimiento en reconocimiento se obtiene cuando el VAD LTCM se usa para WF y FD. Note que FD proporciona mejores resultados para el sistema de reconocimiento entrenado con voz limpia. Esto se debe al hecho de que los modelos entrenados con voz limpia no modelan adecuadamente los procesos de ruido, y normalmente causan errores de inserción de los periodos de silencio. Por lo tanto, eliminar de manera eficiente los frames de silencio provocará una reducción significativa de esta fuente de error. Finalmente, el ruido se modela correctamente cuando los modelos son entrenados usando voz ruidosa, y el sistema de reconocimiento tiende el mismo a reducir el número de errores de inserción en condición múltiple de entrenamiento como muestra la Tabla 5.2(a).

La Tabla 5.3(a) compara los promedios de precisión de palabra para los modos de entrenamiento limpio y de múltiple condición con el límite superior que podría alcanzarse cuando el sistema de reconocimiento pudiera usar la base de datos etiquetada manualmente. Estos resultados muestran que el rendimiento del algoritmo propuesto está muy próximo a la base de datos de referencia. En todos los conjuntos de test, el VAD LTCM supera los VADs estándar obteniendo los mejores resultados comparados con los obtenidos por AFE, AMR2, AMR1 y G.729. La Tabla 5.3(b) extiende esta comparación a otros VADs recientemente publicados [Sohn et al., 1999; Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002].

La Tabla 5.4 representa los resultados de reconocimiento para la base de datos SpeechDat-Car en español cuando usamos WF y FD en el sistema base [ETSI, 2000]. Otra vez, el VAD supera todos los algoritmos usados como referencia proporcionando mejoras significativas en reconocimiento de voz. Esta base de datos usada en los experimentos de AURORA 3 tiene periodos de silencio más largos que los de AURORA 2 y por tanto, los resultados

Tab. 5.3: Precisión promedio de palabra para condición limpia y multi-condición para experimentos con AURORA-2. Comparición con: (a) VADs estandarizados y (b) métodos recientemente publicados.

(a)

	G.729	AMR1	AMR2	AFE	**LTCM**	Hand-labelling
Base + WF	66.19	74.97	83.37	81.57	**84.01**	84.69
Base + WF+ FD	70.32	74.29	82.89	83.29	**85.44**	86.86

(b)

	Woo	Li	Marzinzik	Sohn	**LTCM**	Hand-labelling
Base + WF	83.64	77.43	84.02	83.89	**84.01**	84.69
Base + WF+ FD	81.09	82.11	85.23	83.80	**85.44**	86.86

del VAD son mejores para el sistema de reconocimiento. Este hecho puede apreciarse claramente cuando comparamos el rendimiento del VAD LTCM con por ejemplo el de Marzinzik [Marzinzik and Kollmeier, 2002]. La precisión de palabra de ambos VADs es bastante similar para la tarea AURORA2. Sin embargo, el propuesto en esta sección aporta una mejora significativa sobre el VAD de Marzinzik [Marzinzik and Kollmeier, 2002] en la tarea AURORA 3.

5.7. Conclusiones

Hemos mostrado en esta sección un nuevo algoritmo para mejorar la detección de voz y el reconocimiento del habla en ambientes ruidosos. El VAD propuesto LTCM se basa en el modelado del espacio de ruido usando la técnica Clustering Hard C-medias y emplea información contextual para la for-

	Base	Woo	Li	Marzinzik	Sohn	**LTCM**
WM	92.94	95.35	91.82	94.29	96.07	96.41
MM	83.31	89.30	77.45	89.81	91.64	91.61
HM	51.55	83.64	78.52	79.43	84.03	86.20
Avg.	**75.93**	89.43	82.60	87.84	90.58	**91.41**
	Base	G729	AMR1	AMR2	AFE	**LTCM**
WM	92.94	88.62	94.65	95.67	95.28	96.41
MM	83.31	72.84	80.59	90.91	90.23	91.61
HM	51.55	65.50	62.41	85.77	77.53	86.20
Avg.	**75.93**	75.65	74.33	90.78	87.68	**91.41**

Tab. 5.4: Precisión promedio de palabra (%) para la base de datos SDC en español.

mulación de una regla de decisión basada en un cociente de energía promedio. El VAD realiza una detección avanzada de los comienzos y retrasada de los finales de palabra lo cual, en parte, evita tener que incluir esquemas adicionales de "hangover" ó de bloques de reducción de ruido. Se mostró que incrementando la longitud de la ventana de información de retardo largo se obtiene una reducción significativa de los errores de clasificación. Hemos desarrollado un análisis exhaustivo sobre la base de datos de AURORA mostrando la efectividad de esta aproximación. El VAD propuesto LTCM supera a los recientemente métodos publicados para VAD como el de Sohn, que define un test de cociente de probabilidades sobre una única ventana de observación, y los estandarizados ITU-T G.729, ETSI AMR para el sistema GSM y los VADs ETSI AFE para reconocimiento de voz distribuido. Finalmente, también mejoró la tasa de reconocimiento cuando el VAD se usó para estimación del espectro de potencia, reducción de ruido y frame-dropping en un sistema ASR robusto al ruido.

6. MÁQUINAS DE VECTORES SOPORTE PARA VAD

Esta sección muestra una estrategia nueva y efectiva que emplea las máquinas de vectores soporte (SVM) para la detección de actividad de voz (VAD) en entornos ruidosos. El uso de kernels en el contexto de SVM permite proyectar los datos en otro espacio donde está definido un producto escalar (llamado espacio de características) por medio de una transformación no lineal. El vector de características incluye la relación señal-ruido en subbandas de la señal de voz. En la regla de clasificación se utilizan kernels basados en funciones de base radial (RBF) para el modelo de SVM. Se muestra la capacidad del método propuesto para aprender como la señal queda enmascarada por el ruido acústico y para definir una regla de decisión no lineal muy efectiva. Esta aproximación que trabaja por lotes obtiene grandes mejoras sobre los VADs estandarizados para transmisión discontinua de voz y para reconocimiento distribuido de voz, y otros recientemente publicados VADs.

6.1. Introducción

Desde su introducción en los años setenta [Vapnik, 1982], las Máquinas de vectores soporte ("*Support Vector Machines*" SVMs) han marcado el comienzo de una nueva era en el paradigma del aprendizaje mediante ejemplos. Las SVMs han atraído la atención de la comunidad de reconocimiento de patrones dado el gran número de méritos teóricos y computacionales que se han derivado de la teoría estadística del aprendizaje [Vapnik, 1995b; Vapnik, 1998] desarrollados por Vladimir Vapnik en AT&T. Esta Sección muestra un VAD efectivo basado en SVM para la mejora del rendimiento de los sistemas de procesado de voz que necesitan operar en ambientes ruidosos. El método propuesto combina un proceso de extracción de características robusto junto con un modelo SVM entrenado para clasificación. Los resultados son comparados con otras técnicas estándar y un conjunto de detectores representativos.

6.2. Introducción al aprendizaje con SVM

La SVMs han sido aplicadas recientemente a reconocimiento de patrones en una gran cantidad de aplicaciones por su capacidad de aprender de realizaciones de variables aleatorias. La razón se debe a que las SVMs son mucho más efectivas que los clasificadores paramétricos convencionales. En reconocimiento de patrones basado en SVM, el objetivo es construir una función $f : R^N \longrightarrow \{\pm 1\}$ usando datos de entrenamiento es decir, patrones de dimensión N $\mathtt{x}_i$ y clases y_i:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_\ell, y_\ell) \in R^N \times \{\pm 1\} \tag{6.1}$$

de tal forma que f clasificará correctamente nuevos ejemplos $(\mathbf{x}, y)$.

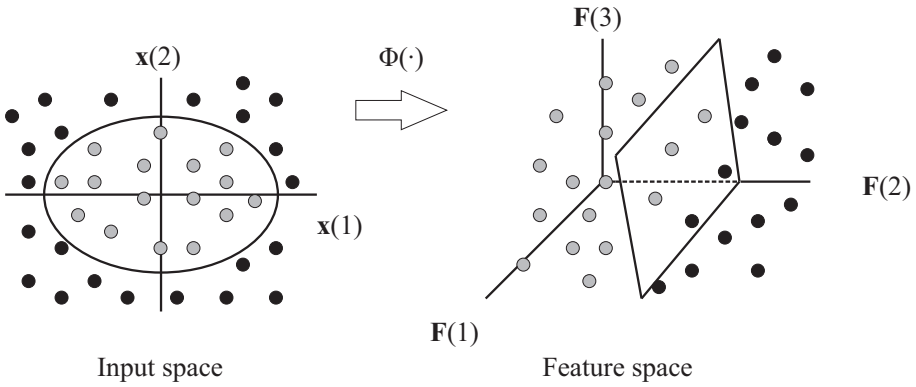


Fig. 6.1: Efecto de la transformación de la entrada al espacio de características donde la superficie de separación se convierte en lineal en este problema de clasificación

Los clasificadores basados en hiperplanos están basados en la clase de funciones de decisión:

$$f(\mathbf{x}) = \text{sign}\{(\mathbf{w} \cdot \mathbf{x}) + b\} \tag{6.2}$$

El hiperplano óptimo se define como el que presenta un margen máximo de separación entre las dos clases. La solución $\mathbf{w}$ del proceso de optimización cuadrática con restricción puede representarse en términos de un subconjunto de patrones de entrenamiento llamados vectores de soporte que están situados en el margen:

$$\mathbf{w} = \sum_{i=1}^{\ell} \nu_i \mathbf{x}_i \tag{6.3}$$

Por lo tanto, la regla de decisión depende sólo del producto vectorial entre patrones:

$$f(\mathbf{x}) = \text{sign}\{\sum_{i=1}^{\ell} \nu_i (\mathbf{x}_i \cdot \mathbf{x}) + b\} \tag{6.4}$$

El uso de kernels en SVM permite transformar los datos a un espacio definido con un producto escalar (llamado espacio de características) F por medio de una transformación no lineal $\Phi : R^N \longrightarrow F$ y aplicar el anterior algoritmo lineal en F. La figura 6.1 ilustra este proceso: el espacio de entrada bidimensional se transforma a uno de características de tres dimensiones en el que los datos son separables linealmente. El kernel está relacionado con la función Φ por $k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}))$. En el espacio de entradas, el hiperplano corresponde a una función de decisión no lineal cuya forma está determinada por el kernel. Existen tres kernels comunes que suelen ser usados por los usuarios de SVM para realizar la transformación no lineal:

- Polinómico

$$k(\mathbf{x}, \mathbf{y}) = [\gamma(\mathbf{x} \cdot \mathbf{y}) + c]^d \tag{6.5}$$

- Función de Base Radial (RBF)

$$k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma||\mathbf{x} - \mathbf{y}||^2) \tag{6.6}$$

- Sigmoidea

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\gamma(\mathbf{x} \cdot \mathbf{y}) + c) \tag{6.7}$$

Por lo tanto, la función de decisión es no lineal en el espacio de entrada aunque lineal en el de características:

$$f(\mathbf{x}) = \text{sign}\{\sum_{i=1}^{\ell} \nu_i k(\mathbf{x}_i, \mathbf{x}) + b\} \tag{6.8}$$

y los parámetros ν_i son las soluciones del problema de programación cuadrática que está usualmente determinado por el conocido algoritmo de Optimización mínima secuencial (SMO) [Platt, 1999]. Muchos problemas de clasificación son siempre separables en el espacio de características y usualmente

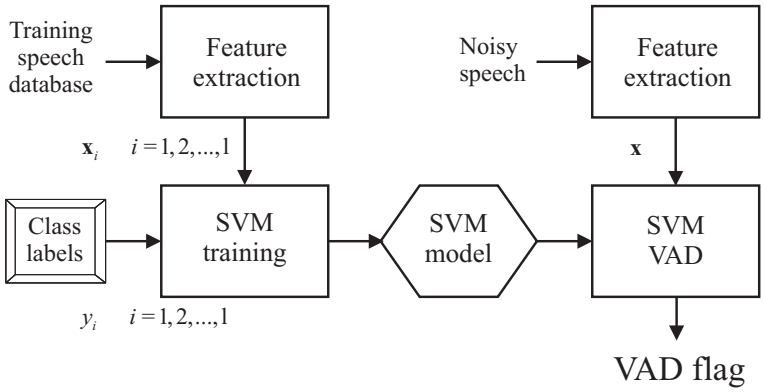


Fig. 6.2: Diagrama de Bloques del VAD basado en SVM

la elección usando kernels RBF obtiene mejores resultados en vez de usar kernels polinómicos o lineales [Clarkson and Moreno, 1999; Ganapathiraju et al., 2004].

6.3. VAD basado en SVM

En la figura 6.2 se muestra un diagrama de bloques del VAD propuesto. El primer paso es el proceso de entrenamiento usando datos seleccionados y sus clases correspondientes. La señal se procesa y el vector de características se extrae para entrenamiento. Una vez que el modelo SVM ha sido entrenado, el algoritmo para VAD basado en SVM consiste en los siguientes pasos: i) la señal de entrada se descompone en frames y se extraen las características para su clasificación, y ii) las características $\mathbf{x}$ son procesados por la función de decisión de SVM entrenada f definida en la ecuación 6.8.

6.3.1. Preprocesado y extracción de características

El algoritmo para extracción de características se define como sigue. La señal de entrada $x(n)$ muestreada a 8 kHz se descompone en frames de 25-ms

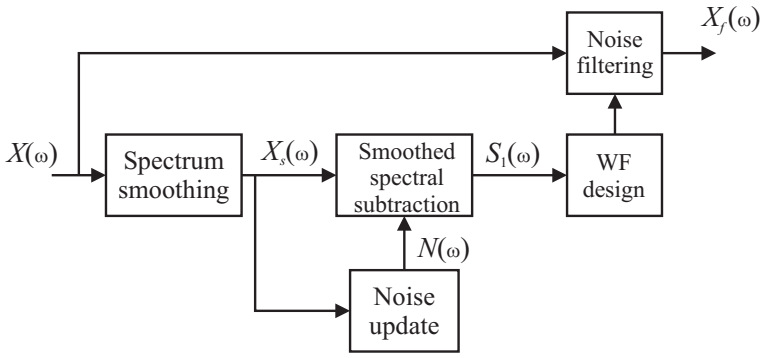


Fig. 6.3: Etapa de Filtrado para extracción de características

solapados con una ventana de desplazamiento de 10-ms. El frame actual que consiste en 200 muestras se rellena con ceros hasta 256 muestras y se calcula el espectro de potencia $X(\omega)$ usando la transformada discreta de Fourier (DFT). Aplicamos un proceso de filtrado basado en filtrado de Wiener para mejorar el rendimiento del VAD en un escenario de alto nivel de ruido. La figura 6.3 muestra el diagrama de bloques del proceso de filtrado. Por lo tanto, el espectro de ruido ha sido estimado durante un corto periodo de inicialización para diseñar el filtro óptimo de Wiener en el dominio de Fourier. El proceso de filtrado se describe como sigue:

i) Suavizado espectral. El espectro de potencia se promedia sobre dos frames consecutivos y dos bandas espectrales adyacentes.

ii) Estimación de ruido. El espectro de ruido $N(\omega)$ se actualiza durante los periodos de silencio por medio de un filtro IIR de primer orden sobre el espectro suavizado $X_s(\omega)$, es decir, $N(\omega) = \lambda N(\omega) + (1 - \lambda)X_s(\omega)$ donde $\lambda = 0{,}99$.

iii) Diseño del filtro de Wiener (WF). Primero se estima la señal limpia, $S_1(\omega)$ se estima combinando suavizado y sustracción espectral:

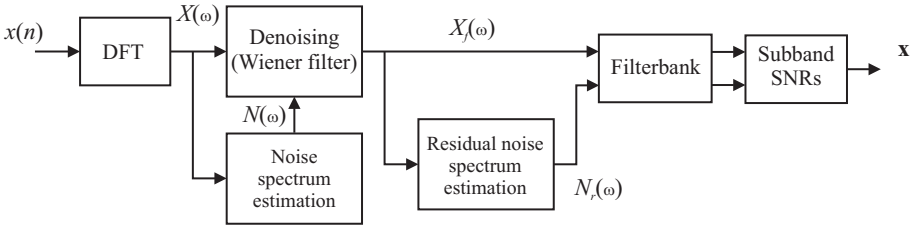


Fig. 6.4: Diagrama de bloques del VAD propuesto basado en SVM

$$S_1(\omega) = \gamma X'(\omega) + (1-\gamma)\,\text{máx}(X_s(\omega) - N(\omega), 0) \tag{6.9}$$

donde γ= 0.98. Entonces, el WF $H(\omega)$ se diseña como:

$$H(\omega) = \frac{\eta(\omega)}{1+\eta(\omega)} \tag{6.10}$$

donde:

$$\eta(\omega) = \text{máx}\left[\frac{S_1(\omega)}{N(\omega)}, \eta_{\text{mín}}\right] \tag{6.11}$$

y $\eta_{\text{mín}}$ se selecciona de tal forma que el filtro H proporciona una atenuación máxima de 20 dB. Note que, $X'(\omega) = H(\omega)X(\omega)$ es el espectro de la señal limpia, asumido nulo en el comienzo del proceso y usado para el diseño WF a través de las Eqs. 6.9 a 6.11. El filtro $H(\omega)$ se suaviza para eliminar cambios rápidos entre frecuencias adyacentes que pueden provocar frecuentemente ruido musical. De esta forma, la varianza del ruido residual se reduce y consecuentemente, la robustez al detectar silencio se realza. El suavizado se consigue truncando la respuesta al impulso del filtro causal FIR a 17 coeficientes usando una ventana de Hanning.

iv) Filtrado en el dominio de la frecuencia. El filtro de suavizado $H_s(\omega)$ se aplica en el dominio de frecuencias para obtener el espectro sin ruido $X_f(\omega) = H_s(\omega)X(\omega)$.

Una vez que la señal ha sido filtrada, el banco de filtros mostrado en la figura 6.4 reduce la dimensionalidad del vector de características a una representación que incluye información en banda ancha adecuada para detección. Esto es, la señal y el ruido residual se pasa a través de un banco de filtros de K bandas definido por:

$$\begin{gathered} E_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} X_f(\omega); \quad N_B(k) = \sum_{\omega=\omega_k}^{\omega_{k+1}} N_r(\omega) \\ \omega_k = \tfrac{\pi}{K}k \qquad k = 0, 1, ..., K-1 \end{gathered} \tag{6.12}$$

y la SNR en cada subbanda se obtiene como

$$\mathrm{SNR}(k) = 20\log_{10}\left(\frac{E_B(k)}{N_B(k)}\right) \qquad k = 0, 1, ..., K-1 \tag{6.13}$$

6.3.2. Entrenamiento de la regla de clasificación basada en SVM

El modelo SVM se entrenó usando la herramienta software LIBSVM [Chang and Lin, 2001]. Se usó un conjunto de entrenamiento de 12 frases de AURORA 3 SpeechDat-Car (SDC) en español. Como sabemos de Capítulos anteriores esta base de datos contiene 4914 grabaciones de más de 160 locutores usando dos micrófonos, uno de proximidad y otro manos libres. Los archivos se clasifican en tres condiciones de ruido: condición limpia, de bajo ruido y de alto ruido, las cuales representan diferentes condiciones de conducción con un promedio de valor de SNR entre los 25dB, y 5dB. Las grabaciones usadas para entrenar las SVMs fueron elegidas entre en los tres condiciones de ruido. La formulación SVM está basada en la clasificación de vectores de soporte C [Cortes and Vapnik, 1995; Vapnik, 1998] y la regla de

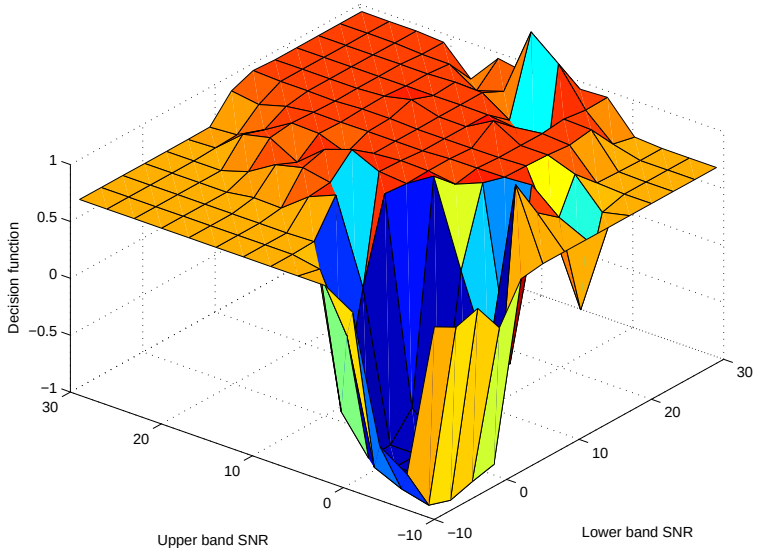


Fig. 6.5: Función de Decisión del modelo SVM entrenado con 2 bandas.

decisión se define usando la ecuación 6.8. Las SNRs en subbandas dadas por la ecuación 6.13 se usan como características discriminativas de voz mientras que el kernel elegido fue de tipo RBF en el proceso de entrenamiento que consiste en la resolución del problema primario:

$$\begin{aligned} &\text{mín}_{\alpha} \quad \tfrac{1}{2}\alpha^T \mathbf{Q}\alpha - \mathbf{e}^T\alpha \\ &0 \leq \alpha_i \leq C, \quad i = 1, 2, ..., \ell \\ &\text{subject} \quad \text{to} \qquad \mathbf{y}\alpha = 0 \end{aligned} \tag{6.14}$$

utilizando LIBSVM [Chang and Lin, 2001], donde $\mathbf{e}$= [1 1 ... 1], $C > 0$ es el límite superior y $Q_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$. Después de este proceso, los vectores de soporte $\mathbf{x}_i$ y los coeficientes α_i requeridos para evaluar la regla de decisión definida en la ecuación 6.8 son extraídos, donde $\nu_i = y_i\alpha_i$.

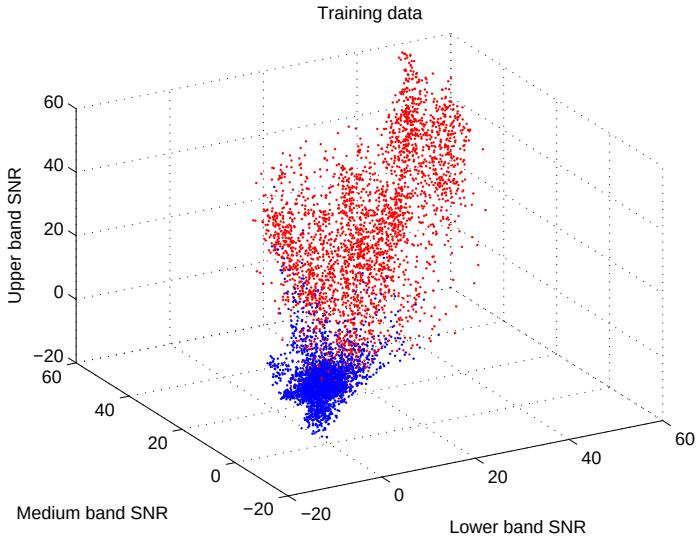
La figura 6.5 muestra la función de decisión de un modelo SVM entrenado a dos bandas. Note como, b puede usarse como el umbral de decisión para el VAD en el sentido que el punto de trabajo del VAD puede elevarse para satisfacer los requerimientos de la aplicación. Esto es crucial para la aplicación

que se está considerando aquí pues una pérdida de frames de voz afecta de forma significativa al rendimiento de la mayoría de los sistemas de procesado de voz. La próxima sección ilustra el comportamiento de la regla de decisión basada en SVM cuando modificamos el umbral del valor de entrenamiento.

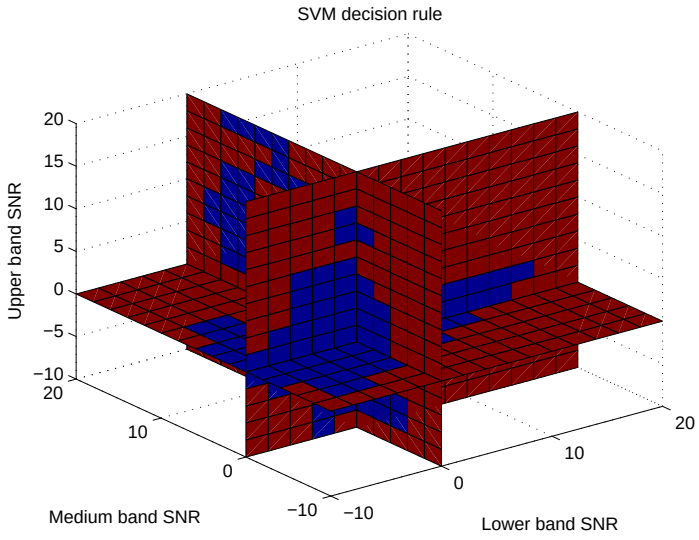
6.4. Análisis y mejoras

Esta sección analiza la regla de decisión en el espacio de entradas y sugiere un algoritmo rápido para la clasificación SVM. La figura 6.6.a muestra los datos de entrenamiento en un espacio de entrada de tres bandas. Se muestra que las dos clases no pueden ser separadas sin error en el espacio de entrada. La figura 6.6.b muestra la regla de decisión de SVM que se obtiene después del proceso de entrenamiento. Note que, *i*) las clases de voz y silencio están perfectamente distinguidas en el espacio de 3-D, y *ii*) el modelo SVM aprende como la señal se enmascara con el ruido y automáticamente define la regla de decisión en el espacio de entrada. La figura 6.6.b también sugiere un algoritmo rápido para calcular la regla de decisión definida por la ecuación 6.8 que se convierte en computacionalmente costosa cuando el número de vectores soporte y/ó la dimensión del vector de características es elevada. Note que toda la información que se necesita para decidir la clase dado un vector de características vector $\mathbf{x}$, reside en la figura 6.6.b. Por lo tanto, el espacio de entrada puede discretizarse sobre las distintas componentes del vector de características $\mathbf{x}$ como

$$\begin{array}{lllllll} \mathbf{x}(1) & m_{\mathbf{x}(1)}, & m_{\mathbf{x}(1)} + \Delta_{\mathbf{x}(1)}, & m_{\mathbf{x}(1)} + 2\Delta_{\mathbf{x}(1)}, & ..., & M_{\mathbf{x}(1)} \\ \mathbf{x}(2) & m_{\mathbf{x}(2)}, & m_{\mathbf{x}(2)} + \Delta_{\mathbf{x}(2)}, & m_{\mathbf{x}(2)} + 2\Delta_{\mathbf{x}(2)}, & ..., & M_{\mathbf{x}(2)} \\ ... & ... & & & & \\ \mathbf{x}(N) & m_{\mathbf{x}(N)}, & m_{\mathbf{x}(N)} + \Delta_{\mathbf{x}(N)}, & m_{\mathbf{x}(N)} + 2\Delta_{\mathbf{x}(N)}, & ..., & M_{\mathbf{x}(N)} \end{array} \tag{6.15}$$



(a)



(b)

Fig. 6.6: Regla de Clasificación en el espacio de entrada después de entrenar un modelo SVM con tres bandas. a) Datos de entrenamiento, b) Regla de clasificación basada en SVM.

y la regla de decisión $f(\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N))$ puede pre-calcularse para la rejilla de datos definido anteriormente y guardada en una tabla de búsqueda N-dimensional. Dado un vector de características $\mathbf{x}= [\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)]$, el primer paso es encontrar el punto más cercano en la rejilla definida y a continuación relaizar una búsqueda en la tabla para asignar la clase (voz y silencio) al vector de características $\mathbf{x}$.

6.5. Marco Experimental

Esta sección analiza el VAD propuesto y compara su rendimiento con otros algoritmos usados como referencia. El análisis está basado en las conocidas curvas ROC, una metodología muy frecuentemente usada en este contexto para describir la tasa de error del VAD. El subconjunto de AURORA de la base de datos original SDC [Moreno et al., 2000] fue usado otra vez para el análisis. Se determinaron las tasas de acierto de segmentos de silencio (HR0) y las de falsa alarma como una función del umbral de decisión siendo los frames de voz y silencio reales determinados por etiquetado manual de la base de datos del micrófono de proximidad.

Antes de mostrar los resultados comparativos, discutimos la selección del número óptimo de subbandas. La figura 6.7 muestra la influencia del bloque de reducción de ruido y el número de subbandas en la curva ROC en condiciones de alto nivel de ruido. Primeramente, la reducción de ruido no se emplea para mostrar mejor la influencia del número de subbandas. Incrementando el número de subbandas se mejora el rendimiento del VAD propuesto con un desplazamiento de la curva ROC en el espacio ROC. Para más de cuatro bandas, el VAD no reporta mejoras adicionales. Este valor proporciona el mejor compromiso entre coste computacional y rendimiento. Por otro lado, el bloque de reducción de ruido incluido en el VAD obtiene un

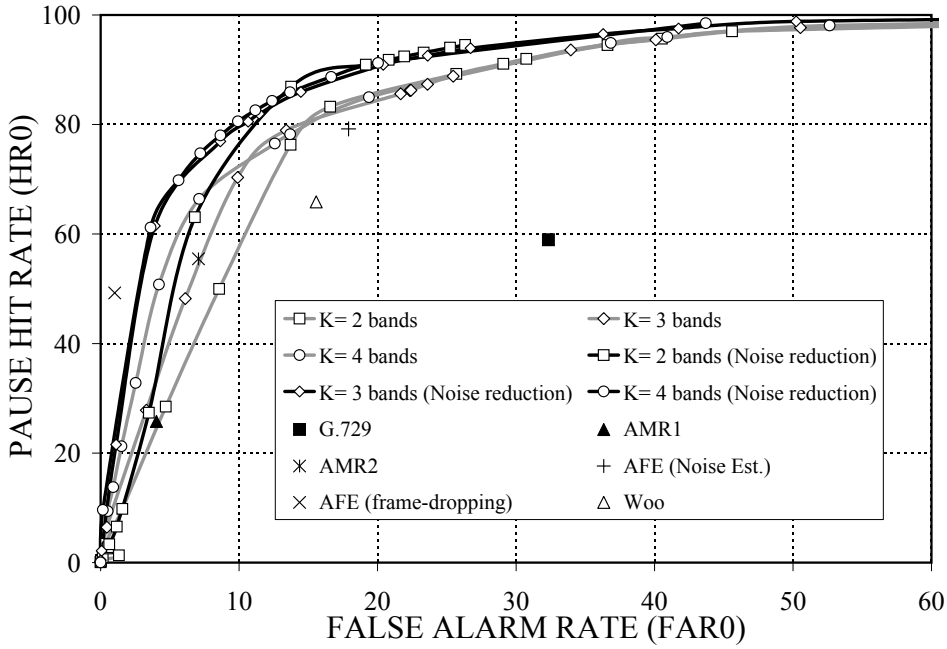


Fig. 6.7: Selección de Subbandas (alto nivel de ruido: alta velocidad en buena carretera y un promedio de SNR de 5 dB).

desplazamiento adicional de la curva ROC como se muestra en la figura 6.7.

La figura 6.8 muestra las curvas ROC del VAD propuesto y otros algoritmos frecuentemente referenciados [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999] para las grabaciones usando el micrófono de manos libres en condiciones de alto nivel de ruido. Los puntos de operación de los VADs ITU-T G.729, ETSI AMR y AFE también están incluidos. Los resultados muestran las ventajas en la precisión de detección sobre los VAD estándar y sobre un conjunto muy representativo de algoritmos VAD [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999]. De entre todos los VAD examinados, el nuestro proporciona las tasas de falsa alarma más bajas para una tasa de acierto fija y también, la tasa de acierto más alta para una tasa de falsa alarma dada. Los beneficios son especialmente importantes sobre ITU-T G.729, el cual es usado junto

con un codificador de voz para transmisión discontinua, y sobre el algoritmo de Li Li et al., 2002, que está basado en un filtro óptimo lineal para la detección de contornos. El VAD propuesto también mejora al VAD de Marzinzik [Marzinzik and Kollmeier, 2002] que captura las envolventes del espectro de potencias, y el de Sohn [Sohn et al., 1999], que formula la regla de decisión por medio de un modelo basado en test de cociente de probabilidades.

Una mejora adicional al VAD propuesto cosiste en la inclusión de información de retardo largo al igual que usan los algoritmos de los capítulos anteriores. En este caso se trata de incluir la información contextual de una ventana de MO en el vector de características:

$$\begin{aligned} E_B(k) &= \sum_{\omega=\omega_k}^{\omega_{k+1}} \hat{X}_f^l(\omega); \\ \omega_k = \tfrac{\pi}{K}k \qquad & k = 0, 1, ..., K-1 \end{aligned} \tag{6.16}$$

donde l denota el frame actual que está siendo procesado y

$$\hat{X}_f^l(\omega) = max\{X_f^j(\omega), j = l-L, \ldots l-1, l, l+1, \ldots l+L\} \tag{6.17}$$

De esta forma incluyendo el contexto de los $2L+1$ frames adyacentes, el rendimiento del VAD se mejora de manera significativa como podemos ver en la figura 6.9. En los siguientes experimentos no se usó la etapa de filtrado de ruido para no enmascarar el efecto de la inclusión de la ventana de múltiple observación. La figura 6.9 aclara las motivaciones para usar características de voz contextuales en el VAD basado en SVM. Para $K = 2$ el espacio 2-D de características se presenta para 12 frases de voz de AURORA 3 SpeechDat-Car en español. Se ve claramente que al incrementar la ventana de observación de $L = 1, \ldots, 8$ frames se produce una mejor separabilidad de los datos en el espacio de características y se permite aplicar un mejor clasificador basado en SVM.

La figura 6.10 muestra la influencia del número de subbandas en la curva ROC en condición de alto nivel de ruido (alta velocidad, buena carretera, SNR promedio de 5 dB) para $L = 8$. Los puntos de trabajo de los VADs ITU-T G.729, ETSI AMR, y ETSI AFE se incluyen también en la gráfica así como las curvas ROC de los métodos de detección más significativos. Incrementando el número de subbandas mejora el rendimiento del VAD como pasaba para una única observación, con un desplazamiento de la curva ROC en el espacio ROC. Para más de cuatro subbandas, el VAD no reporta ningún beneficio adicional. De esta forma, para $K = 4$ subbandas se obtiene el mejor compromiso entre coste computacional y rendimiento.

La figura 6.11 muestra las curvas ROC del VAD propuesto variando L y para $K = 4$ fijo. Se muestra cómo el incremento de la longitud de observación hasta 8 proporciona un desplazamiento de la curva ROC hacia arriba y hacia la izquierda. Por lo tanto los parámetros óptimos para el VAD propuesto son $K = 4$ subbandas y $L = 8$ frames. En este caso los resultados muestran mejoras significativas en detección voz/silencio sobre los VAD estándar y sobre los frecuentemente referenciados[Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999]. Estas mejoras se deben fundamentalmente al uso de información contextual en el vector de características y a la definición de una regla de decisión no-lineal sobre subbandas de los datos lo que permite al clasificador de SVM aprender como la señal de voz se enmascara por ruido acústico presente en el medio.

6.6. Conclusiones

Se ha propuesto un esquema eficiente de modelado para la detección de la presencia de voz en una señal ruidosa. La estrategia combina técnicas de reducción espectral de ruido y máquinas de vectores soporte para aprendizaje

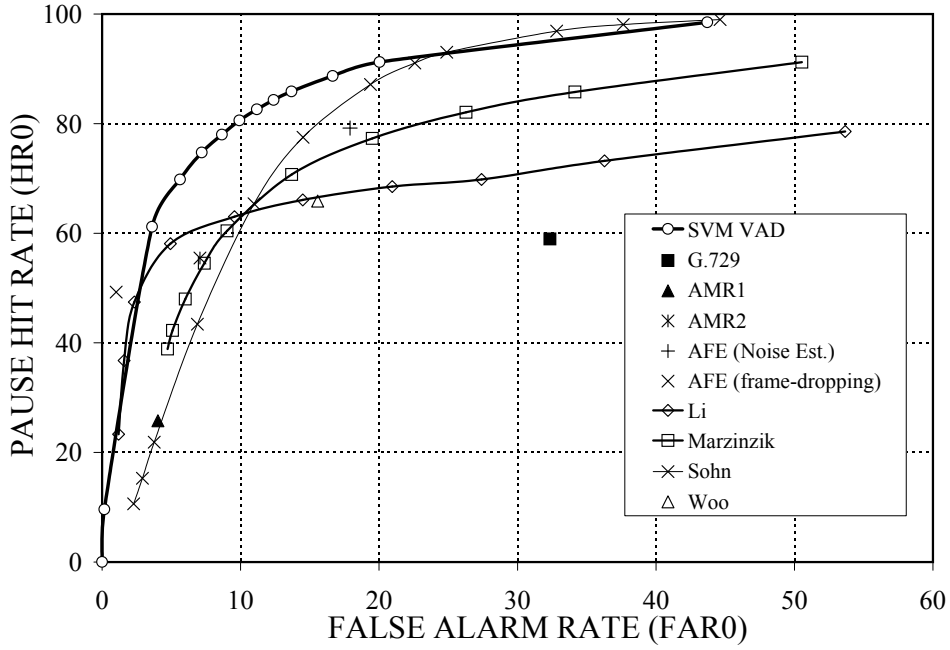


Fig. 6.8: Resultados comparativos con otros métodos para VAD

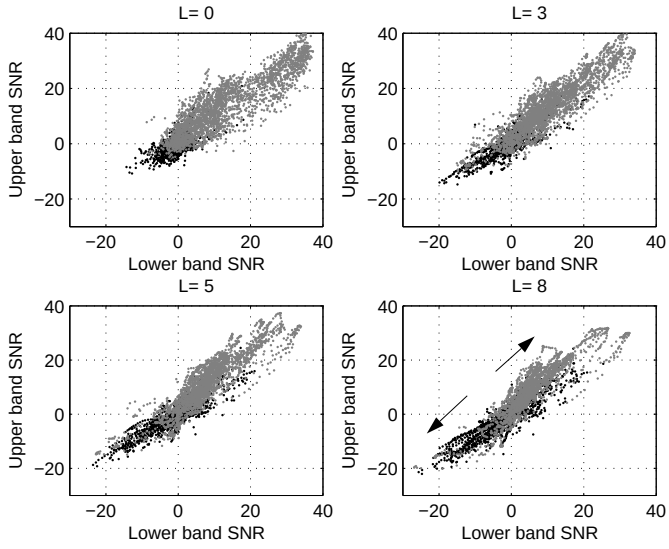


Fig. 6.9: Separación de las características de voz en el espacio de entrada cuando aumentamos el tamaño de la ventana L.

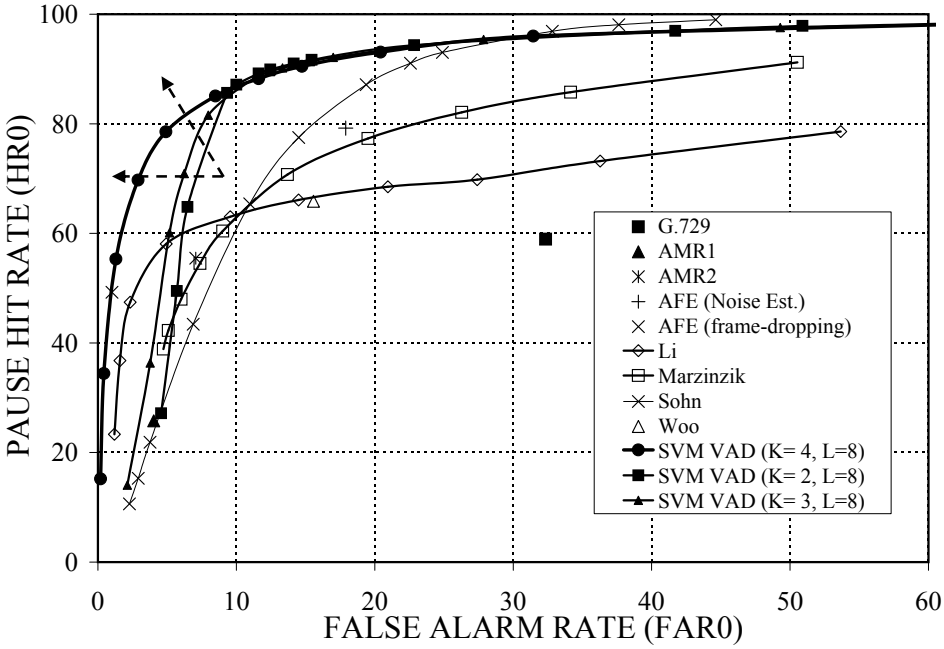


Fig. 6.10: Curvas ROC del VAD propuesto para diferente número de subbandas K (Alto nivel de ruido: alta velocidad en buena carretera a 5 dB de promedio de SNR).

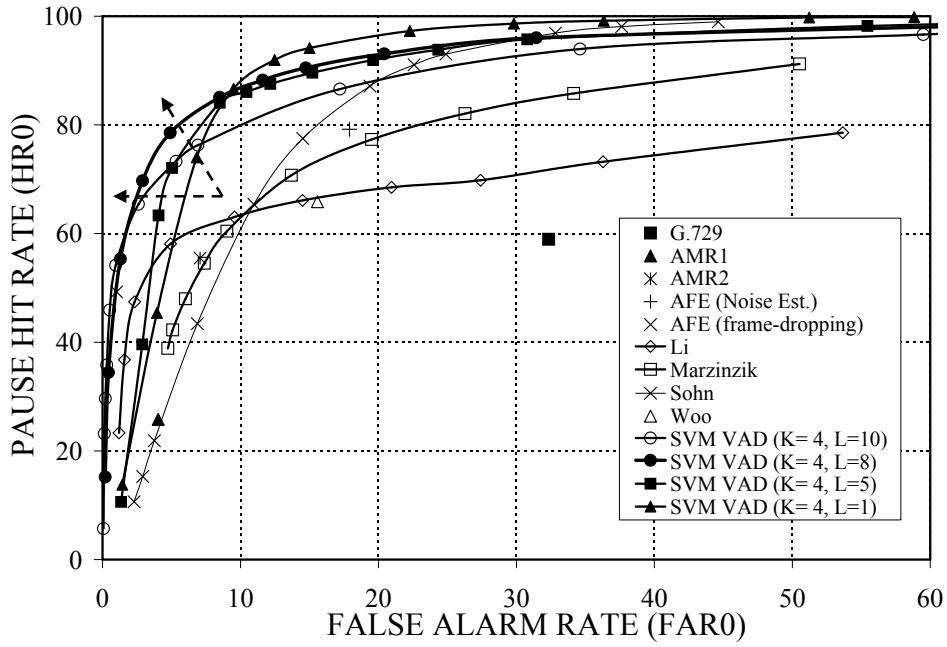


Fig. 6.11: Influencia de la longitud de ventana L en las curvas ROC. Comparición con los VAD estándar y métodos recientes VAD (Alto nivel de ruido: alta velocidad en buena carretera a 5 dB de promedio de SNR).

de las que derivamos una función de decisión no-lineal en el espacio de entrada definida en términos de SNRs en subbandas. Con esta y otras innovaciones el VAD-SVM se muestra más efectivo que los VADs que definen su decisión en términos de los valores promedio de SNR. Se muestra cómo las clases de voz y silencio pueden distinguirse claramente en el espacio 3-D y cómo el modelo SVM aprende la forma en que la señal de voz se enmascara en ruido. Por otro lado aumentando el número de subbandas hasta cuatro mejora el rendimiento del VAD propuesto con un desplazamiento de la curva ROC en el espacio ROC al igual que al aumentar el número de observaciones de la ventana L. Finalmente, múltiples experimentos realizados sobre la base de datos en español SpeechDat-Car mostraron que nuestro VAD mejora a los VADs estándar ITU G.729, ETSI AMR1 y AMR2 y ETSI AFE así como a los recientemente publicados para detección de actividad de voz. Podemos concluir que los experimentos de reconocimiento de voz serán satisfactorios dada la capacidad de detección del método.

7. DISCUSIÓN DE LOS RESULTADOS EXPERIMENTALES

En este Capítulo vamos a comentar los resultados experimentales más importantes discutidos anteriormente para los VADs propuestos a lo largo de los capítulos 3, 4, 5 y 6. Primeramente compararemos los VADs presentados en los dos primeros capítulos, basados en la función biespectral y distintas aproximaciones a ella, para seleccionar el más representativo en cuanto a compromiso entre precisión en la detección y coste computacional. Éste es sin duda el llamado VAD IBI-MO-LRT. A continuación compararemos el anterior VAD con el basado en la heurística clustering que fue ampliamente discutido en el Capítulo 5, el VAD LTCM. Discutiremos, de manera cualitativa y a la vista de los resultados experimentales de los capítulos anteriores, las ventajas e inconvenientes de cada uno de ellos en cuanto a su compromiso rendimiento-coste computacional, cuando son aplicados a reconocimiento robusto de voz, el entorno para los que fueron diseñados.

7.1. Comparativa de los VADs basados en Biespectro

A la vista de los resultados obtenidos en los capítulos 3 y 4 podemos concluir que aunque ambos métodos presentan una precisión elevada en la detección, el IBI-MO-LRT es superior a los VADs basados en biespectro en compromiso entre coste computacional y precisión. De la figuras 3.6 y 4.9b), observamos que en igualdad de condiciones (sin uso de etapa previa de filtrado) ambos VADs presentan una precisión en la detección equivalente para una ventana de observación similar. La diferencia entre estos VADs puede concretarse en:

- El VAD GLRT basado en biespectro presenta un coste computacional muy elevado no siendo práctico para aplicaciones en tiempo real. Aunque usa la magnitud biespectral en la formulación del test GLRT, el uso de la aproximación de "independencia estadística de las distribuciones biespectrales" equipara su rendimiento al uso del promedio biespectral integrado en el IBI-MO-LRT VAD.

- El VAD IBI-MO basa su capacidad en la detección en un promedio integrado del biespectro usando LRT con cálculos rigurosos de la varianza de estimadores, lo que añadido a su implementación recursiva lo hace idóneo para aplicaciones en tiempo real.

La curvas ROC de la figura 4.9 manifiestan la igual precisión de los detectores basados en promedio de bloques y los basados en múltiple observación del biespectro integrado. Esa igualdad se pone de manifiesto cuando comparamos los experimentos de reconocimiento sobre Aurora3 en las Tablas 4.2 y 4.5, donde la tasa de acierto de palabra es equivalente para las dos estrategias (la pequeña diferencia viene del uso de un número distinto de datos y de la selección del umbral). Sin embargo la diferencia esencial entre ellos es

la misma que la anteriormente señalada: el coste computacional. *Por lo que nos decantamos por el VAD IBI-MO-LRT para detección y reconocimiento robusto de voz.*

7.2. Comparativa de los VADs basados en Clustering y SVMs

Estos dos VADs suponen dos nuevas filosofías para la detección de actividad de voz bien distintas. La primera está diseñada para, tras un tiempo de inicialización del modelo de ruido, trabajar en tiempo real y de manera secuencial y adaptativa, filosofía que comparte con los VADs diseñados en los Capítulos 3 y 4. La segunda estrategia presenta una etapa inicial de entrenamiento por lotes o "batch" tras la cual el VAD actúa sin modificar sus parámetros en toda la base de datos. Es lo que se conoce como el procedimiento "estimate&plugg". Este procedimiento tal y como se ha comentado en el Capítulo 6 puede ser optimizado sin más que introducir una versión adaptativa al algoritmo ó modificar el proceso de entrenamiento, lo cual formará parte de las líneas futuras de este trabajo. En cuanto a la calidad de detección de los algoritmos es similar como vemos de las curvas ROC en las figuras 5.14 y 6.11. Los experimentos de reconocimiento de VAD basado en SVM están en camino, aunque de la precisión mostrada en detección sugiere que se obtendrán unos resultados similares a los proporcionados por el algoritmo LTCM en reconocimiento de voz. Los resultados para este último ese muestran en las tablas 5.2, 5.3 y 5.4, resultados que usaremos para establecer la ultima comparativa.

7.3. Comparativa de los VADs basados en Biespectro y Clustering

Por último, en esta sección, vamos a comparar los VADs basados en estadísticos de alto orden (HOS) con un representante de aquellos basados en estrategias heurísticas, dentro del "soft-computing". Como vemos las aproximaciones son bien distintas, así como el nivel de complejidad conceptual. De las curvas ROC, figuras 4.9b) y 5.14, observamos como el VAD LTCM presenta una precisión en detección ligeramente inferior al VAD basado en IBI-MO, aunque uno podría preguntarse hasta qué punto, esa diferencia justificaría usar uno frente al otro. La respuesta a tal pregunta dependerá de la aplicación en la que se usen los VADs en cuestión. Los experimentos de reconocimiento sobre la base de datos Aurora 3 en español ponen de manifiesto tal diferencia en detección, apreciándose una tasa de reconocimiento de palabra superior en el VAD IBI-MO (véase tabla 4.5) sobre el VAD LTCM (tabla 5.4).

8. CONCLUSIONES Y PRINCIPALES APORTACIONES

En este Capítulo presentamos las conclusiones y líneas futuras de este trabajo. En este trabajo hemos presentado un conjunto de VADs basados en nuevas filosofías para detección y su aplicación a reconocimiento robusto del habla. Podemos concluir que el conjunto de los detectores desarrollados es óptimo para aplicaciones de reconocimiento en comparación con los VADs estándar, como el ITU G.729 y GSM AMR para transmisión discontinua y ETSI AFE para reconocimiento distribuido de voz (DSR) y los recientemente publicados [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999]

8.1. Conclusiones y Líneas Futuras

De todos los métodos propuestos y experimentos en detección y reconocimiento de voz podemos concluir que:

1. En los capítulos 3 y 4 se muestran esquemas diferentes para la mejora de la robustez de los detectores de actividad de voz y del rendimiento de los sistemas de reconocimiento de voz en entornos ruidos. Estos métodos están basados en tests estadísticos de cociente de probabilidades definidos sobre el biespectro y su versión integrada de la señal que a su vez está definido como un espectro cruzado entre la señal y su cuadrado, mostrando la capacidad de los estadísticos de orden superior para la detección de señales en ruido. Además estos últimos presentan otras ventajas adicionales: *i*) su cálculo como espectro cruzado conduce a un ahorro computacional significativo, y *ii*) la varianza del estimador es del mismo orden que la del estimador del espectro de potencia. Los métodos propuestos incorporan información contextual a la regla de decisión, una estrategia que ha proporcionado mejoras significativas en la precisión de los detectores de voz y en distintas aplicaciones en sistemas de reconocimiento.

   Las diferentes propuestas se distinguen en la manera en que la ventana de observación se construye o, dicho de otro modo, se procesa la señal para obtener estimadores precisos del biespectro ó biespectro integrado y su varianza. La ventana óptima fue determinada analizando el solapamiento entre las distribuciones de la variable de decisión y la tasa de error de un clasificador de Bayes óptimo. El análisis experimental llevado a cabo sobre la conocida base de datos AURORA ha proporcionado mejoras significativas sobre las técnicas estándares como ITU G.729,

AMR1, AMR2 y ESTI AFE VADs, así como respecto a los mejores VADs recientemente publicados. El análisis valoró: *i*) la precisión en detección de los segmentos de voz/silencio mediante el uso de las curvas ROC obteniendo que nuestro VAD proporcionaba mayores tasas de acierto y menores falsas alarmas cuando lo comparábamos a todos los algoritmos de referencia, y *ii*) la tasa de reconocimiento cuando se consideraba como parte de un sistema global de reconocimiento mostrando un mayor rendimiento en reconocimiento de voz.

2. En el Capítulo 5 hemos mostrado un nuevo algoritmo para mejorar la detección de voz y el reconocimiento del habla en ambientes ruidosos. El VAD propuesto LTCM se basa en el modelado del espacio de ruido usando la técnica Clustering Hard C-medias y emplea la información de largo término para la formulación de una regla de decisión basada en un cociente de energía promedio. El VAD realiza una detección avanzada de los comienzos y retrasada de los finales de palabra lo cual, en parte, evita tener que incluir esquemas adicionales de "hangover" o de bloques de reducción de ruido. Se mostró que incrementando la longitud de la ventana de información de retardo largo se obtiene una reducción significativa de los errores de clasificación. Hemos ralizado un análisis exhaustivo sobre las bases de datos de AURORA mostrando la efectividad de esta aproximación. El VAD propuesto LTCM supera a los recientemente publicados como el de Sohn, que define un test de cociente de probabilidades sobre una única ventana de observación, y los estandarizados ITU-T G.729, ETSI AMR para el sistema GSM y los VADs ETSI AFE para reconocimiento de voz distribuido. Finalmente, también mejoró la tasa de reconocimiento cuando el VAD se usó para estimación del espectro de potencia, reducción de ruido y

frame-dropping en un sistema robusto al ruido ASR.

3. Finalmente, en el Capítulo 6, se ha propuesto un esquema eficiente de modelado basado en SVM para la detección de la presencia de voz en una señal ruidosa. La estrategia combina técnicas de reducción de espectro de ruido y máquinas de vectores soporte para aprendizaje de las que derivamos una función de decisión no-lineal en el espacio de entrada definida en términos de SNRs en subbandas. Con esta y otras innovaciones el VAD-SVM se muestra más efectivo que los VADs que definen su decisión en términos de los valores promedio de SNR. Se muestra cómo las clases de voz y silencio pueden distinguirse claramente en el espacio 3-D y cómo el modelo SVM aprende la forma en que la señal de voz se enmascara por el ruido. Por otro lado aumentando el número de subbandas hasta cuatro mejora el rendimiento del VAD propuesto con un desplazamiento de la curva ROC en el espacio ROC al igual que al aumentar el número de observaciones de la ventana. Realizando múltiples experimentos sobre la base de datos en español SpeechDat-Car hemos mostrado que nuestro VAD mejora a los VADs estándar ITU G.729, ETSI AMR1 y AMR2 y ETSI AFE así como a los recientemente publicados para detección de actividad de voz. Podemos concluir que los experimentos de reconocimiento de voz serán satisfactorios dados la capacidad de detección del método. El trabajo en esta parte esta realizándose en la actualidad y conforma una buena parte de las líneas futuras en el campo de la detección de Actividad de voz.

Los próximos trabajos en VAD irán encaminados a perfeccionar las técnicas presentadas en esta memoria:

- Por un lado, seguiremos perfeccionando el proceso de aprendizaje de

SVM, dado que el número de patrones usados en el entrenamiento así como los parámetros del modelo podrían no ser óptimos y una nueva combinación de éstos podrían aportar mejores resultados en detección. Cabe la posibilidad de aplicar los conceptos estudiados en [Górriz et al., 2006e] al problema de clasificación, para construir un VAD secuencial basado en la Teoría de Regularización equivalente teóricamente a SVM pero con carácter adaptativo (sin entrenamiento). De esta forma los SVs podrían actualizar sus parámetros en tiempo real lo cual es necesario cuando tratamos con señales no estacionarias, con las que un proceso de entrenamiento inicial no captaría la naturaleza dinámica de los datos (recuerde que las SVMs aplicadas aquí aprendían exclusivamente en la fase de extracción de características)

- Queda abierta la línea de investigación consistente en el uso de la fase biespectral (o del biespectro integrado) para la detección de señales de voz, las cuales presentan una fase en esencia distinta al ruido. Podrían derivarse reglas de decisión robustas frente al ruido. En el presente trabajo nos hemos limitado al uso de la magnitud biespectral.

- El cuadrado de una señal presenta características muy interesantes de inmunidad frente al ruido gaussiano. Aprovechando el desarrollo teórico del Apéndice B, se podría derivar una extensión del algoritmo de Sohn con inmunidad extra al ruido gaussiano y con varianza espectral muy parecida a la del espectro de señal.

En definitiva se estudiarán nuevas características de las señales que presenten mayor inmunidad frente al ruido y otros tests estadísticos derivados de estas características para la formulación de reglas de decisión robustas y suaves con un alto poder discriminativo.

Finalmente, consideramos modestamente que el trabajo de investigación presentado en esta memoria constituye una aportación innovadora, eficiente y robusta en un campo de importante aplicación en procesado de señal, como es el de la Detección de Actividad de Voz y el del Reconocimiento Robusto del Habla.

9. CONCLUSIONS AND MAIN CONTRIBUTIONS

In the following Chapter we state some conclusions and remark the forthcoming research lines in this scenario. In a nutshell, we have shown a set of efficient VADs based on several new approaches for voice activity detection and their application to distributed speech recognition. The report concludes that the proposed VADs are optimal in speech recognition experiments when they are compared to the standard VADs, such us ITU G.729 and GSM AMR VAD for discontinuous transmission and ETSI AFE VAD for distributed speech recognition (DSR); and the most representative set of recently reported VADs [Woo et al., 2000; Li et al., 2002; Marzinzik and Kollmeier, 2002; Sohn et al., 1999]

9.1. Conclusions and Research Lines

Based on the set of proposed algorithms presented and detection and recognition experiments carried on in chapters 3 45 and 6, we can conclude that:

1. Chapters 3 y 4 showed two different schemes for improving speech detection robustness and the performance of speech recognition systems in noisy environments. Both methods are based on statistical likelihood ratio tests defined on the integrated bispectrum of the signal which is defined as a cross spectrum between the signal and its square and inherits the ability of higher order statistics to detect signals in noise with many other additional advantages: *i*) its computation as a cross spectrum leads to significant computational savings, and *ii*) the variance of the estimator is of the same order as that of the power spectrum estimator. The proposed methods incorporate contextual information to the decision rule, a strategy that has reported significant improvements in speech detection accuracy and robust speech recognition applications. They differ in the way the window is arranged or the signal is processed in order to obtain precise estimations of the integrated bispectrum and its variance or smoothed likelihood ratio tests. The optimal window size was determined by analyzing the overlap between the distributions of the decision variable and the error rate of an optimum Bayes classifier. The experimental analysis conducted on the well-known AURORA databases has reported significant improvements over standardized techniques such as ITU G.729, AMR1, AMR2 and ESTI AFE VADs, as well as over recently published VADs. The analysis assessed: *i*) the speech/non-speech detection accuracy by

means of the ROC curves with the proposed VAD yielding improved hit-rates and reduced false alarms when compared to all the reference algorithms, and *ii*) the recognition rate when it was considered as part of a complete speech recognition system showing a sustained advantage in speech recognition performance.

2. In chapter 5 a new algorithm for improving speech detection and speech recognition robustness in noisy environments is shown. The proposed LTCM VAD is based on noise modeling using hard C-means clustering and employs long-term speech information for the formulation of a soft decision rule based on an averaged energy ratio. The VAD performs an advanced detection of beginnings and delayed detection of word endings which, in part, avoids having to include additional hangover schemes or noise reduction blocks. It was found that increasing the length of the long-term window yields to a reduction of the class distributions and leads to a significant reduction of the classification error. An exhaustive analysis conducted on the AURORA database showed the effectiveness of this approach. The proposed LTCM VAD outperformed recently reported VAD methods including Sohn's VAD, that defines a likelihood ratio test on a single observation, and the standardized ITU-T G.729, ETSI AMR for the GSM system and ETSI AFE VADs for distributed speech recognition. On the other hand, it also improved the recognition rate when the VAD is used for noise spectrum estimation, noise reduction and frame-dropping in a noise robust ASR system.

3. Finally, in chapter 6, an effective algorithm for detecting presence of speech in a noisy signal is proposed in this paper. The proposed strategy combines spectral noise reduction techniques and support vector

machine learning tools to derive a non-linear decision rule in the input space defined in terms of the subbands SNRs. With these and other innovations the proposed method has shown to be more effective than VADs that define the decision rule in terms of average SNR values. It is shown that the non-speech and speech classes can be clearly distinguished in the 3-D space and that the SVM model learns how the signal is masked by the noise. On the other hand, increasing the number of subbands up to four improves the performance of the proposed VAD by shifting the ROC curve in the ROC space in the same way as when we increase the number of observations in the decision function L. In the last part of the chapter, the experiments conducted on the Spanish SpeechDat-Car database showed that the proposed algorithm outperforms ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards as well as other recently reported VAD methods in speech/non-speech detection performance.

The next methods in VAD will be focused on optimizing the proposed methods presented in this work:

- On one hand, we will complete the SVM learning process, provided that the number of patterns and the model parameters used in the training stage could not be the optimum ones and a different selection of them could achieve better results in the discrimination of pause/speech frames. It could be possible to apply the results achieved in [Górriz, 2003] to speech/pause classification in order to build a sequential Regularization theory-based VAD equivalent to SVM in accuracy but adaptive in nature (without training stage). In this way, SV parameters would be necessarily on-line updated since signals could be non-stationary. In this case, the use of an initial training process could

not learn the dynamics of the input data (note that , in this work, SVM learning was applied in feature extraction process exclusively)

- We haven´t apply an important features such us bispectrum phase (or integrated bispectrum phase) to VAD. The speech bispectrum phase is essentially different from the noise bispectrum phase, thus we could formulate robust decision rules for improving the accuracy and performance of DSR systems in term of this feature. In the present work, we confine ourselves to the use of bispectrum magnitude.

- The "squared" observation signal (clean signal in gaussian noise) usually shows interesting features and properties in terms of HOS. Making use of Appendix B, it could be possible to derive an extension of Sohn's algorithm with interesting properties such us gaussian noise robustness and the variance of the estimator with the same order as that of the power spectrum estimator.

In short, in the future works, we will research on new robust signal features in noisy environments and the application of statistical tests over these new features to formulate smooth and robust decision rules.

Finally, we believe that the research work, presented in this dissertation, constitutes a novel, efficient and robust contribution, in one of the most important application fields in signal processing such us Voice Activity Detection and Robust Speech Recognition.

APÉNDICES

En esta Capítulo listamos todos los trabajos publicados en revistas de impacto, referenciados por el ISI Web of Knowledge en el campo de la Detección de Actividad de Voz y su principal aplicación en Reconocimiento Robusto del Habla y detallamos algunos desarrollos teóricos de los anteriores Capítulos

A. PUBLICACIONES

A.1. Trabajos publicados en el contexto de la Tesis

A continuación destacamos los trabajos más importantes, con índice de impacto, realizados en el contexto de la presente tesis doctoral:

1. **J.M. Górriz**, J. Ramirez, J.C. Segura, S. Hornillo. "Voice Activity Detection based on Higher Order Statistics". Lecture Notes in Computer Science. Vol 3512 Springer 2005, pp 837-844 ISBN 3-540-26208-3.

2. **J.M. Górriz**, J. Ramírez, J.C. Segura, C.G. Puntonet. "Bispectra Analysis-based VAD for Robust Speech Recognition". Lecture Notes in Computer Science. Vol 3562 Springer 2005, pp 567-576 ISBN 3-540-26319-5.

3. **J.M. Górriz**, J. Ramírez, C.G. Puntonet, J.C. Segura. "An Improved MO-LRT VAD Based on a Bispectra Gaussian Model" IEE Electronic Letters Journal Vol. 41 Issue 15 pp 877-879 Jul., 2005.

4. **J.M. Górriz**, J. Ramírez, C.G. Puntonet, F Theis, E. Lang. "Bispectrum-Based Statistical Tests for VAD". Lecture Notes in Computer Science. Vol 3697 Springer 2005, pp 541-546 ISBN 3-540-28755-8

5. **J.M. Górriz**, C.G. Puntonet, J. Ramírez, and J.C. Segura. "Bispectrum Estimators for Voice Activity Detection and Speech Recognition"

Lecture Notes in Artificial Intelligence. Vol 3817 / Springer 2006. pp 174-185 ISBN: 3-540-31257-9.

6. J. Ramírez, P. Yélamos, **J.M. Górriz**, C.G. Puntonet and J.C. Segura: "SVM-enabled Voice Activity Detection". Lecture Notes in Computer Science. Vol 3972, pp. 676-681, Springer (2006).

7. **J.M. Górriz**, J. Ramírez, C.G. Puntonet, and J.C. Segura. "Generalized LRT-based Voice Activity Detector" Accepted for publication in IEEE Signal Processing Letters, 2006.

8. J. Ramírez , **J.M. Górriz**, J. C. Segura, C. G. Puntonet, A. Rubio "Speech/Non-speech Discrimination based on Contextual Information Integrated Bispectrum LRT", IEEE Signal Processing Letters. Aceptado Enero 2006.

9. **J.M. Górriz**, J. Ramírez, C.G. Puntonet, E. Lang, K. Stadlthanner. "Independent Component Analysis applied to Voice Activity Detection" Lecture Notes in Computer Science. Vol 3991, pp. 234-241 Springer (2006)

10. **J.M. Górriz**, J. Ramírez, I. Turias, C.G. Puntonet, J. González E. Lang. "C-means Clustering applied to Speech Discrimination" Lecture Notes in Computer Science. Vol 3991, pp. 649-656, Springer (2006).

11. P. Yélamos, J. Ramírez, **J.M. Górriz**, C.G. Puntonet, J.C. Segura. "Speech Event Detection Using Support Vector Machines" Lecture Notes in Computer Science. Vol 3991, pp. 356-363, Springer (2006).

12. R. Culebras, J. Ramírez, **J.M. Górriz**, J.C. Segura. "Fuzzy Logic Speech/Non-speech Discrimination for Noise Robust Speech Process-

ing" Lecture Notes in Computer Science. Vol 3991, pp. 395-402, Springer (2006).

13. **J.M. Górriz**, J. Ramírez, J.C. Segura, C. G. Puntonet, J.J.G. de la Rosa. "Noise Subspace Fuzzy C-means Clustering for Robust Speech Recognition" Lecture Notes in Computer Science. Vol 3984, pp. 772-779, Springer (2006).

14. J. Ramírez, P. Yélamos, **J.M. Górriz**, J.C. Segura. "SVM-based Speech Endpoint Detection Using Contextual Speech Features" Electronic Letters Vol 42 num. 7 pp 65-66 (2006).

15. **J.M. Górriz**, J. Ramírez, C.G. Puntonet, J.C. Segura "An effective cluster-based model for robust speech detection and speech recognition in noisy environments", The Journal of Acoustical Society of America. Aceptado Marzo 2006.

B. DESARROLLOS

B.1. Cálculo de las varianzas del biespectro integrado

Asumimos en este apartado que la señal limpia $s(t)$ y el ruido $n(t)$ son estacionarios, están centrados ($E[s(t)] = E[n(t)] = 0$) y son estadísticamente independientes. Por lo tanto tenemos las señales están decorreladas:

$$r_x(k) = r_s(k) + r_n(k) \Rightarrow S_{xx}(\omega) = S_{ss}(\omega) + S_{nn}(\omega) \tag{B.1}$$

donde $r_x(k)$ denota la función de correlación de la señal $x(t) = s(t) + n(t)$.

Para el cálculo de las varianzas del biespectro integrado en las hipótesis nula y alternativa, tenemos que evaluar primeramente la secuencia de correlación de la variable $y(t)$, cuadrado de x(t). Se derivarán las expresiones para H_1 que son validas para H_0 sin más que anular la componente de señal limpia $s(t)$. La secuencia de correlación de $y(t) = x^2(t) - E[x^2(t)]$ se define como:

$$r_{yy}(k) = E[y(t)y(t+k)] = E[(x^2(t) - E[x^2(t)])(x^2(t+k) - E[x^2(t+k)])] \tag{B.2}$$

Definiendo $\sigma_x^2 = E[x^2(t)] = E[x^2(t+k)]$ tenemos que:

$$r_{yy}(k) = E[(x^2(t)x^2(t+k)] - \sigma_x^4 \tag{B.3}$$

El segundo término de la parte derecha de B.3 se puede expresar como[1]:

$$\sigma_x^4 = E^2[(s(t)+n(t))^2] = (\sigma_s^2 + \sigma_n^2 + 2\overbrace{E[s(t)n(t)]}^{0})^2 = \sigma_s^4 + \sigma_n^4 + 2\sigma_s^2\sigma_n^2 \tag{B.4}$$

[1] $E[s(t)n(t)] = E[s(t)]E[n(t)] = 0$

El primer término de la parte derecha de la ecuación B.3, definiendo $\bar{y}(t) \equiv x^2(t)$, se puede expresar como:

$$r_{\bar{y}\bar{y}}(k) \equiv E[(s(t)+n(t))^2(s(t+k)+n(t+k))^2] = E[s^2(t)s^2(t+k)]+ \\ E[n^2(t)n^2(t+k)] + E[s^2(t)n^2(t+k)] + 2E[s^2(t)n(t+k)s(t+k)] \\ +E[n^2(t)s^2(t+k)] + 2E[n^2(t)n(t+k)s(t+k)] + 2E[n(t)s(t)s^2(t+k)] \\ +2E[n(t)s(t)n^2(t+k)] + 4E[n(t)s(t)n(t+k)s(t+k)] \tag{B.5}$$

donde podemos notar $r_{s^2s^2} = E[s^2(t)s^2(t+k)]$ y $r_{n^2n^2}$.

Si utilizamos la relación entre momentos y cumulantes de cuarto orden dada por [Nikias and Petropulu, 1993]] para un conjunto de variables aleatorias:

$$C_{x_1,\ldots,x_n} = \sum_{p_1,\ldots,p_m} (-1)^{m-1}(m-1)! \cdot E[\prod_{i\in p_1} X_i]\ldots E[\prod_{i\in p_m} X_i] \tag{B.6}$$

donde $\{p_1,\ldots,p_m\}$ son todas las particiones con $m = 1,\ldots,r$ del conjunto de enteros $\{1,\ldots,r\}$, y particularizamos para cuarto orden tenemos que:

$$C_{x_1,x_2,x_3,x_4} = (-1)^0(0!)E[x_1x_2x_3x_4] + (-1)^1(1!)[E[x_1x_2]E[x_3x_4] + E[x_1x_4]E[x_2x_3] \\ +E[x_1x_3]E[x_2x_4] + E[x_1]E[x_2x_3x_4] + E[x_2]E[x_1x_3x_4] + E[x_3]E[x_1x_2x_4] \\ +E[x_4]E[x_1x_2x_3]] + (-1)^2(2!)[E[x_1]E[x_2]E[x_3x_4] + E[x_1]E[x_3]E[x_2x_4] \\ +E[x_1]E[x_4]E[x_2x_3] + E[x_2]E[x_3]E[x_1x_4] + E[x_2]E[x_4]E[x_1x_3] + E[x_3]E[x_4]E[x_1x_2]] \\ +(-1)^3(3!)E[x_1]E[x_2]E[x_3]E[x_4] \tag{B.7}$$

Asumiendo que las variables aleatorias están centradas (media nula), esta expresión puede simplificarse significativamente:

$$C_{x_1,x_2,x_3,x_4} = E[x_1x_2x_3x_4]-[E[x_1x_2]E[x_3x_4]+E[x_1x_4]E[x_2x_3]+E[x_1x_3]E[x_2x_4]] \tag{B.8}$$

Dado que los cumulantes cruzados de dos señales estadísticamente independientes son nulos y que los momentos conjuntos son iguales al productorio de

momentos, tenemos que los términos cruzados de la ecuación B.5 se pueden simplificar como[2]:

$$
\begin{aligned}
&E[s^2(t)n^2(t+k)] = \overbrace{C_{s,s,n,n} + 2E[s(t)n(t+k)]E[s(t)n(t+k)]}^{0} + E[s^2(t)]E[n^2(t+k)] \\
&2E[s^2(t)n(t+k)s(t+k)] = \\
&2(\overbrace{C_{s,s,n,s} + 2E[s(t)n(t+k)]E[s(t)s(t+k)] + E[s^2(t)]E[s(t+k)n(t+k)]}^{0}) \\
&E[n^2(t)s^2(t+k)] = \overbrace{C_{n,n,s,s} + 2E[n(t)s(t+k)]E[n(t)s(t+k)]}^{0} + E[n^2(t)]E[s^2(t+k)] \\
&2E[n^2(t)n(t+k)s(t+k)] = \\
&2(\overbrace{C_{n,n,n,s} + 2E[n(t)n(t+k)]E[n(t)s(t+k)] + E[n^2(t)]E[s(t+k)n(t+k)]}^{0}) \\
&2E[n(t)s(t)s^2(t+k)] = \\
&2(\overbrace{C_{n,s,s,s} + 2E[n(t)s(t+k)]E[s(t)s(t+k)] + E[n(t)s(t)]E[s^2(t+k)]}^{0}) \\
&2E[n(t)s(t)n^2(t+k)] = \\
&2(\overbrace{C_{n,s,n,n} + 2E[n(t)n(t+k)]E[s(t)n(t+k)] + E[n(t)s(t)]E[n^2(t+k)]}^{0}) \\
&4E[n(t)s(t)n(t+k)s(t+k)] = \\
&4(\overbrace{C_{n,s,n,s} + E[n(t)n(t+k)]E[s(t)s(t+k)] + E[n(t)s(t+k)]E[s(t)n(t+k)]}^{0} \\
&+E[s(t)s(t+k)]E[n(t)n(t+k)])
\end{aligned}
\tag{B.9}
$$

Finalmente tenemos entonces que la ecuación B.5 se puede escribir como:

$$r_{\bar{y}\bar{y}}(k) = r_{s^2s^2}(k) + r_{n^2n^2}(k) + 2\sigma_s^2\sigma_n^2 + r_s(k)r_n(k) \tag{B.10}$$

y por tanto la ecuación B.3 puede expresarse como:

$$r_{yy}(k) = r_{s^2s^2}(k) + r_{n^2n^2}(k) + 4r_s(k)r_n(k) - (\sigma_s^4 + \sigma_n^4) \tag{B.11}$$

Sin más que calcular la transformada discreta de Fourier en esta expresión llegamos a una expresión "lineal" para el espectro de la señal centrada $y(t)$:

$$S_{yy}(\omega) = S_{s^2s^2}(\omega) + S_{n^2n^2}(\omega) + 4S_{ss}(\omega) * S_{nn}(\omega) - 2\pi(\sigma_s^4 + \sigma_n^4)\delta_\omega \tag{B.12}$$

[2] Recuerde que las señales $s(t)$y $n(t)$ están centradas!

donde "$*$" denota convolución en el dominio transformado. Para el término en $n(t)$ no cruzado de la ecuación B.5 (equivalentemente para $s(t)$), asumiendo que es un proceso gaussiano o de cumulante de cuarto orden despreciable frente a su varianza tenemos que, usando la ecuación B.8:

$$E[n^2(t)n^2(t+k)] = \overbrace{C_{n,n,n,n}}^{0} + 2\overbrace{E[n(t)n(t+k)]E[n(t)n(t+k)]}^{r_n(k)r_n(k)} + \overbrace{E[n^2(t)]E[n^2(t+k)]}^{\sigma_n^4} \tag{B.13}$$

o lo que es equivalente:

$$S_{n^2n^2}(\omega) = 2S_{nn}(\omega) * S_{nn}(\omega) + 2\pi\sigma_n^4\delta(\omega) \tag{B.14}$$

Con las expresiones deducidas se pueden derivar completamente las varianzas del biespectro integrado en H_1 y H_0 en las ecuaciones 4.18 y completar así los algoritmos de detección basados en estas propiedades.

BIBLIOGRAFÍA

Anderberg, M. R., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1973). *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY.

Arce, J., Gallagher, N., and T.Nodes (1986). *Advances in Computer Vision and Image Processing*, chapter Median filters: theory and applications. JAI Press, Connecticut.

Benyassine, A., Shlomot, E., Su, H., Massaloux, D., Lamblin, C., and Petit, J. (1997). ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, 35(9):64–73.

Beritelli, F., Casale, S., and Cavallaro, A. (1998). A robust voice activity detector for wireless communications using sof computing. *IEEE Journal of Selected Areas in Communications*, 16(9):1818–1829.

Beritelli, F., Casale, S., Rugeri, G., and Serrano, S. (2002). Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors. *IEEE Signal Processing Letters*, 9(3):85–88.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, pages 208–211.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27:113–120.

Bouquin-Jeannes, R. L. and Faucon, G. (1994). Proposal of a voice activity detector for noise reduction. *Electronics Letters*, 30(12):930–932.

Bouquin-Jeannes, R. L. and Faucon, G. (1995a). Study of a voice activity detector and its influence on a noise reduction system. *Speech Communication*, 16:245–254.

Bouquin-Jeannes, R. L. and Faucon, G. (1995b). Study of voice activity detector and its influence on a noise reduction system. *Speech Communication*, 16:245–254.

Brillinger, D. (1975). *Time series data analysis and theory.* New York: Holt, Rinehart and Winston.

Brillinger, D. and Rossenblatt, M. (1975). *Spectral Analysis of Time Series*, chapter Asymptotic theory of estimates of kth order spectra. Wiley.

Brillinger, D. R. and Rosenblatt, M. (1968). *Spectral Analysis of Time Series*, chapter Computation and interpretation of k-th order spectra. Wiley. New York.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-8:679–698.

Chang, C. and Lin, C. J. (2001). LIBSVM: a library for support vector machines. Technical report, Dept. of Computer Science and Information Engineering, National Taiwan University.

Chengalvarayan, R. (1999). Robust energy normalization using speech/non-speech discriminator for German connected digit recognition. In *Proc. EUROSPEECH*, pages 61–64.

Cho, Y. D., Al-Naimi, K., and Kondoz, A. (2001a). Analysis and improvement of a statistical model-based voice activity detector. *IEEE Signal Processing Letters*, 8(10):276–278.

Cho, Y. D., Al-Naimi, K., and Kondoz, A. (2001b). Improved voice activity detection based on a smoothed statistical likelihood ratio. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 737–740.

Cho, Y. D., Al-Naimi, K., and Kondoz, A. (2001c). Mixed decision-based noise adaptation for speech enhancement. *Electronics Letters*, 37(8):540–542.

Clarkson, P. and Moreno, P. (1999). On the use of support vector machines for phonetic classification. In *Proc. of the IEEE Int. Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 585–588.

Cortes, C. and Vapnik, V. (1995). *Support-vector network*. Machine Learning.

Cox, B. V. and Tinothy, L. K. (1980). Nonparametric rank-order statistics applied to robust voiced-unvoiced-silence classification. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-28(5):550–561.

Culebras, R., Ramírez, J., Górriz, J. M., and Segura, J. C. (2006). Fuzzy logic speech/non-speech discrimination for noise robust speech processing. *Lecture Notes in Computer Science. Springer.*, 3991:395–402.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-32:1109–1121.

Estévez, P. A., Becerra-Yoma, N., Boric, N., and Ramírez, J. A. (2000). Genetic programming-based voice activity detection. *Electronics Letters*, 41(20):1141–1143.

ETSI (1999). Voice activity detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels. *ETSI EN 301 708 Recommendation.*

ETSI (2000). Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. *ETSI ES 201 108 Recommendation.*

ETSI (2002). Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. *ETSI ES 201 108 Recommendation.*

Fisher, D. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 1(2):139–172.

Freeman, D. K., Cosier, G., Southcott, C. B., and Boyd, I. (1989). The voice activity detector for the pan-european digital cellular mobile telephone service. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, pages 369–372.

Furui, S. (2001). *Digital Speech Processing, Synthesis and Recogniztion.* Marcel Dekker, Inc.,New York, ISBN 0 8247 0452 5.

Ganapathiraju, A., Hamaker, J., and Picone, J. (2004). Applications of sup-

port vector machines to speech recognition. *IEEE Transactions on Signal Processing*, 52(8):2348–2355.

Górriz, J. M. (2003). *Algoritmos Híbridos para el Modelado de Series Temporales con Técnicas AR-ICA*. PhD. Servicio de Reprografía de la Universidad de Cádiz, ISBN: 84-96274-12-8.

Górriz, J. M., Puntonet, C. G., Ramírez, J., and Segura, J. C. (2005a). Bispectrum estimators for voice activity detection and speech recognition. *Lecture Notes in Artificial Intelligence. Springer. ISBN 3-540-31257-9.*, 3817:174–185.

Górriz, J. M., Ramirez, J., Segura, J. C., and Hornillo, S. (2005b). Voice activity detection based on higher order statistics. *Lecture Notes in Computer Science. Springer. ISBN 3-540-26208-3.*, 3512:837–844.

Górriz, J. M., Ramírez, J., Puntonet, C. G., Lang, E., and Stadlthanner, K. (2006a). Independent component analysis applied to voice activity detection. *Lecture Notes in Computer Science. Springer.*, 3991:234–241.

Górriz, J. M., Ramírez, J., Puntonet, C. G., and Segura, J. C. (2006b). An effective cluster-based model for robust speech detection and speech recognition in noisy environments. *In press in the Journal of Acoustical Society of America*, XX(X):XXX–XXX.

Górriz, J. M., Ramírez, J., Puntonet, C. G., and Segura, J. C. (2006c). Generalized lrt-based voice activity detector. *In press in the IEEE Signal Processing Letters.*

Górriz, J. M., Ramírez, J., Puntonet, C. G., Theis, F., and Lang, E. (2005c). Bispectrum-based statistical tests for vad. *Lecture Notes in Computer Science. Springer. ISBN 3-540-28755-8.*, 3697:541–546.

Górriz, J. M., Ramírez, J., Segura, J. C., and Puntonet, C. G. (2005d). Bispectra analysis-based vad for robust speech recognition. *Lecture Notes in Computer Science. Springer. ISBN 3-540-26319-5.*, 3562:567–576.

Górriz, J. M., Ramírez, J., Segura, J. C., and Puntonet, C. G. (2005e). Improved MO-LRT VAD based on bispectra gaussian model. *Electronics Letters*, 41(15):877–879.

Górriz, J. M., Ramírez, J., Segura, J. C., Puntonet, C. G., and de la Rosa, J. (2006d). Noise subspace fuzzy c-means clustering for robust speech recognition. *Lecture Notes in Computer Science. Springer.*, 3984:772–779.

Górriz, J. M., Ramírez, J., Segura, J. C., Puntonet, C. G., and García, L. (2006e). Effective speech/pause discrimination using an integrated bispectrum likelihood ratio test. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP06)*.

Górriz, J. M., Ramírez, J., Turias, I., Puntonet, C. G., González, J., and Lang, E. (2006f). C-means clustering applied to speech discrimination. *Lecture Notes in Computer Science. Springer.*, 3991:649–656.

Haigh, J. and Mason, J. (1993). Robust voice activity detection using cepstral features. In *IEEE International Conference on on Computers, Communications, Control and Power Engineering*, volume 3, pages 321–324.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction Series*. Springer Series in Statistics 1st ed.

Hinich, J. (1982). Testing for gaussianity and linearity of a stationary time series. *Journal of Time Series Analisys*, 3:169–176.

Hirsch, H. and Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. In *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France.

Huang, L. S. and Yang, C. H. (2000). A novel approach to robust speech endpoint detection in car environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1751–1754.

Hwang, H. and Haddad, R. (1994). Multilevel non-linear filters for edge detection and noise suppression. *IEEE Transactions on Signal Processing*, 42(2):249–258.

Instruments, T. (2001). *Description and baseline results for the subset of the Speech-Dat Car German database used for ETSI STQ WI008 advanced front-end evaluation.*

Itoh, K. and Mizushima, M. (1997). Environmental noise reduction based on speech/non-speech identification for hearing aids. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 419–422.

ITU (1996). A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70. *ITU-T Recommendation G.729-Annex B.*

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data.* Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ.

Jain, A. and Flynn, P. (1996). Image segmentation using clustering. In *In Advances in Image Understanding. A Festschrift for Azriel Rosenfeld, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ.*, pages 65–83.

Karray, L. and Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse environments. *Speech Communitation*, 40(3):261–276.

Kay, S. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, Inc., New Jersey. ISBN 0 13 042268 1.

Kim, H. I. and Park, S. K. (2000). Voice activity detection algorithm using radial basis function network. *Electronics Letters*, 40(22):1454–1455.

Ko, S. and Lee, Y. (1991). Center weighted median filters and their application to image enhancement. *IEEE Transactions on Circuits and Systems*, 38(9):984–993.

Kohonen, T. (1989). *TSelf Organizing and Associative Memory.* Springer-Verlag, Berlin. 3rd ed.

Lee, I., Stern, H., and Mahmoud, S. (2003). A voice activity detection algorithm for communication systems with dynamically varying background acoustic noise. In *IEEE Vehicular Technology Conference*, volume 2, pages 1214–1218.

Li, Q., Zheng, J., Tsai, A., and Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing*, 10(3):146–157.

Liao, X. and Bao, Z. (1998). Circularly integrated bispectra: Novel shift invariant features for high-resolution radar target recognition. *Electronics Letters*, 34(19):1879–1880.

Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer Academic Publishers, Boston,.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press*, pages 281–297.

Markel, J. and Gray, A. (1976). *Linear Prediction of Speech.* Springer-Verlag, Berlin and New York.

Martin, R. (1993). An efficient algorithm to estimate the instantaneous SNR of speech signals. *Proc. of the Eurospeech Conference*, 1:1093–1096.

Marzinzik, M. and Kollmeier, B. (2002). Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing*, 10(6):341–351.

Moreno, A., Borge, L., Christoph, D., Gael, R., Khalid, C., Stephan, E., and Jeffrey, A. (2000). SpeechDat-Car: A Large Speech Database for Automotive Environments. In *Proceedings of the II LREC Conference.*

Nemer, E., Goubran, R., and Mahmoud, S. (2001). Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Trans. Speech and Audio Proc.*, 9(3):217–231.

Nikias, C. and Petropulu, A. (1993). *Higher Order Spectra Analysis: a Nonlinear Signal Proc. Framework.* Prentice Hall.

Nikias, C. and Raghuveer, M. (1987). Bispectrum estimation: A digital signal processing framework. *Proceedings of the IEEE*, 75(7):869–891.

Nokia (2000). *Baseline results for subset of Speech-Dat Car Finnish database for ETSI STQ WI008 advanced front-end evaluation.*

Pitas, I. and Pitas, A. V. (1993). Application of adaptive order statistics filters on digital image/image sequence filtering. *IEEE International Conference on Circuits and Systems*, 2:327–330.

Platt, J. (1999). *Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola Eds.*, chapter Fast Training of Support Vector Machines using Sequential Minimal Optimization, pages 185–208. MIT Press.

Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition*. PTR Prentice Hall, Inc., New Jersey, ISBN 0 13 015157 2.

Rabiner, L. R. and Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. *Bell System Technical Journal*, 54(2):297–315.

Ramírez, J., Górriz, J. M., Segura, J. C., Puntonet, C. G., and Rubio, A. (2006a). Speech/non-speech discrimination based on contextual information integrated bispectrum lrt. *In press in the IEEE Signal Processing Letters*.

Ramírez, J., Segura, J. C., Benítez, C., de la Torre, A., and Rubio, A. (2005a). An effective subband osf-based vad with noise reduction for robust speech recognition. *IEEE Trans. on Speech and Audio Processing*.

Ramírez, J., Segura, J. C., Benítez, C., García, L., and Rubio, A. (2005b). Statistical voice activity detection using a multiple observation likelihood ratio test. *IEEE Signal Processing Letters*, 12(10):689–692.

Ramírez, J., Segura, J. C., Benítez, M. C., de la Torre, A., and Rubio, A.

(2004a). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4):271–287.

Ramírez, J., Segura, J. C., Benítez, M. C., de la Torre, A., and Rubio, A. (2004b). A new Kullback-Leibler VAD for speech recognition in noise. *IEEE Signal Processing Letters*, 11(2):666–669.

Ramírez, J., Yélamos, P., Górriz, J. M., and Segura, J. C. (2006b). Svm-based speech endpoint detection using contextual speech features. *The IEE Electronic Letters*, 42(7):65–66.

Rasmussen, E. (1992). Clustering algorithms. *In Information Retrieval: Data Structures and Algorithms, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ.*, pages 419–442.

Restrepo, A., Hincapié, G., and Parra, A. (1994). On the detection of edges using order statistic filters. In *Proc. IEEE International Image Processing Conference*, volume 1, pages 308–312.

Salton, G. (1991). Developments in automatic text retrieval. *Science*, 1(253):974–980.

Sangwan, A., Chiranth, M. C., Jamadagni, H. S., Sah, R., Prasad, R. V., and Gaurav, V. (2002). VAD techniques for real-time speech transmission on the Internet. In *IEEE International Conference on High-Speed Networks and Multimedia Communications*, pages 46–50.

Sohn, J., Kim, N. S., and Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 16(1):1–3.

Sohn, J. and Su, W. (1998). A voice activity detector employing soft decision based noise spectrum adaptation. In *Proc. of the International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 365–368.

Subba-Rao, T. (1982). A test for linearity of stationary time series. *Journal of Time Series Analisys*, 1:145–158.

Tanyer, S. G. and Özer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on Speech and Audio Processing*, 8(4):478–482.

Öten, R. and de Figueiredo, R. J. P. (2003). An efficient method for L-filter design. *IEEE Transactions on Signal Processing*, 51(1):193–203.

Tucker, R. (1992). Voice activity detection using a periodicity measure. *IEE Proceedings, Communications, Speech and Vision*, 139(4):377–380.

Tugnait, J. (1993). Two channel tests fro common non-gaussian signal detection. *IEE Proceedings-F*, 140:343–349.

Tugnait, J. K. (1994). Detection of non-gaussian signals using integrated polyspectrum. *IEEE Trans. on Signal Processing*, 42(11):3137–3149.

Tugnait, J. K. (1995). Corrections to detection of non-gaussian signals using integrated polyspectrum. *IEEE Trans. on Signal Processing*, 43(11):2792–2793.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York.

Vapnik, V. (1995a). *The Nature of Statistical Learning Theory.* Springer, N.Y.

Vapnik, V. (1995b). *The Nature of Statistical Learning Theory.* Springer-Verlag, Berlin.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons, Inc., New York.

Woo, K., Yang, T., Park, K., and Lee, C. (2000). Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 36(2):180–181.

Yélamos, P., Ramírez, J., Górriz, J. M., Puntonet, C. G., and Segura, J. C. (2006). Speech event detection using support vector machines. *In press Lecture Notes in Computer Science. Springer.*, XX(X):XXX–XXX.

Young, S., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1997). *The HTK Book*. Cambridge University.

Zhang, X., Shi, Y., and Bao, Z. (2001). A new feature vector using selected bispectra for signal classification with application in radar target recognition. *IEEE Transactions on Signal Processing*, 49(9):1875–1885.