UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS
VICERRECTORÍA DE FORMACIÓN
DIRECCIÓN DE FORMACIÓN POSTGRADUADA

DETECCIÓN DE ANUNCIOS EN EMISIONES DE TELEVISIÓN BASADA EN LA CARACTERIZACIÓN DEL AUDIO

Tesis presentada en opción al título de Máster en Informática Aplicada

Autor: **MsC. Ruber Hernández García**

Tutor: **Dr. Julián Ramos Cózar**

La Habana, Noviembre de 2011

A mi adorada familia, Daisyta y Dialenys...

A mis padres, hermanos y amigos...

« Intenta no volverte un hombre de éxito,

sino volverte un hombre de valor. »

Albert Einstein (1879-1955)

DECLARACIÓN JURADA DE AUTORÍA Y AGRADECIMIENTOS

Yo Ruber Hernández García, con carné de identidad 83011219485, declaro que soy el autor principal del resultado que expongo en la presente memoria titulada “Detección de anuncios en emisiones de televisión basada en la caracterización del audio”, para optar por el título de Máster en Informática Aplicada.

El presente trabajo fue desarrollado individualmente en el transcurso de los años 2009-2010.

En especial deseo agradecer a los profesores Dr. Julián Ramos Cózar y Dr. Nicolás Guil Mata de la Universidad de Málaga, España, por su constante apoyo y orientación en los trabajos de investigación. Además, deseo agradecer al Departamento Arquitectura de Computadores de la Universidad de Málaga, por poner a mi disposición los medios materiales necesarios para el desarrollo de este trabajo. En lo personal, agradezco a mis amadas esposa e hija, Daisyta y Dialenys, por su apoyo incondicional en todo momento y por permitirme robarle tiempo a ellas para dedicarlo a este trabajo sin reprochármelo jamás; así como a otros colegas y amigos que no menciono por razones de espacio, mi más sincero agradecimiento .

Finalmente declaro que todo lo anteriormente expuesto se ajusta a la verdad, y asumo la responsabilidad moral y jurídica que se derive de este juramento profesional.

Y para que así conste, firmo la presente declaración jurada de autoría en La Habana a los ___ días del mes de _____ del año 2011.

_____

Firma del maestrante

RESUMEN

La televisión (TV) como medio de transmisión es una parte fundamental en la vida diaria de millones de hogares en todo el mundo. Numerosas compañías y empresas explotan dicha potencialidad como una importante herramienta de marketing con la transmisión de informaciones en forma de publicidad (anuncios).

La detección de anuncios en una señal de vídeo es un problema crucial que está altamente relacionado con técnicas de procesamiento de vídeo como la segmentación, indexación y recuperación; y en este mismo orden se puede decir que esta problemática en los últimos años ha logrado la atención de la comunidad científica e industrial.

Con anterioridad se han planteado varias técnicas para la detección de anuncios teniendo en cuenta solamente la señal de vídeo o la combinación de audio y vídeo, destacándose aquellas basadas en sistemas híbridos que explotan características tanto del audio como el vídeo. Sin embargo el desafío de todas las técnicas desarrolladas es el mismo: cómo detectar correctamente intervalos de publicidad; y cómo realizar un proceso en el menor tiempo posible.

El presente trabajo se centrará en la detección de anuncios con el objetivo de aumentar las capacidades de almacenamiento de los sistemas de monitoreo e indexación de emisiones de televisión. Durante el desarrollo del trabajo se presentará la propuesta de una técnica para la detección de bloques de publicidad en emisiones televisivas utilizando solo la señal de audio, teniendo en cuenta el estudio y aplicación de diversas técnicas existentes que se utilizan para resolver problemas similares de manera eficiente.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La televisión (TV) como medio de transmisión es una parte fundamental en la vida diaria de millones de hogares en todo el mundo; debido a su forma de atraer al público ha logrado ser durante años el medio de preferencia mundial. Numerosas compañías y empresas explotan dicha potencialidad como una importante herramienta de marketing con la transmisión de informaciones en forma de publicidad (anuncios). Los anuncios normalmente son insertados en la emisión de televisión en forma de bloques de publicidad o cortes de comerciales, de los cuales tanto televidentes como anunciantes no pueden tener un control exacto de su transmisión.

Como se sugiere en [39] se puede decir que existen dos grupos fundamentales interesados en la detección de anuncios en emisiones de televisión. El primero, formado por los anunciantes, desean identificar, monitorear y comprobar que sus comerciales son transmitidos en el momento y con el volumen correcto (identificación de anuncios). Verificando de esta manera que sus contratos con las televisoras sean cumplidos, teniendo en cuenta que el precio de un anuncio depende primeramente de la popularidad del programa de televisión que se interrumpe con el corte comercial, por tanto el dueño se asegurará de que esto sea cumplido.

El otro grupo necesita detectar los bloques de publicidad ya sea para eliminarlos de las grabaciones de televisión o para verificar el cumplimiento de determinadas regulaciones o normas, de esta manera desean conocer cuántos anuncios diferentes (y sus características) son transmitidos durante un determinado período de tiempo (detección de anuncios). Dentro de este grupo se encuentran los televidentes que quieren ver sus programas de televisión grabados sin cortes de publicidad. De igual manera se encuentran los sistemas de bases de datos de vídeo, que requieren almacenar las emisiones de televisión sin anuncios con el objetivo de disminuir el espacio requerido para el almacenamiento de los materiales audiovisuales. [39]

Antecedentes y estado actual del tema

La indexación automática de contenidos audiovisuales es una tecnología emergente en el campo de la informática, la cual cuenta con múltiples aplicaciones. La detección de anuncios en una señal de vídeo es un problema crucial que está altamente relacionado con técnicas de procesamiento de vídeo como la segmentación, indexación y recuperación; y en este mismo orden se puede decir que esta problemática en los últimos años ha logrado la atención de la comunidad científica e industrial.

Los métodos desarrollados para la detección de anuncios se pueden dividir en dos categorías fundamentales: los basados en las características [22, 26] y los basados en el reconocimiento [2, 7, 13, 18, 19, 21, 38]. Mientras que el primer grupo de técnicas utilizan algunas características inherentes a los anuncios de televisión para distinguirlos de otros tipos de vídeo, los métodos de reconocimiento intentan identificarlos buscando en una base de datos que contiene anuncios conocidos previamente.

Con anterioridad se han planteado varias técnicas para la detección de anuncios teniendo en cuenta solamente en la señal de vídeo o la combinación de audio y vídeo, destacándose aquellas basadas en la caracterización de las secuencias de vídeo. Los trabajos de la comunidad investigadora se han centrado en la detección de la presencia de logotipo del canal de televisión [17] y en la detección basada en los fotogramas en negro y en silencio [36]. Sin embargo, las estaciones de televisión actualmente no siempre ocultan sus identificadores durante los anuncios y los fotogramas en negro son insertados de forma aleatoria para determinados propósitos de edición. Con el objetivo de evitar estas limitaciones, las técnicas de procesamiento y clasificación basada en el cambio de tomas de vídeo muestran un gran potencial [8, 22]; pero una de las desventajas de esta se encuentra en la continuidad del contenido del comercial siendo clasificados como falsos negativos.

Estos sistemas donde solo se analiza la señal de vídeo son poco eficientes computacionalmente y poseen resultados inferiores a los que además usan las características del audio. Los métodos más robustos desarrollados consisten en sistemas híbridos que explotan características tanto del audio como el vídeo [6, 8, 22, 36]. Sin embargo el desafío de todas las técnicas desarrolladas es el mismo: cómo detectar correctamente intervalos de publicidad; y cómo realizar un proceso en el menor tiempo posible.

El estudio de las características del audio ha despertado el interés de la comunidad científica durante los últimos años [34, 43, 45]. Se pueden encontrar numerosos trabajos que explotan las características de las señales de audio para la segmentación, indexación, clasificación y recuperación de materiales audiovisuales [9, 14, 27, 41, 43, 45]. La caracterización de la señal de audio de una emisión de televisión igualmente suele ser útil para la detección de anuncios, sin embargo muchos de los métodos creados hasta el momento no utilizan el audio [26, 37], mientras que otros solo usan unas pocas características [36, 39].

Formulación del Problema

El desarrollo de un sistema de monitoreo e indexación de emisiones de televisión requiere de la incorporación de técnicas de avanzada de procesamiento de vídeo. Específicamente las aplicaciones orientadas a mantener archivos audiovisuales con las grabaciones de diferentes cadenas televisivas deben considerar mecanismos que posibiliten el acceso, recuperación y reutilización de los materiales de forma efectiva y eficiente. Con este objetivo se llevan a cabo procesos como la extracción automática de información que permite, entre otras, la detección automática de cambios de escena, segmentación del vídeo, creación de índice visual, generación e identificación de huella digital (*fingerprinting*[1]) de vídeo y audio, extracción de subtítulos de texto sobreimpresos, detección de intervalos de publicidad.

Como parte de la fase conceptual de un producto con estas características, de manera que permita un impacto significativo en el mercado nacional e internacional, y dada la importancia que tiene la temática en numerosas aplicaciones tanto legales como comerciales; el presente trabajo se centra en la detección de anuncios con el objetivo de aumentar las capacidades de almacenamiento de los sistemas de monitoreo e indexación de emisiones de televisión. Teniendo en cuenta el poco desarrollo de técnicas de detección que exploten las características del audio y la importancia que tiene desarrollar métodos cada vez más eficientes, se plantea como **problema de la investigación**, la inserción de anuncios en las emisiones de televisión provoca un uso innecesario de espacio de almacenamiento en los sistemas de monitoreo e indexación de transmisiones televisivas.

El **objeto** en el cual se enmarca el **estudio**, tanto desde el punto de vista teórico como práctico, con vistas a la solución del problema planteado consiste en la detección de anuncios a partir técnicas automáticas. Mientras que el **campo de acción** lo constituye la detección de forma automática de bloques comerciales para aplicaciones de monitoreo e indexación de emisiones de televisión.

Con el fin de brindar una solución efectiva al problema, se plantea como **objetivo general** desarrollar una técnica que permita detectar bloques de publicidad basada en la caracterización del audio para sistemas de monitoreo e indexación de televisión, teniendo en cuenta el estudio y aplicación de diversas soluciones existentes utilizadas para resolver problemas similares de manera eficiente.

1 Dado que la mayor parte de la bibliografía sobre el tema que se trata está en lengua inglesa, de cara a establecer una relación clara entre lo que se expone y las referencias bibliográficas, a lo largo del trabajo se hará referencia expresa a la nomenclatura inglesa de los términos que se consideren más relevantes.

Diversas son las técnicas y algoritmos que se aplican para darle solución al problema descrito o similares. Con la presente investigación, se propone una variante de solución partiendo de la siguiente **hipótesis**:

Si se explotan las características de la señal de audio de las emisiones de televisión de manera que identifiquen las secuencias de anuncios, se logrará desarrollar una técnica para la detección de bloques de publicidad, que utilice una cantidad reducida de información para realizar los análisis estadísticos de clasificación, a la vez que permita reducir el espacio de almacenamiento requerido por los sistemas de monitoreo e indexación de televisión.

Como **tareas de investigación** se proponen las siguientes:

- Identificar las variantes de solución existentes y tendencias actuales para dar solución al problema planteado, a partir del estudio de los referentes teóricos que preceden al presente trabajo.
- Evaluar las técnicas de extracción de características de audio, diseño de clasificadores y sistemas de monitoreo de televisión recogidas en la bibliografía actual, que puedan servir de precedentes para la investigación.
- Desarrollar una técnica para la detección de intervalos de publicidad a partir de la selección de descriptores de audio que representen la secuencia de video y permitan realizar una clasificación eficiente de la información.
- Evaluar los resultados obtenidos usando métricas de referencia para esta problemática, a través de la implementación de una herramienta de software que permita realizar pruebas de la técnica desarrollada.

Estructura del Documento

El presente documento se encuentra dividido en tres capítulos. En el primero de ellos, Fundamentación Teórica, se realiza una caracterización de los principales conceptos asociados al dominio del problema. Además se hace un análisis del estado del arte que precede a la realización de este trabajo, presentando las más importantes técnicas existentes para la detección de anuncios, así como sus resultados.

En el segundo capítulo, Presentación de la Solución Propuesta, se explica el proceso de detección de anuncios desarrollado a partir de la señal de audio de emisiones de televisión, caracterizando cada una de las fases que la componen. Consolidando de esta manera las aportaciones teórico-prácticas de la investigación.

En el tercer capítulo, Análisis de los Resultados, se exponen los resultados obtenidos a partir de pruebas realizadas al método de detección de anuncios propuesto. Se describen las herramientas utilizadas para la realización de los ensayos, las medidas de eficiencia utilizadas y se analizan los resultados.

Finalmente se dan las conclusiones generales y un conjunto de recomendaciones con vistas a trabajos futuros.

Glosario de Términos

Con el objetivo de hacer más fácil la comprensión del trabajo a continuación se definen algunos términos asociados al dominio del problema:

- Clasificación: la acción de asignar segmentos de datos en categorías, nótese que la clasificación siempre incluye una segmentación de los datos.
- Fotograma: cada una de las imágenes que conforman un vídeo.
- *Ground truth*: conjunto de datos clasificados con conocimientos pleno de sus categorías verdaderas, es utilizado para la corrección y evaluación del reconocimiento de patrones.
- Reconocimiento de patrones: la combinación de las acciones de segmentación y clasificación, la segmentación puede ser realizada de forma independiente.
- Segmentación: la acción de dividir un conjunto de datos en subconjuntos más pequeños, aplicado al vídeo puede ser espacialmente o temporalmente.
- Vector de características: es un vector n-dimensional de características que representan un objeto, estás características son comúnmente representadas numéricamente para facilitar el procesamiento y el análisis estadístico realizado en el reconocimiento de patrones.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

Con el objetivo de facilitar la comprensión del alcance de la investigación y la teoría que lo respalda, en el presente capítulo se exponen conceptos asociados al dominio del problema planteado. Se realiza una caracterización de temas fundamentales tales como segmentación, clasificación, descriptores de audio y sistemas de monitoreo e indexación de televisión. Además se hace un análisis del estado del arte que precede a la realización de este trabajo y que contribuye a esclarecer su objeto de estudio; presentando las principales técnicas existentes para la detección de anuncios, así como los resultados logrados por estas.

1.1 Conceptos asociados al dominio del problema

Siempre que se plantea la tarea de realizar una investigación surgen conceptos esenciales asociados al problema abordado, estos constituyen la base fundamental del tema tratado y resulta de suma importancia su comprensión para el desarrollo del trabajo. A continuación se relacionan los principales conceptos asociados al dominio del problema, los cuales facilitarán la comprensión de la presente investigación.

1.1.1 Segmentación de vídeo

El análisis de secuencias de vídeo orientado a la extracción automática de información referente a su contenido es un proceso genéricamente conocido como *video parsing* [1]. El primer paso de dicho análisis consiste habitualmente en llevar a cabo una segmentación temporal de la secuencia de vídeo (*video temporal segmentation*), es decir, una subdivisión o estructuración en unidades homogéneas desde algún punto de vista, ya sea objetivo (luminosidad media, distribución de color, movimiento de la cámara, etc.) o subjetivo (coherencia de contenido) [3]. En primer lugar, ello permite disponer de un índice de la secuencia que posibilita un acceso eficaz a partes de la misma, motivo por el que las técnicas de segmentación temporal también se agrupan bajo el nombre de técnicas de indexación de vídeo o *video indexing*. En segundo lugar, la homogeneidad de las subsecuencias obtenidas facilita la aplicación de técnicas de extracción manual o automática de características (personajes, marcas publicitarias, letreros, primeros planos o paisajes, etc.) orientadas a anotar los contenidos de la secuencia, de ahí que se conozcan como técnicas de catalogación de vídeo o *video annotation*.

Dado que el problema abordado sugiere la división de una secuencia de vídeo en subsecuencias de bloques de anuncios y de no anuncios, para el presente trabajo se adopta la definición de segmentación dada en [4], donde se define como el proceso al cual es sometido un fichero audiovisual digital para ser dividido en partes, teniendo en cuenta criterios de selección para proceder a su posterior análisis con mayor facilidad.

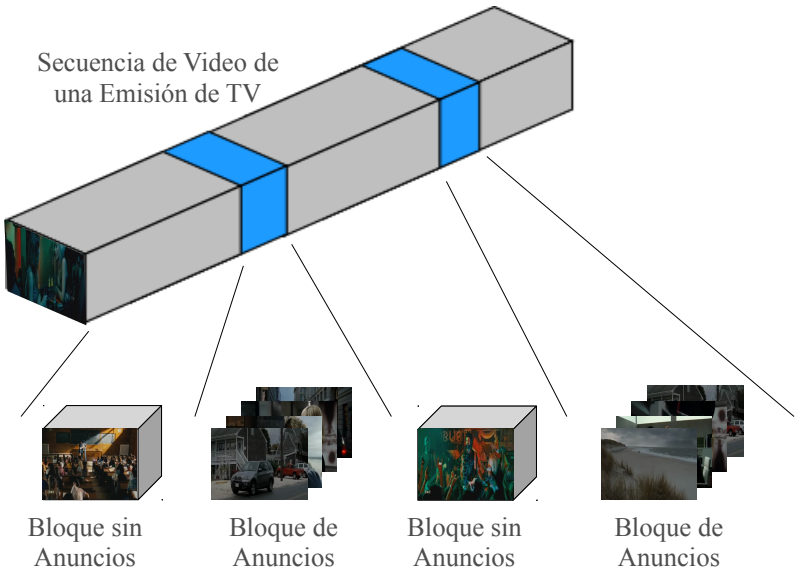

Figura 1.1: Segmentación de Vídeo por Bloques de Anuncios.

1.1.2 Diseño de clasificadores

El problema de la clasificación es uno de los primeros que aparecen en la actividad científica y constituye un proceso consustancial con casi cualquier actividad humana, de tal manera que en la resolución de problemas y en la toma de decisiones la primera parte de la tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología correspondiente y que en buena medida dependerá de esa clasificación. La base principal de la solución a la problemática abordada consiste precisamente en una tarea de clasificación de las secuencias de vídeo en dos clases específicas: ANUNCIO y NO ANUNCIO. (Veáse Figura 1.1)

Se puede decir que un clasificador es una función que permite etiquetar objetos de diferentes clases a partir de un conjunto de atributos de entrada. Una forma muy común para construir un clasificador es tomar un conjunto de ejemplos previamente etiquetados (*training data*) para tratar de definir una regla que pueda asignar un nuevo objeto (*testing data*) a una de las clases, esta técnica es conocida como

machine learning. El principal objetivo de los métodos de *machine learning* es programar las computadoras para el uso de datos de ejemplo o la experiencia para resolver un problema dado. [42]

Las técnicas de aprendizaje de estos métodos de clasificación suelen dividirse en las siguientes categorías:

- *Aprendizaje supervisado:* funciona con ejemplos etiquetados previamente, por lo que a *priori* se conoce las clases existentes, luego dado un conjunto de vectores de características intenta clasificarlos en una de esas clases.
- *Aprendizaje no supervisado:* consiste en el aprendizaje a partir de patrones de entrada sin valores o clases determinadas. El principal problema de esta técnica es cómo tomar una decisión entre los patrones dados. El sistema toma los objetos de entrada como un conjunto de variables aleatorias, construyendo de esa manera un modelo de densidad para el conjunto de datos. Esta técnica es conocida como *clustering*.
- *Aprendizaje semi-supervisado:* se basa en técnicas que combinan las dos anteriores, esto se debe a que en ocasiones resulta muy difícil clasificar todos los datos. La idea es combinar los datos etiquetados y no etiquetados para obtener un modelo de clasificación. Esta técnica toma importancia cuando el sistema diseñado tiene acceso a un número limitado de datos etiquetados.

La clasificación puede ser tratada como dos problemas por separado: la clasificación binaria y la multi-clases. En la clasificación binaria solo están involucradas dos clases, mientras que en la multi-clases un objeto puede ser asignado a una de varias clases [16]. En la literatura se pueden encontrar varios tipos de clasificadores, entre ellos se destacan los mencionados a continuación:

- *Clasificadores Bayesianos*: es un clasificador de patrones basado en las teorías estadísticas de aprendizaje. Se calcula la probabilidad de cada hipótesis de los datos y se realiza la predicción a partir de esta, basándose en la teoría de decisión de Bayes. Se destacan las técnicas *Nearest Neighbor* y *Naive Bayes*.
- *Clasificadores Lineales:* en este caso el resultado del clasificador es equivalente a un conjunto de funciones discriminantes. Resaltan por su simplicidad y por su poca complejidad computacional. Se destacan los algoritmos del *Perceptron, Least Square (LMS) y Mean Square (MSE)*.
- *Clasificadores No Lineales o de Backpropagation:* en algunos casos es necesario utilizar modelos de densidad más sofisticados cuando los modelos simples o de densidad paramétrica no

ofrecen buenos resultados. La redes neuronales son una técnica de aproximación muy útil para construir modelos paramétricos de densidad. Los modelos de redes neuronales usados comúnmente consisten en una red con una capa de entrada con varios nodos de entrada, una capa oculta con un número variable de nodos que dependen de las características del problema, y una capa de salida con varios nodos con posibles resultados.

- *Clasificadores con Análisis de Componentes Principales (PCA, por sus siglas en inglés):* en este caso es usado el método PCA, esta es una técnica que intenta reducir el número de variables mediante una transformación del conjunto original de estas (que tienen cierta información en común) en un conjunto de variables sin redundancia llamado "conjunto de componentes principales". Las nuevas variables son una combinación lineal de las originales.

- *Clasificadores de Máquina de Soporte Vectorial (SVM, por sus siglas en inglés):* SVM crea un hiper-plano de separación óptima a partir de un conjunto de ejemplos de entrenamiento. Puede ser lineal o no lineal dependiendo de la función de optimización utilizada (*kernel function*), por lo que se puede decir que es un caso específico de un clasificador lineal, sin embargo su uso más común suele ser cuando el conjunto de datos no es linealmente separable. Es considerado más fácil de usar que las redes neuronales y su meta es crear un modelo a partir de los datos de entrenamiento. Entre los *kernels* más usados se encuentran el lineal, el polinomial, el sigmoid y la función de base radial.

- *Clasificadores con Dependencia de Contexto:* normalmente en la tarea de clasificación se asume que no existe relación entre las diferentes clases. En otras palabras, teniendo un vector de características de una clase dada, este no puede ser parte de otra clase. En este tipo de clasificadores se elimina esta suposición, asumiendo que las diferentes clases están estrechamente relacionadas. Bajo esta premisa la clasificación de cada vector de características no tiene sentido por separado del resto. La clase de un vector de características es asignada en dependencia de (a) su propio valor, (b) del valor de los demás vectores, y (c) de la relación entre las diferentes clases. La técnica más usada son los Modelos Ocultos de Markov (HMM, por sus siglas en inglés) y suele ser empleada en aplicaciones para las comunicaciones, el reconocimiento de voz y procesamiento de imágenes.

- *Clasificadores de Boosting:* es un meta-algoritmo de *machine learning* para el aprendizaje supervisado que ha alcanzado gran popularidad en los últimos años. La técnica de *boosting* está

basada en la hipótesis de Kearns [24]: puede un conjunto de clasificadores débiles formar un clasificador fuerte. Un clasificador débil es definido como un clasificador el cual es poco correlacionado con la clasificación correcta (puede etiquetar ejemplos mejor que las suposiciones aleatorias). Por lo contrario, un aprendiz fuerte es un clasificador que está arbitrariamente bien correlacionado con la clasificación correcta. Mientras que la técnica de *boosting* no es algorítmicamente restringido, muchos de los algoritmos de *boosting* consisten en el aprendizaje iterativo de clasificadores débiles con determinada distribución para finalmente formar un clasificador fuerte. Existen varios algoritmos de *boosting*. Los primeros fueron los propuestos por Robert Schapire [40] y Yoav Freund [12], los cuales no eran adaptativos y no tomaban una completa ventaja de los clasificadores débiles. El algoritmo AdaBoost es el más popular y es el primero que pudo adaptar los clasificadores débiles. No obstante, hay muchos más recientes tales como LPBoost, TotalBoost, BrownBoost, MadaBoost, LogitBoost, entre otros.

1.1.3 Extracción de características del audio

Un proceso para la clasificación automática de objetos, imágenes o contenido audiovisual en general consiste en una secuencia de pasos. El primero de estos suele ser extraer un conjunto de características o descriptores que identifiquen en forma de un vector de características al objeto como tal y que puedan ser utilizadas para su clasificación. Este conjunto de características normalmente ocupa menos espacio de almacenamiento que el objeto que describen, debido a que se almacenan de forma numérica para facilitar el procesamiento y análisis estadístico posterior. Además los descriptores de características aportan mayor información para agrupar y comparar diferentes objetos.

Este mismo proceso de extracción de características se lleva a cabo para la clasificación de señales de audio y cada una de ellos describen un aspecto diferente de la señal (Veáse Figura 1.2). Existen numerosos descriptores que pueden ser utilizados para la clasificación de señales de audio. Generalmente son clasificados en dos categorías: descriptores de dominio de tiempo y de dominio de frecuencia. Estos descriptores pueden ser extraídos en dos niveles diferentes: a nivel de ventana o a nivel de clip. A continuación se dará una panorámica de los principales descriptores de audio[2] que se utilizan para la clasificación de señales de audio encontrados en la literatura consultada.

2 En algunos casos se mantiene la nomenclatura inglesa de los descriptores para no perder la referencia directa con la literatura especializada y con el estándar MPEG-7.

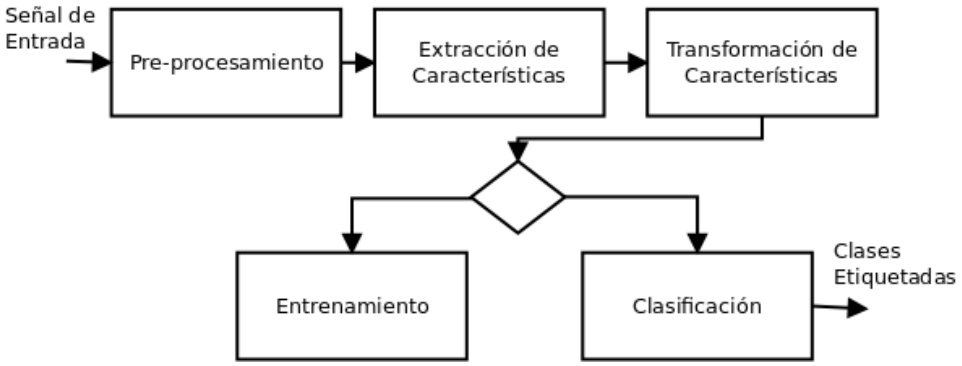

Figura 1.2: Diagrama de bloques de un Sistema Genérico de Clasificación de Audio.

Para la extracción de las características la señal de audio se divide en clips de una duración entre 1 y 3 segundos, estos a su vez se subdividen en un conjunto de *frames* o ventanas. La duración de cada *frame* es de 20~30 ms, debido a que si la duración es muy grande no se obtiene una buena caracterización de los cambios de la señal de audio; mientras que si es muy pequeña no se puede extraer descriptores válidos. Por lo general la longitud de los *frames* (en términos de muestras) es igual a una potencia de 2 (por ejemplo: 128, 256, 512, etc). Si se desea reducir la diferencia entre dos *frames* vecinos, se puede solapar los mismos. Usualmente el solapamiento debe ser de 1/2 a 2/3 de la longitud.

Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs son características espectrales basadas en la percepción auditiva humana, las cuales son extensamente usadas para el procesamiento de audio y reconocimiento del habla. Se pueden consultar varios trabajos relacionados con la clasificación de audio donde utilizan este tipo de descriptor [9, 15, 43].

Para el cálculo de los MFCCs la señal de audio es segmentada y enventanada en pequeños *frames* de 256 muestras. Se calcula la magnitud del espectro para cada *frame* usando la Transformada de Fourier (FFT, por sus siglas en inglés) y son convertidos por un conjunto de filtros (*filter bank*) en la escala mel[3]. Luego se le aplican el logaritmo y la Transformada Discreta del Coseno (DCT, por sus siglas en inglés)

3 Es una escala perceptual del tono a juicio de oyentes equiespaciados. 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.

para obtener los coeficientes. Comúnmente solo son usados los primeros 13 coeficientes como descriptores.

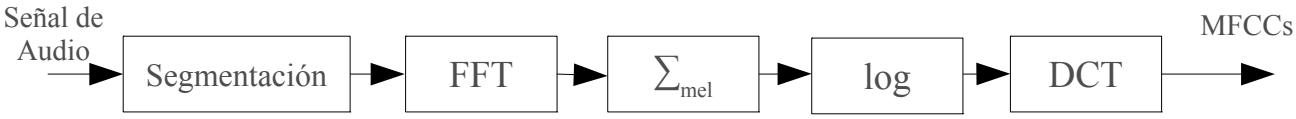

Figura 1.3: Diagrama de flujo para cálculo de MFCCs.

Entropía de la Energía

La entropía de la energía expresa los cambios abruptos en los niveles de energía de la señal de audio. En el cálculo de esta característica los *frames* son divididos en **K** sub-ventanas de igual duración. Luego para cada sub-ventana, se calcula la energía normalizada σ_i^2, por ejemplo la energía de la sub-ventana dividida por la energía de la ventana. El valor de la entropía de la energía es menor en los *frames* con mayores cambios en el nivel de la energía. La fórmula se puede ver a continuación:

$$EE = -\sum_{i=1}^{K} \sigma_i^2 \log_2 \sigma_i^2(l)$$

Short Time Energy (STE)

STE se define como la energía de la señal de audio en un *frame*; su fórmula está dada por:

$$STE_n = \frac{1}{N} \sum_{m=1}^{N} [x(m)w(n-m)]^2$$

Donde **x(m)** es la señal de audio, **n** es el índice del descriptor y **w(m)** es la ventana de longitud **N**.

Esta característica provee una representación de la variación de amplitud sobre el tiempo. Este descriptor puede ser usado como una medida para distinguir el sonido del silencio, además para diferenciar la voz hablada de cuando no existe voz en la señal de audio.

Spectrum Flux (SF)

Spectrum flux es definido como valor medio de la variación del espectro entre dos ventanas adyacentes; a continuación se muestra la fórmula:

$$SF = \frac{1}{(N-1)(K-1)} \times \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log A(n,k) - \log A(n-1,k)]^2$$

Donde **A(n,k)** es la Transformada Discreta de Fourier (DFT, por sus siglas en inglés) para el n-ésimo *frame*:

$$A(n,k) = |\sum_{m=-\infty}^{\infty} x(m)\,\omega(nL-m)\,e^{j\frac{2\pi}{L}km}|$$

x(m) es la señal de audio original, **ω(m)** es la función ventana, **L** es la longitud de la ventana, **K** es el orden de la DFT y **N** es el número total de *frames*.

En general, los valores de SF del habla son mayores que para la música. Este descriptor puede ser usado para la clasificación entre el habla y la música, así como el habla sobre un sonido de fondo. También es usado para determinar el timbre de una señal de audio.

Spectral Centroid (SC)

También conocido como *Frecuency centroid*, es definido como el "centro de gravedad" de la magnitud del espectro, por ejemplo la frecuencia que divide la magnitud espectral en dos porciones iguales aproximadamente.

$$FC(n) = \frac{\sum_{i=0}^{N-1} iS_n(i)}{\sum_{i=0}^{N-1} S_n(i)}$$

Donde $\mathbf{S_n(i)}$ es la muestra i-ésima en el *frame* n-ésimo y **N** es la longitud del *frame*.

Este descriptor presenta un mayor valor para la música que para señales del habla. Este descriptor es comúnmente utilizado como una buena caracterización del timbre.

Zero-Crossing Rate (ZCR)

ZCR describe las veces que una señal de audio cruza el *Eje X* en un *frame*. Es una simple medida de la frecuencia del contenido de la señal. Por la importancia de este tipo de descriptor es usado en la mayoría de las aplicaciones para la clasificación y procesamiento de señales de audio. La fórmula para su cálculo es la siguiente:

$$ZCR = \frac{\sum_{n=1}^{N-1} |sgn[x(n+1)] - sgn[x(n)]|}{(N-1)}$$

Donde **sgn[]** es una función signo y **x(n)** es la señal de audio.

$$sgn[x(n)] = \begin{cases} 1 & (x(n) \geq 0) \\ -1 & (x(n) < 0) \end{cases}$$

Por lo general, las señales de audio del habla están compuestas por una alternación de sonidos con voz y sin voz, mientras que la música no presenta este tipo de estructura. Esta variación provoca que el ZCR sea mucho mayor para señales que contienen habla que las de música.

Descriptores derivados del Volumen

La distribución del volumen de una señal de audio revela la variación temporal de la magnitud de la señal, la cual es importante para la clasificación de escenas. Para calcular el valor del volumen en un *frame* se puede utilizar la media cuadrática (RMS, por sus siglas en inglés) de la magnitud de la señal. Específicamente el volumen del *frame* n-ésimo es calculado por la fórmula:

$$V(n) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} S_n^2(i)}$$

Donde $\mathbf{S_n(i)}$ es la muestra i-ésima en el *frame* n-ésimo y **N** es la longitud del *frame*.

Nótese que el volumen de una señal de audio depende de la ganancia dada en el sistema de grabación y digitalización. De esta manera, la media del volumen en un clip no necesariamente refleja el contenido de las escenas, sin embargo la variación temporal del volumen sí lo hace. [27]

Para medir la variación temporal del volumen se definen dos descriptores basados en la distribución del volumen. El primero es la **Desviación Estándar del Volumen** (VSTD, por sus siglas en inglés), definida por la desviación estándar del clip n-ésimo normalizada por el valor máximo del volumen en el clip.

$$VSTD(n) = \frac{STD(V_n)}{MAX(V_n)}$$

El segundo descriptor es el **Rango Dinámico del Volumen** (VDR, por sus siglas en inglés) definido como:

$$VSTD(n) = \frac{MAX(V_n) - MIN(V_n)}{MAX(V_n)}$$

Como es obvio estas dos características están correlacionadas, pero según el trabajo realizado por [27] ambas presentan información independiente acerca del contenido de las escenas en la secuencia de vídeo.

Ratio de Silencio

El Ratio de Silencio (SR, por sus siglas en inglés) es definido como el ratio de las ventanas en silencio que tiene un clip de audio. Un *frame* es considerado en silencio si su energía total es menor que un determinado umbral. Por lo general el valor de SR del habla es mayor que las señales de música, debido a la continuidad de sonidos que tienen estas últimas.

Una técnica propuesta por [27] para detectar los *frames* en silencio, consiste en comparar el volumen y el ZCR de cada ventana con un umbral determinado. Si ambos descriptores se encuentran por debajo del umbral el *frame* es declarado en silencio.

A partir del resultado de la detección de silencio, se puede calcular el **Ratio de No-Silencio** (NSR, por sus siglas en inglés), el cual consiste en la razón de los intervalos de no-silencio en todo el clip de audio. Esta característica varía significativamente en diferentes tipos de vídeos, por ejemplo en secuencias de noticias presenta pausas debido al habla del reportero; sin embargo en secuencias de anuncios tiene un alto valor provocado por la continuidad de la música de fondo.

Contorno de Pitch

Pitch es el período fundamental de una señal de audio; es un importante parámetro en el análisis y procesamiento de señales del habla. Esta característica representa la tasa de vibración de la señal de audio. Puede ser representado igualmente como la frecuencia fundamental de la señal.

Se pueden encontrar varios algoritmos para determinar el valor de *pitch*, a continuación se muestra uno que utiliza *Average Magnitude Difference Function* (AMDF), la cual se define como:

$$AMDF(l) = \frac{\sum_{i=0}^{N-l-1} |S_n(i+l) - S_n(i)|}{N-l}$$

Donde $\mathbf{S_n}$ es la señal de audio en el *frame* n-ésimo, **N** es la longitud del *frame* y **l** es un parámetro de desplazamiento.

El algoritmo fue propuesto por [20] y la idea es encontrar el primer mínimo local. En este caso se utiliza un mínimo local que satisfaga las restricciones adicionales en cuanto a su valor en relación con el mínimo global y su curvatura. Se realiza la búsqueda de izquierda a derecha, siendo el valor de *pitch* la distancia entre el origen y el primer mínimo local. Después de calcular el *pitch* para cada *frame* se puede obtener el contorno de *pitch* para todo el clip. Para los intervalos donde no se encuentra un mínimo el valor de *pitch* se asume como nulo.

Ratio de Voz o Música

Basado en el resultado de la estimación del *pitch* se puede detectar cuáles *frames* corresponden a voz o música. El ratio de voz o música (VMR, por sus siglas en inglés) es usado como otra característica del audio, el cual es definido como la razón de la longitud de los *frames* de voz o música entre la longitud del clip. Un *frame* es considerado como voz o música si tiene un *pitch* mayor que 50 ms.

Ratio de Ruido

El Ratio de Ruido (NR, por sus siglas en inglés) se define como la razón entre las ventanas de ruido y no ruido en un clip de audio. Se considera un *frame* como ruido si el máximo local de su función correlación normalizada es menor que un umbral determinado. Igualmente se puede determinar si el valor de *pitch* es nulo en un *frame* que no sea de silencio. El valor de NR en señales con sonido ambiente es mayor que en la música.

Descriptores de audio de MPEG-7

El estándar MPEG-7 especifica para el audio estructuras de datos y técnicas para la descripción del contenido. Entre ellos se encuentran los descriptores de bajo nivel (LLDs, por sus siglas en inglés), que son una colección de características que describen propiedades del sonido como la armonía, nitidez, el timbre y el pitch, entre otros.

Los descriptores son aplicados a nivel de *frames* y sus valores pueden ser escalares o vectores. Los LLDs se organizan en seis grupos que se muestran en la Tabla 1.1.

Grupo	Descriptor de Bajo Nivel	Abreviatura
Básico	AudioWaveform	AW
	AudioPower	AP
Básicos Espectrales	AudioSpectrumEnvelope	ASE
	AudioSpectrumCentroid	ASC
	AudioSpectrumSpread	ASS
	AudioSpectrumFlatness	ASF
Base Espectral	AudioSpectrumBasis	ASB
	AudioSpectrumProjection	ASP
Parámetros de la Señal	AudioHarmonicity	AH
	AudioFundamentalFrequency	AFF
Timbre Temporal	LogAttackTime	LAT
	TemporalCentroid	TC
Timbre Espectral	SpectralCentroid	SC
	HarmonicSpectralCentroid	HSC
	HarmonicSpectralDeviation	HSD
	HarmonicSpectralSpread	HSS
	HarmonicSpectralVariation	HSV

Tabla 1.1: Descriptores de audio de bajo nivel de MPEG-7.

1.1.4 Sistemas de Monitoreo e Indexación de TV

En la actualidad existen dos grupos bien definidos interesados en el monitoreo e indexación de las emisiones de televisión. Por una parte se encuentran aquellos que desean controlar las emisiones debido a normativas, legislaciones o intereses económicos; entre estos se encuentran los organismos encargados de supervisar el cumplimiento de las leyes en cada país, además están las compañías y agencias de publicidad que hacen uso de la televisión para sus campañas de marketing. El segundo grupo se puede definir como aquellos interesados en mantener archivos audiovisuales con las grabaciones de diferentes televisoras, ya sea por interés personal o comercial, entre estos se encuentran televidentes, bases de datos audiovisuales, las propias cadenas de televisión, entre otros.

Las empresas y organizaciones que quieran explotar al máximo el vídeo, deben habilitar los mecanismos necesarios para que de forma efectiva y eficiente puedan acceder, recuperar y reutilizarlo. El disponer de un acceso flexible e inteligente a los recursos de vídeo, combinado con métodos asequibles y simples de colaboración y distribución, se traduce en ciclos de producción más cortos y mejoras en los procesos de comunicación . De aquí la necesidad de los sistemas de monitoreo e indexación de televisión.

Management Sciences for Health [30] define el término monitoreo como el proceso de verificación periódica de la situación de un programa, para determinar si las actividades se están cumpliendo en la forma planeada. Mientras que [32] precisa que es proceso continuo de recolección y análisis de datos cualitativos y cuantitativos, con base en los objetivos planteados en un programa o proyecto, que tiene como propósito descubrir fortalezas y/o debilidades para establecer líneas de acción, permitiendo brindar correcciones y reorientaciones técnicas en la ejecución.

La indexación audiovisual consiste en la generación de metadatos (etiquetas) en base a determinados contenidos audiovisuales para su posterior registro con dichas referencias las cuales facilitaran posteriormente posibles búsquedas de información. Indexar se refiere a la acción de registrar ordenadamente información para elaborar su índice. En informática, tiene como propósito ejecutar la elaboración de un índice que contenga de forma ordenada la información, esto con la finalidad de obtener resultados de forma sustancialmente más rápida y relevante al momento de realizar una búsqueda.

Un sistema informático consiste en un conjunto de procedimientos y reglas lógicas escritas en la forma de programas y aplicaciones, que definen el modo de operación de la computadora y están almacenadas en los diferentes tipos de memoria de lectura/escritura [35]. Igualmente se puede enunciar como series de instrucciones codificadas que sirven para que la computadora realice una tarea.

De esta manera, se puede definir un sistema de monitoreo e indexación de televisión, como un conjunto de procedimientos y reglas lógicas escritas en la forma de programas y aplicaciones, que sirven para la recolección y análisis continuo de datos cualitativos y cuantitativos de las emisiones de televisión, permitiendo la generación de metadatos para el almacenamiento de las grabaciones audiovisuales, así como las posteriores búsquedas y recuperación de la información.

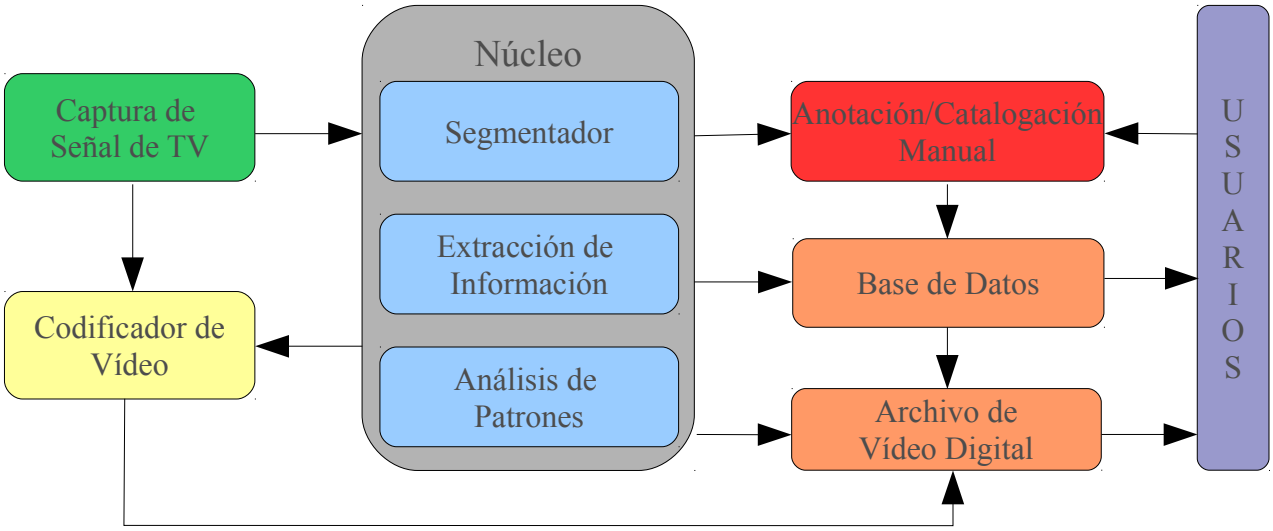

Figura 1.4: Esquema conceptual de un Sistema de Monitoreo e Indexación de TV.

Un sistema de monitoreo e indexación de televisión debe facilitar a los usuarios todos los procesos implicados en la gestión de fondos audiovisuales, desde la entrada o ingesta del vídeo hasta su posterior recuperación, una vez pasada la etapa de procesado del mismo. Funcionalmente están formados por dos grandes bloques: el núcleo o *indexer* y las aplicaciones de gestión de los fondos digitales.

El primero de ellos, el núcleo, toma la salida procedente de la captura de la señal y lleva a cabo procesos para la extracción automática de información que permiten: la detección automática de cambios de escena, segmentación del vídeo, creación de índice visual, generación e identificación de huella digital (*fingerprinting*) de vídeo y audio, extracción de subtítulos de texto sobreimpresos, detección de intervalos de publicidad, entre otros. Todos estos procesos contribuyen a la indexación de la información en la base de datos audiovisual, su posterior recuperación, así como el manejo de alarmas en caso de detectar alguna anomalía.

El segundo bloque es el encargado de la gestión (catalogación y recuperación) del archivo audiovisual que va generando el núcleo y proporciona funcionalidades como la catalogación de los clips generados por el *indexer* (algunos aspectos del contenido pueden extraerse de forma automatizada, pero una representación detallada obliga a que el vídeo sea catalogado o descrito con la suficiente precisión) y la búsqueda y recuperación de secuencias de vídeo.

1.2 Técnicas para la detección de anuncios

En este epígrafe se relacionan las técnicas más destacadas para la detección de anuncios relacionadas en la literatura consultada, la comprensión de las mismas servirán para tener una idea del estado del arte y de los resultados alcanzados por otras investigaciones sobre el tema. Igualmente se caracterizan algoritmos híbridos que integran varias técnicas para lograr mayor robustez y mejores resultados.

Antes de pasar a la caracterización de las diversas técnicas existentes, resultaría beneficioso explicar ciertas características que presentan los anuncios en las emisiones de televisión, las cuales son la base fundamental de muchas de las técnicas que se explicarán más adelante.

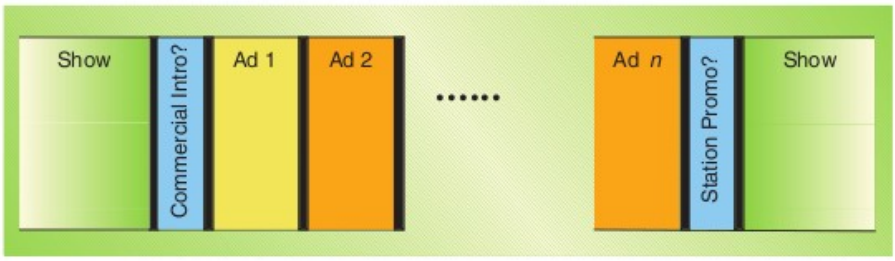

Figura 1.5: Estructura general de un bloque de publicidad.

Entre las principales características en que se basan los diferentes algoritmos desarrollados se encuentran las siguientes:

- **Fotogramas en negro y en silencio**: dos anuncios consecutivos en una secuencia de vídeo normalmente son separados por un corte de 5 a 12 fotogramas en negro y en silencio.
- **Cambio de tomas de vídeo**: la publicidad es más llamativa a la atención de los televidentes, por esta razón se compone de mucho movimiento, animación y acción en la secuencia de vídeo. Esto es posible a una alta frecuencia de cortes (*shots*) y cambios constantes de las escenas mostradas.
- **Presencia del indentificador del canal de televisión**: normalmente las cadenas de televisión ocultan el identificador o logotipo del canal mientras se muestran intervalos de publicidad.
- **Volumen del audio**: es muy común que durante la transmisión de comerciales las televisoras aumenten el volumen de la señal de audio para llamar mucho más la atención de los televidentes.

- **Duración de los anuncios**: la medida de la duración de los spots y bloques de publicidad en muchas ocasiones es restringido de acuerdo a ciertas normativas, de esta manera un anuncio no debe superar los 90 segundos y la duración de los bloques en muchos países está regulada.

1.2.1 Detección basada en *fotogramas* en negro y en silencio

La característica más común que se utiliza para la detección de anuncios es la delimitación de los fotogramas en negro y en silencio. Para la localización de los fotogramas en negro se utiliza un método bien simple, consistente la búsqueda del valor medio de intensidad de los píxeles de la imagen.

La intensidad media es determinada fácilmente, así como plantea [23], y es la base de varias aplicaciones de detección de anuncios. Un fotograma es declarado en negro si el valor medio de intensidad es inferior a determinado umbral. También puede ser reconocido utilizando su histograma de luminancia. Otra técnica propuesta por [26] para la detección de los fotogramas en negro utiliza la desviación estándar de la intensidad acorde a un umbral. En [36] proponen un método para detectar los fotogramas en negro sobre una codificación MPEG sin el costo computacional de la decodificación del vídeo.

De una manera similar los fotogramas en silencio pueden ser determinados examinando el nivel de volumen medio de la señal de audio. Muchas aplicaciones utilizan esta segunda características para rectificar la detección de fotogramas en negro que son irrelevantes. De esta manera un fotograma en negro solo es detectado positivamente si está acompañado de un silencio.

De acuerdo a determinados propósitos de las compañías de televisión son insertados fotogramas en negro de manera aleatoria durante el proceso de edición. Para reducir los falsos positivos muchos algoritmos solo declaran como positivos un número determinado de fotogramas en negro consecutivos, usualmente cinco o seis.

Una vez que los fotogramas en negro son detectados, pueden ser explotados determinados aspectos con relación al tiempo de los anuncios. Cuando se determinan dos secuencias de fotogramas en negro indica que puede haber entre ellos un segmento de comerciales. Algunos algoritmos establecen un tiempo máximo entre las secuencias en negro para que un segmento pueda ser considerado como anuncio. Si el tiempo entre ambas es mayor que el umbral establecido, el segmento de vídeo es considerado como NO ANUNCIO. El algoritmo propuesto por [36] establece este umbral en 90 segundos. Este método igualmente determina qué cantidad de segmentos de anuncios ocurren de manera

consecutiva para saber si se trata de un bloque de publicidad. Además plantea que un bloque de publicidad no contiene menos de tres anuncios, de lo contrario es declarado como NO ANUNCIO.

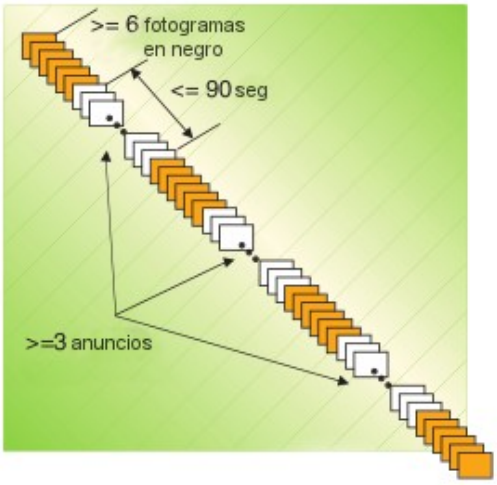

Figura 1.6: Estructura de un bloque de anuncios basado en fotogramas en negro.

Usando 10 grabaciones de diferentes géneros como deporte, noticias y espectáculos, en [36] realizaron la evaluación de su algoritmo con 315 minutos de secuencias de vídeo. Las grabaciones contenían 11 cortes de publicidad y fueron detectadas todas, sin perder ninguno de los contenidos de programación general.

Sin embargo, el algoritmo falló al detectar las partes finales de ciertos bloques, incluyendo incorrectamente partes de NO ANUNCIO. No obstante, de acuerdo a las medidas de *recall,* se puede decir que es una técnica razonablemente buena. De las 11 grabaciones solo una presentó un *recall* por debajo del 85%, mientras que 8 de ellas tuvieron un valor superior al 98%.

1.2.2 Detección basada en la frecuencia de cortes

Otra característica utilizada en la detección de anuncios es la alta frecuencia de cortes o cambios de tomas que se pueden apreciar típicamente en los *spots* comerciales. El problema de determinar la frecuencia de cortes en un segmento de vídeo es básicamente similar a determinar los cambios de tomas. Una vez localizados los cambios de tomas, determinar la frecuencia de cortes se reduce solamente a contar los mismos.

Existen numerosos métodos para la localización de cambios de tomas, muchos se basan en la diferencia del histograma de color o de la detección de bordes entre dos fotogramas consecutivos. Otro

método propuesto en [44] utiliza una métrica de distancia de la Transformada de Wavelet para cuantificar la diferencia entre dos fotogramas e identificar los *shots*.

Un algoritmo muy sencillo propuesto por [26] presenta dos reglas básicas: 1) una secuencia candidata debe tener una frecuencia de cortes superior a 5 cortes por minuto; y 2) la frecuencia de cortes debe ser superior a 30 cortes por minuto en algún punto de la secuencia. Según las pruebas realizadas, este método tuvo un *recall* de 93.43% y una tasa de detecciones falsas de 0.09%, confirmando la idoneidad del uso de la frecuencia de cortes como un pre-filtro para la detección de los bloques de anuncios.

Algunos algoritmos incorporan otras técnicas de edición usadas frecuentemente en los anuncios, tales como disolvencias y *fades*, para indicar la posibilidad de la presencia de comerciales. En [26] se emplean dos métricas adicionales relacionadas con el alto nivel de acción en los anuncios. La primera es el cambio de contornos y la segunda la longitud del vector de movimiento. La detección basada en estas dos métricas presentó un valor de *recall* alrededor del 96%.

Naturalmente, la detección de anuncios basada en las características de estos es más efectiva cuando son consideradas múltiples características en conjunto. En [26] crearon un sistema combinado compuesto por dos fases. Primeramente, se detectan las secuencias de fotogramas en negro y con alta frecuencia de cortes para determinar los segmentos de anuncios candidatos. Luego estos segmentos candidatos son pasados por los detectores de acción (cambio de bordes y longitud del vector de movimiento) para encontrar los límites exactos de los bloques comerciales. La ventaja de este método es que las operaciones más costosas computacionalmente son reservadas para la segunda fase. El empleo de estas fases para la detección de los bloques de anuncios posibilitó obtener valores de *recall* cercanos a un 98%.

1.2.3 Detección basada en la presencia del logotipo del canal

Una práctica generalizada en las emisiones de televisión es la existencia del logotipo o identificador de la cadena de televisión o del programa sobre la señal de vídeo. Normalmente estos identificadores se localizan en una de las cuatro esquinas de la pantalla, aunque lo más común es ubicarlo en la esquina inferior derecha.

Esta técnica se basa en la localización y seguimiento de estos logotipos para distinguir los programas, debido a que por el contrario durante la emisión de los cortes comerciales estos

identificadores no son mostrados. Como declaran en [2], los logotipos suelen ser de tres tipos: animados, opacos o transparentes. Debido a que los primeros no se utilizan mucho, las técnicas desarrolladas solo se han centrado en los otros dos.

Los logotipos opacos son el tipo menos difícil de distinguir. Los valores de sus píxeles se mantienen relativamente constantes, y una vez encontrados son fáciles de monitorear. En [2] proponen un método en el cual los fotogramas del vídeo son inspeccionados para encontrar un área que contenga contornos estables. Ciertos errores son causados por el ruido y pequeños cambios en la posición del identificador. Sin embargo, si un área presenta un contorno estable por un largo período de tiempo se puede asumir la existencia del logotipo en ese lugar. A partir de ese momento se puede generar una máscara del identificador que defina el área del mismo y a partir de esta establecer una medida de similitud para su detección.

La ventaja de esta técnica es que una vez detectada la presencia del identificador durante la emisión de un programa, el problema se reduce solamente a chequear su ausencia para determinar que una sub-secuencia de vídeo es parte de un bloque de anuncio. A pesar de esto presenta una gran desventaja, el problema está dado porque no todas las estaciones de televisión mantienen sus identificadores durante la transmisión de todos sus programas, conllevando entonces a que esta técnica por sí sola falle clasificando los programas como anuncios. Teniendo en cuenta esto se puede decir que esta característica que presentan gran parte de las televisoras en la actualidad, no se pueda explotar únicamente para la detección de publicidad.

Figura 1.7: Ejemplos de identificadores de TV y la detección de uno de ellos.

1.2.4 Detección basada en métodos híbridos

En la literatura consultada se pueden encontrar varios trabajos que plantean métodos híbridos para la detección de anuncios en secuencias de vídeo. Los mismos emplean múltiples técnicas, tales como la integración de múltiples características tanto del vídeo como del audio, características temporales del vídeo y la detección del texto junto al cambio de tomas; siempre en unión a clasificadores de un alto nivel como SVM o HMM. A continuación se dará una breve panorámica de cada uno de ellos, recomendando consultar los respectivos artículos para una mejor comprensión de los mismos.

En [29] plantean detección de anuncios a partir de la fusión sistemática de características temporales del audio y el vídeo. Entre las características utilizadas se encuentran *Audio Class Histogram (ACH)*, *Commercial Pallet Histogram (CPH)*, *Text Location Indicator (TL)*, *Scene Change Rate (SCR)* y *Blank Frame Rate (BFR)*. Para distinguir entre los segmentos de anuncios y programas son empleados clasificadores discriminatorios basados en las múltiples características. Las decisiones son realizadas por diferentes discriminadores fusionados usando SVM. Los experimentos fueron realizados sobre 36 horas de vídeo tomadas de 6 fuentes diferentes. Según los resultados expuestos, el método presenta un 92% de detecciones de anuncios correctas. Sin embargo, presenta la desventaja de haber sido evaluado solo en vídeos compuestos de noticias y comerciales, lo que implica que no necesariamente sea escalable a conjuntos de vídeo diferentes.

En su trabajo [22] definen su método como una detección robusta basada en el aprendizaje. Primeramente son analizadas un conjunto de características básicas que permiten distinguir los comerciales de los programas generales (Veáse Figura 1.8), entre estas se encuentran seis descriptores visuales y cinco referentes al audio, explotando características como el cambio de contornos, la diferencia entre fotogramas, el cambio de tomas y los fotogramas en negro. Luego una serie de características basadas en el contexto, las cuales son más efectivas para la identificación de anuncios, son derivadas de los descriptores básicos, calculando la media de los descriptores básicos para el conjunto de *shots* que se encuentran en el "vecindario" del *shot* que se analiza. Cada uno de estos *shots* es clasificado por un SVM teniendo en cuenta las características derivadas del contexto. Finalmente el resultado de esta clasificación es refinada utilizando post-procesamiento a través de un agrupamiento de escenas y ciertas reglas heurísticas. Los experimentos fueron realizados sobre 10 horas de vídeo de diferentes géneros (noticias, deportes, películas, entretenimiento y programas en general) y diferentes cadenas de televisión (6 horas de NBC, 2 horas de ESPN2 y 2 horas de CNN), mostrando altos niveles de acierto en las detecciones de anuncios. Los resultados generales expuestos para la detección de

anuncios presentan valores de *recall* superiores al 88% sin el post-procesamiento y al 91% realizando el mismo.

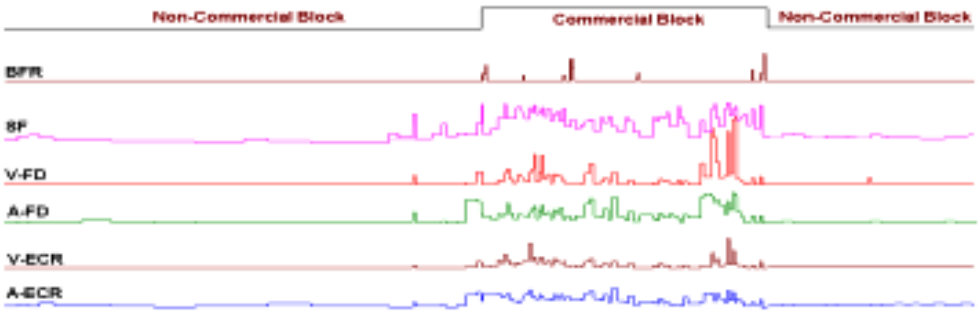

Figura 1.8: Curvas de características usadas en [22].

Un método basado en el cambio de *shot* y la extracción de texto fue propuesto por [28]. Este algoritmo utiliza el cambio de toma como una característica básica para la detección de anuncios. Primeramente se calcula la diferencia de histograma en fotogramas consecutivos y luego se detectan las cuatro transiciones de vídeo más comunes: corte, disolvencia y *fade in/out*. El algoritmo para la detección del texto se basa en la diferencia del gradiente máximo, que permite un rápido filtrado de las líneas sin texto. Según el método desarrollado, un anuncio es detectado si la frecuencia de *shots* es mayor que un umbral o si es detectada información comercial de texto en el vídeo. A pesar que los valores de *recall* expuestos en los resultados (82%) no son tan satisfactorios como los métodos explicados con anterioridad, la principal contribución del este trabajo consiste en explotar la información de texto como una característica más para la detección de anuncios. En la siguiente figura se muestran ejemplos de la detección de texto realizada por el algoritmo descrito.

Figura 1.9: Ejemplos de la detección de texto en anuncios.

1.3 Conclusiones Parciales

Durante este capítulo se desarrolló un análisis de los elementos teóricos que sirven como base a la problemática y objetivos planteados en el presente trabajo. Para ello se realizó un estudio de los principales conceptos asociados al dominio del problema, caracterizando cada uno de ellos de forma tal que se logre comprender el entorno de la investigación y la teoría que la sustenta. Después de analizar las principales técnicas utilizadas en la detección de anuncios, así como sus resultados, se puede resumir que entre las de mayor efectividad se encuentran aquellas que utilizan métodos híbridos donde se fusionan características tanto del vídeo como del audio. Sin embargo, se pudo apreciar la poca cantidad de descriptores de audio que son utilizados, basando sus sistemas en gran medida en características visuales, lo que hace que dichos métodos presenten un alto costo computacional.

CAPÍTULO 2. PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA

La revisión bibliográfica del estado del arte sirvió para adentrarse en los conocimientos del tema y al mismo tiempo como base teórica para la presentación de la solución propuesta. En el presente capítulo se explica el proceso de detección de anuncios desarrollado a partir de la señal de audio de emisiones de televisión. En primer lugar, se describe la técnica propuesta a través de un diagrama de flujo que integra las diferentes fases de la misma. Posteriormente, se caracterizan cada una de las fases del proceso, explicando los métodos utilizados en cada una de ellas. De esta manera se realiza la presentación de la solución a la problemática tratada en el trabajo, consolidando las aportaciones teórico-prácticas de la investigación.

2.1 Descripción de la solución propuesta

La solución al problema de la detección automática de intervalos de publicidad en emisiones de televisión, como ya ha sido explicado, integra en sí misma diferentes fases del procesamiento de vídeo como son captura, segmentación, clasificación y análisis. De igual manera, al ser desarrollada una técnica que utiliza únicamente la señal de audio, se conjugan además los distintos procesos de un sistema genérico de procesamiento de audio (Veáse Figura 1.2).

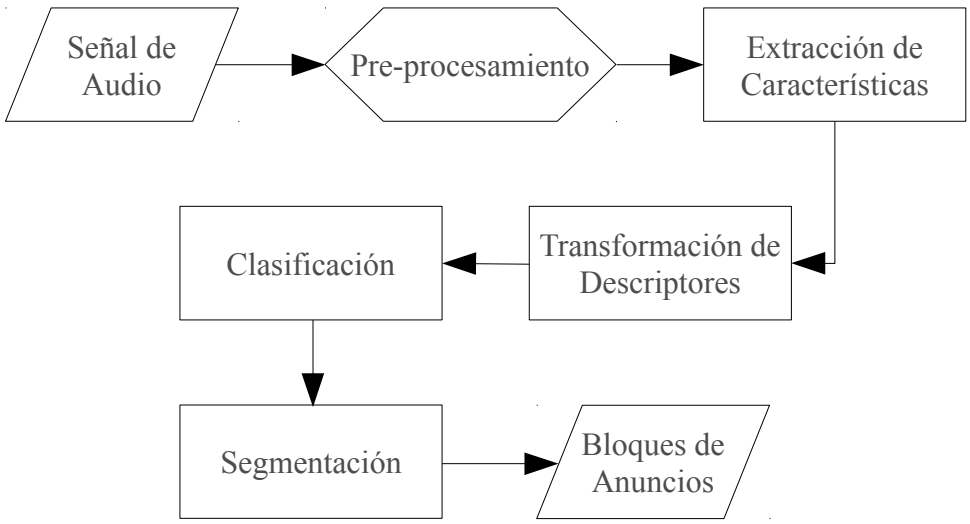

Figura 2.1: Diagrama de flujo de la técnica propuesta.

Vale explicar que al estar orientada a sistemas de monitoreo e indexación de emisiones de televisión, el método propuesto se llevará a cabo en el llamado núcleo del sistema (Veáse Figura 1.4), interviniendo en el mismo los componentes que lo conforman, debido a la estrecha relación de los mismos. Razón por la cual en el diagrama del flujo propuesto se obvian las fases de captura y transcodificación de la señal de vídeo, partiendo entonces de la señal de audio como entrada al sistema.

2.2 Caracterización de las fases del proceso

En la Figura 2.1 se muestra el diagrama de flujo de la técnica propuesta. Está compuesta por cinco fases en las cuales la señal de audio es procesada, analizada, clasificada y segmentada, para dar como salida finalmente los intervalos de los distintos bloques de publicidad. A continuación se caracterizaran las distintas fases, ofreciendo una explicación de las técnicas empleadas en cada una de ellas.

2.2.1 Pre-procesamiento de la señal de audio

La entrada al sistema consiste en una señal de audio sin comprimir en formato WAV, con una frecuencia de muestreo de 24 KHz. Dicha señal necesita pasar por un pre-procesamiento antes que se extraigan las características de las mismas. Esto garantizará que los descriptores obtenidos tengan la calidad requerida para la clasificación de los segmentos de audio.

Como se explicó en el epígrafe relacionado con los descriptores de audio, la señal de audio es susceptible a los medios de transmisión y captura de vídeo. Al utilizar señales provenientes de diferentes fuentes de transmisión, como pueden ser televisión analógica, digital o satelital, así como variadas estaciones de televisión, el audio puede tener variaciones de unas a otras. Por esta razón el primer proceso por el que pasa la señal de audio es la normalización de la misma. La normalización del audio consiste en igualar al máximo el volumen promedio (o la amplitud de la señal) de una misma o varias pistas de audio, para que éstas suenen al mismo volumen general y no se escuchen unos segmentos más altos que otros.

Existen varios métodos para llevar a cabo esta homogeneización de la señal, entre ellos el llamado *Peak Level,* que buscará la máxima amplitud de la onda de audio, le aplicará una reducción para que no distorsione si es necesario y conforme a ese punto de máximo amplificará o reducirá el resto de la onda. Otro de los métodos de normalización es el llamado *Average Output,* en el que se busca la cantidad de sonido de salida para igualar ese volumen en todas las pistas.

Una vez que se ha normalizado la señal de audio, se hace necesario para la extracción de las características, la división de la misma en un conjunto de pequeños clips de audio. En [27] proponen que el audio sea dividido en clips de un segundo de duración, para que posteriormente sean segmentados en ventanas más pequeñas para la extracción de las características básicas a ese nivel de *frame*.

En la Figura 2.2 se muestran las sub-divisiones realizadas en la señal de audio para la posterior extracción de los descriptores. Se utilizan segmentos de audio de un segundo y ventanas de 512 muestras con un solapamiento 384 muestras (3/4 de la longitud).

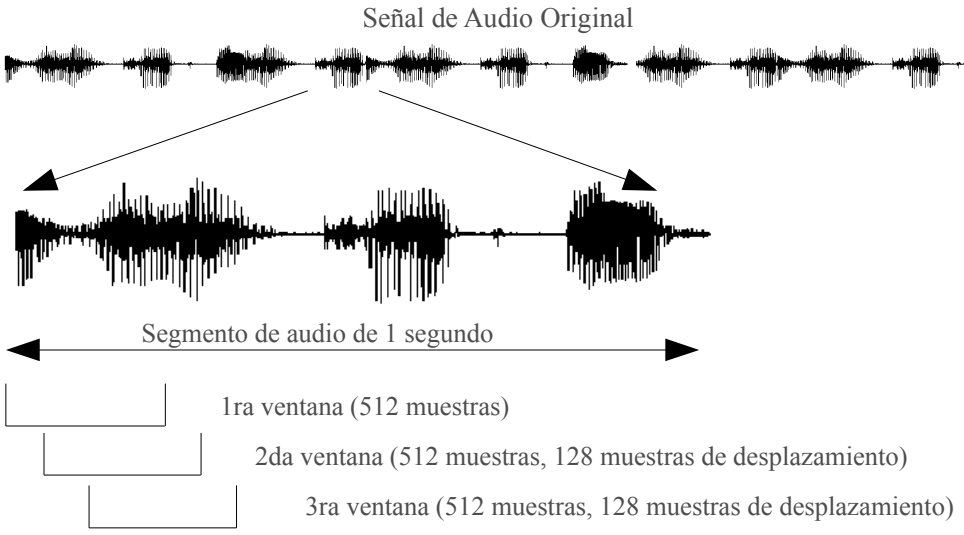

Figura 2.2: Esquema de sub-división de la señal de audio.

2.2.2 Extracción de características

Teniendo en cuenta lo expuesto por varios autores de investigaciones acerca del uso de la señal audio para la segmentación y clasificación [9, 14, 22, 27, 34, 41], y a partir de un estudio previo realizado mediante la caracterización de un conjunto de datos y la realización de pruebas experimentales (Veáse Figuras 2.3 y 2.4), se decidió utilizar catorce descriptores de audio para la caracterización de la señal.

Durante la investigación se realizaron dos fases de pruebas con la utilización de diferentes descriptores. La primera sin tener en cuenta medidas estadísticas de los descriptores, se extrajeron características básicas a nivel de clip y con estos se intentó realizar la clasificación. Esto tenía como objetivo minimizar el costo computacional que implica la extracción de los descriptores a nivel de

ventanas. Sin embargo los resultados no fueron satisfactorios y fue necesario incluir ciertas medidas estadísticas que posibilitaran una mejor caracterización de la señal de audio.

Los descriptores básicos utilizados en la primera fase de pruebas para la extracción de las características fueron los siguientes: *Energy Entropy*, *Short Time Energy*, *Zero Crossing Rate*, *Spectral Flux*, *Spectral Rolloff*, *Spectral Centroid*, *Volume Contour*, *Volume Dynamic Range*, *Pitch Contour* y *Fundamental Frecuency*.

Para la caracterización de la señal de audio fue necesario incluir medidas estadísticas en los descriptores que fueran capaces de mostrar las variaciones que ocurren en la señal a lo largo de su duración. Fueron utilizadas medidas como la media, la varianza, el ratio y la desviación estándar. Los 14 descriptores empleados se describen a continuación.

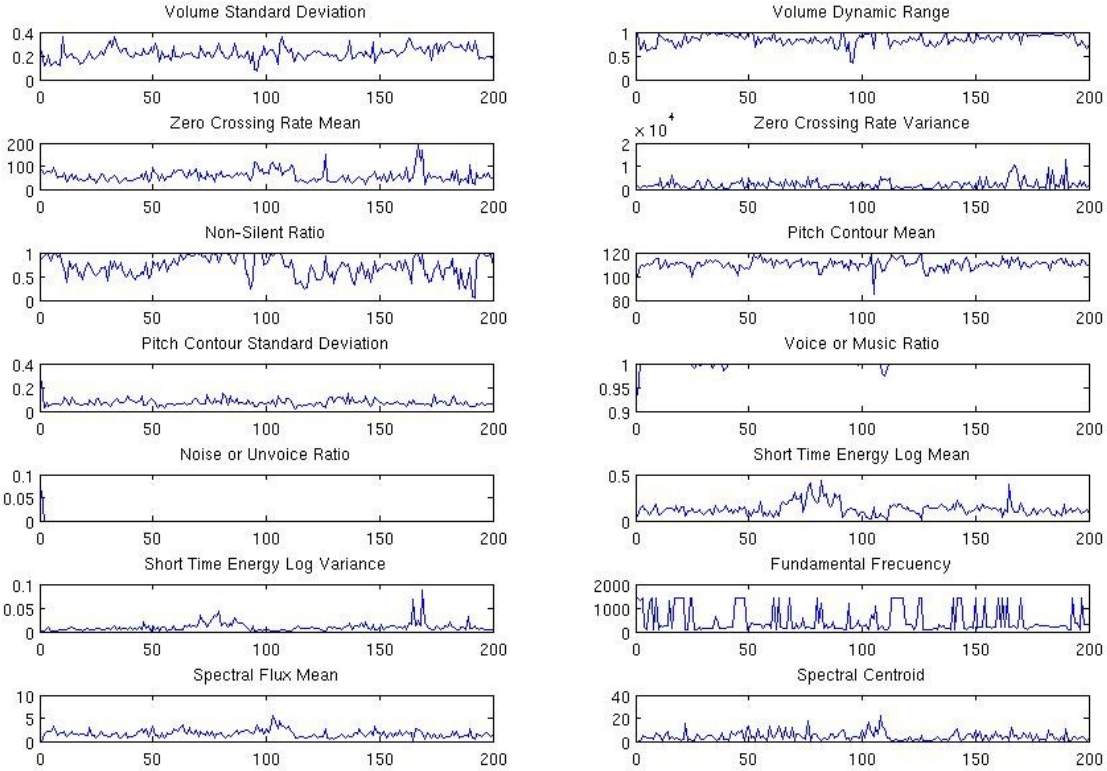

Figura 2.3: Gráficas de los descriptores utilizados de una secuencia de anuncios.

Descriptores basados en el volumen

Una de las características que mejor describe las señales de audio es el volumen. En el caso especial de la detección de anuncios toma un significado mayor debido al aumento de volumen en gran parte de los cortes de publicidad, lo cual implica una variación de este parámetro a lo largo de la señal de audio donde se conjugan bloques de comerciales y de programación en general.

Para medir la variación temporal del volumen se utilizan los dos descriptores propuestos por [27]: *Volume Standard Deviation (VSTD)* y *Volume Dynamic Range (VDR)*.

Descriptores basados en ZCR

Como ya fue explicado el ZCR describe las veces que una señal de audio cruza el Eje X. Esta característica es una simple medida de la frecuencia del contenido de la señal. Por lo general, la variación del ZCR suele ser mucho mayor para señales que contienen habla que las de música. De esta manera esta característica puede ser un buen indicador para la detección de anuncios, además de servir como base para la extracción de otros descriptores.

Fueron calculadas dos medidas estadísticas del ZCR, que a pesar de estar correlacionadas ofrecen informaciones diferentes acerca de la composición del audio. Los descriptores utilizados fueron: *Zero Crossing Rate Mean (ZCRM)* y *Zero Crossing Rate Variance (ZCRV).*

Descriptores basados en la continuidad del audio

La continuidad de la señal de audio es una característica fundamental de los bloques de anuncios. Esto se debe a que normalmente los *spots* comerciales están acompañados de una música de fondo y muchos de ellos tienen descripciones habladas de los productos que se desean promover.

Para caracterizar la continuidad de la señal de audio se utilizaron los descriptores *Non-Silent Ratio (NSR)*, *Voice or Music Ratio (VMR)* y *Noise or Unvoice Ratio (NUR).*

Descriptores basados en la frecuencia

Los anuncios por lo general se componen de sonidos variados caracterizados por el dinamismo. Lo anterior y la corta duración que tienen los *spots* de publicidad hacen que la señal de audio presente frecuencias diferentes a la programación en general.

Los descriptores utilizados para representar las variaciones de la frecuencia de la señal de audio son *Pitch Contour Mean (PCHM)*, *Pitch Contour Standard Deviation (PSTD)* y *Fundamental Frecuency (FF)*.

Descriptores basados en la energía

La energía de una señal de audio describe la variación de la amplitud de la misma. Debido a que esta característica puede ser usado como una medida para distinguir el sonido y el silencio, se propone utilizar la media y la varianza de la energía (*Short Time Energy Log Mean (STEM)* y *Short Time Energy Log Variance (STEV)*) como dos descriptores más para la detección de anuncios.

Descriptores basados en el espectro

Las características espectrales del audio describen el timbre de la señal y son muy utilizadas para distinguir el habla de la música o sonidos en general. Teniendo en cuenta que los anuncios comúnmente están acompañados de música y sonidos se proponen los descriptores *Spectral Flux Mean (SFM)* y *Spectral Centroid (SC)*.

2.2.3 Transformación de los descriptores

A partir de la extracción de los descriptores de características, se obtiene un vector de características para cada clip de un segundo que compone la secuencia de audio en su totalidad. Estos vectores de características serán los que pasarán posteriormente al clasificador, pero antes es necesaria la transformación de esos descriptores.

La gran mayoría de los clasificadores se basan en cálculos estadísticos o matemáticos que solo trabajan en el dominio numérico. Trabajar con datos numéricos en dominios de rangos diferentes puede acarrear algunas dificultades, tales como que las características no sean homogéneas, que se necesite más memoria para su almacenamiento o simplemente que se tengan dificultades para realizar los cálculos. Por estas razones el primer paso a seguir en la transformación de los datos es la normalización o escalado.

Ninguna de las características usadas presenta valores en el dominio de los negativos, por tal razón se emplea un escalado de los datos al rango (0,1). El escalado de los datos implica que se utilice el mismo factor de escalado tanto para los datos de entrenamiento como para los de prueba.

El otro proceso de transformación de los datos llevado a cabo es la sumarización, con el objetivo de reducir la dimensionalidad de los mismos y obtener componentes de un mayor grado de varianza. La reducción de la dimensionalidad de los datos, a pesar que solo se tienen 14 descriptores, posibilita minimizar el costo computacional del proceso de entrenamiento y clasificación. Por otra parte, tener datos con más varianza implica que tengan un mayor poder descriptivo.

Para este propósito se lleva a cabo un Análisis de Componentes Principales (PCA, por sus siglas en inglés), que es un método bien riguroso para la simplificación del conjunto de datos. Este método genera un nuevo conjunto de variables, llamado *componentes principales*. Cada componente de estas es una combinación lineal de las variables originales, sin contener información redundante entre ellas. Las componentes generadas se ordenan por su valor de varianza.

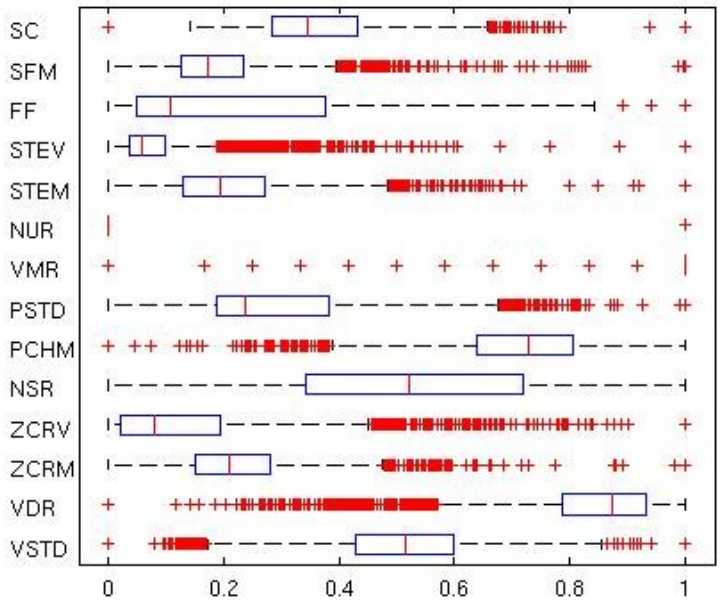

Figura 2.4: Gráfica que muestra la varianza de los descriptores originales.

Después del estudio PCA realizado, de acuerdo a la varianza de las componentes resultantes y a los resultados de los experimentos con diferentes números de componentes, se llegó a la conclusión de utilizar las 8 primeras componentes para el proceso de clasificación.

2.2.4 Clasificación de segmentos de audio

La implementación del método de clasificación de la señal de audio se realizó a partir de los vectores de características obtenidos y los intervalos reales de la detección de publicidad. Como primer

paso del estudio, se implementó un clasificador lineal utilizando el análisis discriminante de Fisher con el objetivo de determinar si los conjuntos eran linealmente separables; resultando que no eran tan separables como para obtener buenos resultados en la clasificación. De esta manera se utilizó entonces un clasificador SVM que posee características más apropiadas para la clasificación de conjuntos de datos que no sean linealmente separables y que son utilizados en varios de los trabajos consultados en el estudio del estado del arte [2, 7, 13, 22, 25, 38].

Análisis discriminante lineal

El análisis discriminante se ocupa de describir los rasgos diferenciales entre las clases. Se trata de encontrar funciones discriminantes o reglas de decisión, cuyos valores en las distintas clases estén lo más separados posible. Es decir, se buscan funciones sencillas que permitan asignar cada uno de los objetos a una categoría concreta minimizando la tasa de error en dicha asignación. En este caso particular se utilizó el discriminante lineal de Fisher [11], donde se utiliza una función lineal como discriminante y es la regla más conocida para estos fines.

Para el discriminador lineal el criterio geométrico, consiste en asignar el objeto ω, con vector de características $x=(x_1,..., x_n)$, a la clase más próxima, utilizando la distancia de Mahalanobis:

$$\delta_M^2(x, \mu_i)=(x-\mu_i)' \sum^{-1} (x-\mu_i), \qquad i=1,2$$

Donde $\boldsymbol{\mu_1}$ y $\boldsymbol{\mu_2}$ son los vectores de medias de las clases $\boldsymbol{\Omega_1}$ y $\boldsymbol{\Omega_2}$ respectivamente; Σ es la matriz de covarianzas común para ambas clases.

La regla de decisión que se emplea es la siguiente:

1. ω se asigna a Ω_1 si $\delta_M(x, \mu_1) < \delta_M(x, \mu_2)$.
2. ω se asigna a Ω_2 en caso contrario.

Fisher por su parte plantea que a partir de la diferencia $\delta_M(x, \mu_2) - \delta_M(x, \mu_1)$, se construye la función discriminante lineal :

$$L(x)=(x-\frac{1}{2}(\mu_1+\mu_2))' \sum^{-1} (\mu_1-\mu_2)$$

Expresando la regla de decisión en función de ésta:

1. ω se asigna a Ω_1 si $L(x) > 0$.

2. en caso contrario se asigna ω a Ω_2.

En las pruebas de clasificación se obtuvo valores de *recall* por debajo de un 55% y de *precision* inferiores al 40%. Estos resultados demostraron que los conjuntos no son linealmente separables, por lo que se decidió el uso de un clasificador más robusto como SVM.

Clasificación basada en SVM

A continuación se dará una breve introducción sobre algunos de los conceptos básicos de SVM. Para más detalles se puede consultar el trabajo de [5]. Como todo clasificador binario, la naturaleza de las Máquinas de Soporte Vectorial se basa en encontrar un hiper-plano discriminante que separe de manera óptima dos clases, a partir de un conjunto de ejemplos positivos y negativos.

Un SVM puede ser lineal o no lineal, para esto se emplea una función *kernel* K(x,v). En este caso los vectores de entrenamiento son mapeados en un espacio de mayores dimensiones donde los datos son linealmente separables. Los investigadores se han dedicado a proponer nuevos *kernels* para estos propósitos, sin embargo existen cuatro básicos que se pueden utilizar fácilmente: el lineal, el polinomial, el sigmoid y la función de base radial.

Para el presente trabajo se utilizó el *kernel* RBF, sobre la base en primer lugar que es el más usado para casos que no son linealmente separables. Otra razón de gran peso para la selección es el número de hiper-parámetros que influye en la complejidad del modelo y las escasas dificultades numéricas; esto a diferencia de otros *kernels* como el polinomial. También se tuvo en cuenta que el número de descriptores utilizados no es muy grande, porque de otro modo RBF no podría ser usado.

Esquema de clasificación

En la fase anterior del modelo propuesto, se obtuvo un vector de ocho características que representa cada segmento de audio de un segundo de duración. Cada uno de estos vectores es empleado para el entrenamiento del SVM y la posterior clasificación de segmentos de prueba.

Una vez seleccionado el *kernel* a utilizar y antes de pasar al entrenamiento del clasificador, es necesario encontrar los valores óptimos de los dos parámetros de RBF (C y γ), para los cuales el clasificador puede predecir correctamente segmentos desconocidos. En SVM se implementa una técnica llamada *cross-validation*, que consiste en dividir los datos de entrenamiento en diferentes subconjuntos para efectuar un entrenamiento del modelo y probarlo al mismo tiempo, prediciendo de este modo el porciento de efectividad del propio modelo.

Para encontrar dichos valores óptimos se utilizó una búsqueda secuencial utilizando la técnica de *cross-validation*. Varios pares de C y γ fueron testeados secuencialmente y los de mejor efectividad fueron los seleccionados. Se realizó entonces una búsqueda exhaustiva, empleando un método muy práctico para identificar estos parámetros; el cual consiste en probar con una secuencia exponencial, por ejemplo: $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ y $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}$. Finalmente se obtuvieron los valores $C = 2048$ y $\gamma = 0.5$; con una efectividad del 95.76%.

Utilizando dichos valores de C y γ fue entrenado el modelo con un conjunto bastante amplio de segmentos de audio provenientes de diferentes fuentes y géneros. A partir de dicho entrenamiento el sistema quedó capacitado para clasificar nuevas secuencias de audio, dando como salida cada uno de los segmentos que la componen clasificados en ANUNCIO o NO ANUNCIO. Por esta razón es que se introduce una última fase en el algoritmo, la cual se ha llamado Segmentación, para que se encargue de la detección de los bloques de publicidad en la secuencia de audio.

2.2.5 Segmentación de intervalos de anuncios

Como se debe saber, es inevitable que ciertos errores sean cometidos en los procesos de clasificación automáticos. Los clasificadores basados en SVM, a pesar de ser muy robustos, no están exentos de estos. Por esta razón, son necesarios mecanismos de post-procesamiento que realicen un refinamiento del resultado de la clasificación antes de definir los bloques finales de anuncios.

Los anuncios se caracterizan por ser transmitidos en forma de bloques continuos que contienen varios *spots* comerciales, apareciendo de manera alterna entre dos bloques de programación general y viceversa (un bloque general aparece entre dos cortes de publicidad). El objetivo del trabajo se centra en la detección de estos bloques de publicidad, debido a esto no es necesaria la segmentación de cada uno de los anuncios que lo conforman, sino del bloque como tal. La duración de un bloque normalmente es de 5 a 15 minutos, por lo que se puede pensar que la detección de bloques de una duración mucho menor o mayor que esta sería incorrecta. De manera análoga sucede con los bloques de programación general, los cuales comúnmente son de una duración mayor a 15 minutos. Teniendo en cuenta lo anterior, en las secuencias resultantes del clasificador no podrían existir bloques similares a los que se muestran en la Figura 2.6.

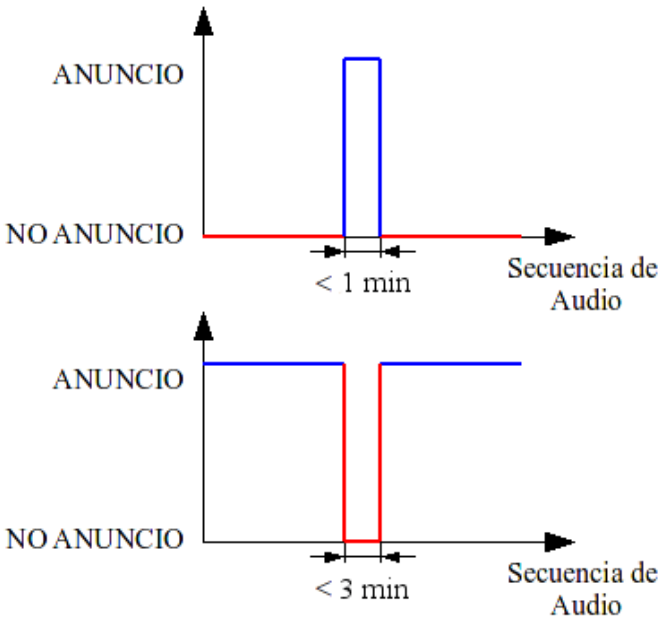

Figura 2.5: Esquemas de secuencias de bloques erróneas.

Basado en estas observaciones, se proponen la siguiente regla de post-procesamiento para refinar los resultados de la detección de los bloques comerciales.

1. *Eliminar los bloques pequeños*: si se encuentra un segmento clasificado con una duración menor a los umbrales que se muestran en la Figura 2.6 (1 minutos para anuncios y 3 minutos para programación general), el mismo será convertido a la clasificación opuesta.

Después de aplicar el post-procesamiento a los resultados de clasificación se está en condiciones de definir la segmentación de la secuencia de vídeo clasificada. En este punto se obtendría como resultado los intervalos de tiempo de los bloques de anuncios que serán determinados. Un bloque de anuncio lo conforman un conjunto consecutivo de segmentos clasificados como anuncio. Por tanto los intervalos de los bloques se determinarán a partir del tiempo relativo a la secuencia de audio clasificada (que coincide con la duración del vídeo) del primer y último segmento que conforma el bloque. De esta manera la secuencia de vídeo podría ser segmentada por el componente para este fin dentro del sistema de monitoreo e indexación de televisión.

2.3 Conclusiones Parciales

Durante la definición de una técnica de reconocimiento de patrones resulta de vital importancia el proceso de selección de descriptores para la representación y comparación de las entidades. En este capítulo se logró definir un conjunto de descriptores capaces de caracterizar una secuencia de audio para

su clasificación en intervalos de anuncio. Logrando además consolidar un conjunto de técnicas para el desarrollo de un método de detección de anuncios sobre señales de audio. La solución al problema planteado ha representado un gran reto para el autor del trabajo, pero sin dudas es solo el inicio de un largo camino por recorrer; habrá que seguir investigando para lograr toda una metodología eficiente y con un bajo costo computacional para la detección de anuncios en emisiones de televisión.

CAPÍTULO 3. ANÁLISIS DE LOS RESULTADOS

Comprobar el comportamiento y la calidad del proceso de clasificación de una técnica propuesta es un paso fundamental en la validación de la misma. En el presente capítulo se exponen los resultados obtenidos a partir de pruebas realizadas al método de detección de anuncios propuesto en la presente investigación. Se plantean las herramientas utilizadas para la realización de los experimentos, las medidas de eficiencia utilizadas y se analizan los resultados.

3.1 Herramientas de ensayo

Para realizar las pruebas prácticas con las técnicas propuestas se utilizó el software MATLAB R2009a sobre Sistema Operativo GNU/Linux Ubuntu 10.04 con *kernel* 2.6.32. El hardware utilizado fue una computadora de escritorio con procesador Intel Core Duo a 1.86 Ghz y 1 Gb de memoria RAM. Para la extracción de los descriptores de las secuencias de audio se utilizó el Superdome del Centro de Supercomputación y Bioinformática (SCBI) de la Universidad de Málaga, que contiene 128 núcleos Itanium2 y 400 Gbytes de memoria compartida entre todos los núcleos. La utilización de este supercomputador posibilitó la paralelización del proceso de extracción de características, debido a la cantidad de horas de medias con que se contaba para los experimentos.

MATLAB es el nombre abreviado de “Matrix Laboratory”, es un programa para realizar cálculos numéricos con vectores y matrices (la mejor y más eficiente forma de trabajar las imágenes y el audio). Como caso particular se puede también trabajar con números escalares, tanto reales como complejos, con cadenas de caracteres y con otras estructuras de información más complejas. Una de las capacidades más atractivas es la de realizar una amplia variedad de gráficos en dos y tres dimensiones. MATLAB tiene también un lenguaje de programación propio.

Dentro de la instalación del MATLAB se cuenta con una serie de herramientas adicionales (*toolbox*) para el trabajo en diferentes campos y aplicaciones matemáticas con una gran variedad de funciones implementadas que optimizan mucho el trabajo para el cual se esté usando el software. La mayoría de las funciones están escritas en el lenguaje abierto MATLAB, que brinda la posibilidad de inspeccionar los algoritmos, modificar el código fuente y crear sus propias funciones personalizadas.

Para la implementación del clasificador SVM se utilizó la librería **LIBSVM** (Versión 2.91) creada por Chih-Chung Chang and Chih-Jen Lin del Departamento de Ciencias de la Computación de la Universidad Nacional de Taiwan. [10]

3.2 Medidas de eficiencia

En las estadísticas, los términos **falso positivo** y **falso negativo** son usados para describir posibles errores en los procesos de decisión estadísticos tales como los clasificadores. En un algoritmo de detección de anuncios se produce un falso positivo cuando se concluye haber detectado un segmento de anuncio que no lo es. Mientras que se produce un falso negativo cuando se concluye que un segmento no es de anuncio y sí lo es.

Análogamente se definen los términos **verdadero positivo** y **verdadero negativo** para describir las decisiones correctas del proceso de clasificación. Un verdadero positivo se produce cuando se detecta un segmento de anuncio correctamente. Por otra parte, cuando se clasifica un segmento de no anuncio de manera correcta se dice que es un verdadero negativo.

Condición del Segmento de Audio		**ANUNCIO**	**NO ANUNCIO**
Resultado de clasificación	**Positivo**	Anuncio + Positivo = **Verdadero Positivo**	No Anuncio + Positivo = **Falso Positivo**
	Negativo	Anuncio + Negativo = **Falso Negativo**	No Anuncio + Negativo = **Verdadero Negativo**

Tabla 3.1: Términos usados para describir los resultados de clasificación .

A pesar que la técnica propuesta tiene como objetivo la detección de intervalos de publicidad, para la medición de la eficiencia se tomará cada uno de los segmentos que componen estos intervalos. De esta manera las mediciones no se realizan como tal al proceso de clasificación, el cual se basa precisamente en los segmentos.

La eficiencia máxima se consigue cuando este es capaz de procesar un conjunto de datos sin producir ningún falso positivo o falso negativo. Esto se traduce en que los intervalos de anuncios detectados son correctos en un 100% y no debe quedar ningún segmento de anuncio sin detectar. En la literatura existen diversas técnicas para medir la eficiencia de estos algoritmos, aunque una de ellas destaca sobre las demás por su amplia utilización.

Se trata de las medidas de *recall* y *precision* definidas como:

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

$$Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

donde *recall* expresa el porcentaje de segmentos de anuncios correctamente detectados de entre todos los existentes, mientras que *precision* ofrece una medida de la fiabilidad de la detección, que significa cuántos de los efectos detectados son realmente correctos y no falsos positivos.

Un algoritmo plenamente eficaz será aquél que consiga un valor de 1 en ambas medidas, aunque dada la dificultad de esta tarea se pueden considerar satisfactorios unos resultados donde *recall* y *precision* estén equilibrados y próximos a 1.

Otras medidas posible son las utilizadas por [31]: *True Negative Rate (TNR)* y *Accuracy*.

$$TNR = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}$$

$$Accuracy = \frac{Verdaderos\ Positivos + Verdaderos\ Negativos}{Total\ de\ Segmentos}$$

La primera refleja una medida de la cantidad de segmentos de programación general detectados sobre el total de segmentos existentes de este tipo; mientras que la segunda métrica expresa un porcentaje de todos los segmentos detectados correctamente sean anuncio o no.

Una medida que combina *recall* y *precision* es la media harmónica de estas, se trata de la conocida medida F_1 y se define como:

$$F_1 = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}$$

Las medidas definidas anteriormente son más apropiadas para caracterizar los resultados de la clasificación, sin embargo no se adecuan totalmente para comprobar los resultados de la segmentación. Por esta razón se utilizó una métrica complementaria para este propósito. La medida empleada para caracterizar la calidad de la segmentación es una adaptación de la definida por [33] y conocida por *WindowDiff (WD)*, la cual es ampliamente utilizada en el campo de la segmentación de texto y es más

apropiada para determinar las discrepancias que aparecen entre el *ground truth* y el resultado de la segmentación.

WindowDiff funciona de la siguiente manera: explora cada posición del conjunto de datos comparando el número de límites de segmentación encontrados en el intervalo (r_i) del *ground truth* frente a la cantidad de límites que son determinados por el algoritmo (a_i). El algoritmo es penalizado si $r_i \neq a_i$ (calculado como $|r_i - a_i| > 0$). *WD* se define formalmente como:

$$WD(G, S) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(G_i, G_{i+k}) - b(S_i, S_{i+k})| > 0)$$

Donde **b(i, j)** representa el número de límites de segmentación entre las posiciones **i** y **j**, **N** es el número total de segmentos y **k** es el tamaño de la ventana utilizada para realizar la búsqueda de los límites.

3.3 Experimentos

En los experimentos realizados para la validación de la técnica propuesta se utilizó un conjunto de datos ofrecidos por la empresa TEDIAL, líder en el mercado de los sistemas de información audiovisual. En total se utilizaron más 30 horas de vídeo grabados de 13 canales de la televisión venezolana (11) y española (2); en la Tabla 3.1 están relacionadas las horas de vídeo utilizadas por canal según el género.

Canal TV	General	Noticia	Entrevista	Deporte	Serie	Documental
Antena 3	1	1	1	0	0	0
AnTV	2	1	0	0	0	0
Canal-I	1	1	1	0	0	0
Globovisión	2	3	0	0	0	0
LaTele	1	0	0	0	2	0
Meridiano	1	0	0	2	0	0
Telecinco	1	0	0	0	1	0
Telesur	1	0	0	0	0	0
Televen	0	0	0	0	2	0
TVES	3	0	0	0	0	0
ValeTV	0	0	0	0	0	2
Venevision	0	0	0	0	1	0
ViVe	0	0	0	0	0	1
Totales	**13**	**6**	**2**	**2**	**6**	**3**

Tabla 3.2: Distribución por canales de las horas de vídeo utilizadas según el género.

Las emisiones de televisión fueron grabadas y segmentadas en secuencias de una hora por el sistema informático TdTARSYS desarrollado por la misma empresa. La señal de audio fue extraída de las secuencias de vídeo con las siguientes características:

- Codificación: PCM
- Canales: Estéreo
- Frecuencia de muestreo: 24000 Hz
- Bits por muestra: 16

Se realizaron tres tipos de experimentos teniendo en cuenta el origen de los datos:

1. Todos los datos disponibles sin tener en cuenta ningún agrupamiento. El conjunto de datos fue divido en 3/4 para entrenamiento (24 horas) y 1/4 para prueba (8 horas). Los datos de entrenamiento contienen 68 bloques de anuncios de diferentes duraciones con un total aproximado 4 horas. En el conjunto de prueba aparecen 1.6 horas de anuncios divididas en 30 bloques. En la Tabla 3.2 se puede apreciar la distribución de segmentos de anuncios y programación general utilizados (Veáse Totales). Se organizaron por género para ofrecer una mayor idea de la representación de los datos.

	Entrenamiento				Prueba			
Género	**Segmentos de Anuncios**	**Bloques de Anuncios**	**Segmentos de No Anuncios**	**Bloques de No Anuncios**	**Segmentos de Anuncios**	**Bloques de Anuncios**	**Segmentos de No Anuncios**	**Bloques de No Anuncios**
General	5596	32	30404	41	2624	12	8176	14
Noticia	826	4	17174	9	641	3	2959	4
Entrevista	513	4	3087	5	521	3	3079	3
Deporte	1172	6	2428	6	682	4	2918	5
Serie	4649	17	13351	22	621	3	2979	4
Documental	1221	5	5979	7	794	5	2806	5
Totales	**13977**	**68**	**72423**	**90**	**5883**	**30**	**22917**	**35**

Tabla 3.3: Cantidad de segmentos y bloques utilizados para los experimentos generales.

2. Se agruparon los datos según la estación de televisión. Los datos fueron divididos en 2/3 para entrenamiento y 1/3 para las pruebas. Solo se pudieron utilizar medias de siete de los canales disponibles por no haber suficientes datos de los restantes para el entrenamiento (2 horas) y las pruebas (1 hora). Este tipo de experimento se realizó para tener una idea de la variabilidad de los resultados dependiendo de la fuente de los datos. Las cantidades de segmentos y bloques utilizados se pueden apreciar en la Tabla 3.3.

	Entrenamiento				Prueba			
Canal TV	**Segmentos de Anuncios**	**Bloques de Anuncios**	**Segmentos de No Anuncios**	**Bloques de No Anuncios**	**Segmentos de Anuncios**	**Bloques de Anuncios**	**Segmentos de No Anuncios**	**Bloques de No Anuncios**
Antena 3	701	2	6499	3	521	3	3079	3
AnTV	138	2	7062	4	81	1	3519	2
Canal-I	1625	8	5575	10	826	4	2774	5
Globovisión	1483	7	5717	9	897	4	2703	5
LaTele	1505	8	5695	10	1132	4	2468	5
Meridiano	1219	8	5981	10	1172	6	2428	6
TVES	1137	7	6063	9	1124	5	2476	6
Totales	**7808**	**42**	**42592**	**55**	**5753**	**27**	**19447**	**32**

Tabla 3.4: Cantidad de segmentos y bloques utilizados de acuerdo a la fuente de emisión.

3. Para verificar la eficacia de la técnica con respecto a diferentes géneros del audio las medias fueron agrupadas según su género. Debido a que de todos los géneros no se tenían la misma cantidad de horas, fue necesario utilizar diferentes cantidades de datos para el entrenamiento y las pruebas realizadas. La Tabla 3.3 muestra las cantidades de datos utilizados para el ensayo según el género, mientras que en la Tabla 3.5 se puede apreciar la cantidad de horas utilizadas por cada género.

Género	Entrenamiento (Hrs)	Prueba (Hrs)
General	10	3
Noticia	5	1
Entrevista	1	1
Deporte	1	1
Serie	5	1
Documental	2	1
Totales	**24**	**8**

Tabla 3.5: Cantidad de horas de audio utilizadas por cada género.

En los siguientes sub-epígrafes se ofrecen algunos detalles acerca de la implementación de las principales fases del algoritmo para la realización de los experimentos.

3.3.1 Extracción de los descriptores

Se implementaron varios *scripts* encargados del procesamiento de las medias, los dos principales fueron **mainComputeMedia** y **mainComputeDir** responsables de dirigir todo el proceso de extracción de los descriptores a una media específica o a un directorio completo respectivamente. Asociados a estos se encuentran **computeAllStatistics**, **allFeatures**, **prevFeatures**, **featuresPCA** y **createLabelVectors** y cuyas funciones pueden ser consultadas en la siguiente tabla.

Nombre	Descripción
computeAllStatistics	Se encarga de calcular el vector de características de cada uno de los segmentos que componen una secuencia de audio. Lee la secuencia en segmentos de un segundo, determina los valores de los descriptores, conforma el vector de característica y los agrupa en anuncio y no anuncio.
allFeatures	Determina los valores de cada descriptor para un segmento de un segundo, realizando los cálculos estadísticos de cada descriptor básico.
prevFeatures	Calcula los descriptores básicos sobre los tamaños de ventanas utilizados para segmentos de un segundo.
featuresPCA	Realiza el PCA para los vectores de características obtenidos.
createLabelVectors	Agrupa los vectores de características en anuncio y no anuncio con el objetivo de ser usados para el proceso de entrenamiento y crea un vector de etiquetas utilizando para ello el *ground truth*.

Tabla 3.6: Funciones implementadas para el proceso de extracción de características.

3.3.2 Entrenamiento

El conjunto de datos para el entrenamiento se obtiene a partir del proceso anterior donde se separan los datos en anuncio y no anuncio para ser utilizados por el clasificador gaussiano, y se crea un vector de etiquetas para ser usado en el clasificador SVM. Se implementaron los *scripts* **gaussianTrain** y **svmTrain** mediante los cuales se determinan los modelos de clasificación a utilizar. Para el entrenamiento del clasificador SVM se realizó una búsqueda exhaustiva de los parámetros C y γ utilizando la técnica *cross-validation* (Veáse Figura 3.1).

En ambos son guardados los modelos matemáticos obtenidos para ser utilizados como entradas de la implementación de los clasificadores.

```
% Best parameter selection - Exhaustively evaluate (coarse) -> -v option
bestcv = 0;
bestc = 0;
bestg = 0;
for log2c =11:40 %21:40 % 0:12%J -1:3,
 for log2g = -1:5 %6  %0:15%J -4:1,
     cmd = ['-q -v 5 -c ', num2str(2^log2c), ' -g ', num2str(2^log2g)]; % '-v' option is specified
     cv = svmtrain_lib(labels, data_scaled, cmd);
     if (cv >= bestcv),
         bestcv = cv;
         bestc = 2^log2c;
         bestg = 2^log2g;
     end
     fprintf('%g %g %g (best c=%g, g=%g, rate=%g)\n', log2c, log2g, cv, bestc, bestg, bestcv);
  end
end
```

Figura 3.1: Fragmento de código que muestra la selección de los parámetros de SVM.

3.3.3 Clasificación

Igualmente que para el entrenamiento se desarrollaron dos funciones **gaussianClassifier** y **svmClassifier**, responsables de los procesos de clasificación por ambos métodos. Para el clasificador gaussiano se utilizó la función del Discriminante Lineal de Fisher explicada en el Epígrafe 2.2.4 (Veáse Figura 3.2). Mientras que para el clasificador SVM se utilizó el modelo RBF obtenido durante el proceso de entrenamiento.

```
for i = 1:length(classX)

    % Función del Discriminante Lineal de Fisher
    L = (classX(i,:) - 1/2*(meanA + meanB))/Sp*(meanA - meanB)';

    if (L > 0)
        classR(i) = 1;
    else
        classR(i) = 0;
    end
end
```

Figura 3.2: Fragmento de código que muestra la utilización de FLD.

Los resultados de la clasificación fueron pasados por el filtrado que se describió en el Epígrafe 2.2.5 con el objetivo de eliminar los bloques pequeños, para este fin se implementó **classificationFilter**. Finalmente se implementó la función **classificationQuality** para calcular las distintas métricas que caracterizan la eficiencia de las pruebas.

3.4 Resultados

Los resultados de los experimentos fueron evaluados utilizando los conceptos relacionados en la Tabla 3.1 y las métricas de eficiencia definidas con anterioridad en el Epígrafe 3.2.

A continuación se reflejan los resultados obtenidos a través de un conjunto de tablas y gráficos, ofreciendo comentarios acerca de los mismos.

	Resultados				Medidas de Eficiencia				
Género	**Verdaderos Positivos**	**Falsos Positivos**	**Verdaderos Negativos**	**Falsos Negativos**	**Recall**	**Precision**	**TNR**	**F1**	**Accuracy**
General	1539	981	7195	1085	0,5865	0,6107	0,8800	0,5984	0,8087
Noticia	297	570	2389	344	0,4633	0,3426	0,8074	0,3939	0,7461
Entrevista	305	540	2539	216	0,5854	0,3609	0,8246	0,4466	0,7900
Deporte	233	685	2233	449	0,3416	0,2538	0,7653	0,2913	0,6850
Serie	288	402	2577	333	0,4638	0,4174	0,8651	0,4394	0,7958
Documental	379	399	2407	415	0,4773	0,4871	0,8578	0,4822	0,7739
Todos	3041	3577	19340	2842	0,5169	0,4595	0,8439	0,4865	0,7771

Tabla 3.7: Resultados con Clasificador Gaussiano teniendo en cuenta el género.

	Resultados				Medidas de Eficiencia				
Género	**Verdaderos Positivos**	**Falsos Positivos**	**Verdaderos Negativos**	**Falsos Negativos**	**Recall**	**Precision**	**TNR**	**F1**	**Accuracy**
General	2321	293	7883	303	0,8845	0,8879	0,9642	0,8862	0,9448
Noticia	566	76	2883	75	0,8830	0,8816	0,9743	0,8823	0,9581
Entrevista	457	77	3002	64	0,8772	0,8558	0,9750	0,8664	0,9608
Deporte	601	101	2817	81	0,8812	0,8561	0,9654	0,8685	0,9494
Serie	548	85	2894	73	0,8824	0,8657	0,9715	0,8740	0,9561
Documental	705	99	2707	89	0,8879	0,8769	0,9647	0,8824	0,9478
Todos	5198	731	22186	685	0,8836	0,8767	0,9681	0,8801	0,9508

Tabla 3.8: Resultados utilizando SVM teniendo en cuenta el género.

En los resultados anteriores se puede apreciar las diferencias sustanciales que existen entre la utilización de un Clasificador Gaussiano y SVM para estos propósitos, siendo el segundo el de mayor efectividad.

Es significativo reflejar que para los distintos tipos de géneros los resultados se mantuvieron relativamente estables. Teniendo en cuenta que los descriptores propuestos ofrecen resultados satisfactorios para diferenciar segmentos de anuncio de segmentos de un género específico, lleva a pensar que los conceptos utilizados para la detección de anuncios en este trabajo pueden ser objeto de estudios futuros para su utilización en la detección de distintos tipos de escenas utilizando igualmente la señal de audio.

	Resultados				Medidas de Eficiencia				
Canal TV	**Verdaderos Positivos**	**Falsos Positivos**	**Verdaderos Negativos**	**Falsos Negativos**	**Recall**	**Precision**	**TNR**	**F1**	**Accuracy**
Antena 3	279	347	2732	242	0,5355	0,4457	0,8873	0,4865	0,8364
AnTV	33	179	3340	48	0,4074	0,1557	0,9491	0,2253	0,9369
Canal-I	453	341	2433	373	0,5484	0,5705	0,8771	0,5593	0,8017
Globovisión	377	567	2136	520	0,4203	0,3994	0,7902	0,4096	0,6981
LaTele	678	513	1955	454	0,5989	0,5693	0,7921	0,5837	0,7314
Meridiano	635	896	1532	537	0,5418	0,4148	0,6310	0,4698	0,6019
TVES	592	197	2279	532	0,5267	0,7503	0,9204	0,6189	0,7975
Valores Medios	3047	3040	16407	2706	0,5113	0,4722	0,8353	0,4790	0,7720

Tabla 3.9: Resultados con Clasificador Gaussiano teniendo en cuenta la fuente de emisión.

	Resultados				Medidas de Eficiencia				
Canal TV	**Verdaderos Positivos**	**Falsos Positivos**	**Verdaderos Negativos**	**Falsos Negativos**	**Recall**	**Precision**	**TNR**	**F1**	**Accuracy**
Antena 3	451	83	2996	70	0,8656	0,8446	0,9730	0,8550	0,9575
AnTV	69	29	3490	12	0,8519	0,7041	0,9918	0,7709	0,9886
Canal-I	719	121	2653	107	0,8705	0,8560	0,9564	0,8631	0,9367
Globovisión	773	163	2540	124	0,8618	0,8259	0,9397	0,8434	0,9203
LaTele	987	186	2282	145	0,8719	0,8414	0,9246	0,8564	0,9081
Meridiano	1021	152	2276	151	0,8712	0,8704	0,9374	0,8708	0,9158
TVES	1015	167	2309	109	0,9030	0,8587	0,9326	0,8803	0,9233
Valores Medios	5035	901	18546	718	0,8708	0,8287	0,9508	0,8486	0,9358

Tabla 3.10: Resultados utilizando SVM teniendo en cuenta la fuente de emisión.

Los experimentos realizados teniendo en cuenta la fuente de emisión de las medias básicamente muestran los mismos patrones descritos con anterioridad. No obstante es apreciable que los resultados para una misma fuente de emisión son un tanto superiores que cuando se emplean medias de diferentes estaciones de TV, aún cuando se utiliza una menor cantidad de datos de entrenamiento.

En la totalidad de las pruebas efectuadas con SVM los valores de *TNR* y *Accuracy* permanecieron por encima del 90%, a pesar de que los valores de *recall* y *precision* sean inferiores. La primera medida refleja la efectividad de la detección de segmentos de programación general, sin embargo la efectividad de la detección de anuncios puede ser mejorada con la integración de descriptores visuales de la secuencia de vídeo. El hecho de tener una medida de *Accuracy* superior al 90% expresa la efectividad general de la técnica propuesta.

Si se comparan los resultados obtenidos en los experimentos realizados con los resultados de otros trabajos citados en el desarrollo de la memoria, es notable que los aquí expuestos son relativamente inferiores cuantitativamente. Sin embargo se debe decir que ningún trabajo anterior utilizó un conjunto de datos tan amplio para las pruebas, ni tampoco que representara una variedad tan amplia de género y canales de televisión, exceptuando el trabajo de [22] que sí expone resultados que se pueden clasificar como excelentes y emplearon una variedad representativa de datos. Otro elemento significativo a tener en cuenta para realizar un análisis comparativo, es el hecho que las técnicas que se comparan con los

resultados alcanzados en el presente trabajo utilizan métodos híbridos en los que se integran características tanto de la imagen como el audio. En este sentido se puede decir que el algoritmo aquí propuesto está en desventaja y puede ser objeto de trabajos posteriores integrar alguna característica visual que permita mejorar los resultados. No obstante en la Figura 3.3 se puede apreciar que los resultados obtenidos no se alejan demasiado del resto e incluso son superiores a una de las técnicas comparadas.

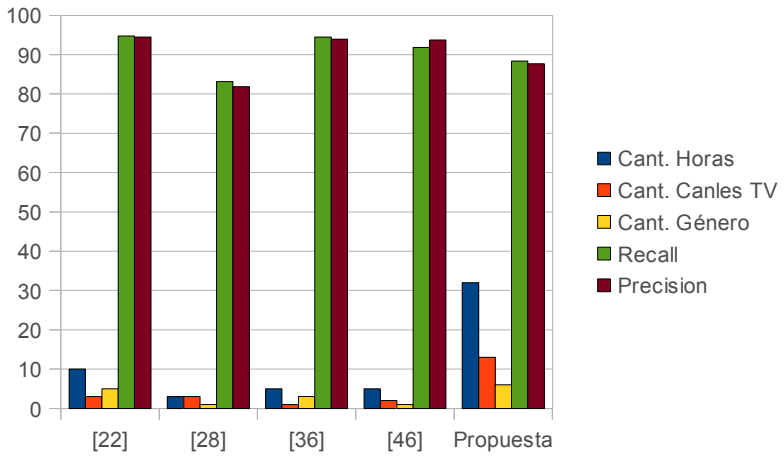

Figura 3.3: Gráfica comparativa de los resultados de diferentes técnicas.

En las siguientes tablas se exponen los resultados de la métrica WindowDiff que caracterizan la calidad de la segmentación, los cuales presentan valores relativamente altos debido a que la segmentación de la secuencia basada en la clasificación realizada presenta problemas con la definición de los límites de los bloques de anuncio y programación general. Dichas indeterminaciones en los límites de los bloques se reflejan en las gráfica de la Figura 3.4.

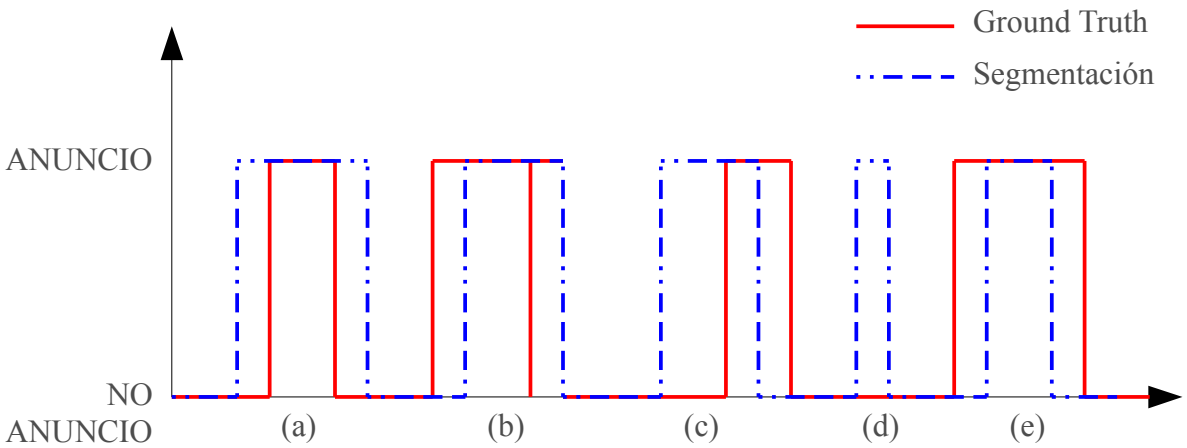

Figura 3.4: Gráfica que muestra las desviaciones de los límites de los bloques.

	Resultados de Segmentación		
Género	**Bloques de Anuncios**	**Bloques de No Anuncios**	**WindowDiff**
General	15	13	0,2818
Noticia	4	4	0,4329
Entrevista	3	2	0,4189
Deporte	4	5	0,3956
Serie	3	4	0,4277
Documental	5	5	0,4005
Todos	30	34	0,4954

Tabla 3.11: Resultados de WindowDiff para los diferentes géneros.

	Resultados de Segmentación		
Canal TV	**Bloques de Anuncios**	**Bloques de No Anuncios**	**WindowDiff**
Antena 3	3	2	0,4531
AnTV	1	2	0,4437
Canal-I	4	5	0,2774
Globovisión	4	5	0,5781
LaTele	4	5	0,4894
Meridiano	6	6	0,3682
TVES	5	6	0,4766

Tabla 3.12: Resultados de WindowDiff para los diferentes canales de TV.

Los problemas en los límites de los bloques de segmentación reflejados en la gráfica anterior sugieren que la técnica propuesta debe evolucionar para mejorar dichas deficiencias. Una posible

solución, como ya se dijo con anterioridad, es integrar algunos descriptores visuales que permitan una mejor definición de los bloques. Un ejemplo que soluciona en gran parte esta problemática es el trabajo de [22], donde utilizan una segmentación por cambios de escenas previamente al procesamiento de los descriptores y al finalizar la clasificación realizan un post-procesado de la misma teniendo en cuenta un conjunto de premisas para mejorar los límites de los bloques de segmentación. El estudio de estas y otras soluciones deberá ser continuado como trabajos futuros de la presente investigación.

3.5 Conclusiones Parciales

Con la realización de los experimentos se pudo validar la técnica propuesta para la detección de bloques de anuncios utilizando la señal de audio. Se logró implementar un conjunto de funciones y *scripts* en MATLAB que sirvieron como base experimental del trabajo y podrán ser reutilizadas en trabajos futuros asociados. Se utilizaron cinco medidas de eficiencia, sus resultados demostraron la efectividad de la investigación y se pudieron establecer comparaciones con otros trabajos sobre el tema. A pesar que los resultados no están al nivel de los métodos más robustos existentes, se puede decir que con la presente investigación se sentaron las bases para trabajos futuros que mejoren estos. Igualmente se demostró la eficacia de los descriptores propuestos para la detección de información sobre la señal de audio.

CONCLUSIONES GENERALES

Después de haber completado el trabajo de investigación del Proyecto de Fin de Máster se puede afirmar que el autor pudo poner en práctica los conocimientos adquiridos en las asignaturas del máster, así como nuevos conocimientos alcanzados durante el proceso de estudio e implementación. El desarrollo del presente trabajo aportó al autor conocimientos relacionados con temas como descriptores de audio, segmentación y clasificación de audio, procesamiento automático de señales de audio, técnicas de clasificación y detección automática de anuncios.

Con el estudio de los principales conceptos asociados al dominio del problema y las técnicas desarrollados para la detección de anuncios, se sentaron las bases teóricas para el desarrollo de la técnica propuesta. La investigación realizada ofrece como aportaciones investigativas un conjunto de descriptores de audio eficientes para la detección de publicidad en emisiones de televisión. Además se logró conceptualizar todo un proceso de clasificación y segmentación de vídeo basado en la detección de anuncios utilizando solo la señal de audio.

La realización de determinados experimentos sobre un conjunto amplio de datos de diferentes géneros y fuentes de emisión, demostraron la efectividad de la investigación y se pudieron establecer comparaciones con otros trabajos sobre el tema. A pesar que los resultados no están al nivel de los métodos más robustos existentes, se puede decir que con la presente investigación se sentaron las bases para trabajos futuros que mejoren estos.

De esta manera y por todo lo anterior se puede concluir que el objetivo general de la presente investigación fue cumplido, lográndose diseñar una técnica que permite detectar bloques de publicidad en emisiones televisivas la cual podrá ser aplicada en sistemas de monitoreo e indexación de televisión, siendo validada y demostrada su funcionalidad.

RECOMENDACIONES

Una vez cumplido los objetivos del presente trabajo y en correspondencia con los resultados obtenidos, los cuales se expusieron en el propio documento, se recomienda como trabajos futuros:

- Analizar en detalle los fallos de las detecciones realizadas para poder identificar su causa y proponer mejoras que aseguren una detección más precisa, centrado principalmente en la definición de las fronteras de los bloques de anuncio/no anuncio.
- Aplicar los descriptores de audio propuestos a otras problemáticas relacionadas con la clasificación y segmentación de señales de audio, con el objetivo de distinguir géneros de programas de televisión.
- Integrar la técnica propuesta a un Sistema de Monitoreo e Indexación de Televisión de manera que se pongan en práctica las aportaciones teóricas realizadas.
- Utilizar descriptores de más alto nivel para la identificación de melodías de cortinillas de televisión que identifican el inicio y fin de un bloque de anuncios.

REFERENCIAS BIBLIOGRÁFICAS

[1] **Akutsu, A.** *Video Indexing Using Motion Vectors.* SPIE Conference on Visual Communications and Image, Boston, MA, USA,1992.

[2] **Albiol, A., Fulla, M. J. & Torres, L.** *Detection Of TV Commercials.* International Conference on Acoustics, Speech and Signal Processing, pp. 541–544, 2004.

[3] **Brunelli, R.** *A Survey on the Automatic Indexing of Video Data.* Journal of Visual Communication and Image Representation, 1997.

[4] **Cheriet, M. & Suen, C. Y.** *A recursive thresholding technique for image segmentation.* IEEE Transactions of Image Recognition, 1998.

[5] **Cortes, C. & Vapnik, V.** *Support Vector Networks.* Machine Learning, pp. 273-297, 1995.

[6] **Covell, M., Baluja, S. & Fink, M.** *Advertisement detection and replacement using acoustic and visual repetition.* In Proc. IEEE 8th Workshop on Multimedia Signal Processing, pp. 461-466, Oct. 2006.

[7] **Duan, L.Y., Wang, J., Zheng, Y., Jin, J.S., Lu, H. & Xu C.** *Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis.* ACM Multimedia, pp. 201-210, 2006.

[8] **Duygulu, P., Chen, M. & Hauptmann, A.** *Comparison and combination of two novel commercial detection methods.* In Proc. ICME, Taiwan, 2004.

[9] **Fadeev, A., Missaoui, O. & Frigui, H.** *Dominant Audio Descriptors For Audio Classification and Retrieval.* International Conference on Machine Learning and Applications, 2009.

[10] **Fan, R.-E., Chen, P.-H. & Lin, C.-J.** *Working set selection using the second order information for training SVM*. Journal of Machine Learning Research, Vol, 6, pp. 1889-1918, 2005.

[11] **Fisher, R.** *The Use Of Multiple Measurements In Taxonomic Problems.* In Annals Of Eugenics, Vol. 7, pp. 179-188,1936.

[12] **Freund, Y.** *Boosting a weak learning algorithm by majority*. In Proc. Third Annual Workshop on Computational Learning Theory, 1990.

[13] **Gauch, J.M. & Shivadas, A.** *Finding and identifying unknown commercials using repeated video sequence detection.* Computer Vision and Image Understanding. Vol. 103, Elsevier Science Inc., New York, NY, USA, pp. 80–88, 2006.

[14] **Giannakopoulos, T., Kosmopoulos, D., Aristidou, A. & Theodoridis, S.** *Violence Content Classification Using Audio Features.* Department of Informatics and Telecommunications, National and Kapodistrian. University of Athens, 2006.

[15] **Guo, G. & Li, S. Z.** *Content-based audio classification and retrieval by support vector machines*. IEEE Trans. Neural Netw., Vol. 14, No. 1, pp. 209–215, Jan. 2003.

[16] **Har-Peled, S., Roth, D. & Zimak, D.** *Constraint Classification for Multiclass Classification and Ranking.* Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference, MIT Press, 2003.

[17] **Hauptmann, A. G. & Witbrock, M. J.** *Story segmentation and detection of commercials in broadcast news video*. In Proc. ADL'98, Santa Barbara, USA, 1998.

[18] **Herley, C.** *Accurate repeat finding and object skipping using fingerprints.* ACM Multimedia, pp. 656–665, 2005.

[19] **Herley, C.** *ARGOS: automatically extracting repeating objects from multimedia streams.* IEEE Transactions on Multimedia, Vol. 8, pp. 115-129, 2006.

[20] **Hess, W.** *Pitch Determination of Speech Signals.* Springer-Verlag, 1983.

[21] **Hua, X., Chen, X.& Zhang, H.** *Robust Video Signature Based on Ordinal Measure.* International Conference on Image Processing (ICIP), pp. 415–423, 1998.

[22] **Hua, X., Lu, L.& Zhang, H.** *Robust Learning- Based TV Commercial Detection.* In Proc. ICME, pp. 149–152, 2005.

[23] **Hurst, R.** *TV Commercial Editor*. Electronics Now, Vol. 65, No. 8, pp.31, Aug. 1994.

[24] **Kearns, M.** *Thoughts on hypothesis boosting.* Unpublished manuscript, 1988.

[25] **Li, Y. & Kuo, C.-C. J.** *Detecting commercial breaks in real TV programs based on audiovisual information.* In Proc. SPIE, pp. 225–236, 2000.

[26] **Lienhart, R., Kuhmnch, C. & Effelsberg, W.** *On the detection and recognition of television commercials.* In Proc. IEEE Conference on Multimedia Computing and Systems, pp. 509–516, Otawa, Canada, 1997.

[27] **Liu, Z., Wang, Y. & Chen, T .** *Audio Feature Extraction And Analysis For Scene Segmentation And Classification.* Journal of VLSI Signal Processing System, 1998.

[28] **Meng, L., Cai, Y., Wang, M., & Li, Y.** *TV Commercial Detection Based On Shot Change And Text Extraction*. Image And Signal Processing, CISP'09. 2nd International Congress, pp.1-5, Oct. 2009.

[29] **Mizutani, M., Ebadollahi, S. & Chang, S.** *Commercial Detection In Heterogeneous Video Streams Using Fused Multi-modal And Temporal Features*. In Proc. Acoustics, Speech, And Signal Processing. ICASSP '05. IEEE International Conference. Vol. 2, pp. 157- 160, March, 2005.

[30] **MSH, Management Sciences for Health.** *Desarrollo de un Vocabulario Común: Glosario de Términos Gerenciales*. [En Línea] [Citado el: 20 de agosto de 2010]. Disponible en: http://erc.msh.org/readroom/espanol/vocab.htm. 2001.

[31] **Olson, D. L. & Delen, D.** *Advanced Data Mining Techniques.* Springer; Fist Edition. pp. 138. Feb. 2008.

[32] **Orozco, I.** *Homogeneización de Conceptos de Servicios de Desarrollo Empresarial (SDE).* [En Línea] [Citado el: 20 de agosto de 2010]. Disponible en: *http://www.infomipyme.com/Docs/GT/Offline/tecnicos/mercadoSDE/sdeprint.htm*. 2003.

[33] **Pevzner, L. & Herst, M.** *A Critique and Improvement of an Evaluation Metric for Text Segmentation*. Computational Linguistics. Vol. 28, No. 1, pp. 19-36, 2002.

[34] **Pohjalainen, J.** *Methods of Automatic Audio Content Classification.* Licentiate's Thesis submitted in partial fulfillment of the requirements for the degree of Licentiate of Science in Technology. Helsinki University Of Technology, Nov. 2007.

[35] **Puentes, S.** *Glosario*. [En Línea] [Citado el: 20 de agosto de 2010]. Disponible en: *http://fbio.uh.cu/bioinfo/glosario.html*. 2005.

[36] **Sadlier, D. A., Marlow, S., O'Connor, N. & Murphy, N.** *Automatic TV Advertisement Detection From MPEG Bitstream.* Journal of the Pattern Recognition Society, Vol. 35 No. 12, pp. 2–15, 2002.

[37] **Sánchez, J. & Binefa, X.** *Audicom: a video analysis system for auditing commercial broadcasts.* In Proc. ICMCS'99, Firenze, Italy, 1999.

[38] **Sánchez, J.M., Binefa, X. & Vitria, J.** *Shot Partitioning Based Recognition of TV Commercials.* In Multimedia Tools Applications, Vol. 18, Kluwer Academic Publishers, Hingham, MA, USA, pp. 233–247, 2002.

[39] **Satterwhite, B. & Marques, O.** *Automatic detection of TV commercials.* Potentials, IEEE, Vol. 23, No. 2, pp. 9- 12, Apr.-May 2004.

[40] **Schapire, R.** *Strength of Weak Learnability*. Machine Learning, Vol. 5, pp 197-227, 1990.

[41] **Song, Y., Wang, W.-H. & Guo, F.-J.** *Feature Extraction And Classification For Audio Information In News Video.* In Proc. International Conference on Wavelet Analysis and Pattern Recognition, Baoding, Jul. 2009.

[42] **Theodoridis, S. & Koutroumbas, K.** *Pattern Recognition. Fourth Edition.* Institute for Space Applications & Remote Sensing, National Observatory of Athens, Greece. Oct. 2008.

[43] **Umapathy, K., Krishnan, S. & Raveendra, K. R.** *Audio Signal Feature Extraction and Classification Using Local Discriminant Bases.* IEEE Transactions On Audio, Speech, And Language Processing, Vol. 15, No. 4, May 2007.

[44] **Wen, X., Huffmire, T. D., Hu, H. H. & Finkelstein, A.** *Wavelet-based Video Indexing and Querying*. Multimedia Systems, Vol. 7, Issue 5, pp. 350-358, 1999.

[45] **Zhang, T. & Kuo, J.** Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. IEEE transactions on speech and audio processing, Vol. 9, No. 4, May 2001.

[46] **Zhang, L., Zhu, Z. & Zhao, Y.** *Robust Commercial Detection System* . IEEE International Conference On Multimedia And Expo, pp. 587-590, Jul. 2007.