UNIVERSIDAD POLITÉCNICA DE MADRID

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS DE TELECOMUNICACIÓN

TESIS DOCTORAL

Técnicas de análisis, caracterización y detección de señales de voz en entornos acústicos adversos

Óscar Varela Serrano

Madrid, 2011

Agradecimientos

Durante estos años son muchas las personas e instituciones que han participado en este trabajo y a quienes quiero expresar mi gratitud por el apoyo y la confianza que me han prestado de forma desinteresada.

Quiero dedicar enteramente este trabajo de Tesis Doctoral a mi querida madre, a la que echo muchísimo de menos desde que no está entre nosotros. También quiero agradecerle su paciencia, su postura y su apoyo cuando me propuse iniciar este largo camino que finalmente ha obtenido su fruto. Gracias Mamá.

Quiero agradecer a mis dos Directores de Tesis, Rubén San-Segundo Hernández y Luís Alfonso Hernández Gómez, su dedicación, su esfuerzo y su enorme paciencia en todo momento. Sin ellos no hubiese sido posible la realización de este emprendedor proyecto. También quiero dar las gracias tanto al departamento de Señales, Sistemas y Radiocomunicaciones como al departamento de Electrónica, ambos de la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universidad Politécnica de Madrid, por su acogida, su apoyo y los medios recibidos.

Igualmente me complace agradecer a José Margineda Puigpelat, catedrático de Microondas de la Universidad de Murcia, sus sabios consejos como tutor en mis primeros dos años de Doctorado.

Tampoco me puedo olvidar de todos mis compañeros de Telefónica I+D, que siempre me han ayudado y apoyado en esta larga travesía, y a la mencionada institución, Telefónica I+D, que me ha facilitado muchos de los medios necesarios durante mi labor investigadora.

Para finalizar, también deseo hacer extensivo este agradecimiento a toda mi familia, en especial a mi padre y a mi mujer que, en las situaciones más complicadas, me han animado a continuar, y, cómo no, a mi pequeña Elenita, siempre con una sonrisa para darme fuerza en el día a día.

RESUMEN

Este trabajo de Tesis ha abordado el objetivo de dar robustez y mejorar la Detección de Actividad de Voz en entornos acústicos adversos con el fin de favorecer el comportamiento de muchas aplicaciones vocales, por ejemplo aplicaciones de telefonía basadas en reconocimiento automático de voz, aplicaciones en sistemas de transcripción automática, aplicaciones en sistemas multicanal, etc. En especial, aunque se han tenido en cuenta todos los tipos de ruido, se muestra especial interés en el estudio de las voces de fondo, principal fuente de error de la mayoría de los Detectores de Actividad en la actualidad. Las tareas llevadas a cabo poseen como punto de partida un Detector de Actividad basado en Modelos Ocultos de Markov, cuyo vector de características contiene dos componentes: la energía normalizada y la variación de la energía. Las aportaciones fundamentales de esta Tesis son las siguientes: 1) ampliación del vector de características de partida dotándole así de información espectral, 2) ajuste de los Modelos Ocultos de Markov al entorno y estudio de diferentes topologías y, finalmente, 3) estudio e inclusión de nuevas características, distintas de las del punto 1, para filtrar los pulsos de pronunciaciones que proceden de las voces de fondo. Los resultados de detección, teniendo en cuenta los tres puntos anteriores, muestran con creces los avances realizados y son significativamente mejores que los resultados obtenidos, bajo las mismas condiciones, con otros detectores de actividad de referencia.

ABSTRACT

This work has been focused on improving the robustness at Voice Activity Detection in adverse acoustic environments in order to enhance the behavior of many vocal applications, for example telephony applications based on automatic speech recognition, automatic transcription applications, multichannel systems applications, and so on. In particular, though all types of noise have taken into account, this research has special interest in the study of pronunciations coming from far-field speakers, the main error source of most activity detectors today. The tasks carried out have, as starting point, a Hidden Markov Models Voice Activity Detector which a feature vector containing two components: normalized energy and delta energy. The key points of this Thesis are the following: 1) feature vector extension providing spectral information, 2) Hidden Markov Models adjustment to environment and study of different Hidden Markov Model topologies and, finally, 3) study and inclusion of new features, different from point 1, to reject the pronunciations coming from far-field speakers. Detection results, taking into account the above three points, show the advantages of using this method and are significantly better than the results obtained under the same conditions by other well-known voice activity detectors.

INDICE

Pág.

Figuras

Figura 7.1. VAD completo con las nuevas técnicas para el filtrado de voces de fondo.

Tablas

CAPÍTULO 1

JUSTIFICACIÓN Y OBJETIVOS

1.1.- Introducción y justificación.

La aparición de nuevas aplicaciones de tecnología del habla, comunicaciones móviles, codificación, telefonía de manos libres, audio conferencia, cancelación de ecos o reconocimiento de voz robusto, hace que se requiera de un sistema de reducción de ruido en combinación con un detector de actividad de voz preciso, a partir de aquí "*VAD*" debido a las iniciales de las palabras en inglés *Voice Activity Detector*. La operación del *VAD* puede considerarse como un problema de clasificación en situaciones del tipo voz-silencio, voz-ruido o voz-voces de fondo. Durante los últimos años se han estudiado distintas estrategias para la detección de voz inmersa en ruido y la influencia de la decisión del *VAD* en sistemas de procesado de voz ([1], [2], [3], [4], [5], [6], [7], [8]). Por otro lado, el algoritmo de detección de "no-voz" es una parte muy importante de los sistemas de reducción de ruido en general [77] y en arrays de micrófonos en particular ([9], [10]).

Son, por tanto, muchos los esquemas propuestos para la detección de la actividad vocal. En el siguiente capítulo (Estado del arte o de la cuestión) se realizará un análisis completo y una revisión de dichos esquemas proponiendo una clasificación de los mismos en función del campo de aplicación donde se utilicen, su modo de funcionamiento, las características utilizadas, el modelado con el que trabajen o la regla de decisión. Es importante mencionar que pueden existir solapes en estas clasificaciones.

Una de las fuentes más importantes de error en sistemas en los que se usa un *VAD* es la detección imprecisa en sus extremos principio y final: incluso se puede hablar de detección de extremos cuando se acotan los segmentos de voz procedentes de pronunciaciones. El proceso de la detección de extremos no resulta complicado cuando la relación señal-ruido (SNR) es bastante alta (alrededor de 30 dB). En estas condiciones, el ruido de fondo, incluidas las voces de fondo, tiene un nivel energético mucho menor que los segmentos más débiles de la palabra tales como fricativas o nasales. Sin embargo, en la práctica, la detección de extremos puede resultar complicada debido a ciertos problemas, especialmente en los fonemas de baja energía (fricativas, oclusivas, etc.):

- Los problemas que tienen que ver con el locutor y con la manera de producir la voz. Por ejemplo, durante la articulación, el locutor produce

pequeños ruidos debidos a golpes de los labios, respiraciones pronunciadas y chasquidos de la boca. Éstos pueden aparecer antes o después de la pronunciación. La energía de los mencionados ruidos puede llegar a tener valores comparables a los de la voz, y, en el caso de la respiración, suele aparecer adherida al final de la palabra dificultando el proceso de detección de extremos. Además, el locutor puede encontrarse en condiciones bastante críticas como por ejemplo estresado, con distintos estados anímicos o en movimiento.

- Debido al ruido del entorno en el que se produce la voz. En unas ocasiones la voz se genera embebida en ruido estacionario (máquinas o coches en funcionamiento), en otras en ruido no estacionario (golpe producido por el cierre brusco de una puerta, movimiento de sillas, etc.) y finalmente, otras veces hay voces de fondo (por ejemplo producido por la TV, radio, conversaciones de fondo, etc.). La mayoría de estas señales interferentes pueden llegar a ser comparables a la propia voz de interés, dificultando de forma acusada la detección de extremos. El gran problema de las voces de fondo será abordado en esta Tesis.
- La distorsión que introduce el sistema de transmisión debida a la diafonía, intermodulación, transitorios, tonos de líneas de la red telefónica conmutada, etc.

También es importante mencionar el problema de los ecos que puedan llegarle al Detector. Los ecos pueden producir disparos indeseados del mismo ya que pueden tener presencia de voz lejana. Para solucionar este problema de presencia de eco se suelen incorporar canceladores de ecos que con la tecnología actual resuelven bastante bien este problema [11].

En cuanto a las características básicas que debe cumplir un buen *VAD* se puede decir lo siguiente [12]:

- Fiable y robusto para no cometer errores en condiciones de trabajo complicadas como, por ejemplo, relaciones señal-ruido variables, niveles de ruido no estacionario, voces de fondo, etc.
- Precisión en sus extremos. Será una característica más o menos importante en función de la aplicación del *VAD*: en reconocimiento de voz

es necesaria mucha precisión, mientras que en codificación se puede recortar un poco de voz sin que afecte a la conversación.

- Adaptativo, especialmente al ruido de fondo variable.
- Simplicidad es otra característica que se le pide al *VAD*, especialmente cuando el algoritmo va a formar parte de un reconocedor de voz u otras aplicaciones que lo requieran.
- Debe de ser capaz de funcionar en tiempo real, cuestión difícil si no se cumpliese el punto anterior.
- Capaz de diferenciar lo mejor posible las voces de fondo de la voz principal. Es con diferencia el caso más crítico de todos ya que los modelos de voz fallan a favor de la existencia de ésta, y es que, una voz de fondo no deja de ser voz.

Con todo ello, no hay que perder de vista que en muchas ocasiones no se pueden cumplir todas las características anteriores, y que los requisitos también dependen de la aplicación que se le dé al Detector. A pesar de que siempre es deseable una precisa detección de extremos, en algunas aplicaciones, por ejemplo las de reconocimiento de voz, el éxito de los resultados depende críticamente de lo buena que haya sido la detección de extremos. En otras aplicaciones como por ejemplo codificación, es menos importante una detección de extremos tan precisa como en el caso anterior ya que los detectores utilizados en codificación se usan para que el codificador transmita a velocidades bajas la señal en ausencia de voz (se reduce el nivel de las interferencias radio).

Una detección de actividad muy precisa hace que haya que utilizar complicados algoritmos de cálculo, sobre todo en entornos con ruido no estacionario (por ejemplo el generado con comunicaciones móviles GSM), con un mayor gasto computacional que puede hacer que el Detector no funcione correctamente en tiempo real. Por lo tanto, los requisitos que debe cumplir un *VAD* son los anteriormente mencionados, pero en función de la aplicación práctica que se les dé se puede hacer alguna flexibilización en algún aspecto. En este trabajo se tendrán en cuenta todos estos requisitos a la hora de realizar comparaciones y de exponer resultados.

La problemática que generalmente caracteriza el diseño de un *VAD* es, en función de su tecnología, la elección de las características que va a utilizar, los

modelos que usan estas características, o las topologías de estos modelos. En general, por cada unidad de tiempo que suele estar comprendida entre 20 y 40 ms, el *VAD* genera una salida binaria (voz/no voz) a partir de una señal de entrada. Algunas características tradicionalmente utilizadas para el diseño de un *VAD* son: energía, tasa de cruces por cero, coeficientes de predicción lineal (LPC), coeficientes cepstrales (cepstrum), los formantes, etc. La selección de un adecuado vector de características para la detección de la señal y una regla de decisión sobre un modelado robusto es un problema desafiante que afecta al funcionamiento del *VAD* cuando opera en condiciones de ruido o voces de fondo. La mayoría de los algoritmos son efectivos en numerosas aplicaciones, pero a menudo tienen errores para niveles bajos de relación señal-ruido (SNR) [5]. Por ejemplo, un simple detector de nivel de energía puede trabajar bien en condiciones de alta relación señal-ruido, pero fallaría de forma significativa en condiciones de baja SNR. Muchos algoritmos han sido propuestos para paliar estas desventajas por medio de la definición de reglas de decisión más robustas.

Resumiendo, los detectores de actividad vocal son sistemas capaces de discriminar entre la ausencia (ruido, silencio) o presencia de voz. Su estudio se lleva realizando desde hace muchos años, como se ha descrito anteriormente, aunque, todavía no se ha llegado a una solución general debido principalmente tanto a la gran variedad de ruidos existentes en los distintos entornos como a las diferentes necesidades de los distintos sistemas donde pueden integrarse. Los autores de los trabajos de investigación, en sus resultados finales, no son capaces de emitir un juicio generalista en los diferentes ambientes o entornos de aplicación en el que mejoren las tasas de error en todos los casos y para todos los tipos de ruido. Normalmente, los algoritmos innovadores que se proponen, mejoran el comportamiento del Detector únicamente en ciertos entornos de ruido, mientras que empeoran en los otros. La complejidad del tema hace necesario por tanto, no sólo invertir un esfuerzo importante en la mejora de la tecnología empleada en la detección de actividad sino también el estudio exhaustivo de nuevas técnicas de medida y evaluación de prestaciones, aspecto este al que se prestará especial interés en el trabajo de Tesis realizado.

1.2.- Objetivos.

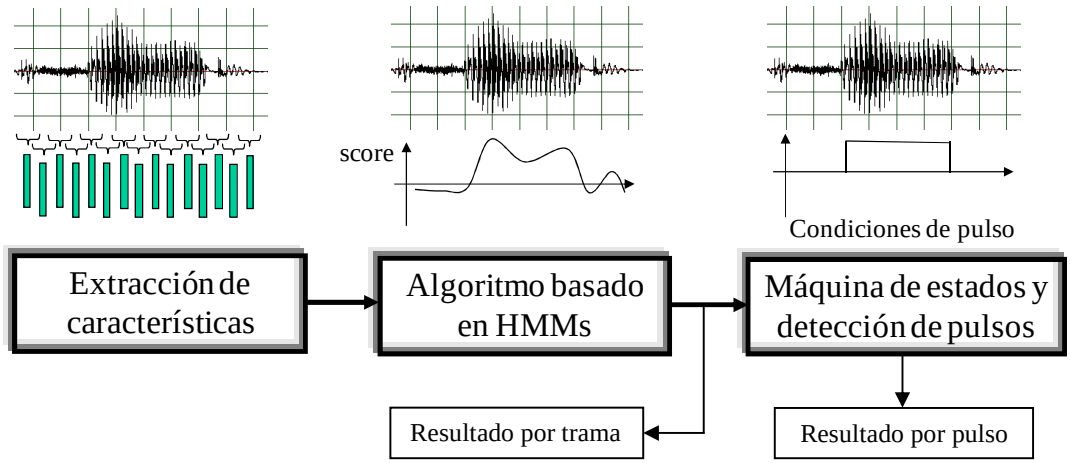

Figura 1.1. Esquema del *VAD* de partida.

El objetivo general de la Tesis es el de disminuir los errores de detección mediante la inserción de diversas características dentro de los algoritmos que usa el *VAD* del que se parte en este trabajo (Fig. 1.1). Se prestará especial interés al problema de detección de las voces de fondo: es una de las principales fuentes de error en la actualidad. Tal y como se refleja en la Fig. 1.1, el *VAD* sobre el que se trabajará en esta Tesis es un Detector basado en el uso de Modelos Ocultos de Markov (HMM). Concretamente se basa en utilizar dos HMMs, uno para representar la clase "voz" y otro para representarla clase "no voz". El vector de características con que trabajan los dos HMMs aprovecha su "inserción" en Sistemas de Reconocimiento, y, por tanto, se estudiarán características de Reconocimiento Automático de Habla (ASR), los cepstrum, en sus dos HMMs, a diferencia de los codificadores que suelen usar los parámetros disponibles en el codificador. Estos HMMs obtienen una puntuación a nivel de trama que corresponde a la verosimilitud de que la trama haya sido generada por el HMM de voz o el HMM de ruido. La diferencia entre estas dos verosimilitudes se utiliza como un valor (o *score*) sobre el cual tomar una primera decisión sobre si la trama contiene voz o ruido. Así, como ilustra la Fig. 1.1, sobre esa diferencia de verosimulitudes o *scores* se genera una secuencia, con un resultado por trama, por ejemplo del tipo: no-voz (trama 1), voz

(trama 2), voz (trama 3), voz (trama 4), no-voz (trama 5), etc. Posteriormente, el *VAD* mediante la máquina de estados y la detección de pulsos (ver Fig. 1.1) es capaz de unir las tramas y formar pulsos, por ejemplo las tramas 2, 3 y 4, o incluso tratar varios pulsos, por ejemplo ensancharlos o juntarlos si la estructura del lenguaje lo requiere. Teniendo en cuenta esta estructura del *VAD* de partida, los tres puntos fundamentales de estudio abordados en este trabajo son los siguientes:

- Ampliación del vector de características ("features" en inglés): Se realizará un estudio exhaustivo que sea capaz de medir el poder de discriminación de cada una de las características tratadas (cepstrum) para poder así realizar una selección adecuada de las mismas. Es una aportación sobre el módulo de extracción de características.
- Ajuste de los Modelos Ocultos de Markov (HMMs) al entorno y estudio de su comportamiento con distintas topologías, por ejemplo, el número de modelos a tener en cuenta, el número de estados por modelo o el número de gaussianas por estado. Es una aportación sobre el módulo de los HMMs para la posterior toma de decisión.
- Inclusión de nuevas características o "features", distintas de las del vector de características utilizadas en los HMMs, para eliminar o filtrar los pulsos que proceden de las voces de fondo. Se presentarán distintos métodos de filtrado: basado en umbrales, en un árbol de decisión y en una red neuronal. Es una aportación sobre el módulo de máquina de estados y detección de pulsos.

También se realizarán las comparaciones experimentales precisas entre el sistema de detección propuesto y otros *VAD* de referencia usando las mismas bases de datos, medidas de evaluación, etc.

En general, se crearán uno o varios detectores que traten de obtener mejores resultados no sólo para tipos de ruido concretos, como la mayoría de los autores hacen, si no intentando cubrir el mayor espectro posible de situaciones habituales en los sistemas de telefonía: voz limpia, voz con ruido estacionario o por el contrario voz con ruidos no estacionarios como por ejemplo las voces de fondo (televisión, conversación, radio, etc.).

1.3.- Detector de actividad propuesto y mejoras en cada uno de sus módulos.

Con el objeto de tener una visión global del resultado de la Tesis, en la Fig.1.2 se presenta de forma esquemática las principales características del *VAD* resultante de este trabajo. Como se puede observar, las aportaciones planteadas afectan a tres módulos perfectamente diferenciados.

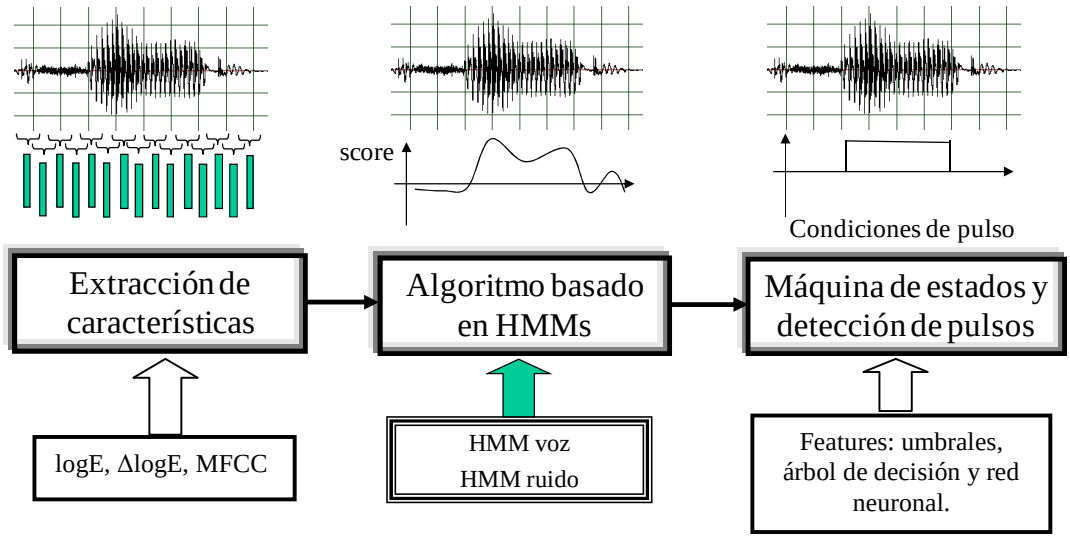

Figura 1.2. Esquema completo del nuevo *VAD*.

A continuación se describe brevemente cada uno de los tres módulos que forman el *VAD* propuesto, y, en cada uno de ellos se resumen las mejoras que se efectúan en este trabajo de Tesis:

1. *Extracción de Características*: Es el encargado de generar el vector de características y está formado por dos componentes: logaritmo de la energía normalizada y derivada de la energía. La energía se normaliza con el fin de que el *VAD* sea invariante ante los cambios de la relación señal a ruido (SNR). En este trabajo se realiza una ampliación del vector de características a cinco componentes: se añaden los tres primeros cepstrum o MFCCs, del inglés Mel Frecuency Cepstral Coefficients, procedentes de un banco de 12 filtros Mel con preénfasis. De esta manera, al vector de

características se le está dotando de información espectral. Además, otra mejora importante en este apartado de extracción de características es la mejora del proceso de normalización de la energía: se usa una versión simplificada del estimador de ruido del códec AMR1 (Adaptive Multi-Rate).

2. *Algoritmo basado en HMMs:* El algoritmo basado en HMMs calcula la diferencia del logaritmo de verosimilitudes de los modelos de voz y de ruido, para obtener a su salida una segmentación temporal en tramas de voz y no voz. En este trabajo, se han evaluado distintas topologías, encontrándose como topología óptima de estos dos modelos acústicos la siguiente: el modelo de ruido es un HMM de izquierda a derecha formado por tres estados y una gaussiana por estado, y el modelo de voz es un HMM de izquierda a derecha formado por cuatro estados y también una gaussiana por estado. Además, es importante hacer hincapié en que estos modelos son caracterizados en el proceso de entrenamiento por tener una gaussiana por estado, que se representa mediante una media y una varianza por cada una de las cinco componentes de vector de características, de forma separada para cada uno de los modelos.
3. *Máquina de estados y detección de pulsos*: La toma de decisión se realiza sobre las puntuaciones correspondientes a la relación de verosimilitudes proporcionadas por el algoritmo basado en HMMs que se menciona en el punto anterior. Tras esta segmentación, el sistema propuesto añade otro tipo de información relevante para detectar pulsos de voz y mejorar así los resultados finales. Esta información está relacionada con la duración de los pulsos, el silencio entre pronunciaciones y tramas adicionales al inicio y fin de pulso. En este trabajo, y como cuestión innovadora, se presentan distintas técnicas para eliminar o filtrar las voces de fondo. Estas técnicas son combinaciones de estadísticos calculados sobre los pulsos de voz, sobre distintas características o “features” calculadas para cada trama de los mencionados pulsos de voz. Estas características son las siguientes: armonicidad de la señal de voz, distancia de Mahalanobis entre los MFCCs obtenidos sobre tramas de voz consecutivas y dos características sobre un LPC residual de orden 10 (kurtosis y máximo valor de la auto-correlación del residuo). El filtrado se realiza tomando decisiones sobre la combinación de

los diferentes estadísticos, mediante la comparación con distintos umbrales, uno por estadístico, o mediante la comparación con un umbral global a través de un árbol de decisión o una red neuronal.

1.4.- Estructura de la Tesis.

En esta sección se presenta la estructura global de este trabajo describiendo el contenido de cada de cada uno de los capítulos que van a constituir esta Tesis Doctoral:

- Capítulo 2: este capítulo recoge el estado del arte o de la cuestión de los Detectores de Actividad. En él se explica la estructura de un Detector de Actividad típico y cuáles son todas sus etapas: preproceso, extracción de características, clasificación, decisión y evaluación final. En cada etapa se describen diferentes propuestas considerando distintos aspectos tales como el tipo tecnología, el ámbito de aplicación, etc. Se presenta especial atención a los usados en reconocimiento de voz, entre los que se tienen por ejemplo, los que se valen del análisis espectral (MFCCs) o los que usan análisis y predicción de espectro.
- Capítulo 3: es el capítulo en el que se exponen las medidas de evaluación y se describen todas las bases de datos usadas: entrenamiento, desarrollo y test. Contienen una gran variedad de casos, desde voz limpia (relación señal a ruido alrededor de 30 dB) hasta voz mezclada con ruidos tanto estacionarios como no estacionarios. Dentro de los ruidos no estacionarios cabe destacar la base de datos Av16.3 [13] que contiene voces de fondo procedentes tanto de uno como de varios locutores. También se tienen en cuenta diferentes canales de comunicaciones para que los experimentos abarquen la mayor casuística posible: red GSM, red de telefonía fija y voz IP. La tecnología más empleada es GSM, ya que, es especialmente crítico usar un *VAD* en entornos abiertos con la presencia de ruidos de todo tipo.
- Capítulo 4: se presenta el *VAD* basado en Modelos Ocultos de Markov. Se explica la estructura de cada uno de los módulos que forman el *VAD*, el de extracción de características formado por un vector de cinco componentes, el algoritmo basado en HMMs, uno de voz y otro de ruido, y, por último, el

módulo de decisión (máquina de estados) y de detección de pulsos. Una de las componentes del vector de características es la energía normalizada: que se obtiene mediante un proceso de normalización basado en una versión simplificada del códec AMR1. Además, se realiza un estudio de las variaciones de las topologías de los modelos en el *VAD* y su comportamiento ante distintas redes de comunicación, GSM, fija y VoIP.

- Capítulo 5: en este capítulo se analiza la inclusión de características, tanto de estructura armónica como de envolvente espectral, dentro del módulo de detección de pulsos del *VAD* con el fin de filtrar los pulsos que provienen de voces de fondo. Este análisis se lleva a cabo mediante la base de datos Av16.3 para medir el poder de discriminación de las características anteriores. También se presenta un resumen, al final del capítulo, que amplia los resultados de los errores de clasificación para pulsos de voces de fondo de distinta naturaleza.
- Capítulo 6: en este capítulo se presentan los resultados finales que se obtienen tras introducir en el módulo de detección de pulsos del *VAD* las características capaces de filtrar los pulsos de voces de fondo. Estos resultados son comparados, sobre las mismas bases de datos de este trabajo, con los obtenidos por otros *VAD* de referencia: AURORA, AMR1, AMR2 y G729 anexo B.
- Capítulo 7: Para terminar, se presentan las conclusiones finales y las aportaciones realizadas en el campo en el que este trabajo de Tesis queda encuadrado. También se plantean las posibles líneas de trabajo futuro.

CAPÍTULO 2

ESTADO DE LA CUESTIÓN

2.1.- Introducción.

En este capítulo se presenta la estructura básica de un Detector de Actividad completo típico y se realiza a su vez una clasificación en función de los módulos que lo forman y, el tipo de aplicación en la que se usa. Para finalizar se presentan sistemas similares al propuesto en este trabajo.

2.2.- Estructura de un detector de actividad típico.

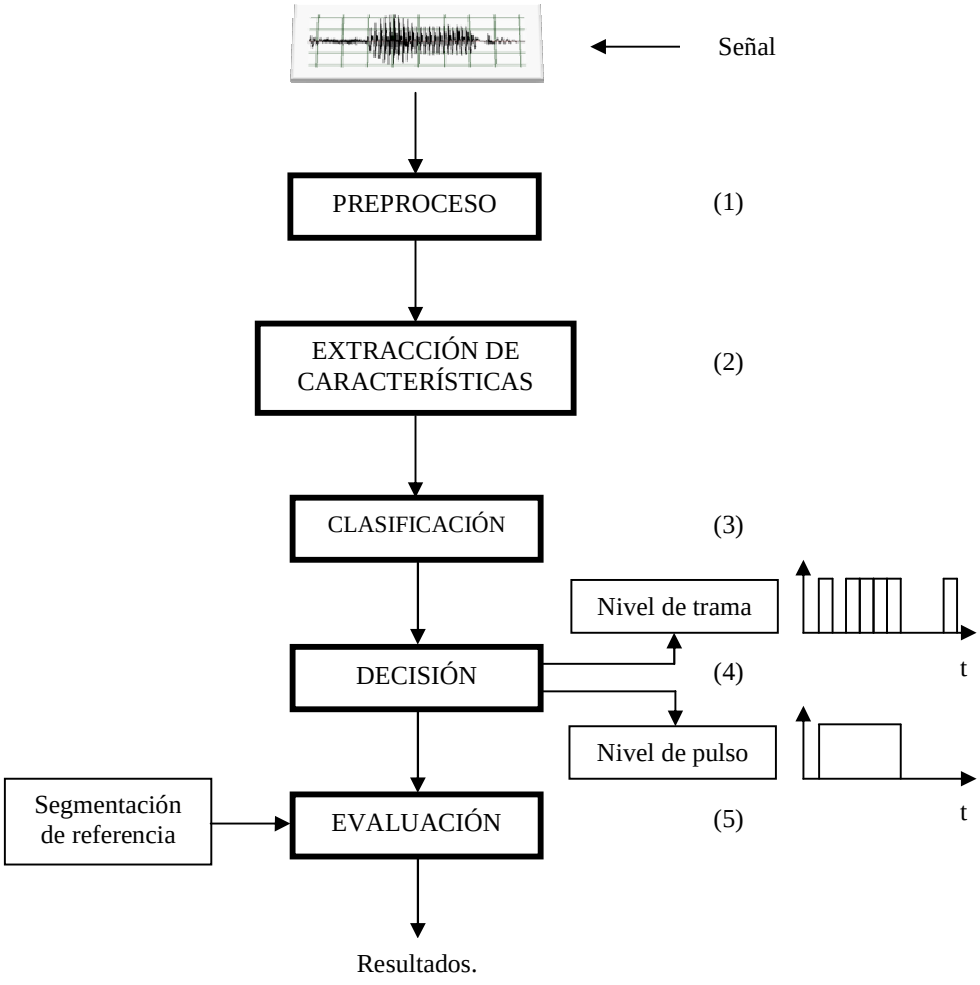

Figura 2.1. Estructura del Sistema de Detección.

En este apartado se plantea el esquema de detección que se propone en esta Tesis estructurado en diagrama de bloques, desde la entrada de la señal hasta la decisión y evaluación final de los resultados. En la Fig.2.1 se presenta un esquema típico de clasificación de patrones. Como se puede observar, junto a cada uno de los bloques aparece un número de identificación que servirá de referencia. Así, cuando en el futuro se hable del bloque (2), se referirá a la extracción de características. La estructura de bloques se describe como sigue:

- Señal (de entrada): se trata de voz embebida en silencio, ruido o incluso voces de fondo.
- Preproceso: aquí se enmarcan los sistemas de reducción o eliminación de ruido o voces de fondo. Pretende mejorar la relación Señal a Ruido (voz principal frente al resto de efectos acústicos).
- Extracción de características: se obtiene la información de la señal, preprocesada. Esta información puede ser la que se obtenga del:
    - Dominio del tiempo: tasa de cruces por cero, variaciones de energía (SNR), correlaciones etc...
    - Dominio de la frecuencia:
        - Análisis y predicción: coeficientes cepstrales (MFCC, Mel Frequency Cepstrum Coefficients), filtros de Wiener (WF) o discriminación de voz y no-voz (ruidos de cualquier tipo o voces de fondo) por bandas.
- Técnicas de clasificación: se utilizan para representar cada una de las clases a clasificar (voz/no voz). Más tarde se utiliza para decidir de qué tipo es la porción de señal recibida y analizada. Algunos ejemplos de estas técnicas:
    - Modelos Ocultos de Markov (HMM) y Modelos de Mezcla de Gaussianas (GMM). Son modelos basados en una aproximación a la función de densidad de probabilidad.
    - Redes Neuronales Artificiales (ARN). Los parámetros de los modelos se estiman haciendo uso del criterio de menor error de clasificación.

  - K-Means, K-Nearest Neighbours (KNN). Conjunto de ejemplos etiquetados y se calculan las distancias a los mismos.
  - Árboles de Decisión. Técnicas de aprendizaje cuyo cometido es extraer reglas a partir de ejemplos disponibles.

- Decisión: proceso en el que se compara las características de una trama con los diferentes modelos, obteniendo una medida de parecido. Estas medidas se pueden comparar con umbrales y obtener así el resultado trama a trama: cada trama obtendría el resultado de “voz” o “no voz”. Si el proceso termina aquí se puede hablar de “Decisión a nivel de trama”. Si por el contrario se avanza un punto más y se tienen en cuenta, para la toma de decisión, otra serie de características como la información de la estructura del habla, se tendrán en cuenta conjuntos de resultados por trama y se podrán tomar decisiones a nivel de pulso (“Decisión a nivel de pulso”). La estructura del habla, por tanto, no es algo que se aplica trama a trama si no que puede tener en cuenta la historia de la voz y del ruido, cualquier tipo de condición sintáctica y semántica, etc.

  También, tanto a nivel de trama como a nivel de pulso, se puede realizar otra subdivisión en función del tipo de decisión a tomar: de forma directa a partir de características sencillas, modelos generativos o técnicas discriminantes.

  Por otro lado, una técnica antigua usada para la toma de decisión es la basada en el ajuste manual de umbrales, típicamente los de energía.

- Evaluación: se comparan los resultados obtenidos en este sistema de detección con los reales, procedentes, por ejemplo, de un etiquetado manual. Se pueden usar diferentes criterios o métricas, como por ejemplo:
  - La precisión de marcas.
  - La respuesta de las curvas ROC (Reciever Operating Characteristics) enfocado a pulsos de voz, es decir, las falsas aceptaciones y los falsos rechazos se calculan a nivel de trama aunque la decisión se toma a nivel de pulso y siempre se tiene en cuenta la estructura del habla.

- Resultados de la tasa de acierto del reconocedor de voz: Tasa de error de palabra (WER≡Word Error Rate) dependiendo de la calidad de la detección.

2.3.- Preproceso.

Es el primer bloque, (1), según el esquema del sistema de clasificación en la Fig.2.1, por el que pasa la señal captada.

En el "preproceso" se intenta mejorar la relación señal/"no voz" para que todo el procedimiento de detección sea más sencillo, fiable y robusto [1,14,15].

Los primeros sistemas de detección funcionaban sin una etapa de preproceso con un consecuente deterioro de su funcionamiento. Sin embargo, trabajos más actuales demuestran la ventaja existente cuando la señal es sometida a un tratamiento de este tipo, sobre todo en entornos ruidosos o con características especiales. En el preproceso se encuadran tanto las técnicas de reducción de ruido como de cancelación de ecos.

Las técnicas de reducción de ruido pueden ayudarse de un detector de voz rápido y sencillo, normalmente distinto del detector robusto del sistema que será más complejo y usará más características, o simplemente ayudarse de características de la voz, como por ejemplo buscar las zonas de máximos y mínimos de energía, y en las zonas con mínimos de energía, se puede realizar una reestimación del ruido ([16], [17]). Así por ejemplo, en el estándar de la ETSI Aurora [77] el preproceso consiste en un sistema de reducción de ruido basado en un análisis en frecuencia. Este análisis se lleva a cabo mediante los llamados filtros de Wiener [1,14]. Los filtros de Wiener y sus variantes han sido diseñados para trabajar en los casos donde el ruido es muy elevado. Este tipo de filtro requiere que conozcamos mucho la señal y las características del ruido añadido a la voz.

Existe un número elevado de publicaciones de algoritmos de reducción de ruido en uno o varios micrófonos consolidando el gran interés en este campo a lo largo de dos o tres décadas. Un punto crucial en este tipo de algoritmos es la estimación concurrente (para las señales capturadas en los distintos micrófonos) del espectro de ruido interferente con la voz principal. Debido a que los entornos ruidosos se caracterizan por la no estacionaridad, es necesario actualizar la

estimación del espectro de ruido tan a menudo como sea posible para mantener una reducción de ruido efectiva. Esto se puede llevar a cabo por ejemplo en ausencia de voz.

Otra cuestión importante es la limitación de complejidad del algoritmo cuando se supone implementado en circuitos digitales. Los requisitos computacionales y de memoria deben ser los más pequeños posibles.

2.4.- Extracción de características.

La extracción de características, que corresponde al bloque (2) del sistema completo, trata de obtener medidas o características cuantitativas que aporten la mayor información posible de la señal, en el dominio espectral o temporal, y utilizarla posteriormente para clasificar las tramas como voz o no voz de la mejor manera posible. Dentro de este marco existen multitud de sistemas de detección en base al uso de distintas características. A continuación se describen los distintos casos encontrados en la literatura. Se puede realizar la siguiente clasificación en función del tipo de características usadas:

- Dominio temporal: en este caso el conjunto de características se obtiene del dominio temporal, tasa de cruces por cero, evolución de la energía, etc...
- Dominio espectral: corresponde al conjunto de características obtenidas en el dominio de la frecuencia (Transformada de Fourier). En cada trama, formada por N muestras, se realiza una Transformada de Fourier. Sobre los valores transformados se obtienen las características del *VAD*. Caerían dentro de esta clasificación los detectores que usen análisis cepstral o los que utilicen la información útil por bandas de frecuencia (las propiedades de la señal hacen que a veces se pueda realizar una clasificación selectiva de sonidos: ruido de coche, voz principal o de fondo, tonos de cabina, etc.).

También pueden existir combinaciones de distintos tipos de características, por ejemplo en [18] se usan tanto parámetros de dominio espectral como de dominio temporal a la vez.

Otros autores [15,16] detectan la actividad mediante la medida de periodicidades en la señal, muestreando datos en un rango de frecuencias entre 200 y 1000 Hz. Esta banda tan estrecha ayuda a reducir la probabilidad de interferencia, aunque un ruido

con energía a estas frecuencias puede interrumpir la toma de decisión del *VAD*. Para distinguir claramente entre los niveles de voz y de ruido se incorpora un control automático de ganancia con adaptación de umbrales mínimos de adaptación.

Las características que usan los detectores de actividad en muchos casos dependen de la aplicación en la que van a ser utilizados, ya que si la aplicación requiere el cálculo de ciertos parámetros, lo más cómodo y rápido es que el Detector también los use a ser posible. Una posible clasificación sería la siguiente:

- Detectores sencillos: usan características de bajo coste computacional. Son normalmente los usados para realizar una reestimación inicial de ruido, por ejemplo los basados en la energía, tasa de cruces por cero, etc. Un ejemplo de este tipo es el *VAD* de Rainer y Sambur [19].
- Detectores usados en reconocimiento de habla: se trata de detectores que usan coeficientes cepstrales o de predicción lineal, entre otras, como características. Shafran por ejemplo en [20] propone un *VAD* para aplicaciones de ASR que utiliza coeficientes de predicción lineal.
- Detectores usados en codificación: se emplean para transmitir la señal a bajas velocidades en las zonas en las que no hay voz. El *VAD* del códec GSM [21] es uno de los más usados, y usa características como la energía o la auto-correlación.
- Detectores usados en sistemas de reducción de ruido (energía por bandas): existe una gran diversidad de detectores propuestos de este tipo: [22], [16], [17] o [23]. En estos casos mediante valores de la energía, ya sea de voz o de ruido, la detección de actividad se realiza sobre el cálculo de la SNR en diferentes bandas de frecuencia.

En los subapartados siguientes se detalla cada uno de los casos y se analizan sus principales características.

2.4.1.- Detectores sencillos y de bajo coste: detector basado en la energía y tasa de cruces por cero.

Tradicionalmente, este tipo de *VAD* [24-26] combina la energía instantánea en una trama, con medidas de tasas de cruce por cero y confían en que un nivel alto de energía es el mejor estimulo para la detección. Se asume que la tasa de cruces

por cero es bien distinta en zonas de voz con fricativas (baja energía) o de ruido de fondo y seria el punto adicional para determinar los límites de la voz.

Para explicar su funcionamiento tomaremos como referencia el descrito por Rainer y Sambur [19]. La señal se filtra en un ancho de banda entre 100 Hz y 4 KHz, con una frecuencia de muestreo de 8 KHz. De forma simple, el algoritmo de detección se basa en medidas de la energía cada 10 milisegundos (80 muestras a 8 KHz.) según la siguiente fórmula:

$$E(n) = \sum_{i=-50}^{50} |s(n+i)| \qquad (2.1)$$

Por otro lado, la tasa de cruces por cero, z(n), también se calcula una vez cada trama de 10 milisegundos. El algoritmo supone que durante los 100 milisegundos primeros de la grabación no hay voz presente. Durante este periodo de silencio inicial se miden, la media, $\overline{IZC}$ (Integrating Zero Crossing), y la desviación estándar, σ_{IZC}, de la tasa de cruces por cero, y la energía media del ruido de fondo (IMN). Una vez obtenida la energía de toda la grabación se establecen los siguientes umbrales:

$$\begin{aligned} IZCT &= MIN(IF, \overline{IZC} + 2\sigma_{IZC}) \\ I1 &= 0.03 \cdot (IMX - IMN) + IMN \\ I2 &= 4 \cdot IMN \\ ITL &= MIN(I1, I2) \\ ITU &= 5 \cdot ITL \end{aligned} \qquad (2.2)$$

El umbral $IZCT$ tiene en cuenta el mínimo entre la tasa de cruces por cero promedio en una zona de silencio ($\overline{IZC}$) más dos veces su desviación estándar y un umbral de tasa de cruces por cero, IF, típicamente con valor 25. $I1$ simboliza la suma entre la energía media del ruido de fondo (IMN) y el 3% de la diferencia entre el valor máximo de la energía de voz (IMX) y la energía media del ruido de fondo (IMN). $I2$ es cuatro veces la enegía media del ruido de fondo, ITL el mínimo entre $I1$ e $I2$, y, finalmente, ITU, cinco veces ITL. A continuación vendría la toma de decisión, cuestión esta de la que se hablará en el apartado 2.6.

En el escenario de detectores sencillos y de bajo coste también encontramos a Kang y Fransen [27], que proponen un esquema muy simple: siempre que la banda baja de energía (0-1 Khz.) de la trama actual de la señal este por debajo de una fracción especifica del rango dinámico de la mencionada banda baja en las tramas anteriores, la trama es considerada como ruido y se usa para actualizar la estimación del espectro de ruido. Obviamente, este proceso posee fuertes limitaciones. Sólo trabajará con altas SNRs y fallará con ruidos de baja frecuencia.

Un algoritmo más elaborado usa umbrales de energía adaptativos [28]. Elberling [29] usó el llamado método síncrono para la estimación espectral de ruido de fondo. Este procedimiento hace uso de una característica específica de sonidos de voz que es que la energía está confinada en frecuencias fundamentales. Basándose en multiplicaciones sucesivas de las envolventes de los pares vecinos de las bandas, se obtiene una medida global de la sincronización de la energía. De esta manera, para clasificar las tramas de la señal de entrada, se asocia la alta sincronización a las tramas de voz mientras que la baja a las de ruido.

Sheikhzadeh [30] propuso un algoritmo de detección basado en la auto-correlación, que fue realizado para señales realzadas (después de la reducción de ruido). Aunque se realizan muchos experimentos, en ninguno de ellos existe mejora. Sin embargo, los autores suponen que no se trabaja bien si las SNRs están por debajo de 0 dB. Dendrinos y Bakamidis [31] presentan un algoritmo con el que determinan los extremos de los segmentos de voz en entornos de ruido coloreado basándose en algunos umbrales determinados experimentalmente. Se obtuvieron mejoras para SNRs mayores de 0 dB.

Otro trabajo a destacar es el estándar de la ETSI Aurora [77]. En este caso el estándar usa dos *VAD*s: uno de ellos es un *VAD* sencillo basado en la energía que se usa dentro del proceso de reducción de ruido.

2.4.2.- Detectores usados en reconocimiento de voz.

Se trata de detectores que se valen de la información espectral para obtener las características que a su vez son usadas por los reconocedores de voz.

En un análisis de habla normal, los coeficientes cepstrales varían con el tiempo, reflejando los distintos tipos de sonidos de voz, cuestión ésta importante para un contexto de reconocimiento de voz. Es importante comentar que los

coeficientes que se usan en reconocimiento son los mismos que se usan para el *VAD*. Todo se basa en las diferencias existentes en las componentes cepstrales de la voz y del ruido. Volviendo al fundamento del *VAD*, en cuanto a la discriminación entre voz y ruido, se puede decir que la voz se puede modelar de manera precisa ya que los espectros de voz varían rápidamente, cuestión que era previsible. De forma más esquemática, las ventajas de los coeficientes espectrales para *VAD* son:

- Los coeficientes espectrales de la voz varían más rápidamente que los del ruido.
- Los rangos de los coeficientes espectrales de la voz son mayores que los del ruido.

Esto quiere decir que estos coeficientes de la voz pueden poseer valores en un amplio rango además de variar de forma rápida. Por el contrario, los coeficientes espectrales del ruido suelen tener rangos de valores más pequeños y cada uno de ellos en concreto varía más lentamente.

Un caso especial de sustracción espectral directa basado en la Transformada de Fourier es [20]. La solución que propone se basa en un algoritmo de funcionamiento de la estimación no-paramétrica del espectro del ruido de fondo usando la estadística mínima de la transformada de Fourier a corto plazo suavizada (STFT). Se muestra que el algoritmo puede operar de forma efectiva bajo condiciones de variación de la SNR. Se obtienen resultados buenos para bases de datos que se diferencian por su estilo de habla, tipo de ruido de fondo y ancho de banda. Y con un retardo de 400 ms resulta ser adecuado para aplicaciones de reconocimiento de voz automático (ASR). Es un método de detección de voz que puede ser aplicado sobre distintas aplicaciones y condiciones sin necesidad de muchos ajustes, cuestión ésta que delimitaba mucho la estructura del *VAD*, y uno de los requisitos de un buen *VAD*: su sencillez.

Dentro de este tipo de detectores se puede a su vez realizar otra clasificación:

- Detectores que usan coeficientes MFCC (Mel Frecuency Cepstral Coefficients).
- Detectores que usan predicción lineal perceptual: PLP (Perceptual Linear Predictive).

- Detectores que usan filtrado RASTA (Relative spectra filtering of log domain coefficients).

2.4.2.1.- Detectores que usan coeficientes MFCC.

Aunque algunos *VAD*s usan el análisis de Fourier para realizar un estudio de la señal por bandas de frecuencia [32], normalmente, los coeficientes MFCC (Mel Frequency Cepstral Coefficients) son los más usados. Los coeficientes MFCC usan una distribución Mel o logarítmica de filtros. La Escala Mel es una escala logarítmica que pretende dar mayor importancia a las frecuencias más bajas. El punto de referencia entre esta escala y la frecuencia normal se define equiparando un tono de 1000 Hz, 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels. Por encima de 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, cuatro octavas en la escala de hercios por encima de 500 Hz se comprimen en dos octavas en la escala Mel. La conversión de hezcios a mels es la siguiente: $m = 1127{,}01 \cdot \log_e(1 + f/700)$.

Una representación gráfica de filtros triangulares en escala MEL o logarítmica se puede observar en [34]. La escala Mel trata de simular la distribución de sensibilidad del oído humano. La voz posee más información a frecuencias bajas que a frecuencias altas. Los MFCC se calculan de la siguiente manera,

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cdot \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \qquad (2.3)$$

donde "N"es el número total de filtros y m_j es la potencia o energía recogida por el filtro *j*. Además, el número de coeficientes MFCC que se pueden calcular depende del número de filtros de la banda, comprendida normalmente entre 400 y 4000 Hz. Por ejemplo en [35] se usan coeficientes MFCC en el cálculo de los HMMs para una adaptación del locutor. Otro trabajo importante es [36], donde se presenta un esquema parecido al del estándar Aurora [77] salvo por el Detector que utiliza. Dicho Detector se basa en un percetrón multicapa que usa varias tramas de 6 coeficientes cepstrales calculados a partir de un filtrado paso bajo de 23 filtros Mel.

La salida del perceptrón da una estimación de la probabilidad de que la trama actual sea de voz o no.

2.4.2.2.- Detectores que usan predicción de espectro: PLP (Perceptual Linear Predictive).

Se usan métodos de predicción lineal (LPC) cuando se quiere parametrizar la envolvente espectral de una señal con un número pequeño de coeficientes con los cuales es posible reconstruirla adecuadamente.

Itoh y Mizushima [33] propusieron una identificación de voz y ruido basada en 4 parámetros distintos. El primero es el máximo valor de la función de auto-correlación de la señal residual obtenida a partir de los coeficientes de predicción lineal (LPC), que representa el grado de periodicidad de la onda. El segundo es un parámetro de la pendiente del espectro, el tercero es un coeficiente de reflexión que se calcula a partir de algunos coeficientes PARCOR, y, por último, el cuarto es la energía. Por otro lado en [20] se usa un reconocedor que usa coeficientes PLP como características de entrada.

2.4.2.3.- Detectores que usan filtrado RASTA (Relative spectra filtering of log domain coefficients).

Al igual que ocurre con todos los métodos de espectro a corto plazo, los coeficientes PLP son vulnerables cuando los valores del espectro a corto plazo son modificados debido a la respuesta en frecuencia del canal de comunicaciones o ruidos. Para paliar en la medida de lo posible este problema, en algunos sistemas de reconocimiento se usan los denominados filtros RASTA (Relative spectra filtering of log domain coefficients) que hace que el análisis PLP sea más robusto a las distorsiones espectrales lineales ya que intentan eliminar el ruido lo mejor posible. Se trata, por tanto, de un método de filtrado que usa información del espectro a corto plazo robusto ante distorsiones espectrales lineales.

En [101] se presenta un *VAD* que usa filtrado RASTA para generar un vector de características que sirve de entrada a un SVM (Support Vector Machine) para la toma de decisión final. Los resultados justifican la robustez del método a bajas SNRs. Otro trabajo en el que se usa filtrado RASTA es [102]. En este se genera un

vector de características RASTA-PLP que sirve de entrada a una red neuronal formada por un perceptrón multicapa.

Es importante comentar que aunque el método más robusto sea el filtrado RASTA, también es el de mayor gasto computacional, y no hay que que perder de vista que la rapidez también es una característica importante en un *VAD*. Por tanto, de nuevo, será importante el tipo aplicación en la se vaya a utilizar este *VAD* para elegir la técnica de filtrado óptima.

2.4.3.- Detectores usados en codificación.

Los detectores usados en codificación se utilizan como herramienta para enviar tramas de "no voz" a una velocidad de transmisión mucho más baja que las tramas de "voz", reduciendo el nivel de las interferencias radio. Esto supone un ahorro de energía grande y evita errores en la decodificación. De nuevo la idea es que los detectores usados en codificación usen los parámetros que el codificador tenga que calcular para su funcionamiento.

Así, algunos codificadores suelen usar detectores que utilizan la información espectral que calculan. Este tipo de *VAD*s los podemos encuadrar dentro del grupo que usan LSF (Line Spectral Frequencies). Es el caso de los codificadores G729B [3] y AMR-2 [106]. El *VAD* del G729B usa como parámetros de entrada para la toma de decisión la energía, la tasa de cruces por cero, la banda baja de energía y una medida espectral. Por otro lado, el diagrama de bloques del *VAD* del codificador AMR-2 se puede visualizar en [32]. En este caso, la señal de entrada se convierte primero al dominio de la frecuencia. Las bandas de frecuencia se agrupan en canales (N_c=16) y se calculan las energías de los mismos (E_{ch}(i,m), i=1,2,...,N_c, m=índice de trama). Dada una estimación de ruido de fondo (E_n(i,m), i=1,2,...,N_c), se estima la SNR del canal (σ(i), i=1,2,...,N_c) donde "i" denota el número de canal y σ(i) el vector donde cada componente se refiere a la SNR en cada canal:

$$\sigma(i) = 10\log_{10}\left(\frac{E_{CH}(m,i)}{E_n(m,i)}\right) \qquad (2.4)$$

A continuación se calculan los índices ($\sigma_q(i)$) de la SNR del canal. Para ello se cuantifica el cálculo de la relación SNR del canal en pasos de 3/8 dB:

$$\sigma_q(i) = \max\{0, \min\{89, round\{\sigma(i)/0.375\}\}\}; \quad 0 \le i \le N_C \quad (2.5)$$

donde los valores de $\sigma_q(i)$ deben ser valores comprendidos entre 0 y 89 ambos inclusive. Con esto se calcula la métrica de voz:

$$v(m) = \sum_{i=0}^{N_C-1} V(\sigma_q(i)) \quad (2.6)$$

donde $V(k)$ es el k-ésimo valor de la tabla de métricas de voz de 90 elementos. Por tanto una función no lineal vincula la SNR del canal a esta métrica de voz, $V(m)$. Además, la SNR del canal se usa también para calcular la SNR de trama y la SNR a largo plazo ($SNR_q(m)$). La métrica de voz ($V(m)$) y el término a largo plazo de SNR proporciona los parámetros primarios para la decisión del *VAD*. También hay un mecanismo de *hangover* en el *VAD* para el tratamiento de los pulsos de voz.

Otro parámetro de entrada directo al *VAD* es sinweave_flag. Este parámetro controla la actualización del ruido de fondo y se basa en la razón del pico promedio del espectro. Cuando la desviación (promediado sobre subbandas, $\Delta E(m)$) se hace muy pequeña, ocurre que se actualiza el ruido bajo circunstancias seguras. Por último, también es importante comentar que la actualización de ruido también depende de la energía total de trama ($E_{tot}(m)$), que se suministra por un estimador de desviación espectral.

Por otro lado, otros codificadores usan términos de predición espectral a largo plazo o LTP (Long Term Prediction). Este es el caso del codificador EFR (GSM) que a continuación se describe.

Detector de voz del estándar GSM

En la recomendación 06.32 dada por la ETSI [21] se explica el Voice Activity Detector (*VAD*) o Detector de Voz del estándar EFR GSM. La función del *VAD* es

indicar si cada trama de 20 milisegundos producida por el codificador contiene voz o no. Para ello el detector podrá utilizar información de las tramas anteriores.

El esquema general del Detector de voz se muestra en la Fig.2.2, en cuya entrada se parte de la auto-correlación en tramas consecutivas (ACF) y se obtiene la decisión de voz final.

La señal de entrada se filtra entre 300 y 3400 Hz, y se muestrea a 8 kHz. Posteriormente se realiza una compensación de offset y un filtrado de preénfasis. Las características usadas son: el pitch en tramas consecutivas de 20 milisegundos (N) sin solapamiento, nueve valores de auto-correlación (ACF) a partir de los cuales se calcula su error de predicción (Pvad), y un umbral (thvad).

Los coeficientes del filtro adaptativo han de actualizarse cuando se detecta una estacionariedad relativa al ruido presente en móviles en periodos largos. Las vocales y los tonos también poseen esta estacionariedad aunque son fácilmente distinguibles del ruido ya que se comparan valores consecutivos del pitch proporcionados por el decodificador.

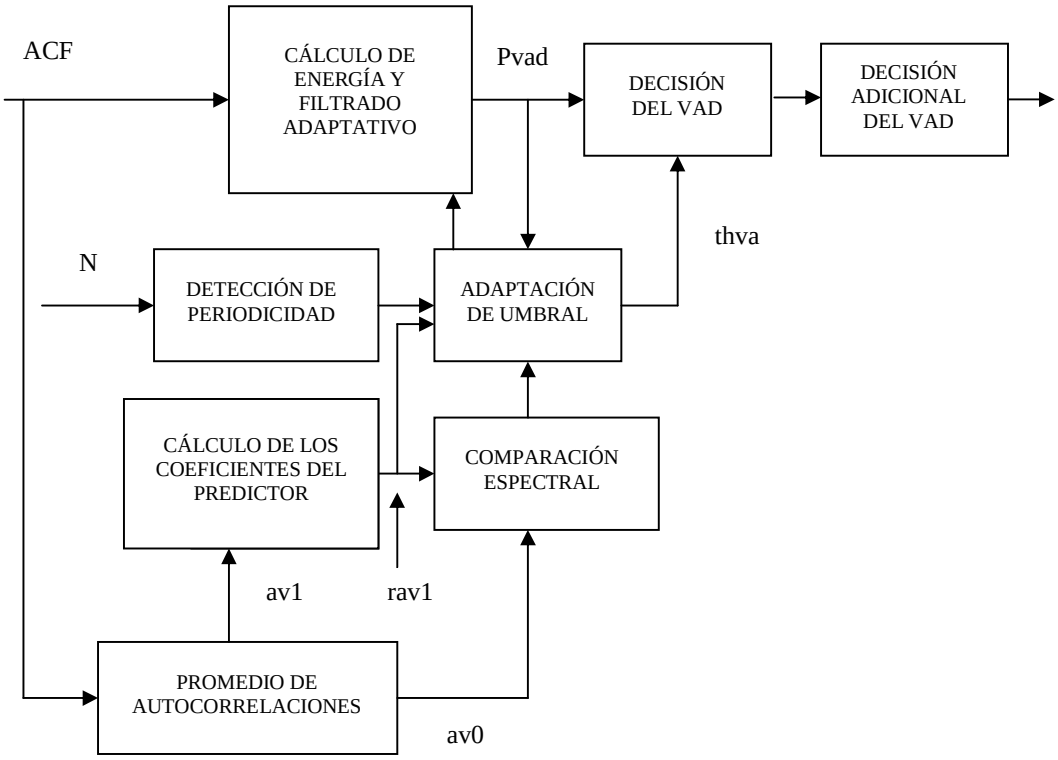

Figura 2.2. Diagrama de bloques del Detector de Voz del estándar GSM.

Además de comprobar el pitch cada 20 milisegundos, también se determina si el umbral, thvad, debe o no cambiar. Esta adaptación tiene lugar en dos situaciones distintas: cuando el término ACF[0] tiene un valor muy bajo y cuando existe una probabilidad muy alta de que no haya voz presente. En el primer caso, el umbral se actualiza si el valor de auto-correlación en el origen, esto es, la energía, es menor que un cierto valor. En el segundo caso, el umbral y los coeficientes del filtro adaptativo serán actualizados sólo si la señal es estacionaria y no periódica, pues en este caso existe una alta probabilidad de no haber pronunciación. La estacionariedad se detecta en el dominio frecuencial, calculando la diferencia entre espectros usando el método ya explicado anteriormente. Cuando esta diferencia espectral es ínfima en un cierto número de muestras, la señal no tiene componente periódica, inherente en los sonidos sonoros y tonos de información, y ocurre la adaptación.

La condición que se evalúa para tomar una decisión referente a si existe voz o no en la trama es comparar el valor de la energía con un cierto umbral:

$$\begin{aligned} Pvad > umbral &\Rightarrow VOZ \\ Pvad < umbral &\Rightarrow NO \quad VOZ \end{aligned} \qquad (2.7)$$

Para finalizar, hay que decir que el Detector de Actividad posee una decisión final en caso de que se tengan picos de voz de corta duración embebidos dentro de ruido.

2.4.4.- Detectores utilizados en sistemas de reducción de ruido.

En este apartado se presentan detectores que se usan en sistemas para reducir el ruido de la señal entrante, por ejemplo, para saber qué tramas son de ruido o de voz cuando se quiere realizar la reestimación de la energía de ruido con éxito.

Se han propuesto distintos algoritmos de reducción de ruido, que necesitan un detector de ruido, que actualizan continuamente la estimación de ruido y evitan la necesidad de detecciones de pausas entre habla explícita. Martín [16], [17] usa el mínimo de la señal de potencia de la subbanda dentro de una ventana de tiempos de aproximadamente 1 segundo. Esta idea fue ya formulada por Paul [22].

Doblinguer [37] propuso un esquema de estimación de ruido continuo similar al de Martín pero es más eficiente computacionalmente. Hirsch [23] y Ehrlicher [38] propusieron un algoritmo que se basaba en la observación de que el valor de la magnitud espectral más común en voz limpia es casi cero. Si tenemos voz con ruido, los algoritmos miden la función de distribución de la magnitud espectral y determina los máximos que son normalmente usados como una estimación de la respectiva magnitud de ruido. Este tipo de algoritmos que evitan la detección de pausas para la estimación de ruido se supone que funcionan mejor frente a los ruidos no estacionarios, debido a que son más rápidos en su adaptación al cambio de niveles de ruido incluso durante la actividad de voz.

Por otro lado, la continua actualización de la estimación de ruido (reducción de ruido) es susceptible de capturar erróneamente energía de voz. Esto, sin embargo, afecta inevitablemente al deterioro de la voz por medio del proceso de reducción de ruido. Ficher y Stahl [39] investigaron sobre un algoritmo de reducción de ruido basado en la sustracción espectral mediante un esquema de actualización de espectro continuo de ruido. Encontraron que la degradación de la estimación de ruido por la voz era demasiado grande como para ser considerada y concluyen que la detección de actividad juega un papel muy importante y no puede ser omitida. Recientemente, en [40] se propuso utilizar estadísticas de cuarto orden de la señal ruidosa para estimar continuamente la voz y la energía del ruido. Muchos ejemplos usan señales de voz con ruido con SNRs positivas y ofrecen resultados prometedores, pero investigaciones adicionales requieren extender estos resultados para SNRs negativas y diferentes clases de ruidos. Como apuntaba Hirsch [23], "este es un problema muy difícil y sin resolver para situaciones reales con variación de nivel de ruido". Muchos estudios evaden el problema usando una detección de pausas ideal mediante señales de voz limpia o mediante pequeñas señales test con un periodo inicial de sólo ruido para la estimación sin necesidad de la actualización de la misma. En algunas aplicaciones como en la restauración de audio, la estimación de ruido puede ser hecha de forma manual. Sin embargo, en otras aplicaciones como la reducción de ruido para comunicaciones móviles requieren una actualización automática de la estimación del espectro de ruido. La mayoría de los autores están de acuerdo en que los detectores de actividad o de pausas,

respectivamente, son muy sensibles y frecuentemente limitan una parte de los sistemas para la reducción de ruido aditivo embebido en voz [31,41].

2.5.- Técnicas de clasificación.

Nos encontramos en la fase (3) del sistema completo de clasificación de patrones, tras la extracción de características (ver Fig.2.1). Se trata de realizar la representación de cada una de las clases consideradas en nuestro problema de clasificación (voz/no voz) [42], a partir de la información que se obtiene procedente de la extracción de características. Estas representaciones son normalmente patrones o modelos estadísticos. Durante años se ha investigado sobre qué técnicas son las que generan modelos más representativos. En este problema de clasificación consideramos al menos dos representaciones o modelos: uno para cada una de las clases consideradas.

La teoría de Bayes se puede usar en cualquier técnica de clasificación. Bayes habla de la teoría de la probabilidad condicionada. Para clasificar una trama con vector de características "Z" se deberán comparar los siguientes productos con probabilidad condicionada a los modelos de "voz" y de "no voz" y podemos calcular entonces las funciones de densidad de probabilidad de "voz" y "no voz":

$$P_{VOZ\_POSTERIORI}(Z) = \frac{P_{VOZ\_APRIORI}(Z) \cdot p\left(Z/W_{VOZ}\right)}{P(TOTAL)} \qquad (2.8)$$

$$P_{NOVOZ\_POSTERIORI}(Z) = \frac{P_{NOVOZ\_APRIORI}(Z) \cdot p\left(Z/W_{NOVOZ}\right)}{P(TOTAL)} \qquad (2.9)$$

donde $P(TOTAL) = P_{VOZ\_APRIORI}(Z) \cdot p\left(Z/W_{VOZ}\right) + P_{VOZ\_APRIORI}(Z) \cdot p\left(Z/W_{NOVOZ}\right)$.

Para clasificar una trama con vector de observación "Z" bastará con comparar los productos con probabilidad condicionada a los modelos de "voz" y de "no voz", es decir, bastaría con comparar los numeradores de las expresiones de ec. 2.8 y ec. 2.9 para la voz y la "no voz" respectivamente (el denominador es común), pudiendo

utilizarse información a-priori de las clases “voz” y “no voz”. Por lo tanto, en la práctica sólo compararemos los mencionados numeradores. Más aún, la tarea de decidir sobre la clasificación de una trama con vector de características “Z” dependerá, en general, de lo estrictos que queramos ser y la aplicación en que deba funcionar el detector de actividad:

- Normal: la trama con vector de observación “Z” será de voz sí y sólo sí $P_{VOZ}(Z) \geq P_{NOVOZ}(Z)$.
- Estricto con voz principal: es el típico caso que puede existir en codificación -> $P_{VOZ}(Z) \geq 1.2 \cdot P_{NOVOZ}(Z)$. Dá igual si se pierde alguna trama de voz.
- Poco estricto con voz principal. Por ejemplo, en verificación del locutor o reconocimiento de voz es preferible no perder tramas de voz: $P_{VOZ}(Z) \geq 0.8 \cdot P_{NOVOZ}(Z)$.

Por otro lado, un caso muy poco usado en la actualidad, pero quizá uno de los primeros cronológicamente hablando, es el de ajuste manual. Se trata de calcular el valor de los umbrales [60], típicamente de energía [61,63], la entropía [62] o de tasa de cruces por cero, a partir de los resultados que se obtienen de una muestra de ejemplos lo suficientemente representativos. El problema de esta ténica sencilla es que es poco robusta a cambios de nivel, tipo de ruido, etc.

En los siguientes subapartados se exponen las distintas técnicas de clasificación que un *VAD* puede utilizar: Modelos Ocultos de Markov (HMM) y Modelos de Mezclas de Gaussianas (GMM), Redes Neuronales Artificiales (ARN), K-Means y K-Nearest Neighbours (KNN) y, finalmente, Árboles de Decisión.

2.5.1.- Detectores que usan HMMs y GMMs.

Son detectores que utilizan un modelado basado en una aproximación a la función de densidad de probabilidad: el procedimiento de entrenamiento más común suele basarse en el uso de técnicas de Estimación de Máximas Verosimilitudes (EM). En este caso el problema de obtención de los modelos consiste en calcular los parámetros estadísticos de los mencionados modelos.

Modelar estadísticamente significa encontrar un patrón estadístico que pueda representar una sucesión de eventos comparables a otros futuros que puedan caer dentro de esa misma clase [43,44]. Por ejemplo, en el caso de reconocimiento de voz sería buscar patrones acústicos que generen una secuencia de símbolos. Y el mismo símbolo tendrá asociado siempre el mismo patrón acústico. En [45] se asume que cada componente espectral de voz y "no voz" posee una distribución Gaussiana compleja en la cual el ruido es aditivo y no correlado con la voz y se pueden calcular las funciones de densidad de probabilidad de ambos en función de sus medias, varianzas para el vector de características dado. Se trata por tanto de proponer modelos estadísticos a los que se ajusten las clases de "voz" y "no voz".

En [35] se utiliza un algoritmo adaptativo Bayesiano para el *VAD* debido tanto a su atractivo matemático tanto como a su uso acertado en otras tareas de reconocimiento. Además, se compara con otros algoritmos. Se supone que hay ***Q*** modelos distintos $\lambda_0, \lambda_1, ..., \lambda_{Q-1}$ que se van actualizando en tiempo real, tomando como modelos base los obtenidos de un entrenamiento previo. También, en [54] se parte de la hipótesis de que la regla de decisión óptima que minimiza la probabilidad de error es el clasificador de Bayes. Se crea un vector de observación que tiene en cuenta varias medidas (MO-LRT ≡ Multiple Observation Probability Likelihood Ratio Test), de tramas anteriores y posteriores, para obtener la clase de la trama actual. La decisión final se realiza en función de un umbral que a su vez depende de las razones de probabilidad anteriormente mencionadas. Por encima de ese umbral se decide que se está ante una trama de voz y por debajo de "no voz".

Otro trabajo también basado en técnicas Bayesianas es [44]. Se propone un criterio de aproximación de clustering e información Bayesiana para estimar los umbrales del *VAD*. Comparando con algoritmos previos, es más robusto y con reglas heurísticas libres. Se trata de un algoritmo de estimación mediante un criterio de aproximación de clustering e información Bayesiana. El algoritmo se puede aplicar a cualquier característica que tenga discriminación entre voz y ruido de fondo. Cada 20 milisegundos se extraen las características de cada trama. Estas características se organizan en clusters mediante medidas de aproximación de clustering. El criterio de información Bayesiana se usa para determinar el mejor

número de cluster. Hay dos clusters por intervalo conteniendo, uno de voz y ruido de fondo, y el otro cluster con sólo ruido de fondo. Y, finalmente, se usan para determinar el umbral del *VAD*. Son reglas libres heurísticas y se necesitan pocos datos para obtener una buena decisión.

Los *VAD* que usan Modelos Ocultos de Markov (HMM) [43,46] se caracterizan en que cada trama genera un vector de observación que a su vez invoca una secuencia de estados dependientes entre sí. Cada modelo puede representarse con uno o varios estados conectados entre sí. A su vez, la función densidad de probabilidad de cada estado del modelo se representa con un conjunto de funciones Gaussianas (mezcla de Gaussianas) con medias y varianzas diferentes (parámetros del modelado). La interacción de estados pertenecientes a modelos diferentes permite, no sólo conocer la clase del vector de características obtenido, sino también la evolución en el tiempo y las transiciones entre clases.
Los GMMs (Modelos de Mezclas de Gaussianas) [103] son un caso particular de HMMs. La particularización es que cada modelo posee un único estado: como cada trama simboliza un modelo, en este caso particular, cada trama vendrá representada por un único estado, y tramas consecutivas llevarán asociadas estados consecutivos independientes. En algunas ocasiones ocurre que además, cada estado contiene una sóla gaussiana: cada trama vendrá simbolizada por una única gaussiana, caracterizada por una media y una variaza.

En todos los casos se trata de proponer el modelo y hallar ciertos parámetros mediante un proceso llamado entrenamiento, como ya se ha comentado anteriormente. La problemática radica en muchas ocasiones en que existen sonidos no estacionarios difíciles de catalogar, por ejemplo voz de fondo. Por ello puede ser interesante tratar en este trabajo de Tesis la posibilidad de considerar una amplia variedad de modelos de “no voz”: modelo de ruido de coche, modelo de voces de fondo, etc.

2.5.2.- Detectores que usan Redes Neuronales Artificiales.

En este caso los parámetros de los modelos se estiman haciendo uso del criterio de menor error de clasificación. En las Redes Neuronales Artificiales, aunque hay parámetros que calcular, como por ejemplo los pesos de un perceptrón, se trata de clasificar ejemplos muy bien conocidos y hallar la función que mejor se ajusta

para clasificar esos ejemplos, aunque no se supone ninguna distribución de los datos. Un ejemplo de *VAD* usado para realizar la clasificación voz/no voz es [47]. En este caso se pretende encontrar los aspectos que discriminan a las clases “voz” y ”no voz” al igual que trata de estudiar la estructura más adecuada para la Red Neuronal.

Una red neuronal o perceptrón es un modelado basado en la interconexión de las neuronas en diferentes capas como aparecen en el cerebro. Esta representación consiste en una matriz de N niveles y M neuronas por nivel que son calculados de la siguiente manera:

- Sean los pesos P_i de la matriz.
- Conocemos las entradas y las salidas en todos los casos.
- Calculamos los P_i de tal manera que minimicemos el error en el conjunto, es decir, ajustamos los pesos para que se tenga la mejor aproximación en cada uno de los casos conocidos. En este proceso se utiliza el algoritmo de retropopagación partiendo de las salidas.

Trabajos muy recientes como [48] o [49] entrenan redes neuronales para obtener una óptima respuesta del sistema en el que se aplica. En el caso del *VAD*, se tratará de entrenar las clases de voz y “no voz” para conocer los pesos anteriormente mencionados de la matriz. Existe una gran variedad de topologías en función del número de capas que use: normalmente la naturaleza del perceptrón viene dada por la complejidad del sistema. Otro tipo de técnicas discriminantes un poco más complejas serían por ejemplo usar planos o hiperplanos, con más parámetros a calcular. Como caso particular nos encontramos los SVM (Support Vector Machines). Un ejemplo de la utilización de un detector basado en un perceptrón multicapa es [42]. En este caso se propone un método para diseñar clasificadores de voz y ruido usando el algoritmo adaptativo de empuje (AdaBoost). El método usa una combinación de clasificadores base simples a través del algoritmo AdaBoost y un grupo de características de voz optimizadas combinadas con la sustracción espectral. Un análisis más detallado de los pesos utilizados para la base de clasificadores reveló la contribución de cada componente característica. El método propuesto fue evaluado usando la base de datos del estándar Aurora para la detección de actividad y la detección de voz robusta.

En [42] se tiene la simplicidad del clasificador de voz/ruido del G.729 anexo B mientras se entrenan los hiperplanos en un principio y de manera automática usando el algoritmo AdaBoost. Todas las características se normalizan para tener media cero y varianza unidad a lo largo de cada eje. Usa un perceptrón como base del clasificador cuya magnitud de salida denota una medida de confianza.

Por último, es importante mencionar que la computación requerida es linealmente proporcional al número de clasificadores base o hiperplanos, que depende del tipo de aplicación.

2.5.3.- Detectores que usan K-Means y K-Nearest Neighbours.

En estos casos se tienen un conjunto de ejemplos etiquetados y se calculan las distancias a los mismos. Se suelen usar las siguientes distancias:

- Distancia euclídea: la distancia euclídea entre dos puntos en el espacio, aunque en general se pueden tener "*n*" coordenadas, es la distancia del segmento que los une. Por ejemplo, dados los puntos en el espacio A(a,b,c) y B(c,d,f), la distancia euclídea entre los dos puntos es la siguiente:

$$d(A,B) = \sqrt{(c-a)^2 + (d-b)^2 + (f-c)^2} \qquad (2.10)$$

- Distancia de Mahalanobis: es una distancia estadística que generaliza la distancia euclídea. En este caso se da la misma relevancia a todas las coordenadas, mientras que con la euclídea ocurre que prevalecen las diferencias de las coordenadas con variaciones más grandes. Muchos programas de matemática estadística proporcionan la distancia de Mahalanobis de los puntos muestrales $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{ik})$ con $i = 1,2,...,n$ al punto medio de la nube regresora $\bar{x}_i = (\bar{x}_{.1}, \bar{x}_{.2}, ..., \bar{x}_{.k})$ donde $\bar{x}_{.j}$, $j = 1,2,...,n$ es la media de los datos de la variable x_j. Esta distancia viene definida de la siguiente forma, en donde S es la matriz de varianzas-covarianzas del vector de variables $(x_1, x_2, ..., x_k)$:

$$d_M^2(\vec{x}_{i.};\bar{x}) = (\vec{x}_{i.} - \bar{x})S^{-1}(\vec{x}_{i.} - \bar{x})^t \qquad (2.11)$$

2.5.3.1.- Clasificación con K-Means.

Es un proceso de clasificación simple en el que se tiene un conjunto de ejemplos etiquetados. Estos ejemplos suelen representarse en un diagrama y calcular las distancias entre los mencionados ejemplos. Si existe un grupo A de ejemplos cercanos entre sí, pero alejados de otro B, se creará la clase de grupo A y se elegirá como muestra de esta clase al ejemplo cuyas distancias a los otros dentro del grupo sea mínima (la misma idea de un ajuste por mínimos cuadrados). Lo mismo se haría con la clase B, etc. Pueden existir casos en los que no este clara la distinción de clases porque las distancias son parecidas entre los ejemplos que se tengan. Sería un caso de difícil clasificación. Un ejemplo de la técnica aplicada a la detección de actividad de voz se tiene en [64].

2.5.3.2.- Clasificación con K-Nearest Neighbours.

En este caso se conocen también un conjunto de ejemplos de voz y no voz etiquetados. Una vez obtenido el nuevo ejemplo “E” a clasificar, se miden las distancias de este a cada uno de los ejemplos anteriores etiquetados y se eligen los K más cercanos. Posteriormente se revisa a qué clases pertenecen estos K ejemplos y la clase que más se repita será la más representativa para nuestro nuevo ejemplo “E”. Para que no haya ambigüedades, K se suele elegir impar. Por ejemplo si K=5 y los vecinos más cercanos de “E” pertenecen 3 a la clase “ruido” y 2 a la clase “voz”, se decide que “E” pertenece a la clase “ruido” ya que 3 > 2. Un ejemplo sería [65] donde se usan clasificadores basados en KNN (K-Nearest Neighbor).

2.5.4.- Detectores que usan Árboles de decisión.

Es una técnica de aprendizaje cuyo cometido es el extraer reglas a partir de los datos o ejemplos disponibles. Para la aplicación de estos sistemas de aprendizaje inductivo se parte de un conjunto de ejemplos. Cada ejemplo debe tener la misma estructura consistente en una conclusión (o decisión) y un número de características o atributos que definen esa conclusión o decisión. El sistema

construye un árbol de decisión que representa la relación existente entre la conclusión-decisión y sus atributos. Es decir, se produce un proceso de generalización de forma que el árbol de decisión generado “clasifica” correctamente los ejemplos dados. Este árbol, además, se caracteriza por ser el óptimo en el sentido que minimiza el número de atributos requeridos para alcanzar la decisión final.

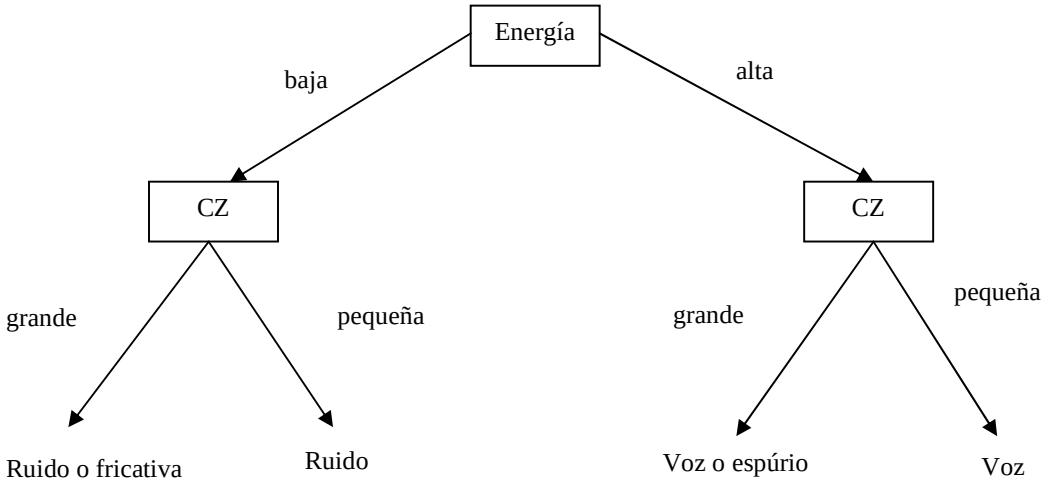

Figura 2.3. Árbol de decisión para un *VAD*.

Por ejemplo, el árbol de decisión que se presenta en la Fig.2.3 posee dos atributos o características:

- Energía (alta, baja).
- Tasa de cruces por cero (grande, pequeña) que lo denotaremos con CZ.

La decisión a tomar variará entre dos valores: voz y ruido.

El árbol representa la dependencia lógica entre la decisión a tomar y los atributos considerados. Serían por tanto cuatro los resultados posibles:

- Ruido o fricativa: energía baja + tasa de cruces por cero grande.
- Ruido: energía baja + tasa de cruces por cero pequeña.
- Voz o espúrio: energía alta + tasa de cruces por cero grande.
- Voz: energía alta + tasa de cruces por cero pequeña.

2.5.5.- Otras técnicas de clasificación.

Dada la complejidad de las técnicas de clasificación en los *VAD*s, existen múltiples aproximaciones. Por su originalidad, algunas aproximaciones dignas de mencionar son el uso de caos y fractales, y el uso de la lógica borrosa. En [50] los estudios giran en torno a la teoría de recurrencia de Poincaré (muy usada en mecánica estadística), que nos dice que si en cierto instante se han dado unas condiciones determinadas, si esperamos un tiempo lo suficientemente grande, se volverán a dar las condiciones originales o, al menos, muy cercanas a las originales. Esto nos habla de la reversibilidad del sistema. La hipótesis es entonces que: "el habla es reversible". [50] usa técnicas de análisis de series en el tiempo tanto lineales como no lineales. En este caso el número medio de los puntos de recurrencia de Poincare para cada bloque diseñado y para cada forma de la onda es considerado como una nueva característica [51,52]. Tras representar la curva característica, se toma de referencia un umbral adaptado para determinar los extremos basado en un modelado simple de la señal. El algoritmo supone varias ventajas: presenta una alta precisión, no empeora demasiado cuando se incrementan los niveles de ruido y, por último, no tiene la necesidad de estimar el ruido de fondo como es comúnmente requerido en otros algoritmos de detección de extremos. Los experimentos ascienden a un total de 600 dígitos aislados en inglés. Todas las locuciones fueron etiquetadas manualmente antes de los experimentos. Para generar los ficheros con ruido, se usaron distintos tipos de ruido disponible procedente de la Base de Información de Procesado de la Señal (SPIB) recogida por la Universidad Rice [53]. Tres clases de ruido fueron considerados: ruido blanco, ruido rosado y ruido de voces de fondo. Para preparar la base de datos con ruido para la evaluación, se añadieron señales de ruido a los datos de voz de forma que se obtuvieron distintos niveles de SNR incluyendo 5, 10, 15 y 20 dB.

La lógica borrosa es básicamente una lógica multievaluada que permite valores intermedios para poder definir evaluaciones convencionales como sí/no, verdadero/falso, negro/blanco, etc. En [47] se usa un Detector basado en Redes Neuronales combinado con lógica borrosa en reconocimiento de voz. También se da un enfoque hacia el reconocimiento de voz ruidosa. En este trabajo se muestra que la decodificación de una onda de voz ruidosa puede ser sencilla si el reconocedor posee conocimientos explícitos de dónde debería hipotéticamente haber voz, y

dónde debería trazar la acústica de los segmentos de ruido. El Detector de voz y ruido usaba su salida como una característica más del front-end en sistemas de reconocimiento. Se muestra que mediante un apropiado peso, la contribución de esta característica en el reconocedor modificando los modelos acústicos por consiguiente, se pueden penalizar las confusiones de voz/ruido (detección Fuzzy (blanda o suave) de la decisión de la voz) y consecuentemente reducir la tasa de error de reconocimiento. Este sistema reduce el error total en una gran variedad de tareas de reconocimiento y ruidos característicos sin degradar su funcionamiento en condiciones de voz limpia.

2.6.- Decisión.

La fase de decisión, bloque (4) del sistema completo de clasificación de patrones (Fig.2.1), consiste en el proceso en el que se obtiene el resultado de clasificación. El resultado de esta clasificación se puede realizar a dos niveles: nivel de trama o nivel de pulso. La decisión a nivel de trama (Fig.2.4) implica la decisión de la trama actual sin tener en cuenta el histórico de resultados de tramas anteriores o la estructura de lenguaje. Esta decisión se puede obtener de tres maneras diferentes:

- Decisión directa a partir de las características: este caso es por ejemplo el de comparar un simple umbral con el valor obtenido en la trama actual.
- Mediante modelos generativos: máxima verosimilitud entre clases o LLR (Log-Lilelihood Ratio). En este caso se analiza la trama actual y se compara con los modelos pertinentes. Es una comparación entre las medidas del parecido del vector a clasificar con el modelo de cada clase. Se elegirá la clase que ofrezca una medida de parecido mayor.
- Mediante técnicas discriminantes. La trama actual se clasifica según el criterio de mínimo error de clasificación (ver apartado 2.5.2).

Por otro lado, la decisión a nivel de pulso sí que tiene en cuenta el histórico de resultados de tramas anteriores y la estructura de lenguaje (Fig.2.5).

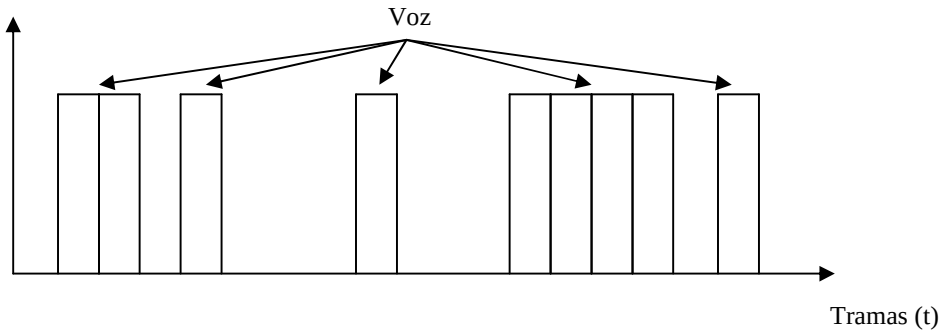

Figura 2.4. Decisión a nivel de trama.

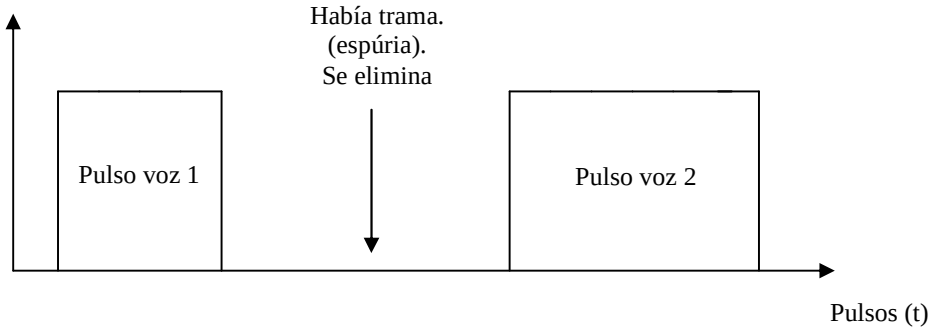

Figura 2.5. Decisión a nivel de pulso.

En la Fig.2.5 aparece el concepto de pulso de voz: agrupación de tramas que trata de simbolizar una pronunciación con un extremo inicial y otro final. De esta manera, la Fig.2.4 con decisión a nivel de trama se podría convertir en la Fig.2.5 a nivel de pulso: tramas con actividad aisladas suelen ser espúrios (golpes, clicks), o por ejemplo, tramas con actividad cercanas a varias tramas con actividad consecutivas suelen ser parte de una pronunciación.

La decisión a nivel de pulso se puede obtener de:

- Características sencillas: parámetros que se basan en las características comunes de las señales de habla, por ejemplo la duración mínima de pronunciación, silencios inter-silábicos, etc.
- Modelos generativos: aplicación de las teorías de Bayes. Consiste en el proceso de clasificación según las medidas de parecido obtenidas en la etapa del apartado 2.5.

- Técnicas discriminantes, como, por ejemplo, la creación de una Red Neuronal con procesamiento a nivel de pulso.

A continuación se realiza una posible clasificación de los VADs en función de cómo obtiene la decisión a nivel de pulso.

2.6.1.- Detectores que usan características sencillas para la decisión a nivel de pulso: recursos lingüísticos.

En este caso nos valemos de las características comunes de las señales de habla para resolver o mejorar nuestro problema de detección, por ejemplo, una palabra debe siempre estar acentuada por lo que es esperable que el nivel de energía sea mayor en una de las sílabas que en el resto, debe haber al menos un golpe de voz, o si se tiene fricativa al final de pronunciación es esperable una tasa de cruces por cero elevada. Veamos dos ejemplos:

Detector "top-down" con condiciones sintácticas y semánticas

El detector "top-down" fue definido por los investigadores Wilpon, Rabiner y Martín [59]. Este *VAD* "top-down" también trabaja con la energía normalizada. El proceso de cálculo comienza por la búsqueda de la trama de máxima energía en la grabación. Partiendo de esta trama, busca la trama anterior donde la energía desciende de un umbral alto k_1, en este caso "k_i" indica la energía en dB, y la posterior que desciende de un umbral bajo k_3, habiendo encontrado las tramas posibles de comienzo y final de pulso. Para eliminar ruidos espúrios se revisa el nivel de energía de las IT1 (recuérdese que IT1 es ITL con L=1 del caso del detector basado en niveles de energía y tasa de cruces por cero con la única diferencia que ahora se utilizan unidades de trama y no de tiempo) primeras tramas y las IT2 (ITL con L=2) últimas tramas (véase apartado 2.4.1). Para que el pulso sea válido, su duración y amplitud deben superar unos umbrales mínimos. El algoritmo se repite para el resto de la locución encontrando un cierto número total de pulsos válidos.

El conjunto de los mencionados pulsos se combinan tomando como referencia el de máxima energía. Para añadir un pulso anterior al actual, la pendiente de la energía de bajada (definida sobre las últimas tramas del pulso) debe ser mayor que un umbral. Además, deben distar menos de un número de tramas predefinido de antemano (condiciones de una oclusiva). De forma análoga se

combinan también pulsos posteriores al actual y su pendiente de subida también ha de ser mayor que un cierto umbral. La duración de los pulsos combinados debe ser menor que un valor máximo para no unir palabras distintas, y mayor que un umbral mínimo de número de tramas.

En función del vocabulario (dígitos del 0 al 9 en este caso) incorpora un par de condiciones sintácticas y una semántica. La primera sintáctica implica que no se puede añadir un pulso después del pulso de máxima energía (1 sola palabra). Por otro lado la condición semántica que se debe añadir es que el número de pronunciaciones debe ser exactamente igual al número de picos de energía máxima de la locución ya que cada palabra lleva un acento.

Detector de umbrales de energía y tasa de cruces por cero

En el detector sencillo que proponen Rainer y Sambur, la decisión viene dada en función de una combinación de umbrales de energía y de tasa de cruces por cero (véase el apartado 2.4.1). El diagrama de energía se puede visualizar en la Fig.2.6.

En este caso IF es una constante de valor 25 (cruces por cero), ajustada manualmente, e IMX es el pico de energía de la grabación incluyendo la voz principal. El comienzo del pulso, trama N_1, es el punto en el que la energía excede ITL (umbral de energía bajo):

- Si se supera ITL, estamos ante un posible pulso pero todavía no se da ninguna confirmación debido a que podemos estar ante un pulso espúrio.
- Una vez superado ITL, si además se supera ITU, ya se da la confirmación de pulso y se confirma también inicio de palabra. Se toma como inicio de pulso el instante de tiempo (N_1) en el que supera ITL. Algo análogo ocurre con el final de pulso N_2.

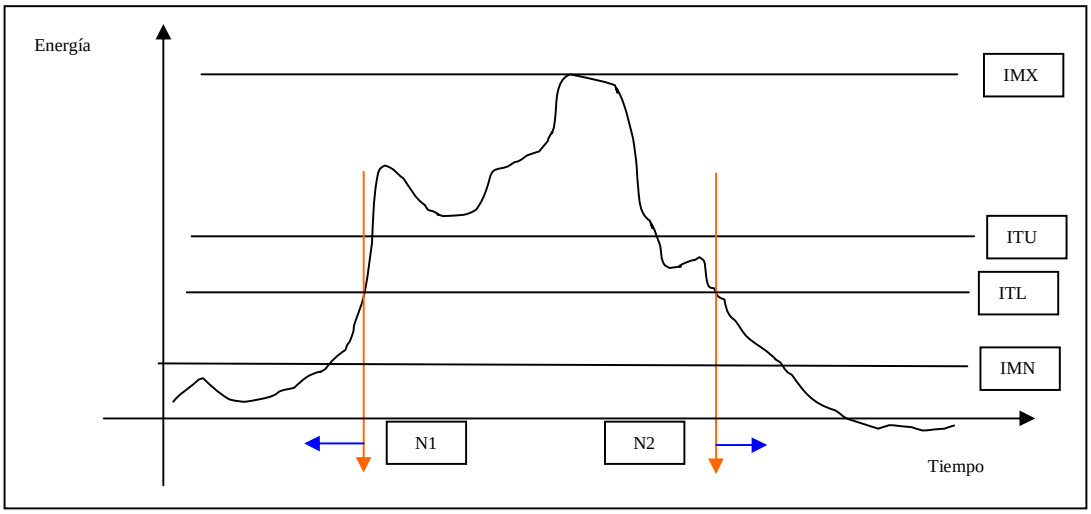

Figura 2.6. Diagrama de energía.

En este caso IF es una constante de valor 25 (cruces por cero), ajustada manualmente, e IMX es el pico de energía de la grabación incluyendo la voz principal. El comienzo del pulso, trama N_1, es el punto en el que la energía excede ITL (umbral de energía bajo):

- Si se supera ITL, estamos ante un posible pulso pero todavía no se da ninguna confirmación debido a que podemos estar ante un pulso espúrio.
- Una vez superado ITL, si además se supera ITU, ya se da la confirmación de pulso y se confirma también inicio de palabra. Se toma como inicio de pulso el instante de tiempo (N_1) en el que supera ITL. Algo análogo ocurre con el final de pulso N_2.

Estos extremos son conservativos y se refina la localización de los extremos con la tasa de cruces por cero. El algoritmo examina desde N_1 hasta N_1 – 25 (250 milisegundos) y si el número de veces que se excede el umbral IZCT es tres o más, el punto de comienzo se retrasa al primer punto (en el tiempo) en el que excede el mencionado umbral. De lo contrario, el comienzo se mantiene en N_1. De igual manera se opera en el intervalo comprendido entre N_2 y N_2 + 25 (250 milisegundos).

El cuanto a los umbrales de energía: se configuran los umbrales en función de lo robusto que se quiera ser ante espúrios (por ejemplo umbrales de energía ITL e ITU más altos). Es importante comentar que se pueden tratar los umbrales de

energía que se quiera, tres, cuatro, en lugar de dos como en el caso del ejemplo (ITL e ITU).

2.7.- Aplicaciones de los detectores de actividad.

Es necesario comentar que existe una diferencia muy importante a tener en cuenta en los detectores de actividad dependiendo de su uso o aplicación:

- Detectores usados en comunicaciones (codificación): por ejemplo los *VAD*s de los códecs EFR GSM [21] o AMR2 [32]. Es un caso en el que se trata de transmitir a velocidades bajas cuando no hay voz y evitar así las interferencias radio. Por tanto, en este caso importa menos si se pierde alguna trama de voz principal.
- Detectores usados en ASR (Reconocimiento Automático de Habla): por ejemplo el *VAD* del Front End AURORA [77]. En este caso se trata de "no" perder segmentos de la voz principal: somos más tolerantes con la voz, aunque esto implique dejar pasar algún ruido..

Por lo tanto, cabe decir que en codificación el objetivo es eliminar el ruido (no-voz) mientras que en reconocimiento el objetivo es no eliminar la voz. Por eso en reconocimiento puede ser crucial perder tramas de voz mientras que en codificación somos más tolerantes.

2.7.1.- Detectores usados en comunicaciones.

En esta sección vamos a realizar una breve introducción de los detectores que se usan en sistemas de comunicaciones [66,69].

Dentro del amplio mundo de los sistemas de comunicaciones móviles GSM, los *VAD* se usan en transmisión discontinua para alargar la vida de las baterías de las unidades portátiles. También se utiliza en CDMA (Code Division Multiple Access) y PCS (Personal Communications Services) en la tasa de bit variable (VBR-Variable bit rate) para reducir las interferencias. El *VAD* es indispensable en cualquier codificador VBR (Variable Bit Rate) para controlar el promedio de la tasa de bit y la calidad del codificador. Recientemente, se han descrito varios procedimientos de detección de actividad para diferentes aplicaciones incluyendo servicios de comunicaciones móviles [2], transmisión de voz en tiempo real vía Internet [7] o

reducción de ruido en dispositivos digitales para el oído [69]. Las investigaciones más importantes han ido encaminadas al desarrollo de algoritmos robustos, con especial atención en el estudio y derivación de las características robustas de ruido y las reglas de decisión. Sohn y Sung (1998) presentaron un algoritmo que usa una adaptación del espectro de ruido empleando reglas de decisión. Una versión posterior, Sohn en 1999, del *VAD* original fue añadirle un esquema de hang-over, el cual considera las observaciones previas de un proceso de modelado de voz de Markov de primer orden. Este algoritmo fue superado por el *VAD* del codificador G.729B (1996), un codificador de alta calidad, en términos de detección y probabilidad de falsas alarmas. La recomendación de este decodificador usa un vector de características que consiste en una predicción lineal del espectro (LP), la energía de toda la banda, la energía de la banda estrecha (definida como la energía que hay entre 0 y 1 KHz) y la tasa de cruces por cero (ZCR). Otras investigaciones presentadas mejoran el algoritmo propuesto por Sohn en 1999: Cho en 2001 [70]. Cho y Kondoz presentaron un sistema que usaba la razón de probabilidad suavizada para disminuir los errores de detección, ofreciendo mejores resultados que el G.729B y con una mejora comparable al AMR. Por otro lado Cho también propuso una regla de decisión mixta basada en adaptaciones de ruido que ofrecía mejores resultados que su anterior técnica de adaptación del año 1998.

Los códecs para VoIP, por ejemplo el H.323, también necesitan *VAD*s. En [104] se comparan distintos algoritmos de detección para VoIP.

2.7.2.- Detectores aplicados a reconocimiento de voz.

Los *VAD*s son una parte imprescindible dentro de los sistemas de Reconocimiento de Habla porque son los que van a decidir qué tramas son las que se van a entregar al Reconocedor de Voz. Como se ha dicho anteriormente, en un sistema de este tipo es deseable no perder ninguna trama de voz principal.

La gran cantidad de servicios vocales que existen en la actualidad hace necesario el estudio continuo de los reconocedores de voz: automatización de servicios de consulta, portales de voz etc. De esta manera, cada vez es más usual encontrar interfaces hombre-máquina controlados por voz. El reconocimiento de voz supone que la señal está representada por una secuencia de palabras, símbolos y fonemas, que se pueden obtener a partir de la forma de onda de la señal.

Recientemente, el ETSI ha aprobado un nuevo estándar (AFE, Advanced Front End) de reconocimiento de voz distribuido que incorpora: métodos de supresión de ruido para la extracción de características, incorpora un *VAD* basado en la energía, filtrado de Wiener para la estimación del espectro de ruido y otro *VAD* más sofisticado para realizar el eliminado de tramas (frame dropping).

Dentro de los detectores que se usan en reconocimiento de voz, otra clasificación, según su modo de funcionamiento, sería la siguiente:

- Explícitos o independientes del reconocimiento (Fig.2.1). Detección de voz independiente de las fases de reconocimiento.
- Implícitos o dependientes del reconocimiento (Fig.2.7). Detección de voz completamente integrada en la fase de reconocimiento [73-76].
- Híbridos o parcialmente integrados en la fase de reconocimiento (Fig.2.8). Lamel propone en [72] un Detector de Actividad Híbrido.

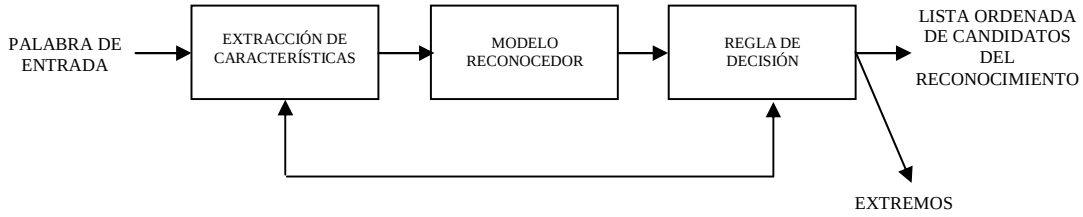

Figura 2.7. Diagrama de bloques de un *VAD* Implícito.

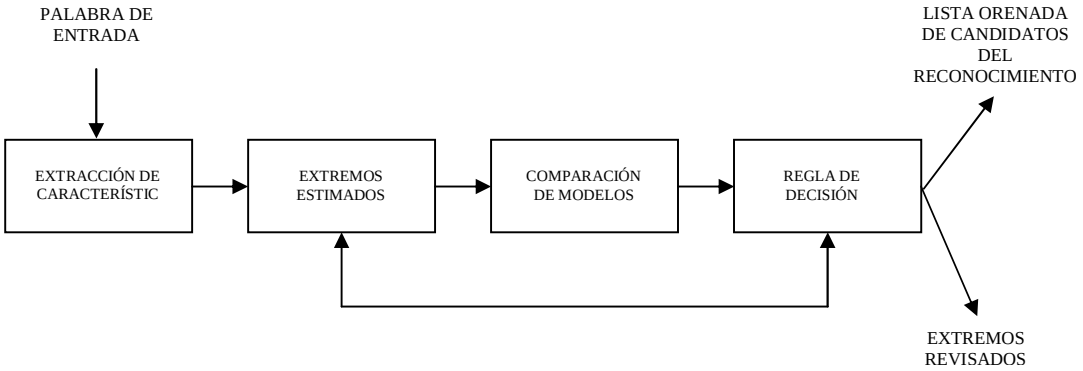

Figura 2.8. Diagrama de bloques de un *VAD* Híbrido.

Las garantías de acierto de un sistema de reconocimiento de voz serán mayores si los segmentos de voz están bien seleccionados. Así pues, en este caso, los objetivos básicos van a ser principalmente:

- Evitar una carga computacional alta del reconocedor en instantes de no actividad para así por ejemplo poder aumentar el número de canales posibles en un servidor de telefonía.
- Evitar los errores o falsas alarmas del reconocedor sin olvidar que no se pueden perder tramas de pronunciación.

Los errores de los detectores de actividad pueden causar, por un lado, la pérdida parcial o total del mensaje pronunciado por el usuario, con nulas probabilidades de éxito en el proceso de reconocimiento, y, por otro, la aceptación de sonidos indeseados capaces de ser confundidos con unidades lingüísticas como fonemas, bifonemas o trifonemas. En [59] se puede visualizar un experimento en el que se observa la tasa de reconocimiento según el error en la localización de los extremos.

Por otro lado, y como ya se apuntaba al principio del capítulo, la buena operatividad del *VAD* es básica para el funcionamiento de los reconocedores de voz en tiempo real, ya que la información que se procesará será únicamente la que provenga de las tramas donde el Detector haya decidido que hay voz, mientras que las tramas de ruido serán desechadas. De esta manera la CPU se libera en los momentos de no existencia de pronunciación, permitiendo, por ejemplo, compartir los recursos de la máquina para atender varias líneas de reconocimiento simultáneas y sobre todo no perder voz.

En cualquier conversación, la media del tiempo en la que se está enviando voz es del 50%, debido a las pausas introducidas entre palabras, intervalos de toma de aire, intervalos de toma de decisiones, etc. En servicios automáticos de voz que existen en la actualidad, el 95% del tiempo de interacción con el sistema es silencio. En este caso el ahorro de tiempo de procesado sería enorme. Por todo esto, en los sistemas de transmisión multicanal analógico se usa una técnica denominada Time Assignment Speech Interpolation (TASI) para asignar un canal no usado sólo cuando el Detector encuentra voz. Así, un medio de transmisión con 96 canales de voz puede atender a unos 235 clientes. Un ejemplo de multiconferencia sobre voIP es [105].

En [71] se trata del problema que existe en el reconocimiento de voz a largas distancias en entornos de ruido y reverberación. El algoritmo propuesto se basa en el mismo análisis de la potencia de cruce del espectro de fase (técnica de estimación del retardo) que el usado para localización de locutor. Es más eficiente que el basado en energía. La aplicación nos habla de la no linealidad de las tramas de “no voz” mediante la actualización dinámica de características. Requiere un modelo de ruido estadístico bastante bueno.

Recientemente ha aparecido el estándar de la ETSI Aurora [77], ETSI ES 202 050 V1.1.1. El estándar se caracteriza por realizar la extracción de características en el terminal y no en el servidor, estando, por tanto, orientado hacia reconocimiento distribuido (DSR). Se usan parámetros MFCC, se comprimen estas características para conseguir una tasa de transmisión más baja y por último se decodifican en el servidor. Además, está preparado para muestreos de 8 y 16 Khz, el último por la aparición de la nueva tecnología UMTS. Respecto al *VAD*, en 8 Khz lo usa en la extracción de características y dentro de un sistema de reducción de ruido. En [36] se realiza una pequeña modificación al mencionado estándar Aurora de la ETSI y se realiza un *VAD* basado en un perceptrón.

2.7.2.1.-VAD utilizando un ASR para la segmentación de tramas.

El uso de detectores de actividad para la segmentación de tramas de voz surge por el hecho de que en ciertas aplicaciones con interacción hombre-máquina, el sistema tiene que ser capaz de saber cuándo el locutor está hablando. Además, el sistema debe tener en cuenta que es importante no enviar las tramas de ruido al decodificador. Existen algoritmos de segmentación robustos que mejoran el funcionamiento del reconocedor de voz, particularmente en ciertas condiciones de ruido difíciles [55, 56]. Por otro lado, en aplicaciones donde el sistema pregunta al usuario, la voz de entrada es naturalmente segmentada en respuestas específicas del usuario (frases). En ASR se graba sólo lo importante (voz principal): un buen algoritmo de segmentación puede reducir sensiblemente la cantidad de datos grabada en el servidor que esta procesando múltiples clientes a la vez. Además, un algoritmo general de segmentación debería ser capaz de operar en condiciones donde la SNR varíe considerablemente y la entrada esté corrupta con ruidos como tonos de teléfono etc.

En ASR, las características de entrada utilizadas para reconocer pueden ser vectores cepstrales (por ejemplo [55]) o características especiales como correlaciones cruzadas normalizadas (por ejemplo [56]). La segmentación se realiza mediante el algoritmo de Viterbi sobre la completa secuencia de vectores de entrada. Esto los hace poco convenientes para aplicaciones de ASR en tiempo real. En ciertas tareas de ASR en tiempo real como los subtítulos de noticias de difusión, es posible evitar la segmentación explícita en el front-end. En lugar de eso, la sentencia completa de entrada se manda al reconocedor, y este usa la última parte del reconocimiento continuo para obtener la palabra reconocida o la secuencia de segmentos [57, 58]. Cuando la entrada contiene ruido de fondo que no ha sido tratado en los datos de entrenamiento, este método podría incrementar potencialmente los errores del reconocedor.

2.8.- Sistemas similares al propuesto en el trabajo de Tesis.

En primer lugar empezaremos hablando del detector más general que usa modelos ocultos de Markov (HMM). Un ejemplo de *VAD* basado en modelos ocultos de Markov (HMM), similar al desarrollado en este trabajo de Tesis, es el expuesto en [46]. La señal de entrada se muestrea a 8 KHz por limitaciones del canal. La voz se analiza en tramas de 32 milisegundos, solapadas 16 milisegundos, con un posterior filtrado de preénfasis que en referencia a nuestro esquema general de detección pertenecería al bloque de preproceso. Se usa una ventana de Hamming para suavizar las transiciones entre tramas sucesivas y el filtrado de preénfasis.

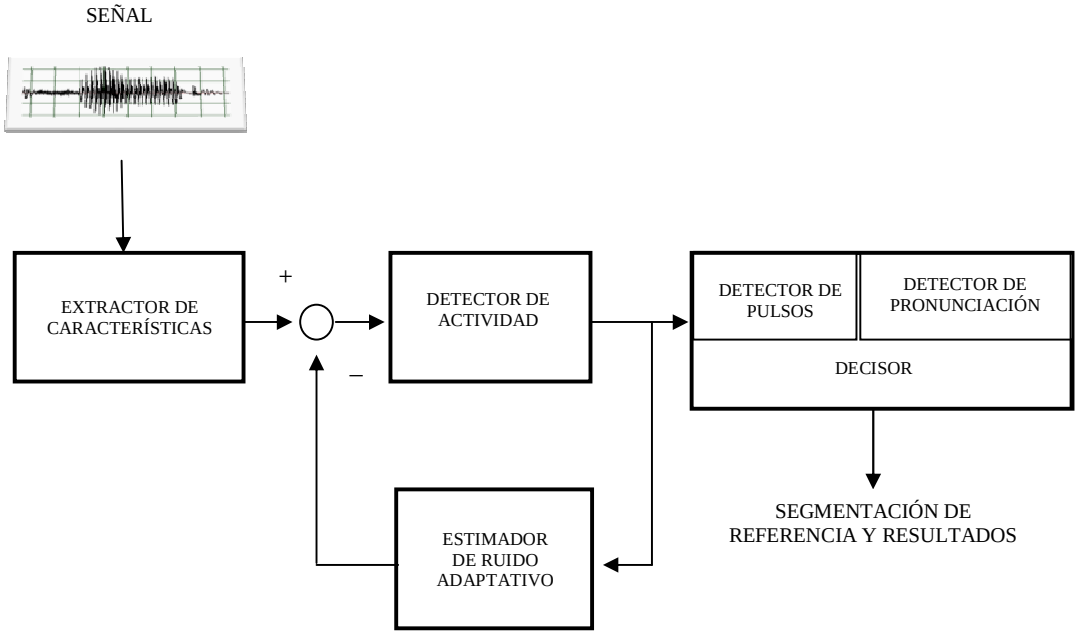

Figura 2.9. Diagrama de bloques similar al de Acero [46]. *VAD* basado en Modelos Ocultos de Markov.

El detector de la Fig.2.9 [46] consta de cuatro módulos: uno extractor de características, otro de detección de actividad, el de adaptación al ruido y el de la toma de decisión de extremos. Se trata de un esquema de detección completo, similar al propuesto en este trabajo. El módulo de extracción de características calcula los parámetros que utiliza el *VAD* en cada trama. Los mencionados coeficientes (coeficientes cepstrales) se obtienen a partir de un filtrado de la voz (banco de filtros Mel). Se usa la energía y su derivada (velocidad). Se usan tramas de 32 milisegundos y 256 muestras (8 KHz). Así pues, el Detector de Actividad decide si la trama corresponde a un segmento de “no voz” o uno de voz. Para ello se modela mediante modelos ocultos de Markov las clases de voz y “no voz” siguiendo los principios básicos de la detección de actividad implícita. Se usa un HMM para modelar ruidos de diversos tipos, estacionarios o no, y otro para modelar la voz. Los HMMs utilizan funciones de densidad de probabilidad continuas en cada estado. Se emplean gaussianas multivariadas para modelos de distribución de un vector de características compuesto por la energía y la variación de la energía. El

modelo de ruido posee tres estados y el de voz cuatro. El entrenamiento de los modelos de Markov se realiza con bases de datos que contienen ficheros de voz. Para decidir el comportamiento del *VAD* sobre la señal de entrada se realiza una búsqueda de Viterbi usando la red mostrada en la Fig.2.10:

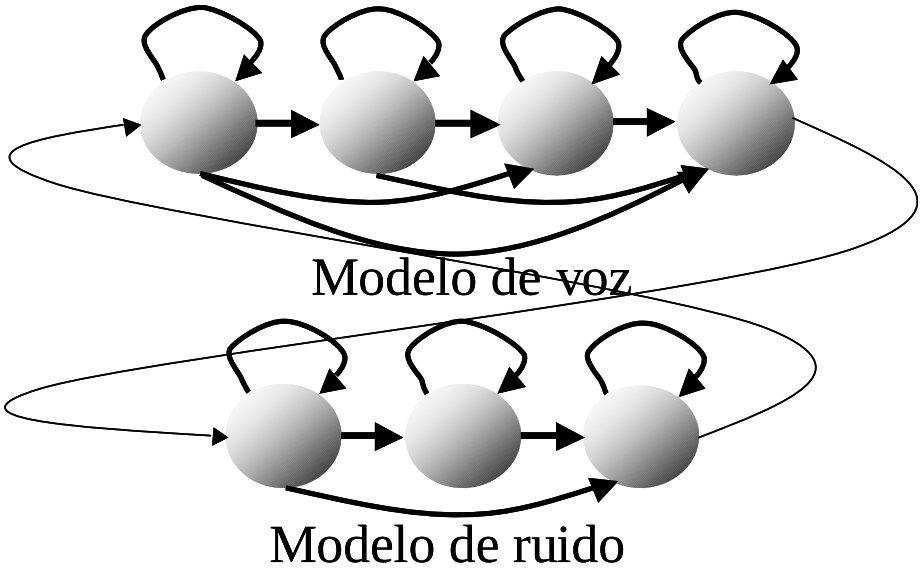

Figura 2.10. Red de HMM con las transiciones entre modelos de voz y de ruido [46].

Se detecta el comienzo de un pulso cuando el modelo de voz gana durante L tramas seguidas. De la misma forma, cuando gana el modelo de “no-voz”, durante K tramas seguidas se habrá detectado el final del hipotético pulso. Para que un pulso de voz se considere válido, se exige que el modelo de voz gane durante un número de tramas determinado al modelo de ruido. De igual manera se impone una duración mínima del pulso de aproximadamente 160 milisegundos. Por último, el detector de pronunciación considera que ésta ha finalizado cuando el silencio que sigue a un pulso excede un umbral predeterminado de antemano. El detector de pulsos hace una detección más fina de extremos mediante el número de cruces por ceros de la señal de score.

En cuanto a la adaptación de características, para aumentar la robustez del *VAD*, se han propuesto numerosas técnicas. En el trabajo de Ramirez [80], los autores solucionan el problema de la independencia de la relación señal a ruido

(SNR) usando la medida de la divergencia de Kullback-Leibler (distancia de Kullback-Leibler). En este trabajo de Tesis también se usan HMMs cuyo funcionamiento también es independiente de la SNR.

En cuanto a los modelos, en [43] también se habla de la obtención de modelos de voz y de ruido con HMM mediante un análisis cepstral. El problema radica en el funcionamiento del sistema en tiempo real. McAulay [78] investigó sobre un *VAD* partiendo del algoritmo de Roberts [79]. La implementación se basaba en una reestimación del modelo de "no voz" durante las tramas de silencio. Sin embargo, el algoritmo de Robert requería un largo periodo de entrenamiento para estimar el ruido y calcular el valor del umbral. Además, con altos niveles de ruido de banda estrecha este algoritmo encontraba dificultad en discriminar entre periodos de voz y de ruido o silencio. Floc'h utilizaba las medidas de las distancias espectrales entre la trama actual y el promedio del espectro de baja energía que son los periodos de silencio o ruido que se espera para calcular el promedio para el ruido. Las medidas de la distancia se calculan mediante este promedio y el espectro de la trama actual, decisión basada en un umbral fijo determinado experimentalmente.

En cuanto a las técnicas de decisión, el uso de la energía es una característica muy efectiva en cualquier condición para la detección de un pulso de voz. El mayor problema es ajustar los umbrales de energía para las diferentes SNRs. Para solucionar este problema, en [81], los autores entrenan diferentes modelos de ruido para las distintas SNRs y se propone un algoritmo de decisión para elegir el modelo basado en la SNR estimada en cada trama. Sin embargo se trata de un proceso de alta carga computacional y difícil de implementar. Como se verá más adelante, el *VAD* propuesto usa sólo un modelo para voz y otro para ruido para todas las SNRs, reduciendo de esta manera la complejidad.

Por otro lado, y mejorando la propuesta de Sheikhzadeh [81], Acero [82] propone la idea de usar el logaritmo de la energía normalizada para evitar entrenar distintos modelos en función de la SNR. El trabajo de Acero es la base del *VAD* propuesto en esta Tesis. El *VAD* de Acero usa un algoritmo basado en HMMs y un mecanismo de detección de pulsos como técnica de post-procesado simple y basada en dos umbrales, en lugar de cuatro como hace el algoritmo usado por Lamel [72].

En el *VAD* propuesto se extiende el vector de características añadiendo tres cepstrum al que usaba Acero, dotando así de información espectral a los modelos. La estimación del logaritmo de la energía se basa en una versión simplificada de la que usa el códec AMR1: el logaritmo de la energía normalizada se usa tanto para el modelo de ruido como el de voz. El nivel de ruido, necesario para el cálculo del logaritmo de la energía normalizada, se adapta en tiempo real durante las tramas de ruido (no durante las tramas de voz). Otro aspecto importante que Acero no consideró fue que el cálculo en pre-proceso del logaritmo de la energía normalizada (antes de entrenar los HMMs) no se puede obtener si dicho cálculo necesita información de post-entrenamiento (la estimación del logaritmo de la energía normalizada de ruido en el *VAD* de Acero necesita información de la detección, y eso es un problema para entrenar los HMMs). El mismo problema ocurre en el trabajo de Qi Li [83] que usa las marcas de final de pulso de voz para normalizar la energía secuencialmente. Otro *VAD* que usa información espectral es el de Zhang [84)]. Zhang, considerando la idea de que la información lingüística juega un papel muy importante en la detección de actividad, presenta un *VAD* basado en HMMs de cinco estados. Además, estos HMMs usan en su vector de características MFCCs, energía a corto plazo y la tasa de cruces por cero, pero sin incluir ninguna información ni del logaritmo de la energía normalizada ni del delta de energía. En [85] se presentan dos técnicas de clasificación, SVM y GMM, usando el retardo de grupo modificado. Dos modelos distintos, uno de voz y otro de ruido, son presentados por el clasificador, parecido al de nuestro trabajo pero con un vector de características distinto.

Es importante insistir en que el *VAD* propuesto tiene como aplicación principal el reconocimiento automático de habla. Por ejemplo Shon [89] propone un *VAD* similar al propuesto en este capítulo: usa un *VAD* que se basa en modelos estadísticos incluyendo un esquema efectivo de hang-over que tiene en cuenta las observaciones previas por medio de un proceso de Markov de primer orden para modelar las regiones de voz.

Finalmente, y como se expondrá más adelante, uno de los aspectos más importantes de este trabajo es el estudio y la incorporación de nuevas características en un *VAD* basado en HMMs para conseguir eliminar los pulsos que provienen de las voces de fondo. En esta línea, trabajos muy actuales vienen siendo

propuestos por los investigadores, por ejemplo, en [86] se presenta un esquema basado en el espectro de la señal para detectar la presencia de un locutor principal distinguiendo entre voz cercana y voz lejana. Otros trabajos recientes usan características parecidas a las propuestas: en [87] se usa el “pitch” como característica para mejorar la calidad de la voz en el códec AMR.

CAPÍTULO 3

BASES DE DATOS Y MEDIDAS DE EVALUACIÓN

3.1.- Introducción.

En este capítulo se detalla la descripción de las bases de datos usadas para la realización de este trabajo de Tesis, así como de las medidas de evaluación y la forma de calcular los errores a la hora de realizar los análisis comparativos pertinentes entre los detectores de actividad estudiados, y otros detectores estándares que nos sirven como referencia de uso para el campo que abordamos.

En primer lugar se establecen una serie de pautas posibles para evaluar un *VAD* con algunas referencias bibliográficas que lo apoyan. Posteriormente se describen las distintas bases de datos a utilizar, clasificadas en tres grandes grupos: bases de datos de entrenamiento, bases de datos de desarrollo y, finalmente, bases de datos test. Para finalizar el capítulo se presentan tanto las medidas de evaluación que serán usadas en este trabajo de Tesis, por ejemplo la tasa de falsas alarmas o la tasa de falsos rechazos, como los cuatro *VAD*s de referencia con los que será comparado el nuestro, el del codificador G729 anexo B, el del estándar AURORA y los de los codificadores AMR1 y AMR2.

3.2.- Evaluación de detectores.

Es importante establecer una serie de pautas a la hora de evaluar cualquier Detector de Actividad. Por ello, en esta sección se recogen los principales métodos y métricas para evaluar los *VAD*s de referencia y usarlos en este trabajo de Tesis:

- Las bases de datos consideradas. Antes de desglosar este punto es conveniente presentar algunos trabajos y bases de datos de referencia. Por ejemplo, en [66] se realizan pruebas en distintos entornos de ruido: autobuses, calle o restaurantes. Además se utiliza la base de datos NOISEX-92 [59,85] que posee entornos con ruido de coche y voces de fondo. El ruido se añade digitalmente a voces limpias terminando con SNRs de 20, 10 y 0 dB. También se usa esta base de datos en [59] y [85]. Otras bases de datos de referencia son AURORA-2 y AURORA-3 [36,14]. Como se indicó anteriormente en este trabajo utilizaremos bases de datos de entrenamiento, desarrollo y test:
    1. Base de datos de entrenamiento etiquetada, en aquellos casos en los que se tenga que entrenar HMMs, GMMs, una red neuronal, un árbol de

decisión etc., con una gran diversidad de muestras de voces y ruidos: voces de hombres, mujeres, de distintas edades, o ruidos de tipo estacionario o no estacionario.

2. Base de datos de desarrollo etiquetada, que contendrá el número de ejemplos necesario para poder realizar los estudios pertinentes. En este trabajo de Tesis estos estudios se basan en el análisis de un conjunto de características que pretenden discriminar entre la voz de fondo y la voz del locutor principal.
3. Base de datos test etiquetada [11], con distintas SNR y con ruidos de diversos tipos (de alto nivel de energía, de bajo nivel de energía, de fondo donde también se incluyen las voces de fondo, etc.). Es la que da las prestaciones finales del sistema de detección.

- Las distintas métricas a través de las que se representa matemáticamente la sensibilidad de cualquier *VAD* son las siguientes:
    1. Elaboración de dos histogramas de la diferencia entre marcas reales e ideales, uno que nos represente la bondad del Detector en el inicio de pulso y otro que lo haga en el fin de pulso de la voz. La forma de operar sería la de comparar los siguientes datos:
        - Marcas de la lista de ficheros de audio “prueba” etiquetados manualmente (inicio y fin de voz), que de alguna forma sería el resultado óptimo para el Detector.
        - Marcas reales que nos da en tiempo real nuestro *VAD*.
    2. Si el Detector se usa en aplicaciones de reconocimiento, otra forma de evaluarlo sería ver la tasa de aciertos del reconocedor, ya que normalmente, un reconocimiento con una alta tasa de aciertos va asociado a una precisión en los extremos del Detector. Se puede por tanto hallar la tasa de error de palabra (WER ≡ Word Error Rate) [20] en función de la detección utilizada. En este caso es necesario que el reconocedor de voz que evalúe sea lo suficientemente robusto y fiable como para que el peso del efecto de degradación de los resultados recaiga en gran medida sobre el Detector.

3. Por último, el método de evaluación más preciso de todos, independiente de la aplicación, es el que calcula los resultados a nivel de trama. Se presentó en [11] como modelo válido y preciso para la evaluación de cualquier *VAD* independientemente del tipo de aplicación. En este caso también hay que realizar un etiquetado manual y posteriormente un procesado de marcas que sea capaz de catalogar trama a trama de qué tipo es: si es de voz con un "1" y si es de ruido o silencio "0". Este procesado de marcas se repite con las etiquetas que obtenga el Detector. Los criterios para calcular los errores fueron los siguientes:
   - Pc_on ≡ Porcentaje de tramas de voz clasificadas como de ruido al principio de la pronunciación.
   - Pc_off ≡ Lo mismo que las anteriores pero al final de la pronunciación.
   - TrueVAF ≡ Número de tramas de voz frente al total.
   - VadVAF ≡ Número de tramas de voz dadas por el *VAD* frente al total.
   - Error Type I ≡ Porcentaje de tramas que son realmente voz y las clasifica como ruido, así que:

$$Type \quad I \quad Error \quad = Pc\_on + Pc\_off \qquad (3.1)$$

   - Error Type II ≡ Porcentaje de tramas que son realmente ruido y las clasifica como voz.

- Comparación de los resultados obtenidos con el *VAD* a evaluar con los obtenidos con otros *VAD*s de referencia manteniendo las mismas métricas y bases de datos test. En el caso de este trabajo los detectores de referencia son los siguientes: G.729 anexo B [3], AMR [106] y AURORA (FD) [77]. Por ejemplo, algunos investigadores [67,42,68] comparan los resultados con los del *VAD* del codificador G.729. En [93] se comparan con los del Detector del G.729 y con el Detector del codificador AMR.

Como ejemplo, en [14] se realizan pruebas experimentales con bases de datos españolas de ruidos de coche y con la base de datos de AURORA-2, con una frecuencia de muestreo de 8 KHz y tramas de 10 milisegundos. La continuación de este trabajo es [1] donde los mismos autores de [14] realizan un análisis comparativo más completo. Se usan como técnicas para su evaluación:

- Análisis de aciertos a nivel de trama.
- Tasa de acierto del reconocedor de voz.

Se llaman valores de referencia al número de tramas que se sabe de qué tipo son, ruido o voz, procedentes de un etiquetado manual. En cuanto a las pruebas de reconocimiento, se puede decir que se realizaron para diferentes SNR, en distintos idiomas y para distintas premisas de entrenamiento. En este caso, el *VAD* se usa en el sistema de reconocimiento tanto para la estimación de ruido, en combinación con el filtrado de Wiener, como para el eliminado de tramas (frame dropping). Las premisas de entrenamiento fueron dos:

- Entrenamiento con voz limpia. Se caracteriza por tener una elevada SNR.
- Entrenamiento multicondición. Se caracteriza por la presencia de distintos ruidos.

En [35] la base de datos de entrenamiento está formada por 286 locutores con un total de 15.3 horas de grabación. La base de datos de test consiste en diálogos con duración de 21 minutos con la mitad de locutores hombres y la mitad mujeres. De todas las grabaciones, el 50% fueron realizadas con teléfonos móviles. Además se han tenido en cuenta los siguientes tipos de ruidos: clicks, respiraciones, ruido de coche, de TV y radio, niño llorando, gritando y titubeando. Además se tuvieron SNRs que variaban desde 5 hasta 40 dB con una SNR media de 20 dB. Se considera la tasa de falsas alarmas P_f, tasa de no detección en presencia de voz P_m y el error total P_e definidos como $P_f = N_{n\to s} / N$, $P_m = N_{s\to n} / N$, y $P_e = P_f + P_m$ donde $N_{n\to s}$ y $N_{s\to n}$ son el número de tramas de ruido y de voz detectadas como falsas y N el número total de tramas test. El ruido se añade digitalmente a voces limpias terminando con SNR de 20, 10 y 0 dB.

Para verificar el funcionamiento del *VAD* en [43] se utilizan dos tipos de dato. En el primero se usa una grabación de la pronunciación "Hello" que dura

aproximadamente 1.5 segundos para demostrar las mejoras del algoritmo con voz limpia y ruidosa. Las tramas iniciales que son de silencio incluyen impulsos artificiales y ruido de un despertador. Estos artefactos son causa del mal funcionamiento de otros algoritmos. Además, la grabación también posee ruido de respiración que es perfectamente excluido del modelo de voz. En el segundo ejemplo, el caso ruidoso se muestra para la misma pronunciación, "Hello", pero con amplias zonas de ruido blanco a lo largo de toda la grabación con una SNR de 0 dB. La decisión de entrada de nuevo obtiene una relativa precisión en los extremos. En casos de la utilización de energía y tasa de cruces por cero seguro que fallaría. En este caso las pruebas son demasiado específicas y por ello, convendría el estudio de esta técnica con una base de datos más amplia.

3.3.- Bases de datos utilizadas.

Formalmente, una base de datos es una colección de información organizada de forma que un programa de ordenador pueda seleccionar rápidamente los fragmentos de datos que necesite: es un sistema de archivos electrónico. En nuestro caso concreto, las bases de datos son un conjunto de ficheros de audio codificados en ley-mu, en ley-a o en lineal, junto con sus respectivas marcas y transcripciones.

A continuación, se efectúa una descripción completa de cada una de las bases de datos usadas. Estas bases de datos las agrupamos en tres tipos diferentes en función del uso que se ha realizado de las mismas: bases de datos de entrenamiento, bases de datos de desarrollo y bases de datos test.

La primera base de datos que se describe es la base de datos Av16.3 [88], la más importante por su uso para el estudio de características para el rechazo de voces de fondo. Esta base de datos se utiliza en todos los ámbitos: entrenamiento, desarrollo y test. Los ficheros de audio de Av16.3 se han dividido aleatoriamente por locutores en ficheros de entrenamiento (TRAIN_AV), de desarrollo (DEV_AV) y de test (TEST_GSM_PREAV y TEST_GSM_POSTAV). A continuación, se va a realizar una descripción detallada de la mencionada base de datos Av16.3.

3.3.1.- Base de datos Av16.3.

La base de datos Av16.3 está formada por señales, tanto de audio como de video, grabadas en una sala como la que se puede ver en la Fig.3.1 y en la Fig.3.2.

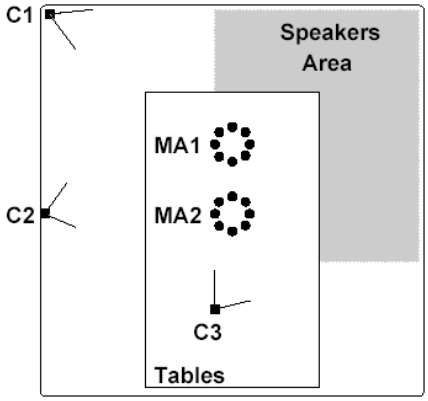

Figura 3.1. MA1 y MA2 son arrays circulares de 8 micrófonos. Observar la "Speakers Area". Esta figura se ha obtenido de [88].

Figura 3.2. Sala de grabación obtenida de [88].

Los ficheros de audio se han grabado con 16 micrófonos perfectamente sincronizados y convenientemente calibrados. Para cada grabación, hay 16 ficheros WAV grabados con dos arrays circulares de 8 micrófonos cada uno (Fig.3.1), muestreados a 16 Khz, y otros ficheros WAV grabados desde micrófonos de la solapa del locutor, también muestreados a 16 Khz. Es importante mencionar que en

algunas de estas grabaciones existe voz de habla simultánea, esto es, distintos locutores hablan a la vez. Los ficheros de voz se nombran en función de las características de los locutores:

- Seq01-1p-0000: 1 locutor parado mientras habla. El locutor no habla constantemente (hace pausas de vez en cuando) y cuando lo hace mira hacia los arrays de micrófonos.
- Seq11-1p-0100: 1 locutor moviéndose la mayor parte del tiempo mientras habla. El locutor habla todo el tiempo y lo hace mirando hacia los arrays de micrófonos.
- Seq15-1p-0100: 1 locutor moviéndose sin rumbo fijo y alternando entre habla y silencios de larga duración.
- Seq18-2p-0101: 2 locutores moviéndose y hablando mientras miran hacia los arrays de micrófonos siempre. Ambos locutores tratan de mantener una conversación lenta.
- Seq24-2p-0111: 2 locutores moviéndose y hablando todo el tiempo mientras cruzan la mirada 2 veces y la ocultan otras 2.
- Seq37-3p-0111: 3 locutores estáticos (no se mueven): 2 de ellos permanecen sentados y otro de pié. La mayoría del tiempo hablan 2 o 3 locutores.
- Seq40-3p-0111: 3 locutores, 2 sentados y otro de pié, hablando todos ellos continuamente y mirando los arrays de micrófonos. El locutor que está de pié anda entre los otros dos locutores que permanecen sentados, es decir, que se pone delante de ellos.
- Seq45-3p-1111: 3 locutores moviéndose, entrando y saliendo del escenario, continuamente hablando y ocultándose la mirada varias veces. Es el caso más complicado de solapamiento de voz.

Ver referencia [88] para más información.

Para concluir esta descripción, es importante decir que todos los ficheros de audio de la base de datos se han submuestreado a 8 Khz (para simular el canal telefónico).

3.3.2.- Bases de datos de entrenamiento.

La nomenclatura a usar en las bases de datos de entrenamiento será la siguiente: TRAIN_”X”, donde “X” simboliza alguna característica descriptiva de la

base de datos en cuestión. Con esto, las bases de datos a tener en cuenta serían las siguientes:

- TRAIN_GSM_MFCC: está formada por 1102 ficheros de voz limpia contaminados con ruido de diversos tipos: estacionarios y no estacionarios. Los ficheros de voz limpia fueron grabados en un laboratorio que contaba con una central GSM y terminales móviles a través de los que 10 locutores, 5 hombres y 5 mujeres, emitían locuciones de frases cortas previamente etiquetadas. En lo que respecta a los ruidos, se trata de 1102 segmentos de medio segundo procedentes de conversaciones de servicios reales (clicks, golpes, coches circulando, bullicio, etc.) procurando que los mismos se encontraran de forma aislada y no solapados con voz. Para finalizar, estos ruidos fueron sumados, sin escalar o modificar el nivel de la señal, a los 1102 ficheros de voz limpia de forma aleatoria. Esta base de datos se ha utilizado para calcular el poder de discriminación de los coeficientes cepstrales (apartado 4.2.1) para diferenciar entre voz y no voz.
- TRAIN_GSM_HMM: está formada por 101.350 ficheros grabados en distintos terminales móviles con tecnología GSM por Telefónica y procedentes de conversaciones reales entre una gran multitud de locutores: 150 hombres y 148 mujeres. Hay que remarcar que la mencionada base datos contenía algunos ruidos no estacionarios como golpes, clicks, de tal manera que, el modelo de ruido, al entrenar, tuviera en cuenta este efecto. El etiquetado de la enorme base de datos fue llevado a cabo gracias al trabajo de 6 personas que contrató Telefónica durante varios meses. Esta base de datos se usa para entrenar los modelos acústicos de telefonía móvil del Detector Base Mejorado (apartados 4.3 y 4.4).
- TRAIN_AV: está formada por un subconjunto de 14 ficheros de larga duración aleatoriamente seleccionados de la base de datos Av16.3. Esta base de datos se ha utilizado para entrenar los umbrales de decisión de las características estudiadas para el rechazo de las voces de fondo (apartado 6.3).
- TRAIN_FIJA: está formada por alrededor de 130.000 ficheros grabados a partir de distintos terminales de telefonía fija por Telefónica y formados por frases cortas emitidas por una gran diversidad de locutores con distintas edades: 162

hombres y 130 mujeres, ambos con edades comprendidas entre 20 y 46 años. El etiquetado de la base de datos fue llevado a cabo gracias a un grupo de 6 personas que contrató Telefónica durante varios meses. Esta base de datos se usa para entrenar los modelos acústicos de telefonía fija del Detector Base Mejorado (apartado 4.6).

- TRAIN_IP: está formada por 23.400 ficheros etiquetados y grabados a partir de diferentes teléfonos IP con una posterior codificación/decodificación mediante el códec G.723. Los locutores que generaban las locuciones fueron 7 hombres y 8 mujeres de mediana edad. Esta base de datos se usa para entrenar los modelos acústicos de voz IP del Detector Base Mejorado (apartado 4.6).

3.3.3.- Bases de datos de desarrollo.

En cuanto a las bases de datos de desarrollo, la más importante por su uso para el estudio de características para el rechazo de voces de fondo es el subconjunto de desarrollo de Av16.3, que también se llamará DEV_AV (del inglés “development”). La nomenclatura a usar en las bases de datos de desarrollo, de forma análoga a las bases de datos de entrenamiento, será la siguiente: DEV_”X”, donde “X” simboliza alguna característica descriptiva de la base de datos en cuestión.

Otras bases de datos de desarrollo usadas en este trabajo son las siguientes:

- DEV_GSM_COCHE: está formada por 1860 ficheros etiquetados y grabados en distintos terminales móviles con tecnología GSM en los que los locutores hablan mientras conducen a distintas velocidades. En este caso los locutores fueron 10 hombres y 9 mujeres. Es por tanto una base de datos en la que la voz de los locutores se encuentra embebida en ruido estacionario de coche. Esta base de datos se usa para estudiar el comportamiento del Detector Base Mejorado sobre ruido estacionario.
- DEV_GSM_RUIDONE: está formada por 778 ficheros etiquetados que contienen conversaciones emitidas por teléfonos móviles con tecnología GSM donde los locutores principales, 7 hombres y 8 mujeres, se encuentran en entornos adversos con ruidos no estacionarios: bares, salas con ruido de fondo de

televisión (tertulias) o simplemente en la calle donde existe ruido de fondo procedente de otros locutores que están a una cierta distancia del locutor principal. Esta base de datos se usa para estudiar el comportamiento del Detector Base Mejorado sobre ruido no estacionario.

- DEV_GSM_LIMPIA: está formada por 2380 ficheros, que contienen frases cortas emitidas por locutores de distintas edades y diferente sexo, etiquetados y grabados en distintos terminales móviles con tecnología GSM. En este caso no existen ruidos de ningún tipo y la media de la relación señal a ruido es mayor de 25 dB (SNR ≥ 25dB). Se puede considerar, por tanto, una base de datos de voz limpia. Esta base de datos se usa para estudiar el comportamiento del Detector Base Mejorado con distintas topologías en sus modelos acústicos.

3.3.4.- Bases de datos test.

La nomenclatura a usar en las bases de datos test, de forma análoga a las bases de datos de entrenamiento, será la siguiente: TEST_"X", donde "X" simboliza alguna característica descriptiva de la base de datos en cuestión. Con esto, las bases de datos a tener en cuenta serían las siguientes:

- TEST_GSM_LIMPIA: está formada por 2500 ficheros, que contienen frases cortas emitidas por locutores de distintas edades y diferente sexo, etiquetados y grabados en distintos terminales móviles con tecnología GSM. En este caso existen ruidos reales pero de niveles bajos: la media de la relación señal a ruido es mayor de 25 dB (SNR ≥ 25dB). Se puede considerar, por tanto, una base de datos de voz real bastante limpia. Esta base de datos se usa en todos los casos para evaluar el comportamiento tanto del Detector Base Mejorado como del Detector Final propuesto en esta Tesis que incluye el rechazo de las voces de fondo.
- TEST_GSM_COCHE: está formada por 2350 ficheros etiquetados y grabados en distintos terminales móviles con tecnología GSM cuando los locutores hablan mientras conducen a distintas velocidades. En este caso los locutores fueron 14 hombres y 11 mujeres. Es por tanto una base de datos en la que la voz de los locutores se encuentra embebida en ruido estacionario de coche, parecida a DEV_GSM_COCHE, pero más amplia y con otro tipo de locuciones. Esta base

de datos se usa para evaluar el comportamiento del Detector Base Mejorado sobre ruido estacionario.

- TEST_GSM_RUIDONE: está formada por 2630 ficheros etiquetados que contienen conversaciones procedentes de servicios reales y emitidas por teléfonos móviles con tecnología GSM donde los locutores principales, 15 hombres y 13 mujeres, se encuentran en entornos adversos con ruidos no estacionarios: bares, salas con ruido de fondo de televisión (tertulias) o simplemente en la calle donde existe ruido de fondo procedente de otros locutores que están a una cierta distancia del locutor principal. Se trata de una base de datos de voz con ruidos no estacionarios. Esta base de datos se usa para evaluar el comportamiento tanto del Detector Base Mejorado como del Detector Final que incluye el rechazo de las voces de fondo sobre ruido no estacionario que, entre otros, contiene ruidos de voces de fondo.
- TEST_GSM_PREAV: se trata de la base de datos TEST_GSM_LIMPIA (voz limpia) contaminada con segmentos de las voces de fondo de la base de datos Av16.3 de test seleccionados de forma aleatoria. Estas voces de fondo fueron sumadas antes de las pronunciaciones de los locutores principales de TEST_GSM_LIMPIA y a diferentes SNRs: 5dB, 10dB, 15dB, 20dB y 25dB. Al tener TEST_GSM_LIMPIA etiquetada, por ende, también se tiene etiquetada TEST_GSM_PREAV. Esta base se usa para evaluar el comportamiento del Detector Final que incluye el rechazo de las voces de fondo sobre voz limpia contaminada con voces de fondo: umbrales de decisión, árbol de decisión y red neuronal.
- TEST_GSM_POSTAV: se trata de la base de datos TEST_GSM_LIMPIA (voz limpia) contaminada con segmentos de las voces de fondo de la base de datos Av16.3 de test seleccionados de forma aleatoria, y distintos de los usados en TEST_GSM_PREAV. Estas voces de fondo fueron sumadas después de las pronunciaciones de los locutores principales de TEST_GSM_LIMPIA y a diferentes SNRs: 5dB, 10dB, 15dB, 20dB y 25dB. Al tener TEST_GSM_LIMPIA etiquetada, por ende, también se tiene etiquetada TEST_GSM_POSTAV. Esta base de datos se usa para evaluar el comportamiento del Detector Final que

incluye el rechazo de las voces de fondo sobre voz limpia contaminada con voces de fondo: umbrales de decisión, árbol de decisión y red neuronal.

- TEST_FIJA: está formada por 2930 ficheros etiquetados convenientemente y que contienen conversaciones procedentes de servicios reales y emitidas desde diferentes terminales de telefonía fija donde los locutores, 16 hombres y 15 mujeres, se encuentran en sus casas o recintos de trabajo. Las mediciones de la SNR media en esta base de datos es 23.8 dB. Se trata, por tanto, de una base de datos bastante limpia. Esta base de datos se usa para evaluar los modelos acústicos de telefonía fija del Detector Base Mejorado (apartado 4.6).
- TEST_IP: está formada por 2650 ficheros etiquetados convenientemente y que contienen conversaciones emitidas desde diferentes PC's con micrófonos para voz IP donde los locutores, 10 hombres y 5 mujeres, de mediana edad, se encuentran en sus recintos de trabajo y teclean mientras hablan. Esta base de datos se usa para evaluar los modelos acústicos de voz IP del Detector Base Mejorado (apartado 4.6).

3.4.- Medidas de evaluación.

En esta sección se definen formalmente cuáles son las medidas de evaluación y de qué forma se calculan los errores que se presentarán más adelante.

El error de detección será a nivel de trama, por ser el que mayor resolución presenta. En nuestro caso, este error tiene en cuenta tanto la tasa de falsas alarmas (en inglés False Alarm Rate), un error de detección en tramas de voz, como la tasa de falsos rechazos (en inglés Miss Rate), un error de detección en las tramas de ruido. El Error de Detección Global, GDE, tiene en cuenta la contribución de las falsas aceptaciones y los falsos rechazos. Matemáticamente se exponen en ec. 3.2-3.4 las tres expresiones que representan los mencionados errores.

$$false\_alarm\_rate(\%) = \left(\frac{N_{Tramas\_ruido \quad como \quad Tramas\_voz}}{N_{Tramas\_ruido}} \right) \times 100 \tag{3.2}$$

$$mis\_rate(\%) = \left(\frac{N_{Tramas\_voz \quad como \quad Tramas\_ruido}}{N_{Tramas\_voz}} \right) \times 100 \tag{3.3}$$

$$GDE = \Pr ob(Error\_Ruido) + \Pr ob(Error\_Voz) \quad (3.4)$$

En estas expresiones, $N_{Tramas\_ruido \;\; como \;\; Tramas\_voz}$ denota el número de tramas que son realmente ruido y se han clasificado erróneamente como voz, $N_{Tramas\_voz \;\; como \;\; Tramas\_ruido}$ simboliza el número de tramas que son realmente voz y se han clasificado erróneamente como ruido, N_{Tramas_ruido} denota el número total de tramas de ruido, y, finalmente, N_{Tramas_voz} simboliza el número total de tramas de voz. Es importe comentar que estas expresiones corresponden a las expuestas en [11] (ver apartado 3.2). El *Error Type 2* sería la tasa de falsas alarmas (*false_alarm_rate*) y el *Error Type 1* sería la tasa de falsos rechazos (*miss_rate*).

Otras expresiones de error que se encontrarán en la literatura son las siguientes: error de clasificación entre dos clases y tasa de aciertos a nivel de pulso. El error de clasificación representa un error en la decisión a nivel de trama, mientras que la tasa de aciertos a nivel de pulso (ec. 3.5) representa un error a nivel de pulso. En este último caso se tiene en cuenta la estructura de lenguaje y otras características que serán capaces de diferenciar los pulsos de voz que proceden de locutores principales y de voces de fondo, eliminando estos últimos. La tasa de aciertos a nivel de pulso se obtendrá al aplicar como método de decisión la red neuronal o el árbol de decisión.

$$Tasa\_de\_aciertos = \frac{n^o\_pulsos\_acertados}{n^o\_pulsos\_totales} \quad (3.5)$$

3.5.- VADs de referencia a comparar.

Para contrastar los resultados obtenidos con el *VAD* creado por nosotros, se realizan estudios comparativos con otros *VAD*s de referencia estándar. Estos *VAD*s de referencia son los siguientes:

- *VAD* del codificador G729 anexo B [3].
- *VAD* frame dropping del estándar AURORA [77].
- *VAD* del codificador AMR opción 1 [106].

- *VAD* del codificador AMR opción 2 [106].

En los experimentos que se irán planteando en los sucesivos capítulos de esta memoria de Tesis, se presentan resultados del Error de Detección Global (GDE) para todos ellos usando las bases de datos “test”, tanto de voz limpia como con ruido estacionario y no estacionario.

CAPÍTULO 4

VAD BASADO EN HMMs

4.1.- Introducción.

En este capítulo se describe paso a paso cada uno de los módulos que forman el *VAD* basado en Modelos Ocultos de Markov (HMM) y se demuestra y justifica el uso de las técnicas utilizadas en cada uno de los módulos. Los aspectos más importantes a tratar son los siguientes:

- Se analizan las distintas características: características de naturaleza espectral y el logaritmo de la energía normalizada que formarán el vector de características de los dos Modelos Ocultos de Markov, uno para representar las tramas de voz y otro para representar las tramas de ruido o no voz.
- Se comparan diferentes topologías de los mencionados HMMs y se obtienen resultados experimentales usando éstas.
- Se estudia la detección de pulsos de voz y se aplican reglas que tienen en cuenta la estructura del habla. Por ejemplo, no pueden existir pronunciaciones o pulsos de voz extremadamente pequeños.
- Se obtienen resultados experimentales para el *VAD* basado en HMMs propuesto usando distintas bases de datos y se comparan con los *VAD*s de referencia.

4.2.- Descripción del VAD basado en HMMs.

Se trata de un *VAD* híbrido basado en Modelos de Markov y capaz de funcionar en tiempo real. Es híbrido porque usa algunos MFCCs del extractor de características del reconocedor. Aún así, en este trabajo se va a considerar el extractor de características como parte del sistema global de detección, en principio independiente de la fase de reconocimiento. Por otro lado, el reconocedor también podría valerse de las marcas de voz/no-voz obtenidas por el *VAD*. Por tanto, el *VAD* propuesto está diseñado para estar integrado dentro de un reconocedor automático de habla, aunque podría ser usado de forma independiente para cualquier aplicación que lo necesitase.

El *VAD* de partida usa características tanto estáticas (logaritmo de la energía normalizada) como dinámicas (delta del logaritmo de la energía normalizada), y es entrenable de tal manera que sus modelos se pueden adaptar a los diferentes canales de telefonía, por ejemplo de telefonía fija, GSM o incluso voz IP. Es

importante comentar que el punto de partida de este trabajo es el *VAD* presentado por Acero [82]. El *VAD* propuesto en esta Tesis difiere del de Acero [82] en que el vector de características que usan los HMMs incluye información espectral (se añaden 3 cepstrum) y en que el algoritmo de reestimación de energía de ruido es diferente. También se tiene la posibilidad, a la hora de realizar la decisión, de poder usar el logaritmo de energía normalizada. Además, como veremos en el capítulo 5, se añadirán características para rechazar las voces de fondo.

El diagrama de bloques del *VAD* se muestra en la Fig.4.1. Como se puede ver en el dibujo, el sistema está formado por 3 módulos perfectamente diferenciados que se explican a continuación. Por otro lado insistir en que el algoritmo basado en HMMs obtiene resultados a nivel de trama mientras que la máquina de estados y detección de pulsos incluye información de la estructura del habla y genera resultados a nivel de pulso.

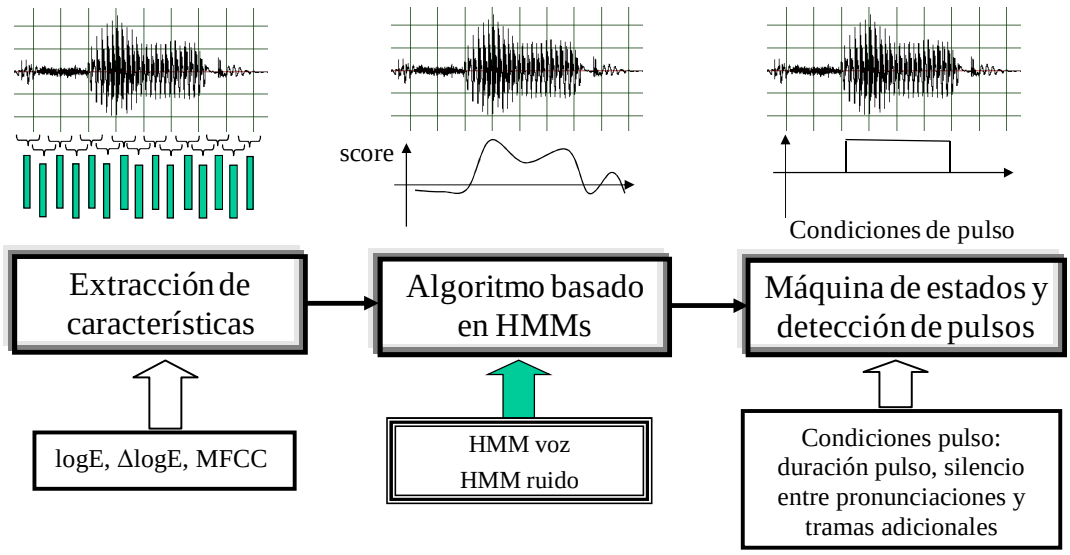

Figura 4.1. Esquema completo del nuevo *VAD* basado en HMMs.

Una vez seleccionada la topología a usar, se realiza el cálculo del LLR teniendo en cuenta las verosimilitudes de cada clase. Se calcula una puntuación a la que llamaremos *score*, procedente directamente del logaritmo del cociente de las verosimilitudes, generado a partir de los modelos de voz y de ruido con los vectores de características calculados trama a trama (ec. 4.1),

$$score = \log\left(L(\varphi_{speech})\right) - \log\left(L(\varphi_{noise})\right) \qquad (4.1)$$

donde $L(\varphi) = \Pr ob(\varphi \mid v(n))$ denota la verosimilitud de un modelo acústico. Otro aspecto importante es que los modelos están realimentados. Es importante señalar que el *VAD* usa tanto información probabilística como información específica del ruido de fondo (logaritmo de la energía normalizada). Si además se quiere desglosar este algoritmo en función de los estados de los modelos, se puede también definir el logaritmo de la verosimilitud de una trama de voz como la media acumulada del logaritmo de la verosimilitud para los N estados del modelo de voz:

$$\ln P(O_t / speech) = \frac{1}{N}\sum_{i=0}^{N-1}\ln(\alpha_t^S(i)) \qquad (4.2)$$

donde $\alpha_t(i)$ representa la probabilidad acumulada en el tiempo t y para el estado i. De igual forma, para el modelo de ruido y M estados:

$$\ln P(O_t / noise) = \frac{1}{M}\sum_{i=0}^{M-1}\ln(\alpha_t^n(i)) \qquad (4.3)$$

La diferencia de las dos expresiones anteriores, tal y como recoge (ec. 4.4), obtiene de nuevo el *score* o puntuación normalizada, considerando las puntuaciones de los estados de los modelos:

$$score[t] = \frac{1}{N}\sum_{i=0}^{N-1}\ln(\alpha_t^S(i)) - \frac{1}{M}\sum_{i=0}^{M-1}\ln(\alpha_t^n(i)) \qquad (4.4)$$

Como ya se ha comentado anteriormente, el *VAD* propuesto no usa ningún tipo de umbral ajustable para clasificar las tramas de voz y de ruido. Por tanto, se trata de un *VAD* de sencilla configuración y rápido. La clasificación, en este caso a nivel de trama, se basa en la diferencia del logaritmo de verosimilitudes de voz y de ruido:

score. Si el *score* es mayor que cero, la trama se clasifica como de voz, mientras que en caso contrario se clasifica como de ruido.

Enlazando con la sencilla clasificación anterior, el *VAD* propuesto, a partir del valor de la relación de verosimilitud entre las dos clases, el *score*, se podría tomar directamente la decisión a nivel de trama: por encima del valor elegido se clasificaría una trama como trama de voz y en caso contrario como de ruido. En este caso no se utilizaría la máquina de estados y la detección de pulsos. Por el contrario, avanzando un paso más, si se añade información adicional de la estructura del habla, se crearían los llamados pulsos de voz (agrupaciones de tramas consecutivas) mejorando los resultados finales. Esta información está basada en anteriores mensajes de voz:

- Duración de pulsos: la duración de un pulso no debe ser muy pequeña ni excesivamente grande. En nuestro caso, si la duración de pulso es menor que 168 ms. (múltiplo del valor de desplazamiento de trama, 12 ms.), el pulso no se considera pulso de voz. Con esta condición el *VAD* evita detectar clicks, toses o risas como voz.
- Silencio entre pronunciaciones: el silencio o el ruido intersilábico entre pronunciaciones no debe ser ni muy pequeño ni excesivamente grande. En el *VAD* propuesto, si el silencio entre pulsos de voz consecutivos es menor que un parámetro, configurable y en milisegundos, los pulsos se conectan para formar uno sólo.
- Tramas adicionales: son también llamadas tramas de salvaguarda (aplicaciones de reconocimiento de voz y codificación). En nuestro caso, el algoritmo añade tres tramas (36 ms.) tanto al inicio como al final de pronunciación para evitar perder tramas de voz de baja energía (oclusivas o fricativas).

4.3.- Extracción de características.

En general, los HMMs suelen usar los parámetros que también utiliza el reconocedor, por ejemplo los MFCCs. Se trata por tanto de estudiar:

1. Cuáles son los MFCCs que mejor discriminan entre las clases “voz” y “no voz”.

2. Qué papel juega la energía y de qué manera podría integrarse dentro del vector de características final.

En el caso de nuestro detector, se trata de obtener un vector *v(n)* formado por las características que mejor discriminen entre las clases “voz” y “no voz”. Para ello, en primer lugar estudiaremos los coeficientes cepstrales o cepstrum y, más tarde, la normalización de la energía. El uso de la energía normalizada, y no de la energía, hace que no se tengan que usar umbrales para distintas SNRs (invariante ante cambios de la SNR), cuestión que no se aborda en otros sistemas parecidos como el de Lamel [72]. Otros autores como Sheikhzadeh en [89] usan distintos modelos de voz y de ruido para distintas SNRs: el problema es que el uso de tantos modelos y parámetros a tener en cuenta hace que sea un *VAD* difícil de implementar y con un retardo considerable en cuanto a tiempo de respuesta se refiere.

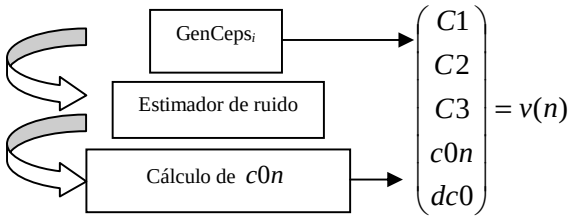

Figura 4.2. Vector de características (Feature Vector Extraction).

En la Fig.4.2 se presenta el módulo generador de cepstrum (GenCeps). Este generador de cepstrum realiza cálculos matemáticos sobre la señal de entrada en el dominio de la frecuencia, es decir, tras aplicar la transformada rápida de Fourier (FFT) a la señal acústica de entrada (dominio temporal). Los cálculos matemáticos se basan en aplicar a la señal de entrada en dominio frecuencial un filtrado en escala Mel en un banco de 12 filtros Mel con preénfasis (elimina el término de continua y enfatiza las altas frecuencias para no perder, por ejemplo, la detección de fricativas). Estos filtros son triangulares y se encuentran solapados a un 50%. La señal de entrada se analiza y se filtra en tramas o segmentos de 24 milisegundos con un desplazamiento de 12 milisegundos, es decir que, por ejemplo, en 48 milisegundos de señal acústica nos encontraríamos con cuatro tramas de la señal

acústica. En total, el generador de cepstrum obtiene ocho coeficientes (*C1-C8*) y la energía (*C0*). De las nueve características, sólo tres serán las usadas más tarde por los dos modelos. Además, nótese que estos tres parámetros usados por el *VAD* son un subconjunto de los que usará el reconocedor de voz, por lo tanto, el hecho de que el *VAD* use cepstrum en sus HMMs no va a suponer pérdida de tiempo de proceso en el sistema de reconocimiento de voz completo.

Para encontrar los cepstrum más significativos, esto es, los coeficientes que discriminan mejor entre las dos clases acústicas (voz y no-voz), se analizaron las funciones de densidad de probabilidad de los primeros nueve parámetros (*C0-C8*) de las dos mencionadas clases acústicas. Este análisis se lleva a cabo mediante una base de datos de entrenamiento, asumiendo que los cepstrum son estadísticamente independientes. Esta base de datos la denominamos TRAIN_GSM_MFCC y está formada por voz limpia contaminada con ruido tanto estacionario como no estacionario. Los cepstrum fueron calculados para todas las tramas, tanto para las de voz como las de ruido.

El poder de discriminación de cada cepstrum se midió como la inversa de la incertidumbre [90]. La incertidumbre (ec. 4.5) mide la probabilidad de fallo al clasificar una trama por un parámetro, en nuestro caso cada cepstrum.

$$\text{incertidumbre}_i = \int_{-\infty}^{umbral\_optimo} p_{voz}(x_i)dx_i + \int_{umbral\_optimo}^{\infty} p_{no-voz}(x_i)dx_i \qquad (4.5)$$

En (4.5), x_i representa el *i'ésimo* cepstrum. p_{voz} y p_{no-voz} denotan las distribuciones de probabilidad del cepstrum x_i para la voz y el ruido (no-voz) respectivamente. Independientemente para cada coeficiente, las distribuciones de probabilidad se estimaron para cada clase acústica (voz y no-voz). Las distribuciones se calcularon sin normalizar los histogramas de cada cepstrum. El error de clasificación (incertidumbre) calculado usando (ec. 4.5) se basa en un umbral óptimo (si x_i es mayor que *umbral_optimo* es voz y si no, ruido) considerando las distribuciones de probabilidad como funciones continuas. Este *umbral_optimo* es el punto de intersección entre las dos distribuciones de

probabilidad, procedentes de cada una de las clases. Es importante comentar que en este caso específico se usan las distribuciones discretas de probabilidad (aproximación por histogramas) y que el *umbral_optimo* es el valor discreto más cercano al punto de intersección entre las funciones de densidad de probabilidad ideales.

La Tabla 4.1 contiene las incertidumbres de cada cepstrum ordenadas de menor a mayor. Los cepstrum con los que se han entrenado los modelos de voz y de ruido para el *VAD* propuesto están marcados con negrita. Los resultados de incertidumbre muestran que los cepstrum que mejor discriminan entre las dos clases (menor incertidumbre) son, en orden, *C3*, *C0*, *C1* y *C2*. Como *C0* se usa para calcular el logaritmo de la energía normalizada (*c0n*), *C3*, *C1* y *C2* son los tres cepstrum seleccionados para ser incorporados dentro de vector de características final de nuestro *VAD*. Nótese que se asume independencia entre los coeficientes los MFCC como es habitual en ASR.

Coeficiente Cepstral	Incertidumbre
3	**0,3428**
0	**0,3606**
1	**0,3623**
2	**0,3686**
4	0,3765
5	0,3898
7	0,4137
6	0,4371
8	0,4495

Tabla 4.1. Incertidumbre de las distribuciones de probabilidad para cada cepstrum.

Es importante comentar que los resultados obtenidos coinciden con los obtenidos por Skorik [107].

En cuanto a la normalización de la energía, es necesario el uso de un algoritmo de estimación de ruido. La normalización de la energía implica eliminar la energía de ruido de la energía total de la señal (que normalmente está contaminada

con ruido). Por tanto, al normalizar se obtiene la energía únicamente que procede de la voz. En este caso el estimador de ruido se basa en una versión simplificada del estimador de ruido del códec AMR1. Su formulación es la siguiente:

$$bckr\ \_est[i+1] = (1.0-\alpha)\cdot bckr\ \_est[i] + \alpha \cdot energy\ [i-1] \tag{4.6}$$

donde i simboliza la trama actual y α toma valores según el siguiente criterio:

$$\left.\begin{array}{l} if \quad bckr\ \_noise[i] < energy\ [i-1], \quad \alpha = 1.0-\lambda \\ else \quad \alpha = \lambda \end{array}\right\} \tag{4.7}$$

En nuestro caso λ toma valor 0.85 (manualmente ajustado mediante algunos ejemplos críticos). Esto implica una adaptación del 85% a caídas de energía provocadas por el silencio o ruido de fondo. Con esto ya se puede obtener el logaritmo de la energía normalizada: se calcula trama a trama como la diferencia entre el logaritmo de la energía o *C0* (procedente de los MFCCs calculados en el módulo GenCeps) y el logaritmo de la energía de ruido de fondo estimada previamente.

Para finalizar, la variación de la energía normalizada se calcula simplemente como la diferencia entre el logaritmo de la energía normalizada de la trama actual y el de la trama anterior, completando así el vector de características.

Se muestra un ejemplo práctico (Fig.4.3.a, Fig.4,3.b y Fig.4.4) de cómo calcula el *VAD* con el método anterior el logaritmo de la energía normalizada del mismo fichero de audio limpio (25dB, Fig.4.3.a), y mezclado con ruido de fondo estacionario (5dB, Fig.4.3.b).

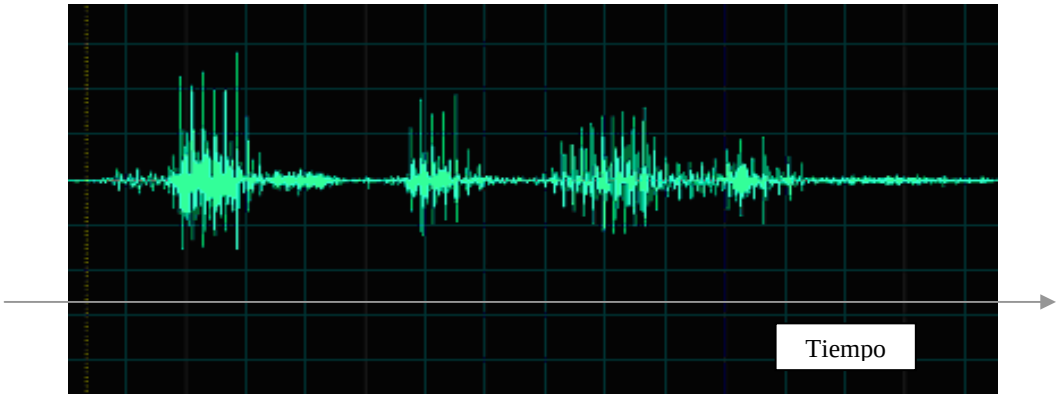

Figura 4.3.a. Fichero de voz limpia analizado (SNR=25dB).

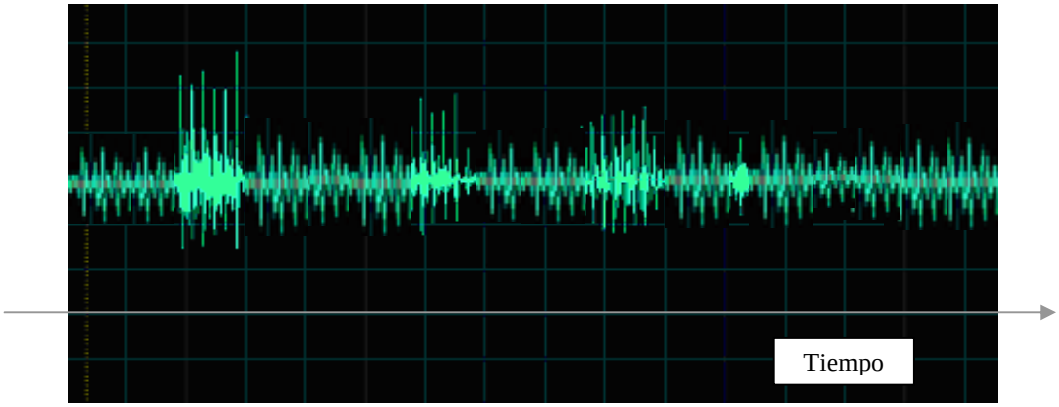

Figura 4.3.b. Fichero de voz limpia con ruido de fondo estacionario (SNR=5dB).

La pronunciación de la Fig.4.3.b es la misma que la pronunciación de la Fig.4.3.a pero con niveles de ruido estacionario bastante más elevados. Un *VAD* de energía sería sensible a estas variaciones de la señal de ruido, y, por tanto, poco robusto: una posible solución podría ser tener en cuenta distintos umbrales para diferentes SNRs. La energía normalizada trata de solucionar este problema al ser invariante ante las variaciones de la relación señal a ruido y como consecuencia no tener que considerar distintos umbrales.

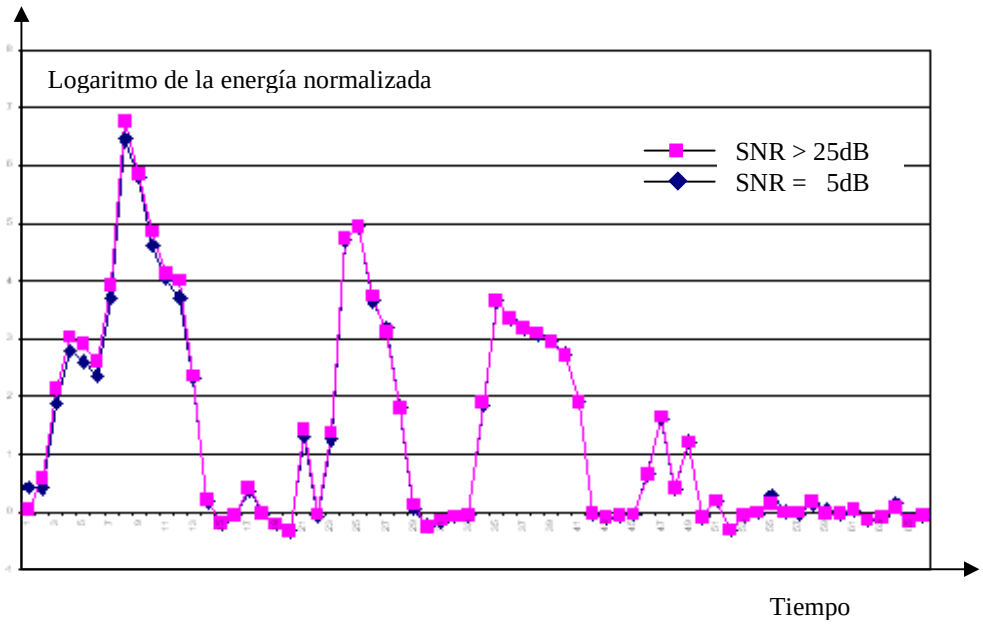

Figura 4.4. Logaritmo de la energía normalizada para diferentes SNRs.

Como se puede observar, la idea de normalización de la energía hace que las distribuciones de energías sean muy parecidas para relaciones señal a ruido (SNR) diferentes.

A continuación, se realiza el estudio comparativo entre usar únicamente como vector de características el logaritmo de la energía normalizada (*c0n*) y la variación del logaritmo de la energía, caso similar al del *VAD* presentado por Acero [82] con sólo dos componentes, y ampliar este vector anterior con 3 cepstrum, caso del *VAD* propuesto en este capítulo de Tesis. En primer lugar se realizará un estudio comparativo con DEV_GSM_COCHE y más tarde el estudio se efectuará a través de la base de datos que contiene ruidos no estacionarios, DEV_GSM_RUIDONE. Para DEV_GSM_COCHE se realiza el estudio comparativo representando las curvas DET para las distintas SNRs (Fig.4.5-4.8).

Por otro lado, los resultados anteriores indican que, para todas las SNRs, el nuevo *score* funciona mucho mejor que el basado en sólo dos componentes, logaritmo de la energía normalizada y delta de energía, similar al usado en el *VAD* de Acero. En la Tabla 4.2 se muestra la tasa de falsas alarmas y la de falsos rechazos teniendo en cuenta que las tramas con *scores* (el *score* incluye la energía

normalizada) positivos son tramas de voz y las tramas con *scores* negativos son tramas de ruido o de no voz:

SNR (dB)	Falsas alarmas. *score*=0 (c0norm + deltac0)	Falsos rechazos. *score*=0 (c0norm + deltac0)	Falsas alarmas. *score*=0 (5 componentes)	Falsos rechazos. *score*=0 (5 componentes)
0	1,58%	67,23%	1,55%	55,67%
5	1,38%	61,53%	1,27%	43,72%
15	2,09%	54,46%	1,83%	32,88%
20	2,62%	52,89%	2,34%	31,25%

Tabla 4.2. Tasa de falsas alarmas y tasa de falsos rechazos para DEV_GSM_COCHE y distintas SNRs.

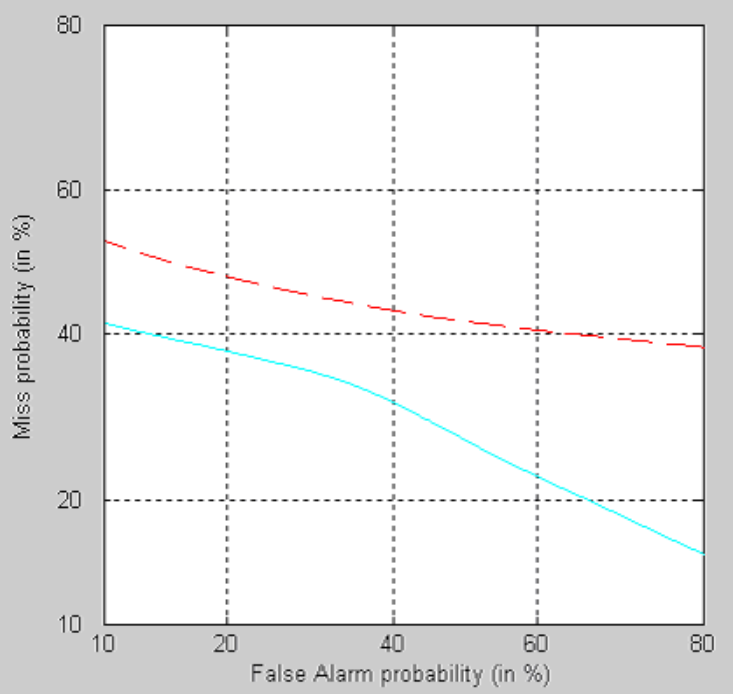

Figura 4.5. Curva DET comparando nuevo *score* (línea continua) con el *score* similar al de Acero (línea discontinua) para DEV_GSM_COCHE con SNR=0dB.

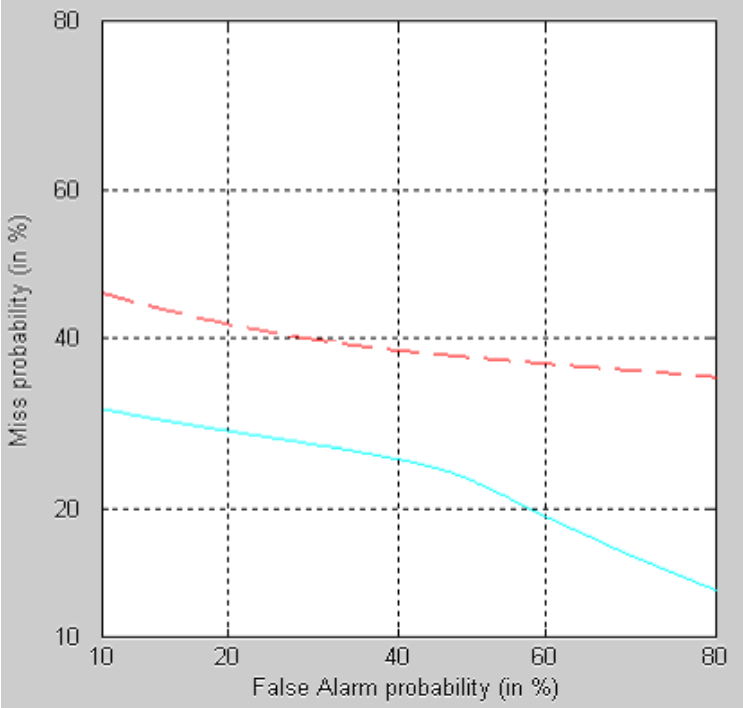

Figura 4.6. Curva DET comparando nuevo *score* (línea continua) con el *score* similar al de Acero (línea discontinua) para DEV_GSM_COCHE con SNR=5dB.

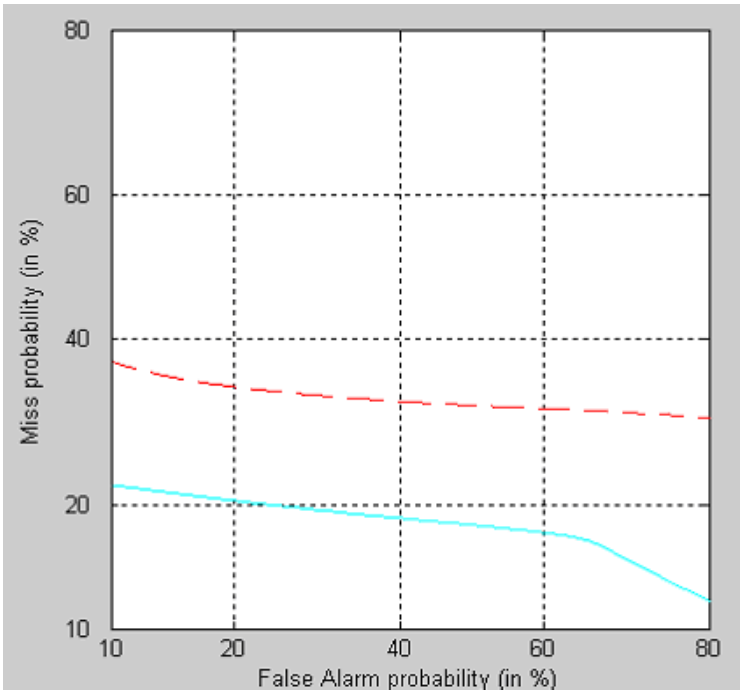

Figura 4.7. Curva DET comparando nuevo *score* (línea continua) con el *score* similar al de Acero (línea discontinua) para DEV_GSM_COCHE con SNR=15dB.

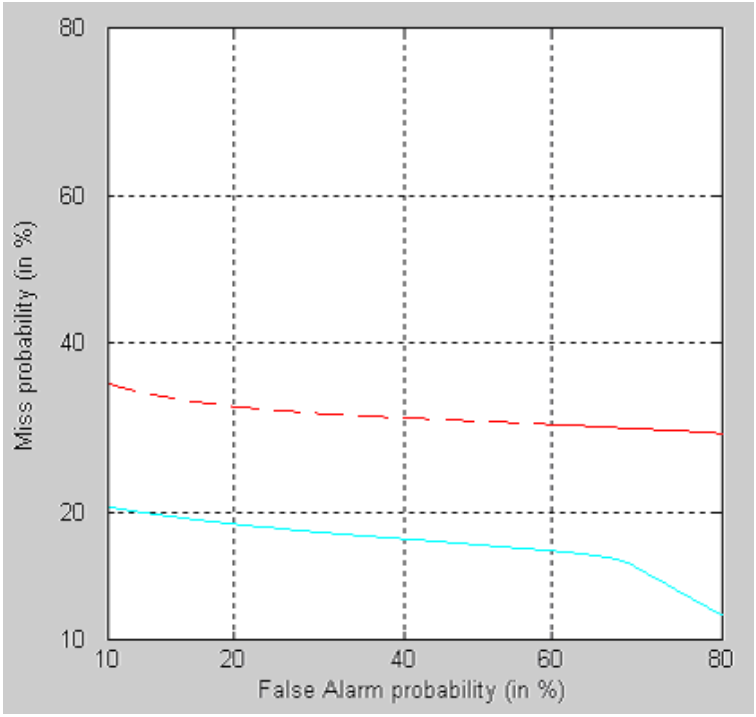

Figura 4.8. Curva DET comparando nuevo *score* (línea continua) con el *score* similar al de Acero (línea discontinua) para DEV_GSM_COCHE con SNR=20dB.

Las curvas DET de las figuras anteriores reflejan que las distribuciones de las clases "voz" y "no voz" poseen varianzas muy distintas: la gaussiana que representa la clase "no voz" es mucho más abrupta que la que representa la clase "voz".

En cuanto a los resultados de la Tabla 4.2, las falsas alarmas se normalizan al número de tramas de ruido y los falsos rechazos al número de tramas de voz. Si por el contrario se normaliza al número total de tramas se obtienen los siguientes resultados:

SNR (dB)	Falsas alarmas. *score*=0. (c0norm + deltac0)	Falsos rechazos *score*=0. (c0norm + deltac0)	Falsas alarmas. *score*=0 (5 componentes)	Falsos rechazos *score*=0. (5 componentes)
0	0,63%	40,26%	0,66%	33,34%
5	0,55%	36,84%	0,51%	26,18%
15	0,84%	32,61%	0,73%	19,39%
20	1,05%	31,67%	0,94%	18,71%

Tabla 4.3. Tasa de falsas alarmas y tasa de falsos rechazos respecto del número total de tramas para DEV_GSM_COCHE. Distintas SNRs.

Como se puede observar, la tasa de falsas alarmas es muy baja y constante para todas las SNRs tanto en el caso de dos características como en el de cinco, y cabe destacar que, aunque las diferencias son pequeñas, el vector de cinco componentes muestra mejores resultados. Sin embargo, las diferencias son bastante más acusadas en el caso de falsos rechazos.

A continuación se muestra el estudio del comportamiento de las clases “voz” y “no voz” con una base de datos que contiene ruidos no estacionarios, DEV_GSM_RUIDONE. En este caso se muestra de nuevo la tasa de falsas alarmas y la de falsos rechazos con el mismo criterio de decisión que en el caso de DEV_GSM_COCHE:

SNR (dB)	Falsas alarmas *score*=0. (c0norm + deltac0)	Falsos Rechazos *score*=0. (c0norm + deltac0)	Falsas alarmas *score*=0. (5 componentes)	Falsos rechazos *score*=0 (5 componentes)
5	15.88%	82.12%	14.72%	82.03%
10	14.7%	77.63%	13.6%	76.11%
15	12.65%	72.61%	10.5%	66.89%
20	10.3%	66.63%	7.74%	54.75%
25	5.41%	55.95%	3.68%	35.1%

Tabla 4.4. Tasa de falsas alarmas y tasa de falsos rechazos para DEV_GSM_RUIDONE. Distintas SNRs.

En los resultados de la Tabla 4.4, las falsas alarmas se normalizan al número de tramas de ruido y los falsos rechazos al número de tramas de voz.

Como era de esperar, los resultados para ruido no estacionario son peores que para ruido estacionario: la tasa de falsas alarmas aumenta debido a la existencia de voces de fondo. En cualquier caso, los resultados con el nuevo vector de características de cinco componentes son mejores, tanto en la tasa de falsas alarmas como en la tasa de falsos rechazos, sobre todo para SNRs mayores.

4.4.- Algoritmo basado en HMMs y estudio de diferentes topologías.

Una vez seleccionado el vector de características $v(n)$, se trata de modelar las distribuciones condicionadas para obtener el LLR (Log-likelihood Ratio). Antes del cálculo del LLR debemos seleccionar la topología óptima.

Un importante grado de libertad dentro de la naturaleza de los modelos ocultos de Markov es su topología: número de estados de los modelos, transición entre estos estados y número de gaussianas por estado. Típicamente, la complejidad de la aplicación o el número de estados a tratar son factores que pueden afectar directamente sobre el funcionamiento de una topología específica. En este apartado se realizan una serie de pruebas para encontrar la topología que obtiene mejores resultados de clasificación de las clases “voz” y “no-voz”.

La topología de los modelos es un aspecto importante en cualquier aplicación, en especial las que funcionan en tiempo real. Es esperable que topologías más complejas generen un *GDE* (Global Detection Error) menor, sin embargo, este hecho incrementa el tiempo de proceso del *VAD*. Por lo tanto, hay que encontrar el punto óptimo en el que se tenga una topología sencilla y un tiempo de retardo en proceso pequeño. Son por tanto estos los dos parámetros con los que tendremos que jugar en la medida de lo posible.

Los HMMs a tratar son dos, uno que representa al modelo de voz, y otro que representa al modelo de ruido. Dado que el modelo de voz es más complicado de modelar dada su variabilidad acústica, para nuestro estudio, dotaremos al modelo de voz con un estado más que al modelo de ruido (ya lo hacía Acero en el *VAD* que propone [82]). Las topologías estudiadas son las siguientes:

- Topología 1: 2 estados de voz y 1 estado de ruido.
- Topología 2: 3 estados de voz y 2 estados de ruido.
- Topología 3: 4 estados de voz y 3 estados de ruido.
- Topología 4: 5 estados de voz y 4 estados de ruido.
- Topología 5: 6 estados de voz y 5 estados de ruido.

En Fig.4.9-4.11 se realiza una representación gráfica de algunas de las topologías enumeradas. Es importante comentar que en cada una de las distintas topologías, cada estado posee una única gaussiana caracterizada por su media y varianza. Además, topologías con diferente número de estados se pueden considerar como estados con diferente número de gaussianas.

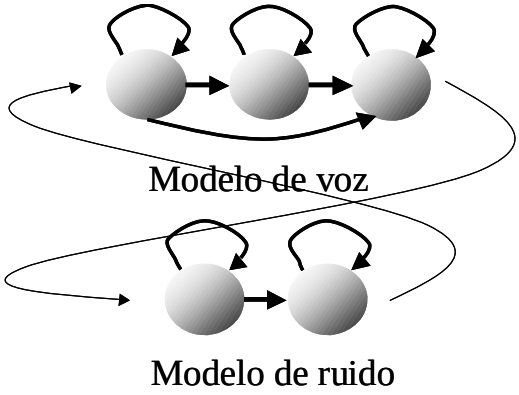

Figura 4.9. Topología 2: 3 estados de voz y 2 estados de ruido.

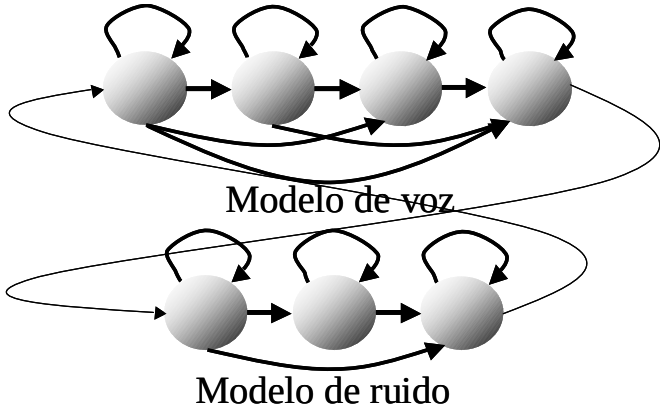

Figura 4.10. Topología 3: 4 estados de voz y 3 estados de ruido.

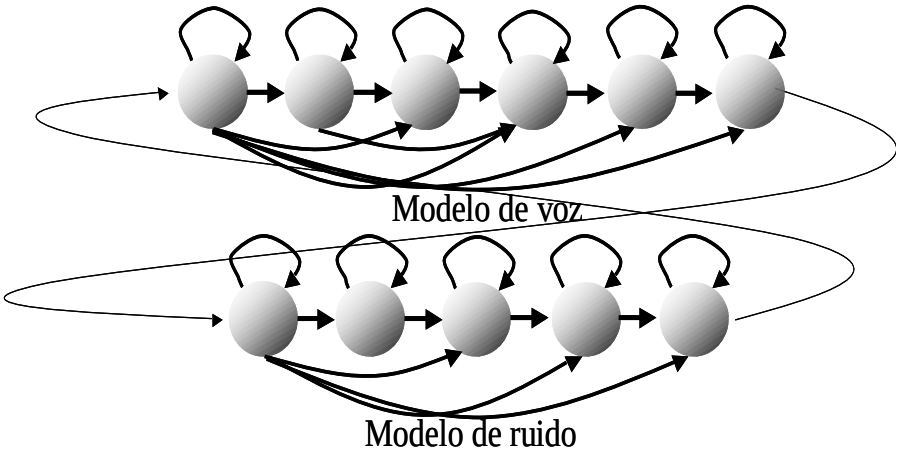

Figura 4.11. Topología 5: 6 estados de voz y 5 estados de ruido.

Como se puede observar en las figuras anteriores, en todos los casos, se trata de modelos de ocultos de Markov de izquierda a derecha y además se realimentan, esto es, cuando finaliza el modelo de ruido puede comenzar el de voz o el de ruido.

Se realizaron distintos procesos de entrenamiento, cada uno de ellos referente a cada una de las topologías contempladas, mediante la base de datos TRAIN_GSM_HMM. Se presenta en la Fig.4.12, de forma gráfica, tanto el *GDE* como la tasa de falsas alarmas (FA) para la base de datos DEV_GSM_LIMPIA (ruidos reales pero con niveles bajos). En ella se puede observar la disminución tanto del *GDE* como de la tasa de falsas alarmas conforme aumenta el número de estados de las topologías de los modelos a tener en cuenta.

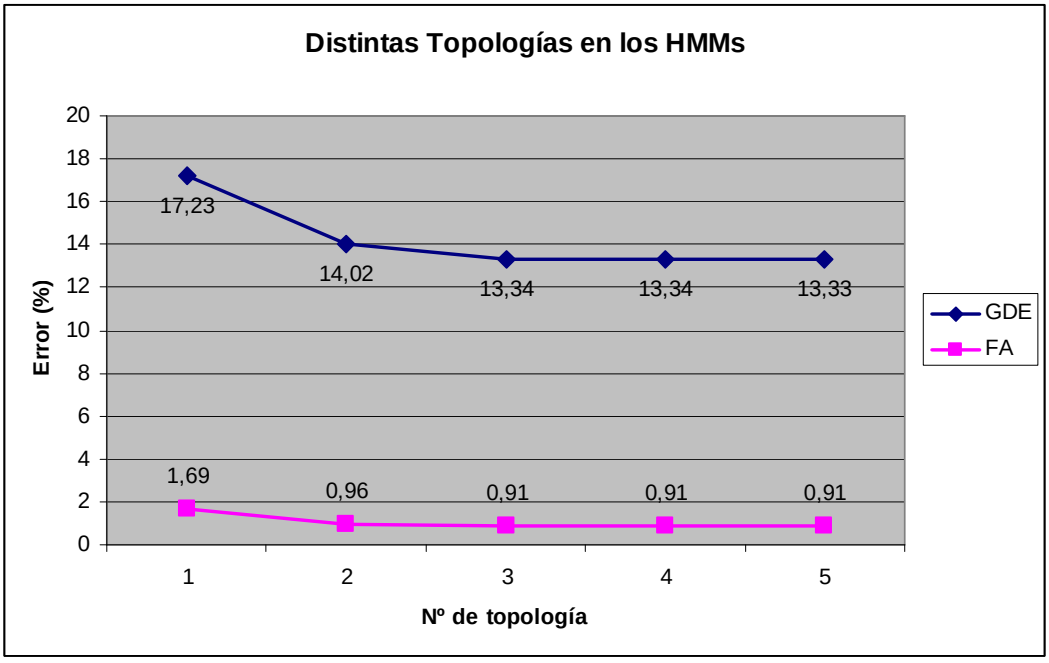

Figura 4.12. *GDE* y FA: distintas topologías en los HMMs y usando DEV_GSM_LIMPIA.

Como se puede observar en la Fig.4.12, a partir de la topología 3, y, aunque se amplíe el número de estados, los resultados prácticamente no mejoran, y es por esto por lo que elegimos como topología "óptima" la topología 3: una topología más compleja incrementaría el tiempo de retardo del *VAD* sin mejorar prácticamente nada el *GDE* final. Por tanto, la topología elegida es la de la Fig.4.10. Los dos son

Modelos de Markov de izquierda a derecha con 3 y 4 estados para los modelos de ruido y voz respectivamente, y una gaussiana por estado. Comentar que el *VAD* presentado por Acero [82] también usa exactamente misma topología.

4.5.- Combinación de puntuaciones de verosimilitud con valores de la energía.

En este apartado se estudia la posibilidad de establecer un criterio de decisión a nivel de trama que pueda tener en cuenta, además de la puntuación de verosimilitudes de las dos clases (LLR), alguna información de la energía, como se hace habitualmente en detectores simples basados en energía. Para ello, en esta sección, se realizan los siguientes estudios:

1. Análisis de las distribuciones de las clases "voz" y "no voz" a partir de las puntuaciones de verosimilitud en distintas situaciones.
2. Comparación del "*score*" con los valores del logaritmo de la energía.
3. Comparación del "*score*" con los valores del logaritmo de la energía normalizada.
4. Combinación del "*score*" y del logaritmo de la energía normalizada en el criterio de decisión a nivel de trama.

En los siguientes subapartados se desarrollan los cuatro aspectos anteriores.

4.5.1.- Puntuaciones de verosimilitud en distintas situaciones.

Teniendo en cuenta el algoritmo de clasificación a nivel de trama en función del cociente de verosimilitud de las dos clases (*score*) se va a realizar el estudio de cómo se comporta esta puntuación en distintas situaciones. Para ello, se ha usado la base de datos de entrenamiento TRAIN_GSM_HMM: voz limpia con ruidos no estacionarios. También fue necesario etiquetar todos sus ficheros en segmentos de voz y de ruido para que HTK [34] pudiese calcular los modelos convenientemente.

Por otro lado, y para evaluar el nuevo *VAD*, los resultados que a continuación se muestran se calculan sobre otras dos bases de datos etiquetadas manualmente: DEV_GSM_COCHE (voz limpia con ruido de coche) y DEV_GSM_RUIDONE (voz limpia con ruidos no estacionarios).

Como primera etapa, se representarán las distribuciones de los "*scores*" o puntuaciones de las que se dispone de la etiqueta de voz y de ruido anotados manualmente para el caso de la base de datos de voz con ruido de coche y distintas SNRs (DEV_GSM_COCHE) en Fig.4.13-4.16. En estas figuras se puede observar que, el modelo de ruido, se mantiene muy estable para todas las SNRs, mientras que el de voz funciona mejor conforme aumenta la mencionada SNR. Esto se traduce en que la media y la varianza de la distribución del *score* en las tramas de ruido no cambia para distintas SNRs, por tanto es esperable una tasa de falsas alarmas constante y pequeña por ejemplo para un *score* igual a cero. Por el contrario, en la distribución del *score* en las tramas de voz ocurre lo siguiente: la media toma valores mayores conforme aumenta la SNR (desplazamiento a valores de *score* mayores) mientras que la varianza aumenta proporcionalmente menos. Por tanto, aunque la distribución de ruido se mantenga estable, el área de solapamiento entre las dos clases disminuye conforme aumenta la SNR y por tanto el error de clasificación es menor para SNRs altas: en la Fig.4.13 el área de solapamiento es claramente mayor que en la Fig.4.16.

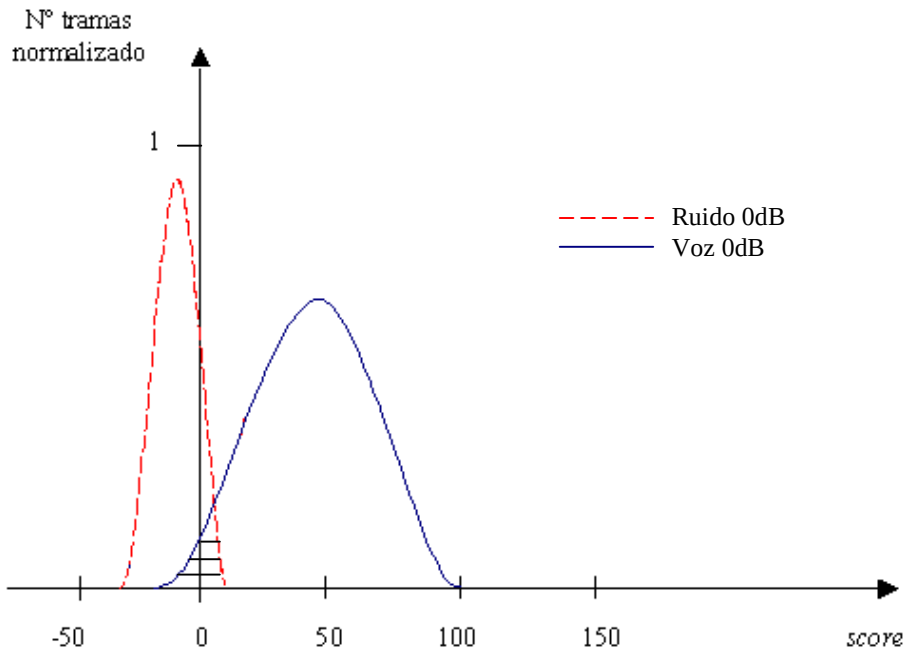

Figura 4.13. Distribución del nuevo *score* en tramas de ruido y de voz para DEV_GSM_COCHE. SNR=0dB.

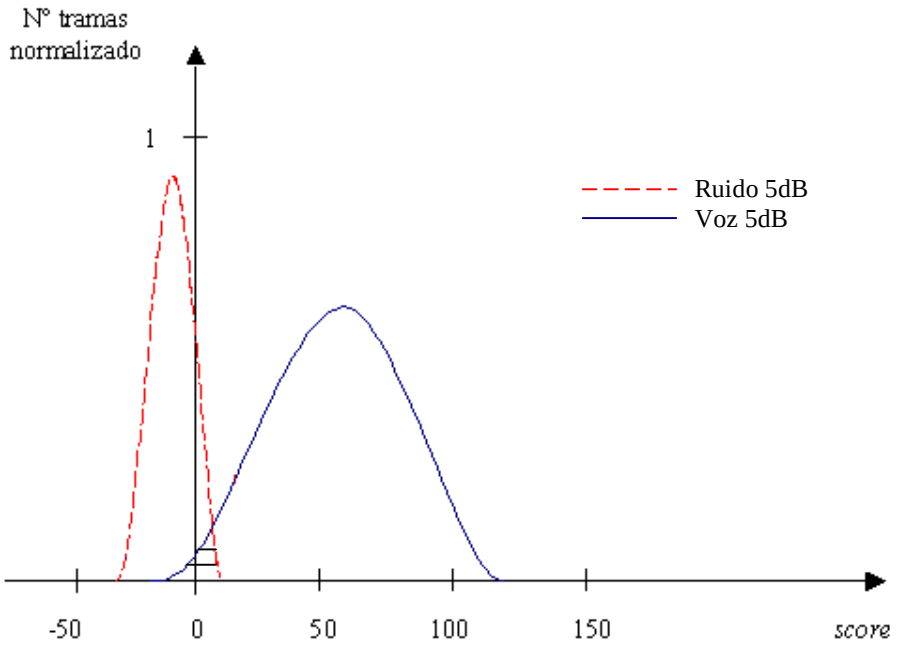

Figura 4.14. Distribución del nuevo *score* en tramas de ruido y de voz para DEV_GSM_COCHE. SNR=5dB.

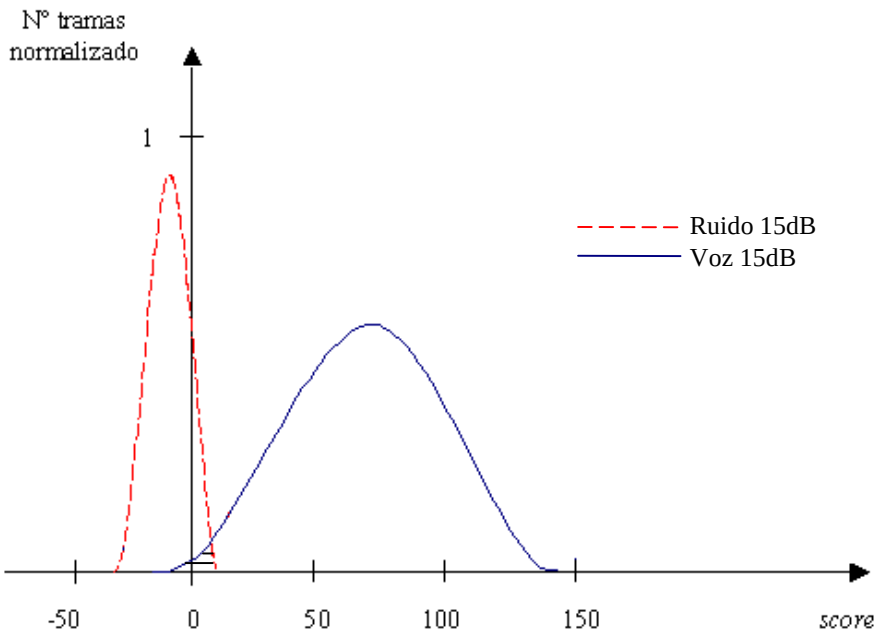

Figura 4.15. Distribución del nuevo *score* en tramas de ruido y de voz para DEV_GSM_COCHE. SNR=15dB.

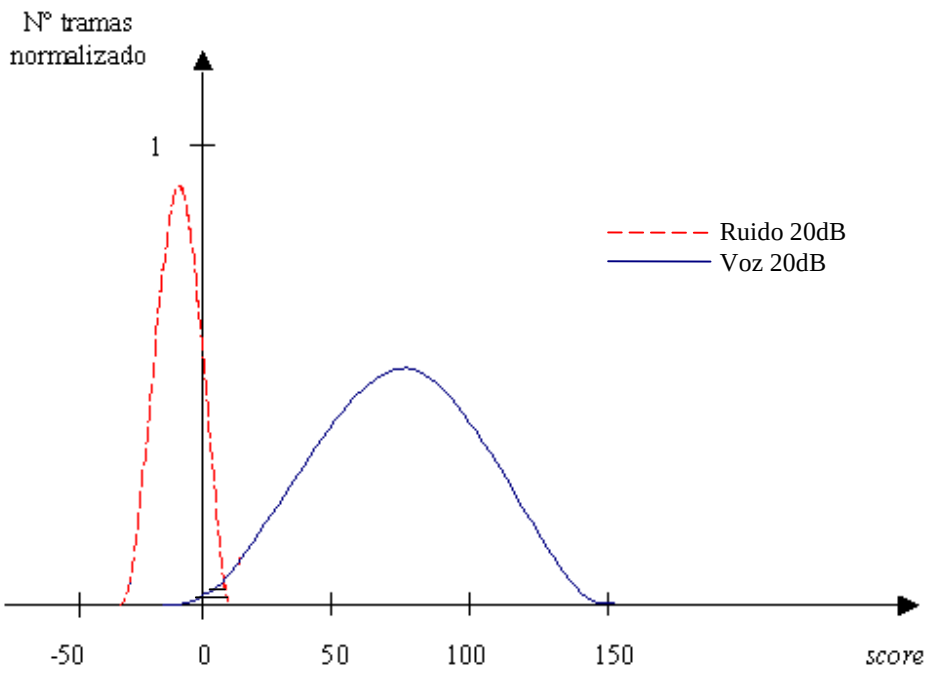

Figura 4.16. Distribución del nuevo *score* en tramas de ruido y de voz para DEV_GSM_COCHE. SNR=20dB.

Una vez representadas las distribuciones de *score* para ruidos estacionarios, y que, como era de esperar, funciona bien para SNRs bajas, vamos a realizar el estudio comparativo con el logaritmo de la energía y el logaritmo de la energía normalizada.

4.5.2.- Comparación del "*score*" con los valores del logaritmo de la energía.

En este subapartado se presentan las curvas DET usando la base de datos DEV_GSM_COCHE y comparando el *score* del nuevo *VAD* con el logaritmo de la energía (Fig.4.17 y Fig.4.18). Como se puede observar en la Fig.4.17, cuando el umbral de energía se ajusta a una SNR determinada, funciona mejor como característica que el *score*. El problema resulta cuando se cambian las condiciones de entorno (diferente SNR). Además habría que considerar tener distintos modelos para las diferentes SNRs o distintos umbrales de energía, cuestión que complica el algoritmo de decisión.

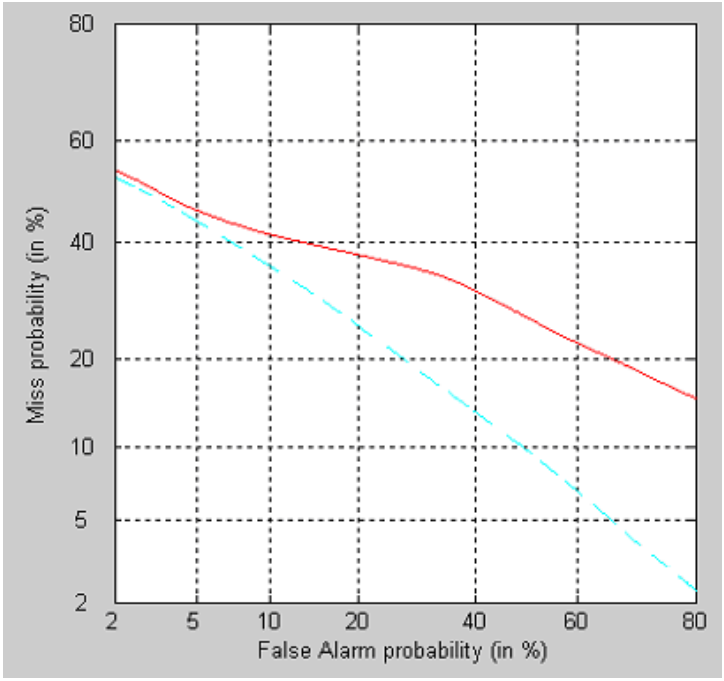

Figura 4.17. Curva DET comparando nuevo *score* (línea continua) y logaritmo de la energía ajustado (línea discontinua) para DEV_GSM_COCHE con SNR=0dB.

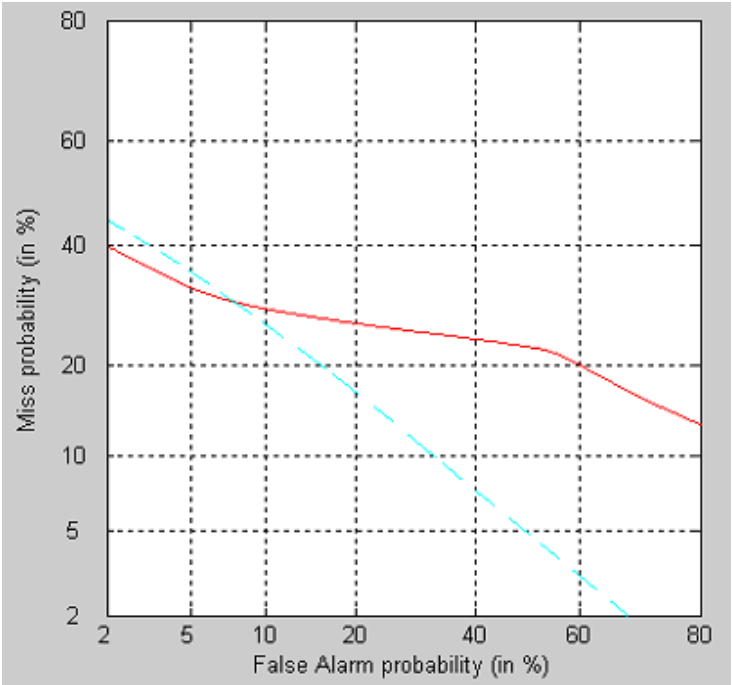

Figura 4.18. Curva DET comparando nuevo *score* (línea continua) y log. de la energía ajustado (línea discontinua) para DEV_GSM_COCHE: distintas SNRs (0, 10, 15 y 20dB).

Por otro lado, en la curva DET de la Fig.4.18, que tiene en cuenta todas las SNRs, el *score*, incluso en la región izquierda del dibujo funciona mejor como característica que la energía, incluso esta última funcionando con distintos umbrales de ajuste para cada una de las SNRs.

4.5.3.- Comparación del "*score*" con los valores del logaritmo de la energía normalizada.

En cuanto al funcionamiento del logaritmo de la energía normalizada como característica, el *score* siempre es mejor en todos los casos (Fig.4.19 y Fig.4.20). Por tanto, no se puede pensar en usar el logaritmo de la energía normalizada como única característica para la toma de decisión, esto es, un *VAD* basado únicamente en el logaritmo de la energía normalizada.

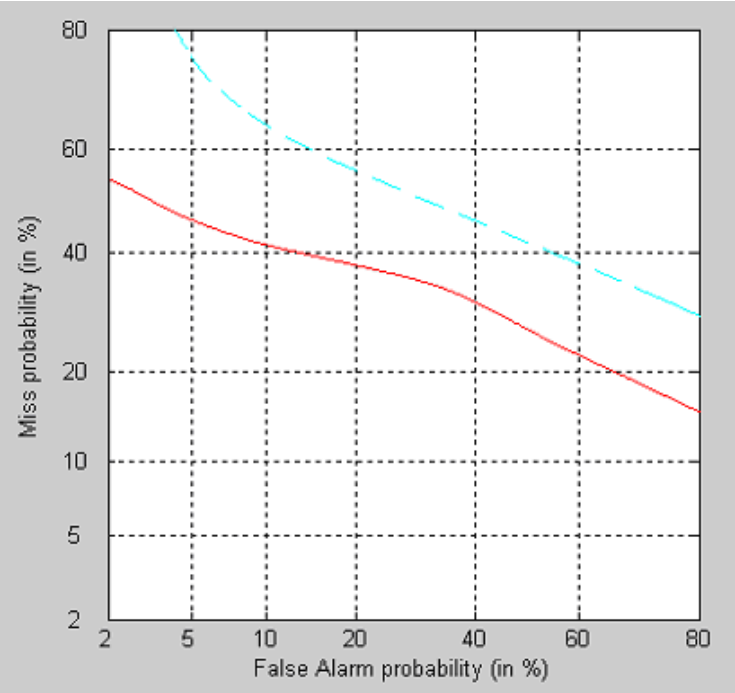

Figura 4.19. Curva DET comparando nuevo *score* (línea continua) y log. de la energía normalizada (línea discontinua) para DEV_GSM_COCHE con SNR=0dB.

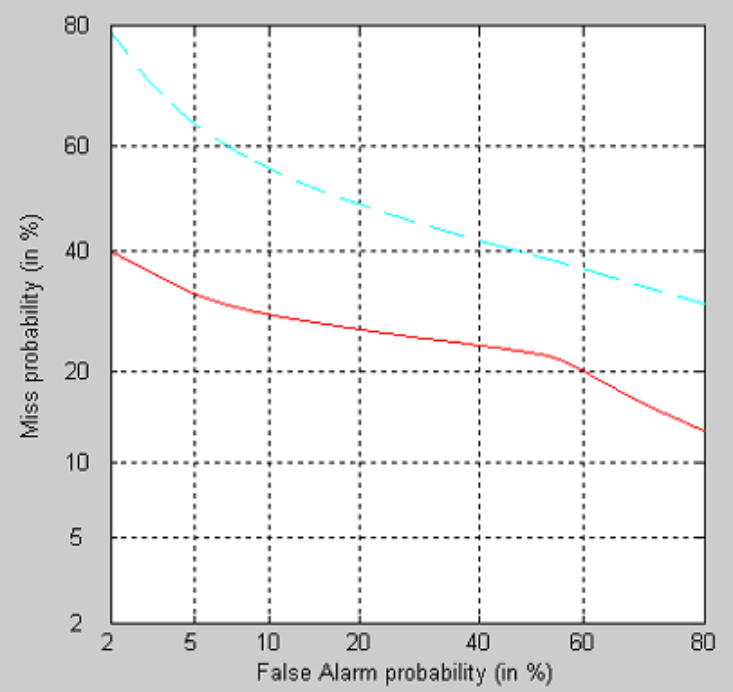

Figura 4.20. Curva DET comparando nuevo *score* (línea continua) y log. de la energía normalizada (línea discontinua) para DEV_GSM_COCHE: distintas SNRs (0, 10, 15 y 20dB).

Llegado a este punto, y, demostrada la eficacia del *score* como característica, se demuestra que, el valor del *score* de EER (Equal Error Rate) permanece invariante ante las distintas SNRs. Este valor se sitúa en *score*=-10 (se representan las falsas alarmas y los falsos rechazos en Fig.4.21-4.24 realizando un barrido con los valores de *score*). El EER en estas figuras es el punto de intersección entre la línea punteada (falso rechazo) y la línea contínua (falsas aceptaciones). Por ejemplo, en la Fig.4.21, siendo el eje X el valor del *score*, el EER es del 35% aproximadamente, correspondiendo a un valor de *score*=-10. Además, se puede ver que para el *score* igual a 0 (máxima verosimilitud), prácticamente ya no existen falsas aceptaciones: este efecto también se puede visualizar en Fig.4.22-4.24. Por último, comentar que el EER va disminuyendo conforme aumenta la SNR, desde el 35% para una SNR=0dB (Fig.4.21) hasta el 20% para SNR=20dB (Fig.4.24).

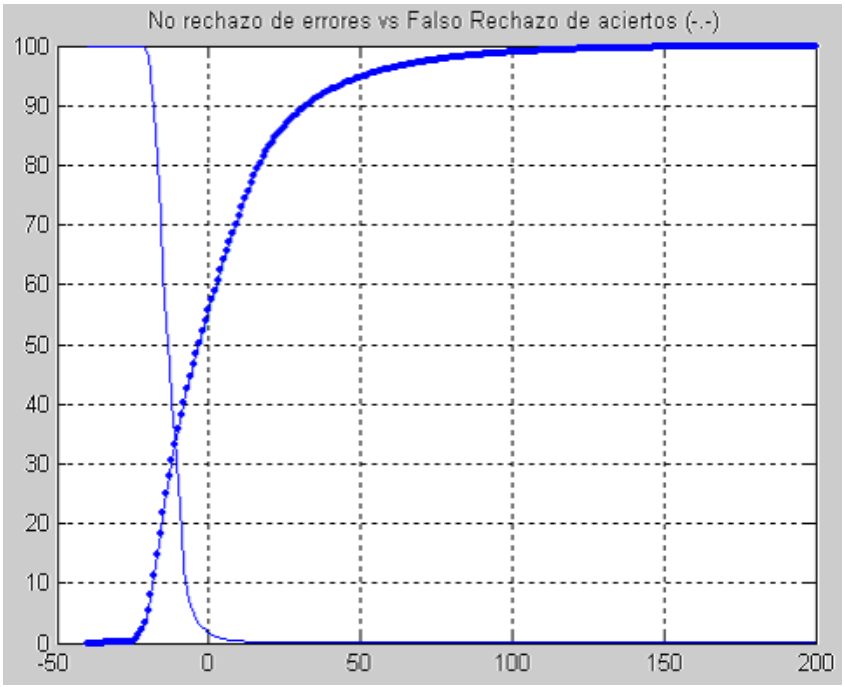

Figura 4.21. Barrido con valores de *score* (eje X) para DEV_GSM_COCHE con SNR=0dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

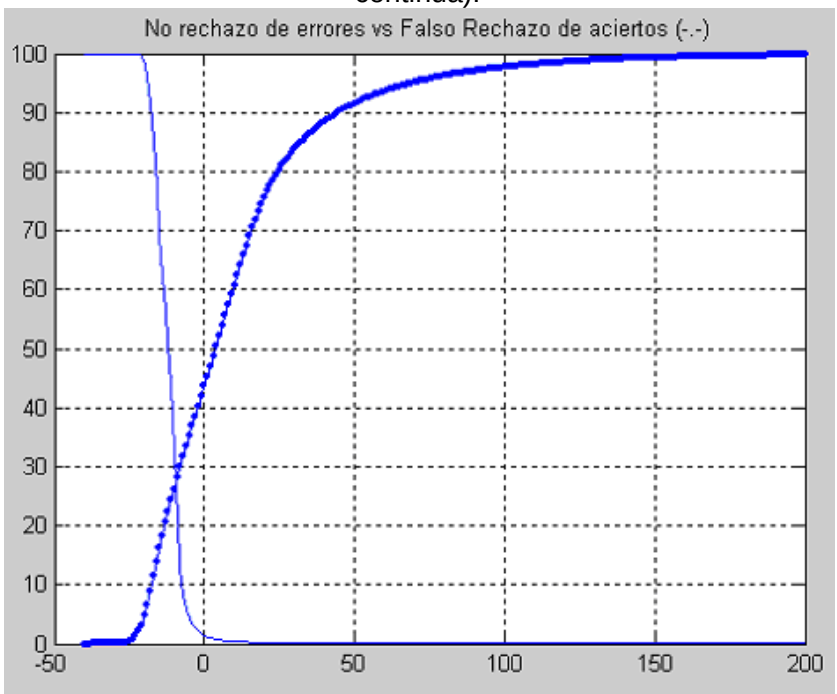

Figura 4.22. Barrido con valores de *score* (eje X) para DEV_GSM_COCHE con SNR=5dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

Figura 4.23. Barrido con valores de *score* (eje X) para DEV_GSM_COCHE con SNR=15dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

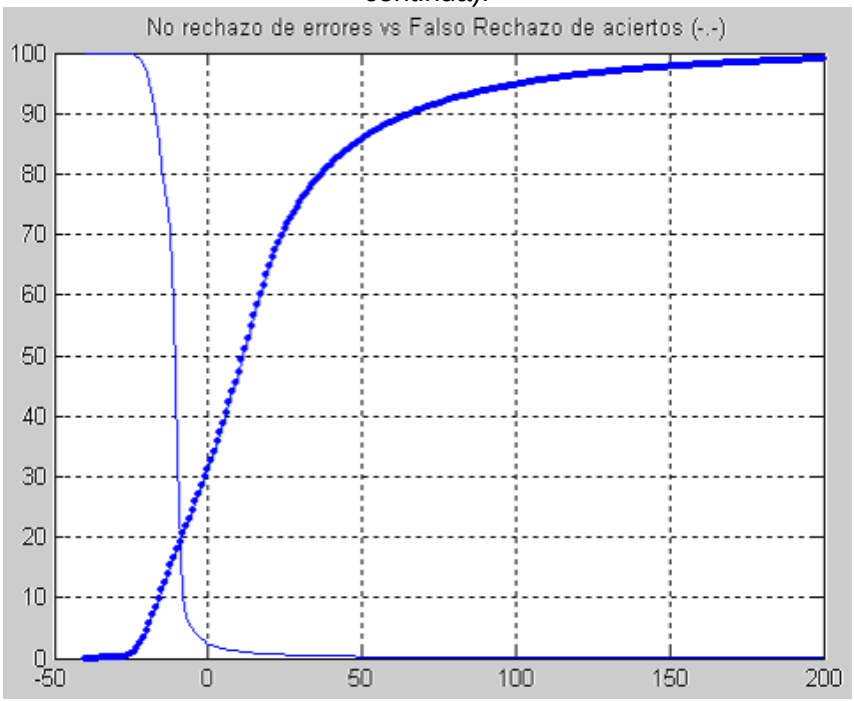

Figura 4.24. Barrido con valores de *score* (eje X) para DEV_GSM_COCHE con SNR=20dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

4.5.4.- Combinación del "score" y del logaritmo de la energía normalizada en el criterio de decisión a nivel de trama.

En este subapartado se estudia la combinación del "*score*" y del logaritmo de la energía normalizada en el criterio de decisión a nivel de trama. El punto de partida, de igual forma que en el apartado 4.4, es la máxima verosimilitud entre clases, esto es, se considera únicamente un valor (umbral) de *score* igual a cero para la toma de decisión a nivel de trama: independiente de la SNR. Ahora, con ánimo de realizar mejoras sobre los resultados obtenidos, se estudiará el efecto de añadir información del logaritmo de la energía normalizada para realizar la toma de decisión final. La decisión basada en *score* conlleva a errores de clasificación, sobre todo provocado por las falsas aceptaciones. Los criterios de decisión quedarían de esta manera:

- $score \geq 0$ y logaritmo de la energía normalizada $\geq 0 \Rightarrow$ Trama de voz.
- $score \geq 0$ y logaritmo de la energía normalizada $< 0 \Rightarrow$ Trama de ruido.
- $score < 0 \Rightarrow$ Trama de ruido.

A continuación se va a realizar el estudio combinando *score* y logaritmo de la energía normalizada para la base de datos DEV_GSM_COCHE (distintas SNRs). En las Fig. 4.25-4.28 se representa, realizando un barrido para distintos valores de umbral del logaritmo de la energía normalizada (eje X), la tasa de falsas alarmas y la tasa de falsos rechazos sobre lo que se ha dicho que es voz tras imponer la condición $score \geq 0$. En este caso ya no se puede mejorar la tasa de falsos rechazos obtenida cuando el sistema se equivoca al clasificar con la condición $score < 0$, sino que sólo se trabajará para intentar mejorar la tasa de falsas aceptaciones. Para ello se tendrán que evaluar dos aspectos:

1. Con la condición $score \geq 0$ se tiene una tasa de falsas aceptaciones determinada. Al imponer la segunda condición es esperable que algunas tramas que se han clasificado "mal" como voz se clasifiquen finalmente como ruido, partiendo de un total del 100% (100 en el eje Y de Fig.4.25-4.28) de tasa de falsas aceptaciones que se tenía con la primera condición ($score \geq 0$). En Fig.4.25-4.28 la línea continua representa la mencionada tasa de falsas aceptaciones.

2. Con la condición *score* ≥ 0 se habrán clasificado bien un cierto número de tramas, las que son realmente de voz. Al imponer la segunda condición es esperable que algunas tramas que son realmente voz se terminen clasificando finalmente como ruido, generando de esta manera una tasa de falsos rechazos sobre lo que se clasificó bien con la primera condición. Por tanto se parte en este caso de un 0% (0 en eje Y de Fig.4.25-4.28) de tasa de falsos rechazos que se tenía con la primera condición (*score* ≥ 0). En Fig.4.25-4.28 la línea punteada representa la mencionada tasa de falsos rechazos.

Por tanto, se trata de evaluar cuánto se puede disminuir la tasa de falsas aceptaciones variando el umbral de decisión sobre el logaritmo de la energía normalizada, y por ello, cuánto pierdo (tasa de falsos rechazos) de lo que ya se tenía bien clasificado con la primera condición al imponer la segunda.

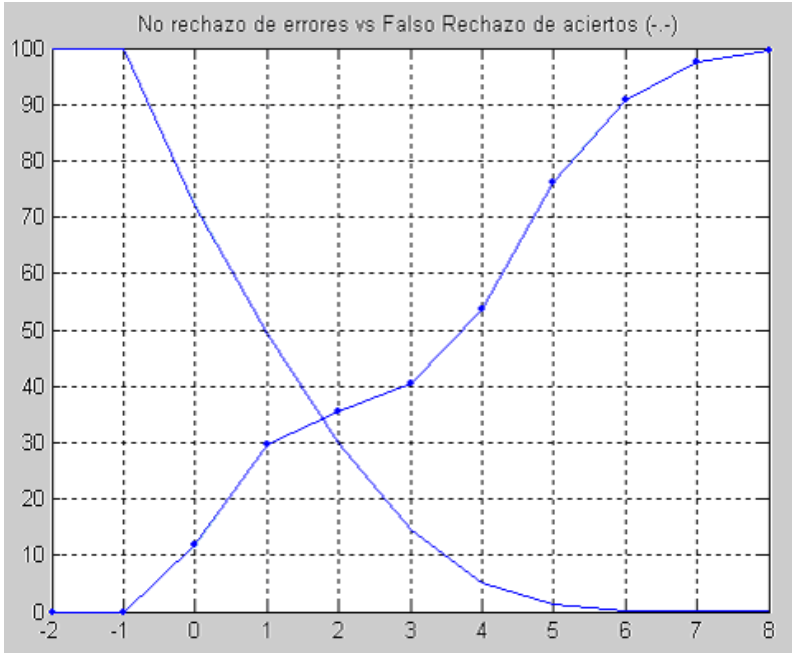

Figura 4.25. Barrido de umbrales del logaritmo de la energía normalizada para DEV_GSM_COCHE con SNR=0dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

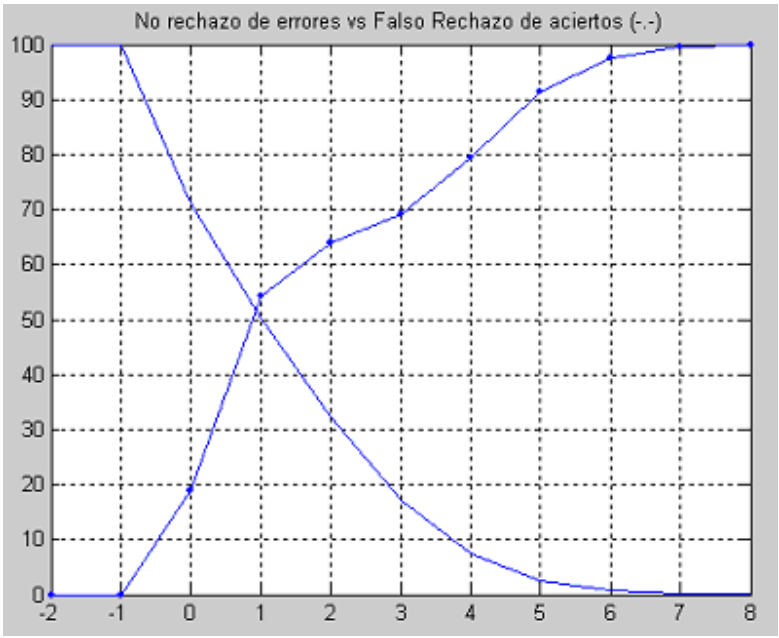

Figura 4.26. Barrido de umbrales del logaritmo de la energía normalizada para DEV_GSM_COCHE con SNR=5dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

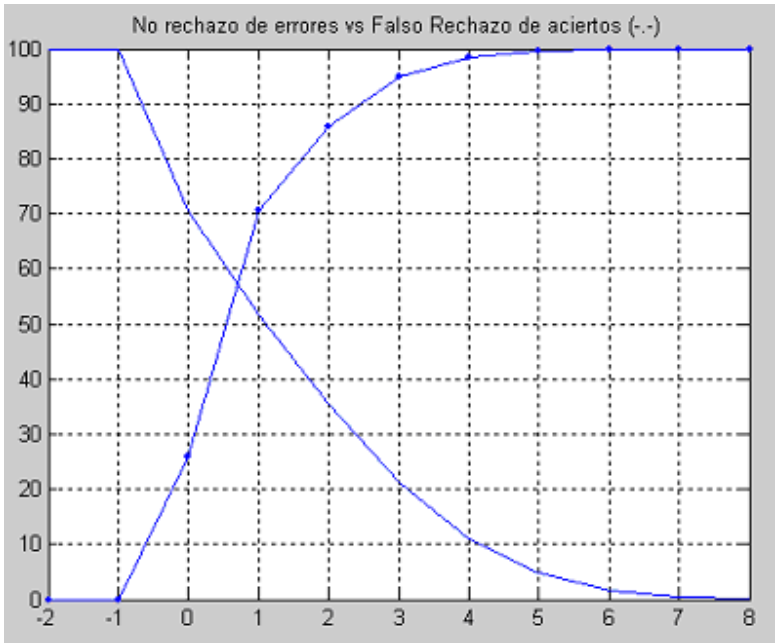

Figura 4.27. Barrido de umbrales del logaritmo de la energía normalizada para DEV_GSM_COCHE con SNR=15dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

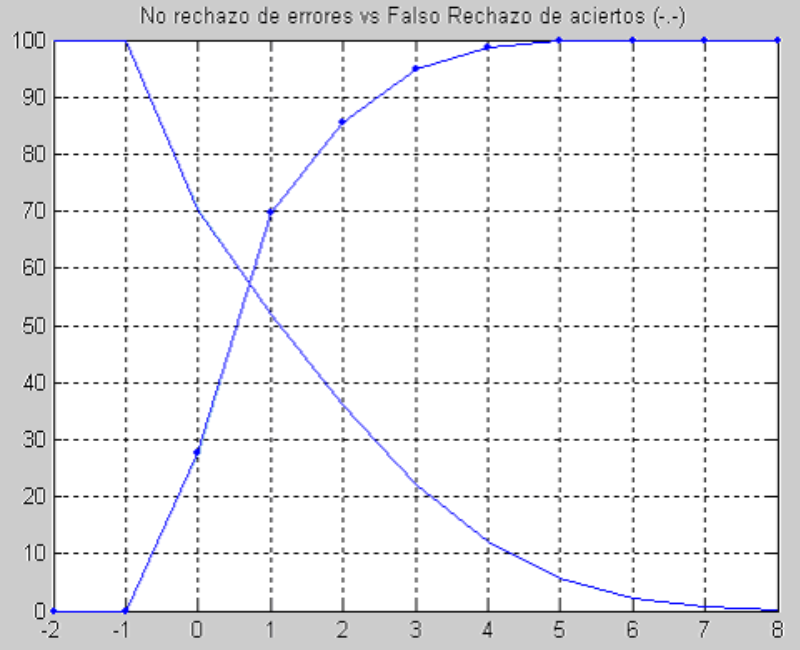

Figura 4.28. Barrido de umbrales del logaritmo de la energía normalizada para DEV_GSM_COCHE con SNR=20dB. Errores de falso rechazo (línea punteada) y falsas aceptaciones (línea continua).

Antes de continuar, se va a explicar el significado de lo obtenido en la Fig.4.25. Eligiendo cero como valor de referencia de logaritmo de la energía normalizada, se puede ver que disminuyen los errores de tasa de falsas alarmas, y lo hacen en un 28.3% relativo, respecto de la tasa de falsas aceptaciones que se obtenía al imponer sólo la condición $score \geq 0$. En cuanto a la tasa de falso rechazo, se empeora un 11.6% relativo (al empeorar a partir de ahora se hablará de valor negativo: -11.6%) respecto al número total de tramas de voz acertadas con la condición $score \geq 0$. Esto quiere decir que en las tramas de ruido que puntuaron con *score* positivo se mejora en un 28.3% y que en las tramas de voz que puntuaron con *score* positivo algunas de ellas puntuarían con logaritmo de la energía normalizada negativo (algunas tramas de voz bien clasificadas por el *score* serían rechazadas al imponer la condición del umbral de la energía normalizada), y se empeoraría un 11.6%. Este hecho se debe a la existencia de segmentos de voz de baja energía. De todas formas, hablando en términos relativos se tendría una mejora sustancial: 28.3% - 11.6% = 16.7%. Sin embargo, en nuestro caso, hablando en términos absolutos, es mucho mayor el número de falsos rechazos, por lo tanto no interesa

aplicar esta segunda decisión. De esta manera, todo va a depender del cociente entre número de falsas alarmas y número de falsos rechazos: este cociente en nuestro caso está cerca de cero. Si por el contrario el *VAD* se encontrara en otro punto de trabajo, por ejemplo en el EER (Equal Error Rate) para la decisión del *score*, el cociente del que se hablaba anteriormente sería uno y sí que habría una mejora también en términos absolutos (términos absolutos = términos relativos). Y por supuesto, ni que decir tiene que si el cociente es mayor que uno, se mejoraría más todavía. En nuestro trabajo nos interesa perder el mínimo número de tramas de voz posible ya que, el *VAD* está orientado a trabajar en aplicaciones de reconocimiento automático de habla. Por tanto, no nos interesa aplicar esta segunda condición. Sin embargo, en otro tipo de aplicaciones en las que no importe mucho perder algunas tramas de voz, será muy interesante aplicarla.

En forma de tabla, y, numéricamente, se exponen los errores relativos, respecto a los errores obtenidos a partir de la condición de clasificar como trama de voz la que obtiene un *score* positivo, de Fig.4.25, Fig.4.26, Fig.4.27 y Fig.4.28:

SNR (dB)	Mejora en tasa de falsas alarmas	Mejora en tasa de falsos rechazos
0	28.3%	-11.6%
5	29.3%	-18.7%
15	29.5%	-25.8%
20	29.9%	-27.7%

Tabla 4.5. Mejoras relativas tras imponer la condición del umbral de la energía normalizada respecto a los resultados obtenidos sobre las tramas puntuadas con *score* positivo.

Se puede observar en la Tabla 4.5 que, globalmente, se mejora más cuando la SNR es más pequeña, aunque en todos los casos la mejora en la tasa de falsas alarmas es mayor que lo que empeora la tasa de falsos rechazos.

Se concluye que, para aplicar la segunda condición, y para la toma de decisión será muy importante tener en cuenta el punto de trabajo en el que se encuentre el *VAD* (referido a la decisión del *score*), y esto en la mayoría de las ocasiones se encuentra condicionado por el tipo de aplicación en la que se vaya a utilizar el *VAD*.

4.6.- Resultados globales del VAD.

En esta sección se presentan los resultados finales del sistema de detección completo incluyendo la información a nivel de pulso (decisión a nivel de pulso: máquina de estados y detección de pulsos) y se comparan con los de otros *VAD*s estándar de referencia: AMR2, AURORA(FD), G729 anexo B y AMR1. Para ello se han usados tres bases de datos, distintas a la del apartado anterior, TEST_GSM_LIMPIA, TEST_GSM_COCHE y TEST_GSM_RUIDONE:

En las figuras que se presentan a continuación, se muestra el error global de detección (*GDE*: Global_Detection_Error) que recoge, como ya se explicó en el apartado 3.4 del capítulo anterior, la media de los errores de detección de las tramas de voz y de ruido o no-voz, esto es, la media de la tasa de falsas alarmas y de la tasa de falsos rechazos.

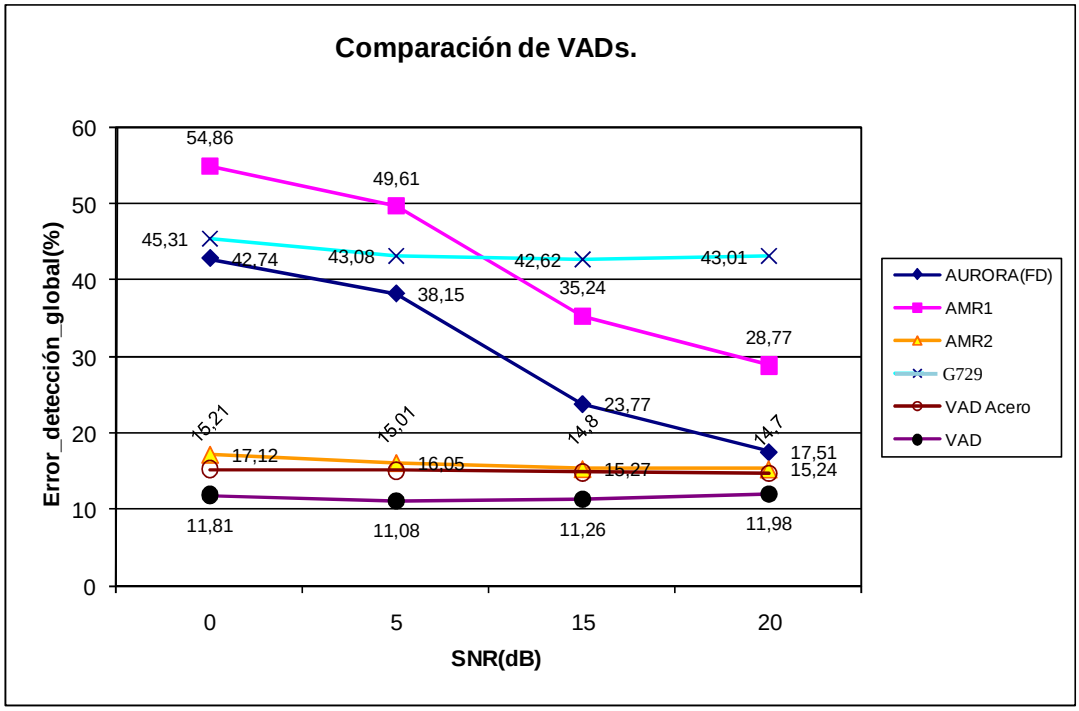

Figura 4.29. Error de detección global para diferentes SNRs para TEST_GSM_COCHE. Ruido estacionario.

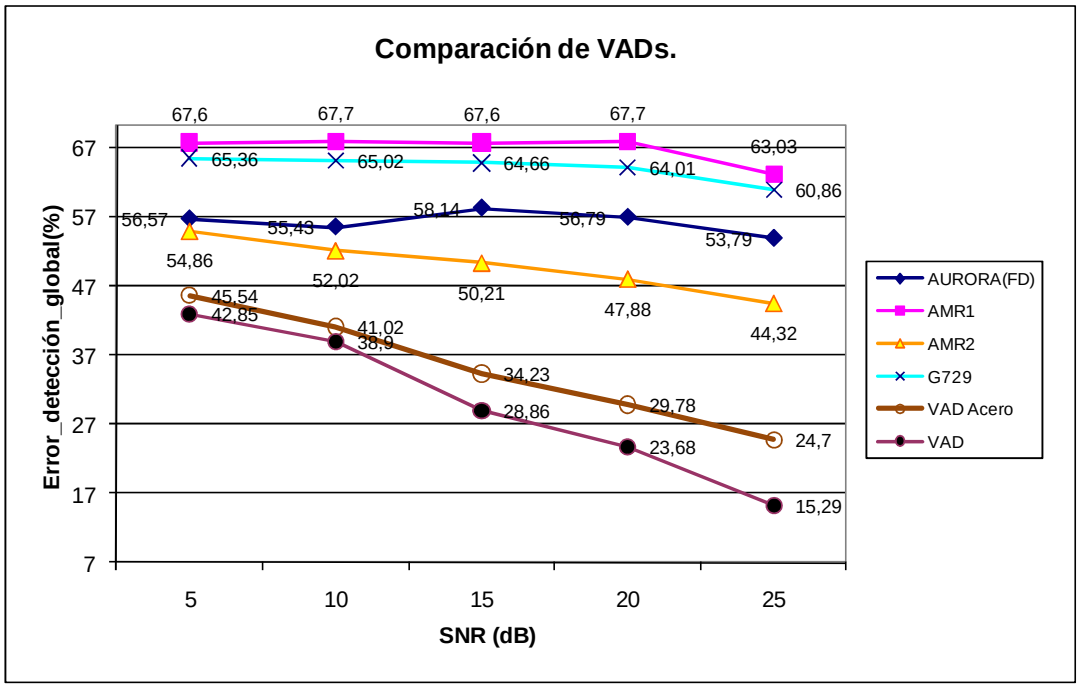

Figura 4.30. Error de detección global para diferentes SNRs para TEST_GSM_RUIDONE. Ruido no estacionario.

Para el caso de la base de datos de voz limpia, TEST_GSM_LIMPIA, los resultados se muestran en la Tabla 4.6.

GDE	AURORA(FD)	AMR1	AMR2	g729b	VAD
%	40.59	22.38	13.55	32.98	**13.34**

Tabla 4.6. Error de detección global (*GDE*) para TEST_GSM_LIMPIA.

El *VAD* propuesto obtiene los mejores resultados con las tres bases de datos (TEST_GSM_LIMPIA, TEST_GSM_COCHE y TEST_GSM_RUIDONE), seguido del AMR2, en términos generales. Particularmente, en voz limpia el *VAD* del estándar AURORA es el que peores resultados obtiene. Sin embargo, tanto para ruido de fondo estacionario como no estacionario obtiene el mejor tercer resultado, y es destacable su funcionamiento lineal en el caso de ruidos no estacionarios, cuestión que ocurre con el *VAD* propuesto con ruidos estacionarios (menos del 12% para todas las SNRs), gracias al uso de la información espectral y del logaritmo de la energía normalizada, mientras que disminuye el error global de detección cuando se

incrementa la SNR para el caso de la base de datos que contiene ruidos no estacionarios.

Numéricamente, y tomando de referencia el *VAD* de este capítulo, se ha obtenido una mejora relativa importante de error global de detección. La Tabla 4.7 muestra algunos ejemplos dignos de destacar. En cada casilla se obtiene la mejora relativa del *GDE* del *VAD* expuesto en este capítulo, frente a otros *VAD* de referencia, teniendo en cuenta 2 parámetros: base de datos y SNR utilizada por un lado y el *VAD* con el que comparar por otro.

Base de datos	AMR2	AURORA	G729b	AMR1
TEST_GSM_COCHE con 0dB	31.02%	72.37%	73.94%	78.47%
TEST_GSM_COCHE con 5dB	30.97%	70.96%	74.28%	77.67%
TEST_GMS_RUIDONE con 5dB	21.89%	24.6%	34.44%	36.61%
TEST_GSM_RUIDONE con 10dB	25.22%	29.82%	40.17%	42.54%
TEST_GSM_LIMPIA	1.55%	40.39%	59.55%	67.13%

Tabla 4.7. Mejora relativa del *GDE* del *VAD* expuesto frente a otros *VAD* de referencia.

Por ejemplo, la primera línea de la Tabla 4.7 habría que interpretarla de la siguiente manera: con la base de datos que contiene ruido de coche para una SNR de 0dB, la mejora relativa del *GDE* de nuestro VAD es del 31.02%, 72.37%, 73.94% y 78.47% si lo comparamos con los *VAD* del AMR2, AURORA, G729 y AMR1 respectivamente.

4.7.- Resultados de detección usando otras redes telefónicas.

Hasta el momento, se ha realizado un fuerte estudio del *VAD* de partida para probar su robustez ante todo tipo de ruidos usando tecnología GSM, la más utilizada en la actualidad y la que por su uso es la más compleja de tratar (un teléfono móvil se puede encontrar en los ambientes más adversos: bares, calle, etc.). Para completar el estudio del *VAD* de partida desarrollado en el capítulo anterior se han realizado pruebas con otro tipo de redes telefónicas: telefonía fija y voz IP.

En telefonía fija, el entorno es generalmente menos ruidoso, y es esperable que los resultados sean mejores que en telefonía móvil. En cuanto a voz IP, una de

las cosas más importante a tener en cuenta son los clicks del teclado que un usuario puede generar mientras habla, también tratables por los HMMs y eliminables mediante el método de detección de pulsos por duración (los clicks de un teclado de ordenador suelen tener una duración reducida).

Desde la aparición de la telefonía fija hace muchos años, las comunicaciones por medio de la conmutación de circuitos ha evolucionado mucho (PSTN, Red Telefónica Pública Conmutada o en inglés Public Switched Telephone Network). Los problemas de nivel debido a la distancia y como consecuencia a la atenuación de la señal de voz entrante se han solucionado mediante la incorporación de amplificadores adecuados. Además, en telefonía fija, el entorno en el que se habla generalmente es menos ruidoso que el de los móviles. Los niveles de ruido serán bastante pequeños. Por tanto, es esperable que los resultados sean mejores que en telefonía móvil.

El Detector de Actividad usado en las pruebas realizadas es el *VAD* de partida, es decir, el *VAD* completo descrito en este capítulo (Fig.4.1). Para entrenar los modelos de voz y de ruido para voz fija se usó la base de datos TRAIN_FIJA, formada por alrededor de 130.000 ficheros de audio emitidos por una gran diversidad de locutores y capturados sobre una red de telefonía fija. Los resultados se obtuvieron usando la base de datos TEST_FIJA, con una SNR media de 23.8 dB y formada por 2930 ficheros de conversaciones reales, etiquetados y capturados, obviamente, sobre la red de telefonía fija. La tasa de falsas alarmas fue del 29.18%, mientras que la de falsos rechazos del 3.81%. Teniendo en cuenta los valores anteriores se obtiene el error de detección global, *GDE*, igual al 16.46% (Tabla 4.8).

Base de datos	FAR	MR	*GDE*
TEST_FIJA	29.18%	3.81%	16.46%

Tabla 4.8. Resultados de detección para TEST_FIJA (Telefonía fija).

La voz sobre Protocolo de Internet, también llamado Voz IP, VozIP, VoIP (por sus siglas en inglés: Voice over Internet Protocol), es un grupo de recursos que hacen posible que la señal de voz viaje a través de Internet empleando un protocolo IP. Esto significa que se envía la señal de voz en forma digital, en paquetes, en

lugar de enviarla en forma analógica, a través de circuitos utilizables sólo para telefonía fija, como se ha comentado anteriormente.

La principal ventaja de este tipo de servicios es el ahorro de costes: se evita pagar precios elevados, principalmente en comunicaciones a larga distancia, usuales en la Red Pública Telefónica Conmutada (PSTN). Este ahorro parte del uso de una misma red para llevar voz y datos, especialmente cuando los usuarios tienen sin utilizar toda la capacidad de una red ya existente, la cual podría ser usada para VoIP sin un coste adicional. En contraposición, los problemas más importantes de la VoIP provienen de la pérdida de paquetes, que se puede traducir en pérdidas de tramas de voz.

El desarrollo de codecs para VoIP ha permitido que la voz se codifique en paquetes de datos de cada vez menor tamaño. Esto deriva en que las comunicaciones de voz sobre IP requieran anchos de banda muy reducidos. Los codificadores más utilizados son el G.729 y el G.723, y lo hacen en ley a. Junto con el avance permanente de las conexiones ADSL en el mercado residencial, este tipo de comunicaciones están siendo muy populares para llamadas internacionales (larga distancia), por ejemplo con el uso de Skype.

En cuanto a los experimentos realizados para probar esta tecnología en base al *VAD* de partida, se puede decir que, fueron llevados a cabo gracias a las siguientes bases de datos:

- TRAIN_IP: Se usa para entrenar los modelos de voz y de ruido para Voz IP. Procede de la emisión de locuciones a través de teléfonos IP con una posterior codificaficación/decodificación mediante el códec G.723.
- TEST_IP: Se usa para obtener tanto el error global de detección (GDE) como la tasa de falsas alarmas y la tasa de falsos rechazos. También procede de la emisión de locuciones a través de teléfonos IP con una posterior codificaficación/decodificación mediante el códec G.723.

Los resultados para voz IP sobre TEST_IP son los siguientes: se obtiene una tasa de falsas alarmas del 4.23%, una tasa de falsos rechazos del 32.05% y un error global de detección (GDE) final del 18.16% (Tabla 4.9). Por tanto, y teniendo en cuenta que la tasa de falsas alarmas es muy pequeña, resulta que prácticamente todos los clicks han sido eliminados por nuestro *VAD* de partida.

Base de datos	FAR	MR	*GDE*
TEST_IP	4.23%	32.05%	18.16%

Tabla 4.9. Resultados de detección para TEST_IP (Voz IP).

Se concluye que el *VAD* expuesto también es capaz de trabajar de forma adecuada con otro tipo de tecnología, con el único requisito de usar los modelos correspondientes (HMMs) para cada una de las tecnologías tratadas: GSM, telefonía fija y voz IP. Con esto se pretende decir que por ejemplo, los modelos de fija tendrían peores prestaciones si se usan para voz IP o GSM, y es recomendable que cada tecnología use sus modelos específicos.

Por otro lado queda demostrado que la topología óptima a usar por el *VAD* de partida es la que contiene cuatro modelos de voz y tres modelos de ruido, ya que topologías más complejas no mejoraban los resultados y sí que incrementaban el tiempo de retardo del *VAD*.

4.8.- Conclusiones.

En este capítulo se han realizado estudios para la creación de un *VAD* basado en HMMs robusto ante situaciones críticas en las que el locutor principal se puede encontrar en entornos tanto de ruido estacionario como de ruido no estacionario. Estos estudios comprenden desde la elección de las características que forman el vector de los HMMs hasta la obtención final de una pronunciación en forma de pulso de voz, pasando por el análisis de las distintas topologías de los HMMs o de la ventaja que supone usar el "*score*" para la toma de decisión a nivel de trama en comparación con otras características como la energía, ya sea normalizada o sin normalizar.

Para finalizar, es importante comentar que el *VAD* expuesto en este capítulo será llamado a partir de aquí *VAD* Base Mejorado. Así constituirá el punto de partida para la búsqueda e integración de nuevas característica con el fin de poner solución a la problemática de los ruidos no estacionarios (voces de fondo). Con ello se intentará reducir la tasa de falsas alarmas y el *GDE* final. Aún así remarcar que, el *VAD* expuesto en este capítulo supera con creces el comportamiento de otros detectores de referencia en todos los casos y que se intentará en la medida de lo posible mejorar su comportamiento en entornos de ruido no estacionario.

CAPÍTULO 5

ESTUDIO DE CARACTERÍSTICAS PARA EL RECHAZO DE VOCES DE FONDO

5.1.- Introducción.

En este capítulo se aborda el problema que tienen muchas aplicaciones basadas en reconocimiento automático del habla cuando otros locutores, estáticos o en movimiento, distintos del locutor principal, hablan a cierta distancia de él. Así pues, este problema de las voces de fondo (habla lejana) es especialmente importante en aplicaciones telefónicas basadas en reconocimiento de voz, sobre todo en aplicaciones de telefonía móvil: el locutor principal puede encontrarse en un entorno abierto o en una sala de reuniones, en ambos casos con la posible existencia de voces de fondo. El Detector de Actividad juega un papel muy importante en este sentido, ya que, es un sistema capaz de discriminar entre la ausencia (ruido, silencio) o presencia de voz para que un reconocedor automático de habla use esta información de forma adecuada. En la actualidad, los modelos de voz de los detectores de actividad tradicionales no pueden evitar reconocer estas voces de fondo como parte del diálogo hombre-máquina, situación que da lugar a un error de reconocimiento que hace que el sistema de diálogo falle. El objetivo es el de analizar nuevos parámetros con el fin de obtener más información que permita, en la medida de lo posible, rechazar esas voces de fondo.

En algunos trabajos previos, se han usado algunos parámetros acústicos para técnicas de dereverberación similares a las que se presentan en este capítulo. En [92], por ejemplo, los autores usan la idea de reverberación para reconstruir la voz degradada, debido a los rebotes del sonido producidos en una habitación, con la medida de dos micrófonos. Una técnica de dereverberación que usa el “pitch” como principal característica es [93]. Este método inicialmente estima el “pitch” y la estructura armónica de la señal de voz y obtiene un operador de dereverberación. Más tarde, el mencionado operador, basado en una operación de filtrado inverso, amplifica la señal. Por otro lado [94] propone un nuevo método de dereverberación con enventanado en la función de auto-correlación de ciertas tramas inteligentemente elegidas.

Bees en [95] muestra una técnica que reduce la reverberación en salas. Esta técnica consiste en una deconvolución cepstral compleja y en el comportamiento que posee la respuesta del impulso en una sala. Se usan filtros cuadráticos inversos para recuperar la voz resultando una reducción importarte de la reverberación.

Yegnanarayana propone en [96] un método para obtener el retardo entre dos señales de voz recogidas por dos micrófonos. El retardo se estima usando información espectral a corto plazo (amplitud, fase o ambas a la vez) ya que la reverberación y el ruido degradan la señal de la voz y las características espectrales se ven afectadas por este efecto. Es importante señalar que aunque este trabajo usa técnicas de audio multicanal, en los estudios y resultados obtenidos en este trabajo de Tesis se considera la limitación de usar un único canal.

En este capítulo se presenta el análisis, sobre la parte de desarrollo de la base de datos Av16.3 (DEV_AV), de distintas características o parámetros acústicos para clasificar habla de campo cercano, habla de campo lejano y habla simultánea de distintos locutores. Se trata, por tanto, de evaluar el poder de discriminación que poseen estas características para rechazar las voces de campo lejano (procedentes de uno o varios locutores) [98]. Las principales características analizadas son: la distancia de Mahalanobis entre los MFCCs de tramas de voz consecutivas, la armonicidad, y diversas medidas sobre un LPC residual de orden 10. Estas nuevas características se pueden clasificar en tres grupos diferentes:

- Características de estructura armónica: la auto-correlación a nivel de trama a partir del rango de frecuencias usado durante el cálculo del “pitch”. Esta característica trata de detectar cómo desaparece la estructura armónica en el caso de voces de campo lejano (como consecuencia de la reverberación) o voces procedentes de varios locutores.
- Características de envolvente espectral: la distancia de Mahalanobis entre los MFCCs de tramas consecutivas. Esta característica se usa para medir cómo de rápido cambia el espectro. Se supone que en caso de voz reverberante o voz procedente de varios locutores debería de ser más rápido que en el caso de la voz de campo cercano.
- Características mixtas: LPC residual de orden 10. Esta característica contiene tanto información armónica como de la envolvente espectral.

A continuación, en los siguientes apartados, se muestra el estudio de cada una de las características anteriormente mencionadas.

5.2.- Características de estructura armónica. Armonicidad.

En este apartado se estudia el uso del máximo de la función de auto-correlación $R(k)$ normalizada a $R(0)$ en cada trama de voz sonora como característica para la discriminación entre voz procedente de un locutor principal y voz procedente de un locutor de habla lejana o de varios locutores. Este máximo se calcula entre dos extremos, k_{min} y k_{max}, que se corresponden a un intervalo de frecuencia de f_{max}=320 Hz a f_{min}=50 Hz respectivamente. Es importante comentar que este análisis pretende obtener resultados en la etapa de decisión a nivel de pulso, y, por tanto, el estudio se realiza sobre varias tramas consecutivas para generar información adicional sobre los pulsos de voz obtenidos tras aplicar las tres reglas enunciadas en el apartado 4.2.

Los análisis se llevan a cabo con la base de datos Av16.3 y se obtienen los errores de clasificación teniendo en cuenta un futuro funcionamiento en tiempo real.

5.2.1.- Estudio de la capacidad de discriminación del máximo de la función de auto-correlación.

En primer lugar es importante decir que el mencionado valor máximo sólo se calcula sobre las tramas de voz. Las tramas de voz se clasifican por medio del *VAD*, que se describe en el tema anterior, en forma de pulsos de voz. Posteriormente, para asegurar una buena segmentación, las marcas de los pulsos de voz obtenidas de forma automática por el mencionado *VAD*. Con esto ya se tenían seleccionadas todas las tramas de voz. Más tarde, sobre esas tramas de voz, se calcula el máximo de la función de auto-correlación $R(k)$ normalizada a $R(0)$ (ec. 5.1), cada trama con 256 muestras, entre el intervalo de frecuencias f_{min}=50 Hz y f_{max}=320 Hz. Se asume que el "pitch" está contenido siempre en ese intervalo de frecuencias, por lo que se espera encontrar el valor de auto-correlación más alto en ese intervalo. Además, sólo se han tenido en cuenta las tramas cuyo valor de auto-correlación normalizado a $R(0)$ es mayor que 0.5.

$$\max\_autocorr(i) \quad = \quad \max_i \left\{ \frac{R_i(k)}{R_i(0)} \right\} \quad / \quad 50Hz \le f = \frac{f_s}{k} \le 320Hz \qquad (5.1)$$

En ec. 5.1 "*i*" denota la posición del mínimo en el número de trama, "*f*" es la frecuencia en hercios cercana a la frecuencia del "pitch" y "f_s" es la frecuencia de muestreo.

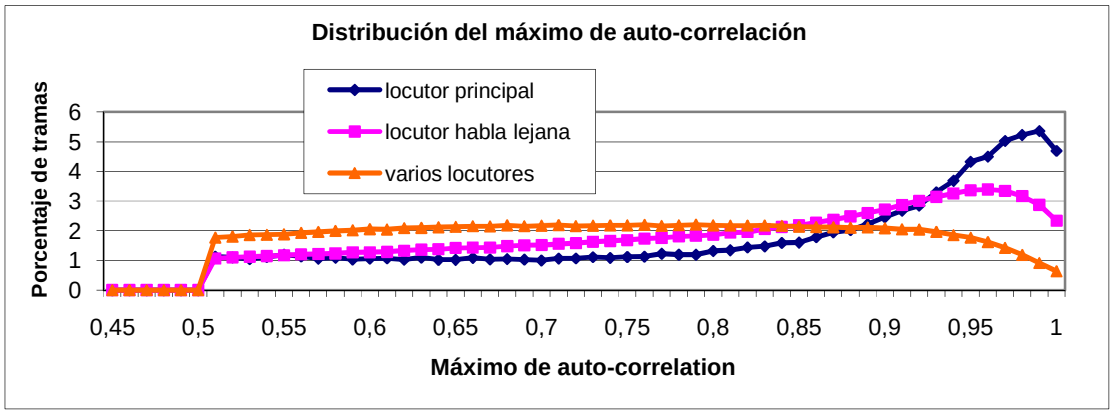

Figura 5.1. Distribución del máximo de auto-correlación para un locutor principal, un locutor de habla lejana y varios locutores.

La Fig.5.1 presenta la distribución (histograma) de los valores máximos de la función de auto-correlación para las tramas de voz que superan el valor 0.5 para un locutor principal, un locutor de habla lejana y varios locutores. Los comportamientos del máximo valor de la función de auto-correlación son muy diferentes en los tres casos a comparar, especialmente mayor frecuencia de valores de auto-correlación máxima superior a 0.9. Hay más tramas en el caso de voz procedente de un locutor principal y muy pocas en el caso de voz procedente de varios locutores. Este efecto es debido a la reverberación para el caso de un locutor de voz lejana y a la mezcla de señales con reverberación para el caso de varios locutores de voz lejana. En el caso de voz lejana procedente de un locutor, suponiendo que las señales de error en amplitud y fase provocan un efecto de reverberación, una señal de campo lejano $S_n^{far-field}(t)$ puede ser escrita como la suma de la señal limpia sin reverberación $S_n^{near-field}(t)$ y el error producido por la mencionada reverberación $S_n^{error}(t)$.

Por otro lado, para el caso de voz procedente de varios locutores de habla lejana, la señal de audio se puede considerar como la generada por varias señales reverberantes (ec. 5.2).

$$S_n^{multispea\,ker-far-field}(t) = \sum_{spea\,ker-i} S_n^{far-field(i)}(t) = \sum_{spea\,ker-i} S_n^{near-field(i)} + \sum_{spea\,ker-i} S_n^{error(i)} \qquad (5.2)$$

Tras considerar este efecto, se analiza un segmento de N tramas (para filtrar un pulso de voz de N tramas): se calcula para los tres casos el porcentaje de tramas, en un pulso de voz de N tramas, con un valor máximo de auto-correlación superior a 0,9. Este estudio se lleva a cabo seleccionando pulsos consecutivos de voz de N tramas, y sin solape, sobre el conjunto total de tramas de voz obtenidas en el proceso de decisión a nivel de trama sobre la base de datos DEV_AV.

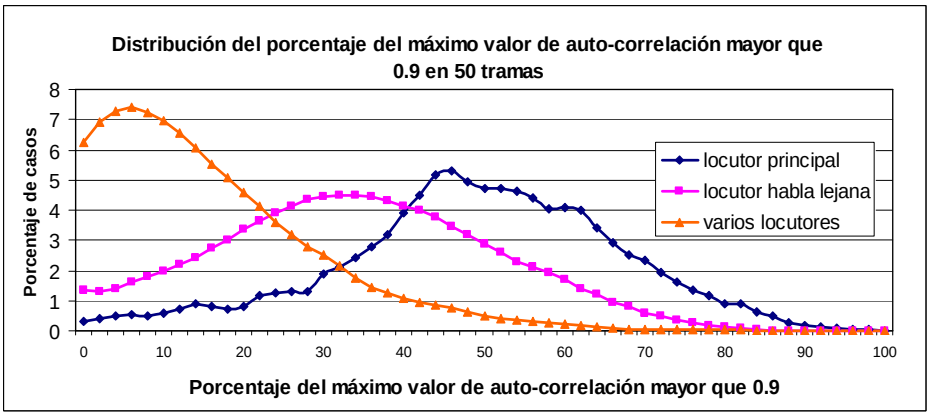

Figura 5.2. Distribución del porcentaje de tramas con un valor máximo de auto-correlación mayor del 90% (0.9) para pulsos de 50 tramas.

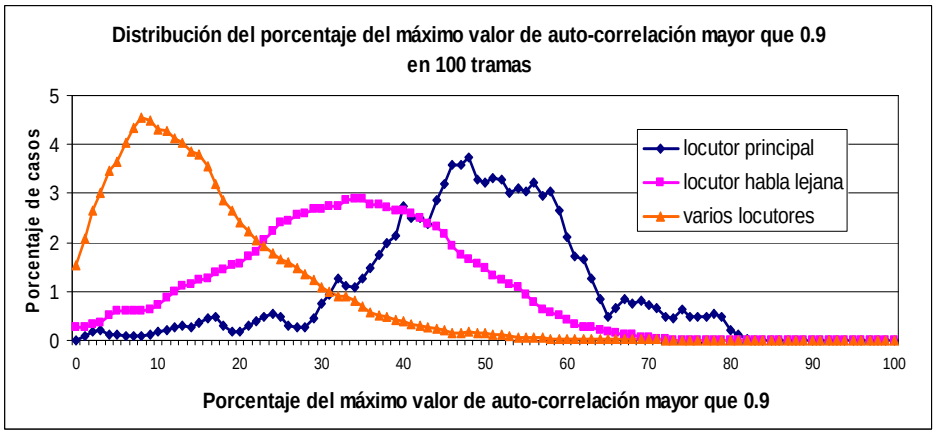

Figura 5.3. Distribución del porcentaje de tramas con un valor máximo de auto-correlación mayor del 90% (0.9) para pulsos de 100 tramas.

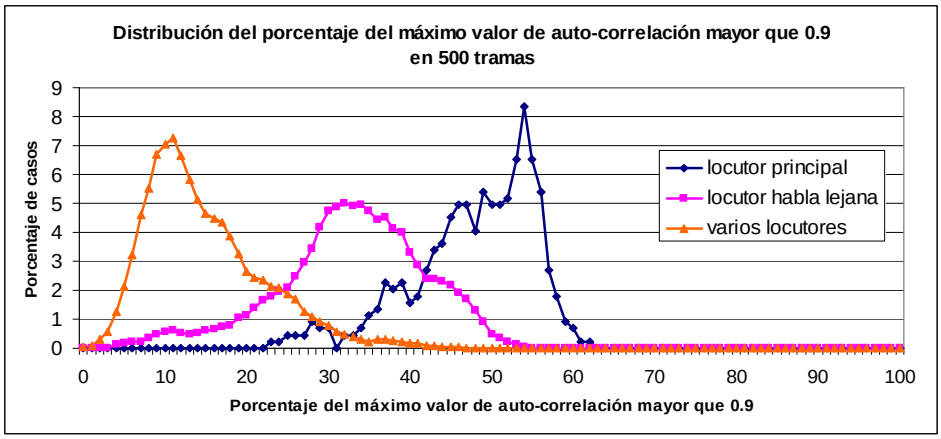

Figura 5.4. Distribución del porcentaje de tramas con un valor máximo de auto-correlación mayor del 90% (0.9) para pulsos de 500 tramas.

En Fig.5.2-5.4 se presentan las distribuciones del porcentaje de tramas con valores del máximo de auto-correlación mayores de 0,9 para la voz de un locutor principal, la de un locutor de habla lejana y la de varios locutores para pulsos de voz de 50, 100 y 500 tramas respectivamente. Como se puede apreciar, el porcentaje sobre pulsos de voz de N tramas es menor para el caso de habla de varios locutores. Este parámetro acústico discrimina muy bien entre la voz procedente del locutor principal y la que procede de varios locutores: el error de clasificación es menor del 15%, del 10% y del 3,5% para pulsos de 50, 100 y 500 tramas respectivamente. Los errores de clasificación se calculan como el área de

intersección entre las dos distribuciones a comparar, dividido por dos (a partir de aquí siempre se adoptará este criterio). Conforme se aumenta el número de tramas considerado, mejor es el resultado y el poder de discriminación aumenta. El poder de discriminación entre la voz de un locutor principal y la de un locutor de campo lejano no es tan bueno como el de un locutor principal y varios locutores de habla lejana. Para este caso se obtiene un error de clasificación menor del 33,5%, 29% y 19% considerando pulsos de 50, 100 y 500 tramas respectivamente.

En el siguiente punto se extiende el estudio a otras medidas dentro del cálculo del máximo de auto-correlación mayor que 0.9.

5.2.2.- Medidas sobre el máximo de la función de auto-correlación.

Además del porcentaje de tramas con valor del máximo de la función de auto-correlación mayor que 0.9 (ec. 5.3), también se han estudiado las siguientes medidas: media (ec. 5.4), varianza (ec. 5.5) y kurtosis (ec. 5.6).

$$Perc(N) = \frac{1}{N} n^{o}\_times\{\max\_autocorr(i) \geq 0.9\}_{i=1}^{N} \quad (5.3)$$

$$aver\{\max\_autocorr(i)\}_{i=1}^{N} \quad (5.4)$$

$$\mathrm{var}\{\max\_autocorr(i)\}_{i=1}^{N} \quad (5.5)$$

$$kurt\{\max\_autocorr(i)\}_{i=1}^{N} \quad (5.6)$$

Teniendo en cuenta estas medidas, se representa gráficamente cómo varían los errores de clasificación comparando un locutor principal con un locutor de habla lejana (Fig.5.5) y un locutor principal con varios locutores de habla lejana (Fig.5.6) donde se varía el número de tramas que forman los pulsos de voz (10, 50, 100 y 500). Como se puede observar en ambas figuras, los errores de clasificación siguen la misma tendencia, aunque en el caso de un locutor de habla cercana y varios locutores de habla lejana las diferencias son mayores y los resultados mejores (errores de clasificación más pequeños):

- Media, varianza y kurtosis: en los tres casos los errores siguen la misma tendencia aunque la media obtiene mejores resultados de clasificación que las otras dos medidas. El error de clasificación disminuye cuando el número

de tramas por pulso de voz (N) aumenta, así que las medidas funcionan mejor cuanto mayor sea el número de tramas del pulso considerado.

- Porcentaje de tramas con un valor de la función de auto-correlación mayor que 0.9: al igual que con las medidas anteriores, el error de clasificación disminuye cuando el número de tramas por pulso de voz (N) aumenta. Esta medida resulta ser la que mejor resultados de clasificación obtiene.

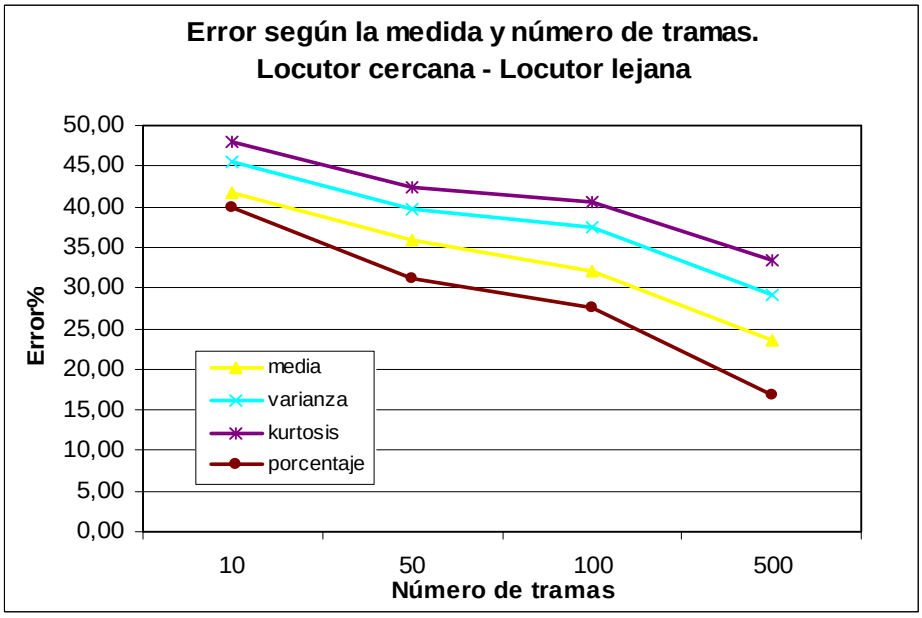

Figura 5.5. Errores de clasificación para un locutor de habla cercana y un locutor de habla lejana.

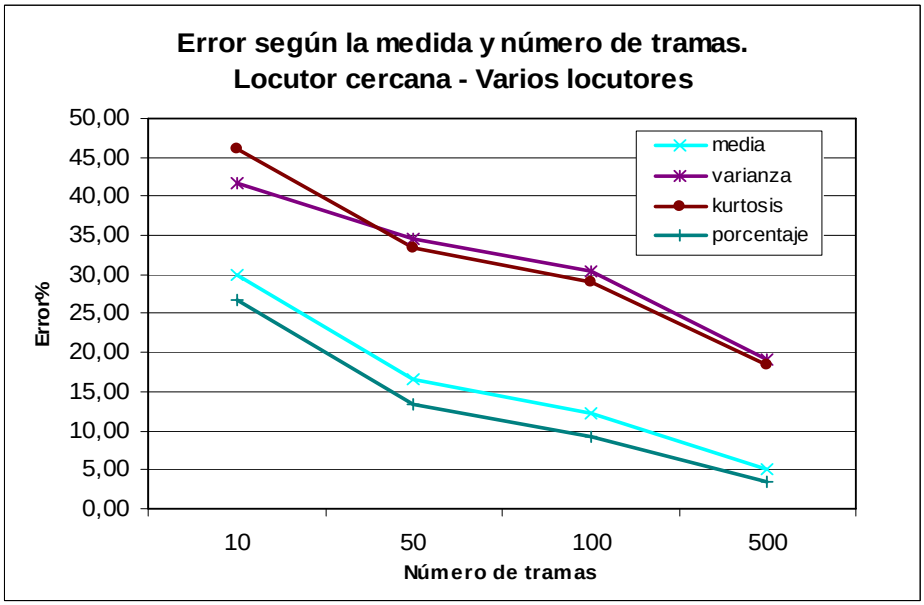

Figura 5.6. Errores de clasificación para un locutor de habla cercana y varios locutores de habla lejana.

Por tanto, la medida que mayor poder de discriminación presenta es el porcentaje de tramas con un valor de la función de auto-correlación mayor que 0.9. A continuación se procede a extender el estudio a valores entorno a 0.9 considerando ahora este valor como un umbral (*THR* en ec. 5.7).

$$Perc(N) = \frac{1}{N} n^{o}\_times\{\max\_ autocorr(i) \geq THR\}_{i=1}^{N} \qquad (5.7)$$

Se obtuvieron los errores de clasificación haciendo variar *THR* desde 0.8 a 1.0 (80% y 100% respectivamente en Fig.5.7 y Fig.5.8) considerando de nuevo pulsos de voz con un número de tramas de 10, 50, 100 y 500.

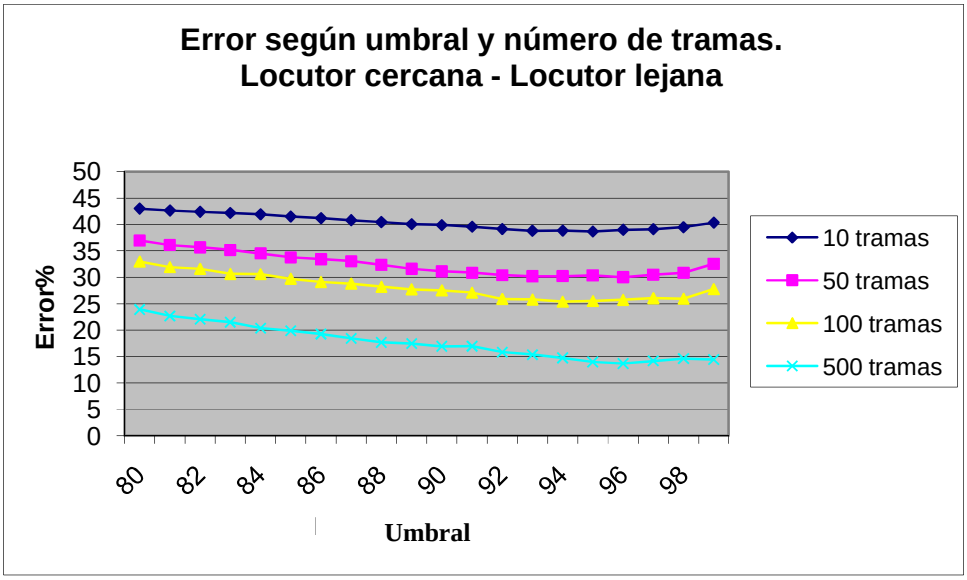

Figura 5.7. Errores de clasificación para el porcentaje de tramas con un valor de la función de auto-correlación mayor que un umbral THR (Loc. cercana – Loc. lejana).

Para los dos casos a comparar (Fig.5.7 y Fig.5.8), el valor de THR que hace mínimo los errores de clasificación se sitúa en torno a 0.96. Para este valor se obtienen errores de clasificación menores del 15% para un locutor de habla cercana y un locutor de habla lejana y del 3% para un locutor de habla cercana y varios locutores de habla lejana. Si nos centramos en el caso de N=50, un compromiso bastante razonable entre "enventanado de número de tramas consideradas no muy grande (válido para aplicaciones en tiempo real)" y un "error de clasificación aceptable", se puede observar que los errores de clasificación permanecen prácticamente constantes para valores de THR entre 0.90 y 0.97 (THR=0.9 por tanto es una buena aproximación, como ya se percibía a simple vista en la Fig.5.1).

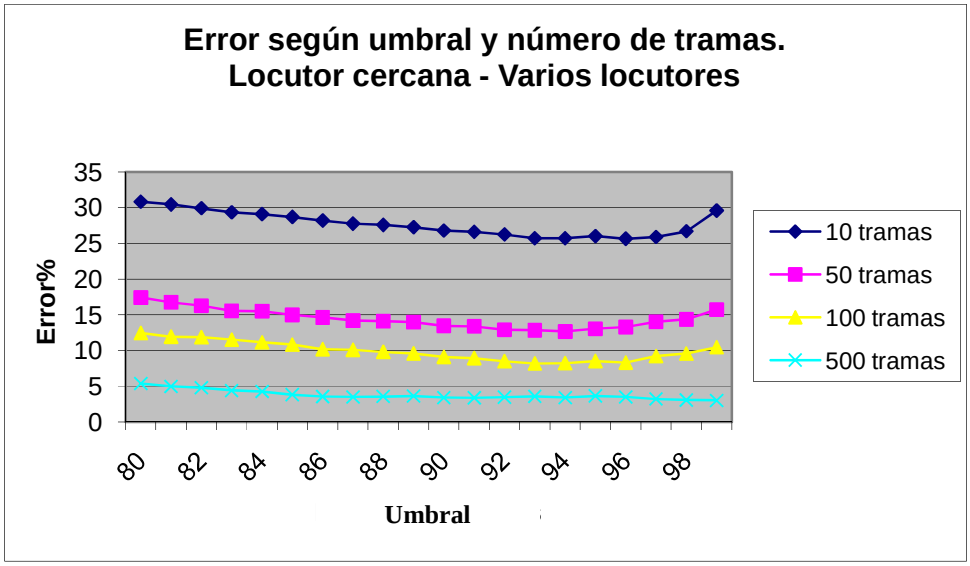

Figura 5.8. Errores de clasificación para el porcentaje de tramas con un valor de la función de auto-correlación mayor que un umbral THR (Loc. cercana – Varios loc.).

Otra forma de representación, y ya centrada en un enventanado con un número de tramas de 50 (pulsos de voz de 50 tramas) orientado para el funcionamiento de esta medida en tiempo real, se representan las curvas DET (Fig.5.9) variando de nuevo el umbral THR de igual manera que en las Fig.5.7 y Fig.5.8 pero particularizado a N = 50. Como se puede observar en la Fig.5.9, se pueden distinguir perfectamente dos tipos o agrupaciones de curvas DET, unas para un locutor de habla cercana vs. un locutor de habla lejana, con mayor error de clasificación, y las otras para un locutor de habla cercana vs. varios locutores de habla lejana. En cada agrupación se pueden distinguir diferentes líneas: cada una de ellas corresponde a un valor del umbral THR del mencionado barrido de valores. En ambos casos, “locutor cercana – locutor lejana” y “locutor cercana – varios locutores lejana”, se corroboran los resultados con los de la Fig.5.7 y los de la Fig.5.8, los errores de clasificación más pequeños se obtienen tanto para falsas alarmas como falsos rechazos para un valor de THR de 0.96:

Comparando un locutor de habla cercana con un locutor de habla lejana: se obtiene un EER (Equal Error Rate) del 20% aproximadamente. Además, las curvas se mantienen prácticamente simétricas respecto de la línea de EER, así que, el error total se mantiene aunque se eleve el error de tasa de falsas alarmas (se decrementa en la misma medida el error de falsos rechazos) y viceversa.

Comparando un locutor de habla cercana con varios locutores de habla lejana: se obtiene un EER del 12% aproximadamente. En este caso las curvas DET no son simétricas y la tasa de falsas alarmas tiene un peor comportamiento si se quiere tener una tasa de falsos rechazos pequeña. La forma escalonada de las curvas DET se debe a que el número de ejemplos es limitado: número de pulsos de voz de 50 tramas que se pueden formar en la base de datos.

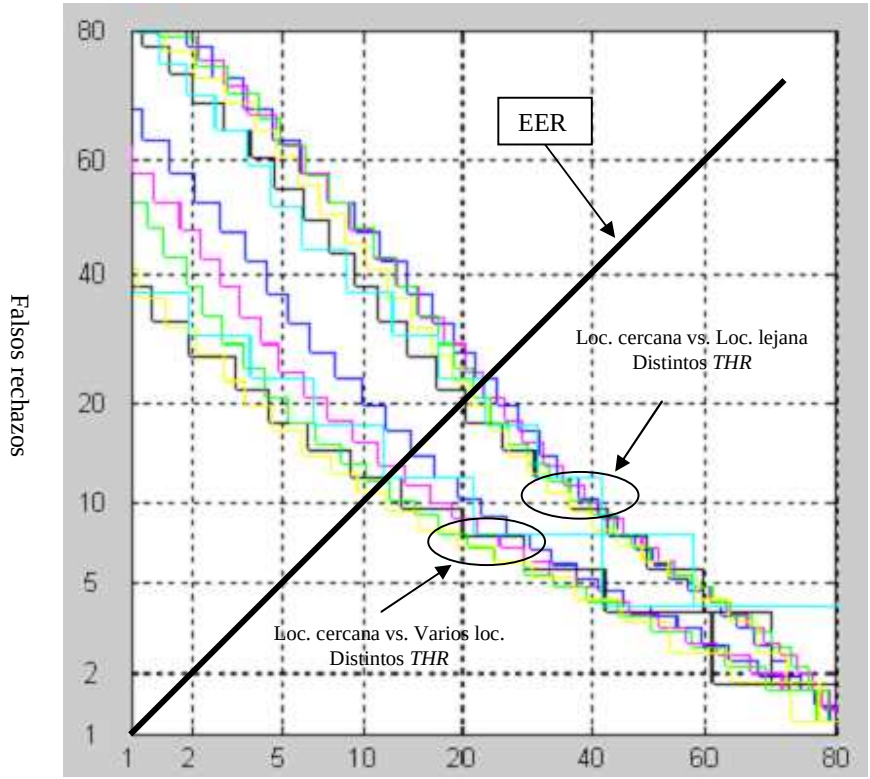

Figura 5.9. Curvas DET para loc. cercana vs. loc. lejana y loc. cercana vs. varios loc. para el porcentaje de tramas con un valor de la función de auto-correlación mayor que un umbral THR y con un enventanado de N = 50 tramas.

5.2.3- Análisis de algunos ejemplos ilustrativos de Av16.3.

En este subapartado se va a realizar un estudio de la función de auto-correlación a nivel de fichero de audio, aprovechando que la base de datos Av16.3 tiene las mismas locuciones de audio grabadas en los micrófonos de solapa (habla cercana) y en los arrays de micrófonos (habla lejana).

Ejemplo 1. Comparación habla cercana – habla lejana

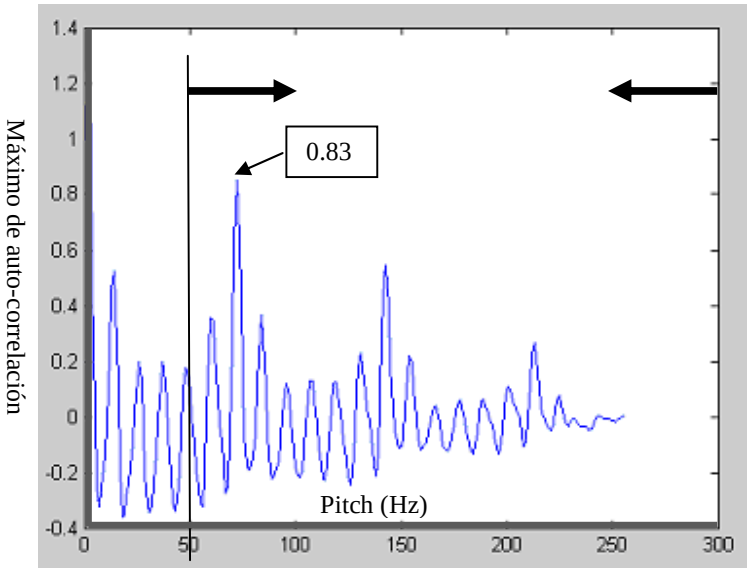

Figura 5.10. seq01-1p-0000_lapel1. Entre 50 Hz y 320 Hz.

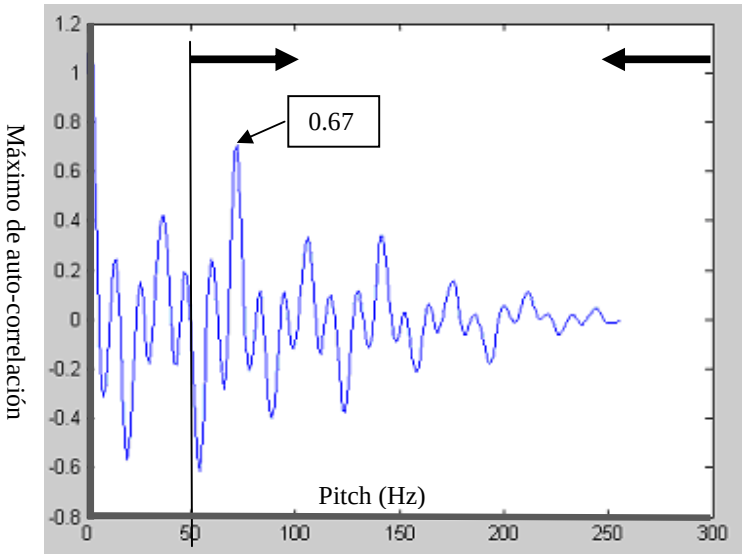

Figura 5.11. seq01-1p-0000_array1_mic1. Entre 50 Hz y 320 Hz.

Tanto la Fig.5.10 como la Fig.5.11 representan el valor de la auto-correlación en una zona sonora de una locución concreta en la que habla un solo locutor. Simplemente se diferencian en que en la primera la grabación se realiza desde un micrófono de solapa (habla cercana procedente de un locutor), la del locutor que emite la locución, mientras que en la segunda, la grabación la realiza otro micrófono (del

array) a cierta distancia del locutor principal (habla lejana procedente de un locutor). En ambos casos el pitch se encuentra en unos 75 Hz, sin embargo, el valor de auto-correlación es mayor para el caso del micrófono que graba desde muy cerca (habla cercana). Por tanto, parece que, cuando tenemos un locutor de habla lejana, efectivamente la reverberación hace que la auto-correlación de la señal se parezca más a un ruido, y la amplitud del máximo en un entorno del pitch se atenúa. En otros ejemplos estudiados el comportamiento es similar: el valor del máximo de auto-correlación es más alto en el caso de un locutor de habla cercana que el de un locutor de habla lejana.

Ejemplo 2. Comparación habla cercana – varios locutores de habla lejana

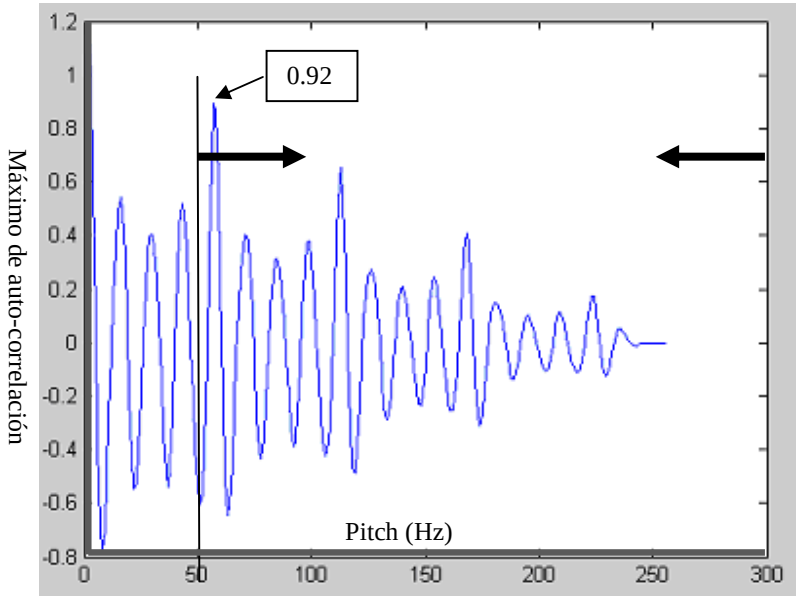

Figura 5.12. seq39-3p-0111_lapel1. Entre 50 Hz y 320 Hz.

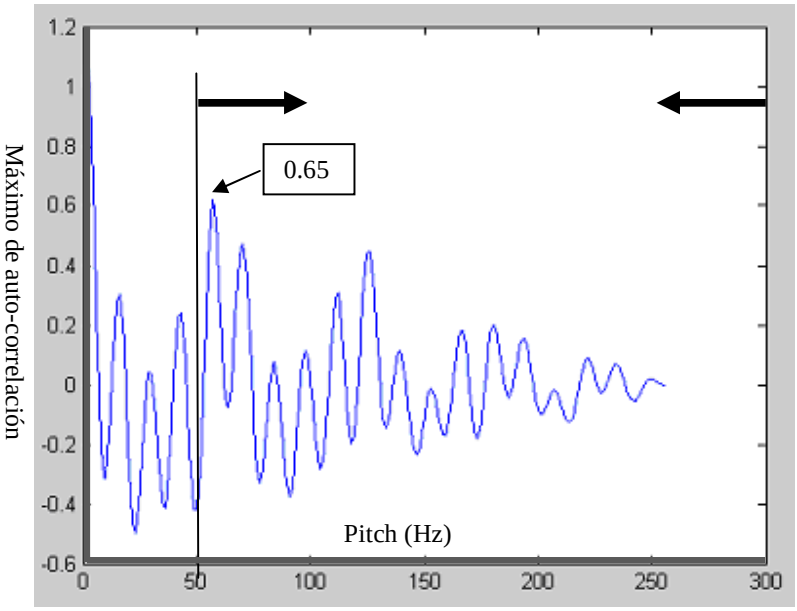

Figura 5.13. seq39-3p-0111_array2_mic1. Entre 50 Hz y 320 Hz.

Tanto la Fig.5.12 como la Fig.5.13 representan el valor de la auto-correlación en una zona sonora de varias locuciones concretas simultáneas procedentes de distintos locutores: en este caso se tuvo en cuenta la zona sonora de uno de los locutores, el resto no se encontraban en una zona sonora. Los dos casos se diferencian en que en la Fig.5.12 la grabación se realiza desde un micrófono de solapa (habla cercana procedente de un locutor), la del locutor que emite la locución en zona sonora, mientras que en la segunda, la grabación la realiza otro micrófono (del array) a cierta distancia de los locutores que hablan de forma simultánea (habla lejana procedente de varios locutores). En ambos casos el pitch se encuentra en unos 60 Hz, sin embargo, el valor de auto-correlación es mayor para el caso del micrófono que graba desde muy cerca (habla cercana), incluso con más diferencias que en el caso del ejemplo 1. Cuando tenemos varios locutores de habla lejana, efectivamente la reverberación hace que la auto-correlación de la señal se parezca más a un ruido, y la amplitud del máximo en un entorno del pitch se atenúa mucho. Al igual que en el ejemplo anterior, otros casos han sido estudiados y el comportamiento ha sido similar: el valor del máximo de auto-correlación es más alto en el caso de un locutor de habla cercana que en el caso de varios locutores de habla lejana que hablan simultáneamente.

5.3.- Características de envolvente espectral. Distancia de Mahalanobis entre coeficientes MFCC.

Esta característica se basa en el cálculo de la distancia de Mahalanobis entre vectores de componentes MFCC, a partir de un banco de 12 filtros Mel con un filtrado de pre-énfasis previo para suavizar la señal, y calculados en tramas de voz consecutivas. Cada uno de estos vectores está formado por los primeros 8 MFCC, el logaritmo de la energía normalizada y la delta del logaritmo de la energía. La distancia de Mahalanobis se usa para determinar la similitud entre variables multidimensionales aleatorias (5.9).

$$d_M(\vec{x}_i;\vec{x}_j)=\sqrt{(\vec{x}_i-\vec{x}_j)S^{-1}(\vec{x}_i-\vec{x}_j)^t} \qquad (5.9)$$

En ec. 5.9 S denota la matriz de covarianza del vector variable $(x_1,x_2,...,x_k)$.

Las distribuciones de la distancia de Mahalanobis entre tramas consecutivas para un locutor principal, un locutor de voz lejana y varios locutores se presentan en la Fig.5.14. Como se puede ver, la voz del locutor principal presenta la menor distancia, mientras que la mayor es para el caso de voz procedente de varios locutores. Es importante comentar que la distancia se normaliza al máximo valor de distancia de Mahalanobis obtenido, considerando todos los valores calculados en cada trama. Este máximo valor de distancia representaría el 100 en la Fig. 5.14. En general, a partir de ahora, cuando se hable de normalización (siempre en el eje de abscisas) realmente se realiza entre 0 a 1, y posteriormente se multiplica por 100 para conseguir una visualización gráfica más cómoda y entendible. En estos casos, el valor mínimo, $V_{\min}$, correspondería al 0 en la nueva escala, por tanto, se realiza un desplazamiento por valor igual a ese mínimo, mientras que el valor máximo, $V_{\max}$, correspondería al 1. Matemáticamente, la transformación de un valor real, $V(i)$, a su valor normalizado, $V_n(i)$, quedaría de la siguiente manera:

$$V_n(i)=\frac{V(i)-V_{\min}}{V_{\max}-V_{\min}}*100 \qquad (5.10)$$

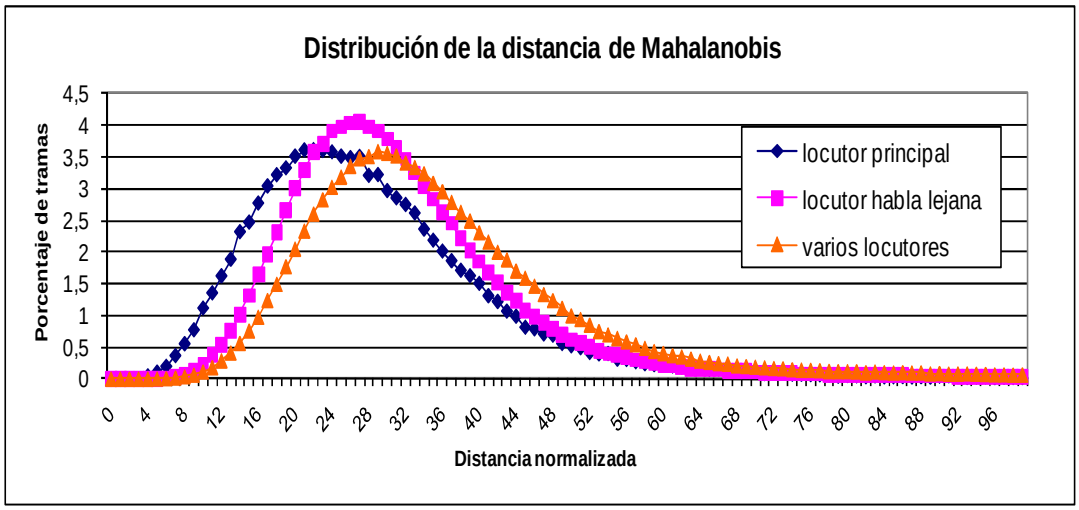

Figura 5.14. Distribución de la distancia de Mahalanobis normalizada (de 0 a 100) para un locutor principal, un locutor de habla lejana y varios locutores.

Teniendo en cuenta el análisis anterior, se considera el análisis para pulsos de voz de un número de N tramas (N = 50, N = 100 y N = 500 tramas) calculando la distancia mínima (de Mahalanobis) en las mencionadas N tramas. Las figuras Fig.5.15-5.17 muestran las distribuciones de la mínima distancia para los tres casos: locutor principal, locutor de habla lejana y varios locutores. Además, la distancia se normaliza en el eje X,

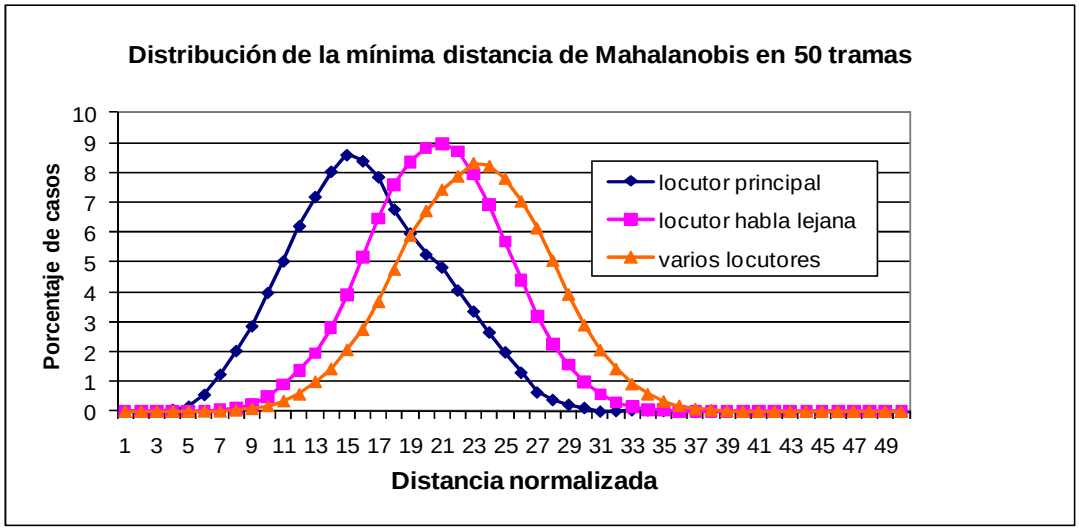

Figura 5.15. Distribución de la mínima distancia de Mahalanobis normalizada (de 0 a 100) para pulsos de 50 tramas (N = 50).

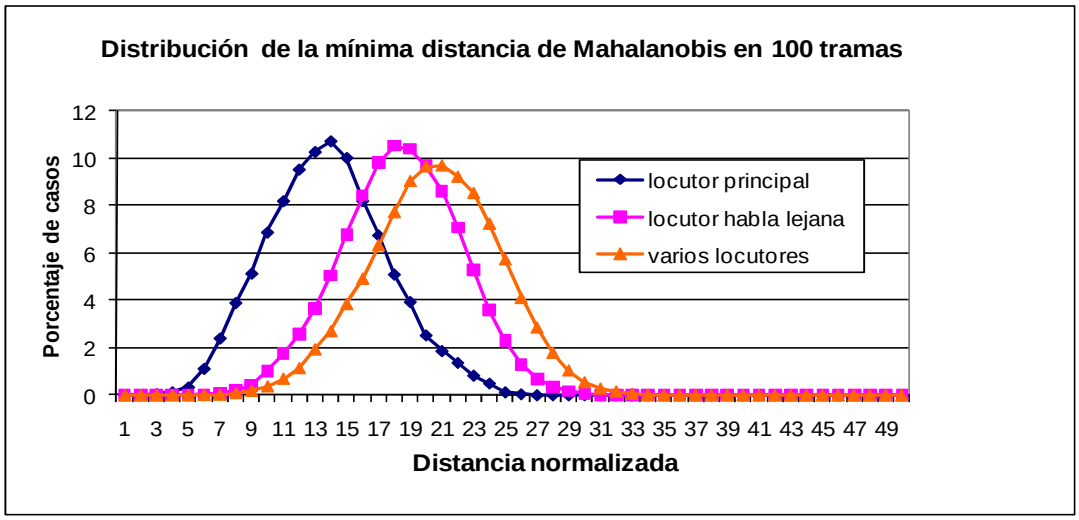

Figura 5.16. Distribución de la mínima distancia de Mahalanobis normalizada (de 0 a 100) para pulsos de 100 tramas (N = 100).

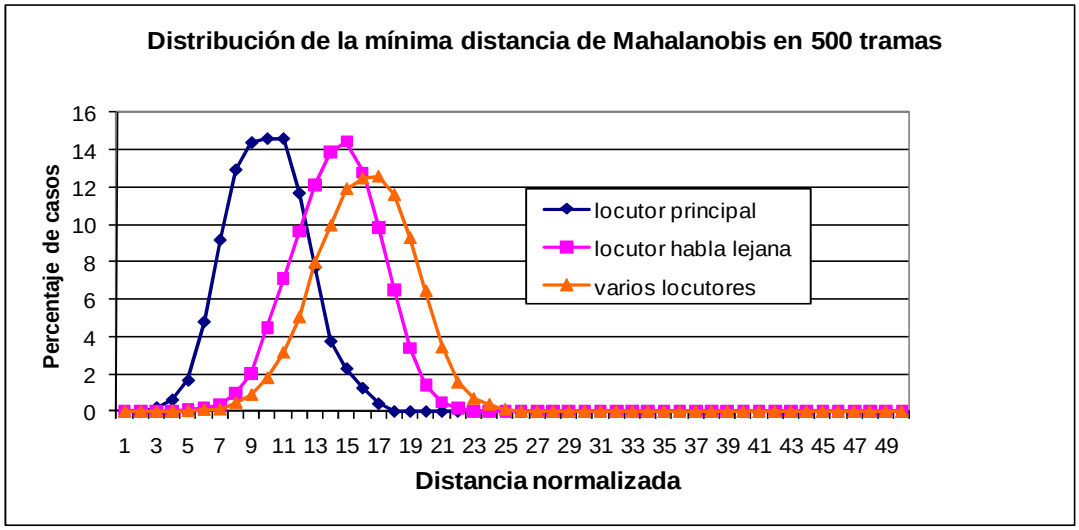

Figura 5.17. Distribución de la mínima distancia de Mahalanobis normalizada (de 0 a 100) para pulsos de 500 tramas (N = 500).

Como se puede observar en las figuras Fig.5.15-5.17, la mínima distancia en pulsos de voz a partir de N tramas es mayor para el habla que procede de varios locutores. Esta característica puede también discriminar muy bien entre la voz de un locutor principal y las voces procedentes de varios locutores. Para ese caso, el error de clasificación es menor del 24%, 20% y 14% para pulsos de voz de 50, 100 y 500 tramas respectivamente. Además, cuando aumenta en número de tramas por pulso consideradas desde 50 hasta 500 para el cálculo de esta mínima distancia, los

resultados son mejores y el poder de discriminación aumenta. Por otro lado, el poder de discriminación entre la voz de un locutor principal y la voz de un locutor de campo lejano es algo peor que para el caso de varios locutores que hablan simultáneamente. En este caso, el error de clasificación es menor del 35%, del 32% y del 27% para pulsos de 50, 100 y 500 tramas respectivamente.

5.4.- Características mixtas. LPC Residual de orden 10.

Las últimas medidas sobre el estudio que se lleva a cabo en este capítulo de Tesis están relacionadas con el residuo procedente de un LPC residual de orden 10. Se trata de una característica mixta, contiene tanto información de estructura armónica como de la envolvente espectral. Como en el resto de casos anteriores, el residuo sólo se calcula sobre las tramas de los pulsos de voz (evaluación de las pronunciaciones).

La predicción lineal calcula una serie de coeficientes que se van reestimando (o prediciendo) para la muestra venidera $y'[n]$ (ec. 5.11) dado un vector de muestras de entrada $x[n]$:

$$y'[n] = e[n] + \sum_{k=0}^{p} a_k x[n-k] \qquad (5.11)$$

donde a son los coeficientes de predicción y p el orden del residuo (en nuestro caso 10). El error residual se presenta en ec. 5.12.

$$e[n] = y'[n] - y[n] \qquad (5.12)$$

El residuo o error residual es la base para todos los cálculos realizados en esta sección. Aunque se han considerado muchas medidas sobre el residuo en este estudio, sólo la kurtosis y la auto-correlación obtienen resultados interesantes dependiendo del tipo de pronunciación considerada. La kurtosis mide la calidad de la voz [109]: concentración de la distribución entorno a los formantes. En el caso de mezclas de voces o voz mezclada con ruido ocurre que el modelo excitación-filtro no

representa de forma apropiada a la señal de audio haciendo que la distribución del error de predicción sea más próxima a una distribución gaussiana (kurtosis = 3), y por tanto los valores de kurtosis sean más pequeños que para señales de un único locutor (con distribuciones más abruptas). En cuanto a la auto-correlación del residuo, se puede decir que, el comportamiento debería ser bastante parecido al de la auto-correlación de la señal ya que al calcular del residuo únicamente desaparecería la envolvente espectral.

Los resultados muestran diferencias importantes en la distribución de la kurtosis para un locutor de habla cercana, un locutor de habla lejana y varios locutores, sobre todo para valores de kurtosis mayores que 5.

$$Perc(N) = \frac{1}{N} n^{o}\_times\{kurt\_residuo(i) \geq kurt\_th\}_{i=1}^{N} \qquad (5.13)$$

Matemáticamente (ec. 5.13) el mínimo solapamiento para las distribuciones se obtiene para un valor de *kurt_th* igual a 5. Teniendo esto en cuenta, se calcula el porcentaje de tramas de voz, sobre pulsos de voz consecutivos de N tramas (al igual que con las características anteriores), cuyos valores de la kurtosis son mayores que 5. En la Fig.5.18 se muestra la distribución del porcentaje con valores de kurtosis del residuo mayores que 5 para la voz de un locutor principal, la voz de un locutor de voz lejana y voces de varios locutores que hablan simultáneamente para pulsos de voz de 50 tramas.

Como se puede ver en la Fig.5.18, el porcentaje es menor para las voces procedentes de varios locutores que hablan a la vez. Esta característica discrimina muy bien entre el habla del locutor principal tanto con voces de varios locutores como el habla de un locutor de campo lejano. Para ambos casos, el error de clasificación ronda el 18% considerando pulsos de voz de 50 tramas, así que el poder de discriminación entre la voz de un locutor principal y la de un locutor de habla lejana es el mejor si se compara con el resultado de las medidas anteriormente expuestas, aunque un poco peor para el caso del habla de un locutor principal y la de varios locutores si se compara con el resultado con la medida del porcentaje de tramas por pulso de voz con una valor de auto-correlación mayor que 0.9.

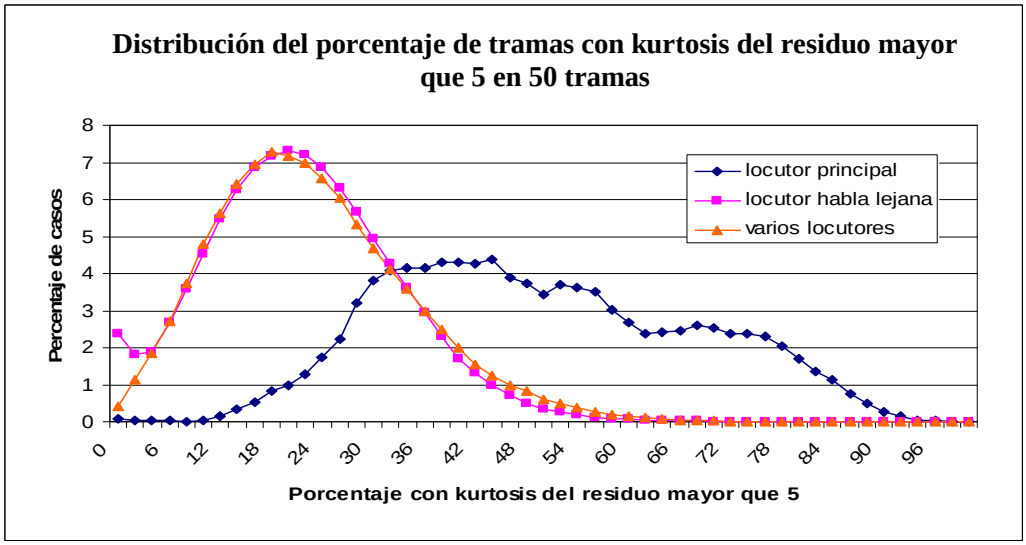

Figura 5.18. Distribución del porcentaje de tramas con kurtosis del residuo mayor que 5 para pulsos de 50 tramas (N = 50).

La otra medida interesante también basada en el LPC residual es el máximo de auto-correlación del residuo en las tramas de voz. En este caso, las medidas con las que se tenía el mayor poder de discriminación fueron las siguientes:

- El porcentaje de tramas (N) en cada pulso de voz que cumplen que el máximo de la auto-correlación del residuo (por trama) es mayor que un umbral de auto-correlación *res_autocorr_th* (ec. 5.14). En este caso, la mejor clasificación sobre los datos se obtiene para un valor de *res_autocorr_th* igual a 0.425. A partir de ahora *LPC_Autocorr_estadístico_1*.

$$Perc(N) = \frac{1}{N} n^{o}\_times\{\max\_autocorr\_residuo(i) \geq res\_autocorr\_th\}_{i=1}^{N} \quad (5.14)$$

- El máximo valor (ec. 5.15) en cada pulso de voz sobre el máximo de la auto-correlación del residuo (por trama). A partir de ahora *LPC_Autocorr_estadístico_2*.

$$\max\{\max\_autocorr\_residuo(i)\}_{i=1}^{N} \quad (5.15)$$

- La media (ec. 5.16) en cada pulso de voz calculada sobre el máximo de la auto-correlación del residuo (por trama). A partir de ahora *LPC_Autocorr_estadístico_3*.

$$aver\{\max\_autocorr\_residuo(i)\}_{i=1}^{N} \quad (5.16)$$

- La varianza (ec. 5.17) en cada pulso de voz calculada sobre el máximo de la auto-correlación del residuo (por trama). A partir de ahora *LPC_Autocorr_estadístico_4*.

$$\mathrm{var}\left\{\max\_\ autocorr\ \_residuo\ (i)\right\}_{i=1}^{N} \tag{5.17}$$

En Fig.5.19-5.22 se muestran las distribuciones para las 4 medidas sobre el máximo de auto-correlación del residuo anteriormente expuestas considerando pulsos de 50 tramas. En estos 4 casos los errores de clasificación resultan ser bastante distintos si se compara la voz de un locutor principal con un locutor de habla lejana o con varios locutores de hablan simultáneamente. Se muestran los resultados en la Tabla 5.1.

Medida	*Error de clasificación. Loc. cercana - Loc. lejana*	*Error de clasificación. Loc. cercana – Varios loc. lejana*
LPC_Autocorr_estadístico_1	**26.8%**	**18.8%**
LPC_Autocorr_estadístico_2	**29.2%**	**19.6%**
LPC_Autocorr_estadístico_3	**26.2%**	**19.4%**
LPC_Autocorr_estadístico_4	**28.1%**	**14.9%**

Tabla 5.1. Errores de clasificación para 4 estadísticos sobre la auto-correlación del LPC de orden 10 para pulsos de 50 tramas.

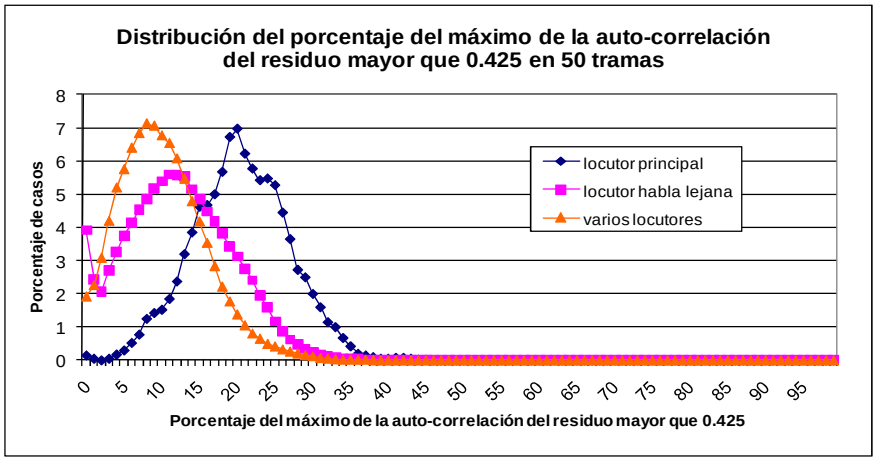

Figura 5.19. Distribución del porcentaje con auto-correlación del residuo mayor que 0.425 para pulsos de 50 tramas (N = 50).

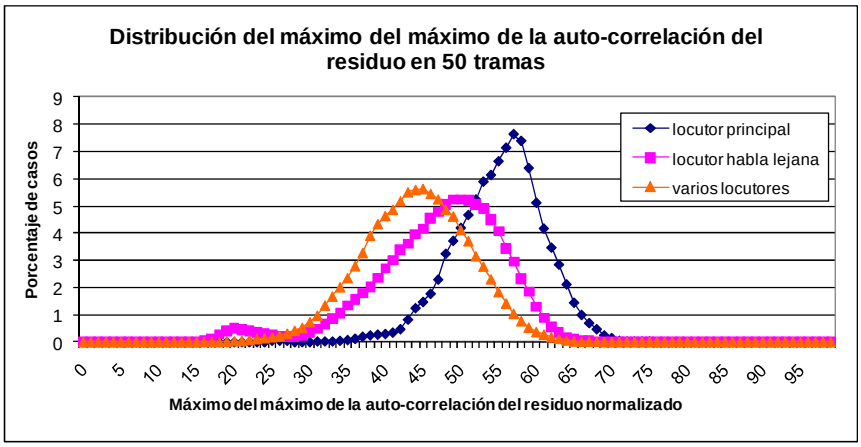

Figura 5.20. Distribución del máximo en un pulso del máximo de auto-correlación en cada trama del residuo normalizado (de 0 a 100) para pulsos de 50 tramas.

Es importante señalar el uso en el eje de abscisas de los valores normalizados en Fig.5.20-5.22 (recuérdese ec. 5.10).

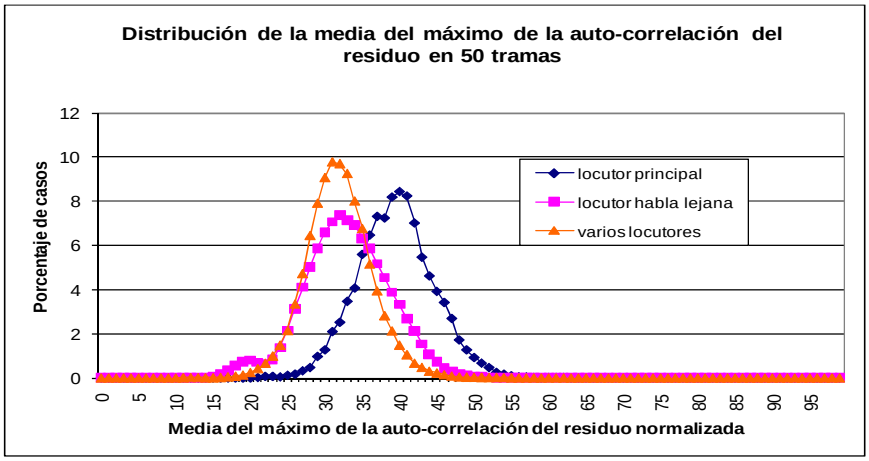

Figura 5.21. Distribución de la media en el pulso del máximo de auto-correlación en cada trama del residuo normalizada (de 0 a 100) para pulsos de 50 tramas.

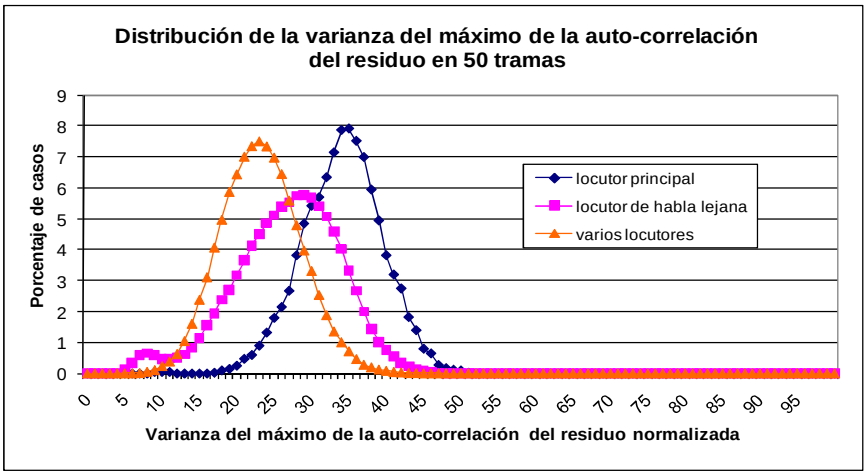

Figura 5.22. Distribución de la varianza en el pulso del máximo de auto-correlación en cada trama del residuo normalizada (de 0 a 100) para pulsos de 50 tramas.

5.5.- Resumen.

En este último apartado se resume y a la vez se amplía de forma esquemática (Tabla 5.2) los errores de clasificación de todos los estadísticos sobre las características estudiadas. Se amplían los resultados obtenidos en apartados anteriores en cuanto al número de tramas de pulso de voz consideradas. En la Tabla 5.2, “EC 1c-1l“ denota el error de clasificación de un locutor de habla cercana frente a un locutor de habla lejana, mientras que “EC 1c-nl“ denota el error de clasificación de un locutor de habla cercana frente a varios locutores de habla lejana. En negrita se seleccionan los estadísticos con mejores resultados, es decir, los que menores errores de clasificación obtienen y que serán usados e incluidos dentro del módulo de detección de pulsos dentro del sistema completo del Detector de Actividad que se mostrará en el siguiente capítulo. De las características seleccionadas, se puede decir que, poseen comportamientos similares la auto-correlación y la auto-correlación del residuo, mientras que la kurtosis del residuo mide la calidad de la voz detectada y la distancia de Mahalanobis nos informa de la velocidad con que varía el espectro de la voz.

Medida		*Número de tramas por pulso*	*EC 1c–1l (%)*	*EC 1c–nl (%)*
Porcentaje de tramas con máximo de auto-correlación mayor que 0.9		**50**	**33.5**	**14.8**
		100	28.8	9.7
		500	18.5	3.2
Distancia de Mahalanobis entre coeficientes MFCC		**50**	**34.9**	**23.7**
		100	31.6	19.7
		500	26.8	13.9
Porcentaje de tramas con una kurtosis del residuo mayor que 5		**50**	**17.4**	**18.8**
		100	11.9	13.8
		500	5.4	8.1
Mínimo de la varianza del residuo en un pulso de voz		50	25.6	19.2
		100	17.1	12.0
		500	7.6	1.3
Mínimo de máximo del residuo en un pulso de voz		50	30.7	24.5
		100	23.6	19.0
		500	10.6	10.3
Máximo de auto-correlación del residuo	**Porcentaje de tramas con el máximo de auto-correlación mayor que 0.425**	**50**	**26.8**	**18.8**
		100	23.2	12.7
		500	16.5	3.6
	Máximo	50	29.2	19.6
		100	26.3	16.3
		500	21.6	11.6
	Media	50	26.2	19.4
		100	23.1	14.3
		500	16.3	3.3
	Varianza	**50**	**28.1**	**14.9**
		100	24.3	10.2
		500	20.2	2.7

Tabla 5.2. Errores de clasificación por estadístico.

CAPÍTULO 6

SISTEMA FINAL DE DETECCIÓN DE ACTIVIDAD DE VOZ Y RESULTADOS CON VOCES DE FONDO

6.1.- Introducción.

En este capítulo se presentan distintas técnicas de integración, que usan toda la información obtenida en el Capítulo 5, a nivel de pulso:

1. Fusión simple de características a nivel de decisión sin entrenamiento: umbrales de decisión. Esta técnica tiene en cuenta varios estadísticos cuyos umbrales se ajustan convenientemente mediante una serie de ficheros de la base de datos de desarrollo. En este caso se realizan pruebas tanto usando estos estadísticos de forma individual como de forma conjunta.
2. Fusión de características a nivel de decisión con entrenamiento. En este caso se realiza el entrenamiento, usando todos los estadísticos seleccionados, mediante una base de datos de entrenamiento que contiene voces de fondo. Se consideran las siguientes técnicas de combinación de medidas:
    a. Árbol de decisión: durante el entrenamiento del árbol de decisión se realiza una serie de preguntas sobre la secuencia de medidas y los nodos se dividen para maximizar la detección de pulsos de voz lejana. Las preguntas para la bifurcación de las tramas se realizan variando los valores de cada estadístico y la pregunta óptima (best-question) se usa para dividir el t-ésimo nodo.
    b. Red neuronal. Se trata de una red neuronal de tres capas (multicapa) en la que la entrada es un vector formado por los estadísticos seleccionados, y la salida está formada por dos neuronas, las pertenecientes a las dos clases consideradas, voz procedente de un locutor principal y voz procente de las voces de fondo. La capa oculta del perceptrón está formada por treinta neuronas [108].

Se trata de usar estas técnicas para rechazar los pulsos de voz de habla lejana a través de una toma de decisión controlada y basada en algoritmos o estadísticos que parten de las medidas estudiadas en el Capítulo 5. También se realiza un estudio detallado de las mismas con sus correspondientes resultados y, a su vez, estos resultados se comparan con los resultados de otros *VAD*s estándar, el AMR1 [106], AMR2 [106], G729 anexo b [3] y AURORA (FD) [77]. Por otro lado, es

importante comentar que aunque la toma de decisión se realiza a nivel de pulso de voz, los errores finales se muestran también a nivel de trama para así obtener resultados con la mayor resolución posible.

6.2.- Consideraciones previas.

Antes de comenzar a describir los algoritmos para la toma de decisión, es importante aclarar cuáles van a ser los estadísticos sobre las características usadas por los mismos. En nuestro caso, dichos estadísticos no serán todos los contemplados en el capítulo anterior, si no sólo los que obtienen un menor error de clasificación en el estudio previo sobre la parte de desarrollo de la base de datos Av16.3 (DEV_AV). Así pues, los estadísticos seleccionados son los siguientes:

1. Armonicidad (máximo valor de auto-correlación cuando se calcula el “pitch”). ARMON. En este caso se considera el porcentaje de tramas, dentro de un pulso de voz, con un valor máximo de auto-correlación mayor que 0.9.
2. LPC Residual de orden 10:
    a. Kurtosis del residuo. LPCKURT. El porcentaje de tramas, dentro de un pulso de voz, con una kurtosis del residuo mayor que 5.
    b. Máximo valor de la auto-correlación del residuo en cada trama:
        i. LPCMPORC. Porcentaje de tramas, dentro de un pulso de voz, con un valor máximo de auto-correlación del residuo mayor que 0.425.
        ii. LPCMVAR. Varianza, dentro de un pulso de voz, del máximo de auto-correlación del residuo.
3. Distancia de Mahalanobis entre los MFCCs obtenidos sobre tramas de voz consecutivas. DISTMAH. En este caso, se considera la mínima distancia dentro de cada pulso de voz.

Por lo tanto, son 5 los estadísticos a considerar y éstos serán los usados por los algoritmos de decisión que serán descritos en los siguientes apartados. Además siempre se van a usar las mismas bases de datos para poder realizar estudios comparativos en cada unos de los casos.

6.3.- Fusión simple de características a nivel de decisión sin entrenamiento: umbrales de decisión.

En este apartado se realiza el estudio de la incorporación de nuevas características para el rechazo de voz de fondo mediante decisiones basadas en umbrales. Estos umbrales han sido ajustados convenientemente mediante un subconjunto de ficheros, TRAIN_AV, aleatoriamente seleccionados, de la base de datos Av16.3. El mencionado ajuste implica la obtención de un umbral para cada uno de los estadísticos, sobre las características estudiadas. Estos umbrales delimitarán las dos clases a clasificar: voz de un locutor principal (el pulso de voz se acepta) y voz lejana de uno o varios locutores (el pulso de voz se rechaza). Los umbrales ajustados, junto con la condición requerida para que se considere pulso de voz válido, se presentan en la Tabla 6.1. Es importante comentar que los umbrales de la Tabla 6.1 usan valores absolutos y no normalizados, lo que explica la aparición de valores elevados.

<table>
<tr><th colspan="3">Medida</th><th>Umbral</th></tr>
<tr><td colspan="3">Porcentaje de tramas con un valor del máximo de auto-correlación mayor que 0.9 sobre un pulso de voz</td><td>≥1.0%</td></tr>
<tr><td rowspan="3">LPC Residual de orden 10</td><td colspan="2">Kurtosis del residuo: porcentaje de tramas con un valor de la kurtosis del residuo mayor que 5 sobre un pulso de voz</td><td>≥4.0%</td></tr>
<tr><td rowspan="2">Máximo valor de la auto-correlación del residuo en cada trama</td><td>Porcentaje de tramas con un valor mayor que 0.425 sobre un pulso de voz</td><td>≥4.0%</td></tr>
<tr><td>Varianza sobre un pulso de voz</td><td>≥0.08</td></tr>
<tr><td colspan="3">Distancia de Mahalanobis entre vectores MFCC obtenidos en tramas de voz consecutivas: mínima distancia sobre un pulso de voz</td><td>≤300.0</td></tr>
</table>

Tabla 6.1. Estadísticos sobre características seleccionadas y umbrales ajustados.

Los umbrales de la Fig.6.1. fueron ajustados de forma independiente, es decir, se buscaba un valor del umbral que hiciese mínimo el error de detección a nivel de trama para cada una de las medidas de forma separada, y así, cada medida generaba un error de detección mínimo. Los resultados con estos umbrales y sobre

las bases de datos de test utilizadas, distintas a la de ajuste, se obtuvieron para diferentes casos:

- Para cada estadístico de forma independiente.
- Mediante un algoritmo basado en el número de condiciones cumplidas por cada estadístico. Por tanto es un caso de combinación o fusión de estadísticos con sus correspondientes umbrales.

Para el caso del primer punto, los estadísticos de forma independiente, se muestran los errores de detección (*GDE*) para las distintas bases de datos contempladas en el estudio en Tabla 6.2-6.4 con diferentes SNRs.

	GDE (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
ARMON ≥ 1%	39.62%	37.36%	31.76%	25.25%	19.35%
LPCKURT ≥ 4%	37.15%	35.68%	31.18%	25.11%	19.67%
LPCMPORC ≥ 4%	32.63%	31.90%	28.29%	23.53%	19.24%
LPCMVAR ≥ 0.08	36.82%	35.72%	30.99%	25.14%	19.71%
DISTMAH ≤ 300	40.39%	38.48%	33.36%	26.81%	20.76%

Tabla 6.2. *GDE* de cada estadístico de forma independiente para distintas SNRs. Base de datos con voces de fondo antes de pronunciación (TEST_GSM_PREAV).

La base de datos usada para obtener los resultados de la Tabla 6.2 está formada por ficheros de voz limpia contaminados con algunas voces de fondo de la base de datos Av16.3 antes de pronunciación y a diferentes SNRs: 5dB, 10dB, 15dB, 20dB y 25dB. Además, se puede observar en la Tabla 6.2 que la medida que obtiene un *GDE* más pequeño para todas las SNRs, significativamente, es el estadístico LPCMPORC, esto es, el porcentaje de tramas con un máximo valor de auto-correlación del residuo mayor que 0.425 en un pulso de voz.

La base de datos usada para obtener los resultados de la Tabla 6.3 está formada por ficheros de voz limpia contaminados con algunas voces de fondo de la base de datos Av16.3 después de pronunciación y a diferentes SNRs: 5dB, 10dB, 15dB, 20dB y 25dB. Se puede observar en la Tabla 6.3 que es también el estadístico LPCMPORC el que mejor resultados obtiene para todas las SNRs, pero no con tantas diferencias como en el caso anterior (voces de fondo antes de pronunciación).

	GDE (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
ARMON ≥ 1%	24.99%	23.98%	22.61%	20.17%	17.94%
LPCKURT ≥ 4%	26.03%	24.99%	23.52%	20.95%	18.59%
LPCMPORC ≥ 4%	24.30%	23.14%	22.16%	19.48%	17.45%
LPCMVAR ≥ 0.08	25.64%	24.50%	23.24%	20.62%	18.43%
DISTMAH ≤ 300	26.33%	25.24%	23.95%	21.43%	19.20%

Tabla 6.3. *GDE* de cada estadístico de forma independiente para distintas SNRs. Base de datos con voces de fondo después de pronunciación (TEST_GSM_POSTAV).

Para finalizar el estudio de los estadísticos sobre las características de forma independiente, se muestran los resultados con una base datos basada en conversaciones de servicios reales (TEST_GSM_RUIDONE). Está formada por 2630 ficheros etiquetados que contienen conversaciones procedentes de servicios reales y emitidas por teléfonos móviles con tecnología GSM donde los locutores principales, 15 hombres y 13 mujeres, se encuentran en entornos adversos con ruidos no estacionarios: bares, salas con ruido de fondo de televisión (tertulias) o simplemente en la calle donde existe ruido de fondo procedente de otros locutores que están a una cierta distancia del locutor principal. En este caso los resultados de la Tabla 6.4 arrojan que es el estadístico LPCKURT es el que obtiene un *GDE* más pequeño, esto es, el porcentaje de tramas con una kurtosis de residuo mayor que 5 en un pulso de voz.

	GDE (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
ARMON ≥ 1%	40.70%	27.74%	24.82%	17.86%	14.01%
LPCKURT ≥ 4%	36.30%	23.95%	21.96%	16.14%	13.95%
LPCMPORC ≥ 4%	41.24%	28.41%	25.44%	18.23%	14.04%
LPCMVAR ≥ 0.08	41.79%	28.81%	25.90%	18.65%	14.29%
DISTMAH ≤ 300	40.74%	29.00%	26.14%	19.07%	14.75%

Tabla 6.4. *GDE* de cada estadístico de forma independiente para distintas SNRs. Base de datos basada en servicios conversacionales reales en entornos adversos (TEST_GSM_RUIDONE).

Parece, por tanto, que de forma independiente, compiten para obtener el menor *GDE* dos medidas:

- El porcentaje de tramas con un máximo valor de auto-correlación del residuo mayor que 0.425 en un pulso de voz: LPCPORC.

- El porcentaje de tramas con una kurtosis de residuo mayor que 5 en un pulso de voz: LPCKURT.

Con el fin de homogeneizar los resultados, y reducir aún más el *GDE*, se realiza a continuación una fusión a nivel de decisión de las condiciones de las medidas de la Tabla 6.1 en cuanto al número de éstas que ha de cumplirse, al menos, para aceptar o rechazar un pulso de voz. Por ello, se realizaron pruebas sobre las mismas bases de datos anteriores y para las distintas SNR para 5 casos posibles: cuando se cumple al menos una de las condiciones, al menos dos, al menos tres, al menos cuatro y cuando se cumplen las cinco. De aquí se deberá encontrar que uno de los casos obtiene el *GDE* mínimo, que además debería ser menor que el mejor caso que se obtenía con las medidas de forma individual. En Tabla 6.5-6.7 se muestran estos resultados para cada una de las tres bases de datos anteriores.

Se cumple:	*GDE* (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
Al menos 1	40.84%	38.76%	33.53%	26.86%	20.80%
Al menos 2	40.28%	38.38%	33.31%	26.65%	20.63%
Al menos 3	38.48%	36.94%	31.79%	25.56%	19.85%
Al menos 4	34.07%	33.14%	29.02%	23.50%	18.55%
Las 5	**31.93%**	**30.98%**	**27.01%**	**22.34%**	**18.01%**

Tabla 6.5. *GDE* cuando se cumplen, al menos, las n(1,2,3,4 o 5) condiciones para distintas SNRs. Base de datos TEST_GSM_PREAV (voces de fondo antes de pronunciación).

Se cumple:	*GDE* (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
Al menos 1	26.35%	25.28%	23.97%	21.47%	19.22%
Al menos 2	26.22%	25.16%	23.85%	21.34%	19.10%
Al menos 3	25.65%	24.64%	23.37%	20.77%	18.47%
Al menos 4	25.08%	23.94%	22.59%	19.90%	17.64%
Las 5	**24.02%**	**22.88%**	**21.78%**	**19.25%**	**17.27%**

Tabla 6.6. *GDE* cuando se cumplen, al menos, las n(1,2,3,4 o 5) condiciones para distintas SNRs. Base de datos TEST_GSM_POSTAV (voces de fondo después de pronunciación).

Se cumple:	*GDE* (5dB)	*GDE* (10dB)	*GDE* (15dB)	*GDE* (20dB)	*GDE* (25dB)
Al menos 1	42.35%	29.41%	26.47%	19.17%	14.75%
Al menos 2	42.10%	29.17%	26.24%	18.99%	14.11%
Al menos 3	41.23%	28.46%	25.52%	18.41%	14.11%
Al menos 4	39.00%	26.80%	24.02%	17.36%	13.72%
Las 5	**35.42%**	**23.54%**	**21.31%**	**15.16%**	**12.92%**

Tabla 6.7. *GDE* cuando se cumplen, al menos, las n(1,2,3,4 o 5) condiciones para distintas SNRs. Base de datos TEST_GSM_RUIDONE (servicios conversacionales reales en entornos adversos).

Como se puede observar en las tablas anteriores, se obtiene un *GDE* más pequeño cuando se cumplen las 5 condiciones, siendo así más exigentes con la calidad del pulso, y además ocurre así para las tres bases de datos. Queda así por tanto unificado el criterio de selección de pulso para todas las bases de datos y con un comportamiento idéntico y homogéneo para cada una de ellas. Además, como era de esperar, los mejores resultados para las medidas de forma individual son superados por el caso en el que se tienen que cumplir todas las condiciones (para la aceptación de pulso). Esto quiere decir que, cuando entran en juego todas las medidas, con sus correspondientes umbrales, se rechazan más pulsos de voces lejanas (falsas alarmas que antiguamente se consideraban como pulsos de voz válidos) sin perjudicar a los pulsos de voz procedentes de locutores principales.

6.3.1.- Comparación del nuevo VAD basado en umbrales de decisión con estándares de referencia.

Una vez seleccionado, del apartado anterior, el caso para el que se obtiene el menor *GDE*, se pueden realizar los estudios comparativos con otros detectores de actividad de referencia: AMR1 [106], AMR2 [106], AURORA(FD) [77] y G729 anexo b [3]. Para ello, calculamos los valores de *GDE* para el resto de detectores mencionados y con las mismas tres bases de datos con las que ya se ha calculado el *GDE* para el caso del Detector de Actividad basado en umbrales propuesto por nosotros. Es importante comentar que el punto de trabajo elegido para estos detectores es la implementada por el estándar, es decir, no se ha realizado ninguna modificación en sus parámetros, se entiende que la configuración por defecto es la óptima. Teniendo en cuenta lo anterior, se presenta de forma gráfica la comparación

del *GDE* obtenido por cada *VAD* en Fig.6.1-6.3. y para cada una de las tres bases de datos que contienen voces de fondo.

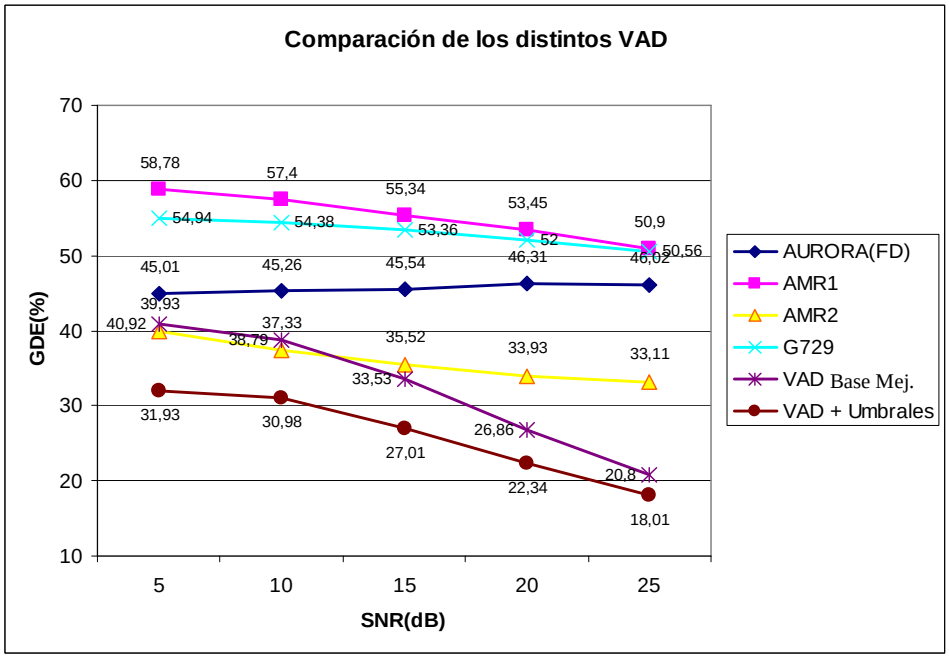

Figura 6.1. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_PREAV (voces de fondo antes de pronunciación). Distintas SNRs.

Como se puede observar en la Fig.6.1, la técnica basada en umbrales de decisión aplicada en el *VAD* Base Mejorado (*VAD* + Umbrales) es la que genera el menor *GDE* para todas las SNRs y, además, las diferencias son notorias tanto respecto del *VAD* que usábamos como punto de partida como del resto de detectores estudiados. Es importante comentar también que el *VAD* de Motorola (AMR2) es mejor que el *VAD* Base Mejorado para SNRs de 5dB y 10dB y sin embargo cuando se le aplica el nuevo método le supera con creces. La mejora relativa del *VAD* basado en umbrales (*VAD* + Umbrales) mediante el nuevo método propuesto para 5dB en comparación con el segundo mejor caso, el del AMR2, es del 20% aproximadamente. En general, para SNRs bajas, el peor caso es para el AMR1, pero para SNRs más altas el comportamiento del AURORA(FD), AMR1 y G729 anexo b es bastante parecido. La mejora general del *VAD* propuesto se debe principalmente a la disminución de la tasa de falsas alarmas generadas por los

pulsos creados por las voces de fondo que provoca a su vez una disminución del Error de Detección Global. A continuación, en la Fig.6.2, se presentan los resultados para la base de datos TEST_GSM_POSTAV, en la que las voces de fondo se encuentran después de la pronunciación de los locutores principales.

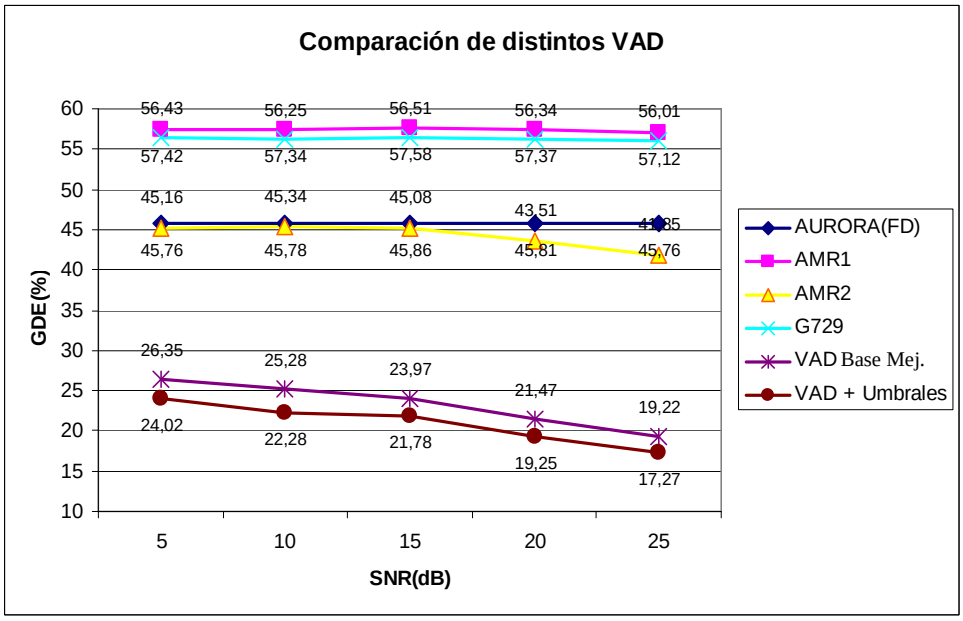

Figura 6.2. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_POSTAV (voces de fondo después de pronunciación). Distintas SNRs.

En la Fig.6.2, de nuevo es el *VAD* basado en umbrales de decisión (*VAD* + Umbrales) el que genera los mejores resultados, aunque, a diferencia del caso de la Fig.6.2, el *VAD* Base Mejorado también consigue unos resultados mejores que AMR2, AURORA(FD), G729 anexo b y AMR1 para todas las SNRs. Además, las diferencias entre el *VAD* basado en umbrales de decisión y el *VAD* Base Mejorado son bastantes significativas en comparación con el resto. Los *GDEs* más altos los obtienen en este caso los *VAD* de los codificadores AMR1 y G729 anexo b con resultados bastante parecidos en todas las SNRs. También es importante resaltar que los detectores AMR1, G729 anexo b y AURORA(FD) muestran un *GDE* con comportamiento bastante plano cuando varía la SNR. Sin embargo para el AMR2, el *VAD* Base Mejorado y el *VAD* basado en umbrales, el *GDE* disminuye cuando aumenta la SNR. De nuevo, la disminución del *GDE* para el *VAD* propuesto es

consecuencia de la disminución de la tasa de falsas alarmas en presencia de voces de fondo.

En la Fig.6.3 se muestran los resultados con la base de datos TEST_GSM_RUIDONE: basada en servicios conversacionales reales en entornos adversos como bares, salas con la TV encendida o en la calle.

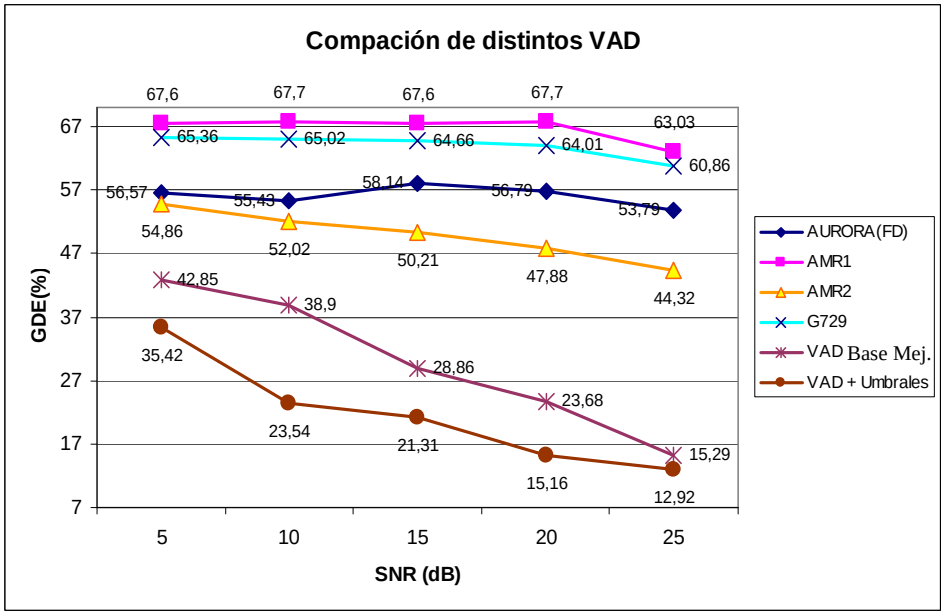

Figura 6.3. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_RUIDONE (servicios reales conversacionales en entornos adversos). Distintas SNRs.

Como se puede observar, la nueva técnica (*VAD* + Umbrales) obtiene el menor *GDE* para todas las SNRs. El segundo mejor caso, sin contar el VAD Base Mejorado, es para el AMR2, aunque este genera resultados significativamente peores para todas las SNRs. Como se ha comentado anteriormente, la disminución del *GDE* se produce gracias a la disminución de las falsas alarmas procedentes de las voces de fondo. Si se compara el *VAD* basado en umbrales de decisión con el *VAD* Base Mejorado es fácil verificar que la mejora es notoria, sobre todo para SNRs bajas, y se ratifica, de nuevo, la gran ventaja de usar esta nueva técnica basada en medidas para el filtrado de pronunciaciones de voz. La mejora relativa se encuentra entre el 40% y el 15%, variabilidad debida al distinto comportamiento para las diferentes SNRs.

Adicionalmente, y con el fin de visualizar el punto de trabajo de los detectores de las gráficas anteriores, se muestra gráficamente la tasa de falsas alarmas (FAR, del inglés False Alarm Rate) en Fig.6.4-6.6 para las tres bases de datos tratadas anteriormente, recuérdese que en este caso se normalizan los errores de inserción al número de tramas de ruido (ec. 3.2). Además es importante comentar que en estas bases de datos el número de tramas de voz y de ruido son prácticamente el mismo.

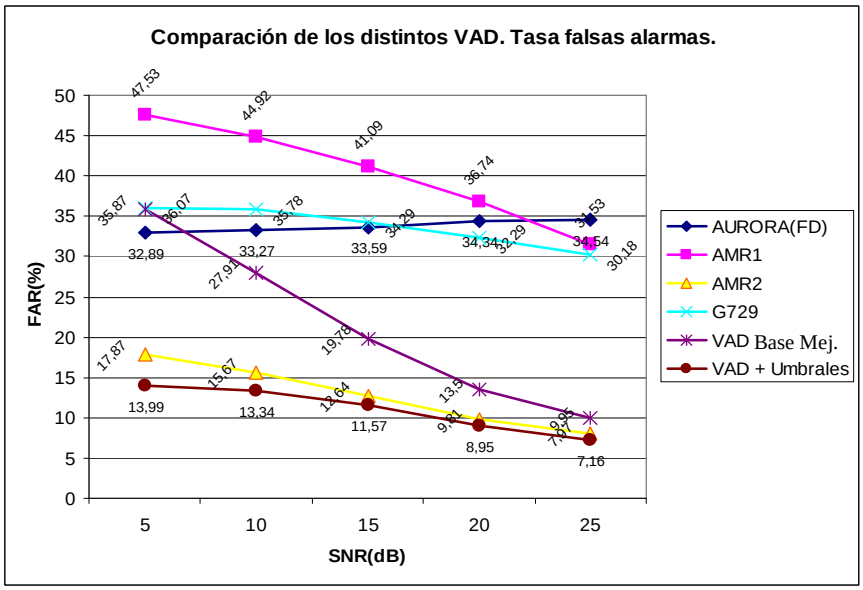

Figura 6.4. Tasa de falsas alarmas (FAR) para la base de datos TEST_GSM_PREAV (voces de fondo antes de pronunciación). Distintas SNRs.

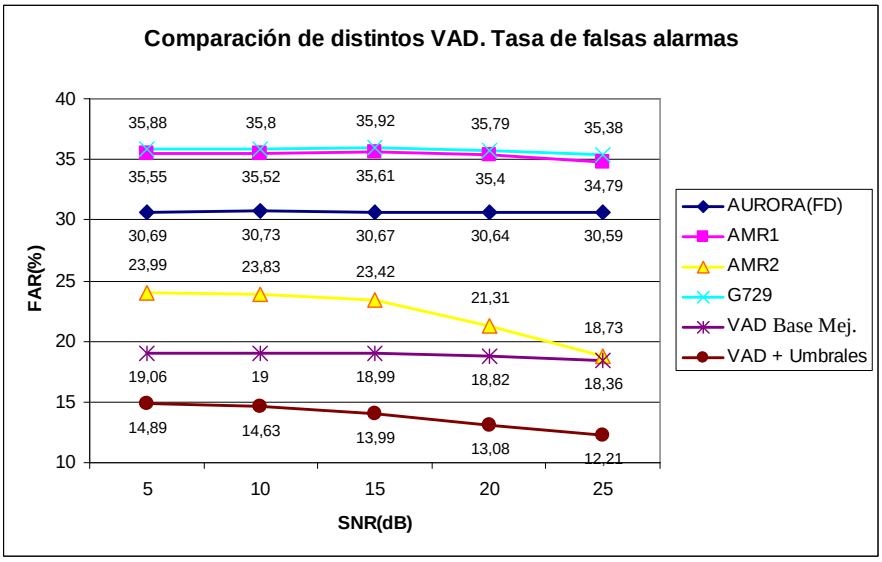

Figura 6.5. Tasa de falsas alarmas (FAR) para la base de datos TEST_GSM_POSTAV (voces de fondo después de pronunciación). Distintas SNRs.

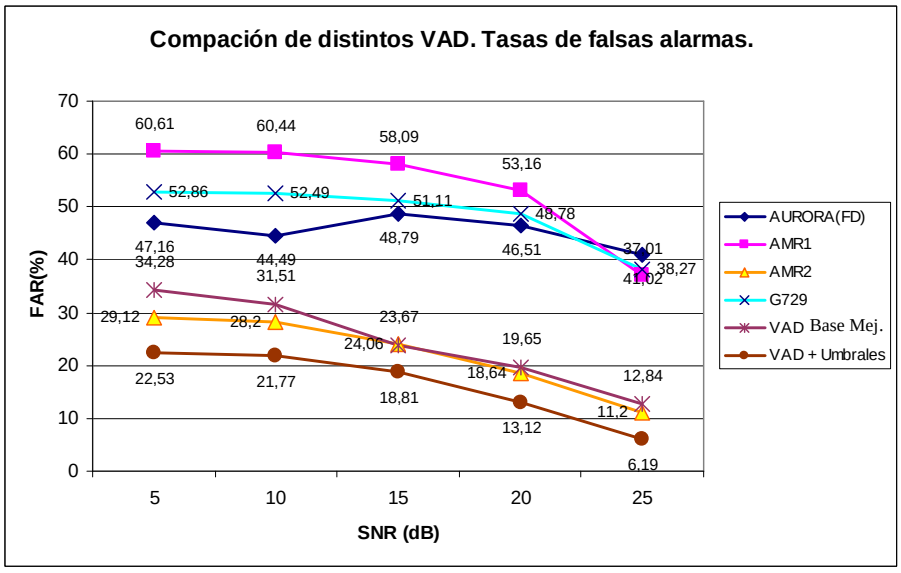

Figura 6.6. Tasa de falsas alarmas (FAR) para la base de datos TEST_GSM_RUIDONE (servicios reales conversacionales en entornos adversos). Distintas SNRs.

Como ya se comentaba anteriormente, la tasa de falsas alarmas disminuye, frente al *VAD* Base Mejorado, cuando se usa el método de los umbrales. Realmente, al usar la nueva técnica basada en umbrales se mejora mucho más en falsas alarmas de lo que se pierde en la tasa de falsos rechazos, y como consecuencia, el *GDE*

disminuye considerablemente. Por ejemplo, en el caso de TEST_GSM_RUIDONE para 5dB ocurre lo siguiente:

- *GDE* (VAD Base Mej.) = 42.85% y *GDE* (VAD + umbrales) = 35.42%.
- FAR (VAD Base Mej.) = 34.28% y FAR (VAD + umbrales) = 22.53%.
- MR (VAD Base Mej.) = 8.45% y MR (VAD + umbrales) = 12.76%.

Esto quiere decir que aunque se haya empeorado un 4.32% (12.76% - 8.45% = 4.31%) en la tasa de falsos rechazos (ec.3.3), se ha mejorado un 11.75% (34.28% - 22.53% = 11.75%) en la tasa de falsas alarmas (ec. 3.2). Como consecuencia, se tiene una mejora del GDE del 7.43% (42.85% – 35.42% = 7.43%). El que se pierda un poco en la tasa de falsos rechazos es debido a que las clases “voz cercana” y “voz lejana de uno o varios locutores” se solapaban y como consecuencia existía un error de clasificación, como ya se veía en el capítulo anterior.

Por último, y para finalizar el apartado, es conveniente mostrar los resultados del GDE para el caso de voz limpia (≥ 25dB), es decir, TEST_GSM_LIMPIA, y comprobar que el nuevo algoritmo basado en umbrales no empeora los resultados en ese caso. De nuevo se compara el resultado con el de los detectores también estudiados con las tres bases de datos anteriores (Tabla 6.8).

DET_ERROR	AURORA(FD)	AMR1	AMR2	G729b	VAD Base Mejorado	VAD + Umbrales
GDE(%)	35.16	17.97	17.28	29.62	14.27	14.46

Tabla 6.8. Error de Detección Global (GDE) para TEST_GSM_LIMPIA: umbrales.

Los resultados (GDE) de la Tabla 6.8 muestran que el *VAD* basado en umbrales de decisión (*VAD* + Umbrales), el *VAD* Base Mejorado obtenido del capítulo 4, el AMR1 y AMR2 obtienen resultados similares, mientras que el G729b y el AURORA(FD) son peores, significantemente, en comparación con el resto.

6.4.- Fusión de características usando un árbol de decisión.

En el apartado anterior se tomaba una decisión basada en umbrales y en el número de condiciones, basadas en los mismos, que tenían que cumplirse para obtener el menor GDE posible. En el caso que trata este apartado se propone un algoritmo basado en un árbol de decisión. Esta técnica se aplica también dentro del módulo de detección de pulsos dentro del sistema completo de detección.

En las teorías de complejidad computacional, un modelo de árbol de decisión es un modelo computacional en el que el algoritmo considerado es básicamente un árbol de decisión, esto es, un conjunto de operaciones de ramificación basadas en las comparaciones de valores en forma de vector de diferentes umbrales. El árbol de decisión trata de obtener en las ramas del árbol la mayor concentración de una clase concreta. En este sentido, el algoritmo en cuestión puede tratarse como el cálculo de una función booleana (ec. 6.1).

$$f : \{0,1\}^n \rightarrow \{0,1\} \qquad (6.1)$$

donde la entrada es una serie de condiciones y la salida es la decisión final, en nuestro caso también booleana porque consideramos las clases pulso de voz cercana (“1”) y pulso de voz lejana (“0”). Existen distintas variantes de modelos de árbol de decisión. En este trabajo usamos un árbol de decisión estocástico.

El árbol de decisión necesita un vector de entrada compuesto por los cinco estadísticos, que también se usaban en la técnica del apartado anterior, por ser los que mejores resultados obtenían en el análisis realizado sobre la base de datos DEV_AV de AV16.3:

- El porcentaje de tramas con el máximo valor de auto-correlación mayor de 0.9 en un pulso de voz de N tramas.
- La mínima distancia de mahalanobis sobre coeficientes MFCC de tramas consecutivas en un pulso de voz de N tramas.
- El porcentaje de tramas con una kurtosis del residuo mayor que 5 en un pulso de voz de N tramas.

- El porcentaje de tramas con una auto-correlación máxima del residuo mayor que 0.425 en un pulso de voz de N tramas.
- La varianza del máximo de auto-correlación del residuo en un pulso de N tramas.

Se calcula este vector de cinco componentes en todos los pulsos de voz, tanto del locutor principal como de las falsas alarmas provocadas por locutores de habla lejana, y el árbol de decisión toma la decisión de forma global. Para esto, es necesario tener en cuenta una serie de premisas:

1. La base de datos de ajuste es etiquetada manualmente con un “1” si la trama es una trama de voz procedente de un locutor principal y con un “0” si procede de locutores de habla lejana.
2. Del etiquetado manual anterior, los pulsos de voz verdaderos son bien conocidos para la salida del árbol de decisión teniendo en cuenta lo siguiente:
    - Si el número de tramas etiquetadas como “1” es mayor que el número de tramas etiquetadas como “0” en los pulsos de voz, la salida del pulso de voz será “1” (se tipifica como clase “1”), y en caso contrario será “0” (se tipifica como clase “0”).

Como era de esperar, las decisiones también se tomarán a nivel de pulso, como se puede discernir de las premisas anteriores, pero siempre conociendo específicamente cuántas tramas, dentro de un pulso tratado como de la clase voz (“1”), son de voz y cuántas de ruido, ya que el GDE siempre se muestra a nivel de trama y no de pulso.

El árbol de decisión se implementa usando las puntuaciones de cada medida, refiriéndonos a los cinco estadísticos del vector de entrada. Durante el entrenamiento del árbol de decisión se realiza una serie de preguntas sobre la secuencia de medidas y los nodos se dividen para maximizar la detección de pulsos de voz lejana. Por ejemplo, “¿Q es el porcentaje de tramas con un valor máximo de auto-correlación mayor que X?”. Las preguntas para la bifurcación de las ramas se realizan variando los valores de cada una de las 5 medidas y la pregunta óptima

(best-question) se usa para dividir el t-ésimo (t^{th}) nodo. La siguiente fórmula se utiliza para analizar la impureza del nodo $I(t)$,

$$I(t) = 2\, p(CORRECT/t)\, p(INCORRECT/t) \qquad (6.2)$$

donde $p(CORRECT/t)$ es la probabilidad de obtener un pulso de voz de campo cercano en el nodo t y $p(INCORRECT/t)$ es la probabilidad de obtener un pulso de voz de campo lejano o ruido en el nodo t. La división termina cuando una de estas dos condiciones se cumple:

- El número de vectores entrenados en un nodo dado es menor de 10.
- Las inserciones en un nodo producen un nuevo nodo sin vectores.

Una vez construido el árbol *T0*, se ajusta para obtener el óptimo sub-árbol. En nuestros experimentos usamos el ajuste de coste de complejidad mínimo (Minimal Cost-Complexity Pruning). Se calcula una secuencia de sub-árboles *T1*, *T2*,...,*Tn*, que minimiza el coste de complejidad hasta que se alcanza al nodo raíz. Posteriormente se evalúan estos sub-árboles, usando el criterio de validación y seleccionando el mejor caso. Para más detalles consultar [99,100].

A continuación, en Fig.6.7-6.9, se muestran los resultados comparativos, de forma idéntica al apartado anterior (usando las mismas bases de datos), de los resultados obtenidos con el método de decisión presentado en este apartado. En este caso, el proceso de entrenamiento no se realiza mediante una base de datos previa, sino que el programa de entrenamiento creado selecciona un 70% de los ficheros de cada una de las bases de datos de forma independiente, esto es, de TEST_GSM_RUIDONE, TEST_GSM_PREAV y TEST_GSM_POSTAV. Del 30% restante de cada una de ellas, un 15% se destina para realizar los ajustes pertinentes y el otro 15% para los resultados de validación cruzada. Estos últimos son los que se muestran en las figuras anteriormente mencionadas.

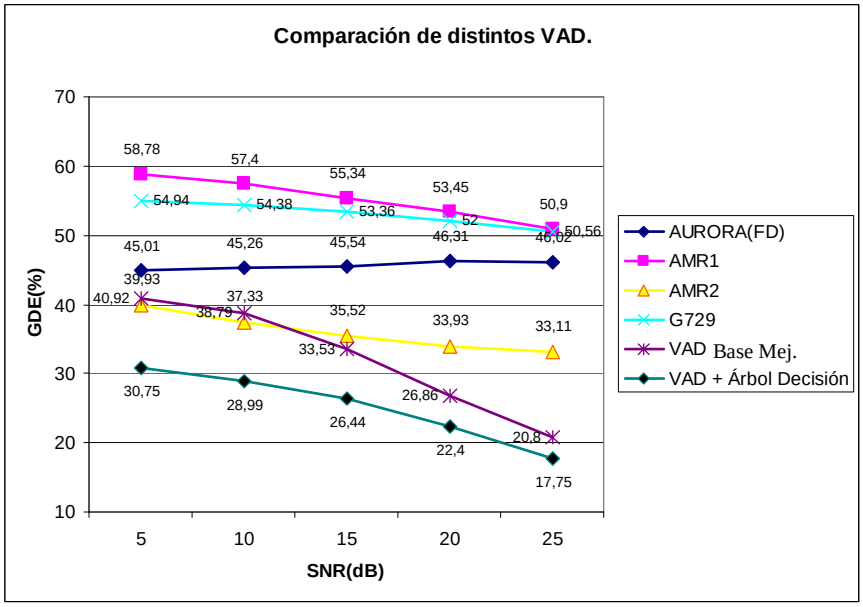

Figura 6.7. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_PREAV (voces de fondo antes de pronunciación). Distintas SNRs.

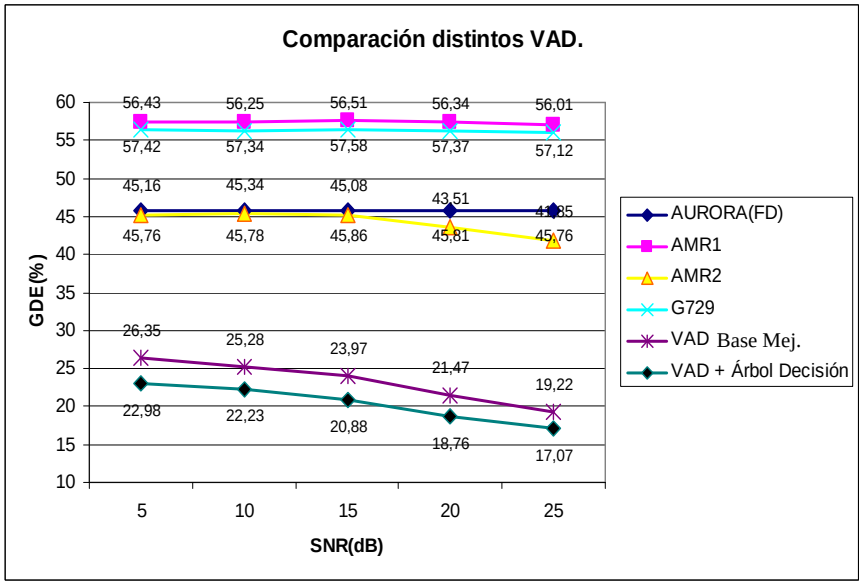

Figura 6.8. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_POSTAV (voces de fondo después de pronunciación). Distintas SNRs.

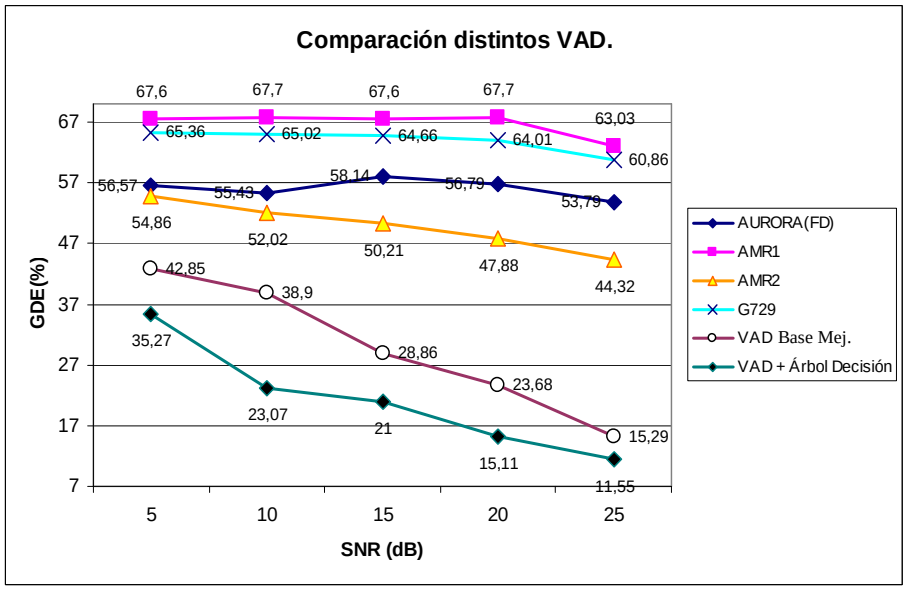

Figura 6.9. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_RUIDONE (servicios reales conversacionales en entornos adversos). Distintas SNRs.

La Fig.6.7 muestra que el *VAD* que menor *GDE* obtiene para todas las SNRs, con mucho, es el que usa el árbol como técnica de decisión, con una mejora relativa alrededor del 23 % respecto el *VAD* del códec AMR2 para una SNR=5dB. En general, para SNRs pequeñas, el AMR1 es el peor caso, mientras que para SNRs altas el comportamiento del AURORA(FD), AMR1 y G729 anexo b es bastante parecido. Además, se puede apreciar la mejora que se obtiene frente al *VAD* Base Mejorado: el *VAD* que usa el árbol de decisión tiene una mejora relativa del 24.85% para una SNR=5dB. La tasa de falsas alarmas para el *VAD* que usa el árbol de decisión es del 13.05%, 11.74%, 11.12%, 8.99% y 6.95% para 5, 10, 15, 20, y 25 dB respectivamente.

La Fig.6.8 también muestra, al igual que ocurre en la Fig.6.7, que el *VAD* del árbol es el que genera un error de detección global más pequeño. El *VAD* de Motorola (AMR2) es el segundo mejor caso (sin incluir el *VAD* de partida) pero con apreciables diferencias respecto del nuestro. Los peores resultados los obtienen el AMR1 y el G729 anexo b de forma bastante similar para todas las SNRs. También es importante remarcar el comportamiento lineal que ofrecen los *VAD* AMR1, G729 anexo b y AURORA(FD) para todas las SNRs. Sin embargo, el AMR2, el *VAD* Base

Mejorado y el que usa el árbol de decisión los errores de detección decrecen cuando se incrementa la SNR. La mejora del *VAD* del árbol respecto al base se debe principalmente a la disminución de tasa de falsas alarmas en presencia de voces de fondo: en este caso del 14.06%, 14.59%, 13.27%, 12.69% y 12.05% para 5, 10, 15, 20 y 25 dB respectivamente.

En la Fig.6.9 de nuevo el *VAD* con árbol de decisión es el que mejor resultados obtiene. Además, el AMR2 obtiene resultados significativamente peores para todas las SNRs. Como hemos comentado anteriormente, la reducción del *GDE* se debe a la disminución de la tasa de falsas alarmas: FAR(5dB) = 22.41%, FAR(10dB) = 21.39%, FAR(15dB) = 18.56%, FAR(20dB) = 13.08% y FAR(25dB) = 5.09%. Comparándolo con el *VAD* Base Mejorado, es fácil comprobar la ventaja de usar la nueva técnica, en general, mayor para SNRs pequeñas. La mejora relativa abarca desde 40.69%(SNR=10dB) hasta 17.69%(SNR=5dB).

Por tanto, y de forma genérica, se puede decir que el *VAD* que usa como técnica el árbol de decisión funciona mejor que la técnica basada en umbrales.

Al igual que en el apartado anterior, es conveniente mostrar los resultados del *GDE* para el caso de TEST_GSM_LIMPIA (voz limpia) y comprobar que el nuevo algoritmo basado en un árbol de decisión mantiene buenos resultados, y cómo se comporta en comparación con el resto de detectores estudiados. Los resultados se muestran en la Tabla 6.9.

DET_ERROR	**AURORA(FD)**	**AMR1**	**AMR2**	**G729b**	**VAD Base Mejorado**	**VAD+Árbol Dec.**
GDE(%)	35.16	17.97	17.28	29.62	14.27	14.39

Tabla 6.9. Error de Detección Global (*GDE*) para TEST_GSM_LIMPIA: Árbol de decisión.

Los resultados (*GDE*) de la Tabla 6.9 muestran que el *VAD* basado en el árbol de decisión, el *VAD* Base Mejorado, el AMR1 y AMR2 obtienen resultados similares, mientras que el G729b y el AURORA(FD) son peores, significativamente, en comparación con el resto.

Para finalizar el apartado, en la Tabla 6.10, se van a exponer también los resultados a nivel de pulso, con TEST_GSM_RUIDONE, TEST_GSM_PREAV y TEST_GSM_POSTAV, para poder así tener una idea de cuánto se equivoca el árbol

de decisión en este caso específico. Se medirá la tasa de aciertos tras la toma de decisión (recuérdese ec. 3.5).

Base de datos	TEST_GSM_PREAV	TEST_GSM_POSTAV	TEST_GSM_RUIDONE
Tasa aciertos (5dB)	**72.74 %**	**78.43 %**	**66.33 %**
Tasa aciertos (10dB)	**75.22 %**	**79.28 %**	**84.02 %**
Tasa aciertos (15dB)	**77.20 %**	**82.05 %**	**84.77 %**
Tasa aciertos (20dB)	**82.25 %**	**86.30 %**	**92.59 %**
Tasa aciertos (25dB)	**89.55 %**	**89.17 %**	**95.34 %**

Tabla 6.10. Tasa de aciertos del árbol de decisión para distintas SNRs y con las tres bases de datos que contienen voces de fondo.

Como se puede observar en la Tabla 6.9, la tasa de aciertos aumenta cuando aumenta la SNR, y aunque para 5dB los peores resultados se consiguen con la base de datos basada en servicios conversacionales reales (TEST_GSM_RUIDONE), para el resto de SNRs es la que mejor tasa de aciertos obtiene.

6.5.- Fusión de características usando una red neuronal.

Las redes de neuronas artificiales (denominadas habitualmente como RNA o en inglés "ANN") son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

Las redes neuronales consisten en una simulación de las propiedades observadas en los sistemas neuronales biológicos a través de modelos matemáticos recreados mediante mecanismos artificiales (como un circuito integrado, un ordenador o un conjunto de válvulas). El objetivo es conseguir que las máquinas den respuestas similares a las que es capaz de dar el cerebro. Estas respuestas se caracterizan por su generalización y su robustez.

Una red neuronal se compone de unidades llamadas neuronas. Cada neurona recibe una serie de entradas a través de interconexiones y emite una salida. Esta salida viene dada por tres funciones:

1. Una función de propagación (también conocida como función de excitación), que por lo general consiste en el sumatorio de cada entrada multiplicada por el peso de su interconexión (valor neto). Si el peso es positivo, la conexión se denomina excitatoria; si es negativo, se denomina inhibitoria.
2. Una función de activación, que modifica a la anterior: puede no existir, siendo en este caso la salida la misma función de propagación.
3. Una función de transferencia, que se aplica al valor devuelto por la función de activación: se utiliza para acotar la salida de la neurona y generalmente viene dada por la interpretación que queramos darle a dichas salidas.

La aproximación basada en las RNA parte de un conjunto de datos de entrada suficientemente significativo y el objetivo es conseguir que la red aprenda automáticamente las propiedades deseadas. En este sentido, el diseño de la red tiene que ver con cuestiones tales como la selección del modelo de red, la de las variables a incorporar y el preprocesamiento de la información que formará el conjunto de entrenamiento. El proceso por el que los parámetros de la red se adecuan a la resolución del problema se denomina entrenamiento neuronal.

Una primera clasificación de las redes de neuronas artificiales que se suele hacer es en función del patrón de conexiones que presenta. Así se definen tres tipos básicos de redes:

- Dos tipos de redes de propagación hacia delante o acíclicas en las que todas las señales van desde la capa de entrada hacia la salida sin existir ciclos, ni conexiones entre neuronas de la misma capa.
    - Monocapa: perceptrón monocapa.
    - Multicapa: perceptrón multicapa.
- Las redes recurrentes que presentan al menos un ciclo cerrado de activación neuronal.

En nuestro caso, la red neuronal a usar es una red neuronal de tres capas (multicapa) o perceptrón multicapa. Las características del perceptrón son las siguientes:

- La función de activación es una sigmoide con valores comprendidos entre 0 y 1.
- Las cinco entradas a la red son las siguientes:
    - El porcentaje de tramas con el máximo valor de auto-correlación mayor de 0.9 en un pulso de voz de N tramas.
    - La mínima distancia de mahalanobis sobre coeficientes MFCC de tramas consecutivas en un pulso de voz de N tramas.
    - El porcentaje de tramas con una kurtosis del residuo mayor que 5 en un pulso de voz de N tramas.
    - El porcentaje de tramas con una auto-correlación máxima del residuo mayor que 0.425 en un pulso de voz de N tramas.
    - La varianza del máximo de auto-correlación del residuo en un pulso de N tramas.
- Las anteriores 5 medidas han sido normalizadas entre 0 y 1 mediante el proceso de normalización z-score.
- La capa oculta del perceptrón está formada por 30 neuronas: la elección de 30 neuronas para la capa oculta suele ser habitual. Valores por debajo de 30 hacen que el error sea grande [108].
- La capa de salida está formada por 2 neuronas, las pertenecientes a las 2 clases consideradas: la que representa a la voz procedente de un locutor principal y la que no. Esta última clase representaría al ruido y a la voz de campo lejano.
- Al igual que en caso del Árbol de decisión, el proceso de entrenamiento no se realiza mediante una base de datos previa, sino que el programa de entrenamiento creado selecciona un 70% de los ficheros de cada una de las bases de datos usada de forma independiente, esto es, de TEST_GSM_RUIDONE, TEST_GSM_PREAV y TEST_GSM_POSTAV. Del 30% restante de cada una de ellas, un 15% se destina para realizar los ajustes pertinentes y el otro 15% para los resultados de evaluación finales.

Al igual que ocurría con el árbol de decisión, la red neuronal se aplicará sobre el módulo de detección de pulsos dentro del sistema completo de detección. También es importante comentar que el vector de cinco medidas de entrada a la red se calcula en todos los pulsos de voz, tanto los procedentes de un locutor principal como de las falsas alarmas provocadas por locutores de habla lejana o ruidos, y la red neuronal toma la decisión de forma global. Para esto, es necesario tener en cuenta una serie de premisas, similares a las que se hacían también con el árbol de decisión:

- La base de datos de entrenamiento se etiqueta manualmente con un “1” si la trama es una trama de voz procedente de un locutor principal y con un “0” si procede de locutores de habla lejana.
- Del etiquetado manual anterior, los pulsos de voz verdaderos son bien conocidos para la salida del perceptrón multicapa teniendo en cuenta lo siguiente:
    - Si el número de tramas etiquetadas como “1” es mayor que el número de tramas etiquetadas como “0” en los pulsos de voz, la salida del pulso de voz será “1” (se tipifica como clase “1”), y en caso contrario será “0” (se tipifica como clase “0”).

A continuación, en Fig.6.10-6.12 se muestran los resultados comparativos, de forma idéntica al apartado anterior (usando las mismas bases de datos), de los resultados obtenidos mediante la red neuronal como otro de los métodos usables para la toma de decisión.

Las tres figuras, Fig.6.10-6.12, muestran que el *VAD* que menor GDE obtiene para todas las SNRs, de forma significativa, es el que usa la RNA como técnica de decisión. Por ejemplo con TEST_GSM_PREAV, la mejora relativa es de alrededor del 26 % respecto el *VAD* del códec AMR2 y del 28.01% respecto del *VAD* de partida, para una SNR=5dB en ambos casos. En TEST_GSM_RUIDONE también es fácil comprobar la ventaja de usar la nueva técnica, en general, mayor para SNRs pequeñas: la mejora relativa abarca desde 41.70%(SNR=10dB) hasta 18.88%(SNR=5dB). De nuevo es la disminución de la tasa de falsas alarmas (Tabla 6.11) la que hace que disminuya el GDE.

Base de datos	TEST_GSM_PREAV	TEST_GSM_POSTAV	TEST_GSM_RUIDONE
FAR (5dB)	**12.01 %**	**13.87 %**	**22.00 %**
FAR (10dB)	**10.50 %**	**13.67 %**	**21.08 %**
FAR (15dB)	**10.19 %**	**12.73 %**	**18.27 %**
FAR (20dB)	**8.45 %**	**12.36 %**	**12.81 %**
FAR (25dB)	**6.82 %**	**11.64 %**	**4.98 %**

Tabla 6.11. Tasa de falsas alarmas (FAR) para TEST_GSM_RUIDONE, TEST_GSM_PREAV y TEST_GSM_POSTAV en el caso de *VAD* + Red Neuronal.

Por tanto, y de forma general se puede decir que el *VAD* que usa como técnica de decisión la RNA funciona mejor que la técnica basada en el árbol del decisión del apartado anterior y, por ende, del método de decisión basado en umbrales. También es importante comentar que las diferencias son mayores para SNRs bajas.

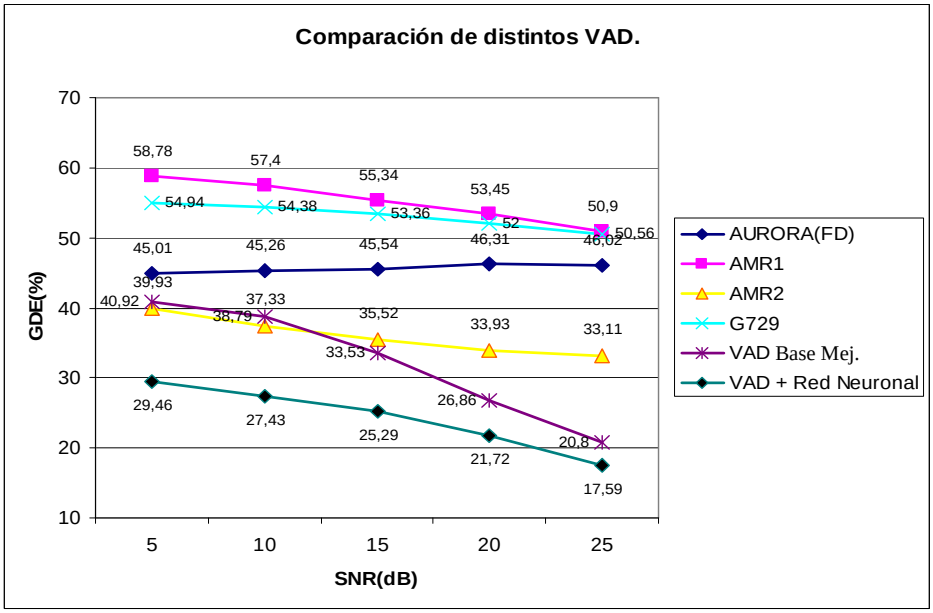

Figura 6.10. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_PREAV (voces de fondo antes de pronunciación). Distintas SNRs.

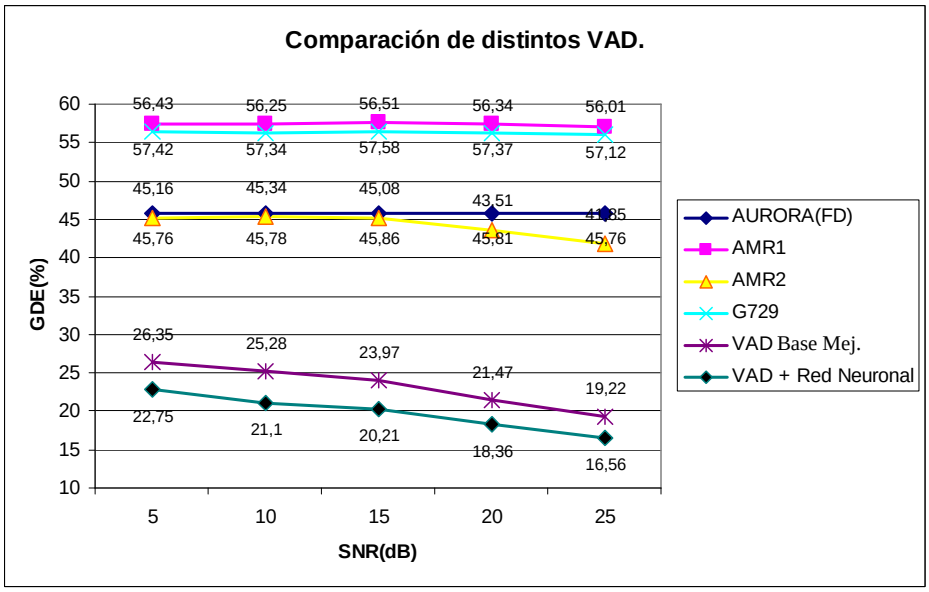

Figura 6.11. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_POSTAV (voces de fondo después de pronunciación). Distintas SNRs.

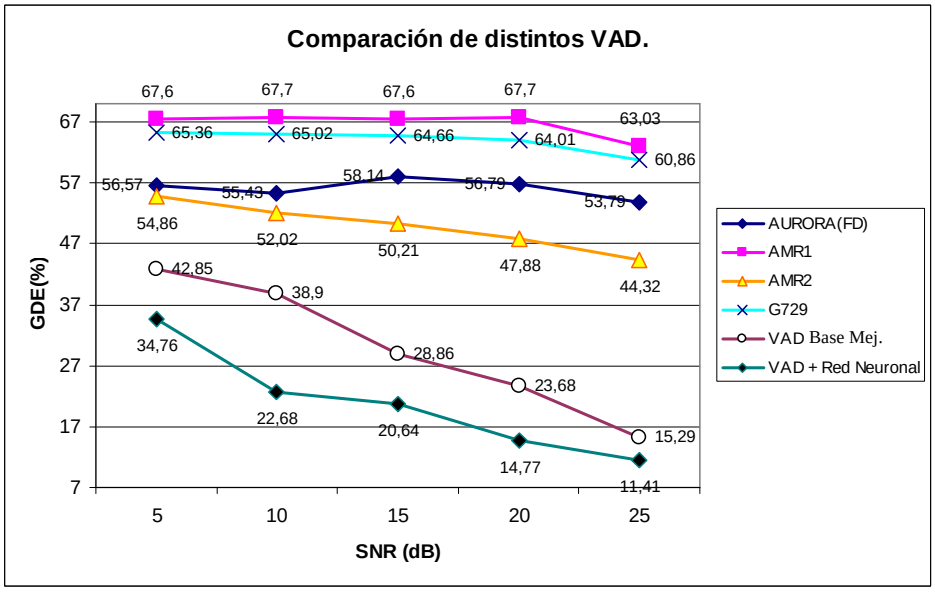

Figura 6.12. Error de Detección Global (*GDE*) para la base de datos TEST_GSM_RUIDONE (servicios reales conversacionales en entornos adversos). Distintas SNRs.

Al igual que en el apartado anterior, es conveniente mostrar los resultados del *GDE* para el caso de TEST_GSM_LIMPIA (voz limpia) y comprobar que el nuevo

algoritmo basado en la RNA mantiene buenos resultados, y cómo se comporta en comparación del resto de detectores estudiados. Los resultados se muestran en la Tabla 6.12.

DET_ERROR	AURORA(FD)	AMR1	AMR2	G729b	VAD Base Mejorado	VAD+Red Neuronal
GDE(%)	35.16	17.97	17.28	29.62	14.27	14.31

Tabla 6.12. Error de Detección Global (GDE) para TEST_GSM_LIMPIA: red neuronal.

Los resultados (*GDE*) de la Tabla 6.12 muestran que el *VAD* basado en la RNA, el *VAD* Base Mejorado, el AMR1 y AMR2 obtienen resultados similares, mientras que el G729b y el AURORA(FD) son peores, con mucho, en comparación con el resto.

Para terminar el apartado, en la Tabla 6.13, se van a exponer también los resultados a nivel de pulso, con las tres bases de datos de voces de fondo que venimos usando, esto es, TEST_GSM_RUIDONE, TEST_GSM_PREAV y TEST_GSM_POSTAV, para poder así tener una idea de cuánto se equivoca la RNA en este caso específico. Se medirá la tasa de aciertos tras la toma de decisión.

Base de datos	TEST_GSM_PREAV	TEST_GSM_POSTAV	TEST_GSM_RUIDONE
Tasa aciertos (5dB)	**76.47**	**79.95**	**67.75**
Tasa aciertos (10dB)	**79.40**	**83.03**	**84.99**
Tasa aciertos (15dB)	**80.30**	**83.78**	**86.17**
Tasa aciertos (20dB)	**84.50**	**87.59**	**93.79**
Tasa aciertos (25dB)	**89.80**	**91.23**	**95.94**

Tabla 6.13. Tasa de aciertos a nivel de pulso de la red neuronal para distintas SNRs y con las tres bases de datos que contienen voces de fondo.

Como se puede observar en la Tabla 6.13, la tasa de aciertos aumenta cuando aumenta la SNR, y aunque para 5dB los peores resultados son para la base de datos basada en servicios conversacionales reales, para el resto de SNRs es la que mejor tasa de aciertos obtiene. En general, la tasa de aciertos, para todos los casos, es algo mejor para el caso de la RNA que para el caso del árbol de decisión. A su vez esto se refleja también para el *GDE* final.

6.6.- Conclusiones.

A continuación, se muestran, en la Tabla 6.14, los resultados finales de detección a nivel de trama (*GDE*) para distintas SNRs con TEST_GSM_RUIDONE, la más relevante dada su naturaleza.

GDE-SNR	AURORA (FD)	AMR1	AMR2	G729 b	VAD Base M.	VAD Umbrales	VAD Árbol	VAD RNA
GDE(%)-5dB	**56,57**	**67,60**	**54,86**	**65,36**	**42,85**	**35,42**	**35,27**	**34,76**
GDE(%)-10dB	**55,43**	**67,70**	**52,02**	**65,02**	**38,90**	**23,54**	**23,07**	**22,68**
GDE(%)-15dB	**58,14**	**67,60**	**50,21**	**64,66**	**28,86**	**21,31**	**21,00**	**20,64**
GDE(%)-20dB	**56,79**	**67,70**	**47,88**	**64,01**	**23,68**	**15,16**	**15,11**	**14,77**
GDE(%)-25dB	**53,79**	**63,03**	**44,32**	**60,86**	**15,29**	**12,92**	**11,55**	**11,41**

Tabla 6.14. Comparación final de todos los detectores para los diferentes métodos de decisión propuestos sobre TEST_GSM_RUIDONE (distintas SNRs).

Es notoria la mejora que se produce, sobre el *VAD* basado en HMMs de partida (*VAD* Base Mejorado), a partir de los tres distintos métodos de decisión propuestos para rechazar los pulsos de voz generados por la voz lejana. El método que mejor funciona es el *VAD* que usa la RNA multicapa, seguido por el que utiliza el árbol de decisión y finalmente el más simple de todos, el basado en umbrales de decisión. A continuación, en la Tabla 6.15, se presenta la mejora relativa de cada uno de los tres métodos sobre el *VAD* Base Mejorado para distintas SNRs y de nuevo sobre TEST_GSM_RUIDONE.

SNR	Mejor relativa Umbrales	Mejora relativa Árbol	Mejora relativa RNA
5dB	**17.34%**	**17.69%**	**18.88%**
10dB	**39.49%**	**40.69%**	**41.70%**
15dB	**26.16%**	**27.81%**	**28.48%**
20dB	**35.98%**	**36.19%**	**37.63%**
25dB	**15.50%**	**24.46%**	**25.38%**

Tabla 6.15. Mejora relativa de los tres métodos de decisión propuestos respecto al caso del *VAD* Base Mejorado para distintas SNRs y sobre TEST_GSM_RUIDONE.

Los *VAD*s usados en codificación (G.729, AMR1 y AMR2) están destinados a transmitir el mínimo número de tramas de ruido posible, así que el punto de trabajo tratará de evitar las falsas alarmas: no es tan importante perder algunas tramas de voz si a cambio evito una alta tasa de falsas alarmas. Sin embargo, el *VAD* del estándar AURORA se usa para reconocimiento automático de habla, por lo tanto sí que será muy importante no perder tramas de voz, y, el punto de trabajo será tal que se prefiera tener una tasa de falsos rechazos baja: esto conlleva el tener un tasa de falsas alarmas alta.

El *VAD* del estándar AURORA es muy completo ya que posee un sofisticado estimador de ruido, usa filtros Wiener y MFCCs, con multitud de umbrales, todos ellos con valores orientados a perder pocas tramas de voz, es decir, poco restrictivos en ese sentido: es por esto que se obtenga una tasa de falsas alarmas alta y como consecuencia un GDE final relativamente elevado. Es por ello que este, a pesar de su sofisticación, tenga una tasa de falsas alarmas elevada en comparación con por ejemplo el AMR2 o el desarrollado en este trabajo. Los códecs G.729 y AMR1 son los más simples y por ello son los que mayor error obtienen:

- El *VAD* del G.729 anexo b no usa MFCCs: usa características sencillas como la tasa de cruces por cero, la energía global o la energía de baja frecuencia. Todo esto hace que sea poco robusto ante ruidos no estacionarios.
- El VAD del códec AMR1, aunque tiene un estimador de ruido sencillo, útil y eficaz para ruidos estacionarios, tampoco usa MFCCs, y, al igual que ocurre con el G.729 anexo b, usa características sencillas y de bajo coste computacional como el “pitch”, las energías en diversas subbandas o la información de la detección de tonos. De nuevo se tiene un detector poco robusto ante ruidos no estacionarios.

Por otro lado, el *VAD* del códec AMR2 es el *VAD* cuyo error de detección más se asemeja en valor al *VAD* propuesto en este trabajo. Este hecho es debido a que, por un lado, el punto de trabajo tiende a ser restrictivo para evitar una tasa de falsas alarmas alta, y por otro que posee un algoritmo de estimación de ruido y cálculo de la SNR de la señal bastante sofisticado y apoyado por el uso de otras características como la energía por subbanda o el cálculo de la métrica de la voz. Es importante remarcar que el *VAD* del códec AMR2 tampoco usa MFCCs dada su no aplicación

en reconocimiento automático de habla. Del *VAD* propuesto en este trabajo se puede decir que en conjunto se trata de un *VAD* bastante sofisticado que contiene un estimador de ruido sencillo y que combina numerosas características (MFCCs, variación de la energía, armonicidad, distancia de Mahalanobis o LPC residual) con un cierto coste computacional. Eligiendo un punto de trabajo óptimo se obtienen los mejores resultados para todos los tipos de ruido existentes, y en especial para los ruidos no estacionarios (voces de fondo).

CAPÍTULO 7

CONCLUSIONES FINALES Y APORTACIONES

Este es el capítulo que pone fin a este trabajo de Tesis doctoral. En él se presentan las conclusiones, se resumen las aportaciones y finalmente se enumeran las posibles líneas futuras de este trabajo de investigación.

Se ha presentado un *VAD* completo, cuyo esquema se presenta gráficamente en la Fig.7.1, especialmente diseñado para rechazar las voces de fondo, y adecuado para integrarse en sistemas de reconocimiento automático de habla, aunque puede ser utilizado en un amplio espectro de aplicaciones.

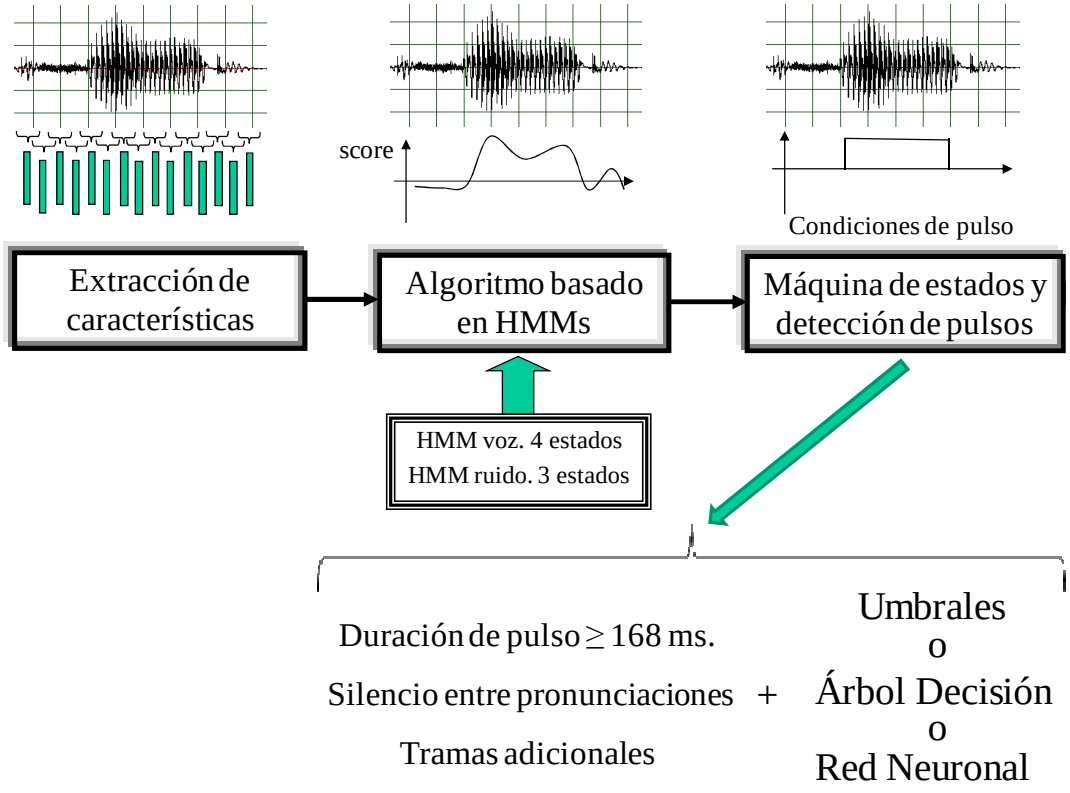

Figura 7.1. *VAD* completo con las nuevas técnicas para el filtrado de voces de fondo.

Los resultados justifican el método, consiguiendo el menor error de detección de todos los detectores de actividad evaluados tanto en entornos de ruido estacionario como de ruido no estacionario. Además, los resultados con voz limpia (SNR superior

a 25 dB) siguen siendo buenos ya que se sigue teniendo una tasa de falsos rechazos moderada.

Las aportaciones específicas obtenidas en este trabajo son las siguientes:

1. Se ha diseñado un *VAD* basado en HMMs, al que llamamos *VAD* Base Mejorado, con algunas mejoras respecto a [82] (*VAD* Base o de partida inicial), cuyo vector de características está formado por:
    a. Los cepstrum que mejor discriminan entre la clase voz y la clase ruido (o clase no-voz): $C1$, $C2$ y $C3$.
    b. La energía normalizada: hace que el *VAD* sea robusto ante niveles altos de ruido y, por ende, robusto ante bajas SNRs.
    c. El delta de energía: mide los cambios de energía y es muy útil en todo proceso de detección.
2. El *VAD* Base Mejorado usa un algoritmo basado en HMMs que obtiene una puntuación con la que se determina si la trama actual es de voz o de ruido. Los modelos usados son dos: uno modela la voz y el otro al ruido. La topología de los HMMs es la siguiente:
    a. El modelo de voz está formado por cuatro estados donde se permiten transiciones de izquierda a derecha y de cada estado consigo mismo. Cada estado está compuesto por una gaussiana.
    b. El modelo de ruido está formado por tres estados donde se permiten transiciones de izquierda a derecha y de cada estado consigo mismo. Cada estado está compuesto por una única gaussiana.
    c. La transición entre modelos es de tal forma que uno realimenta al otro y vice-versa, es decir, cuando termina el modelo de voz prosigue el de ruido o el de voz también y vice-versa.

    Posteriormente y como técnica de post-proceso (decisión a nivel de pulso), mediante un conjunto de reglas, tres en este caso, se logra mejorar los resultados finales detección: duración de pulsos de voz, silencio entre pronunciaciones y tramas adicionales.
3. Se presentan cinco estadísticos capaces de discriminar, a nivel de pulso, entre voz procedente de un locutor principal y voz de fondo procedente de uno o varios locutores:

   a. El porcentaje de tramas con el máximo valor de auto-correlación mayor de 0.9 en un pulso de voz de N tramas.
   b. La mínima distancia de Mahalanobis sobre coeficientes MFCC de tramas consecutivas en un pulso de voz de N tramas.
   c. El porcentaje de tramas con una kurtosis del residuo mayor que 5 en un pulso de voz de N tramas.
   d. El porcentaje de tramas con una auto-correlación máxima del residuo mayor que 0.425 en un pulso de voz de N tramas.
   e. La varianza del máximo de auto-correlación del residuo en un pulso de N tramas.
4. Con el fin de rechazar los pulsos de voz procedentes de locutores de habla lejana, la información sobre las medidas anteriores se introduce en el módulo de detección de pulsos del *VAD* Base Mejorado de tres formas distintas:
   a. En forma de restricciones impuestas por la comparación con umbrales, de los cinco estadísticos, previamente ajustados.
   b. Mediante un Árbol de Decisión de tipo estocástico cuya entrada es un vector formado por los cinco estadísticos.
   c. Mediante una Red Neuronal multicapa (3 capas) o Perceptrón multicapa cuya entrada también es el vector formado por los mencionados cinco estadísticos.

Por otro lado, las conclusiones más relevantes que se desprenden de este trabajo son las siguientes:

1. La inclusión de información espectral al *VAD* de partida inicial [82] basado en HMMs muestra una mejora en los resultados significativa: se añaden tres cepstrum, $C1$, $C2$ y $C3$.
2. El uso de la energía normalizada en el vector de características del punto anterior hace que el *VAD* sea invariante ante una variación de la SNR (umbral fijo).
3. La topología óptima para los HMMs es la siguiente: 4 estados para el modelo de voz y 3 estados para el modelo de ruido, y una gaussiana por estado. Aumentar el número de estados no mejora los resultados. Estos HMMs,

además, funcionan correctamente en todas las redes telefónicas: telefonía móvil (gsm), telefonía fija y voz IP.

4. Los resultados a nivel de trama (máxima verosimilitud entre dos clases: la clase voz y la clase ruido) mejoran si se introduce información de la estructura del habla: duración de pulsos, silencio entre pronunciaciones y tramas adicionales. Se trata de una decisión a nivel de pulso de voz: se crean los pulsos de voz. En este momento se crea lo que llamamos el *VAD* Base Mejorado, que posee todas las características enunciadas en estos 4 puntos.
5. El *VAD* Base Mejorado obtiene buenos resultados para voz limpia y voz contaminada por ruidos estacionarios. Sin embargo, en el caso de voces de fondo (ruidos no estacionarios), la tasa de falsas alarmas hace que los resultados no sean tan buenos. Aún así, en comparación con otros detectores de referencia, obtiene en general los mejores resultados tanto para ruidos estacionarios como para ruidos no estacionarios.
6. Con el fin de solucionar el problema de las falsas alarmas, provocadas por las voces de fondo que genera el *VAD* Base Mejorado, se realiza el estudio de diversas características: Armonicidad, Distancia de Mahalanobis entre coeficientes MFCC de tramas consecutivas y un LPC residual de orden 10.
7. Del estudio de las características del punto anterior, los cinco estadísticos sobre las mismas que mejor funcionaron fueron los siguientes:
    a. El porcentaje de tramas con el máximo valor de auto-correlación mayor de 0.9 en un pulso de voz de N tramas.
    b. La mínima distancia de Mahalanobis sobre coeficientes MFCCs de tramas consecutivas en un pulso de voz de N tramas.
    c. El porcentaje de tramas con una kurtosis del residuo mayor que 5 en un pulso de voz de N tramas.
    d. El porcentaje de tramas con una auto-correlación máxima del residuo mayor que 0.425 en un pulso de voz de N tramas.
    e. La varianza del máximo de auto-correlación del residuo en un pulso de N tramas.
8. Con el fin de rechazar los pulsos de voz procedentes de locutores de habla lejana, la información sobre las medidas anteriores se introduce en el módulo

de detección de pulsos del *VAD* Base Mejorado (decisión a nivel de pulso) de tres formas distintas:

a. En forma de restricciones impuestas por la comparación con umbrales, de los cinco estadísticos, previamente ajustados.
b. Mediante un Árbol de Decisión de tipo estocástico cuya entrada es un vector formado por los cinco estadísticos.
c. Mediante una Red Neuronal multicapa (3 capas) o Perceptrón multicapa cuya entrada también es el vector formado por los mencionados cinco estadísticos.

9. Los tres métodos propuestos logran mejoras significativas respecto del ya sofisticado *VAD* Base Mejorado. Los mejores resultados los obtiene el *VAD* basado en la red neuronal, seguido por el que usa el árbol de decisión y finalmente por el VAD que utiliza umbrales, aunque con diferencias poco significativas.
10. Aunque el *VAD* que usa umbrales de decisión es el método que consigue mejoras más pequeñas de los tres, es importante destacar tanto su fácil y cómoda integración en cualquier sistema de detección como el bajo consumo de tiempo de ejecución que genera.

7.1.- Difusión y publicaciones.

En cuanto a la difusión y publicaciones derivadas de esta Tesis se pueden enumerar las siguientes:

- Revistas indexadas:
    1. Varela O., San-Segundo R. and Hernandez L., “Combining pulse-based features for rejecting far-field speech in a HMM-based Voice Activity Detector”, Computers & Electrical Engineering, vol. 37, Issue 4, pp. 589-600. July 2011.
    2. Varela O., San-Segundo R. and Hernandez L., "Robust Speech Recognition System for Air Traffic Control in noisy environments" IEEE Aerospace & Electronic Systems Society. Aceptada.
    3. Varela O., San-Segundo R. and Hernandez L., “Robust Voice Activity Detection for rejecting background voices in telephone

conversations", International Journal of Computer Systems Science & Engineering. En proceso de revisión.

4. Varela O., San-Segundo R. and Hernandez L., "Using a Neural Network for improving Voice Activity Detection response in non-estationary noice environments". International Journal of Computational Intelligence Systems. En proceso de revisión.

- Comunicaciones a congresos:
  1. Varela, O., San-Segundo, R., Hernandez, L., "Nuevos parámetros acústicos para la clasificación de voz cercana, voz lejana y voz procedente de varios locutores", Telecom I+D, Octubre 2008, Bilbao, Spain.
  2. Varela, O., San-Segundo, R., Hernandez, L., 'New features for improving VAD when dealing with far-field and multi-speaker speech', Jornadas de Tecnologias del Habla, Noviembre 2008, Bilbao, Spain.
- Informes técnicos:
  1. Reconocimiento y Multicodificación fase I. Telefónica Investigación y Desarrollo. Año 2001.
  2. Reconocimiento y Multicodificación fase II. Telefónica Investigación y Desarrollo. Año 2002.
  3. Ajuste de los Parámetros del Detector de Extremos para E-moción Voz. Telefónica Móviles España. Año 2002.
  4. Ajuste de los Parámetros del Detector de Extremos a las Pronunciaciones "si" y "no" en Servicios de Telefonía Móvil. Año 2002.
  5. Tecnología del Habla para Desarrollo de Servicios de Telefónica Móviles España. Apartado 1.1.2: Técnicas Avanzadas del Detector de Extremos. Año 2002.
  6. Incorporación de las Tecnologías del Habla en el Interior de Vehículos para Telefónica Móviles España. Año 2003.
  7. Técnicas de Robustez para Reconocimiento de Voz para Telefonica Móviles España (comparación con el estándar AURORA). Año 2003.

7.2.- Líneas futuras.

Para finalizar este capítulo, y con él este trabajo, es importante enunciar las posibles líneas futuras describiendo a su vez brevemente cada una de las mismas:

- Estudio de la eficiencia computacional. Se abordaría la optimización computacional y evaluación del tiempo de proceso que lleva cada uno de los módulos que forman de *VAD* completo (Fig.7.1) y a su vez compararlo con los de referencia. Una forma posible de medir estos tiempos de proceso es mediante el factor Unidad de Tiempo Real o RTU (del inglés Real Time Unit).
- Implantación en sistemas de reconocimiento multimodal. En este caso se estudiarían las mejoras que podrían alcanzarse combinando los resultados obtenidos por el *VAD* con los resultados de otros sistemas que añaden información adicional al primero. Un ejemplo sería la información generada por un reconocedor de caras: gesticular, mover la boca, etc.
- Incorporación de nuevas características para el rechazo de las voces de fondo. Se pueden seguir añadiendo características en el módulo de detección de pulsos para rechazar las voces de fondo, como por ejemplo información de la Normalización del Tracto Vocal (VTLN del inglés Vocal Tract Length Normalization). En general la incorporación de nuevas características depende directamente de la aplicación para la que se use el *VAD*, ya que, en muchos casos, el incorporar una nueva característica que no trate el sistema o la aplicación puede suponer aumentar demasiado el tiempo de proceso.
- Aplicación en sistemas de transcripción automática y diarización. El *VAD* creado puede ser utilizado para la búsqueda de locutores específicos, o para saber cuándo están hablando varios locutores de forma simultánea. Esta información puede ser muy útil por ejemplo en sistemas de transcripción automática de reuniones, ponencias, cursos etc.
- Aplicación en sistemas multicanal. También se puede usar la información generada por el *VAD* para controlar la directividad de un array de micrófonos y si el micrófono es móvil, poder orientarle para conseguir la

menor reverberación posible y la máxima directividad. En caso opuesto, si se detectan voces de fondo, se podría buscar una mayor cancelación del audio recogido en esa dirección.

- Técnicas discriminativas para los HMMs del *VAD*. Aplicar técnicas de entrenamiento discriminativo para generar los modelos HMMs de voz y "no voz", así como en la selección de características que forman el vector de los modelos como se ha hecho en este trabajo.
- Inclusión de la estructura melódica en el *VAD*. Se trataría de incluir información relativa a la estructura melódica del habla, como por ejemplo información de la prosodia o información de la entonación.
- Entrenamiento de modelos específicos para distintos tipo de ruido. Esto implicaría la obtención de diferentes modelos de ruido, por ejemplo modelo de ruido de coche, modelo de ruido de la calle, modelo de voces de fondo, etc. En este caso sería muy importante el algoritmo de elección del modelo en función de la aplicación o el algoritmo que se use para la toma de decisión cuando compitan los modelos conjuntamente.
- Detección de fuentes de ruido como la música. En algunas ocasiones puede ser interesante detectar si existe ruido de música de fondo o de cualquier otro tipo específico para identificar de forma precisa el entorno acústico del usuario de un sistema interactivo.

BIBLIOGRAFÍA

[1] J. Ramirez, Efficient voice activity detection algorithms using long-term speech information. In Speech Communication, 42, pp. 271 – pp. 287, 2004.

[2] Freeman, D.K., Cosier, G., Southcott, C.B., Boyd, I., 1989.The voice activity detector for the PAN-European digital cellular mobile telephone service. In: Internat. Conf. On Acoust. Speech Signal Process., Vol. 1, pp. 369–372.

[3] ITU-T recommendation G.729-Annex B, 1996. A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

[4] Sohn, J., Sung, W., 1998. A voice activity detector employing soft decision based noise spectrum adaptation. In: Internat. Conf. on Acoust. Speech Signal Process., Vol. 1, pp. 365–368.

[5] ETSI EN 301708 recommendation, 1999. Voice activity detector (VAD) for adaptive multi-rate (AMR) speech traffic channels.

[6] Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. 10 (2), 109–118.

[7] Sangwan, A., Chiranth, M.C., Jamadagni, H.S., Sah, R., Prasad, R.V., Gaurav, V., 2002. VAD techniques for realtime speech transmission on the Internet. In: IEEE Internat. Conf. on High-Speed Networks and Multimedia Comm., pp. 46–50.

[8] Karray, L., Martin, A., 2003. Towards improving speech detection robustness for speech recognition in adverse environment. Speech Comm. 40 (3), 261–276.

[9] Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Internat. Conf. On Acoust. Speech Signal Process., pp. 208–211.

[10] Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

[11] Filiz Basbug, Member IEEE, Kumar Swaminathan, Senior Member IEEE and Srinivas Nandkumar. “Noise Reduction and Echo Cancellation Front-End for Speech codecs”. IEEE Trans. Vol. 11.Nº 1. January 2003.

[12] M. H. Savoji. “A robust algorithm for accurate endpointing of speech signals”. Speech Communication. Nº 8, 1989. Pp. 45-60.

[13] AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking.

[14] José C. Segura, Javier Ramírez, CarmenBenítez, Ángel de la Torre, Antonio Rubio, “Improved feature extraction based on spectral noise reduction and nonlinear feature nonlinear feature normalization”, Eurospeech 2003, pp. 353-356.

[15] M. Marzinzik and B. Kollmeier, “Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics”, IEEE Trans. on Speech and Audio Proc., Vol. 10, No. 2, February 2002.

[16] R. Martin, “An efficient algorithm to estimate the instantaneous SNR of speech signals”, in Proc Eurospeech´93, Vol. 1, 1993.

[17] R. Martín, “Spectral subtraction based on minimum statistics”, in Signal Processing VII, Theories and Applications. Proceedings of EUSIPCO-94, Vol. 1, M. J. J. Holt, C. F. N. Cowan, P. M. Grant, and W. A. Sandham, Eds. Lausanne, Switzerland, 1994.

[18] J. A.Haigh & J. S. Mason, “Robust Voice Activity Detection using Cepstral Features”, University College Swansea.

[19] L. R. Rabiner, M. R. Sambur. “An algorithm for determining the endpoints of isolated utterances”. The Bell System Technical Journal. Volumen 54, nº 2, Febrero 1975. Pp. 297-315.

[20] Izhak Shafran & Richard Rose, “Robust Speech Detection and Segmentation for Real-Time ASR Applications”, ICASSP 2003, pp. 432-434.

[21] Special Mobile Group (GSM). Technical Commite of the European Telecommunications Standards Institute (ETSI). “Voice Activity Detector”. ETSI Secretariant. France. Enero 1991.

[22] D. B. Paul, “The spectral envelope estimation vocoder”, IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, pp. 786-794, April 1981.

[23] H. G. Hirsch, “Estimation of noise spectrum and its application to SNR-estimation and speech enhancement”, Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. TR-93-012,1993.

[24] L. R. Rabiner and M. R. Sambur, “An algorithm for determing the endpoints of isolated utterances”, The Bell System Technical Journal, Vol. 54, No. 2, pp. 267, February 1975.

[25] H. H. Lee and C. K. Un, “A study of On-Off characteristics of conversational Speech”, IEEE Transactions on Communications, Vol. COM-34, No. 6, pp. 630, June 1986.

[26] J. A. Jankowski, “A new digital voice-activated switch”, Comstat Technical Review, Vol. 6, No. 1, pp. 159, June 1976.

[27] G. S. Kang and L. J. Fransen, “Quality improvement of LPC-processed noisy speech by using spectral subtraction”, IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, pp. 930-942, June 1989.

[28] S. Van Gerven and F. Xie, “A comparative study of speech detection methods”, in Proc. 5th Eur. Conf. Speech Communication Technology, EUROSPEECH´97, Rhodes, Greece, 1997.

[29] C. Elberling, C. Ludvigsen, and G. Keidser, “The design and testing of a noise reduction algorithm based on spectral subtraction”, Scand. Audiol., Vol. Suppl. 38, pp. 39-49, 1993.

[30] H. Sheikhzadeh, R. L. Brennan, and H. Sameti, “Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications”, in Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing 1995, pp. 808-811.

[31] M. Dendrinos and S. Bakamidis, “Voice activity detection in clored-noise environment through singular value descomposition”, in Proc. 5th Int. Conf. Signal Processing Applications and Technology. Waltham, MA: DSP Associates, 1994, Vol.1, pp. 137-141.

[32] E. Cornu, H. Sheikhzadeh, R. L. Brennan, H. R. Abutalebi, “ETSI AMR-2 VAD: Evaluation and ultra low-resource implementation”.

[33] K. Itoh and M. Mizushima, “Environmental noise reduction based on speech/nonspeech identification for hearing aids,” in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing 1997, Conference Proceedings. Los Alamitos, CA: IEEE Comput. Soc. Press, 1997, pp. 419–422.

[34] The HTK Book (for HTK Version 3.1). Página 60.

[35] C. Lee, C. Lin and B. Juang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models”, IEEE Trans. Signal Processing, 39 (4): 806-814, 1991.

[36] Andre Adami, Lukas Burget. “Qualcomm-Icsi-Ogi features for ASR”. Qualcomm Inc., San Diego, California, USA.

[37] G. Doblinger, “Computationally efficient speech enhancement by spectral minima tracking in subbands “, in Proc. 4th Eur. Conf. Speech Communication Technology EUROSPEECH´95. Madrid, Spain, Sept. 1995, pp. 1513-1516.

[38] H. G. Hirsch and C. Ehrlicher, “Noise estimation techniques for robust speech recognition”, in Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing 1995, Vol. 1, 1995, pp. 153-156.

[39] A. Fischer and V. Stah, “On improvement measures for spectral subtraction applied to robust automatic speech recognition in car environments”, in Proc. Workshop Robust Methods Speech Recognition Adverse Conditions, Tampere, Finland, May 1999, pp. 75-78.

[40] E. Nemer, R. Goubran, and S. Mahmoud, “SNR estimation of speech signals using subbands and fourth-order statistics”, IEEE Signal Processing Lett., Vol 6, pp. 171-174, July 1999.

[41] P. Sovka and P. Pollák, “The study of speech/pause detectors for speech enhancement methods”, in Proc. 4th Eur. Conf. Speech Communication Technology EUROSPEECH´95. Madrid, Spain: ESCA, September 1995, pp. 1575-1578.

[42] Oh-Wook Kwon and Te-Won Lee, “Optimizing speech/non-speech classifier design using adaboost”, ICASSP 2003, pp. 436-438.

[43] S.S. Kajarekar and H. Hermansky. “Analysis of information in speech and its application in speech recognition”, TSD´2000, Brno, Czech Republic, September 2000.

[44] Ye Tian, Ji Wu, Zuoying Wang, and Dajin Lu, “Fuzzy clustering and bayesian information criterion based threshold estimation for robust voice activity detection”, IEEE Trans., ICASSP 2003, pp. 444-447.

[45] Yong Duk Cho, Khaldoon Al-Naimi, and Ahmet Kondoz, “Improved voice activity detection based on a smoothed statistical likelihood ratio”, University of Surrey.

[46] C. Crespo Casas, C. De la Torre Munilla, J. C. Torrecilla Merchán. “Detector de extremos para reconocimiento de voz”. Comunicaciones de Telefónica I+D. Volumen 5, Nº 2, julio-diciembre 1994. Pp. 89-96.

[47] Françoise Beaufays, Daniel Boies, Mitch Weintranb, Qifeng Zhu. “Using speech/non speech detection to bias recognition search on noisy data”, IEEE Proc. ICASSP 2003, pp. 424-427.

[48] Katidiotisa, A., Tsagkaris, K., and Demestichasa, P., “Performance evaluation of artificial neural network-based learning schemes for cognitive radio systems”, Computers & Electrical Engineering, vol. 36, pp. 518-535, May 2010.

[49] Nabavi-Kerizia, S.H., Abadia, M., and Kabir, E., “A PSO-based weighting method for linear combination of neural networks”, Computers & Electrical Engineering, vol. 36, pp. 886-894, September 2010.

[50] Lingyun Gu, Jianbo Gao and John G. Harris, “Endpoint detection in noisy environment using a Pointcaré Recurrence Metric”, ICASSP 2003, pp. 428-431.

[51] J. B. Gao*, Detecting Nonstationary and State Transitions in a Time Series*. In Physical Review E. Vol. 63, pp. 0062021 – pp. 0662028, May, 2001.

[52] J. B. Gao, Recurrence Time Statistics for Chaotic Systems and Their Applications. In Physical Review Letters, Vol. 83, No. 16, pp. 3178 – pp. 3181, Oct, 1999.

[53] http://spib.rice.edu/spib/select_noise.html

[54] Javier Ramirez, José C. Segura, 2005. "Statistical Voice Activity Detector Using a Multiple Observation Likelyhood Ratio Test". IEEE Signal Process. Vol 12, nº 10.

[55] Venkata R. Gadde, Andreas Stolcke, Dimitra Vergyri, Jing Zheng, Kemal Sonmez, and Anand Venkatraman, "Building an ASR system for noisy environment: SRI´s 2001 SPINE evaluation system", Proc. Int´l Conf. on Spoken Lanuage Processing (ICSLP), 2002.

[56] Brian Kingsbury, George Saon, Lidia Mangu, Mukund Padmanabhan, and Ruhi Sarikaya, "Robust speech recognition in noisy environments: The 2001 IBM SPINE evaluation system", Proc. Int´l Conf. on Acoustic, Speech and Signal Processing, 2002.

[57] Murat Saraclar, Michael Riley, Enrico Bocchieri, and Vincent Goffin, "Towards automatic closed captioning: low latency real time broascast news transcription", Proc. Int´l Conf. on Spoken Language Processing (ICSLP), 2002.

[58] J. C. Spohrer, P. F. Hochschild, and J. K. Baker, "Partial backtrace in continuous speech recognition", Proc. Int´l Conf. on Systems, Man, and Cybernetics, pp. 36-42, 1980.

[59] J. G. Wilpon, L. R. Rabiner, T. Martín. “An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints”. AT&T Bell Laboratories Technical Journal. Volumen 63, Nº 3, marzo 1984. Pp. 479-498.

[60] S. G. Tanyer and H. Ozer, “Voice activity detection in nonstationary noise”, IEEE Trans. Acoust., Voice, Signal Processing, v8, pp. 478-482, Jul. 2000.

[61] L. F. Lamel, L. R. Rabiner, A. E. Rosemberg, and J. G. Wilson, “An improved endpoint detector for isolated word recognition”, IEEE Trans. Acoust., Voice, Signal Processing, v29, pp. 777-785, Aug. 1981.

[62] J. L. Shen, J.W. Hung, and L. S. Lee, “Robust entropy based endpoint detection for voice recognition in noisy environments”, in Proc. ICSLP´96, 1996.

[63] K. H. Woo, T. Y. Yang, K. J. Park, and C. Y. Lee, “Robust voice activity detection algorithm for estimating noise spectrum”, Electronics Letters, v36, pp. 180-181, Jan. 2000.

[64] Hartigan, J., & Wang, M. (1979). “A K-means clustering algorithm” *Applied Statistics*, 28, 100–108.

[65] Y. Jiang, Z.H. Zhou. Editing training data for kNN classifiers with neural network ensemble. I International Symposium on Neural Networks (ISNN'04). Lecture Notes in Computer Science 3173, No Data 2004, Dalian (China, 2004) 356-361

[66] Khaled El-Maleh and Peter Kabal. “Comparison of voice activity detection algorithms for wireless personal communications systems”. Proc. IEEE pp. 470-473. May 1997.

[67] J. Zhang, W. Ward, B. Pellom, X. –Yu and K. Hacioglu, “Improvements in Audio Processing and Language Modeling in the CU Communicator”, Eurospeech, 3: 2209-12, Denmark, 2001.

[68] M. Marzinzik, B. Kollmeier, “Speech pause detection for noise spectrum estimation by tracking power envelope dynamics”, In IEEE Trans. Speech and Audio Processing, vol. 10, Nº 2, pp. 109-110, Feb. 2002.

[69] Itoh, K., Mizushima, M., 1997. Environmental noise reduction based on speech/non-speech identification for hearing aids. In: Internat. Conf. on Acoust. Speech Signal Process., Vol. 1, pp. 419–422.

[70] Cho, Y.D., Kondoz, A., 2001. Analysis and improvement of a statistical model-based voice activity detector. IEEE Signal Process. Lett. 8 (10), 276–278.

[71] Luca Armani, Marco Mattassoni, Maurizio Ornologo, Piergiorgio Sraizer. “Use of a CSP-based voice activity detector for distant-talking ASR”, EUROSPEECH 2003, GENEVS. Pp. 501-504.

[72] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon. “An improved endpoint detector for isolated word recognition” IEEE Transactions on Acoustics, Speech, and Signal Processing. Volumen ASSP-29, Nº 4, junio 1981. Pp. 777-785.

[73] Y. Gong, “Speech Recognition in Noisy Environments: A Survey”, Speech Communication, 16, pp. 261-291, 1995.

[74] Y. M. Cheng, D. Macho, Y. Wei, D. Ealey, H. Kelleher, D. Pearce, W. Kushner, T. Ramabadran, “A Robust Front-End for Distributed Speech Recognition”, Proc. Eurospeech´01.

[75] B. Andrassy, D. Vlaj, Ch. Beaugeant, “Recognition Performance of the Siemens Front-End with and without Frame Dropping on the Aurora 2 Database”, Proc. Eurospeech´01.

[76] C. Benitez, L. Burger, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, S. Sildavas, “Robust ASR Front-End Using Spectral-Based and Discriminant Features: Experiments on the Aurora Tasks”, Proc. Eurospeech´01.

[77] ETSI ES 202 050 V1.1.1 (2002-07). ETSI Standard.

[78] R. J. McAulay and M. L. Marpass, “Speech enhancement using a soft-decision noise suppresion filter”, IEEE Trans., ASSP, Vol. 28, Nº 2, pp. 137-147, April 1980.

[79] J. Roberts, “Modification of piecewise LPC”, MITRE Working Paper WP-21752, May 1978.

[80] J. Ramírez, Student Member, IEEE, J. C. Segura, Senior Member, IEEE, C. Benítez, Member, IEEE, A. Torre, and A. J. Rubio, Member, IEEE, “A New Kullback–Leibler VAD for Speech Recognition in Noise”, IEEE Signal Processing Letters, Vol. 11, No. 2, February 2004. pp. 266-269.

[81] H. Sheikhzadeh, R. L. Brennan and H. Sameti, “Real-Time implementation of HMM-Based MMSE Algorithm for speech enhancement in hearing aid applications”, 1995 IEEE. Pp. 808-811.

[82] A. Acero, C. Crespo, C. Torre and J. C. Torrecilla, “Robust HMM-Based endpoint detector”, EUROSPEECH 1993, 1551-1554.

[83] Qi Li, Jingsong Zheng, Augustine Tsai and Qiru Zhou, *Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker*

Recognition, in IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 3, pp. 146 – pp. 157, Mar, 2002.

[84] Jianping Zhang, Wayne Ward and Bryan Pellom. Center for spoken Language Research University of Colorado at Boulder Boulder. “Phone based voice activity detection using online bayesian adaptation with conjugate normal distributions”. Colorado 80309-0594, USA.

[85] B. S. Atal, L. S. Rabiner. “A pattern recognition approach to voiced-unvoiced-silence classification with aplications to speech recognition”. IEEE Transactions on Acoustics, Speech, and Signal Processing. Volumen ASSP-24, Nº 3, junio 1976. Pp. 201-212.

[86] Yang, J., Yu, S., Zhou, J. and Gao, Y., “A new error concealment method for consecutive frame loss based on CELP speech”, Computers & Electrical Engineering, vol. 36, pp. 1014-1020, September 2010.

[87] Yue, W. and Zheng, B., “Spectrum sensing algorithms for primary detection based on reliability in cognitive radio systems”, Computers & Electrical Engineering, vol. 36, pp. 469-479, May 2010.

[88] AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking.

[89] J. Sohn, Student Member, IEEE, N. S. Kim, Member, IEEE, and W. Sung, “A Statistical Model-Based Voice Activity Detection”, IEEE Signal Processing Letters, Vol. 6, No. 1, January 1999. pp 1-3.

[90] P. Tarapeik, J. Labuda and B. Fourest, “Measurement uncertainty distributions and uncertainty propagation by the simulation approach”, 3rd EURACHEM Workshop, September 1999, Bratislava.

[91] HTK Speech Recognition Toolkit V3.1.

[92] Petropulu, A. P., and Subramaniam, S.,"Cepstrum based deconvolution for speech dereverberation", IEEE Trans. Speech and Audio Proc. pp. 9-12,1994.

[93] Nakatani, T. and Miyoshi, M.,“Blind dereverberation of single channel speech signal based on harmonic structure”, pp. 92-95, ICASSP 2003.

[94] Ohta, K. and Yanagida, M., “Single channel blind dereverberation based on auto-correlation functions of frame-wise time sequences of frecuency components”, Iwaenc 2006 – Paris – September 12-14, 2006.

[95] Bees, D., Kabal, P., and Blostein, M., “Application of complex cepstrum to acoustic dereverberation”, Proc. Biennial Symp. Commun. (Kingston, ON), pp. 324-327, June 1990.

[96] Yegnanarayana, B., Mahadeva Prasana, S. R., Duraiswami, R. and Zontkin, D., “Processing of Reverberant Speech for Time-Delay Estimation”, IEEE Trans. Speech and Audio Proc., pp. 1110-1118, vol. 13, nº 6, November 2005.

[97] Cournapeau, D. And Kawahara, T., “Evaluation of Real-Time Activity Detection based on High Order Statistics”, pp. 2945-2948, Interspeech 2007.

[98] Varela, O., San-Segundo, R., Hernandez, L., ‘New features for improving VAD when dealing with far-field and multi-speaker speech’, Jornadas de Tecnologías del Habla, 2008, Bilbao, Spain.

[99] Yuan Y. and Shaw M.J., “Induction of fuzzy decision trees”, Fuzzy Sets and Systems, vol. 69, pp.125-139, January 1995.

[100] Breiman L., Friedman J.H., Olshen R. A., Stone C.J., “Classification and Regression Trees” Ed. Wadsworth & Brooks/Cole advanced books & software. 1984.

[101] Yu-xin K., Xiao-ning J. and Hang Y., “Voice Activity Detection Algorithm Based on RASTA and SVM” Microcomputer Information. 2009.

[102] Gemello R., Mana F. and Mori R., “Non-Linear Estimation of Voice Activity to Improve Automatic Recognition of Noisy Speech” Interspeech'2005 – Eurospeech. 4-8 September. 2005.

[103] Padmanabhan R., Sree Hari Krishnan P. and Murthy A., “A pattern recognition approach to VAD using modified group delay”. In: National Conference on Communications: NCC-2008.

[104] Prasad R., Sangwan A., Jamadagni H. and Chiranth M., “Comparison of voice activity detection algorithms for voip”, in Proc. IEEE Symposium on Computer and Communications, vol. 5, July 2002, pp. 530-535.

[105] Aguirre J., Álvarez R., Sánchez J. and Zamora A., “Silence Detection in Secure P2P VoIP Multiconferencing”, proceedings of the 5th WSEAS Int. Conference on Information Security and Privacy, Venice, Italy, November 20-22, 2006.

[106] ETSI TS 126 094 V4.0.0 (2001-03).

[107] Skorik S. and Berthommier F., “On a Cepstrum-Based Speech Detector Robust To White Noise,” Specom 2000, St. Petersbourg.

[108] Broggi C. J., Goujon D. J. and Herrmann R. A., “Comparación de Redes Neuronales Utilizados en Sistemas de Soporte de Decisiones”, Universidad Tecnológina Nacional, Facultad Regional Resistencia, Córdoba (Argentina), Mayo de 2007.

[109] Yeoun J. and Hahn M., “Automatic Assessment of Pathological Voice Quality Using Higher-Order Statistics in the LPC Residual Domain” EURASIP Journal on Advances in Signal Processing (2009).