Centro de Investigación Científica y de Educación Superior de Ensenada, Baja California

Programa de Posgrado en Ciencias en Ciencias de la computación

---

Reconocimiento de contexto usando información auditiva

Tesis

para cubrir parcialmente los requisitos necesarios para obtener el grado de

Doctor en Ciencias

Presenta:

Jessica Beltrán Márquez

Ensenada, Baja California, México

2015

Tesis defendida por

Jessica Beltrán Márquez

y aprobada por el siguiente Comité

Dr. Edgar Leonel Chávez González
Codirector del Comité

Dr. Jesús Favela Vara
Codirector del Comité

Dr. Hugo Homero Hidalgo Silva

Dr. José Antonio García Macías

Dra. Marcela Deyanira Rodríguez Urrea

Dra. Ana Isabel Martínez García
Coordinador del Programa de Posgrado en Ciencias de la computación

Dra. Rufina Hernández Martínez
Director de Estudios de Posgrado

Resumen de la tesis que presenta Jessica Beltrán Márquez como requisito parcial para la obtención del grado de Doctor en Ciencias en Ciencias de la computación.

Reconocimiento de contexto usando información auditiva

Resumen aprobado por:

Dr. Edgar Leonel Chávez González
Codirector de Tesis

Dr. Jesús Favela Vara
Codirector de Tesis

La información que provee el contexto puede utilizarse para desarrollar aplicaciones dedicadas al cuidado de adultos mayores. Por ejemplo, para brindar asistencia en sus actividades de la vida diaria, fomentar su integración social y desarrollar estrategias para disminuir riesgos provocados por algunos medicamentos. Para obtener información del contexto se pueden utilizar diferentes sensores. En particular, la captura y el análisis del audio tienen la ventaja de proveer flexibilidad para la recolección de datos ya que los micrófonos están embebidos en dispositivos móviles. Los retos que presenta el reconocimiento del contexto usando audio son la existencia de sonidos traslapados, las diferencias entre sonidos de un mismo tipo, el ruido y las limitaciones de espacio y tiempo de procesamiento. Los métodos existentes permiten reconocer eventos de sonidos cuando están grabados sin que exista traslape con otros sonidos. Recientemente se han propuesto métodos que son capaces de identificar sonidos aunque estén mezclados con ruido de fondo, siempre que se pueda modelar el ruido a priori. En esta tesis, se presenta una representación del audio diseñada para reconocer eventos de sonidos ambientales sin necesidad de modelar el ruido de fondo. El método propuesto consiste en extraer características de las bandas de frecuencia de los sonidos a nivel de trama y posteriormente obtener la primera y segunda derivada en cada una de las bandas. La representación final está formada por un conjunto de histogramas, uno por cada banda. Nuestra propuesta tiene la ventaja de ser ligera tanto en su cálculo como en su representación final. Se muestra evidencia experimental que avala la eficiencia del esquema para el reconocimiento de sonidos ambientales y se compara contra el único trabajo con un enfoque similar al nuestro que considera la misma restricción de no conocer a priori el ruido de fondo. Los resultados obtenidos mejoran en velocidad, precisión y exhaustividad a los métodos en el estado de arte. También se presentan evaluaciones sobre dos casos de uso relacionados con aplicaciones dirigidas al apoyo de adultos mayores cuyos resultados indican evidencia de su eficacia para usar el reconocimiento automático de sonidos como herramienta. Además, se describe el uso del análisis del audio para desarrollar soluciones en escenarios distintos a la asistencia de adultos mayores.

Palabras Clave: **Reconocimiento de contexto auditivo, Firmas de audio, Análisis de audio, Clasificación de sonidos ambientales.**

Abstract of the thesis presented by Jessica Beltrán Márquez as a partial requirement to obtain the Doctorate in Sciences degree in Computer Sciences.

Context recognition using auditory information

Abstract approved by:

Dr. Edgar Leonel Chávez González
Thesis Co-Director

Dr. Jesús Favela Vara
Thesis Co-Director

The information given by the context can be used to develop applications to support older adults. For example, by providing assistance in their daily activities, by increasing their social interaction and by providing strategies to decrease the risk produced by some medications. Different types of sensors can be used to obtain the context information. Particulary, the capture and analysis of audio has the advantage of flexibility in data collection and that microphones are included in mobile devices like smart phones. The challenges of context recognition through audio are the presence of mixed sounds in nature, the differences between sounds from the same class, the noise and the limitations of space and processing time. The current techniques allow classify sound events when they were captured with no overlap with other sounds. Recently, new methods have been proposed that are capable to identify sounds even if they are mixed with background noise, however a priori noise model in necessary. In this thesis, we present the development of a method for sound representation and classification designed to recognize environmental sound events without modelling the background noise. Our proposal consist first in the extraction of features in the frequency bands from the audio frames. Then, we obtain the first and second derivative in each of the bands to finally represent the sounds through a set of histograms, one for each band. Our proposed method has the advantage of being light and easy to calculate. We show experimental evidence that validates the efficiency of the method to recognize sounds events and we compare against the only approach that considers the same restriction of not modelling a priory the background noise. The results in the experiment show an improvement in the processing times, precision and recall compared with the state of the art techniques. Also, we present evaluations over two use cases related with applications to support older adults, which provides evidence of its efficacy for automatic recognition of sounds as a tool for this purpose. Finally, we describe the use of audio analysis to develop solutions in different escenarios besides the support of older adults.

Keywords: **Auditory context recognition, Audio fingerprint, Audio analysis, Environmental sound classification.**

Dedicatoria

Agradecimientos

A toda mi familia por su apoyo incondicional. Mis padres, hermanos, mis hermosos sobrinos y al resto de familia.

A mi asesor Edgar, porque además de ser mi sensei fuiste mi amigo. A mi asesor Jesús por haberme mostrado con la práctica las virtudes que podemos tener los seres humanos. A los miembros de mi comité: Hugo, Tony, Marcela, muchas gracias por su apoyo, sus aportaciones y por haberme dejado encontrar en ustedes unas grandes personas.

A Ana M. por su apoyo en momentos difíciles y la generación de buenos momentos.

A todas las secretarias que pasaron por Ciencias de la computación durante mi estancia, especialmente a Caro y Lydia.

A los indispensables de mis últimos años: Karla, Karina, Héctor, Valeria, René, Santos, Raymundo, Eduardo, Fabiola y Victor por haber colaborado para que todo el tiempo que pasé en mis estudios de doctorado lo viviera feliz, siempre estarán en mis corazones por ser tan importantes para mi. Mi eterno agradecimiento por ser ustedes.

A mis compañeros y maestros de CICESE de todas las generaciones con las que me tocó compartir, UMICH y CREATE-NET, gracias por estar en mi vida.

A las personas que colaboraron en el desarrollo de esta tesis: Cathy y Galatea, y mi maestro J. Antonio Camarena.

A los que colaboran y hacen posible los sitios freesound.org y stackoverflow.

Al Centro de Investigación Científica y de Educación Superior de Ensenada.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico para realizar mis estudios de doctorado.

Tabla de contenido

Página

Lista de figuras

Figura Página

Figura Página

Figura Página

Lista de tablas

Capítulo 1. Introducción y motivación

El término contexto, en el ámbito de los sistemas computacionales, esta definido como "cualquier información que pueda utilizarse para caracterizar la situación de una entidad, donde entidad pueda ser una persona, un lugar o un objeto que pueda ser relevante para la interacción entre el usuario y una aplicación, incluyendo incluso al usuario y a la aplicación misma"(Abowd *et al.*, 1999).

Actualmente existe una amplia gama de dispositivos que poseen la capacidad de sensar el entorno y comunicarse entre si de forma inalámbrica (Meyer and Rakotonirainy, 2003). Esto ha permitido avances en una área de conocimiento denominada cómputo consciente del contexto, donde se busca desarrollar sistemas que se adapten a la información asociada al contexto (Chen and Kotz, 2000).

Una primera generación de aplicaciones conscientes del contexto se basaba en la ubicación para adaptar el funcionamiento de las aplicaciones (Kirk and Newmarch, 2005). Un ejemplo está dado por un sistema que, mediante el uso de dispositivos infrarrojos, detecta el movimiento de los empleados de una oficina para permitir que sean contactados al teléfono más cercano a su ubicación actual dentro de todo un edificio (Want *et al.*, 1992). Otros ejemplos de aplicaciones para este paradigma de cómputo son el desarrollo de tecnologías de información que ofrecen servicios adaptables al contexto del usuario para entregar respuestas más inteligentes (Ma *et al.*, 2003). Una más consiste en un modelo que selecciona estrategias dinámicas de notificación de mensajes basándose en la prioridad del mensaje, el nivel de uso y la ubicación actual (Sawhney and Schmandt, 1999). Otras investigaciones relacionadas con el cómputo consciente del contexto, se enfocan en hacer identificación de actividades (Lukowicz *et al.*, 2004) (Oliver *et al.*, 2004) (Baldauf *et al.*, 2007) e identificación de comportamiento.

Un campo de aplicación para la detección de actividades y comportamientos que ha generado interés en los últimos años, consiste en el cuidado de adultos mayores. Esta es una tarea demandante que normalmente es llevada a cabo por familiares cercanos quienes se enfrentan a condiciones de estrés o *burnout* mientras brindan asistencia 24/7 (Rialle *et al.*, 2008). El concepto en inglés *"Ambient Assisted Living (AAL)"*, consiste en

usar la información y las tecnologías de comunicación en las actividades de la vida diaria de personas para permitir que continúen activas, se mantengan socialmente conectadas y vivan de forma independiente en su edad avanzada (Active and Programme, 2015). Por ejemplo, se pueden notificar automáticamente a los cuidadores sobre las actividades realizadas por los adultos mayores, para identificar oportunamente las situaciones en las que se requiera brindar apoyo, ya sea en condiciones de riesgo o cuando está sucediendo algo inusual (Morris *et al.*, 2003).

1.1. AAL para adultos mayores y sus cuidadores

La oportunidad de los adultos mayores de prolongar su independencia para continuar viviendo en sus propios hogares en vez de ingresar a una residencia geriátrica es conocido como *"Aging in Place" (AiP)*. No es suficiente proporcionar aparatos tecnológicos para brindar la asistencia, sino que es importante conocer el contexto, de forma que el apoyo se lleve a cabo adecuadamente (Morris *et al.*, 2003).

El incremento de la esperanza de vida de la población y la disminución de la natalidad han provocado que la población de adultos mayores tenga un crecimiento importante en los últimos años (Zúñiga and Vega, 2004). Este crecimiento poblacional motiva a atender retos sociales, económicos y de salud con el objetivo de mejorar la calidad de vida del adulto mayor (Acree *et al.*, 2006)

Algunas propuestas de cómputo consciente del contexto para asistir AiP pueden cubrir distintos aspectos de la vida del adulto mayor. Por ejemplo, con la detección automática de sus actividades se puede notificar a los miembros de su familia para mantenerlos involucrados y al tanto de situaciones donde puedan brindar apoyo (Wilson, 2005). Otro aspecto de la vida del adulto mayor donde se puede usar el cómputo consciente del contexto es mediante el diseño de aplicaciones que fomenten su integración social para promover un envejecimiento saludable y mantener su bienestar físico y mental (Cornwell *et al.*, 2008). Un ejemplo se da en (Miluzzo *et al.*, 2008) donde utilizan teléfonos inteligentes para inferir actividades automáticamente y notificarlas mediante redes sociales, esto puede propiciar que los adultos mayores y sus familiares de generaciones más recientes mejoren su comunicación (Cornejo *et al.*, 2013).

Otra área de apoyo consiste en monitorizar las actividades y comportamientos de los pacientes para determinar las intervenciones farmacológicas o no farmacológicas adecuadas (Wilson, 2005). Algunos medicamentos, tales como antipsicoticos y antidepresivos proveen evidencia de ayudar con el trato a enfermedades en adultos mayores, tal como la demencia, sin embargo sus efectos son modestos y producen efectos secundarios severos como accidentes cerebro-vasculares e incremento en el riesgo de mortalidad, esto resalta la importancia de usar alternativas no farmacológicas para disminuir estos síntomas. Mediante el análisis del contexto es posible desarrollar un Sistema de Intervenciones Asistidas (SIA), el cual se puede diseñar para detectar automáticamente la ubicación de adultos mayores que deambulan, detectar si se encuentran impacientes, agitados o en general con un comportamiento problemático y con esta información decidir el uso de una estrategia no farmacológica adecuada para mejorar el estado del adulto mayor.

Se pueden utilizar las lecturas de diversos tipos de sensores (Lukowicz *et al.*, 2004) (Oliver *et al.*, 2004) (Baldauf *et al.*, 2007) para inferir el contexto, ya que éstos permiten la captura de variables que podrían estar involucradas en la manifestación de un comportamiento o la realización de una actividad. Para poder realizar la inferencia, se establece una relación entre los cambios de las variables de entorno, las cuales se capturan directamente por los sensores, y la derivación del conjunto de actividades o comportamientos distinguibles (Meyer and Rakotonirainy, 2003).

Un reto en el diseño para la detección del contexto consiste en identificar cuales variables o características definen mejor a las actividades o situaciones que se desee inferir (Mihailidis and Fernie, 2002). Otro reto está dado por la diversidad con que las personas manifiestan sus comportamientos o ejecutan actividades (Surie *et al.*, 2007). Por ejemplo, la actividad *aspirar la sala* puede ejecutarse de distinta manera por dos personas. Además, la diversidad se puede presentar incluso en un mismo individuo al realizar una misma actividad en entornos o situaciones diferentes. Un individuo puede *lavarse los dientes* en la mañana en el baño de su recámara y el mismo individuo se lava los dientes a mediodía en el baño de su oficina. Por lo tanto, para desarrollar aplicaciones se requiere adaptar a las dinámicas del entorno y ajustar a las necesidades de cada usuario, es decir

que sea capaz de funcionar adecuadamente con las variantes que se dan con diferentes usuarios y condiciones ambientales, (Choudhury *et al.*, 2008).

Es improbable que un sólo tipo de sensor sea adecuado para todo tipo de aplicaciones y normalmente se combinan multiples sensores para incrementar de manera significativa la inferencia del contexto (Meyer and Rakotonirainy, 2003). Es de interés particular para esta investigación el utilizar el micrófono ya que el audio ambiental provee información significativa que se puede utilizar para discriminar la situación de un usuario, y por lo tanto puede ser usada en aplicaciones conscientes del contexto (Potamitis and Ganchev, 2008).

1.2. Reconocimiento de contexto usando audio

El reconocimiento automático del audio ambiental se ha vuelto relevante en el cómputo consciente del contexto, especialmente donde la identificación del audio puede usarse para detectar y monitorizar actividades de la vida diaria (Chen *et al.*, 2012). Una ventaja del audio sobre otro tipo de sensores se da debido a que los micrófonos pueden capturar información en todas direcciones y es altamente robusto ante los cambios de posición y orientación. Esto permite mayor flexibilidad para la recolección de datos siendo menos intrusivo para el usuario, a diferencia de otros sensores como las cámaras las cuales requieren de ciertas condiciones como el enfocar la escena de interés, tener una iluminación apropiada y la necesidad de analizar una mayor cantidad de datos. Otra ventaja de usar audio es que no se requiere de infraestructura centralizada ni de hardware adicional al micrófono, los cuales usualmente vienen ya incluidos en los dispositivos móviles tales como teléfonos celulares (Ma *et al.*, 2006) (Perttunen *et al.*, 2008). Por ejemplo, por medio del análisis del audio capturado por un solo micrófono se puede inferir la ubicación (Ma *et al.*, 2006), mientras que al utilizar otros sensores como RFID (Tesoriero *et al.*, 2010) o dispositivos *WiFi* (Paciga and Lutfiyya, 2005) entonces se requiere instalar una red para lograr la inferencia de la ubicación.

Existen limitantes para inferir contexto mediante el análisis del audio, por ejemplo resulta imposible detectar objetos o ambientes que no emiten ningún tipo de sonido (Ellis and Lee, 2006a). Además, existen distintos retos para llevar a cabo el análisis automático

del audio ambiental tal como la existencia de sonidos traslapados, heterogeneidad para producir los sonidos y el ruido presente al hacer la captura. A continuación se explican a detalle estos retos.

- **Sonidos traslapados.** El audio capturado por un micrófono está formado por mezclas de sonidos que provienen de diferentes fuentes. Aunque los seres humanos poseen la capacidad de separar y poner atención únicamente al sonido que es de su interés, esta tarea se mantiene como un problema abierto para ser realizada automáticamente.

- **Heterogeneidad intra-clase.** La heterogeneidad se da cuando existen diferencias entre los sonidos que provienen de una misma categoría de sonido. Para ilustrar la heterogenedad intra clase, considérese el sonido de *lavarse las manos*; un sistema debería ser capaz de identificar este sonido aunque esto implique que es realizado por diferentes personas, a pesar de que se utilizan diferentes llaves, distinta cantidad de agua y en general se presenten diferentes condidiciones.

- **Similitud inter-clase.** Entre la inmesurable cantidad de sonidos presentes en la naturaleza, existen sonidos que provienen de diferentes fuentes pero se escuchan de forma similar. Por ejemplo, el sonido de una pelota botando o el sonido de tocar una puerta. Un sistema de reconocimiento automático de sonido podría confundirse y predecir incorrectamente que se ha producido un sonido en vez de otro.

- **Ruido.** Otros problemas relacionados con el análisis automático son la existencia del **ruido** introducido por la transmisión de datos o por los dispositivos utilizados para la captura.

- **Representación compacta y fácil de calcular.** Un aspecto importante dentro del proceso del reconocimiento del audio ambiental consiste en obtener una representación apropiada del audio, la cual se denomina *firma de audio*, que se utiliza para conjuntar representaciones de sonidos similares y para distinguir entre sonidos que no están relacionados. Se busca tener una firma de audio que ocupe poco espacio y sea fácil de calcular y de comparar. Estos requerimientos están motivados por que

los dispositivos en donde se realizaran los cálculos tienen limitantes de poder de cómputo y memoria disponible.

Los trabajos en reconocimiento de contexto basados en audio pueden agruparse en dos maneras de describir las clases o categorías del entorno auditivo; *escenas auditivas* y *eventos de sonido*. Se puede considerar a las escenas auditivas como composiciones musicales, en donde los eventos de sonido son los instrumentos musicales. Existen experimentos que determinan que el proceso cognitivo de los seres humanos para el reconocimiento o similaridad de escenas auditivas consiste en la identificación de las fuentes de sonido físicas. Por ejemplo la escena auditiva reconocida como `parque` está compuesta por eventos de audio localizados tales como `ave cantando` y `niños jugando`. Sin embargo, investigaciones recientes muestran que en condiciones en donde existen demasiados sonidos no característicos el ser humano posee estrategias holísticas para poder procesar escenas auditivas (Aucouturier *et al.*, 2007).

Al determinar las escenas auditivas, cada escena puede traducirse a alguna actividad, por ejemplo en la propuesta presentada en (Ma *et al.*, 2006) la identificación de la escena `lavandería` se traduce a que el usuario se encuentra `lavando ropa` mientras que la identificación de la escena `oficina` se traduce a que el usuario está `trabajando en la oficina`. Por otro lado, si se utilizan clases de eventos de sonido, la identificación de `cafetera` o `microondas` se pueden traducir a las actividades `preparando café` y `calentando en microondas` respectivamente. Ambos tipos de clases proveen información contextual y se pueden complementar para lograr un mejor entendimiento del contexto y una mejor clasificación. Existen estudios que demuestran que cuando los seres humanos identifican sonidos no solo utilizan la información auditiva sino que además usan conocimiento del contexto como auxiliar para desambiguar entre sonidos parecidos (Niessen *et al.*, 2008). Esta conclusión es especialmente importante para los métodos de reconocimiento automático; ya que para poder determinar que se ha identificado una clase de alto nivel, tal como una actividad, es necesario establecer una relación entre características de bajo nivel extraídas de la señal auditiva y el contexto en el cual se ha producido dicha señal.

La detección de eventos de sonido puede ayudar a conocer actividades y describir

características del comportamiento de los adultos mayores. Un ejemplo es el sonido *lavándose los dientes* que indica si el usuario está realizando su limpieza dental. Otro ejemplo es el evento de sonido *golpeteo en la silla* que se puede traducir a que un paciente esta inquieto. La detección del sonido *grito* puede indicar que el adulto mayor está enojado, mientras que con el sonido *Tosiendo* se puede conocer su estado de salud.

En esta tesis se busca encontrar una firma de audio que sea capaz de representar correctamente eventos de sonidos mezclados y contaminados con ruido. También, se contribuye al estado del arte al reconocer sonidos que poseen alta diversidad interclase y alta similitud intraclase. Buscamos tener una firma de audio que ocupe poco espacio, sea fácil de calcular y de comparar. Estos requerimientos están motivados por los desarrollos de aplicaciones en dispositivos móviles en donde existen limitantes de poder de cómputo y memoria disponible.

1.3. Preguntas de investigación

Las preguntas de investigación que surgen respecto al análisis del contexto por medio del sonido y la necesidad de brindar asistencia ambiental a adultos mayores son las siguientes:

- ¿Puede una firma basada en entropía ser adecuada para inferir contexto?
- ¿Se puede diseñar una firma compacta, fácil de calcular y evaluar para el reconocimiento de contexto?
- ¿Que capacidad de generalización a distintos usuarios puede tener una firma basada en la entropía?
- ¿Cuál es la utilidad del reconocimiento del contexto usando audio para apoyar al cuidado de adultos mayores?

1.4. Objetivo de investigación

Diseñar un esquema de extracción de firma de audio y clasificación para eventos de sonido dirigido a un sistema de reconocimiento de contexto que sea robusto ante

condiciones ruidosas, ocupe poco espacio y sea fácil de calcular y comparar. Además probar su utilidad en escenarios de adultos mayores.

1.5. Metodología

Buscando responder las preguntas de investigación para lograr el objetivo planteado se siguió la siguiente metodología la cual se ilustra en la Figura 1.

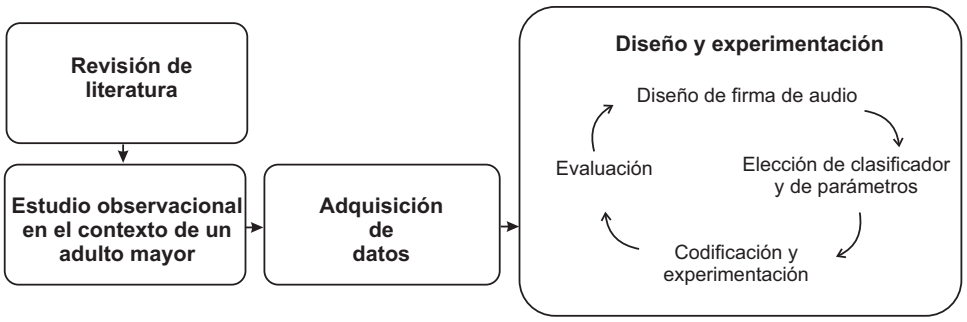


Figura 1: Fases de la metodología.

1.5.1. Revisión de literatura

La primera fase consistió en una revisión exhaustiva de la literatura relacionada con el reconocimiento de contexto utilizando información auditiva. Se prestó atención a las representaciones del audio y a las técnicas de clasificación utilizadas. También se analizaron las creaciones de las bases de datos y las métricas usadas para la evaluación de los métodos propuestos. Además se siguió la evolución de las propuestas para tratar los retos del análisis del audio.

1.5.2. Estudio observacional en el contexto de un adulto mayor

Se realizó un estudio observacional en una residencia geriátrica para desarrollar escenarios de uso de cómputo consciente del contexto. Se estudió si es posible utilizar la información auditiva para realizar un reconocimiento automático de actividades y comportamientos de los adultos mayores.

1.5.3. Adquisición de datos

Se crearon las bases de datos necesarias para realizar los experimentos. El audio se obtuvo utilizando micrófonos y posteriormente se etiquetó manualmente indicando cada

tipo de sonido. También se descargaron sonidos de Internet los cuales también fueron etiquetados manualmente.

1.5.4. Diseño y experimentación

El diseño y experimentación es una fase cíclica la cual consta de 2 etapas. Cada ciclo busca incrementalmente resolver los retos relacionados con el análisis de audio.

Diseño de firma de audio En esta etapa se hace el análisis de la señal de audio buscando una representación que permita clasificar sonidos adecuadamente. Se buscó que la firma de audio discrimine mejor entre diferentes tipos de sonido mejorando al estado del arte. También se consideró que sea robusta al ruido, utilice poco espacio utilizado y sea rápida de calcular.

Experimentación Se codificaron las firmas de audio propuestas en cada ciclo y se realizaron experimentos para su evaluación usando métodos y métricas utilizadas en la literatura.

1.6. Organización de la tesis

En el capítulo 2 se presentan los fundamentos teóricos de los sistemas de clasificación usando audio. En el capítulo 3 se presentan los métodos desarrollados en el estado del arte para el reconocimiento de contexto usando audio con aplicaciones para apoyar a los adultos mayores. En el capítulo 4 se describe las firmas de audio propuestas. En el capítulo 5 se describen los experimentos y resultados. En el capitulo 6 se presentan otros aportes usando el reconocimiento automático de sonidos. Finalmente, el capítulo 7 describe las conclusiones y trabajo futuro.

Capítulo 2. Sistemas de reconocimiento de audio

2.1. Sistemas de reconocimiento de audio

Sonido Un sistema de reconocimiento de audio consiste de tres etapas principales (ver Figura 2). Primero, se captura el **sonido** por medio de micrófonos (aquí se debe tener cuidado con la elección de la **frecuencia de muestreo** y la **resolución en bits** ya que estos valores definen la calidad del audio y la cantidad de datos a capturar). Posteriormente se extrae una firma de audio que se introduce a un clasificador que finalmente indica la clase a la que pertenece el sonido (Stäger, 2007).

La principal diferencia entre los trabajos de investigación en reconocimiento de audio, esta dada por las firmas y clasificadores que se utilizan. Para comparar cual propuesta es mejor para una aplicación específica, se realizan evaluaciones donde se analizan medidas que indican la eficiencia de la clasificación. A continuación se describen detalles y ejemplos de firmas de audio, clasificadores y evaluaciones en sistemas de reconocimiento de audio.

2.2. Firmas de audio

La extracción de firmas de audio se refiere al procedimiento de obtener una representación, basada en el contenido de la señal, que se utiliza para medir similitud entre segmentos de sonidos. La firma de audio se forma con características perceptuales del sonido, de forma que las firmas de dos sonidos que pertenecen a una misma clase deben tener alta similitud. Un objetivo de la firma de audio es reducir la cantidad de información a almacenar y el tiempo de procesamiento utilizado para comparar señales.

La obtención de la firma de audio es una parte muy importante en un sistema de reconocimiento automático. La capacidad de la firma para discriminar entre distintos tipos

Figura 2: Etapas de un sistema de reconocimiento de audio.

de sonidos es más importante que el tipo de clasificador utilizado, ya que si los sonidos no están correctamente representados ningún clasificador será capaz de entregar resultados correctos. Las firmas de audio deben cumplir con las siguientes propiedades para que representen adecuadamente a los sonidos ambientales.

2.2.1. Robustez a degradaciones

En la Figura 3.a se ilustra la forma de onda de una señal acústica. Si se desea comparar entre las formas de onda de dos señales, entonces se requiere comparar muestra por muestra en cada una de las señales, ver Figura 3.b. Esto implica muchos cálculos y no proporciona robustez a variaciones y degradaciones de la señal. Por ejemplo, en la Figura 3.c se observa que si se tienen dos señales muy parecidas pero con un pequeño desfasamiento en el tiempo, entonces al comparar las muestras correspondientes se indicaría que ambas señales son totalmente diferentes. Este resultado también se presentaría si se compara la señal original contra una versión con distinto volumen, ver Figura 3.d, y contra una versión con contaminación de ruido, ver Figura 3.e.

Una característica importante de la firma de audio es su capacidad para representar correctamente señales que pueden estar sujetas a una variedad de degradaciones tales como contaminación por ruido, cambios de volumen y desfasamiento en el tiempo. Esto implica que la firma de un sonido no debe ser muy diferente a una versión degradada del mismo sonido, es decir, la firma de audio debe ser *robusta ante degradaciones.*

2.2.2. Transparencia o robustez al traslape de sonidos

En ambientes no controlados, la señal capturada por micrófonos consta de una mezcla de sonidos que provienen de distintas fuentes. Por ejemplo, en una cocina se puede estar lavando los trastes al mismo tiempo que se esta guisando y usando la licuadora. Se busca poder reconocer un sonido en específico, por ejemplo el sonido de la licuadora, a pesar de que la señal original contiene a todos los sonidos traslapados. Denominamos *transparencia* a la habilidad de la firma de audio para reconocer todos los sonidos individuales presentes en una mezcla.

Si en una mezcla, se tiene que un sonido contribuye con mayor **potencia**, se le llama

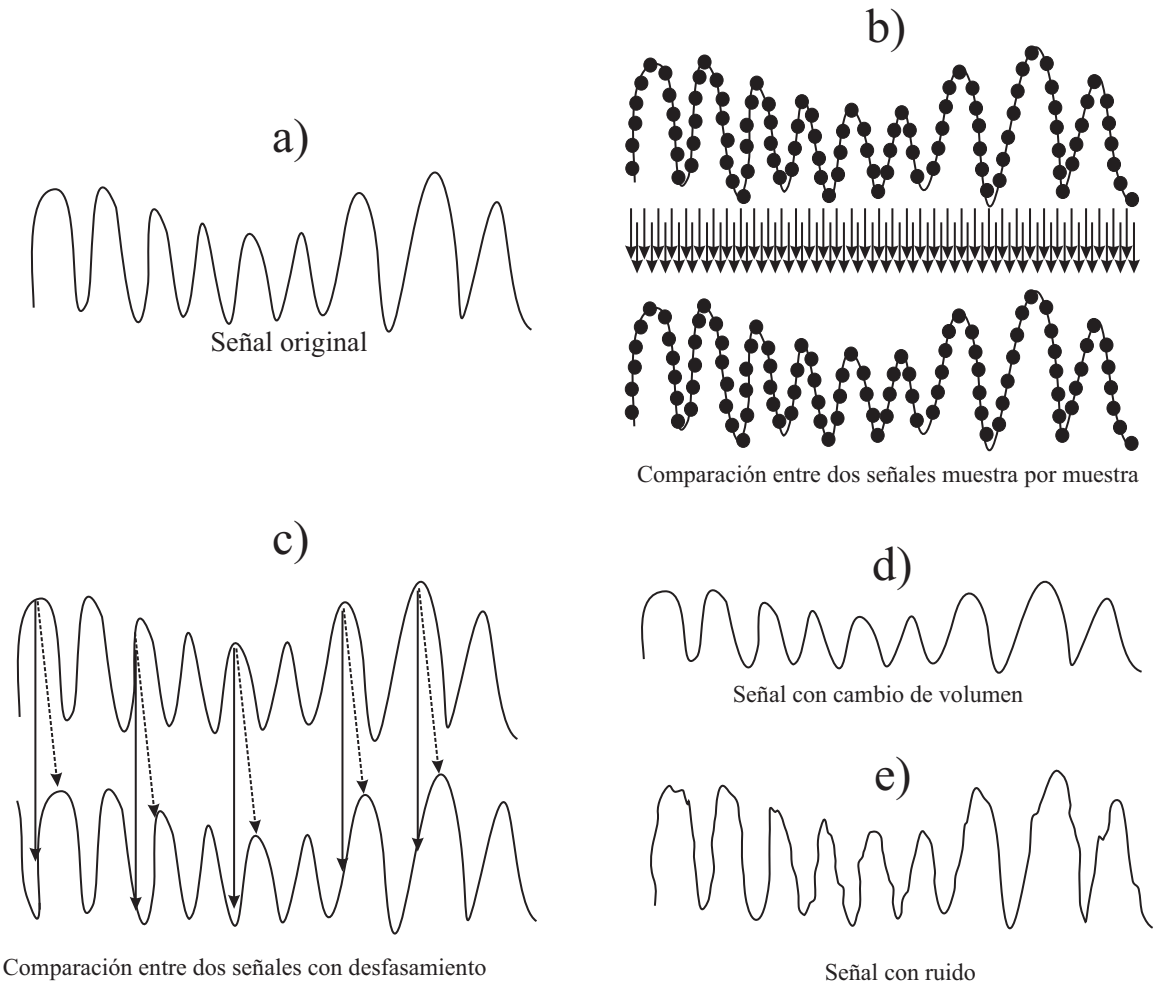


Figura 3: a) Forma de onda de una señal de audio, b) Comparación entre dos formas de onda, c) Desfasamiento en el tiempo en dos señales, d) Cambio de volumen, e) Señal contaminada con ruido.

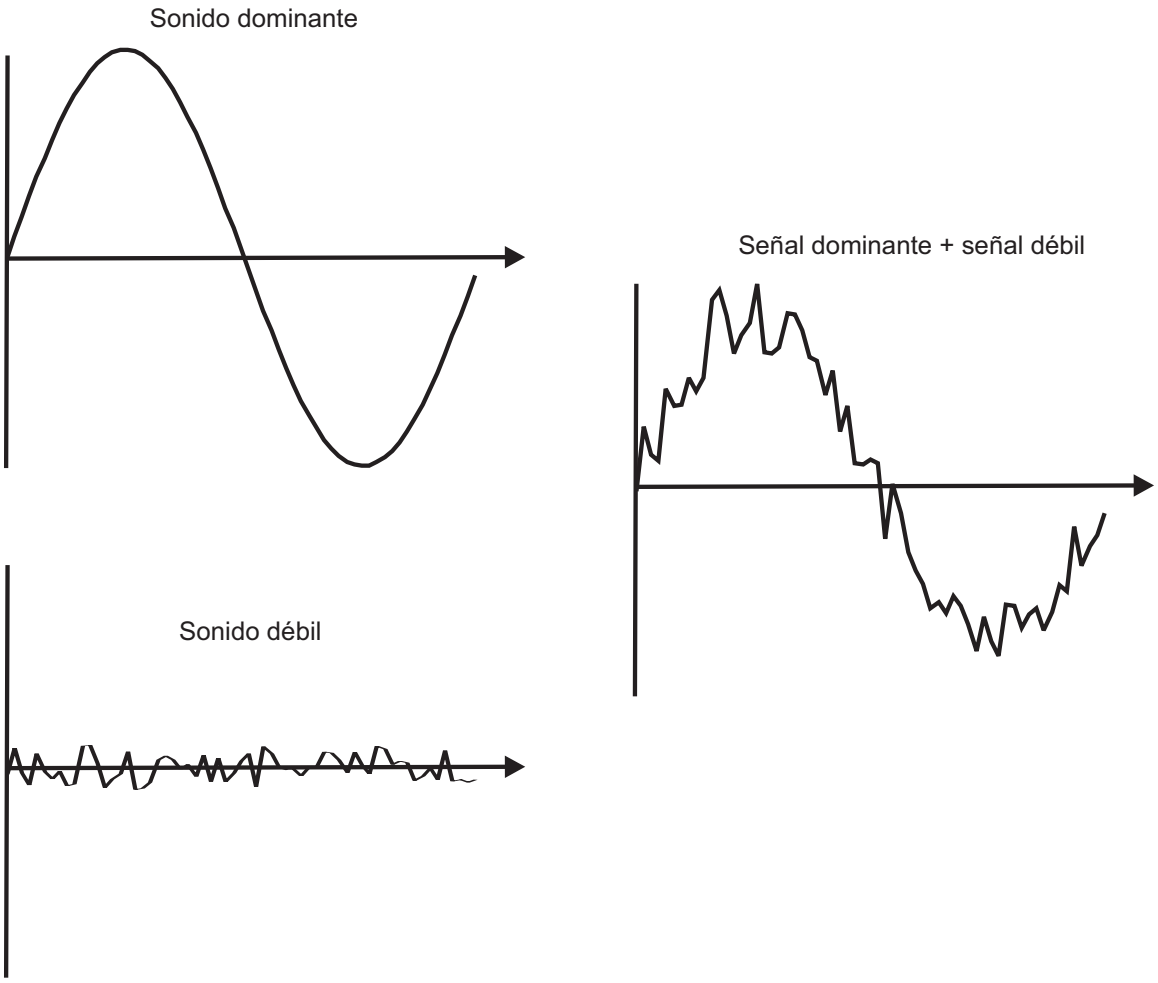


Figura 4: Ejemplo de una mezcla con una señal dominante y una señal débil.

sonido *dominante*, a comparación de un sonido *débil*, esto produce que al escuchar la mezcla, el sonido dominante se escuche más fuerte que el sonido débil , ver Figura 4.

Mediante la Relación Señal Ruido (SNR por sus siglas en ingés) se expresa la relación que existe entre el sonido dominante y el débil. El valor de SNR es una especificación que mide el nivel de sonido deseado presente en una señal auditiva comparado contra el nivel de ruido que también esta presenta en la señal.

$$SNR_{dB} = 10log\frac{P_{senal}}{P_{ruido}} \quad (1)$$

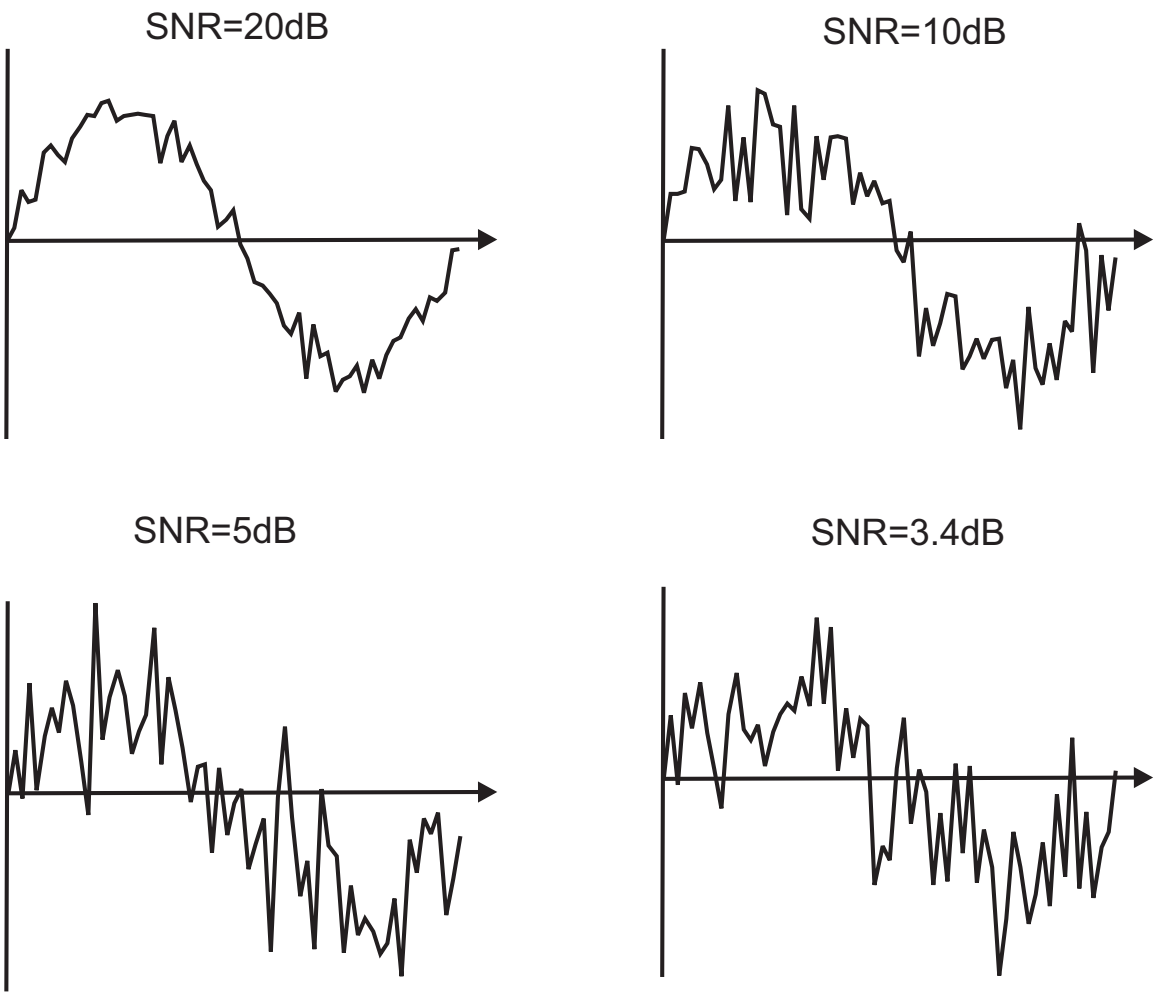


Figura 5: Cuatro ejemplos de mezclas usando distintos valores de SNR.

donde P es la potencia de la señal.

$$P = \frac{1}{N}\sum_{i=1}^{n}(x_i)^2 \tag{2}$$

En la Figura 5 se ilustran ejemplos de cuatro mezclas entre un sonido dominante y un sonido débil usando diferentes valores de SNR. Se observa que con valores altos de SNR la señal dominante resulta menos modificada por la señal débil. Una forma de conocer la transparencia de una firma de audio es evaluando la clasificación sobre mezclas con distintos valores de SNR.

2.2.3. Compacta

Es necesario que una firma de audio sea compacta debido a las limitaciones de memoria existentes en los dispositivos donde correrán las aplicaciones conscientes del contexto. Por ejemplo 3 segundos de una señal de audio crudo necesitan 216kB en memoria, mientras que esos 3 segundos representados con una firma de audio requieren solo 10kB. Una forma de permitir comparaciones rápidas, consiste en almacenar las firmas en la memoria inmediata RAM y asi evitar los costos de accesos a la memoria secundaria. Otra ventaja de tener una firma compacta se presenta si la aplicación requiere realizar transmisión de datos por medio de Iternet.

2.2.4. Facilidad de cálculo y de comparación

Para que un sistema consciente del contexto ejecute una acción oportuna o notifique un evento que esta sucediendo actualmente, se requiere que este tenga una respuesta rápida. Para esto, la firmas de audio se deben de calcular y comparar fácilmente. Adicionalmente, aunque la capacidad de procesamiento continúa incrementándose día con día, es importante optimizar este recurso para reducir el consumo de batería, permitir que se ejecuten otros procesos y que el audio se pueda procesar continuamente.

2.2.5. Escalabilidad

La escalabilidad se refiere a la capacidad que tiene una firma de audio para poder generalizarse ante sonidos de un mismo tipo a pesar de que exista heterogeneidad. Esta heterogeneidad puede producirse por la idiosincrasia entre distintos usuarios para producir un tipo de sonido, por variaciones en las condiciones ambientales o por diferencias entre los dispositivos utilizados para la captura de datos.

2.2.6. Procedimiento para calcular la firma de audio

El procedimiento para obtener firmas de audio generalmente consiste en los pasos ilustrados en la Figura 6. La primera etapa es un pre-procesado sobre el audio entrante que consiste en operaciones básicas, como el suavizado o filtrado de datos, que puede ser útil para identificar si el audio entrante contiene o no información relevante y por lo

tanto es conveniente continuar con el procedimiento o para convertir señales tipo estéreo a mono. La siguiente etapa consiste en obtener sub-segmentos de la señal entrante, los cuales se denominan *tramas*, ver Figura 7.a. Posteriormente se extraen características perceptuales a partir de cada trama. El resumen de las características de todas las tramas construye una firma de audio. Finalmente se aplica un post-procesamiento que puede consistir en una normalización o reducción de la dimensión de la firma final.

Normalmente se traslapan tramas adyacentes con el objetivo de mejorar la descripción de la señal al reducir las variaciones de la firma extraída (ver Figura 7.b). Además, el cálculo de algunas características perceptuales requiere pasar del dominio del tiempo hacia el dominio de la frecuencia utilizando la transformada de Fourier. Un requisito para evitar un fenómeno denominado **Leakage** al realizar el cambio de dominio consiste en multiplicar cada trama por una **función ventana**, ver Figura 7.c.

A continuación se describen algunas características perceptuales comunmente usadas en la literatura y posteriormente se explica el cálculo de dos firmas de audio.

2.2.7. Características perceptuales

Las características perceptuales, basadas en la percepción del oído humano, deben representar la información relevante para la tarea de clasificación, mientras desechan la información no relevante. Algunas ejemplos de características son los siguientes:

2.2.7.1. Tasa de cruce por cero

Esta característica indica el número de cruces por cero cada segundo en el dominio del tiempo. Es una forma simple de medir la frecuencia dominante de una señal sin pasar al dominio de la frecuencia (Kedem, 1986).

2.2.7.2. Energía en corto tiempo (STE)

Esta característica describe la envoltura de la señal la cual se consigue calculando la media de la energía por cada trama (Zhang and Kuo, 2001). La energía en una señal de duración finita $x(n)$ está definida como:

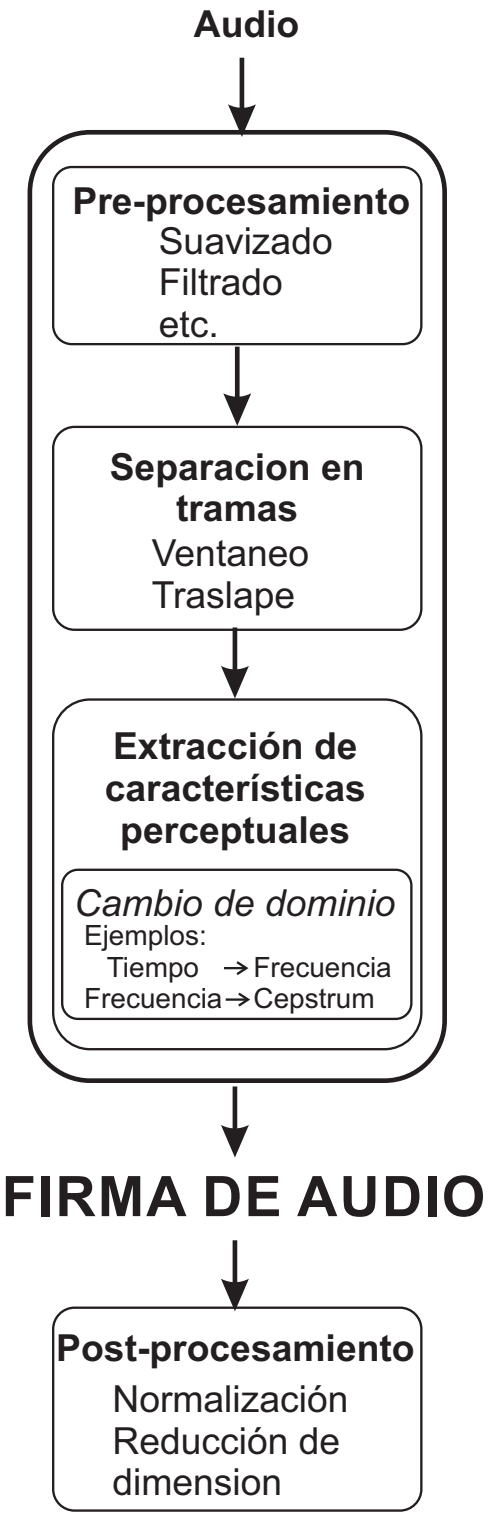


Figura 6: Procedimiento para obtener una firma de audio

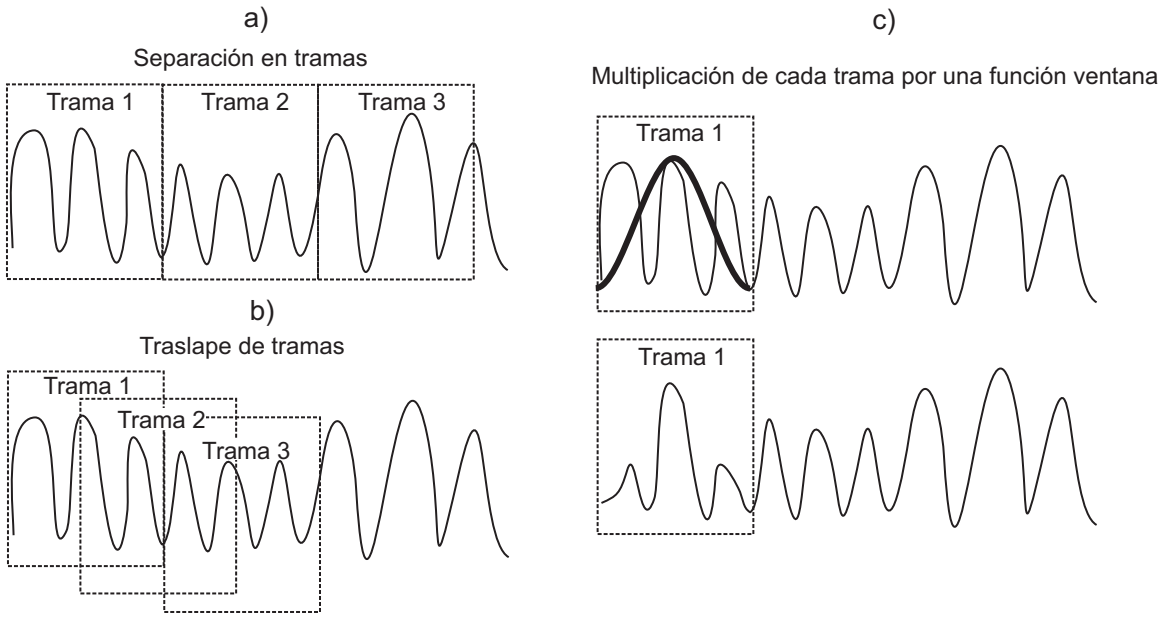


Figura 7: a) Separación en tramas de una señal de audio, b)Traslape de tramas, c) Multiplicación de una tramana por una función ventana.

$$E = \sum_{n=1}^{N} |x(n)|^2 \tag{3}$$

2.2.7.3. Flujo espectral

Es una característica que cuantifica los cambios del espectro a través del tiempo (Scheirer and Slaney, 1997). Se consigue obteniendo la distancia euclideana de las amplitudes del espectro en dos tramas contiguas. Las señales que muestran pocas variaciones en sus propiedades espectrales, por ejemplo el ruido, poseen un flujo espectral bajo mientras que las señales con cambios abruptos (como notas musicales) tienen un flujo espectral alto.

2.2.7.4. Planaridad espectral

Esta característica se utiliza para determinar si existe una distribución similar o planitud en todas las bandas del espectro de frecuencia. Para su cálculo se divide la media geométrica sobre la media aritmética de la potencia del espectro. Un valor cercano a uno representa una distribución similar, por ejemplo en señales de ruido blanco, mientras que un valor bajo indica que hay concentración de potencia en sólo algunas bandas de

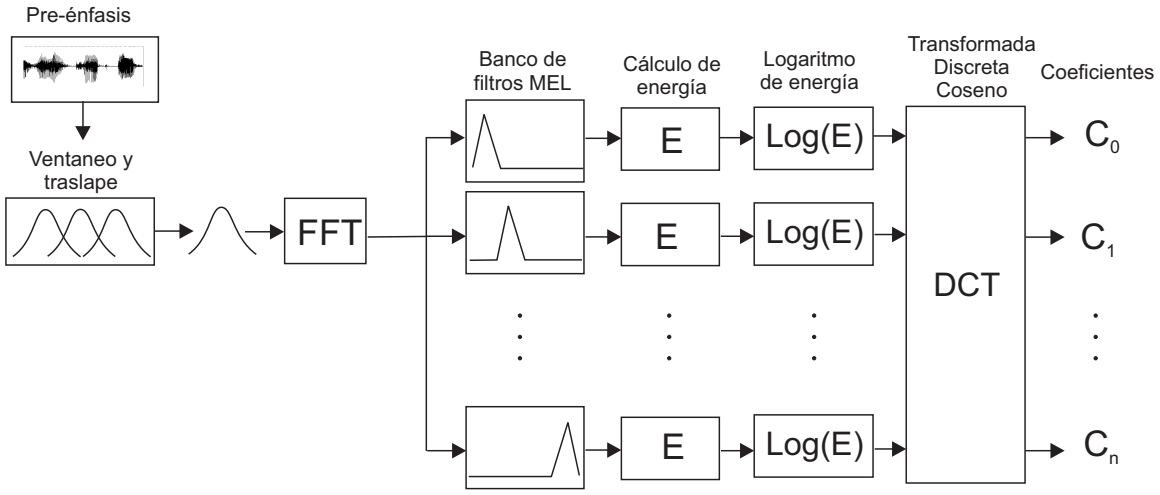


Figura 8: Procedimiento para calcular MFCC.

frecuencia, por ejemplo en tonos puros.

2.2.7.5. Mel Frequency Cepstral Coeficients (MFCC)

Esta característica surgió originalmente en el área de reconocimiento del habla y se mantiene como una de las características más populares en el análisis de audio ambiental (Sahidullah and Saha, 2012). Los coeficientes representan una aproximación a la envoltura espectral de una señal. El procedimiento para la obtención de los MFCC, ilustrado en la figura 8, inicia aplicando un filtro de pre-énfasis pasa altas a la señal de entrada. Posteriormente, se realiza el análisis sobre una ventana traslapada de tipo Hamming de tamaño N muestras. Cada ventana se transforma al dominio de la frecuencia mediante una Transformada de Fourier (FFT) de tamaño N. Después se aplica un banco de filtros triangulares en la escala de frecuencias **mel**. Para finalizar, las características del sonido en cada trama corresponden a los coeficientes obtenidos al aplicar la **Transformada del Coseno (DCT)** sobre el logaritmo de la energía de las bandas.

2.2.7.6. Multi-Band Spectral Entropy (MBSE)

Esta característica de audio surgió en el area de recuperación de música probando ser efectiva para encontrar canciones en colecciones grandes de datos aunque los segmentos de consulta contengan ruido, ecualización, filtrado, cambios de volumen y compresión. Para obtener la firma se utiliza el proceso mostrado en la figura 9. El análisis se realiza sobre una ventanas traslapada de tipo Hann con tamaño N muestras. Cada

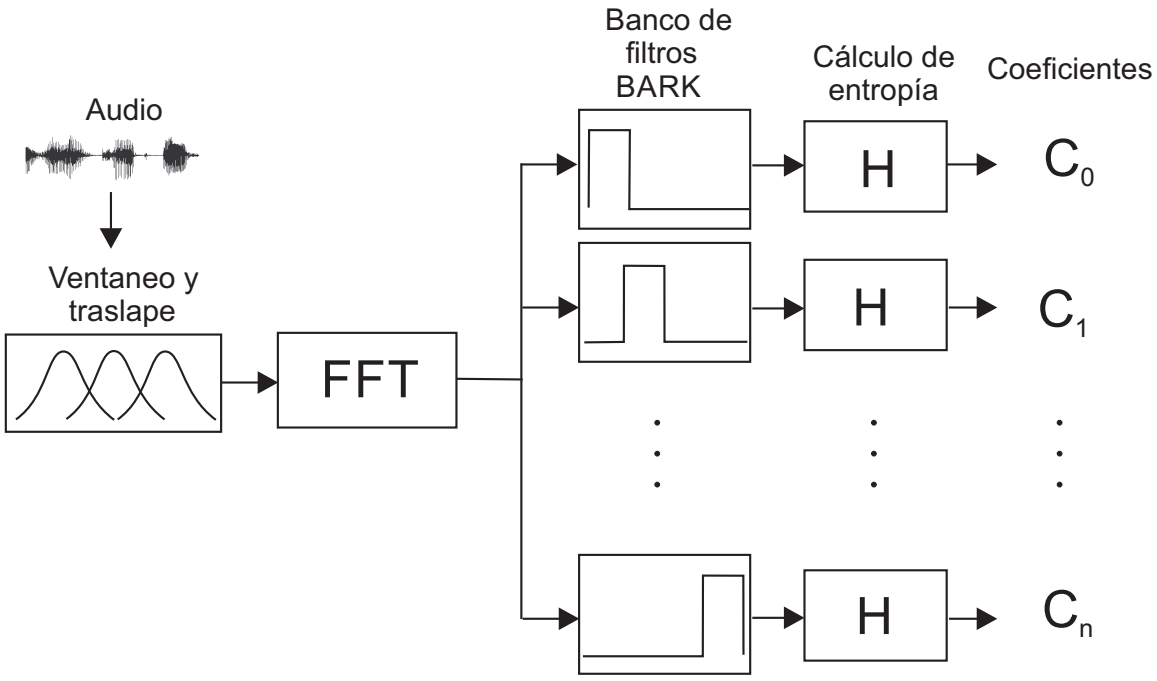


Figura 9: Procedimiento para calcular MBSE.

ventana se transforma al dominio de la frecuencia mediante una FFT de tamaño N. A continuación se aplica un banco de filtros rectangulares en las escala de frecuencias **Bark** desde 0 hasta 15500Hz. A cada banda se le determina la **entropía** usando la ecuación $H = \ln(2\pi e) + \frac{1}{2}\ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)$. Donde σ_{xx} and σ_{yy} son las varianzas de la parte real e imaginaria, y σ_{xy} es la covarianza entre las partes real e imaginaria de los coeficientes obtenidos con la FFT de las bandas correspondientes. Ver Figura 9.

En la Figura 10 se muestran dos ejemplos de firmas obtenidas usando los MFCC y las MBSE. Como se aprecia, se conjuntan las características obtenidas en cada una de las tramas.

2.3. Clasificadores

Un clasificador es un conjunto de algoritmos diseñados para identificar a cual de un grupo de clases pertenece una nueva observación. En la etapa de entrenamiento se tiene un algoritmo que consiste en encontrar los parámetros de un modelo basándose en instancias de las clases que se van a reconocer. La etapa de clasificación consiste en evaluar las nuevas observaciones en el modelo para determinar a la clase que pertenece.

A continuación se explican ejemplos de clasificadores, posteriormente se describe a

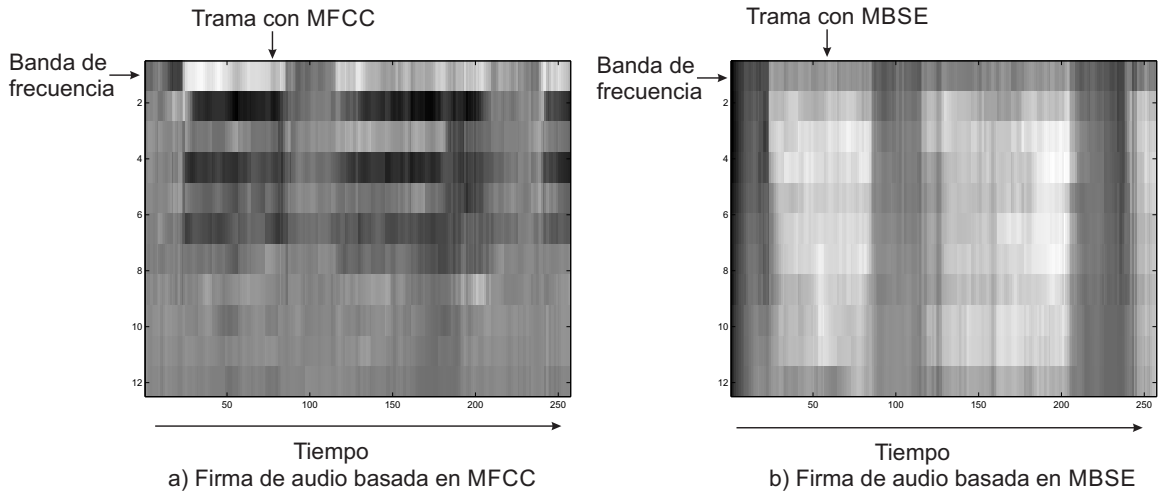


Figura 10: Ejemplos de firmas de audio basadas en características perceptuales MFCC y MBSE.

detalle las evaluaciones indicando las medidas utilizadas.

2.3.1. Máquina de soporte vectorial

2.3.1.1. Definición

La máquina de soporte vectorial (SVM por sus siglas en inglés), es un algoritmo de aprendizaje para clasificación y regresión. La SVM esta basada en el concepto de proyectar un conjunto de datos a un espacio de características de dimensión muy alta y entonces determinar hiperplanos óptimos para separar los datos de clases diferentes (Muller *et al.*, 2001).

2.3.1.2. Intuición

Dados N puntos de entrenamiento con dimension d, los cuales provienen de dos clases distintas $y_i = -1\ o\ +1$,, es decir:

$$\{x_i, y_i\}\ \ donde\ i = 1...N,\ \ y \in \{-1, 1\}\ ,\ \ x \in \mathbb{R}^d \tag{4}$$

El objetivo del algoritmo SVM consiste en encontrar un hiperplano con una orientación que permita separar ambas clases maximizando un margen entre el hiperplano y los puntos más cercanos de cada clase. En la figura 11 se ilustra un ejemplo en donde se supone que puntos de dos clases son linealmente separables por el hiperplano $\mathbf{w} \cdot x + b =$

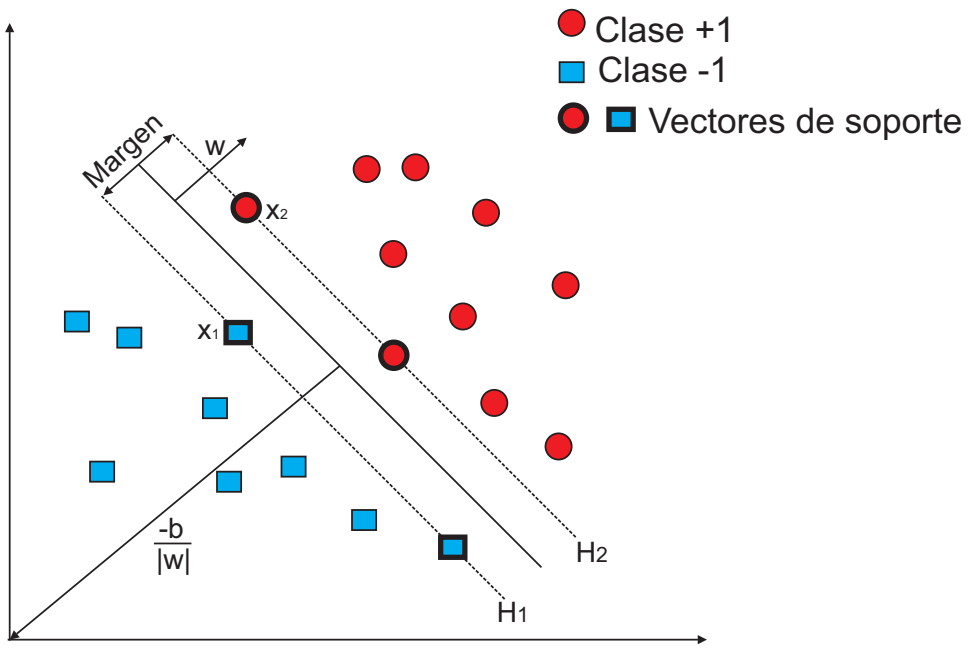


Figura 11: Hiperplano separando puntos de dos clases linealmente

0, donde **w** es normal al hiperplano y $\frac{b}{|\mathbf{w}|}$ es la distancia perpendicular desde el hiperplano hasta el origen. Cada uno de los semi-espacios definidos por este hiperplano corresponde a una clase. Para encontrar a cual clase pretence un punto x se puede evaluar usando: $f(x) = sgn(w \cdot x + b)$. El margen que se busca maximizar esta ilustrado en la figura como el espacio entre las líneas punteadas H_1 y H_2.

2.3.1.3. Hiperplano

El hiperplano se encuentra seleccionando las variables **w** y b de forma que los datos de entrenamiento cumplan con las siguientes ecuaciones.

$$\mathbf{w} \cdot x_i + b \geq +1 \quad para \quad y_i = +1 \tag{5}$$

$$\mathbf{w} \cdot x_i + b \geq -1 \quad para \quad y_i = -1 \tag{6}$$

Estas ecuaciones combinadas producen la siguiente ecuación:

$$y_i \, (\mathbf{w} \cdot x_i + b) - 1 \geq 0 \tag{7}$$

A los puntos que se encuentran más cercanos al hiperplano, se les conoce como

vectores de soporte. Los planos H_1 y H_2 donde caen los vectores de soporte se pueden describir como:

$$\mathbf{w} \cdot x_i + b = +1 \quad para \quad H_1 \tag{8}$$

$$\mathbf{w} \cdot x_i + b = -1 \quad para \quad H_2 \tag{9}$$

En la figura 11 los planos H_1 y H_2 se ilustran con líneas punteadas y los vectores de soporte se representan con los puntos con linea gruesa. Nótese que H_1 y H_2 son paralelos y que ningún punto de entrenamiento cae entre ellos.

Si se consideran dos vectores de soporte x_1 y x_2 de distintas clases con $(\mathbf{w} \cdot x_1) + b \geq +1$ y $(\mathbf{w} \cdot x_2) + b \geq +1$ respectivamente, el margen esta dado por la distancia entre estos dos puntos medidos perpendicularmente al hiperplano, es decir; $\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (x_1 - x_2) = \frac{2}{\|\mathbf{w}\|}$. El maximizar este margen es equivalente al problema de optimización de encontrar el mínimo de $\mathbf{w}$.

$$\min \|\mathbf{w}\| \quad \text{tal que} \quad y_i (\mathbf{w} \cdot x_i + b) - 1 \geq 0 \tag{10}$$

Veáse el Apendice para ver más detalles de la SVM.

2.3.1.4. Modelos Ocultos de Markov

2.3.1.5. Definición

Los modelos ocultos e Markov (HMM por sus siglas en inglés) son una técnica estadística que modela datos temporales y secuenciales que se pueden utilizar para caracterizar las propiedades estadísticas de una señal. Esta técnica tiene aplicaciones en el área de reconocimiento de patrones. El sistema a ser modelado se supone como un proceso de Markov con estados no observables (ocultos).

2.3.1.6. Proceso de Markov

Un proceso de Markov discreto es un proceso estocástico que satisface la propiedad de Markov, esta propiedad indica que se pueden hacer predicciones del futuro basándose solamente en el estado presente sin necesidad de conocer la historia completa de todo el proceso.

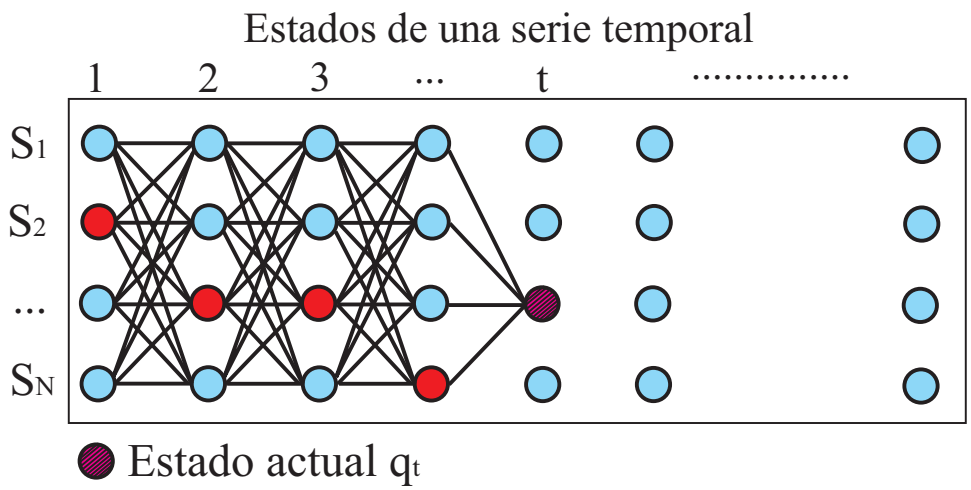


Figura 12: Proceso de Markov representando una serie temporal.

Para ejemplificar un proceso de Markov, supóngase una serie temporal la cual dado un tiempo t puede tomar algún estado de un conjunto finito de N estados distintos $S_1, S_2, ..., S_N$ como se ilustra en la figura 12. Al estado que ocurre en el tiempo t se le conoce como q_t.

Para describir un nuevo estado mediante una descripción completa de la secuencia temporal, se requiere que además de conocer el estado actual, se tenga información de todos los estados anteriores.

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...\right] \tag{11}$$

Sin embargo,un proceso de Markov discreto de primer orden indica que esta descripción se puede truncar y estar completamente determinada considerando solo el estado actual y el estado inmediato anterior

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...\right] = P\left[q_t = S_j | q_{t-1} = S_i\right] \tag{12}$$

Un modelo de Markov contiene las probabilidades de cambiar entre los estados, las cuales se consideran constantes en el tiempo. Una forma de representar estas probabilidades es mediante la matriz de transición de probabilidades A; cada elemento a_{ij} de A indica la probabilidad de pasar del estado S_i al estado S_j. La figura 13 ilustra un ejemplo.

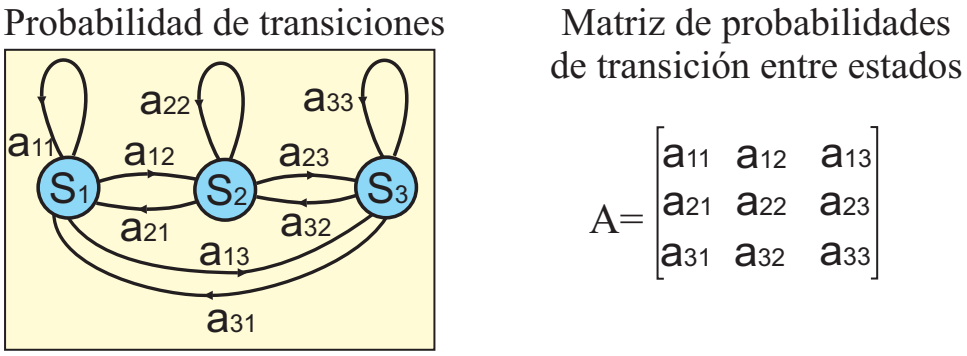


Figura 13: Probabilidades de transición entre estados en un modelo de Markov.

Cada estado indica un evento observable físico, por ejemplo el estado del clima (nublado, soleado, lluvioso) o elementos de un diccionario (A,B,C,D). Cuando se ha observado una serie de estados consecutivos, entonces se tiene una secuencia de observación O, por ejemplo $O = \{S_2, S_1, S1, S_3\}$ correspondientes a $t = 1, 2, 3, 4$. La probabilidad de una observación particular O dada un modelo se expresa:

$$P(O|\text{Modelo}) = P[S_2, S_1, S1, S_3|\text{Modelo}] \tag{13}$$

Hasta este momento, se ha descrito un modelo de Markov observable, ya que los estados son directamente observables en cada instante de tiempo y los únicos parámetros son las probabilidades de transición, tal como se muestra en la figura 13. A continuación se explican los modelos ocultos de Markov.

2.3.1.7. Modelos ocultos de markov discretos

En los modelos ocultos de Markov los estados se encuentran ocultos mientras lo visible corresponde a lo que se conoce como las emisiones de los estados. El ejemplo clásico para explicar HMM discretos es el modelo de la urna y las pelotas, (Rabiner, 1989). Se supone que existen N urnas las cuales no son visibles para un observador. Cada una de las urnas guarda cierta cantidad de pelotas coloredas con M colores distintos. Un genio tiene acceso a las urnas y elige una de acuerdo un proceso aleatorio de selección de urnas. Posteriormente extrae una pelota de dicha urna y le señala al observador el color de la pelota. Después regresa la pelota a la urna de la que fue seleccionada. A continuación selecciona nuevamente una urna de acuerdo al proceso aleatorio de selección de urnas

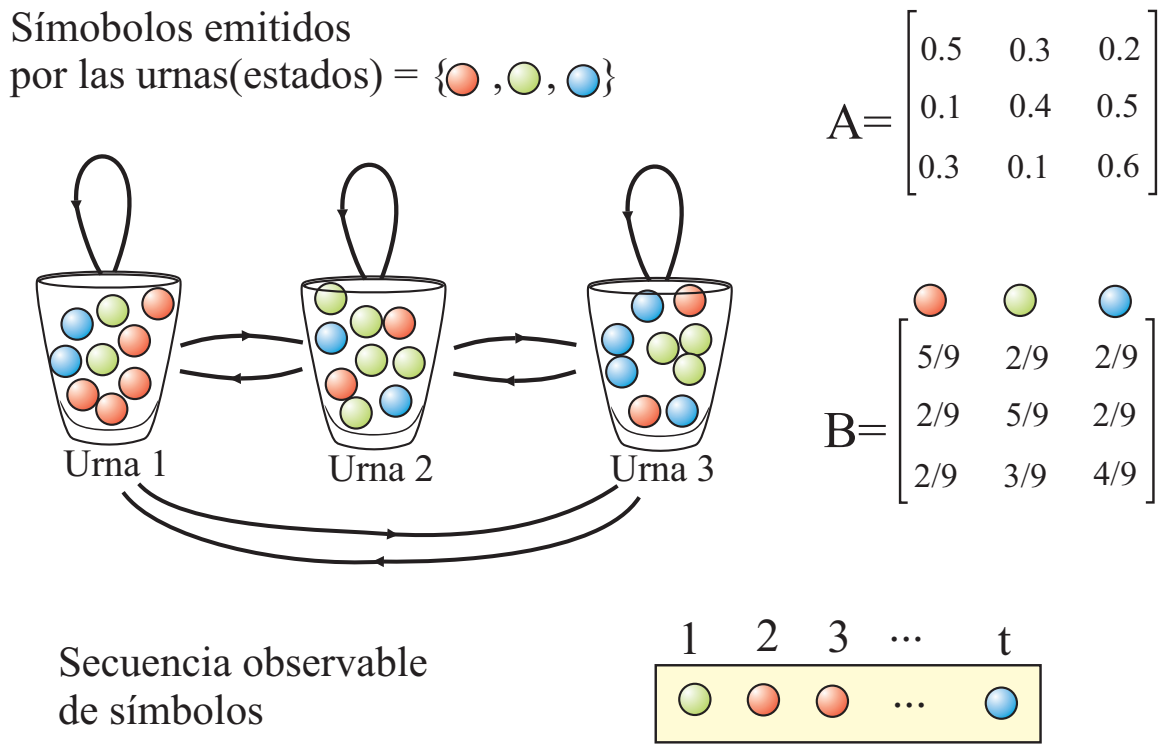


Figura 14: Elementos de los modelos ocultos de Markov.

y vuelve a extraer una pelota. El proceso se repite y produce como resultado una observación finita de colores. La elección de las urnas no depende de las urnas que fueron elegidas con anterioridad, a esto se le conoce como proceso de Markov, sin embargo la eleccion de las urnas esta oculta al observador y lo único que puede ser observado son los símbolos emitidos por las urnas, es decir, los colores.

En la figura 14 se ilustran N=3 urnas que contienen pelotas con M=3 colores distintos (rojo, verde y azul). Las urnas son los estados de un proceso de Markov con su respectiva matriz de probabilidades de transición A. En los HMMs Se introducen los valores $B = \{b_j(k)\}$ que indican la distribución de probabilidad de los símbolos observados en cada estado, para este ejemplo clásico, la matrix B indica la probabilidad de encontrar un color especifico en una urna dada. La secuencia temporal observable $O = O_1 O_2 ... O_T$ contiene los símbolos o colores emitidos por los estados.

Los elementos que conforman un modelo oculto de Markov son los siguientes:

1. N, el numero de estados en el modelo $S_1, S_2, ..., S_N$.

2. M, la cantidad de símbolos observados por estado.

3. A, la matriz de distribución de probabilidad de transiciones entre estados.

4. B, la matriz de distribución de probabilidad de observaciones de los símbolos en cada estado.

5. π, una distribución inicial de los estados.

Un HMM λ se representa mediante estos elementos: $\lambda = (A, B, \pi)$.

Para la aplicación práctica de los HMMs, existen tres problemas algorítmicos los cuales se explican a continuación.

Problema de evaluación. Dada una secuencia de observación $O = O_1 O_2 ... O_T$ y un modelo particular λ, se requiere calcular $P(O|\lambda)$. Este problema se resuelve con un algoritmo conocido como *Forward-Backward Procedure*. La solución de este problema es útil cuando existen modelos para cada clase previamente entrenados y se quiere aquella clase con la cual una observación tiene mayor similitud.

Problema de decodificación. Calcular la secuencia de estados óptima que mejor explica una secuencia de observaciones. Este problema se resuelve con el algoritmo *Viterbi*. La obtención de la secuencia optima se puede usar para aprender cual es la estructura del modelo u obtener estadísticas acerca de los estados individuales.

Problema de aprendizaje. Ajustar los parámetros de un modelo para maximizar $P(O|\lambda)$. Para estimar la probabilidad maxima de la observación dado un modelo se utiliza un algoritmo iterativo llamado *Baum-Welch* o su equivalente *Expectation-Maximization (EM)*. El ajuste de los parámetros del modelo se utiliza en la etapa de entrenamiento en donde se generan modelos para las secuencias temporales de entrenamiento provenientes de cada clase.

Para conocer los algoritmos que se utilizan en cada uno de los tres problemas y extender el problema discreto al problema continuo, veáse el Apéndice.

Algunas diferencias entre distintos tipos de clasificadores son; la cantidad de datos que

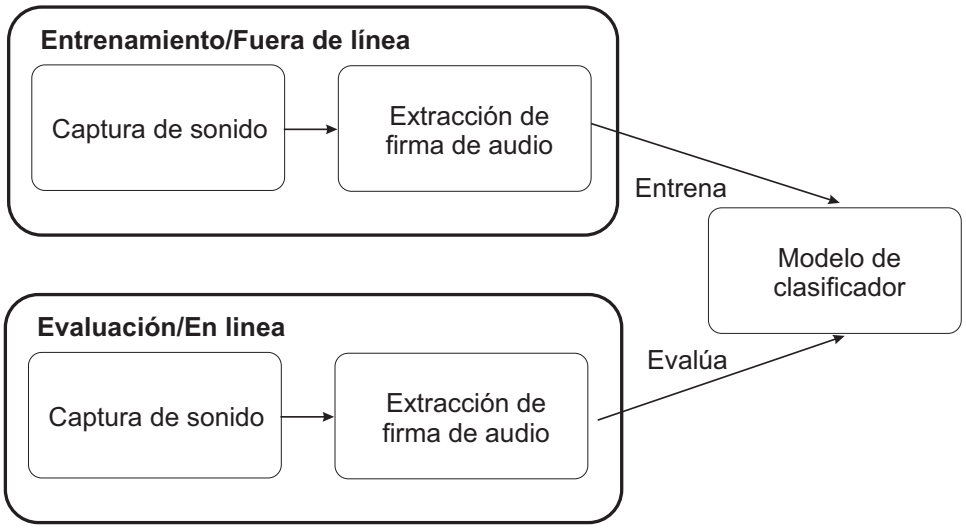


Figura 15: Procedimiento de entrenamiento y evaluación usado en sistemas de reconocimiento de audio.

se requiere para el entrenamiento, la convergencia del algoritmo de entrenamiento y las capacidades de generalización. Una forma de saber que clasificador elegir se consigue haciendo evaluaciones y comparando los resultados.

2.4. Evaluaciones en sistemas de clasificación de audio

Para que los clasificadores puedan asignar una clase al audio capturado, primero se deben de entrenar modelos a partir de muestras de sonidos. El entrenamiento es un proceso que normalmente consume mucho tiempo de cómputo, sin embargo, este se realiza fuera de línea y no afecta el desempeño de los sistemas en tiempo real. El procesamiento que se realiza en línea, consiste en extraer la firma del sonido entrante y evaluar en los modelos de clasificación previamente entrenados. En la Figura 15 se ilustra el proceso de entrenamiento y de prueba para un sistema de reconocimiento de audio.

Al conjunto de los sonidos que se utilizan para el entrenamiento y para la evaluación se le llama *base de datos*. Si se lleva a cabo una **clasificación supervisada** entonces se debe indicar explícitamente a que clase pertenece cada sonido de la base de datos, a esto se le conoce como *etiquetado*. En la Figura 16 se ilustra un ejemplo de una base de datos con 24 muestras de sonidos etiquetadas con 3 clases y su uso para entrenar modelos. Además, la Figura 16 ilustra que los sonidos designados para la evaluación se prueban en todos los modelos y posteriormente se toma una decisión que indica la clase.

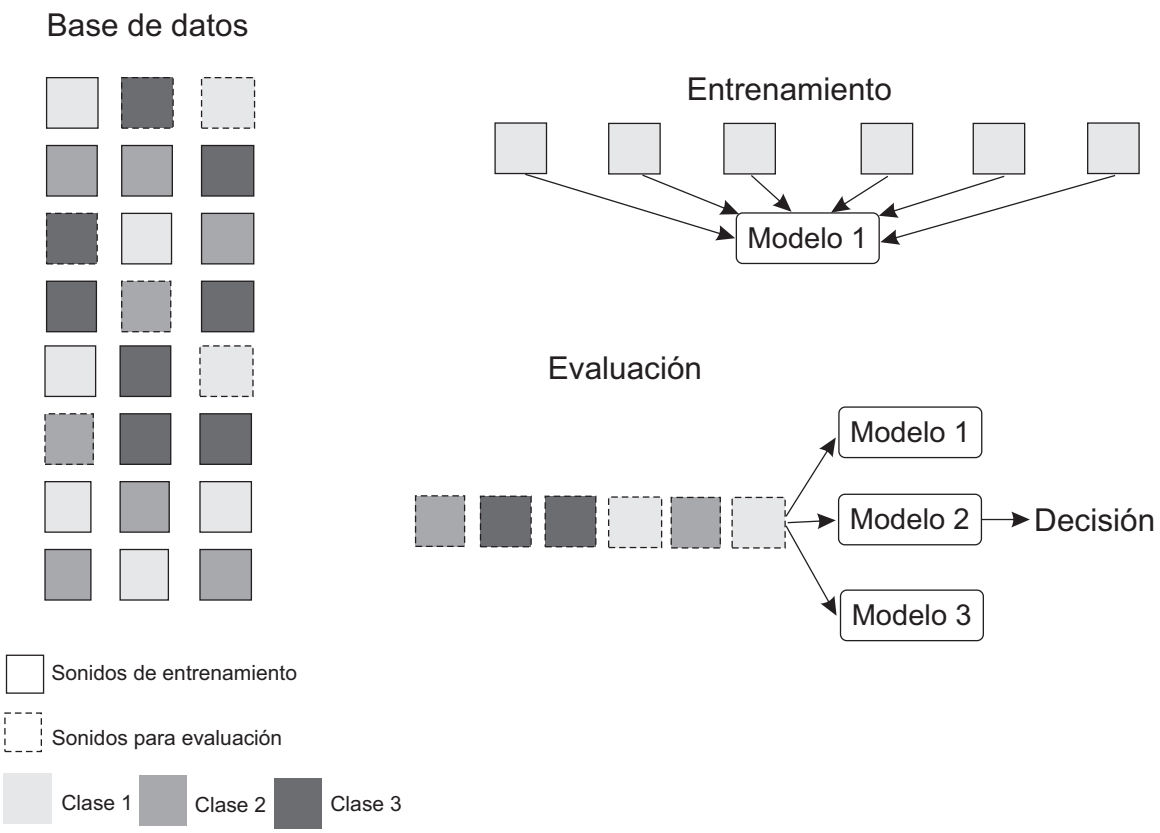


Figura 16: Ejemplo de partición de bases de datos para entrenamiento y evaluaciones para HMMs.

Los sonidos usados en el entrenamiento no deben de utilizarse para la evaluación ya que es necesario conocer la capacidad de generalización del sistema. Normalmente, la base de datos se parte en dos secciones, donde un porcentaje se designa para el entrenamiento y el resto para la evaluación. Existe una técnica llamada *validación cruzada* que consiste en hacer varias particiones de la base de datos y calcular la media aritmética de las medidas de evaluación sobre las diferentes particiones. Como se ilustra en la Figura 17 en una validación cruzada de K iteraciones, la base de datos se parte en K subconjuntos, donde (K-1) subconjuntos se utilizan para el entrenamiento y un subconjunto se usa como datos de prueba.

Como se menciona en esta sección, cuando se introducen los sonidos de prueba ya se conoce de antemano a que clase pertenecen. Por lo tanto, se puede establecer una comparación entre las clases predichas por el clasificador y las clases reales. A continuación se explican algunas medidas que utilizan estas comparaciones y sirven para conocer la eficiencia del esquema firma-clasificador utilizado.

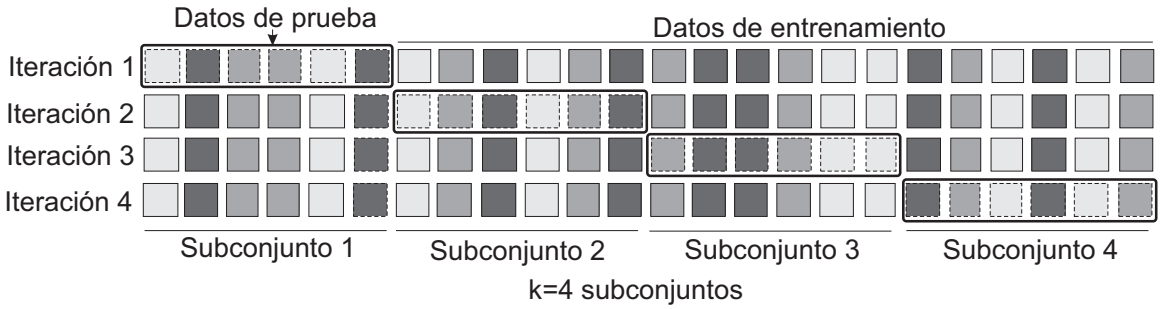


Figura 17: Ejemplo de particiones para hacer una validación cruzada con 4 subconjuntos.

Verdaderos Positivos Los verdaderos positivos (VP) se dan cuando se predice correctamente que un sonido pertenece a una clase. Por ejemplo, se predice que un sonido corresponde a la clase 1 y efectivamente pertenece a esa clase.

Verdaderos negativos Los verdaderos negativos (VN) se dan cuando se predice correctamente que un sonido no pertenece a una clase. Por ejemplo, se predice que un sonido no corresponde a la clase 1 y efectivamente no pertenece a esa clase.

Falsos positivos Los falsos positivos (FP) se dan cuando se predice incorrectamente que un sonido pertenece a una clase. Por ejemplo, se predice que un sonido corresponde a la clase 1, pero en realidad pertenece a la clase 2.

Falsos negativos Los falsos negativos (FN) se dan cuando se predice incorrectamente que un sonido no pertenece a una clase. Por ejemplo, se predice que un sonido no corresponde a la clase 1, pero en realidad si pertenece a la clase 1.

Precision La precisión, indica la relación entre los VP sobre todos los resultados positivos. Ayuda a conocer que fracción de todos los sonidos etiquetados como positivos son correctos.

$$Precisión = \frac{VP}{VP + VN} \tag{14}$$

Exhaustividad La exhaustividad, indica la relación entre los VP y todos los elementos positivos de la base de datos. Ayuda a conocer la fracción de todos los ejemplos positivos de la base de datos que el clasificador pudo predecir correctamente.

$$Exhaustividad = \frac{VP}{VP + FN} \tag{15}$$

Medida F1 La precisión y el exhaustividad por si solas no reflejan completamente el comportamiento del sistema. Por ejemplo, una exhaustividad de 1 indica que todos los sonidos positivos de la base de datos fueron predichos como positivos por el clasificador, sin embargo, esto se pudo haber conseguido haciendo que el clasificador prediga como positivos a todos los sonidos, lo que disminuye la precisión. Por tal motivo, para conocer de una mejor manera el desempeño del sistema se requiere reportar ambos valores. También se puede utilizar otra medida, denominada Medida F1 la cual combina a la precisión y al exhaustividad como se muestra en la siguiente ecuación.

$$MedidaF1 = 2\frac{Precisión \cdot Exhaustividad}{Precisión + Exhaustividad} \tag{16}$$

Capítulo 3. Trabajo previo en sistemas de reconocimiento de sonidos ambientales

El reconocimiento automático de sonidos ambientales es un área relativamente nueva en comparación de la atención que han recibido las áreas de reconocimiento de voz y música. Hasta antes de un par de décadas los sonidos ambientales se descartaban por considerarse como ruido mientras se estudiaban aplicaciones para voz y música. Sin embargo; en los últimos años se ha encontrado que los sonidos ambientales producen información importante para describir el contexto.

Resulta complicado comparar los trabajos que abordan el reconocimiento de sonidos, debido a que los autores de diversas investigaciones utilizan diferentes bases de datos e intentan clasificar diferentes eventos de sonido. Por ejemplo, mientras un trabajo de investigación busca reconocer los sonidos *cepillarse, lavándose las manos y rasurarse* (Min *et al.*, 2008a), otro trabajo investiga el reconocimiento de los sonidos *tosiendo, aplaudiendo, tocando la puerta y teléfono sonando* (Schroeder *et al.*, 2011). En la actualidad no existe una base de datos multi-contexto diseñada para realizar comparaciones, motivo por el cual, cada grupo de investigación graba y etiqueta su propia base de datos (Heittola *et al.*, 2013).

La evolución de los trabajos de investigación ha permitido que cada vez se aborden problemas más complejos relacionados al reconocimiento de contexto, por ejemplo la existencia de ruido o los sonidos mezclados. También se ha permitido explorar esquemas de reconocimiento considerando las propiedades de los sonidos ambientales, ya que en un inicio se utilizaban características perceptuales que originalmente fueron diseñadas para el reconocimiento de voz, la cual es más estructurada que los sonidos ambientales. Además, los aportes se han dirigido hacia el diseño de esquemas más escalables y se ha explorado el uso de micrófonos en dispositivos móviles en lugar de colocarlos en lugares predefinidos que buscan reconocer actividades específicas que dependen del lugar en donde se ubica el micrófono.

Las diferencias en los esquemas firma de audio-clasificador, las clases de sonidos y la cantidad y ubicación de los micrófonos propuestas por distintos aportes complica la

comparación entre ellos. En la Tabla 1 se presenten ejemplos que ilustran esta diversidad.

Tabla 1: Comparación de propuestas para el reconocimiento de sonidos ambientales

Título	Clases	Ubicación micrófonos	Características	Clasificador	Resultados
Implementation and Evaluation of a Low-Power Sound-Based User Activity Recognition System (Stager *et al.*, 2004)	Martillando, Serruchando, Barrenando, Moliendo y Limando	Muñeca y pecho	Tasa de cruce por cero, Fluctuación de amplitud y Relación de energía en 4 sub-bandas logarítmicas	K-Vecinos más cercanos (KNN)	73 % Tasa de reconocimiento
Bathroom Activity Monitoring Based on Sound (Chen *et al.*, 2005)	Bañándose, Orinando, Descargando el retrete, Lavándose las manos y Suspirando	Cerca del lavamanos	MFCC	HMM	Tasa de exactidud sobre 84 % en todos los casos
Sensing from the Basement: A Feasibility Study of Unobtrusive and Low-Cost Home Activity Recognition (Fogarty *et al.*, 2006)	Lavadora de ropa, Lavadora de trastes, Regadera, Retrete y Lavamanos	Micrófonos en tuberías	Media cuadrática, tasa de cruce por cero, entropía	SVM	81.3 % de exhaustividad
Early morning activity detection using acoustics and wearable wireless sensors (Min *et al.*, 2008b)	Lavado de dientes, Lavado de dientes eléctrico, Lavado de cara, Rasurado manual, Rasurado eléctrico	Micrófono en muñeca	No especifica	Modelos Gaussianos Mezclados	90 % exactitud
Advanced Patient or Elder Fall Detection based on Movement and Sound Data (Doukas and Maglogiannis, 2008)	Caídas y Correr	Micrófono en el pie	Short Time Fourier Transform (STFT)	SVM	100 % en caídas y 96.72 % en correr

En cuanto al problema del ruido, los primeros aportes, se realizaban sobre señales que sólo contenían información de la fuente de sonido sin estar contaminadas con otros sonidos (B. Gygi and Watson, 2004). Sin embargo, en los últimos años, se ha considerado analizar la robustez al ruido, por ejemplo en (Heittola *et al.*, 2011) (Dennis *et al.*, 2013) (Dennis *et al.*, 2012) donde se generan mezclas artificiales agregando sonidos a la señal original. También, se ha experimentado identificar sonidos mezclados naturalmente en el ambiente, por ejemplo, usando una estrategia, llamada *beamforming*, que consiste en utilizar varios micrófonos y después separar las fuentes de sonido individual estimando las señales y los ángulos que llegan a cada micrófono (Drake *et al.*, 2002).

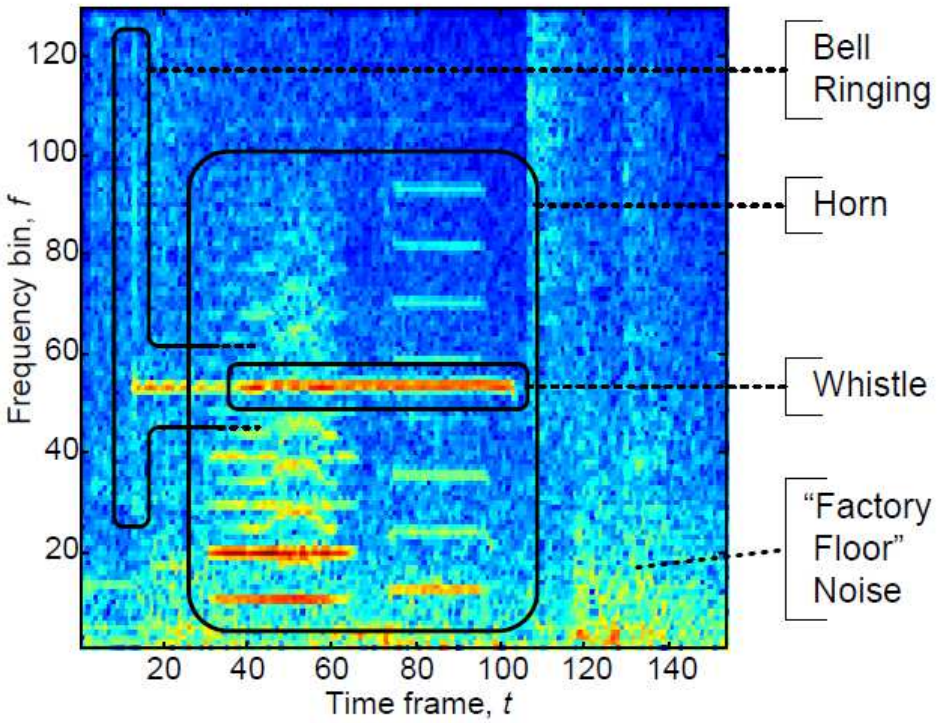
Al usar sólo un micrófono, se permite más flexibilidad y escalabilidad, pero trae consigo mayor complejidad para analizar el contexto auditivo. Sin embargo, existen trabajos de investigación que buscan reconocer eventos de sonidos grabando con un micrófono. Hay resultados y avances importantes descritos a continuación.

Un ejemplo de este tipo de trabajos se presenta a continuación, donde se describe el

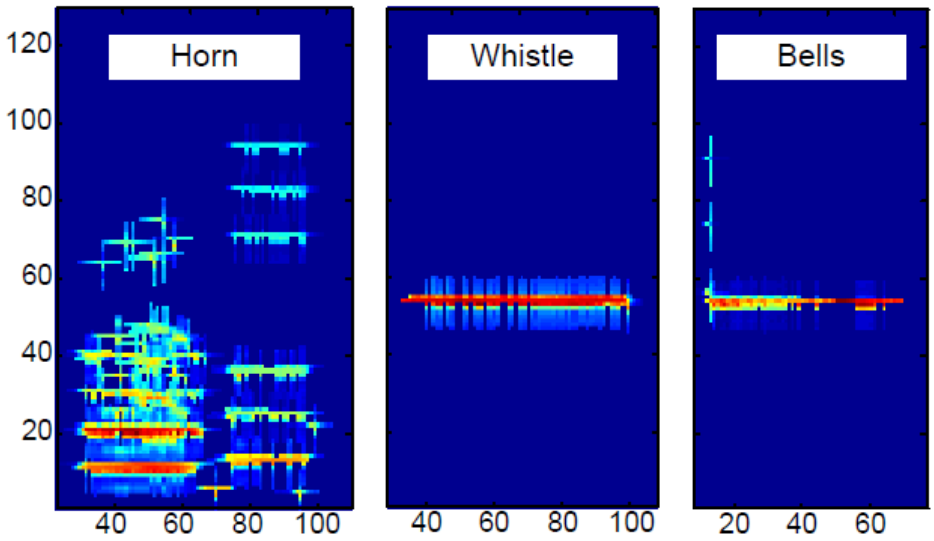
esquema propuesto por Dennis *et al.* (2013). En esta propuesta se busca reconocer mezclas que contienen a los eventos de sonido de tipo impulsivo *timbre de teléfono, bocina de coche, chiflido, soplido, campana y toque a una botella* (Dennis *et al.*, 2013) obtenidos de la base de datos *Real World Computing Partnership (RWCP)* (Ono *et al.*, 2015). Además, cada mezcla se contamina con un sonido de fondo no estacionario usando diferentes valores de SNR, en la Figura 18.a se muestra un ejemplo de sonidos mezclados y un ruido de fondo. El objetivo es identificar cada uno de los sonidos contenidos en las mezclas por medio de un análisis del **espectrograma** de donde se detectan puntos de interés mediante características locales y la transformada de Hough, ver Figura 18.b. En esa propuesta, se utilizan 20 instancias de sonidos de cada una de las 5 diferentes clases para el entrenamiento y para la evaluación se generan 50 mezclas de sonidos obtenidas a partir de traslapar cada una de 15 combinaciones posibles entre sonidos provenientes de las 5 clases. Como resultados, se reporta una exactitud de reconocimiento (relación entre las detecciones correcta y el número de mezclas que contienen a la clase) de 98 % y una tasa de clasificación incorrecta (relación de detecciones incorrectas y el número de clips que no contienen esa clase) de 0.8 %. Estos resultados aunque resultan altos, son para este tipo de sonidos cortos, impulsivos y con poca similitud inter-clase.

El trabajo de investigación que consideramos el estado del arte antes de nuestra contribución (Heittola *et al.*, 2011), propone separar la señal original, que contiene una mezcla de sonidos, en pistas que contienen cada una de las fuentes que conforman la mezcla mediante el algoritmo **Non Negative Matrix Factorization (NMF)**, ver Figura 19. Una vez que se tiene la separación, se aplica el esquema de clasificación estándar MFCC-HMM sobre cada una de las pistas. Sin embargo, debido a que el algoritmo de separación es ciego y no supervisado, las pistas resultantes no corresponden a las fuentes originales. Esto produce que se escuchen como versiones distorsionadas del sonido original, lo cual afecta el desempeño final de la propuesta.

La base de datos que utilizan los autores esta compuesta de 61 clases de sonido. Para evaluar el sistema se obtiene la medida F1 comparando los sonidos detectados contra los sonidos etiquetados en bloques de 30 segundos. Esta métrica esta diseñada para aplicaciones que requieren baja resolución poniendo más atención en detectar un



a) Ejemplo de una mezcla con tres clases de sonidos traslapados más ruido no estacionario. Aquí, un componente de frecuencia del sonido de la campana *(Bell ringing)* esta traslapado con el sonido del silbido *(Whistle),* el cual también esta siendo traslapado por el sonido de la bocina *(Horn).* El sonido *"Factory floor"* se considera



b) Ejemplo del reconocimiento de tres sonidos traslapados mediante la segmentación del espectrograma usando los puntos de interés detectados.

Figura 18: Imagen tomada del artículo (Dennis *et al.*, 2013)

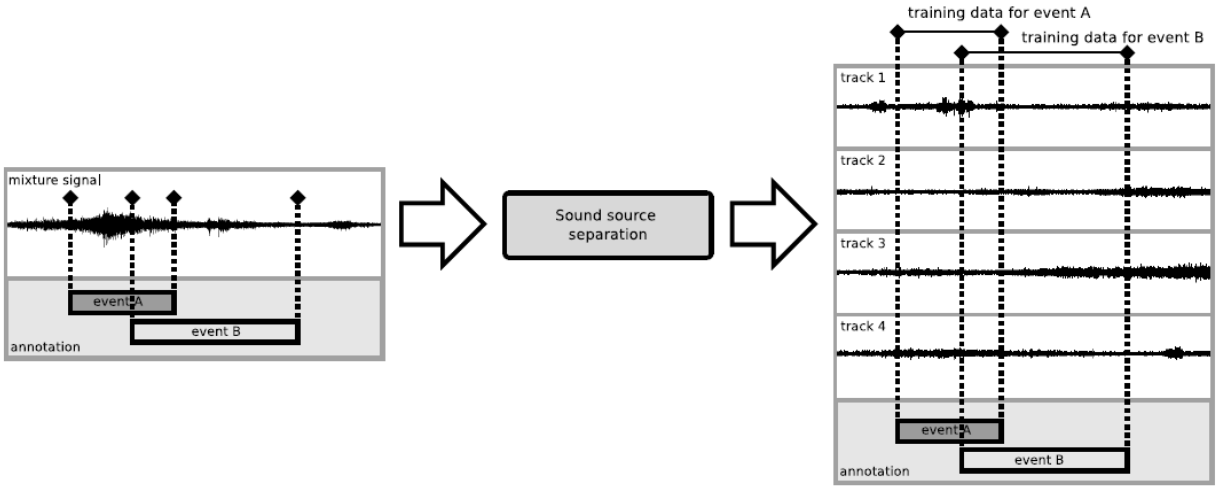


Figura 19: Procedimiento de separación en pistas de una señal original. Imagen tomada de (Heittola *et al.*, 2011)

evento que en su ubicación exacta. En la Figura 20 se muestra un ejemplo para obtener la precisión y la exhaustitivdad. Por ejemplo, en el bloque 1 se tienen etiquetados los sonidos A,B,C y D mientras que el sistema predice a A, C y E. Debido a que A y C están detectados correctamente, esto indica que la precisión es (2/3) y la exhaustividad es (2/4).

Otro ejemplo esta dado por un trabajo inspirado en técnicas de procesamiento de imágenes (Dennis *et al.*, 2012). El procedimiento consiste en generar una imagen formada con histogramas que provienen de la distribución de las sub-bandas de frecuencia sobre la duración del sonido. Posteriormente se aplican estadísticas sobre los sub-bloques de estas imágenes y se genera un vector que contiene a las medias y varianzas (ver Figura 21). El sonido se representa con vectores que se clasifican mediante el algoritmo de vecinos más cercanos. En los experimentos desarrollados en este trabajo, se utilizan 80 muestras de 50 eventos de sonido obtenidos de la base de datos RWCP. Aunque cada muestra se contamina con ruido estacionario sólo se intenta clasificar a los eventos de sonido sin existir traslapes y sin ningún interés en identificar el fondo. Además, se utiliza el recurso de modelar el ruido de fondo para permitir una mejor clasificación. Este esquema se asemeja a la propuesta de esta tesis debido a que utilizamos histogramas para representar los sonidos. Los resultados arrojan un 96 % de precisión de reconocimiento y muestran el potencial de utilizar una firma compacta para conseguir el reconocimiento de eventos de sonidos.

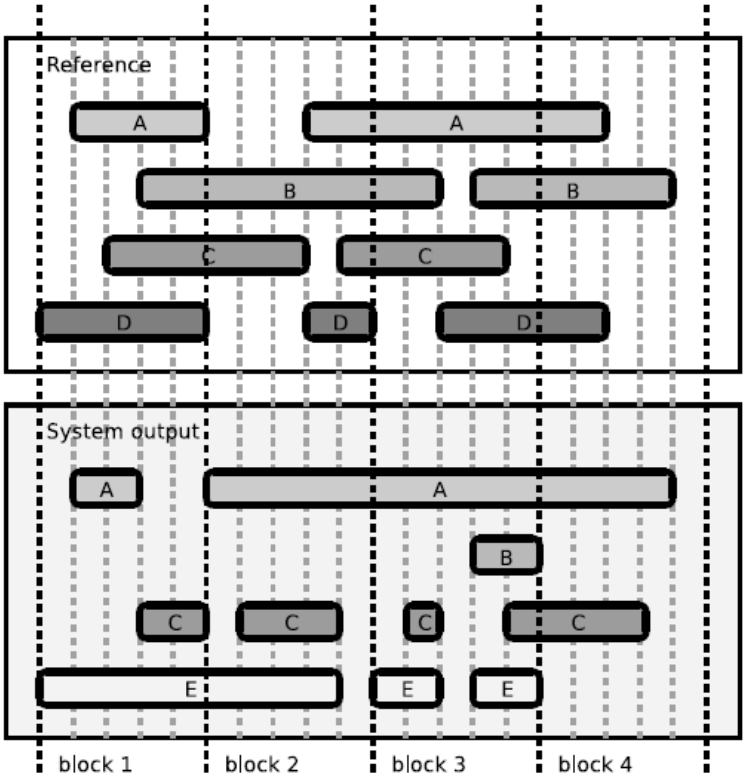


Figura 20: Ejemplo de obtención de la precisión y la exhaustividad para la clasificación de eventos de sonido. Imagen tomada de (Heittola *et al.*, 2011)

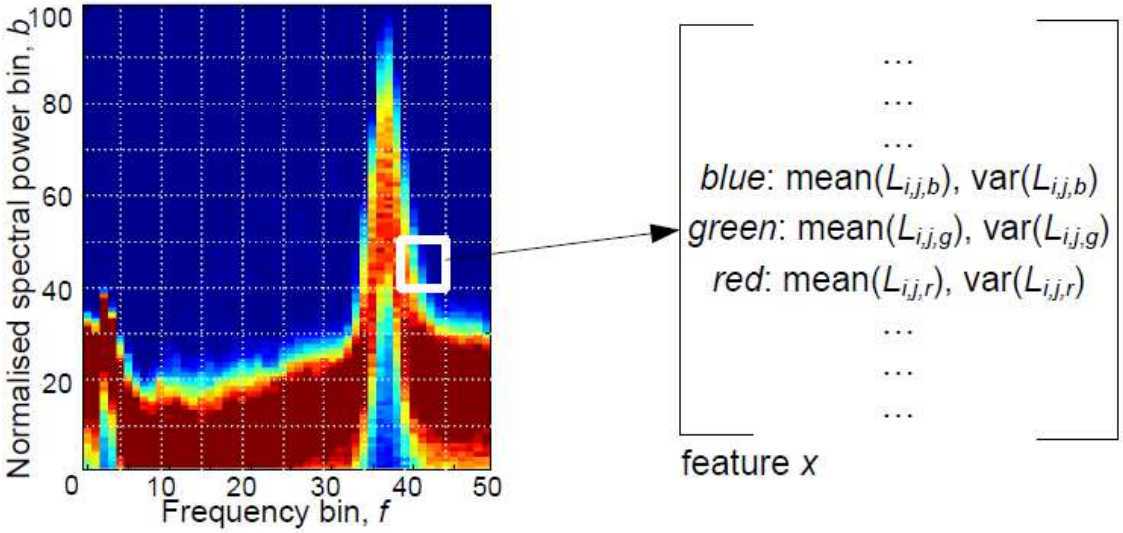


Figura 21: Partición de la imagen en sub-bloques y obtención de las estadísticas media y varianza sobre cada sub-bloque para generar el vector representante x. Imagen tomada de (Dennis *et al.*, 2012)

En las evaluaciones desarrolladas en la presente tesis, hacemos la comparación con el trabajo de (Heittola *et al.*, 2011), que para nosotros representa el único enfoque semejante al aquí presentado, es decir, busca resolver el problema considerando las mismas restricciones. Durante el desarrollo de esta tesis también intentamos comparar con el trabajo de (Dennis *et al.*, 2013) y (Dennis *et al.*, 2012) pero estas técnicas no son adecuadas dado que no buscan reconocer sonidos producidos por actividades de la vida diaria de adultos mayores (ej. Lavado de manos, Lavado de dientes, tosiendo). Con estos algoritmos no fue posible encontrar puntos de interés útiles en el espectrograma ya que los sonidos que buscamos identificar contienen un espectro relativamente plano y no son impulsivos.

Los clasificadores, aún los basados en el enfoque **difuso**, están diseñados para asignar sólo una clase a la entrada (ej. la entrada con mayor probabilidad). Esto es apropiado para sonidos aislados pero no resuelve directamene el problema de sonidos mezclados. Un recurso utilizado es entrenar los clasificadores con clases que contienen a los sonidos aislados, y además incluir clases con ejemplos de sonidos ya mezclados (Temko and Nadeu, 2009a); sin embargo, este enfoque requiere muestras de las combinaciones posibles de las clases y esto implica un entrenamiento con un número exponencial de clases.

3.0.1. Tamaño de las firmas

Otro aspecto importante es el tamaño de la representación final del audio. Para reducir el tamaño se utiliza un procedimiento tradicional que implica aplicar una transformación a través de proyectar el vector con un espacio de dimension alta a un espacio de menor dimensión conservando o permitiendo una clasificación eficiente. Existe un compromiso entre el limite de la reducción y la perdida de información (Potamitis and Ganchev, 2008). Algunos ejemplos de estas transformaciones incluyen al análisis discriminante lineal y el análisis de componentes principales (Eronen *et al.*, 2006) (Ntalampiras *et al.*, 2008). Ambas técnicas representan tiempo de procesamiento adicional además del cálculo de la característica.

Capítulo 4. Contribuciones de diseño de firmas de audio basada en entropía

En esta capitulo se presenta el diseño de la firma de audio para sonidos ambientales, el cual se realizó en dos fases. En la primera etapa se consideraron los retos de ruido y la transparencia para sonidos mezclados. En la siguiente etapa se trató la heterogeneidad intra-clase y similitud inter-clase. Adicionalmente se buscó representar sonidos de distinto tamaño mientras se mantiene la robustez al ruido y la transparencia.

4.1. Diseño de firmas de audio para sonidos ambientales mezclados

En la primera fase, se diseñaron cinco firmas de audio; MEL-MBSES, MEL-CMBSES, B-MFCC, B-MEL-MBSES y B-MEL-CMBSES. Estas firmas fueron obtenidas a partir de información espectral ya que entre sonidos ambientales se presentan diferencias en su distribución espectral. Para su diseño, se tomó en cuenta utilizarlas en aplicaciones con sonidos ruidosos y mezclados.

Mel Multiband Spectral Entropy Signature (MEL-MBSES)

El procedimiento para calcular la MEL-MBSES, que es una firma basada en la entropía espectral de bandas de frecuencia mel (ver figura 22), consiste en dividir la señal en ventanas traslapadas de tipo Hann con tamaño de N muestras. Posteriormente, cada ventana se transforma al dominio de la frecuencia mediante una FFT de tamaño N. A continuación se aplica un banco de 12 filtros rectangulares en la escala de frecuencias MEL desde 0 hasta 15500Hz. Se calcula la entropía sobre la información espectral de cada banda usando la ecuación $H = \ln(2\pi e) + \frac{1}{2}\ln(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)$. Donde σ_{xx} and σ_{yy} son las varianzas de la parte real e imaginaria, y σ_{xy} es la covarianza entre las partes real e imaginaria de los coeficientes obtenidos con la FFT de las bandas correspondientes. Con este proceso se obtiene una firma análoga al espectrograma, el cual calcula la cantidad de energía en tiempo y frecuencia, a la que denominamos *entropigrama*. Esta firma se puede visualizar como una imagen, ver Figura 23.a, y nos brinda la cantidad de información a través del tiempo para cada banda en la escala MEL.

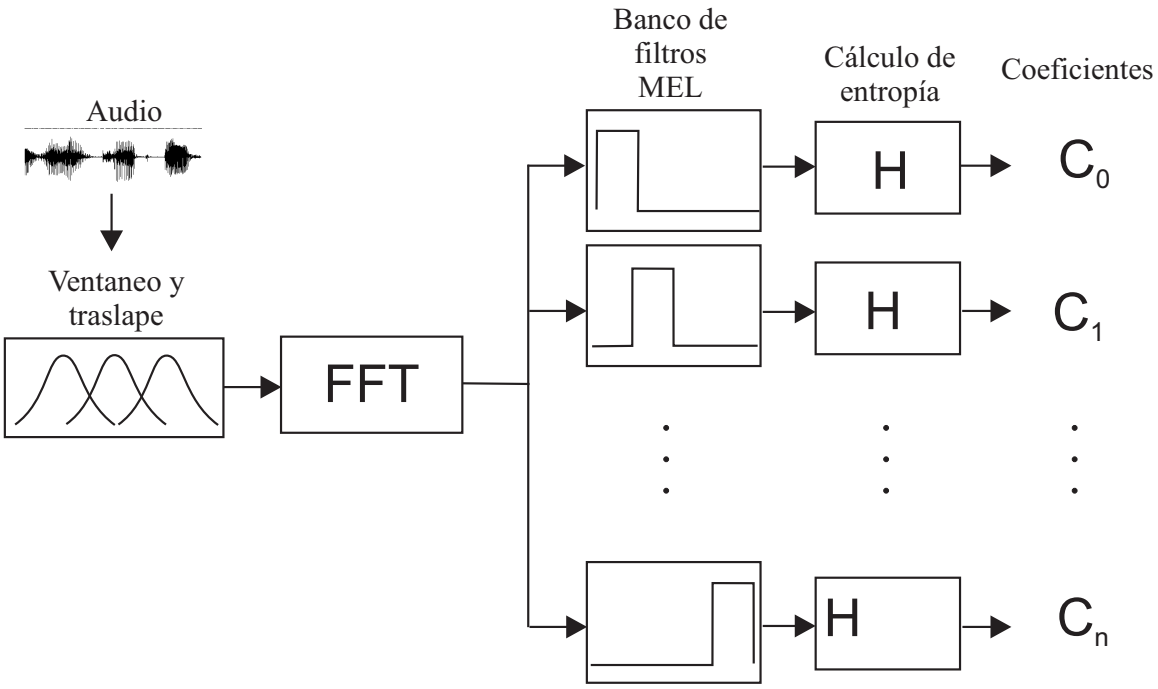


Figura 22: Procedimiento para obtener la firma MEL-MBSES.

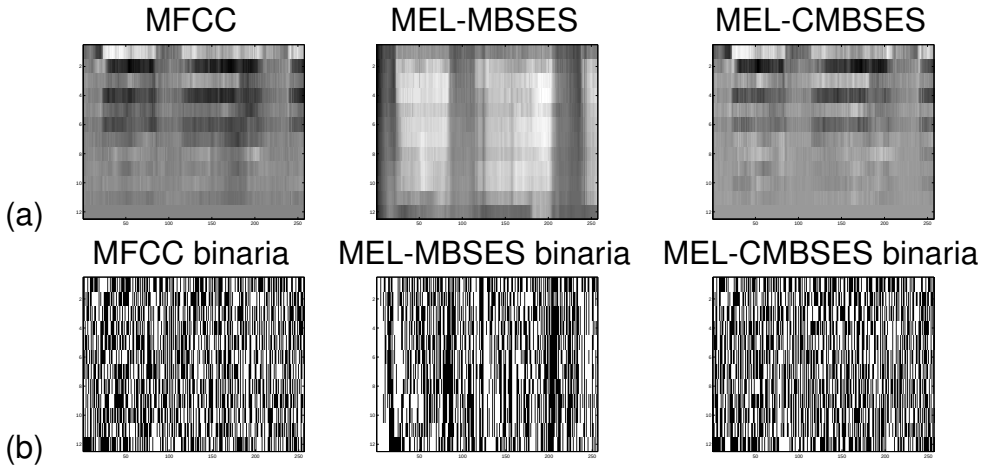


Figura 23: MFCC, MEL-MBSES and MEL-CMBSES (a), y las versiones binarias B-MFCC, B-MEL-MBSES y MEL-CMBSES del sonido *llanto de bebe*. (b)

(a) Volumen (b) Filtrado pasabajas de 1kHz (c) Ecualización

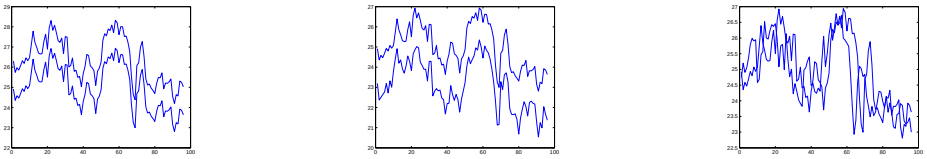


Figura 24: Comparación de la 5ta banda de la firma MEL-MBSES de la canción "Diosa del cobre"(Miguel Bosé y Ana Torroja) contra la 5ta banda de otras versiones de la misma canción.

La firma MEL-MBSES esta basada en la firma MBSES, la cual se obtiene de forma similar a la firma MEL-MBSES pero usando la escala Bark. La MBSES se ha utilizado en recuperación de música siendo robusta ante ruido, ecualización y cambios de volumen (Camarena-Ibarrola, 2008). Se realizó el cambio a la escala MEL, debido que al separar el espectro en bandas usando la escala Bark, las primeras bandas contienen pocas muestras, lo cual degrada el cálculo de la entropía en cada banda. Este problema se soluciona usando la escala MEL, ya que las bandas de frecuencia quedan traslapadas permitiendo más muestras por banda.

El calculo de la firma MEL-MBSES sobre una versión transformada de un sonido (ej. ecualizado, filtrado, cambio de volumen) produce que los valores de la entropía tengan un desplazamiento con respecto a la versión original. En la figura 24 se muestran ejemplos en donde se comparan los valores de entropía en una banda de frecuencia de un sonido contra otras versiones del mismo sonido pero con cambio de volumen, filtrado y ecualización. Debido a que el perfil de los valores de entropía es casi idéntico, es posible utilizar la derivada para capturar el desplazamiento del perfil e ignorarlo. Al calcular la derivada sobre el entropigrama, se obtiene una nueva firma robusta ante las transformaciones de ecualización, filtrado y cambio de volumen. Además, si únicamente se toma en cuenta el signo de la derivada, se consigue que esta nueva firma este codificada de forma binaria lo cual reduce significativamente su tamaño de almacenamiento.

Binary Multiband Spectral Entropy Signature (B-MEL-MBSES)

La firma MEL-MBSES se puede hacer más robusta ante rudio, filtrado y ecualizaión mediante un procedimiento de binarización, ver figura 25. Este proceso elimina el desplazamiento de entropía producido en diferentes versiones de un sonido. Con la binarización

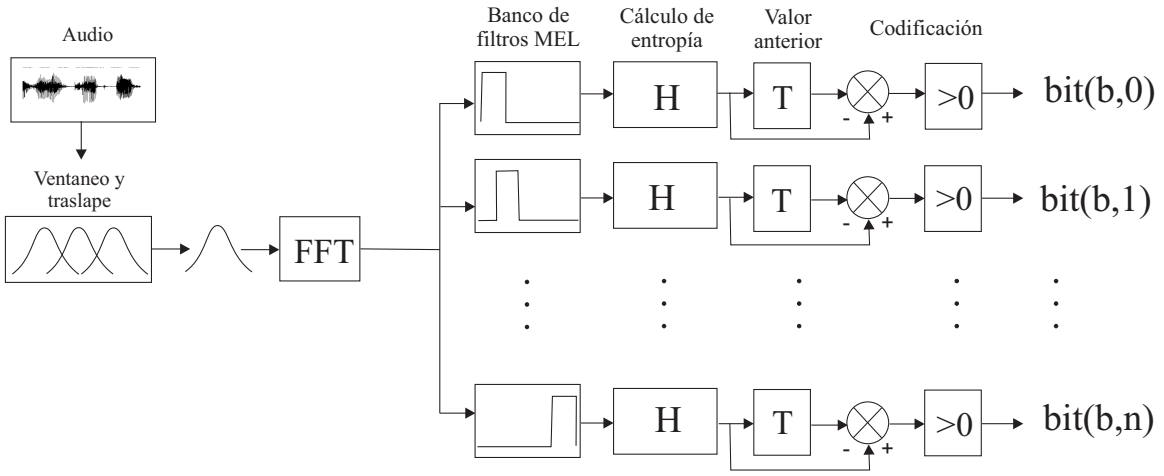


Figura 25: Procedimiento para obtener la firma binaria MEL-MBSES.

se modela el perfil de la firma y se reduce considerablemente el tamaño total de la firma.

La binarización se lleva a cabo tomando el signo de la derivada, como se indica en la ecuación 17, donde el bit correspondiente a la banda b y a la trama n, $bit(n, b)$ se determina con el signo producido al restar $H(n, b)$ y $H(n-1, b)$.

$$bit(n,b) = \begin{cases} 1 & if \quad H(n,b) - H(n,b-1) > 0 \\ 0 & if \quad H(n,b) - H(n,b-1) \leq 0 \end{cases} \tag{17}$$

Una vez realizada la binarización, el sonido queda representado como una cadena de bits cuya longitud depende de la duración del sonido. Debido a que la representación esta en forma binaria, se puede utilizar la distancia de Hamming para comparar entre firmas de distintos sonidos. La figura 23.b muestra una version de la firma del sonido de un bebé llorando.

Cosine Multiband Spectral Entropy Signature (MEL-CMBSES)

Para obtener la firma MEL-CMBSES, se calcula la firma MEL-MBSES como se describe arriba y posteriormente se obtiene la transformada del coseno sobre los 12 valores de entropía en cada trama, ver figura 26. Este paso es similar al último paso de MFCC pero en vez de utilizar el logaritmo de la energía, se utiliza la entropía. Teóricamente, este paso suprime el desplazamiento provocado al obtener la firma en diferentes versiones de un mismo sonido. En la figura 23.a se muestra un ejemplo de la firma MEL-CMBSES sobre

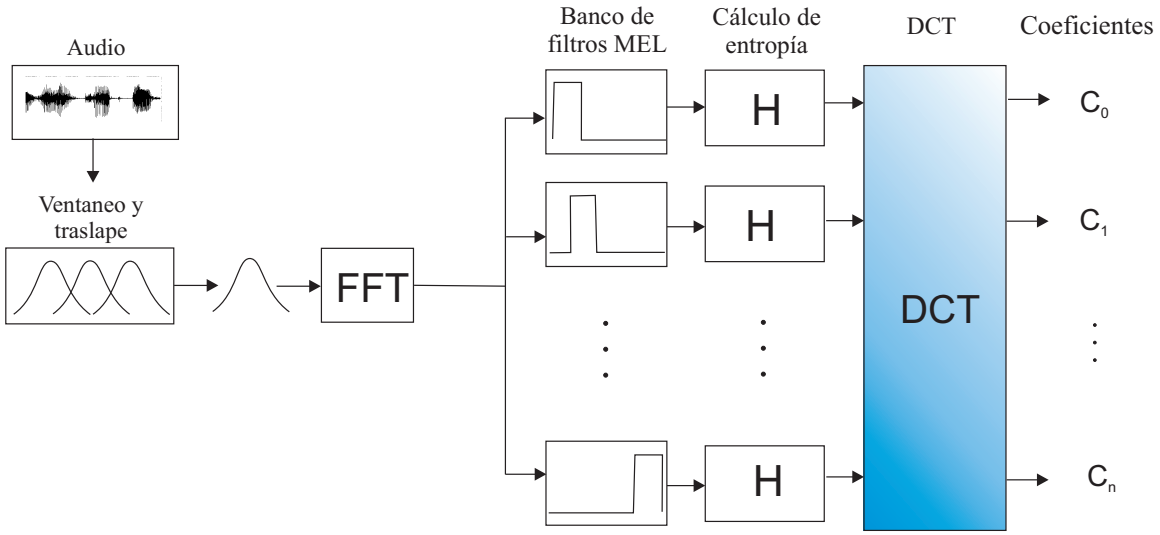


Figura 26: Procedimiento para obtener la firma MEL-CMBSES.

un segmento de sonido.

B-MFCC y B-MEL-CMBSES

Con el objetivo de conocer el comportamiento de las firmas MFCC y MEL-CMBSES al aplicar la binarización, se aplica el mismo procedimiento para producir las firmas B-MFCC y B-MEL-CMBSES. Otra ventaja de las firmas binarias es la reducción del espacio utilizado. En la figura 23.b se muestra un ejemplo de las firmas B-MFCC y B-MEL-CMBSES sobre un segmento de sonido.

En la sección de experimentos se describe el uso y la evaluación de las cinco firmas propuestas en la primera fase de diseño. También se describen las debilidades de estas firmas para reconocer sonidos ambientales heterogéneos y se detectan las necesidades para el diseño de una nueva firma la cual se diseña en una segunda fase.

4.2. Diseño de firma de audio para sonidos ambientales mezclados, heterogéneos y con problemas de alineamiento

La siguiente firma de audio que se propone en esta tesis, surge de la necesidad de obtener una firma que sea robusta ante desfasamientos, ruido y heterogenidad, mientras es capaz de distinguir los sonidos incluidos en una mezcla. Tomando en cuenta estas consideraciones de diseño, se propone una firma de audio que consiste en un vector sucinto formado por histogramas. Con esta firma se puede representar un sonido de cualquier

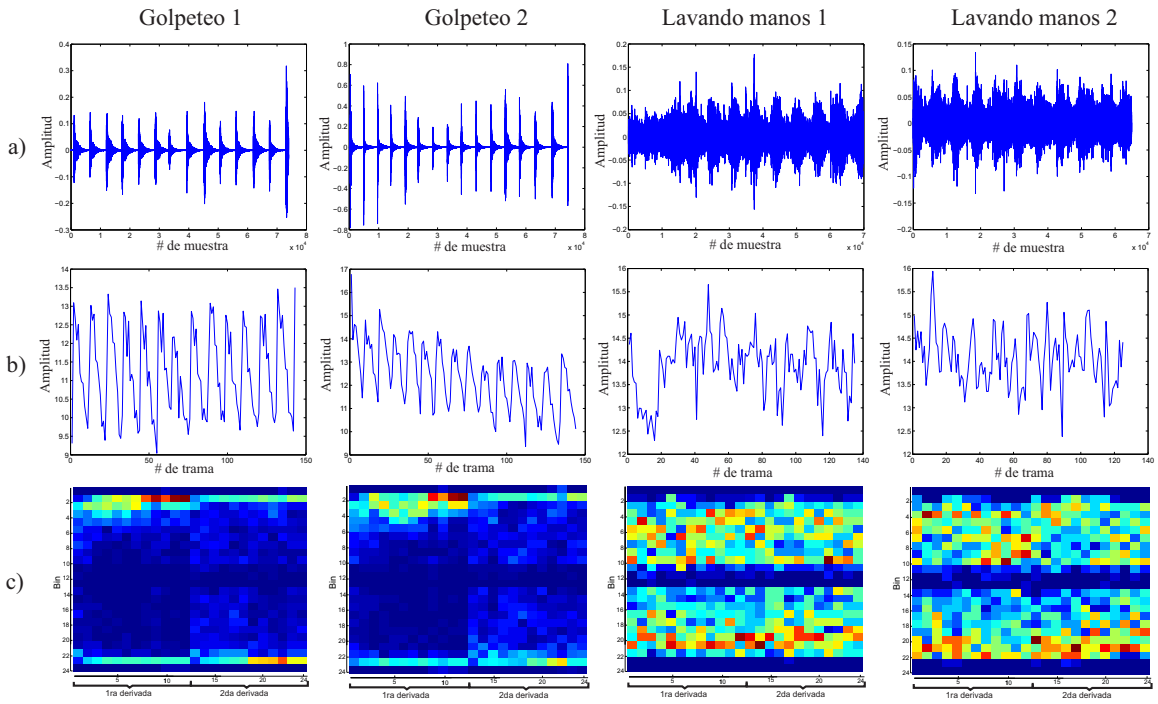


Figura 27: a Amplitudes de dos ejemplos de sonidos de golpeteos y lavado de manos. b Derivadas de la primera banda de la firma MEL-MBSES de los ejemplos de sonidos. c Firmas H1dH2d-MEL-MBSES de los ejemplos de sonidos.

duración en un vector de tamaño fijo el cual se puede evaluar rápidamente mediante un clasificador SVM que requiere entradas de tamaño fijo. Los resultados obtenidos proporcionan mejores resultados que la técnica tradicional usada en el reconocimiento de sonidos ambientales; MFCC-HMM.

La motivación para usar histogramas se debe a que capturan la distribución de cada banda de frecuencia sobre todo el sonido. En la figura 27.c, se ilustra un ejemplo de la firma propuesta aplicada en dos muestras de dos clases de sonidos (golpeteo y lavando las manos). Como se observa, la firma de audio permite distinguir entre sonidos de distintas clases ya que mantiene una estructura similar para los sonidos que pertenecen a una misma clase. Sin embargo, se observa que existe heterogeneidad incluso en sonidos provenientes de una misma clase. Estas variaciones están dadas por las diferencias entre amplitudes, frecuencias y desplazamientos temporales producidos por la forma intrínseca de cada sonido. En la figura 27.a se muestran las amplitudes en el tiempo de las dos muestras de ambas clases de sonidos para ejemplificar las diferencias mencionadas.

El procedimiento para obtener la firma propuesta, denominada H1dH2d-MEL-MBSES,

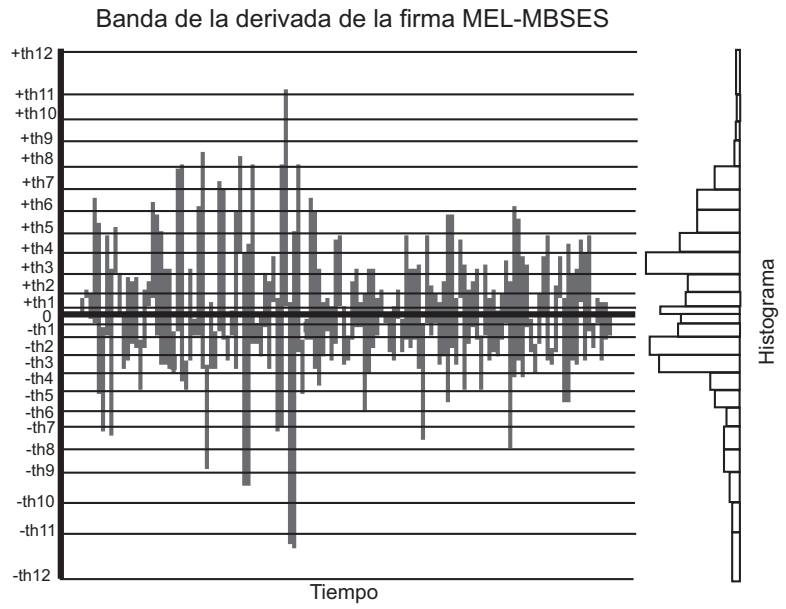


Figura 28: Ejemplo de la obtención de un histograma de una banda de la derivada de la firma MEL-MBSES de un sonidos.

incluye como primer paso la obtención de la firma MEL-MBSES como en la sección anterior. El siguiente paso consiste en calcular derivadas numéricas sobre cada una de las bandas de la firma MEL-MBSES. Con los valores de las derivadas se genera un histograma por cada banda que representa la distribución de los valores de la derivada a lo largo de toda la duración del sonido. En el trabajo (Camarena-Ibarrola, 2008) se mostró que las derivadas de las firmas de audio funcionan mejor ante condiciones de ruido y cambios de volumen en los sonidos. En la figura 27.b se observa que las derivadas proveen información importante de los sonidos, por ejemplo, las derivadas de los sonidos golpeteo y lavando las manos tienen distintos valores. La figura 28 ilustra el procedimiento para obtener el histograma en una banda.

El último paso consiste en calcular segundas derivadas numéricas en cada una de las bandas de la firma MEL-MBSES para nuevamente generar histogramas por cada banda. Estos histogramas representan la distribución de los valores de la segunda derivada a lo largo de toda la duración del sonido. La firma resultante, ilustrada en la figura 27.c se conforma con todos los histogramas de las primeras y segundas derivadas de todas las bandas. Si la firma MEL-MBSES se obtiene con 12 bandas de frecuencia, entonces la firma resultante consta de 24 histogramas.

Para generar los histogramas, es necesario establecer fronteras que separen cada intervalo. Debido a que la mayoría de los valores de la derivada son valores pequeños, utilizamos una técnica adaptativa para encontrar los niveles de cuantización. Esta técnica consiste en encontrar un conjunto de cuantiles a partir de los datos de entrenamiento. Primero se calculan las firmas de audio en todos los sonidos de entrenamiento. Posteriormente se separan los datos positivos y negativos y se obtienen los *cuantiles* sobre cada subconjunto. Los valores obtenidos con los cuantiles, más $\pm$ infinito conforman un total de 24 intervalos.

Una vez que se tienen definidos los limites de los intervalos, se procede a construir los histogramas de la siguiente manera:

$$H_{1D}(b, \mathbb{B}_j) = \frac{1}{N} \sum_{n=1}^{N} 1_{\mathbb{B}_j}(f_b^{'}) \tag{18}$$

$$H_{2D}(b, \mathbb{B}_j) = \frac{1}{N} \sum_{n=1}^{N} 1_{\mathbb{B}_j}(f_b^{''}) \tag{19}$$

Donde $f_b^{'}$ es la primera derivada de la banda b de la firma de audio MEL-MBSES y $f_b^{''}$ es la segunda derivada de la banda b de la firma de audio. N es el número de muestras del segmento de audio. $1_{\mathbb{B}_j}(\cdot)$ es la función indicadora, la cual vale uno si el valor esta dentro del j^{th} bin $\mathbb{B}_j$ y cero en cualquier otro caso. H_{1D} es el histograma obtenido con la primera derivada y H_{2D} es el histograma obtenido con la segunda derivada.

4.2.1. Cálculo de complejidad de la firma H1dH2d-MEL-MBSES

La firma propuesta H1dH2d-MEL-MBSES tiene, además de su desempeño en la clasificación, dos ventajas sobre la técnica usada en el estado del arte NMF-MFCC-HMM. Estas ventajas son el espacio que ocupa la firma y el tiempo de procesamiento para su cálculo y evaluación. En la sección de experimentos se describe a detalle el procedimiento para la clasificación usando la firma propuesta y el estado del arte. En esta sección, con el objetivo de resaltar estas ventajas se comparan los tiempos de procesamiento y el espacio requerido. La tabla 2 muestra el número estimado de operaciones que se requie-

ren para calcular la firma MFCC sobre un segundo de audio. Las operaciones incluidas para la obtención de los MFCC comprenden el cálculo de la energía de las magnitudes en las bandas filtradas de frecuencia, la DCT del logaritmo de los valores de la energía y el cálculo de las derivadas. Las operaciones incluidas para la obtención de H1dH2d-MEL-MBSES comprenden el cálculo de la entropía con la ecuación sobre las bandas filtradas de frecuencia, el cálculo de la primera y segunda derivada y la comparación con fuerza bruta con los umbrales para generar los histogramas.

La tabla 2 muestra el número estimado de operaciones para evaluar la firma MFCC de un sonido usando una HMM continua de 3 estados con 16 gaussianas mezcladas por cada una de las 7 clases. Para evaluar un vector de entrada x en modelos de HMM continuos, la función de densidad de probabilidad de un estado j, $b_j(x)$ esta representada por (Rabiner, 1989).

$$b_j(x) = \sum_{m=1}^{M} c_{jm} N(x, \mu_{jm}, \Sigma_{jm}) \qquad 1 \leq j \leq S \tag{20}$$

donde N denota una función de probabilidad multi-dimensional Gaussiana (pdf) con un vector de media μ y matriz de covarianza Σ. Aquí, c_{jm} son los coeficientes para la mezcla m_{th} en el estado j.

Donde la función de probabilidad multi-dimensional Gaussiana esta dada por

$$N_x = \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} exp\left[-\frac{1}{2}(o_t - \mu_i)\Sigma_i{}^{-1}(o_t - \mu_i)^t\right] \tag{21}$$

La multiplicacion matricial entre cada trama de dimensión 24 con la matriz de covarianza de dimension 24x24 es la operación que requiere más tiempo procesamiento.

En la tabla 2, también se muestra el número de operaciones estimadas para evaluar la firma H1dH2d-MEL-MBSES, que consiste en un vector de dimension 24x24=576, usando SVMs. Para evaluar un SVM lineal, se tiene que la función de decisión esta dada de la forma [Hearst:1998:SVM]:

Tabla 2: Complejidad de operaciones

Firma	No. ops	Clasificador	No. ops	Separación de fuentes	No. ops	Total	No. ops	Tamaño en kilo Bytes
X2 MFCC	7.4M	HMM	33.2M	NMF	43.8M	total	84.4M	8kB
H1d+H2d	7.2M	SVM	0.02M	NO	0	total	7.2M	.9kB
						tasa	11.7	8.8

No. ops =Número de operaciones
M = Millones

$$f(x) = sign(\sum_{i=1}^{l} v_i k(x, x_i) + b) \tag{22}$$

donde x es el vector a ser evaluado, x_i son los vectores de soporte, v_i son los pesos para los vectores de soporte y k es el kernel utilizado. En esta tesis usamos libsvm, la cual entrena $K(K-1)/2$ clasificadores binarios para llevar a cabo una clasificación K-multiclase one-vs-one. Esto implica que para clasificar un sonido de entrada, este se debe comparar con $K(K-1)/2$ clasificadores binarios mediante la operación producto interno entre el vector de entrada y cada hiperplano.

Adicionalmente, se muestran los cálculos requeridos para separar las fuentes usando el algoritmo NMF. Los cálculos sobre la firma MFCC y HMM se consideran sobre ambas fuentes separadas. La factorizacion NMF es un problema NP. Sin embargo, su solución se puede aproximar con un algoritmo polinomial.

Para finalizar, se muestra el espacio requerido por cada una de las firmas, que como se puede observar es 8.8 veces menor que usando MFCC.

Capítulo 5. Evaluaciones

Se hicieron dos fases de experimentos para evaluar la capacidad de las firmas para tratar con los retos del reconocimiento de sonidos ambientales. En ambas fases se compara con el estado de arte.

- El objetivo de la primera fase es comparar el desempeño de las firmas de audio descritas en la sección de métodos para identificar sonidos individuales cuando están traslapados con otros sonidos. Para ello, se usan 9 sonidos ambientales y se obtienen mezclas de tripletas de estos sonidos usando distintos valores de SNR.
- En la segunda fase, se busca evaluar la capacidad de la firma H1dH2d-MEL-MBSES para representar sonidos heterogéneos de diversas clases y con diferente tamaño. Para ello, se experimenta con la clasificación de muestras de 7 clases de sonidos ambientales.

5.1. Identificación de sonidos individuales en segmentos mezclados

En el experimento de la primera fase se compararon las firmas MEL-MBSES, MEL-CMBSES, B-MFCC, B-MEL-MBSES y B-MEL-CMBSES. El experimento consiste en identificar las ocurrencias de un sonido en una base de datos la cual se describe a continuación.

Base de datos

Se descargaron nueve segmentos de sonidos ambientales (llanto de bebé, llaves, sirena de policía, canto de ave, lavado de dientes, música con voz, música sin voz, voz de hombre, voz de mujer) de la página `http://www.freesound.org` con 44100 Hz de frecuencia de muestro y 16 bits de profundidad. Todos los sonidos se recortaron manualmente a una duración de tres segundos. Se generaron mezclas formadas de tres sonidos a partir de combinaciones de los nueve segmentos originales. Antes de generar las mezclas de tres sonidos, se llevo a cabo un paso preliminar donde se generaron mezclas usando solo dos sonidos. En estas mezclas de pares, los dos sonidos incluídos se escuchan con un

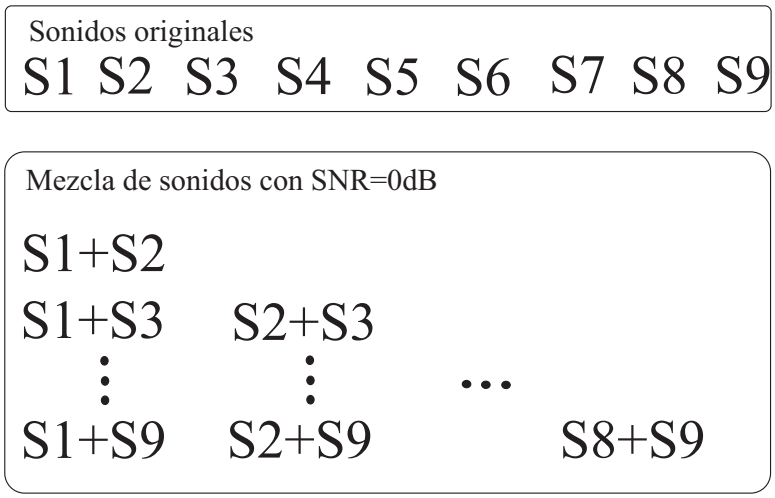


Figura 29: Combinaciones de todas las mezclas con pares de sonidos con un SNR=0dB.

mismo volumen. Esta característica se consigue utilizando un SNR de 0dB que implica que ambos sonidos contribuyen con potencias iguales. Le llamamos “mezcla_A”, (ver figura 29) a todas las combinaciones posibles de pares de sonidos a partir de los nueve sonidos originales.

Para generar las mezclas con tres sonidos, se hacen todas las combinaciones posibles entre los 9 sonidos originales y los elementos de la “mezcla_A”. Sin embargo, para generar estas mezclas, ahora se toma un sonido dominante, el cual se percibe con mayor volumen, y un sonido débil. Esto se consigue usando un SNR mayor de 0dB que implica que el sonido dominante contribuye con una potencia mayor. El sonido dominante es uno de los nueve sonidos originales mientras el sonido débil proviene de “mezcla_A”. Además, se evita que se repita un mismo sonido dentro de la mezcla, ver figura 30.a. Con el propósito de explorar la capacidad de reconocimiento sobre sonidos mezclados, se generan mezclas usando 4 diferentes valores de SNR. Denominamos base de datos de 3.4dB a todas las mezclas generadas usando 3.4dB, y de la misma forma nombramos a las base de datos de 5dB, 10dB y 20dB, ver figura 30.b. Cada base de datos contiene 254 elementos. En cada base de datos, cada uno de los 9 sonidos originales se encuentra en 84 mezclas ya sea como sonido dominante o como sonido débil.

En la figura 31 se aprecian ejemplos de las firmas MFCC, MEL-MBSES, MEL-CMBSES, B-MFCC, B-MEL-MBSES y B-MEL-CMBSES del sonido llanto de bebé. En 31.(b)(e) se muestran las firmas para una mezcla formada de tres sonidos con un SNR=20dB y

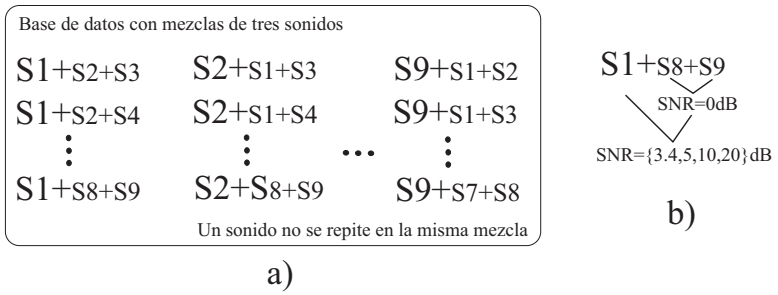


Figura 30: Mezclas de tripletas de sonidos.

en 31.(c)(f) se muestran las firmas para una mezcla formada de tres sonidos con un SNR=3.4dB. El sonido llanto de bebé es la señal dominante mientras que los sonidos canto de ave y música instrumental son la señal débil. Como se observa, la estructura de las firmas de las mezclas con SNR=20dB se asemeja más a la firma del sonido dominante que las firmas con el SNR=3dB.

Para tener un punto de comparación, se pidió a 48 participantes con edades entre 21 y 30 años escuchar los mismos sonidos usados en el experimento. Esto se realizó a través de una página web (http://sound.natix.org) en donde cada participante debía escuchar con audifonos 21 mezclas de sonidos. Las mezclas fueron asignadas aleatoriamente a partir todas las cuatro base de datos (3.4dB, 5dB,10dB, 20dB). A un mismo participante no se le asignaron mezclas repetidas aunque tuvieran SNR diferente. Los participantes primero escucharon por separado cada uno de los sonidos originales y a continuación se les presentaron sus 21 mezclas correspondientes. Se les pidió indicar mediante botones *check* aquellos sonidos originales que creían escuchar en cada mezcla. A ninguno de los participantes se les informó la cantidad de sonidos incluídos en las mezclas y tenían la oportunidad de escuchar cada sonido las veces que necesitaran. En promedio, cada participante escucho 2.52 veces cada mezcla con SNR=3.4dB, 2.47 veces con SNR=5dB, 2.61 con SNR=20dB y 3.53 veces con SNR=20dB.

Experimento

El experimento consiste en usar cada mezcla como consulta y realizar una búsqueda para determinar cuales de los 9 sonidos originales conforman dicha mezcla. Las búsquedas se repiten para cada una de las 4 bases de datos (3.4dB, 5dB, 10dB y 20dB), y para cada una de las 6 firmas mostradas en la sección anterior. En cada búsqueda, se obtie-

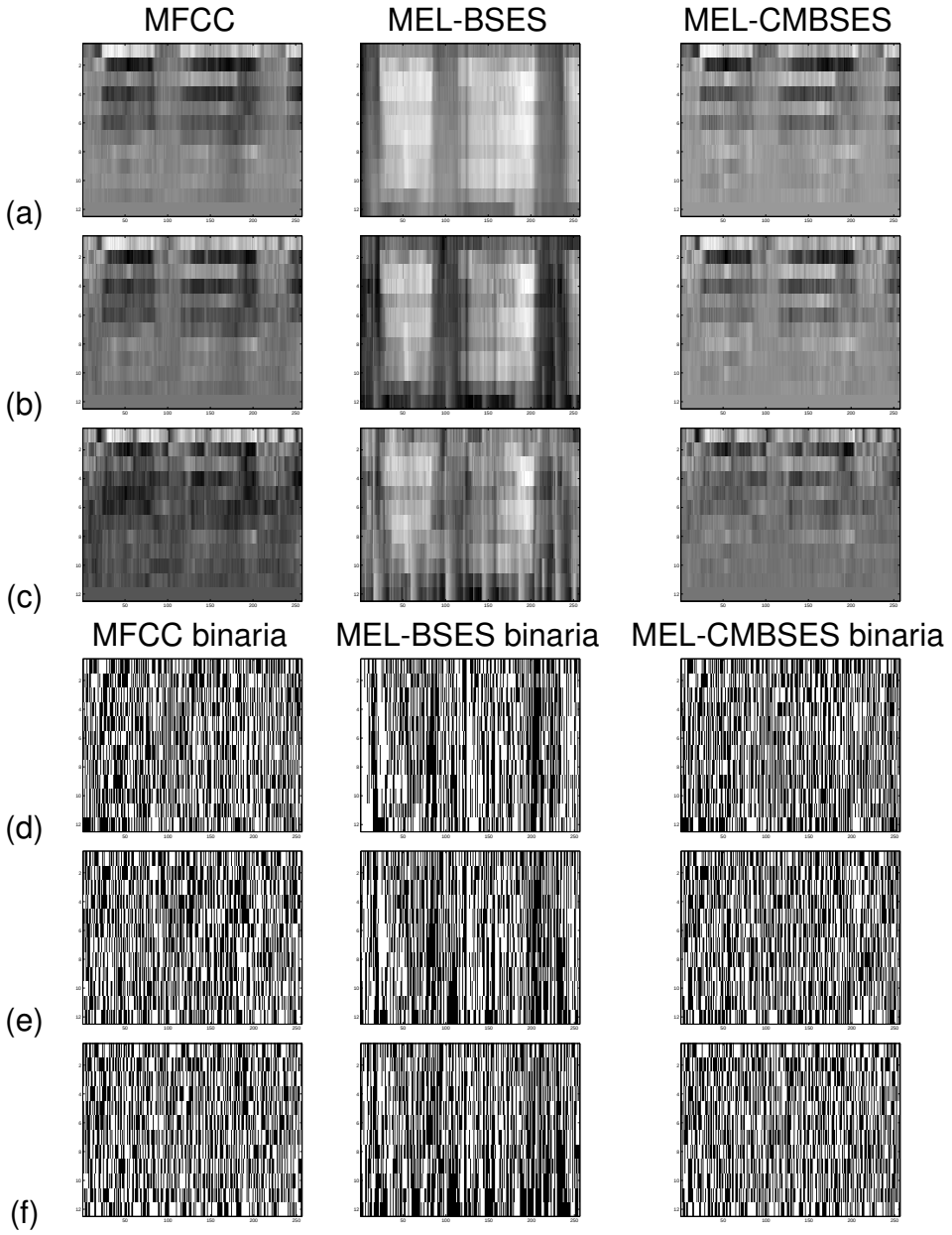


Figura 31: Firmas MFCC, MEL-MBSES and MEL-CMBSES (a), y sus versiones binarias del sonido *Llanto de bebé* (d), y las mezclas de sonido que incluyen *Llanto de bebé* con SNR = 20dB (b)(e) y SNR = 3.4dB (c)(f).

ne la distancia de la mezcla con cada uno de los 9 sonidos originales y se ordenan de menor a mayor. Una búsqueda exitosa es aquella en la que los sonidos que conforman la mezcla resultan los tres sonidos con menor distancia. Para evaluar las firmas se utiliza la exhaustividad dado que ya se conoce que cada sonido esta presente en 84 mezclas. En la Tabla 3 se muestran los resultados para cada una de las bases de datos usando las firmas propuestas y además se muestran los resultados obtenidos cuando los participantes escucharon los sonidos en la página web. Las distancias utilizadas en este experimento son: la distancia Euclideana para los MFCC y MEL-CMBSES, la distancia Euclideana con desplazamiento para la firma MEL-MBSES y la distancia de Hamming para las firmas binarias.

Tabla 3: Resultados para la primera fase de diseño. Las líneas representan los sonidos Llanto de bebé (i), canto de ave(ii), llaves (iii), sirena (iv), lavado de dientes (v), música con voz(vi), música sin voz (vii), voz de hombre (viii), voz de mujer (ix). Las columnas corresponden a las firmas utilizadas (todas con un mismo clasificador) MFCC (a), MEL-MBSES (b),MEL-CMBSES (c), MFCC binario (d) MEL-MBSES binario (e) MEL-CMBSES binario (f). La columna (g) es el promedio de los resultados obtenidos con los 48 participantes. Las cuatro tablas corresponden a las bases de datos con un valor de SNR diferente.

	SNR=3.4dB						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)
(i)	28.57	11.90	23.81	98.81	**100.00**	**100.00**	96.42
(ii)	47.62	57.14	52.38	**100.00**	**100.00**	**100.00**	96.42
(iii)	29.76	84.52	29.76	**100.00**	**100.00**	**100.00**	95.23
(iv)	9.52	**100.00**	5.95	96.43	94.05	**100.00**	90.47
(v)	97.62	**100.00**	92.86	**100.00**	**100.00**	**100.00**	97.61
(vi)	95.24	89.29	96.43	**100.00**	**100.00**	**100.00**	92.85
(vii)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	94.04
(viii)	20.24	5.95	17.86	98.81	**100.00**	96.43	97.61
(ix)	96.43	58.33	84.52	**100.00**	97.62	**100.00**	97.61
Promedio	58.33	67.46	55.95	99.34	99.07	**99.60**	95.37

	SNR=5dB						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)
(i)	32.14	11.90	29.76	97.62	**100.00**	97.62	98.08
(ii)	45.24	57.14	44.05	**100.00**	**100.00**	**100.00**	97.61
(iii)	29.76	84.52	32.14	98.81	98.81	**100.00**	**100**
(iv)	11.90	**100.00**	7.14	91.67	91.67	**100.00**	91.66
(v)	94.05	**100.00**	84.52	**100.00**	**100.00**	**100.00**	98.80
(vi)	92.86	84.52	94.05	**100.00**	**100.00**	**100.00**	90.47
(vii)	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	96.42
(viii)	25.00	7.14	17.86	95.24	**100.00**	94.05	95.23
(ix)	94.05	58.33	80.95	98.81	97.62	**100.00**	96.42
Promedio	58.33	67.06	54.50	98.02	98.68	**99.07**	96.16

	SNR=10 dB						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)
(i)	34.52	27.38	33.33	88.10	89.29	88.10	**95.23**
(ii)	39.29	44.05	36.90	**100.00**	98.81	**100.00**	98.80
(iii)	42.86	79.76	42.86	92.86	91.67	**97.62**	95.23
(iv)	22.62	**100.00**	19.05	83.33	76.19	96.43	73.80
(v)	72.62	98.81	69.05	**100.00**	**100.00**	**100.00**	96.42
(vi)	82.14	82.14	83.33	**100.00**	98.81	**100.00**	94.04
(vii)	**100.00**	96.43	**100.00**	**100.00**	**100.00**	**100.00**	90.47
(viii)	33.33	23.81	26.19	89.29	94.05	86.90	**96.42**
(ix)	80.95	50.00	65.48	90.48	88.10	**97.62**	95.23
Promedio	56.48	66.93	52.91	93.78	92.99	**96.30**	92.85

	SNR=20dB						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)
(i)	34.52	33.33	33.33	69.05	79.76	76.19	**94.04**
(ii)	40.48	39.29	33.33	**97.62**	90.48	**97.62**	89.28
(iii)	41.67	63.10	41.67	91.67	91.67	**95.24**	89.28
(iv)	33.33	**100.00**	33.33	67.86	60.71	64.29	46.42
(v)	57.14	78.57	53.57	**100.00**	**100.00**	**100.00**	86.90
(vi)	82.14	70.24	82.14	**100.00**	88.10	91.67	72.61
(vii)	**100.00**	90.48	**100.00**	98.81	**100.00**	92.86	83.33
(viii)	33.33	33.33	33.33	61.90	**78.57**	67.86	77.38
(ix)	61.90	36.90	58.33	79.76	71.43	79.76	**82.14**
Promedio	53.84	60.58	52.12	**85.19**	84.52	85.05	80.15

Como se observa en la Tabla 3 , los resultados son mejores con valores de SNR bajos. Esto sucede debido a que en las mezclas con 3.4dB todas los sonidos que forman la mezcla se pueden escuchar de forma similar. En el caso de SNR=20dB sucede lo contrario debido a que los sonidos que conforman la señal débil se escuchan muy poco en comparación del sonido dominante. Esto último también se aprecia con el número de veces que los participantes escucharon las mezclas con 20dB (3.53) en comparación con otros valores de SNR.

En esta fase de diseño, solo se usa un sonido por cada clase. Sin embargo, como se aprecia en la sección de método, se observa que diferentes instancias de una misma clase pueden presentar diferencias de estructura en el tiempo y en la frecuencia. Por tal motivo, se diseñó la firma H1dH2d-MEL-MBSES que posee flexibilidad para reconocer las clases de sonidos cuya evaluación se describe en la siguiente sección.

5.2. Reconocimiento de clases de sonidos: Evaluación de escalabilidad y robustez ante alta heterogeneidad y ruido

El experimento de la segunda fase consiste en evaluar la firma H1dH2d-MEL-MBSES ante condiciones de ruido y su capacidad para generalizar su respuesta cuando se usa con diferentes usuarios y dispositivos de grabación. De igual manera que en la fase anterior, se construye una base de datos con las condiciones necesarias para la evaluación.

Base de datos

Para la construcción de las bases de datos utilizadas en esta fase, se grabaron seis clases de sonidos (balón botando, lavado de dientes, grillo, lavado de manos, llaves y tecleando) usando un teléfono móvil LG-E510f. Además, se descargo de www.freesounds.org la clase llanto de bebé. Los sonidos fueron producidos por cuatro participantes donde cada uno contribuyó con 10 muestras para cada clase. Todos los sonidos se grabaron en condiciones ambientales naturales y cada sonido se produjo sin seguir instrucciones específicas. Por ejemplo, cada sujeto se cepillo los dientes como normalmente lo hace en su ambiente especifico y los rebotes de balón se produjeron usando diferentes balones y superficies. Se construyó la *base de datos A* con las 280 muestras capturadas de las 7

clases. Todos los sonidos fueron capturados con una frecuencia de muestreo de 44100Hz y 16 bits de profundidad. La duración de los sonidos es variable entre 1.2 a 5.7 segundos.

Con el propósito de evaluar el desempeño de la firma propuesta ante variaciones intra-clase y evaluar que tan generalizable es a otros usuarios, se crearon dos bases de datos adicionales. Se formó la *base de datos B* al grabar con el mismo teléfono cuatro muestras por cada clase de cuatro participantes adicionales. Además, se creó la *base de datos C* al bajar de internet (www.freesounds.org y www.youtube.com) ocho sonidos por cada clase que provienen de sujetos y ambientes diferentes. Es importante mencionar que los sonidos descargados de internet fueron capturados con diferentes dispositivos de los cuales no tenemos ningún control. De la misma manera, no tenemos control sobre los sujetos que produjeron los sonidos ni los ambientes en donde fueron capturados, lo que produce una base de datos heterogenea.

Para evaluar el desempeño de la firma ante condiciones de ruido y conocer la capacidad de encontrar todos los sonidos que conforman una mezcla, creamos la *base de datos D*. Esta base de datos se generó mezclando dos sonidos de dos clases diferentes. De igual forma que en el experimento anterior, usamos cuatro valores distintos de SNR lo cual produce que los sonidos se perciban con distinto volumen. Por cada clase y por cada valor de SNR se obtuvieron seis mezclas en donde la clase contribuye como sonido dominante y sonidos de otras clases contribuyen como señal débil. Por ejemplo, para la clase *lavado de dientes*, se generaron las seis mezclas : *lavado de dientes-balón botando*, *lavado de dientes-grillo*, *lavado de dientes-lavado de manos*, *lavado de dientes-llaves*, *lavado de dientes-tecleando*,*lavado de dientes-llanto de bebé*. En este ejemplo, el sonido lavado de dientes se escucha dominante con 2dB,5dB,10dB y 20dB mientras que los otros sonidos se escuchan débiles. Cabe mencionar que además ya existe ruido presente en todas las grabaciones debido a que fueron obtenidos sin ningún control sobre el ambiente.

La tabla 4 resume las bases de datos utilizadas en esta fase y las características que evalúan de la firma.

Tabla 4: Resumen de las bases de datos usadas en esta fase

	Base de datos A	Base de datos B	Base de datos C	Base de datos D
No. Elem. x clase A evaluar	280 sonidos limpios	112 generalización a usuarios	56 generalización a usuarios y dispositivos	42 por valor SNR robustez al ruido

Tabla 5: Fscore para los experimentos sobre bases de datos A,B and C

	Base de datos A	Base de datos B	Base de datos C
NMF-MFCC-HMM	**.98**	.79	.26
H1d SVM	.94	.72	.51
H1d + H2d SVM	.96	**.80**	**.56**

Experimento

En el desarrollo del experimento, se comparó la firma propuesta H1dH2d-MEL-MBSES contra NMF-HMM-MFCC. La primera evaluación consistió en llevar a cabo validación cruzada con 10 particiones sobre los elementos de la base de datos A. Para el resto de las evaluaciones se entrenaron modelos usando todos los elementos de la base de datos A para evaluar los el desempeño de las firmas sobre las bases de datos B,C y D.

De la tabla 5 se observa que nuestra propuesta H1dH2d-MEL-MBSES produce resultados similares al método base NMF-HMM-MFCC usando las bases de datos A y B. Además, la tabla muestra resultados al usar la firma H1d-MEL-MBSES, es decir, usando solo las primeras derivadas. Como se puede observar, los resultados mejoran cuando se usan las segundas derivadas. Cabe mencionar que aunque los resultados para las bases de datos A y B son similares, nuestra propuesta requiere menor espacio y es más rápida para evaluar.

Los resultados al evaluar sobre la base de datos C, mostrados en la tabla 5 indican que nuestra propuesta posee mejor generalización a distintos usuarios y dispositivos de grabación, ya que los resultados son significativamente mejores. Recordamos que la base de datos C corresponde a los sonidos descargados de internet.

La evaluación sobre la base de datos D refleja la capacidad de nuestra propuesta para reconocer dos sonidos que forman una mezcla. Aquí cabe mencionar que no tenemos ninguna información a priori sobre los fondos de los sonidos ya que pueden ser

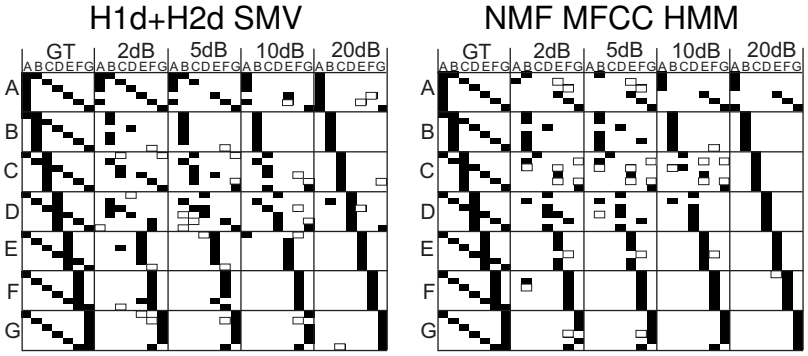


Figura 32: Classificación en mezclas, el eje y indica al sonido dominante en la mezcla mientras en cada columna se indica el otro sonido que forma la mezcla. Las clases de sonidos son: A=balón botando, B=lavándose los dientes, C=grillo, D=llanto, E=llaves, F=lavándose las manos, G= tecleando.

de cualquier clase. La figura 32 muestra la matriz de confusión al clasificar los sonidos mezclados. Los cuadros negros significan que el sonido se identificó correctamente como parte de la mezcla, es decir los VP. Los cuadros blancos significan que que el sonido se clasifico de forma incorrecta como parte de la mezcla, es decir, FP. Las etiquetas en el eje vertical indicas a cual clase de sonido pertenece la señal dominante en la mezcla mientras que las etiquetas en el eje horizontal indican a que clase de sonido pertenece la señal débil. GT se refiere al ground truth. Los valores de SNR usados para formar las mezclas se indican con 2dB, 5dB, 10dB y 20dB. Cada base de datos esta formada por 42 sonidos, donde cada sonido contiene 2 clases, por lo tanto, deberían de existir 84 VP.

En la tabla 6 se muestran los VP, FP y el F1Score al reconocer los sonidos que componen cada mezcla. Como se aprecia, nuestra propuesta resulta mejor que el método base. La firma NMF-MFCC-HMM produce mejores resultados cuando se clasifican sonidos individuales sin ruido, como se muestra en la tabla 5 y en la tabla 6 con un SNR=20dB. Sin embargo, se aprecia de la tabla que es poco probable que NMF-MFCC-HMM encuentre a los dos sonidos que forman la mezcla. Se detectó que incluso al utilizar separación de fuentes con NMF, los dos sonidos separados frecuentemente se clasifican como de la misma clase o de lo contrario existirían más FP.

5.3. Discusión

Los resultados de los experimentos de esta sección indican que es posible utilizar una firma de audio compacta y fácil de calcular para reconocer eventos de sonido. Mediante el primer experimento se identificó que el usar las derivadas de las firmas de audio in-

Tabla 6: Verdaderos Positivos(VP), Falsos Positivos (FP) y F1Score en sonidos mezclados.

Test data D	H1d+H2d SVM			NMF-MFCC-HMM		
	VP	FP	F1Score	VP	FP	F1Score
SNR = 2dB	46	10	**0.65**	39	10	0.59
SNR = 5dB	45	9	**0.65**	38	10	0.57
SNR = 10dB	42	7	**0.63**	40	7	0.61
SNR = 20dB	45	5	**0.67**	43	1	0.67

crementa la eficiencia del reconocimiento para sonidos mezclados mientras que ayuda a disminuir el tamaño de la firma. Además, la respuesta de nuestra propuesta se compara con el desempeño de reconocimiento de sonidos mezclados de los seres humanos.

La evaluación del segundo diseño de firma de audio indica que es posible reconocer sonidos con alta heterogeneidad, con duración distinta y con condiciones de ruido usando una representación compacta y fácil de calcular. Los resultados indican una mejoría en el desempeño comparando con el estado del arte y muestran la viabilidad para utilizar nuestra propuesta en aplicaciones conscientes del contexto con eventos de sonido en condiciones realistas.

En el siguiente capitulo se presentan esquemas de reconocimiento de eventos de sonido para dos casos de uso de aplicaciones para adultos mayores.

Capítulo 6. Evaluación de las firmas en dos casos de estudio

Se presentan evaluaciones en dos casos de estudio relacionados con la asistencia ambiental a adultos mayores.

- En el primer caso, la evaluación consiste en probar la eficiencia sobre datos grabados en un ambiente real de una residencia geriátrica. Aunque en esta evaluación los sonidos son segmentados manualmente, estos se encuentran contaminados con ruido y están traslapados con otros sonidos de forma natural por el ambiente.
- En el segundo caso, se evalúa la capacidad de la firma para clasificar audio continuo que no ha sido segmentado manualmente y es producido en un ambiente real que consiste en un departamento que habita un adulto mayor.

6.1. Caso de estudio 1: Reconocimiento de comportamientos disruptivos audibles en un escenario de una residencia de adultos mayores

Este caso de estudio consiste en reconocer comportamientos disruptivos mediante audio. Los adultos mayores con problemas de demencia y/o Alzheimer, presentan Síntomas de Demencia Psicológicos y de Comportamiento (SDPC). Estos síntomas son trastornos en la percepción que se manifiestan a través de la satisfacción, el humor y el comportamiento (Burns *et al.*, 2012). Los síntomas psicológicos de demencia se relacionan con la ansiedad, depresión y psicosis. Algunos ejemplos de estos síntomas incluyen agresión, apatía, agitación, comportamientos desinhibidos, comportamiento deambulante, alteraciones nocturnas y Comportamientos Disruptivos Audibles (CDA). Aproximadamente el 90 por ciento de los pacientes con demencia del tipo Alzheimer presentan estos problemas de comportamiento haciendo que cuidarlos sea una tarea complicada. Por ejemplo, el CDA, que se manifiesta mediante gritos, maldiciones o con preguntas inapropiadas y repetitivas produce estrés emocional severo a los cuidadores y a otros habitantes de residencias geriátricas. Además, el personal de enfermería de residencias geriátricas expresa más frustración, ansiedad y enojo hacia pacientes que presentan CDA e incluso suelen distanciarse de ellos (Fick *et al.*, 2014).

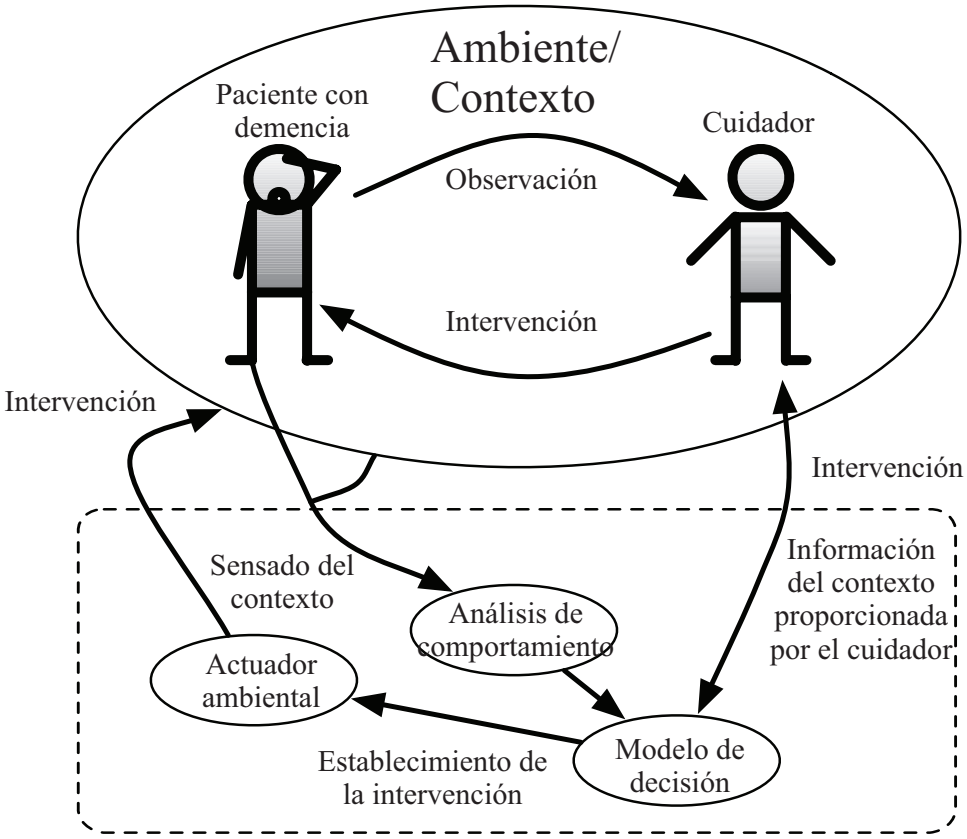


Figura 33: Sistema de Intervenciones Asistidas. El sistema analiza la información del contexto sensada a partir del usuario, la información introducida por el cuidador y el ambiente al cual el usuario se encuentra expuesto. Imagen basada de (Navarro *et al.*, 2014).

En los últimos años, se ha determinado que el tratamiento a la demencia debe incluir Intervenciones No Farmacológicas (INF) cuando se manifiesten comportamientos disruptivos (Sadowsky and Galvin, 2012). Algunos medicamentos, tales como antipsicóticos y antidepresivos proveen evidencia de ayudar con los SDCP, sin embargo sus efectos son modestos y producen efectos secundarios severos como accidentes cerebro-vasculares e incremento en el riesgo de mortalidad, lo cual resalta la importancia de usar alternativas no farmacológicas para disminuir estos síntomas (Jalbert *et al.*, 2008). Mediante el análisis automático del contexto es posible desarrollar un Sistema de Intervenciones Asistidas (SIA) cuyo objetivo sea el sugerir estrategias para tratar comportamientos problemáticos (ver figura 33) (Navarro *et al.*, 2014). Este sistema puede utilizar la información contextual para mejorar la calidad de vida de los pacientes con demencia debido a que identifica automáticamente cuando existe presencia de SDPC y decide una intervención no farmacológica apropiada ya sea modificando el ambiente o persuadiendo al paciente o al cuidador de llevar a cabo una sugerencia.

En esta sección, se evalúa la detección automática de CDA's usando grabaciones de audios recabados en una residencia geriátrica, que es un ambiente naturalista donde habitan personas que producen este tipo de comportamientos disruptivos.

Base de datos

La base de datos para este experimento surgió de un estudio observacional en una residencia geriátrica. Se seleccionaron 5 residentes diagnosticados con demencia (4 mujeres y 1 hombre, con edades entre los 81 y 94 años, promedio=90.2) para participar en el estudio observacional de SDCP. Los criterios de inclusión de los sujetos fueron: 1) deben estar diagnosticados con demencia; 2) estar catalogados por los cuidadores con un puntaje de frecuencia de al menos 2 en algún punto del inventario Neuropsiquiátrico [1] (Cummings *et al.*, 1994) en el mes pasado; 3) Poder conversar en Inglés/Español; y 4) ser capaz de participar independientemente o con alguna asistencia en al menos dos ADL (ej. bañarse, vestirse, alimentarse). Durante 4 semanas, se observó de cerca a los participantes por 5 horas diarias por 3 días cada semana para documentar las manifestaciones de agitación de acuerdo al Inventario de Agitaciones Cohen-Mansfield (IACM) (Cohen-Mansfield and Martin, 2010)(Cohen-Mansfield, 1997). Los comportamientos agitados siempre son socialmente inapropiados y se pueden manifestar de tres formas: a) puede ser abusivo o agresivo hacia si mismo o hacia otras personas, b) puede ser un comportamiento apropiado, pero realizado con una frecuencia inapropiada, por ejemplo preguntar cosas constantemente, y por último c) puede ser inapropiado de acuerdo a estándares sociales para una situación específica, por ejemplo quitarse la ropa en un cuarto público. Se documentaron manualmente los signos que mostraron los pacientes cada vez que producían algún incidente. La observación se realizó en las áreas publicas de la residencia geriátrica. Además, también anotamos las interacciones sociales con los trabajadores y con otros residentes.

De este estudio, se documentaron 181 incidentes de SDCP en 11 de las 29 categorías incluídas en el IACM. La mayoría de los incidentes (81.22 %) tienen una manifestación

[1] El Inventario Neuropsiquiátrico fue desarrollado por Cummings *et al.* (1994), con el fin de recoger información sobre la presencia de síntomas neuropsiquiátricos y conductuales en pacientes con alteraciones cerebrales. Aunque inicialmente se diseñó para ser aplicado en pacientes con Enfermedad de Alzheimer y otras demencias, puede resultar útil para la valoración de cambios conductuales en otras situaciones.

Tabla 7: BPSD con manifestaciones audibles documentadas en el estudio observacional

Descripción	Incidentes	%
Agresión verbal o malas palabras	29	19.73
Preguntas o frases repetitivas	33	22.45
Haciendo ruidos extraños	43	29.25
Quejas	6	4.08
Manerismos repetitivos	36	24.49

auditiva. En la tabla 7 se presenta una categorización de estos incidentes.

Para la grabación de sonidos, los participantes usaron el dispositivo de grabado *Sansa Clip+MP3 Player*. Las sesiones se grabaron con el formato WAV con una frecuencia de muestreo de 24KhZ y 16 bits de profundidad. La Figura 34 muestra dos ejemplos que ilustran a dos residentes usando el dispositivo en la cintura.

Figura 34: Ubicación de los micrófonos en dos adultos mayores que habitan en la residencia geriátrica.

A partir del estudio, se identificaron dos escenarios que ilustran los problemas que se asocian a los CDA's. Los escenarios se narran a continuación como si ya estuviera implementado el sistema SIA resaltando con ello la motivación de realizar un reconocimiento automático de estos comportamientos. Esto además puede servir como guía para futuras sesiones de diseño y para probar prototipos.

Escenario A

Rose es una adulta mayor de 94 años de edad que habita en la residencia geriátrica y esta diagnosticada con la enfermedad de Alzheimer. Sus síntomas de comportamiento y neuropsiquiatricos incluyen explosiones de agresiones expontáneas que provocan estrés

en sus compañeros residentes. Una tarde, Rose esta comiendo en el comedor, cuando su compañera Doris, empieza a gritar y a producir balbuceos. En la figura 35 se ilustra esta interacción. Esto molesta a Rose y provoca que empiece a gritar *Shut up* repetidamente y a golpear la mesa. Los balbuceos de Doris son capturado por los micrófonos del SIA y se detectan automáticamente por el componente de balbuceos. Con la detección, es posible cambiar anticipadamente el fondo musical de acuerdo a las preferencias de Rose para calmar su estado de ánimo. Si el problema persiste, el SIA conduce a Rose a su recamara o a algún lugar más calmado de la residencia para evitar su incomodidad. Las guías se dan mediante estímulos visuales indicándole el camino. Al mismo tiempo, se les informa a los cuidadores sobre la situación y reciben un mensaje cuando Rose llega finalmente a su recamara.

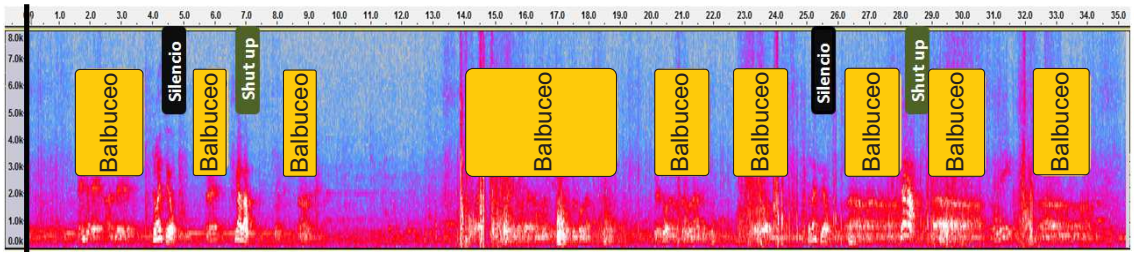


Figura 35: Espectrograma que representa la interacción entre dos pacientes de la residencia geriátrica. Se resaltan las palabras silencio, shut up y el balbuceo.

Escenario B

Frank es un adulto mayor de 90 años que habita en la residencia y está diagnosticado con demencia moderada. Él utiliza una silla de ruedas para poder trasladarse dentro de la residencia. Cada vez que necesita algo, grita *enfermera! enfermera!*. Si no se le atiende rápidamente, entonces empieza a llorar, y esto molesta a otros residentes. El día de hoy, Frank está merodeando la entrada de la residencia geriátrica mientras espera una oportunidad para salirse. Entre más tiempo espera, empieza a agitarse y cuando Frank esta agitado, su comportamiento disruptivo consiste en golpear repetidamente el descansabrazo de la silla de ruedas. Aunque los signos de agitación son manifestados en diferentes formas por cada individuo (Cohen-Mansfield and Martin, 2010) (Navarretta, 2014), este tipo de comportamiento no es raro dentro de los pacientes con demencia que con frecuencia expresan su ansiedad con movimientos repetitivos (Cohen-Mansfield, 1997). Si el sonido que produce Frank, es detectado por los micrófonos y sensores de

ubicación del SIA, entonces se interpretan como un comportamiento deambulante. En respuesta, el sistema enciende la televisión en una área cercana y reproduce el juego de futbol (grabado) favorito de Frank con el objetivo de desviar su atención de la puerta.

Arquitectura del Sistema de Intervenciones Ambientales

Para que el sistema realice las intervenciones adecuadas, este debe ser capaz de detectar automáticamente aquellos sonidos que se asocian con comportamientos disruptivos. El sistema propuesto (ver figura 36), incluye un componente que detecta los CDA's sobre los sonidos que recibe de entrada. El clasificador de CDA's esta ajustado para reconocer tres tipos de sonidos. El primer tipo corresponde a palabras que son repetidas con frecuencia por alguno de los residentes tales como *Shut up* , *silencio* o *nurse*. El segundo tipo de sonidos reconocidos consiste en balbuceos o palabras que no se entienden como las producidas por Doris en el escenario A. El último tipo de sonidos corresponde a sonidos ambientales. Algunos ejemplos de este tipo de sonidos son: golpeteos en la silla de ruedas, como los que produce Frank en el escenario B cuando está agitado, o el sonido del excusado descargando agua. Una vez que se detecta un CDA, el sistema selecciona la intervención apropiada basándose en información de contexto adicional capturada por otros sensores o por conocimiento almacenado en la ontología (Navarro *et al*., 2014).

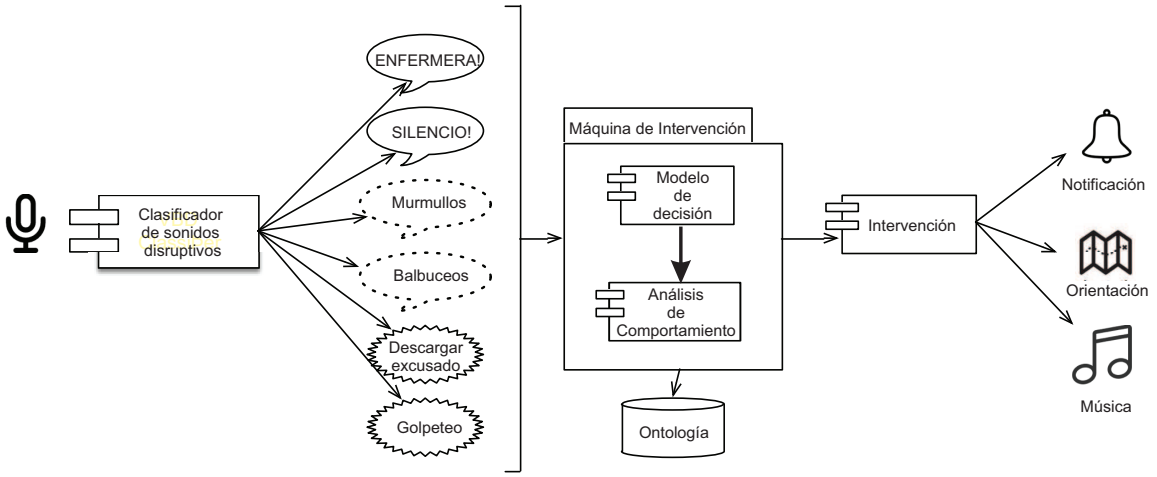


Figura 36: Sistema de Intervenciones Ambientales Asistido basado en la detección de comportambientos disruptivos Audibles y la activación de intervenciones no farmacológicas.

6.1.1. Detector de palabras clave

Para detectar las palabras clave, se construyó un modelo de un clasificador para cada palabra usando muestras de entrenamiento. Las muestras utilizadas provienen del audio capturado en la residencia geriátrica donde habitan pacientes con demencia. Para este problema, se optó por utilizar MFCC con 24 coeficientes como firma de audio y HMM para la clasificación, ya que estos corresponden a la configuración del estado del arte para reconocimiento del habla.

A partir del audio capturado de Rose en el estudio de campo, se encontraron 19 muestras de la palabra *shut up*, 12 muestras de silencio, 9 muestras de c*hicken shit* y 3 muestras de *stop it*. Debido a que se tienen 60 horas registradas de Rose, se estima que ella produce estas palabras claves con una tasa de 0.7 veces por hora. Esta tasa indica que estas expresiones son características de Rose de manera que tiene sentido modelar un clasificador para detectar exclusivamente estas palabras. Esto por un lado simplifica el problema al reducir las posibles palabras, aunque al tratarse de un ambiente realista no controlado existen los problemas de ruido y la heterogeneidad con que ella misma produce las palabras.

En este experimento, se construyeron HMMs continuos para las palabras shut up, y silencio usando los sonidos grabados en la campaña de sensado. La evaluación se realizó usando validación cruzada con partición de tres de donde se obtuvo como resultado un F1Score del 91 %.

Con estos escenarios se identifican manifestaciones audibles de distinta naturaleza. Por un lado se tienen sonidos no verbales tales como los golpeteos. Por otro lado se tienen los sonidos verbales como las palabras y los balbuceos. Todos estos sonidos deben detectarse bajo condiciones ruidosas compuestas de sonidos de fondo no estacionarios. Cada cuarto de la residencia tiene su propio sonido de fondo y adicionalmente las enfermeras y pacientes también producen sonidos aleatorios. Esto implica que el sonido de fondo no puede ser modelado a priori debido a su naturaleza no estacionaria. Este reto en particular se atiende en esta tesis y nuestros resultados muestran que se pueden detectar los sonidos propuestos usando micrófonos de baja calidad y usando menos recursos de

Tabla 8: Matriz de confusión para balbuceos y otros sonidos

	Balbuceos(%)	Otros (%)
Balbuceos	81.24	18.75
Otros	38.28	61.17

cómputo que el estado del arte.

6.1.2. Detector de balbuceos

Para la detección de balbuceos, se optó por utilizar la firma MEL-MBSES con un clasificador HMM ergodico continuo con una gaussiana y 3 estados. El trabajo más parecido en la literatura, hasta donde conocemos, detecta balbuceos de bebés [ref] pero no encontramos nada relacionado con el balbuceo de adultos mayores.

De las 203 horas grabadas en el estudio de campo, se analizaron 113 horas de las cuales se encontraron 198 muestras de balbuceo. Esto indica una tasa de 1.4 balbuceos por hora. Se decidió detectar este CDA automáticamente ya que una intervención adecuada podría disminuir problemas entre los residentes y los cuidadores.

La base de datos para este experimento de reconocimiento de balbuceo se formó segmentando manualmente 43 instancias de balbuceos grabados de Doris durante el estudio de campo. Además, para modelar el resto de los sonidos, recortamos otros 43 ejemplos aleatorios de sonido que también fueron grabados durante el estudio de campo. La duración de los sonidos es variable con una duración promedio 1.75 segundos y una desviación estándar de 0.95. La clasificación se evaluó usando validación cruzada. La matriz de confusión obtenida se muestra en la Tabla 8. Los resultados obtenidos muestran que el balbuceo fue estimado correctamente en promedio 81.24 % sobre todas las validaciones (VP), mientras que fue clasificado incorrectamente un 18.75 % ya que en estos casos fue predicho como la clase otros (FN). Como se observa, los falsos positivos fueron más frecuentes, lo que indica que el algoritmo identificó incorrectamente otros sonidos como balbuceos.

Además de los experimentos sobre los datos segmentados, también se realizó un experimento en donde se clasificaron 6 minutos de segmentos de audio continuo grabado en la residencia. Los segmentos tienen una duración que va desde los 7 hasta los 35

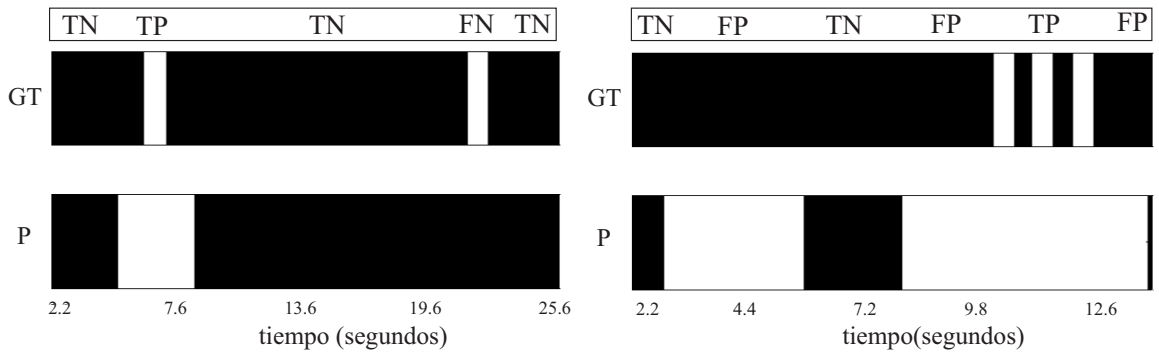


Figura 37: Clasificación de balbuceo en segmentos de audio continuo.

segundos e incluyen sonidos de balbuceo dentro de su contenido. Estos segmentos no fueron segmentados manualmente y están mezclados con otros sonidos de fondo producidos en la residencia.

Se generó un modelo con las 43 instancias segmentadas de balbuceo y las 43 instancias de otros sonidos para entrenar HMMs. Se usaron estos modelos para clasificar los 6 minutos de segmentos continuos. El procedimiento para clasificar los segmentos continuos consiste en tomar ventanas deslizantes de dos segundos del audio de entrada y calcular la MEL-MBSES. Posteriormente, se evalúa la firma en los HMMs para obtener la predicción de los dos segundos. Después se toma otra ventana de dos segundos con un traslape del 80 % con la ventana anterior y esta también se clasifica para obtener la nueva predicción. Se continúa con este procedimiento hasta alcanzar el final del segmento de audio. Se obtuvo un F1Score del 89 %. La figura 37 muestra un ejemplo de los resultados al clasificar dos audios continuos. La etiqueta GT se refiere a Ground Truth y P es la predicción de los HMMs. El color blanco representa los balbuceos y el color negro representa a los otros sonidos de fondo. La figura 37 muestra ejemplos de Verdaderos Positivos (VP), es decir, cuando la predicción indica exitosamente el balbuceos. También se muestran los Falsos positivos (FP), es decir, cuando la predicción indica erróneamente un balbuceo. Adicionalmente se muestran los verdaderos negativos (VN), es decir, cuando la predicción indica correctamente que no hay balbuceo. Para finalizar, se indican los Falsos Negativos (FN), que representan cuando la predicción falla para detectar el balbuceo.

6.1.3. Detector de sonidos ambientales

La manifestación de un comportamiento disruptivo incluye sonidos ambientales tales como golpeteo o descarga del excusado. Para este problema, utilizamos la firma H1DH2S-MEL-MBSES la cual se propone en la sección tal para detectar sonidos ambientales como lavado de manos o cepillados de dientes resultando en una clasificación con buen desempeño.

Debido a que se obtuvieron pocas instancias etiquetadas del sonido de golpeteo en la silla de ruedas, se optó por reproducir el sonido para grabar más instancias para que nuestros resultados sean estadísticamente significativos. Se grabaron 50 sonidos los cuales se segmentaron manualmente. Se calculó la firma H1DH2D-MEL-MBSES para generar un modelo de SVMs lineal con c=0. Además se generaron modelos SVMs para tres diferentes fondos típicos que se producen en la residencia: fondo con voz, fondo con murmullos y música y fondo tranquilo. Para la evaluación se utilizó validación cruzada con una partición de 10 y se obtuvo un F1Score del 93 %.

También se realizó un experimento con 4.6 minutos de porciones de audio continuo grabado en la residencia. Las porciones varían desde 12 hasta 49 segundos de duración e incluyen el sonido de golpeteo en la silla de ruedas. No se realizó una segmentación manual sobre las porciones de audio las cuales también contienen otros sonidos de fondo producidos en la residencia.

Se generaron modelos SVM's con las 50 instancias segmentadas de golpeteo y los tres tipos de fondo. Se usaron estos modelos para clasificar los 4.6 minutos de segmentos continuos. El procedimiento para clasificar los segmentos continuos consiste en tomar ventanas deslizantes de dos segundos del audio de entrada y calcular la H1DH2D-MEL-MBSES. Posteriormente, se evalúa la firma en los SVM's para obtener la predicción de los dos segundos. Después se toma otra ventana de dos segundos con un traslape del 80 % con la ventana anterior y esta también se clasifica para obtener la nueva predicción. Se continúa con este procedimiento hasta alcanzar el final del segmento de audio. Se obtuvo un F1Score del 87 %. La figura 38 muestra un ejemplo de los resultados al clasificar dos audios continuos. La etiqueta GT se refiere a Ground Truth y P es la predicción de

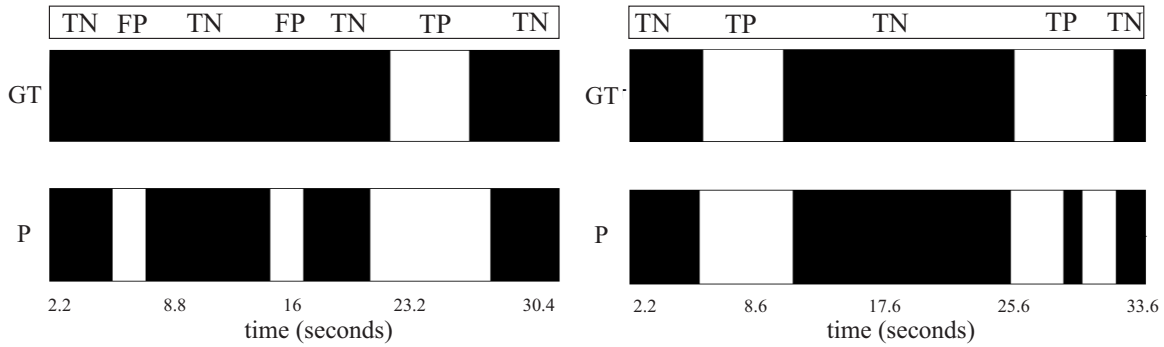


Figura 38: Clasificación de golpeteo en segmentos de audio continuo.

los SVM's. El color blanco representa los golpeteos y el color negro representa a los otros sonidos de fondo. La figura 38 muestra ejemplos de Verdaderos Positivos (VP), es decir, cuando la predicción indica exitosamente el golpeteo. También se muestran los Falsos positivos (FP), es decir, cuando la predicción indica erróneamente un golpeteo. Adicionalmente se muestran los verdaderos negativos (TN), es decir, cuando la predicción indica correctamente que no hay golpeteo. Para finalizar, se indican los Falsos Negativos (FN), que representan cuando la predicción falla para detectar el golpeteo.

6.2. Caso de estudio 2: Reconocimiento de sonidos continuos no segmentados producidos en un ambiente realista en un departamento habitado por un adulto mayor

El objetivo de esta evaluación, es reconocer un número mayor de clases de sonidos ambientales los cuales se producen naturalmente en un departamento habitado por un adulto mayor y otro miembro de su familia. La evaluación se realiza sobre segmentos de audio continuo sin ninguna segmentación manual. Además, se evalúa la capacidad de encontrar los sonidos presentes en una mezcla formada naturalmente (no sintética) en este ambiente.

Base de datos

Se grabaron 20 clases de sonidos en el departamento de un adulto mayor relacionados con alguna actividad y tres tipos de fondo los cuales se presentan en la tabla 9

Para la etapa de entrenamiento, se segmentaron y etiquetaron manualmente 40 segmentos de audio de aproximadamente 3 segundos para cada una de las 23 clases de

Tabla 9: Clases de sonidos para caso de estudio 2

Eventos de Sonidos	Fondos
Lavadora	Fondo de día
Licuadora	Fondo de noche
Aspiradora	Fondo con voz
Lavado de dientes	
Ducha	
Lavado de trastes	
Tosiendo	
Tocando la puerta	
Bicicleta estacionaria	
Descargando excusado	
Agua Hirviendo en tetera	
Tocando el timbre	
Teléfono	
Guisando	
Secadora	
Alarma	
Aplauso	
Tecleando	
Máquina de coser	
Tronando los desos	

sonidos. Para la etapa de evaluación, se grabaron 118 minutos de audio continuo sin segmentar, asegurando que cada clase de sonido apareciera al menos una vez. Los sonidos que forman parte de esta base de datos implican retos en el dominio del tiempo y de la frecuencia. Estos retos se deben a que varias clases tienen forma de onda similar y comparten información en bandas de frecuencia. En la figura 39 se muestran cuatro espectrogramas de los segmentos de audio continuo, donde se observa que existen sonidos que aunque pertenezcan a clases distintas comparten la información en varias bandas de frecuencia, e incluso existen clases de sonidos de actividades que tienen un parecido con la estructura espectral del fondo. Además, en las bases de datos con sonidos mezclados y con segmentos continuos, al menos dos clases están contribuyendo con sus propiedades en el tiempo y en la frecuencia para producir un nuevo sonido traslapado.

6.2.1. Detección de sonidos ambientales continuos

Los experimentos se realizan usando dos configuraciones de firma-clasificador con el propósito de hacer una comparación entre ellas. Las dos configuraciones son: H1dH2d-MEL-MBSES con SVM y NMF-MFCC con HMM. Estas configuraciones se utilizaron en la

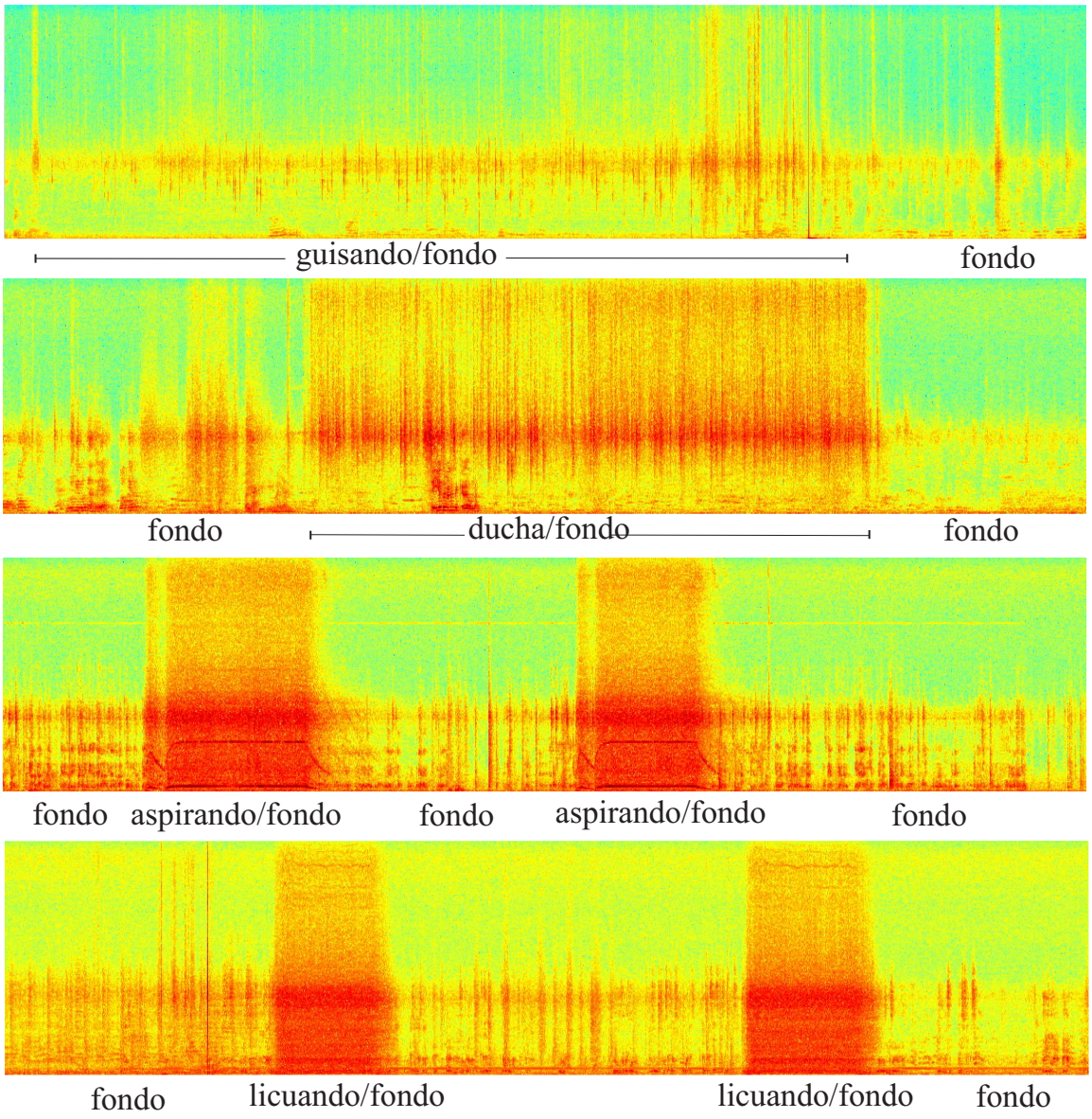


Figura 39: Ejemplo de cuatro clases de eventos de sonidos en segmentos continuos.

segunda sección *Reconocimiento de clases de sonidos*.

Para la evaluación de las firmas en los sonidos continuos, se sigue el siguiente procedimiento. Se utiliza una ventana deslizante con un traslape del 80 % y por cada ventana se extrae una firma, ya sea H1dH2d-MEL-MBSES o NMF-MFCC. Posteriormente, cada firma se clasifica usando SVM o HMM según sea el caso. Cada vez que un sonido se clasifica correctamente se tiene un VP y cada vez que un fondo se clasifica como fondo se tiene un VN. Los sonidos utilizados en este experimento únicamente traslapan alguna de las clases de audio mencionadas con alguno de los fondos, es decir, no se traslapan dos clases de audio provenientes del alguna actividad. Con esta evaluación se obtiene

un F1Score del 41 % cuando se utiliza el método base NMF-MFCC-HMM y 58 % cuando se utiliza la firma propuesta H1dH2d-MEL-MBSES la cual claramente supera al método base.

6.3. Discusión

Los experimentos llevados a cabo en este capítulo indican la viabilidad de usar el reconocimiento automático de eventos de sonidos para utilizarse como herramienta en un sistema dirigido apoyar la vida diaria de adultos mayores y de sus cuidadores.

Con el primer caso de uso se evalúa la firmas de audio y clasificación en sistemas de asistencia ambiental para detectar y responder ante comportamientos disruptivos audibles en condiciones realistas. Se clasifica sobre porciones de audio continuas obteniendo F1Score mayor al 87 %. Estos resultados indican la capacidad de la propuesta para tratar con entradas de audio continuo en ambientes ruidosos y heterogeneos para detectar comportamientos disruptivos en una residencia geriátrica.

Considerando que el sistema esta diseñado para cuidadores, para nosotros es importante que se les informe acerca de eventos cuando existe seguridad de que haya sucedido y así reducir la carga de trabajo con falsos positivos. Esto permite diseñar el esquema de firma-clasificador ajustando los parámetros para evitar falsos positivos aunque esto implique también la existencia de falsos negativos.

Con el segundo caso de uso se evalúa nuestra propuesta H1dH2d-MEL-MBSES en condiciones ambiciosas y realistas debido a que este experimento se realiza sobre 118 minutos de porciones de audio continuo con 23 clases de sonidos capturados en condiciones ruidosas. El escenario de captura corresponde a un departamento donde habita un adulto mayor quien ejecuta sus actividades de la vida diaria. Los resultados, indican la posibilidad de utilizar nuestra propuesta en aplicaciones conscientes de contexto.

Debido a que los sonidos ambientales proveen información de contexto, su reconocimiento automático se puede utilizar en diferentes aplicaciones. En el siguiente capitulo se discuten otras instancias del uso de firmas de audio y clasificadores que han probado ser de utilidad en otros escenarios.

Capítulo 7. Otros aportes de reconocimiento de sonidos

A continuación se describen 3 proyectos que se realizaron en el desarrollo de esta tesis aplicando el conocimiento obtenido de procesamiento y clasificación de sonido. Con estos proyectos se explora el análisis del audio para reconocer diversos elementos de contextos adicionales a eventos de sonidos ambientales. El primer proyecto consistió en la detección automática de llanto de bebés donde se explora el uso de la información del tiempo para incrementar la eficacia de la clasificación. El segundo proyecto consistió de un detector de voz el cual se implementa en un teléfono inteligente. En el tercer proyecto se utilizó el audio de varios micrófonos para determinar si dos personas se encuentran en un mismo lugar y con esto detectar socialización y para optimizar recursos en campañas de sensado.

7.1. Reconocedor de llanto

Con el propósito de entender mejor el lenguaje de los bebés, se puede utilizar una herramienta que almacena los comportamientos de los bebés que suceden antes de una situación crítica de llanto *sostenido*, es decir, llanto fuerte de larga duración. Mediante la observación de estos comportamientos, los padres pueden analizar los movimientos o los sonidos previos a la situación crítica y usar estos datos posteriormente para relacionar estos comportamientos con las necesidades del bebé.

Se trabajó en un proyecto de detección de llanto sostenido de recién nacidos con integrantes del departamento de Ciencias de la Computación de CICESE de donde resultó el artículo de congreso (Rincon *et al.*, 2013).

7.1.1. Experimentos

Para este problema, se capturan segmentos de audio de 30 segundos los cuales se procesan y clasifican para determinar si en cada segmento existe llanto sostenido. El procedimiento para la detección de llanto sostenido se muestra en la figura 40. Se obtienen los coeficientes de la firma MEL-MBSES cada medio segundo y se evalúan en una SVM previamente entrenada para detectar llanto. Si se detecta que existe llanto, se incremente un contador indicando que por medio segundo más se ha detectado llanto. Se repite este

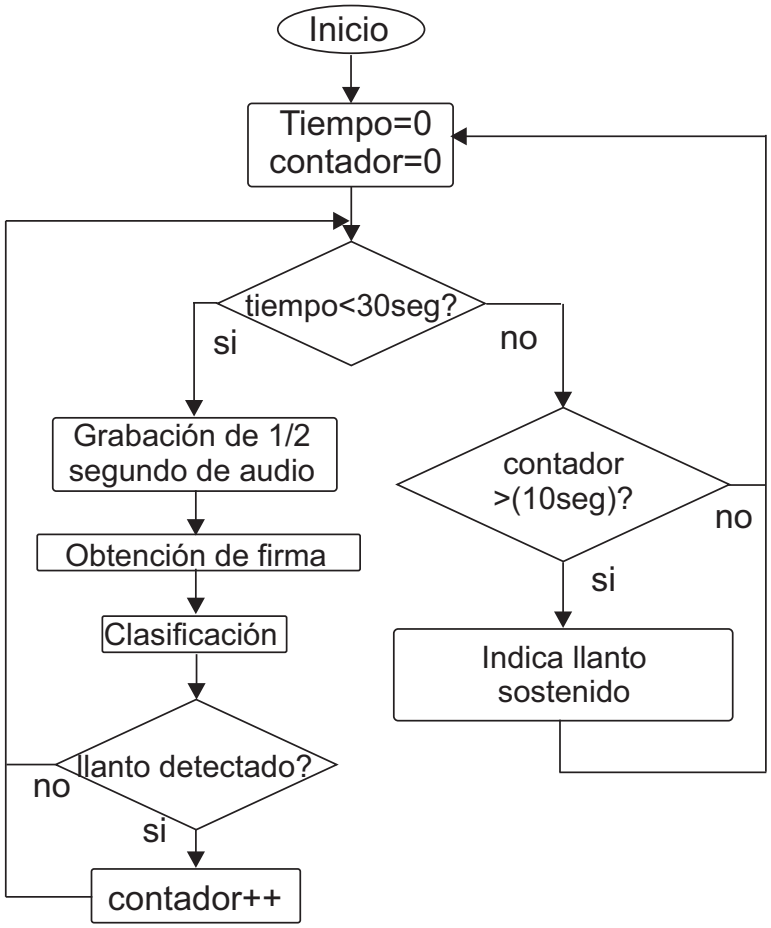


Figura 40: Procedimiento para detectar llanto sostenido.

procedimiento hasta que han pasado 30 segundos. Si el contador indica que existen más de 10 segundos con llanto entonces el segmento se clasifica como llanto sostenido.

Se realizó una comparación entre la firma propuesta y un método base basado en MFCC. Para la clasificación, se entrenaron modelos SVMs para las clases: llanto, silencio y sonidos de fondo. Las SVMs se entrenan y se evalúan usando la biblioteca LibSVM con MATLAB (Chang and Lin, 2015). Los modelos se entrenaron usando kernels RBF con los parametros gamma=2 y C=5 para los MEL-MBES y gamma=2 y C=5 para los MFCC.

7.1.1.1. Base de datos

Para formar la base de datos, se descargaron de Internet 3000 muestras de sonido con una duración de medio segundo. Los sonidos fueron capturados con una frecuencia de muestreo de 44100Hz y 16 bits de profundidad. La base de datos se divide con 1000 sonidos de llantos de bebé, 1000 sonidos de silencio y 1000 sonidos de fondo que no

Tabla 10: Resultados de la clasificación de llanto sostenido.

	MFCC		MELMBSES	
	Precisión (%)	Exhaustividad (%)	Precisión (%)	Exhaustividad (%)
Sonidos de medio segundo	97.99	**98**	**98.01**	**98**
Llanto sostenido	77	70	**100**	**90**

incluyen llanto. Se uso el 85 % de los sonidos de la base de datos para entrenar el clasificador y el 15 % para las evaluaciones. Además, se descargaron de *http://youtube.com* 8 segmentos de audio con una duración entre 30-90 segundos y se etiquetaron cada 30 segundos indicando si existe o no llanto sostenido. Los sonidos en estos segmentos se consideran condiciones realistas ya que provienen de diferentes bebes y de diferentes dispositivos de grabación sin ningún control de las condiciones de grabado.

7.1.1.2. Evaluación

Primero se evaluó el detector de llanto usando los 3000 sonidos descargados de Internet. En la tabla 10 se muestran los resultados usando las firmas MFCC y MEL-MBSES. Posteriormente, se clasificó el detector de llanto sostenido usando los 8 segmentos de audio descargados de Internet. En la tabla 10 se muestran los resultados al comparar los MFCC y MEL-MBSES.

Como se aprecia en la tabla 10, la firma MEL-MBSES produce resultados considerablemente mejor que los MFCC en escenarios ruidosos y realistas tales como los segmentos bajados de internet con llanto sostenido. Un 100 % de precisión indica que no se cometió ningún error al clasificar un segmento como llanto sostenido. Sin embargo, la exhaustividad de 90 % indica que una instancia de llanto sostenido no fue identificada.

Dependiendo de las necesidades de la aplicación, es posible ajustar los parámetros de forma que se modifiquen tanto la precisión como la exhaustividad. Sin embargo, como regla general esto trae consigo que el incremento de uno implique el decremento del otro. Para balancear correctamente estas dos medidas se requiere conocer mas detalles, por ejemplo, es posible que en la aplicación sea crítica la detección de llanto de forma que se prefieran FP a FN.

7.1.2. Discusión

En este trabajo se presenta una comparación de la firma MEL-MBSES contra la firma más utilizada para tratar con voz: MFCC. Los resultados indican que nuestra propuesta supera al método base particularmente en condiciones ruidosas. Con este trabajo incursionamos en el reconocimiento en segmentos continuos al reconocer el llanto sostenido sobre las porciones de audio. También se utilizó un esquema de clasificación que considera otros aspectos además del resultado directo del clasificador, por ejemplo el requerimiento de tener 10 segundos de llanto para considerarse como llanto sostenido.

7.2. Reconocimiento de voz: Implementación del algoritmo en un teléfono inteligente

Se realizó un proyecto con el Centro de Investigación *Create-Net*, en Trento, Italia, que consistió en la implementación de un algoritmo detector de actividad voz (DAV) en un teléfono inteligente de donde resultó el artículo de conferencia (Ferdous *et al.*, 2015). La implementación del DAV se utilizó dentro de un proyecto de monitoreo de personas para detectar estrés en su trabajo. El propósito de la DAV es contabilizar las interacciones verbales para determinar si existe una correlación entre la cantidad de voz detectada con los niveles de estrés percibidos por los propios sujetos de estudio.

El DAV procesa el sonido continuamente en el teléfono sin almacenar ni transmitir el sonido por razones de privacidad. De esta manera, la aplicación solo transmite a un servidor la decisión binaria si el sonido corresponde o no a voz humana. Con esto se permite detectar si el usuario se encuentra participando en una conversación. El rango de captura del micrófono incluido en el teléfono esta limitado a sonidos cercanos, lo que asegura que solo se detecten las conversaciones cercanas al usuario asegurando así su participación.

Para la captura de la información auditiva se utilizó el teléfono Samsung Galaxy 3 mini, que contiene un CPU de 1GHz dual-core Cortez A9, GPU Mali-400 y el sistema operativo Android 4.1.

El algoritmo consiste en la evaluación de dos características de la señal de audio; la

primera es el pitch y la segunda es una versión ligera de la firma MEL-MBSES.

El **pitch** esta relacionado con la medida de la frecuencia fundamental de una señal de voz F0 (Hedelin and Huber, 1990). El pitch se puede utilizar para el reconocimiento de voz, debido a que la frecuencia fundamental de la voz humana se encuentra en el rango de los 40Hz a los 600Hz. En este trabajo, se calcula el pitch usando el algoritmo YIN propuesto en (de Cheveigné and Kawahara, 2002). Este algoritmo utiliza la autocorrelacion para el cálculo del pitch. Se eligió esta implementación ya que provee cualidades importantes para su uso en dispositivos móviles: exactitud, robustez al ruido y eficiencia de procesamiento.

La versión de la firma MEL-MBSES implementada en el teléfono móvil se calcula como se describe en la sección 4.1, pero usando solo 8 bandas de frecuencia en la escala MEL distribuidas desde los 0Hz hasta los 3500Hz. No es necesario utilizar frecuencias mayores ya que la voz humana no rebasa los 3500Hz. Esto a su vez reduce el tamaño de la firma y disminuye el tiempo de procesamiento al usar solo 8 coeficientes.

Para determinar si se trata de voz, se deben de satisfacer dos condiciones: i) el pitch se debe encontrar en el rango de voz humana (40Hz-600Hz), y ii) la evaluación de los 8 coeficientes de una trama de la firma MEL-MBSES en un modelo de SVM, debe indicar que corresponde a voz. Se entrenó la SVM usando muestras de la firma MEL-MBSES sobre tramas de audio provenientes de 3 minutos de sonido de voz y 3 minutos de sonidos de fondo. El calculo de los coeficientes de SVM se obtuvieron usando el algoritmo lineal SMO incluido en la aplicación (of Waikato, 2015).

Se utilizó una frecuencia de muestreo de 44100Hz. El tamaño de ventana se fijó a 1024 muestras. El pitch y la firma MEL-MBSES se calculan para cada ventana. Si se satisfacen las condiciones i y ii en al menos 7 de cada 30 tramas (aproximadamente 0.7 segundos), entonces se determina que existe voz en ese segmento de audio, ver Figura 41.

Se realizó un estudio donde participaron 38 trabajadores que fueron monitoreados durante 6 semanas. Mediante teléfonos inteligentes se capturó la interacción verbal usando el micrófono y se pedía a los usuarios contestar encuestas donde indicaran los niveles

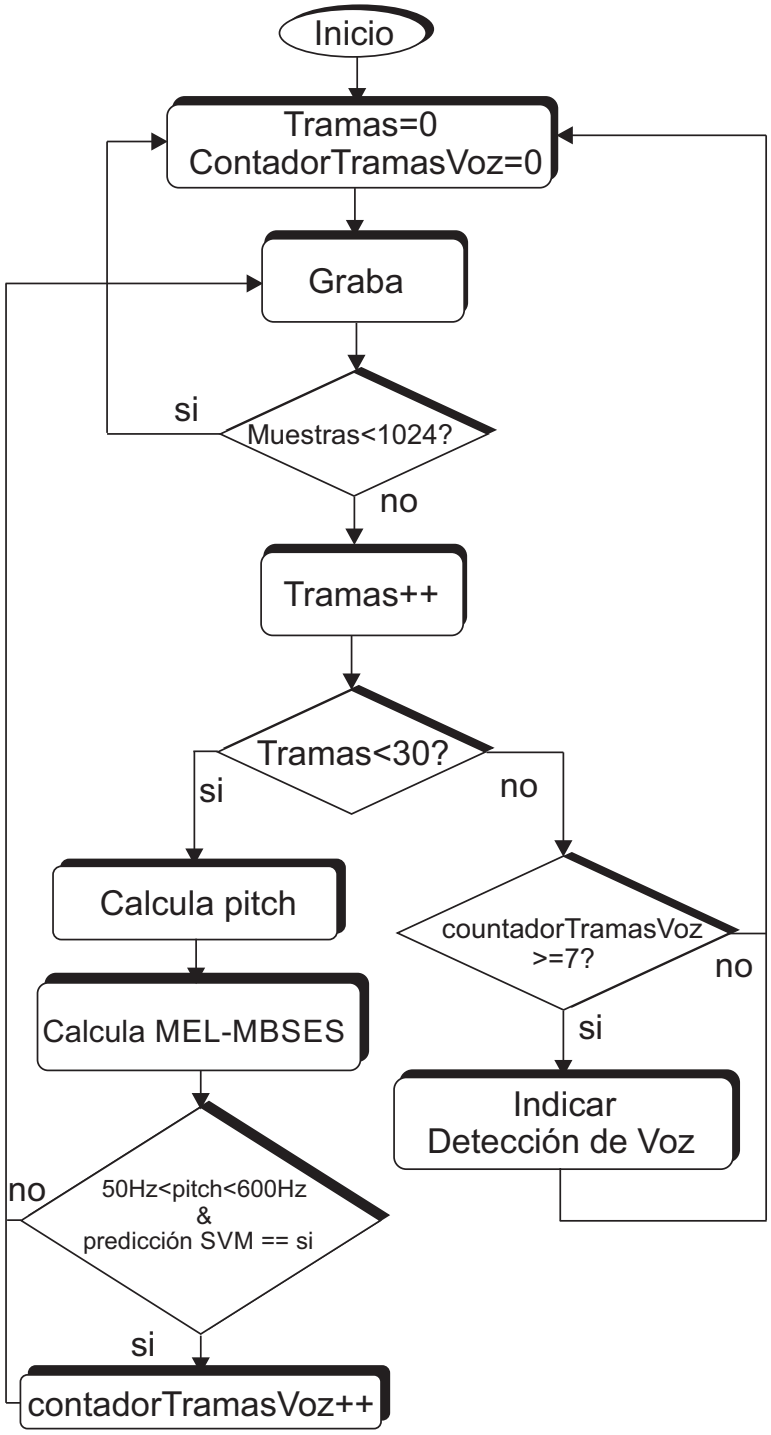


Figura 41: Procedimiento para detectar actividad de voz.

de estrés percibidos. Los resultados indican que no existe correlación entre el nivel de estrés con las interacciones verbales. Sin embargo, existe una correlación positiva del 91.67 % entre la interacción verbal y los niveles de estrés percibidos por aquellos sujetos que reportan altos niveles de estrés (12 sujetos).

7.2.1. Discusión

El desarrollo de este proyecto resultó útil para nuestra investigación debido a que se realizó una implementación de una versión de la firma en un teléfono inteligente. Esto ayudó a conocer las diferencias de programación en PCs y en dispositivos móviles. También ayudo para detectar las limitantes de estos dispositivos móviles, por ejemplo, que la firma se tuvo que programar sin hacer traslape ya que esto imposibilitaba una entrega de respuestas rápida.

7.3. Detección de personas en el mismo lugar

Al detectar si dos personas se encuentran en la misma área, es posible optimizar los recursos computacionales de procesamiento, batería y de memoria en dispositivos móviles usados para sensado. Una forma de determinar si dos personas están en la misma ubicación se consigue al determinar si los dispositivos usados por los usuarios se encuentran grabando el mismo sonido. Se trabajó en un proyecto que se enfoca a determinar si dos adultos mayores viviendo en una residencia geriátrica se encuentran en la misma ubicación de donde surgió el artículo.

En la figura se muestra un ejemplo de la configuración de un sistema para detectar proximidad entre dos sujetos usando información auditiva. El sistema se divide en dos componentes, el primero procesa la información en los dispositivos móviles de cada usuario y el otro componente se encuentra en un servidor local en la residencia geriátrica. En el componente móvil, se captura el audio del ambiente y se calcula una firma de audio denominada *Time Entropy Sigature (TES)* (Ibarrola and Chavez, 2006). A continuación se envían los datos al servidor el cual contiene un *filtro de proximidad* el cual compara las firmas recibidas de los dispositivos móviles y determina si las grabaciones contienen información similar. Finalmente, se tiene un filtro llamado *nodo óptimo* que determina cual dispositivo móvil se mantendrá grabando y envía un aviso a los otros dispositivos para

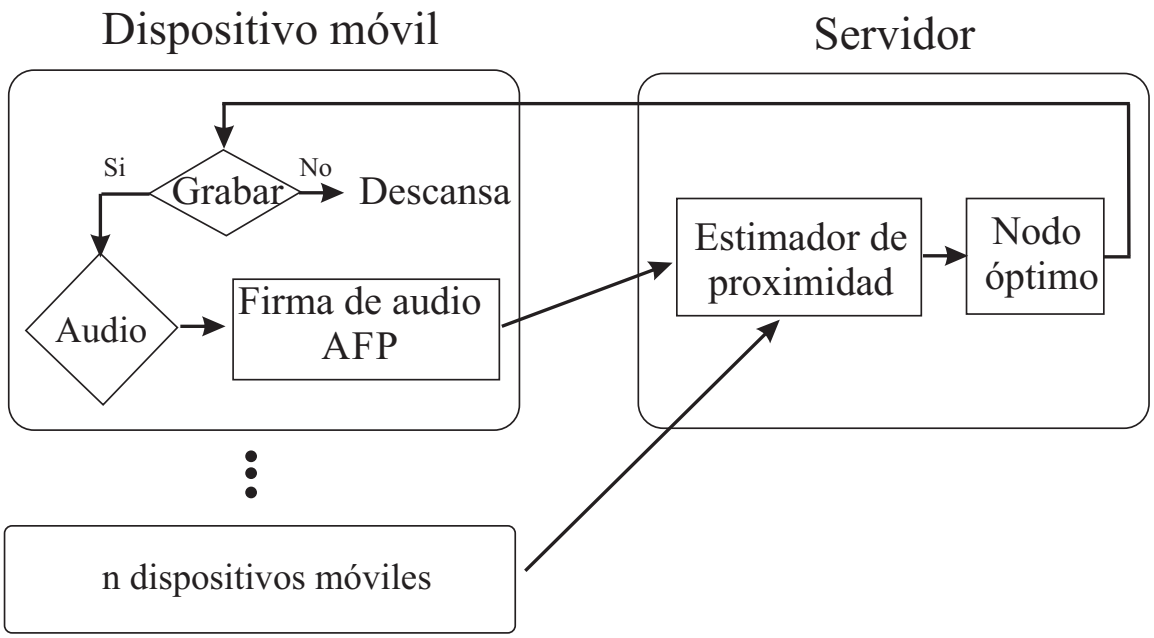


Figura 42: Ejemplo de configuracion de un sistema para detectar proximidad.

que dejen de grabar. De esta forma se establece una estrategia de sensado colaborativo la cual se basa en los recursos utilizados de batería, almacenamiento y transmisión. Por ejemplo, supongamos que se detecta que dos dispositivos se encuentran juntos y uno de ellos tiene poca batería, en ese caso, el *nodo óptimo* indicara que ese teléfono deje de procesar audio por una cantidad determinada de tiempo mientras el trabajo es realizado por el otro dispositivo con mayor disponibilidad de recursos.

La firma de audio *Time Entropy Signature* (TES), usada en este proyecto, proporciona dos características importantes para esta aplicación. La primera es la reducción de los recursos utilizados en la transmisión de los datos al servidor y la segunda es que al obtener distancias entre dos firmas se permite determinar si dos dispositivos están grabando una misma escena.

La TES utilizada en este proyecto consiste en un vector que contiene las entropías de la señal de audio. Se mostró en (Ibarrola and Chavez, 2006) que el obtener la cantidad de información (entropía) de una señal produce una firma de audio robusta ante cambios de energía de la señal, filtrados, compresión y pequeñas cantidades de ruido blanco. El procedimiento detallado para obtener la TES se puede encontrar en (Ibarrola and Chavez, 2006). Brevemente explicado, el procedimiento para obtener la TES consiste en calcular la

entropía sobre tramas de 370ms con un traslape del 87.5 %. Posteriormente, el vector de entropías resultante se codifica de forma binaria de forma que solo refleje la modulación de la entropía en el tiempo, de forma similar a como se mostró en la sección pero sobre el tiempo. La TES es compacta ya que usa solo un bit por trama. Para un audio en formato wav de 6 minutos de duración, el tamaño de la TES representa el 0.012 % de su tamaño original y toma 540 ms para calcularse (Se utilizó una computadora Dell Precision T5500, con procesador Intel(R) Xeon(R) CPU 2.00GHZ). La comparación entre dos TES también utiliza pocos recursos ya que se utiliza la distancia Hamming entre los bits de las firmas. Adicionalmente, la TES proporciona privacidad a los usuarios ya que no es necesario enviar los audios originales al servidor y no es posible recuperar el audio original con los bits que forman la firma.

Para evaluar el funcionamiento de la TES en esta aplicación, se hicieron experimentos utilizando dos bases de datos. La base de datos 1 se grabó en los pasillos del departamento de Ciencias de Computación en CICESE y la base de datos 2 se grabó en las instalaciones de la residencia geriátrica *Serena*. Todos los sonidos se capturaron utilizando una frecuencia de muestreo de 24000Hz con 8 bits de profundidad. La base de datos 1 consta de 18 minutos de audio capturado con 3 dispositivos diferentes: A, B y C. En este escenario se obtienen 5 combinaciones diferentes en las grabaciones. A, B y C se encuentran juntos, A y B están juntos pero no C, A y C están juntos pero no B, B y C estan juntos pero no A, y finalmente, todos los dispositivos están en diferentes ubicaciones por lo que graban distinta información. La base de datos 2 consta de 6 minutos con 30 segundos de audio capturado usando 2 dispositivos para grabación: A y B). En este escenario se tienen las siguientes combinaciones de datos, A y B están juntos, A y B están en diferentes ubicaciones.

Para comparar dos firmas se usan segmentos de 15 segundos. Se comparan 15 segundos provenientes de un dispositivo de grabación contra los 15 segundos correspondientes del segundo dispositivo de grabación. Además, se toma en cuenta un desplazamiento de 300ms. La necesidad de usar un desplazamiento se da porque cuando se transmiten las firmas desde los dispositivos hacia el servidor, existe la posibilidad que estas se reciban desplazadas en el tiempo. Esto puede ser ocasionado por las diferen-

tes posiciones relativas con que los dispositivos capturan el audio o por cuestiones de transmisión de datos.

Cada vez que se comparan los segmentos de las firmas de 15 segundos, se fija el segmento de un archivo mientras que el otro segmento se desliza adelantándose y atrasándose bit por bit hasta 300ms con respecto al segmento fijo. Una vez que se realizan todas las comparaciones posibles, la distancia entre ambos segmentos corresponde a la distancia Hamming mínima después de considerar todos los desplazamientos. Debido a que se utilizan traslapes, cada vez que se obtiene una distancia se representan 1.7s de audio.

Las distancias entre las firmas sirven para determinar si dos dispositivos se encuentran en la misma área o separados. Una distancia pequeña se da cuando dos firmas son similares, es decir, representan el mismo sonido. Por otro lado, cuando una distancia es grande se indica que las firmas son diferentes pues provienen de sonidos distintos. Es por ello que se estableció un umbral bajo para establecer cuando se considera que una distancia entre dos firmas es pequeña y un umbral alto para establecer cuando se considera que se tiene una distancia grande. Ambos umbrales se obtuvieron experimentalmente. Si una distancia es menor que el umbral bajo entonces se predice que en ese momento los dos dispositivos están juntos. Si una distancia es mayor que el umbral alto entonces se predice que en ese momento los dispositivos están separados. Si una distancia se encuentra en medio de los dos umbrales entonces la información de ese momento queda sin categoría. Sin embargo, se lleva a cabo una segunda etapa usando un umbral en el tiempo para que no existan elementos sin clasificar. Esta etapa es posible debido a que los elementos sin categoría suceden de forma dispersa como se aprecia en la figura 43 y se puede aplicar una especie de filtro pasabajas. Para ejemplificar el uso del umbral en el tiempo supóngase que se ha clasificado que dos dispositivos están juntos por 10 ocasiones continuas, después se tienen dos clasificaciones indeterminadas y luego se continúa clasificando que están juntos. Un umbral en el tiempo de 4 indica que se requieren al menos 5 clasificaciones continuas para poder cambiar de categoría, de esta forma las dos clasificaciones indeterminadas quedan absorbidas y todo el segmento sería clasificado como juntos. El umbral en el tiempo también sirve para eliminar valores atípicos,

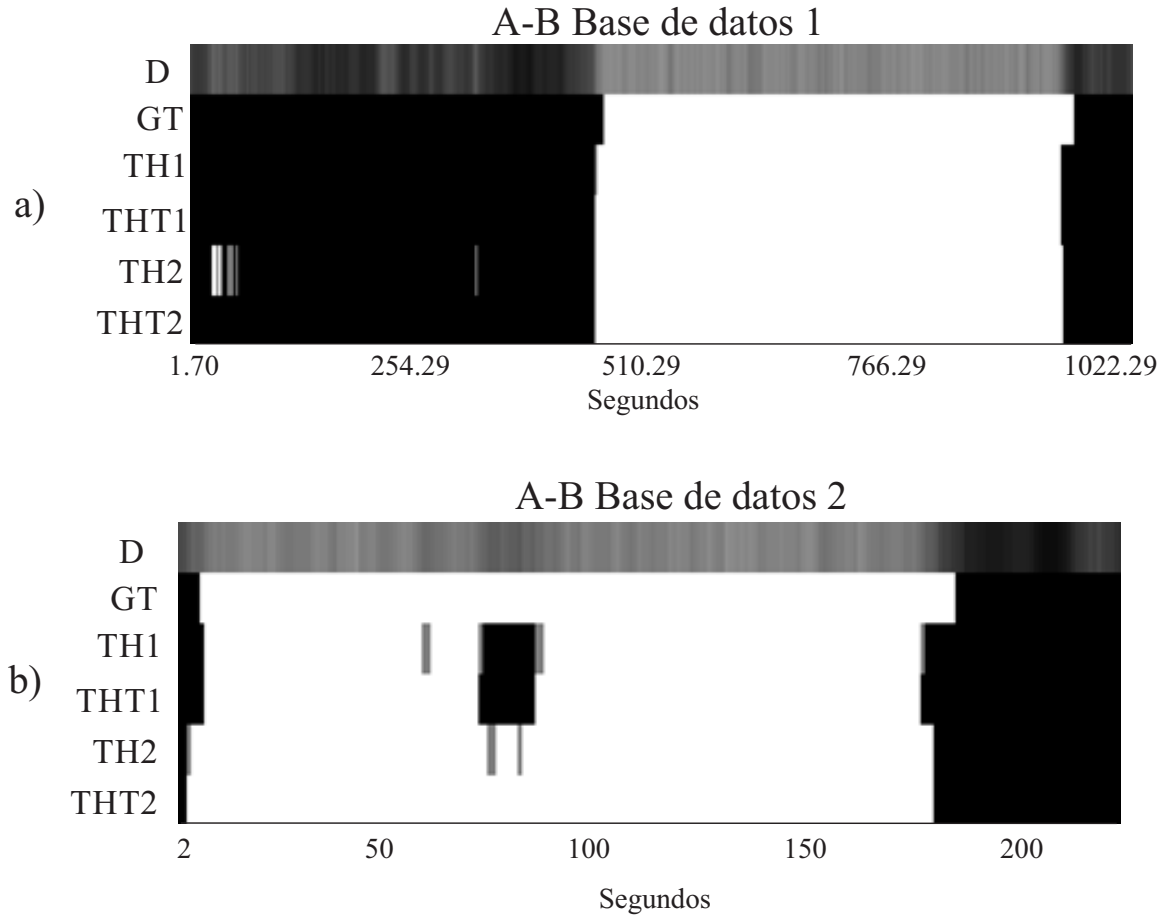


Figura 43: Resultados al comparar los sonidos grabados por los dispositivos A y B en las bases de datos 1 y 2.

por ejemplo, que dada una clasificación donde se considera que los dispositivos estén juntos, exista un instante en medio que indique que están separados.

Se probaron dos conjuntos de umbrales alto y bajo, al primer par le llamamos TH1 y al segundo TH2. En la figura 43.a se muestra la comparación de los sonidos grabados por los dispositivos A y B de la base de datos 1 y en la figura 43.b se muestra la comparación entre los sonidos grabados por los dispositivos A y B de la base de datos 2. El renglón D muestra las distancias, donde un color más oscuro indica mayor similitud (es decir, los dispositivos están cerca) y un color más claro indica que existe una diferencia (es decir, los dispositivos están en diferentes ubicaciones). El renglón GT corresponde a la ground truth, aquí, el color negro indica que ambos dispositivos está juntos y el color blanco indica que están separados. Los renglones TH1 y TH2 corresponden a la clasificación al usar los dos conjuntos de umbrales, aquí el color negro se refiere a la predicción de que ambos dispositivos se encuentran en la misma ubicación, blanco significa que están en lugares

Tabla 11: Resultados de la clasificación de llanto sostenido

	Dispositivos de grabación	VP (%)		FN (%)	
		TH1	TH2	TH1	TH2
Base de datos 1	A-B	97.92	98.08	2.57	1.92
	A-C	94.37	84.20	3.11	1.91
	B-C	98.08	97.35	1.43	0.95
Base de datos 2	A-B	93.75	95.17	12.5	2.84

distintos y el color gris representa una clasificación no determinada. Las líneas THT1 y THT2 corresponden a utilizar umbrales en el tiempo sobre los resultados de TH1 y TH2 para eliminar los valores indeterminados o atípicos.

En la tabla 11 se muestran los VP y los FN al usar umbrales en el tiempo sobre TH1 y TH2. Usando nuestra propuesta logramos obtener una exactitud mayor al 90 %. Para esta aplicación, se busca disminuir los falsos negativos (i.e. el sistema indica que están juntos cuando están separados) aunque esto implique aumentar los falsos positivos (i.e. el sistema indica que están separados cuando están juntos) ya que se prefiere no perder grabaciones a optimizar recursos de batería.

Los FN se producen principalmente cuando dos dispositivos se están acercando o separando. Este error se produce por la discretización del tiempo, recordamos que se utiliza una ventana deslizante de 15 segundos la cual se puede reducir con el objetivo de obtener mejores predicciones en las fronteras. Para esta aplicación, se sugiere no dejar de grabar hasta que se ha detectado que dos dispositivos han estado juntos por suficiente tiempo.

Uno de los retos principales de este experimento se da por el retardo que tiene una audio en llegar a los dispositivos dependiendo de su posición relativa con respecto a la fuente del sonido. Otro reto se da con la reverberación que se produce en lugares cerrados. Además, se detectó otro problema en la residencia de adultos mayores ya que los cuartos se encuentran conectados por un pasillo lo que provoca que en algunas ocasiones un audio producido en un cuarto se pueda escuchar en cuartos adyacentes. Por ejemplo, si una persona se encuentra produciendo un ruido fuerte en el pasillo enfrente de las puertas de dos cuartos esto podría provocar que el sistema piensa que los cuartos separados en realidad son la misma ubicación. Otro problema puede surgir si dos resi-

dentes se encuentran viendo un mismo programa de televisión en cuartos separados ya que el sistema estaría escuchando el mismo sonido.

7.3.1. Discusión

Este proyecto es un ejemplo del uso de la información auditiva para conocer aspectos de la vida de los adultos mayores donde no se requieren detectar eventos de sonido específicos. Mediante el audio capturado por sus dispositivos es posible detectar su socialización la cual podría utilizarse para mejorar sus relaciones interpersonales por medio del sistema de intervenciones asistidas. Como se plantea, los beneficios otorgados con el análisis del audio por medio de la firma también ayudan a optimizar recursos de espacio, procesamiento y batería. Un algoritmo similar es utilizado por la compañía FOX como prueba de compra de los servicios de cable para activar un servicio adicional gratuito llamado foxplay (http://www.foxplay.com/mx/).

Capítulo 8. Conclusiones y trabajo futuro

En este trabajo de investigación se desarrolló una firma de audio para reconocer sonidos ambientales. El objetivo de la firma consistió en mejorar el desempeño del estado del arte para abordar los retos de: robustez al ruido, transparencia, generalización a otros usuarios, tamaño compacto y facilidad de cálculo.

En una primer etapa, se evaluó el uso de la entropía en bandas de frecuencia mediante la MEL-MBSES y su versión codificada binaria. A partir de los experimentos realizados en esta etapa, se determinó que es posible usar una firma de audio compacta para representar sonidos ambientales contaminados con ruido. También se identificó que el uso de las derivadas en las firmas de audio incrementa la eficiencia para el reconocimiento de sonidos mezclados utilizando poco espacio de almacenamiento.

Debido a las variaciones de estructura en el tiempo y frecuencia que se producen entre distintas instancias de sonidos de una misma clase, fue necesaria una siguiente etapa de diseño para obtener una firma de audio transparente, robusta ante desfasamientos, ruido y heterogeneidad. Por tal motivo, se diseñó la firma H1DH2D-MEL-MBSES la cual consiste en un vector sucinto formado por histogramas. Esta firma permite representar sonidos de duración variable en un vector de tamaño fijo que se puede evaluar rápidamente usando un clasificador SVM. La evaluación de esta firma indicó una mejoría al comparar con el estado del arte en todos los retos presentados.

Una vez que se consiguió satisfacer los retos en el desarrollo de la firma de audio, se evaluó la viabilidad de utilizar el reconocimiento automático de sonidos ambientales para desarrollar aplicaciones conscientes del contexto con el objetivo de asistir a adultos mayores y a sus cuidadores. Esta evaluación se realizó mediante experimentos usando sonidos capturados en condiciones reales en una residencia geriátrica y de un departamento habitado por un adulto mayor. Mediante el reconocimiento automático de actividades y comportamientos es posible informar oportunamente sobre la condición actual de los adultos mayores a sus cuidadores. En esta tesis, además se desarrolló un componente detector de comportamientos disruptivos audibles para un sistema de intervenciones no farmacológicas. Con este sistema se busca actuar ante las necesidades de los adultos

mayores evitando el uso de fármacos, los cuales pueden producir efectos secundarios.

Adicionalmente, se desarrollaron otros proyectos donde se realiza reconocimiento de contexto por medio del audio. En estos proyectos se explora el análisis de audio para reconocer otros elementos del contexto distintos a los sonidos ambientales. Por ejemplo, se implementó un detector de voz, el cual puede servir para inferir la socialización de los usuarios. Otro proyecto consistió en utilizar el audio grabado por varios micrófonos para detectar si dos personas se encuentran en un mismo lugar. Finalmente, también se trabajo con identificación de llanto de bebés. Con estos proyectos se resalta la riqueza de la información que provee el audio que se puede utilizar para desarrollar diferentes aplicaciones.

8.1. Contribuciones

A continuación se enumeran las principales contribuciones de este trabajo de investigación:

- El diseño de una firma de audio compacta, robusta al ruido, transparente, generalizable a otros usuarios y fácil de calcular.
- La evaluación de la firma de audio propuesta con sonidos reales provenientes de escenarios habitados por adultos mayores.
- La creación de una base de datos con sonidos heterogéneos y mezclados para evaluar las propuestas de firma de audio y clasificación.
- El desarrollo de un componente de reconocimiento de comportamientos disruptivos audibles dirigido a un sistema de intervenciones no farmacológicas.

Adicionalmente, se realizaron publicaciones en diferentes foros las cuales se indican a continuación:

- Publicaciones en revistas arbitradas:
    - Scalable Identification of Mixed Environmental Sounds, Recorded from Heterogeneous Sources, Pattern Recognition Letters, (2015), pp. 153-160.

- Publicaciones en congresos:

  - Abnormal Behavioral Patterns Detection from Activity Records of Institutionalized Older Adults, 6th International Workshop on Human Behavior Understanding-Behavior Analysis for the Eldery (Ubicomp), Osaka, Japan, September 7th 2015.
  - Investigating Correlation between Verbal Interactions and Perceived Stress, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, August 25-29th 2015.
  - Detecting Disruptive Vocalizations for Ambient Assisted Interventions for Dementia, Workshop on User and Ambient Adaptive Gerontechnologies (WAGER) in the 8th International Conference on Ubiquitous and Ambient Intelligence (UCAMI), Belfast, Ireland, December 2-5th 2014.
  - Collaborative Opportunistic Sensing of Human Behavior with Mobile Phones, Workshop Smart Health systems and Applications (Ubicomp), Seattle, United States, September13-17th 2014.
  - A Context-Aware Baby Monitor for the Automatic Selective Archiving of the Language of Infants, Mexican International Conference on Computer Science, Morelia Michoacán, México, October 30th 2013.
  - Activity Recognition Using a Spectral Entropy Signature (Doctoral Colloquium), 14th ACM International Conference on Ubiquitous Computing, Pittsburgh, Pennsylvania, United States, September 5-8th 2012.
  - Environmental Sound Recognition by Measuring Significant Changes in the Spectral Entropy, 4th Mexican, 4th Mexican Conference on Pattern Recognition, Huatulco, México, June 27-30th 2012.
  - Auditory Scene Analysis for Activity-Aware Computing Using Audio Features as Text Cues (Doctoral Consortium), Pervasive, San Francisco, CA. United States, June 12-15th 2011.

8.2. Trabajo futuro

Si bien se lograron buenos resultados con la firma propuesta, aún existe buen margen de mejora. Por ejemplo, se puede mejorar el desempeño de la clasificación sobre sonidos mezclados. Estos sonidos tienen propiedades aditivas que provocan que la señal corresponda a más de una sola clase. Aunque una forma de mejorar los resultados es mejorando la transparencia de la firma, se sugiere proponer diferentes técnicas de clasificación.

En la actualidad, los clasificadores reportados en la literatura no consideran este tipo de señales mezcladas ya que están diseñados para asignar una y solo una clase a los sonidos de entrada; sin embargo, para este tipo de sonidos se requiere de clasificadores que se asignen múltiples etiquetas a los sonidos de entrada ya que estos no se pueden describir como de un solo tipo.

Con el objetivo de diseñar aplicaciones en tiempo real sobre dispositivos móviles, se requiere un buen desempeño en la obtención de la firma y la clasificación. Una forma de incrementar el desempeño se puede conseguir al indexar las firmas de audio. Para ello, como trabajo futuro se sugiere proponer métricas y procedimientos adecuados para convertir el problema de reconocimiento de sonidos ambientales en un problema de recuperación de información.

Con el objetivo de desarrollar sistemas conscientes del contexto para apoyar a adultos mayores, el siguiente paso consiste en incluir la firma H1dH2d-MEL-MBSES en dispositivos móviles. Además, para conseguir mejores resultados de clasificación, se sugiere construir modelos a partir de una base de datos más completa que incluya más instancias de sonidos y más casos donde se presenten sonidos mezclados correctamente etiquetados para mejorar las evaluaciones.

Lista de referencias bibliográficas

Abowd, G. D., Dey, A. K., Brown, P. J., Davies, N., Smith, M., and Steggles, P. (1999). Towards a better understanding of context and context-awareness. En: *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, London, UK, UK. Springer-Verlag, HUC '99, pp. 304–307.

Acree, L., Longfors, J., Fjeldstad, A., Fjeldstad, C., Schank, B., Nickel, K., Montgomery, P., and Gardner, A. (2006). Physical activity is related to quality of life in older adults. *Health and Quality of Life Outcomes*, **4**(1): 37.

Active and Programme, A. L. (2015). Active and assisted living programme. Recuperado de http://www.aal-europe.eu.

Anusuya, M. and Katti, S. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, **6**(3): 181–205.

Aucouturier, J.-J., Defreville, B., and Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *The Journal of the Acoustical Society of America*, **122**(2): 881–891.

B. Gygi, G. K. and Watson, C. (2004). Spectral-temporal factors in the identification of environmental sounds. *Acoustic Society of America*, **115**(3): 1252 – 1265.

Baldauf, M., Dustdar, S., and Rosenberg, F. (2007). A survey on context aware systems. *Int. J. Ad Hoc Ubiquitous Comput.*, **2**(4): 263–277.

Beranek, L. L. (1949). *Acoustical Measurements*. John Wiley and Sons Inc.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, **2**(2): 121–167.

Burns, K., Jayasinha, R., Tsang, R., and Brodaty, H. (2012). Behavior management, a guide to good practice: Managing behavioral and psychological symptoms of dementia. *Dementia Collaborative Research Centre. Retrieved June*, **3**: 1–190.

Camarena-Ibarrola, J. A. (2008). *Identificación automática de señales de audio*. Tesis de doctorado, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, México.

Chang, C.-C. and Lin, C.-J. (2015). Libsvm. Recuperado de https://www.csie.ntu.edu.tw/ cjlin/libsvm/.

Chen, G. and Kotz, D. (2000). A survey of context-aware mobile computing research. Reporte técnico, Dartmouth College, Hanover, NH, USA.

Chen, J., Kam, A., Zhang, J., Liu, N., and Shue, L. (2005). Bathroom activity monitoring based on sound. En: H.-W. Gellersen, R. Want, and A. Schmidt (eds.), *Pervasive Computing*, Vol. 3468 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 47–61.

Chen, L., Nugent, C. D., and Wang, H. (2012). A knowledge-driven approach to activity recognition in smart homes. *IEEE Trans. on Knowl. and Data Eng*, **6**(24): 961–974.

Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., Legrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., Klasnja, P., Koscher, K., Landay, J., Lester, J., Wyatt, D., and Haehnel, D. (2008). The mobile sensing platform: An embedded activity recognition system. *Pervasive Computing, IEEE*, **7**(2): 32–41.

Cohen-Mansfield, J. (1997). Conceptualization of agitation: results based on the cohen-mansfield agitation inventory and the agitation behavior mapping instrument. *International Psychogeriatrics*, **8**(S3): 309–315.

Cohen-Mansfield, J. and Martin, L. S. (2010). Assessment of agitation in older adults. *Handbook of assessment in clinical gerontology*, pp. 381–404.

Cornejo, R., Tentori, M., and Favela, J. (2013). Ambient awareness to strengthen the family social network of older adults. *Comput. Supported Coop. Work*, **22**(2-3): 309–344.

Cornwell, B., Laumann, E., and Schumm, L. (2008). The social connectedness of older adults: A national profile. *American sociological review*, **73**(2): 185–203.

Cummings, J. L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D. A., and Gornbein, J. (1994). The neuropsychiatric inventory comprehensive assessment of psychopathology in dementia. *Neurology*, **44**(12): 2308–2308.

de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, **111**(4): 1917–1930.

Dennis, J., Dat, T., and Chng, E. (2012). Image feature representation of the subband power distribution for robust sound event classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, **PP**(99): 1.

Dennis, J., Tran, H., and Chng, E. (2013). Overlapping sound event recognition using local spectrogram features and the generalised hough transform. *Pattern Recognition Letters*, **34**(9): 1085 – 1093.

Doukas, C. and Maglogiannis, I. (2008). Advanced patient or elder fall detection based on movement and sound data. En: *Pervasive Computing Technologies for Healthcare, 2008. PervasiveHealth 2008. Second International Conference on*, Jan. pp. 103–107.

Drake, L., Katsaggelos, A., Rutledge, J., and Zhang, J. (2002). Sound source separation via computational auditory scene analysis-enhanced beamforming. En: *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, Aug. pp. 259–263.

Ellis, D. P. and Lee, K. (2006a). Accessing minimal-impact personal audio archives. *IEEE Multimedia*, **13**(4): 30–38.

Ellis, D. P. W. and Lee, K. (2006b). Accessing minimal-impact personal audio archives. *IEEE MultiMedia*, **13**(4): 30–38.

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T., Lorho, G., and Huopaniemi, J. (2006). Audio-based context recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, **14**(1): 321–329.

Ferdous, R., Osmani, V., Beltran-Marquez, J., and Mayora, O. (2015). Investigating correlation between verbal interactions and perceived stress. En: *Proceedings of IEEE Engineering in Medicine and Biology Society (EMBS)*, August, Milan.

Fick, W., van der Borgh, J., Jansen, S., and Koopmans, R. (2014). The effect of a lollipop on vocally disruptive behavior in a patient with frontotemporal dementia: a case-study. *International Psychogeriatrics*, **26**(12): 2023–2026.

Fogarty, J., Au, C., and Hudson, S. E. (2006). Sensing from the basement: A feasibility study of unobtrusive and low-cost home activity recognition. En: *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, New York, NY, USA. ACM, UIST '06, pp. 91–100.

Györbíró, N., Fábián, A., and Hományi, G. (2009). An activity recognition system for mobile phones. *Mob. Netw. Appl.*, **14**(1): 82–91.

Handte, M., Iqbal, U., Apolinarski, W., and Marrón, P. J. (2010). Challenges in ubiquitous context recognition with personal mobile devices. En: *Proceedings of the 4th ACM International Workshop on Context-Awareness for Self-Managing Systems*, New York, NY, USA. ACM, CASEMANS '10, pp. 6:40–6:45.

Hartmann, W. M. (1998). *Signals, Sound, and Sensation*. Springer Science+Business Media.

Hearst, M. A. (1998). Support vector machines. *IEEE Intelligent Systems*, **13**(4): 18–28.

Hedelin, P. and Huber, D. (1990). Pitch period determination of aperiodic speech signals. En: *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, Apr. pp. 361–364 vol.1.

Heittola, T., Mesaros, A., Virtanen, T., and Eronen, A. (2011). Sound event detection in multi-source environments using source separation. En: *Workshop on machine listening in Multisource Environments*. pp. 36–40.

Heittola, T., Mesaros, A., Eronen, A., and Virtanen, T. (2013). Context-dependent sound event detection. *EURASIP Journal on Audio, Speech, and Music Processing*, **2013**(1).

Ibarrola, A. and Chavez, E. (2006). A robust entropy-based audio-fingerprint. En: *Multimedia and Expo, 2006 IEEE International Conference on*, July. pp. 1729–1732.

Institute, A. N. S. (1999). American national standards institute. Accessed: 2014-03-04.

Jalbert, J. J., Daiello, L. A., and Lapane, K. L. (2008). Dementia of the alzheimer type. *Epidemiologic reviews*, **30**(1): 15–34.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, **74**(11): 1477–1493.

Kern, N. and Schiele, B. (2003). Context-aware notification for wearable computing. En: *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, Washington, DC, USA. IEEE Computer Society, ISWC '03, pp. 223–.

Kern, N., Schiele, B., and Schmidt, A. (2007). Recognizing context for annotating a live life recording. *Personal Ubiquitous Comput.*, **11**(4): 251–263.

Kirk, R. and Newmarch, J. (2005). A location-aware, service-based audio system. En: *Consumer Communications and Networking Conference, 2005. CCNC. 2005 Second IEEE*, Jan. pp. 343–347.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A survey of mobile phone sensing. *Comm. Mag.*, **48**(9): 140–150.

Lindsay, P. H. and Norman, D. A. (1977). *Human information processing: An introduction to psychology*. Academic Press.

Lu, L. and Hanjalic, A. (2008). Audio keywords discovery for text-like audio content analysis and retrieval. *Multimedia, IEEE Transactions on*, **10**(1): 74–85.

Lukowicz, P., Ward, J., Junker, H., Stäger, M., Tröster, G., Atrash, A., and Starner, T. (2004). Recognizing workshop activity using body worn microphones and accelerometers. En: A. Ferscha and F. Mattern (eds.), *Pervasive Computing*, Vol. 3001 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 18–32.

Ma, L., Smith, D., and Milner, B. (2003). Environmental noise classification for context-aware applications. En: V. Mařík, W. Retschitzegger, and O. Štěpánková (eds.), *Database and Expert Systems Applications*, Vol. 2736 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 360–370.

Ma, L., Milner, B., and Smith, D. (2006). Acoustic environment classification. *ACM Trans. Speech Lang. Process.*, **3**(2): 1–22.

Maslach, C., Schaufeli, W. B., and Leiter, M. P. (2001). Job burnout. *Annual review of psychology*, **52**(1): 397–422.

Meyer, S. and Rakotonirainy, A. (2003). A survey of research on context-aware homes. En: *Proceedings of the Australasian Information Security Workshop Conference on ACSW Frontiers 2003 - Volume 21*, Darlinghurst, Australia, Australia. Australian Computer Society, Inc., ACSW Frontiers '03, pp. 159–168.

Miao, Z. and Yuan, B. (2005). Discussion on pervasive computing paradigm. En: *TENCON 2005 2005 IEEE Region 10*, Nov. pp. 1–6.

Mihailidis, A. and Fernie, G. R. (2002). Context-aware assistive devices for older adults with dementia. *Gerontechnology*, **2**(2): 173–188.

Miluzzo, E., Lane, N. D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S. B., Zheng, X., and Campbell, A. T. (2008). Sensing meets mobile social networks: The design, implementation and evaluation of the cenceme application. En: *Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems*, New York, NY, USA. ACM, SenSys '08, pp. 337–350.

Min, C.-h., Ince, N. F., and Tewfik, A. H. (2008a). *Early Morning Activity Detection Using Acoustics and Wearable Wireless Sensors*.

Min, C.-H., Ince, N. F., and Tewfik, A. H. (2008b). Early morning activity detection using acoustics and wearable wireless sensors. En: *Signal Processing Conference, 2008 16th European*. IEEE, pp. 1–5.

Morris, M., Lundell, J., Dishman, E., and Needham, B. (2003). New Perspectives on Ubiquitous Computing from Ethnographic Study of Elders with Cognitive Decline. En: *Ubicomp*, Berlin. pp. 227–242.

Muller, K., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, **12**(2): 181–201.

Navarretta, C. (2014). The automatic identification of the producers of co-occurring communicative behaviours. *Cognitive Computation*, **6**(4): 689–698.

Navarro, R. F., Rodriguez, M., and Favela, J. (2014). Intervention tailoring in augmented cognition systems for elders with dementia. *Biomedical and Health Informatics, IEEE Journal of*, **18**(1): 361–367.

Niessen, M., van Maanen, L., and Andringa, T. (2008). Disambiguating sounds through context. En: *Semantic Computing, 2008 IEEE International Conference on*, Aug. pp. 88–95.

Ntalampiras, S., Potamitis, I., and Fakotakis, N. (2008). Automatic recognition of urban environmental sounds events. En: *International Association for Pattern Recognition Workshop on Cognitive Information Processing, EURASIP*.

of Waikato, T. U. (2015). Weka.

Oliver, N., Garg, A., and Horvitz, E. (2004). Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, **96**(2): 163–180.

Ono, N., Yamagish, J., Ohsuga, T., Itahashi, S., and Ishimoto, Y. (2015). Real world computing partnership.

O'Shaughnessy, D. (1999). *Speech Communications: Human and Machine*. Wiley-IEEE Press.

Paciga, M. and Lutfiyya, H. (2005). Herecast:an open infrastructure for locationbased services using wifi. En: *Wireless And Mobile Computing, Networking And Communications, 2005. (WiMob'2005), IEEE International Conference on*, Aug. Vol. 4, pp. 21–28.

Perttunen, M., Van Kleek, M., Lassila, O., and Riekki, J. (2008). Auditory context recognition using svms. En: *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2008. UBICOMM '08. The Second International Conference on*, Sept. pp. 102–108.

Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D. J., Fox, D., Kautz, H., and Hahnel, D. (2004). Inferring activities from interactions with objects. *IEEE Pervasive Computing*, **3**(4): 50–57.

Potamitis, I. and Ganchev, T. (2008). Generalized recognition of sound events: Approaches and applications. En: G. Tsihrintzis and L. Jain (eds.), *Multimedia Services in Intelligent Environments*, Vol. 120 de *Studies in Computational Intelligence*. Springer Berlin Heidelberg, pp. 41–79.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2): 257–286.

Rialle, V., Ollivet, C., Guigui, C., and Hervé, C. (2008). What do family caregivers of alzheimer's disease patients desire in smart home technologies? *Methods of Information in Medicine*, **47**: 63–69.

Rincon, E., Beltran, J., Tentori, M., Favela, J., and Chavez, E. (2013). A context-aware baby monitor for the automatic selective archiving of the language of infants. En: *Computer Science (ENC), 2013 Mexican International Conference on*, Oct. pp. 60–67.

Rothwell, J. D. (2009). *In the Company of Others: An Introduction to Communication*. Oxford University Press.

Sadowsky, C. H. and Galvin, J. E. (2012). Guidelines for the management of cognitive and behavioral problems in dementia. *The Journal of the American Board of Family Medicine*, **25**(3): 350–366.

Sahidullah, M. and Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in $\{$MFCC$\}$ computation for speaker recognition. *Speech Communication*, **54**(4): 543 – 565.

Sawhney, N. and Schmandt, C. (1999). Nomadic radio: Scaleable and contextual notification for wearable audio messaging. En: *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 96–103.

Scheirer, E. and Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. En: *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, Apr. Vol. 2, pp. 1331–1334 vol.2.

Schroeder, J., Wabnik, S., Hengel, P., and Goetze, S. (2011). Detection and classification of acoustic events for in-home care. En: R. Wichert and B. Eberhardt (eds.), *Ambient Assisted Living*. Springer Berlin Heidelberg, pp. 181–195.

Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, **5**(1): 3–55.

Stager, M., Lukowicz, P., and Troster, G. (2004). Implementation and evaluation of a low-power sound-based user activity recognition system. En: *Wearable Computers, 2004. ISWC 2004. Eighth International Symposium on*, Oct. Vol. 1, pp. 138–141.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, **8**(3): 185–190.

Surie, D., Lagriffoul, F., Pederson, T., and Sjölie, D. (2007). Activity recognition based on intra and extra manipulation of everyday objects. En: H. Ichikawa, W.-D. Cho, I. Satoh, and H. Youn (eds.), *Ubiquitous Computing Systems*, Vol. 4836 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 196–210.

Tan, L. and Jiang, J. (2013). *Digital signal processing: fundamentals and applications*. Academic Press.

Temko, A. and Nadeu, C. (2009a). Acoustic event detection in meeting-room environments. *Pattern Recognition Letters*, **30**(14): 1281 – 1288.

Temko, A. and Nadeu, C. (2009b). Acoustic event detection in meeting-room environments. *Pattern Recogn. Lett.*, **30**(14): 1281–1288.

Tesoriero, R., Tebar, R., Gallud, J., Lozano, M., and Penichet, V. (2010). Improving location awareness in indoor spaces using {RFID} technology. *Expert Systems with Applications*, **37**(1): 894 – 898.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, **88**(1): 97–100.

Vacher, M., Portet, F., Fleury, A., and Noury, N. (2010). Challenges in the processing of audio channels for ambient assisted living. En: *e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on*, July. pp. 330–337.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press.

Want, R., Hopper, A., Falcão, V., and Gibbons, J. (1992). The active badge location system. *ACM Trans. Inf. Syst.*, **10**(1): 91–102.

Wilson, D. H. (2005). *Assistive Intelligent Environments for Automatic Health Monitoring*. Tesis de doctorado, Carnegie Mellon University, Pittsburgh, PA, USA.

Wold, E., Blum, T., Keislar, D., and Wheaton, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, **3**(3): 27–36.

Wood, J. T. (2012). *Interpersonal Communication: Everyday encounters*. 4th, Belmont, CA: Wadsworth.

Zhang, T. and Kuo, C. C. (2001). *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*. Kluwer Academic Publishers. Norwell, MA, USA.

Zúñiga, E. and Vega, D. (2004). El envejecimiento en la población mundial. D.F. Consejo Nacional de Población.

Zwicker, E. (1961). Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen). *Journal Acoustical Society of America*, **33**(2): 248.

Apéndice

En esta sección se extiende mayor detalle sobre los clasificadores: Máquina de soporte vectorial y modelos ocultos de Markov. Se repite la introducción de ambos temas para evitar confusión.

Máquina de soporte vectorial

Definición La máquina de soporte vectorial (SVM por sus siglas en inglés), es un algoritmo de aprendizaje para clasificación y regresión. La SVM esta basada en el concepto de proyectar un conjunto de datos a un espacio de características de dimensión muy alta y entonces determinar hiperplanos óptimos para separar los datos de clases diferentes (Muller *et al.*, 2001).

Intuición Dados N puntos de entrenamiento con dimension d, los cuales provienen de dos clases distintas $y_i = -1\ o\ +1$,, es decir:

$$\{x_i, y_i\}\ \ donde\ i = 1...N,\ y \in \{-1, 1\},\ x \in \mathbb{R}^d \tag{23}$$

El objetivo del algoritmo SVM consiste en encontrar un hiperplano con una orientación que permita separar ambas clases maximizando un margen entre el hiperplano y los puntos más cercanos de cada clase. En la figura 44 se ilustra un ejemplo en donde se supone que puntos de dos clases son linealmente separables por el hiperplano $\mathbf{w} \cdot x + b = 0$, donde **w** es normal al hiperplano y $\frac{b}{|\mathbf{w}|}$ es la distancia perpendicular desde el hiperplano hasta el origen. Cada uno de los semi-espacios definidos por este hiperplano corresponde a una clase. Para encontrar a cual clase pretence un punto x se puede evaluar usando: $f(x) = sgn(w \cdot x + b)$. El margen que se busca maximizar esta ilustrado en la figura como el espacio entre las líneas punteadas H_1 y H_2.

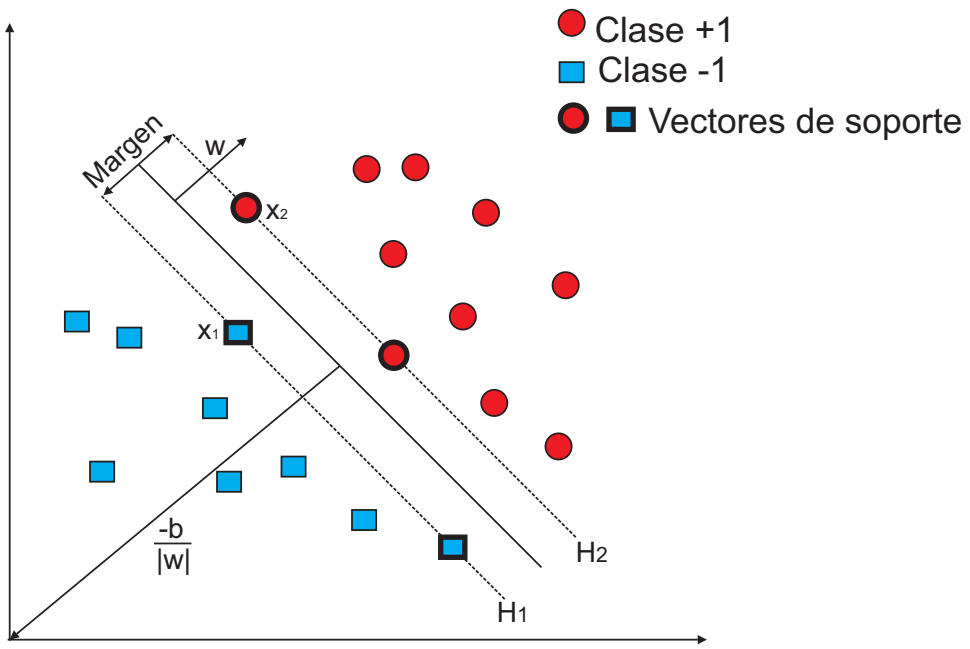


Figura 44: Hiperplano separando puntos de dos clases linealmente

Hiperplano El hiperplano se encuentra seleccionando las variables **w** y *b* de forma que los datos de entrenamiento cumplan con las siguientes ecuaciones.

$$\mathbf{w} \cdot x_i + b \geq +1 \quad para \quad y_i = +1 \tag{24}$$

$$\mathbf{w} \cdot x_i + b \geq -1 \quad para \quad y_i = -1 \tag{25}$$

Estas ecuaciones combinadas producen la siguiente ecuación:

$$y_i \left(\mathbf{w} \cdot x_i + b\right) - 1 \geq 0 \tag{26}$$

A los puntos que se encuentran más cercanos al hiperplano, se les conoce como vectores de soporte. Los planos H_1 y H_2 donde caen los vectores de soporte se pueden describir como:

$$\mathbf{w} \cdot x_i + b = +1 \quad para \quad H_1 \tag{27}$$

$$\mathbf{w} \cdot x_i + b = -1 \quad para \quad H_2 \tag{28}$$

En la figura 11 los planos H_1 y H_2 se ilustran con líneas punteadas y los vectores de soporte se representan con los puntos con linea gruesa. Nótese que H_1 y H_2 son paralelos

y que ningún punto de entrenamiento cae entre ellos.

Si se consideran dos vectores de soporte x_1 y x_2 de distintas clases con $(\mathbf{w}\cdot x_1)+b \geq +1$ y $(\mathbf{w} \cdot x_2) + b \geq +1$ respectivamente, el margen esta dado por la distancia entre estos dos puntos medidos perpendicularmente al hiperplano, es decir; $\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (x_1 - x_2) = \frac{2}{\|\mathbf{w}\|}$. El maximizar este margen es equivalente al problema de optimización de encontrar el mínimo de $\mathbf{w}$.

$$\min \|\mathbf{w}\| \quad \text{tal que} \quad y_i (\mathbf{w} \cdot x_i + b) - 1 \geq 0 \tag{29}$$

Problema de optimización Dado que la norma involucra el cálculo de la raíz cuadrada, la optimización de $\|\mathbf{w}\|$ es difícil de resolver. Sin embargo, se puede alterar la ecuación sustituyendo $\|\mathbf{w}\|$ por $\frac{1}{2} \|\mathbf{w}\|^2$ por conveniencia, esto debido a que minimizar el primer término es equivalente a minimizar el segundo. Con esta sustitución es posible usar optimización de Programación Cuadrática. Ahora el problema consiste en encontrar:

$$\min \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{tal que} \quad y_i (\mathbf{w} \cdot x_i + b) - 1 \geq 0 \tag{30}$$

Se utiliza la formulación Lagrangiana sobre este problema. Para ello se incluyen los multiplicadores de Lagrange α donde $\alpha_i \geq 0$ los cuales son más fáciles de manejar. Además, con esta reformulación los datos de entrenamiento solo apareceran en la forma de producto punto entre vectores. Esto último es una propiedad crucial que permite generalizar el procedimiento a casos no lineales.

$$L_P \equiv \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i y_i (\mathbf{w} \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i \tag{31}$$

Para más información sobre como resolver este problema de optimizar L_P, dirigirse a (Burges, 1998).

Problemas no lineales En la práctica, usualmente los puntos de distintas clases no se pueden separar linealmente. Por este motivo, los datos se mapean a otro espacio

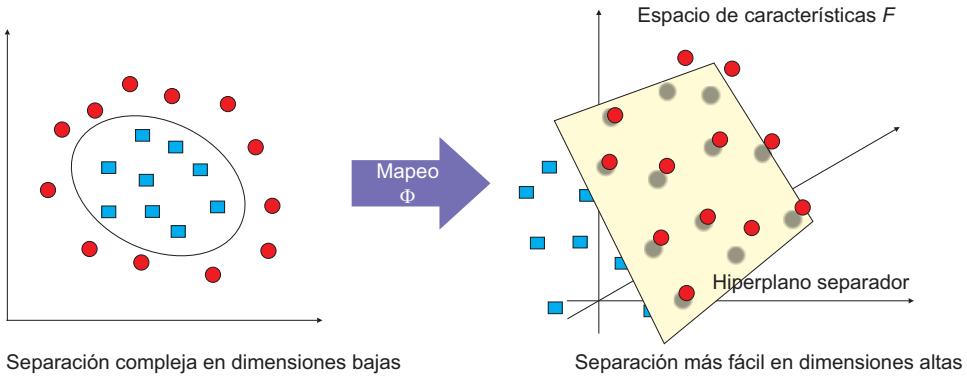


Figura 45: Hiperplano separando puntos de dos clases linealmente en un espacio de características de mayor dimensión

denominado *feature space* mediante un mapeo no lineal.

$$\begin{aligned} \Phi : \mathbb{R}^N &\to F \\ x &\mapsto \Phi(x) \end{aligned} \tag{32}$$

Cada dato $x_1, ..., x_d \in \mathbb{R}^d$ se mapea a un espacio de características F con una dimension mucho mayor. Ahora se trabaja con el problema de aprendizaje en el espacio F en vez de $\mathbb{R}^d$, es decir, ahora se trabaja con la muestra:

$$(\Phi(x_1), y_1), ..., (\Phi(x_d), y_d) \in F \times Y \tag{33}$$

En la figura 45 se ilustra un ejemplo de datos de dos dimensiones que no se pueden separar, sin embargo al ser mapeados a una tercera dimensión entonces si es posible encontrar el hiperplano separador lineal. El mapeo queda expresado de la siguiente manera.

$$\begin{aligned} \Phi : \mathbb{R}^2 &\to \mathbb{R}^3 \\ (x_1, x_2) \mapsto (z_1, z_2, z_3) &:= (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{aligned} \tag{34}$$

En este ejemplo, $F \in \mathbb{R}^3$, pero si la dimensión es mucho más grande entonces la operación producto punto es costosa de calcular. Además, el concepto estadístico de *maldición de la dimensionalidad*, que indica que la estimación de un problema incremen-

ta drásticamente con la dimensión N del espacio, introduce dudas respecto a la idea de transformar los datos a un espacio de características de mayor dimensión para hacer el aprendizaje. Sin embargo, la teoría de aprendizaje estadístico indica que el aprendizaje en F puede ser más simple si se usa baja complejidad, es decir, decisiones de clasificación simples, tales como los clasificadores lineales. Para poder tener un función de clasificación eficiente se utiliza el tructo del kernel.

En vez de mapear cada uno de los datos al nuevo espacio de mayor dimensión, lo cual puede resultar muy caro computacionalmente (Hearst, 1998), el truco del kernel permite obtener el producto punto entre dos vectores que se supone que ya estan mapeados a F pero sin necesidad de haber hecho las operaciones de mapeo.

$$k(\mathbf{x}, \mathbf{y}) := (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})) \tag{35}$$

Ilustremos la ventaja de usar el truco de kernel usando un ejemplo de kernel polinomial, el cual se representa como:

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d \tag{36}$$

Si reformulamos el calculo de producto punto entre dos vectores en el espacio de características en términos de una función kernel k.

$$\begin{aligned}
(x_1, x_2) \mapsto (z_1, z_2, z_3) &:= (x_1^2, \sqrt{2}x_1x_2, x_2^2) \\
(\Phi(x), \Phi(y)) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)(y_1^2, \sqrt{2}y_1y_2, y_2^2) & \\
= ((x_1, x_2)(y_1, y_2)^\top)^2 & \\
= (\mathbf{x} \cdot \mathbf{y})^2 & \\
=: \mathrm{k}(\mathbf{x}, \mathbf{y}) &
\end{aligned} \tag{37}$$

Existen varios kernels, los cuales se pueden utilizar en el mapeo, en la siguiente tabla se presentan sus fórmulas.

Gaussian RBF	$\mathrm{k}(\mathbf{x},\mathbf{y}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{c}\right)$
Polinomial	$((\mathbf{x}\cdot\mathbf{y})+\Theta)^d$
Sigmoindal	$\tanh(k(\mathbf{x}\cdot\mathbf{y})+\Theta)$
Multicuadrático inverso	$\frac{1}{\sqrt{\|\mathbf{x}-\mathbf{y}\|^2+c^2}}$

Ahora se tienen las herramientas para construir clasificadores no lineales, para ello se sustituye $\Phi(x_i)$ por cada punto de entrenamiento x_i y se obtiene el hiperplano óptimo en F. Debido a que se utilizan kernels, la función de decisión no lineal tiene la siguiente forma

$$f(\mathbf{x}) = sign(\sum_{i=1}^{L} v_i \cdot \mathrm{k}(\mathbf{x},\mathbf{x}_i) + b) \tag{38}$$

cuyos parámetros v_i se pueden solucionar con el problema de programación cuadrática.

Modelos Ocultos de Markov

Definición Es una técnica estadística que modela datos temporales y secuenciales que se pueden utilizar para caracterizar las propiedades estadísticas de una señal. Esta técnica tiene aplicaciones en el área de reconocimiento de patrones. El sistema a ser modelado se supone como un proceso de Markov con estados no observables (ocultos).

Proceso de Markov Un proceso de Markov discreto es un proceso estocástico que satisface la propiedad de Markov, esta propiedad indica que se pueden hacer predicciones del futuro basándose solamente en el estado presente sin necesidad de conocer la historia completa de todo el proceso.

Para ejemplificar un proceso de Markov, supóngase una serie temporal la cual dado un tiempo t puede tomar algún estado de un conjunto finito de N estados distintos $S_1, S_2, ..., S_N$ como se ilustra en la figura 46. Al estado que ocurre en el tiempo t se le conoce como q_t.

Para describir un nuevo estado mediante una descripción completa de la secuencia temporal, se requiere que además de conocer el estado actual, se tenga información de

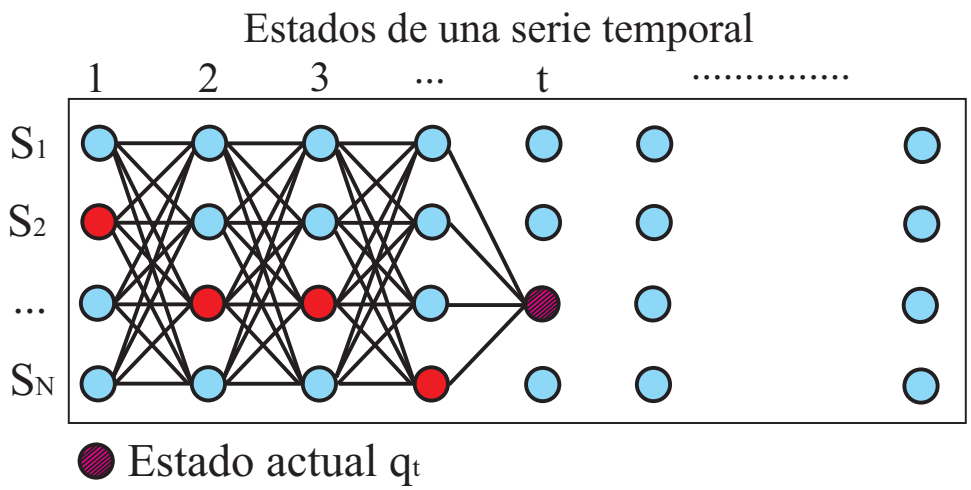


Figura 46: Proceso de Markov representando una serie temporal.

todos los estados anteriores.

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...\right] \tag{39}$$

Sin embargo,un proceso de Markov discreto de primer orden indica que esta descripción se puede truncar y estar completamente determinada considerando solo el estado actual y el estado inmediato anterior

$$P\left[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, ...\right] = P\left[q_t = S_j | q_{t-1} = S_i\right] \tag{40}$$

Un modelo de Markov contiene las probabilidades de cambiar entre los estados, las cuales se consideran constantes en el tiempo. Una forma de representar estas probabilidades es mediante la matriz de transición de probabilidades A; cada elemento a_{ij} indica la probabilidad de pasar del estado S_i al estado S_j. La figura 47 ilustra un ejemplo.

Cada estado indica un evento observable físico, por ejemplo el estado del clima (nublado, soleado, lluvioso) o elementos de un diccionario (A,B,C,D). Cuando se ha observado una serie de estados consecutivos, entonces se tiene una secuencia de observación O, por ejemplo $O = \{S_2, S_1, S1, S_3\}$ correspondientes a $t = 1, 2, 3, 4$. La probabilidad de una observación particular O dada un modelo se expresa:

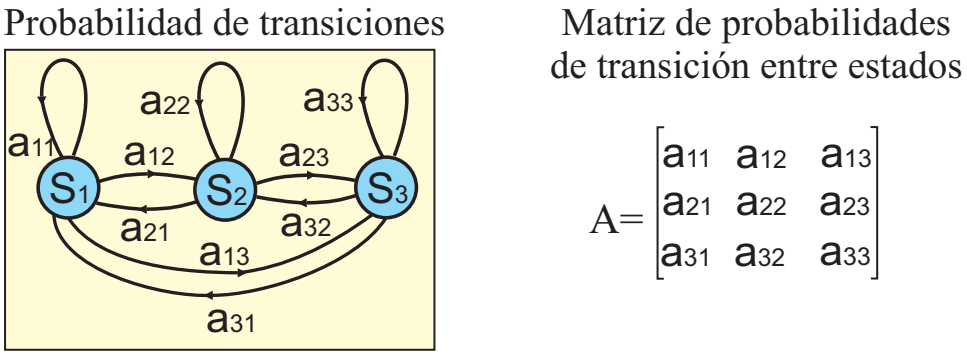


Figura 47: Probabilidades de transición entre estados en un modelo de Markov.

$$P(O|\text{Modelo}) = P[S_2, S_1, S1, S_3|\text{Modelo}] \tag{41}$$

Hasta este momento, se ha descrito un modelo de Markov observable, ya que los estados son directamente observables en cada instante de tiempo y los únicos parámetros son las probabilidades de transición, tal como se muestra en la figura 13. A continuación se explican los modelos ocultos de Markov.

8.2.0.1. Modelos ocultos de markov discretos

En los modelos ocultos de Markov los estados se encuentran ocultos mientras lo visible corresponde a lo que se conoce como las emisiones de los estados. El ejemplo clásico para explicar HMM discretos es el modelo de la urna y las pelotas, (Rabiner, 1989). Se supone que existen N urnas las cuales no son visibles para un observador. Cada una de las urnas guarda cierta cantidad de pelotas coloredas con M colores distintos. Un genio tiene acceso a las urnas y elige una de acuerdo un proceso aleatorio de selección de urnas. Posteriormente extrae una pelota de dicha urna y le señala al observador el color de la pelota. Después regresa la pelota a la urna de la que fue seleccionada. A continuación selecciona nuevamente una urna de acuerdo al proceso aleatorio de selección de urnas y vuelve a extraer una pelota. El proceso se repite y produce como resultado una observación finita de colores. La elección de las urnas no depende de las urnas que fueron elegidas con anterioridad, a esto se le conoce como proceso de Markov, sin embargo la eleccion de las urnas esta oculta al observador y lo único que puede ser observado son los símbolos emitidos por las urnas, es decir, los colores.

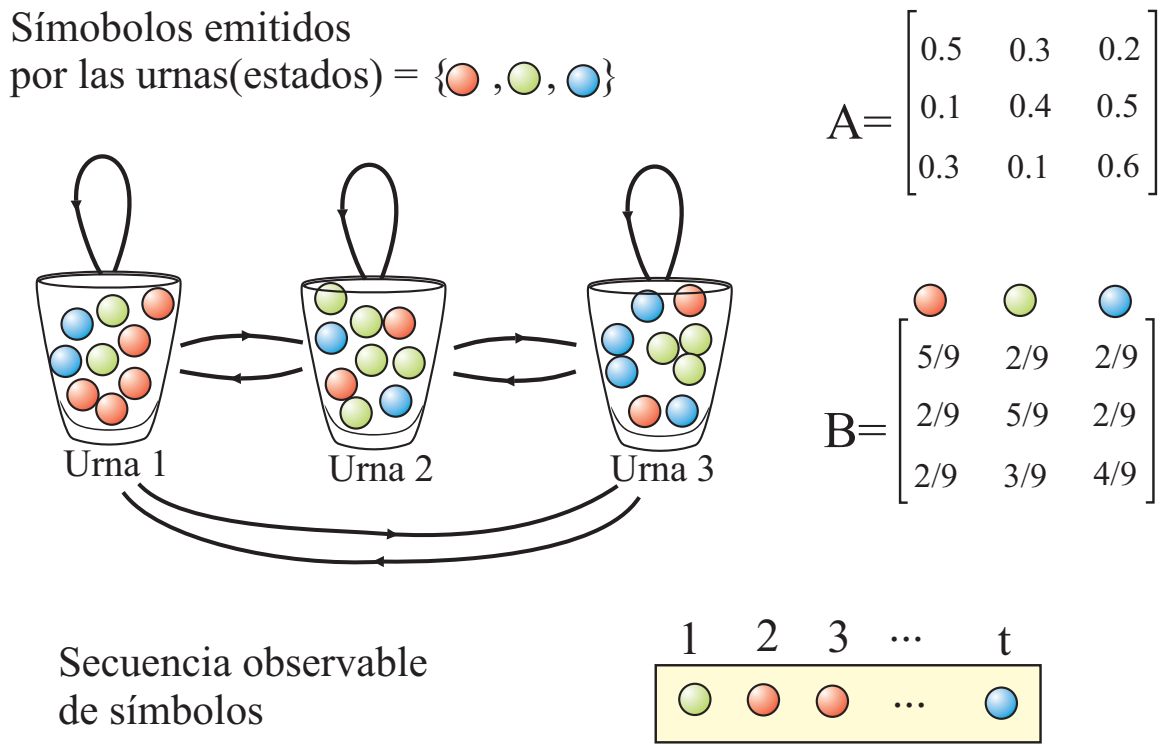


Figura 48: Elementos de los modelos ocultos de Markov.

En la figura 48 se ilustran N=3 urnas que contienen pelotas con M=3 colores distintos (rojo, verde y azul). Las urnas son los estados de un proceso de Markov con su respectiva matriz de probabilidades de transición A. En los HMMs Se introducen los valores $B = \{b_j(k)\}$ que indican la distribución de probabilidad de los símbolos observados en cada estado, para este ejemplo clásico, la matrix B indica la probabilidad de encontrar un color especifico en una urna dada. La secuencia temporal observable $O = O_1 O_2 ... O_T$ contiene los símbolos o colores emitidos por los estados.

Los elementos que conforman un modelo oculto de Markov son los siguientes:

1. N, el numero de estados en el modelo $S_1, S_2, ..., S_N$.
2. M, la cantidad de símbolos observados por estado.
3. A, la matriz de distribución de probabilidad de transiciones entre estados.
4. B, la matriz de distribución de probabilidad de observaciones de los símbolos en cada estado.
5. π, una distribución inicial de los estados.

Un HMM queda totalmente caracterizados por $\lambda = (A, B, \pi)$

Para la aplicación practica de los HMMs, existen tres problemas algorítmicos los cuales se explican a continuación.

Problema de evaluación. Dada una secuencia de observación $O = O_1 O_2 ... O_T$ y un modelo particular λ, se requiere calcular $P(O|\lambda)$. Este problema se resuelve con un algoritmo conocido como *Forward-Backward Procedure*. La solución de este problema es útil cuando existen modelos para cada clase previamente entrenados y se quiere aquella clase con la cual una observación tiene mayor similitud.

Problema de decodificación. Calcular la secuencia de estados optima que mejor explica una secuencia de observaciones. Este problema se resuelve con el algoritmo *Viterbi*. La obtención de la secuencia optima se puede usar para aprender cual es la estructura del modelo u obtener estadísticas acerca de los estados individuales.

Problema de aprendizaje. Ajustar los parámetros de un modelo para maximizar $P(O|\lambda)$. Para estimar la probabilidad maxima de la observación dado un modelo se utiliza un algoritmo iterativo llamado *Baum-Welch* o su equivalente *Expectation-Modification (EM)*. El ajuste de los parámetros del modelo se utiliza en la etapa de entrenamiento en donde se generan modelos para las secuencias temporales de entrenamiento provenientes de cada clase.

En el caso de los datos continuos, cuyos valores pueden ser infinitos, se lleva a cabo una quantization vectorial. Para ello, se elige un diccionario de símbolos y cada dato continuo se compara contra todos los símbolos, el valor continuo toma el símbolo con el que resulta mas cercano.

Solución al problema de evaluación Se desea calcular la probabilidad de que una secuencia de observación $O = O_1 O_2 ... O_T$ es generada por un modelo λ, es decir:

$$P(O|\lambda) = \sum_{toda\ Q} P(O|Q, \lambda) P(Q|\lambda) \tag{42}$$

Donde Q indica una secuencia particular $Q = q_1, q_2, ..., q_T$ de todas las posibles secuencias de longitud T existentes.

El primer término de la ecuación **??** indica que se debe calcular la probabilidad de O dada cada una de las posibles Q.

$$P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda) \tag{43}$$

Al asumir independencia estadística en las observaciones entonces se tiene:

$$P(O|Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdot \cdot \cdot b_{q_T}(O_T) \tag{44}$$

Que indica el producto de las densidades de probabilidad de las observaciones dada una secuencia Q.

El segundo término de la ecuación 42 indica la probabilidad de una secuencia de estados específica Q dado el modelo λ. Se puede rescribir de la siguiente manera:

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} ... a_{q_{T-1} q_T} \tag{45}$$

Que indica el producto de las probabilidades de transición entre estados de la secuencia Q

La complejidad computacional del algoritmo descrito arriba es de orden $O(N^T T)$. Debido a que por cada tiempo $t = 1, 2, ..., T$ existen N estados posibles, el calculo de operaciones no es factible incluso para valores pequeños, por ejemplo, para N=5 y T=100, se requieren $5^{100} \cdot 100 \approx 7.8E^{71}$ cálculos.

Afortunadamente existe un algoritmo llamado procedimiento de avance o *Forward-Bakcward procedure* que permite resolver el problema de evaluación en un tiempo factible para aplicaciones prácticas.

Algoritmo de avance (Forward) Considérese la variable de avance $\alpha_t(i)$ definida como

$$\alpha_t(i) = P(O_1\ O_2\ ...\ O_t, q_t = S_i|\lambda) \tag{46}$$

donde se tiene la secuencia $O_1\ O_2\ ...\ O_t$ que corresponde a una observación parcial (hasta un tiempo t) y un estado S_i en un tiempo t dado un modelo λ. Para resolver $\alpha_t(i)$ de forma inductiva se sigue el siguiente algoritmo.

1. Inicialización:
   $\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \leq i \leq N$
2. Inducción:
   $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \quad 1 \leq t \leq T-1 \quad 1 \leq j \leq N$
3. Terminación:
   $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Para calcular $\alpha_t(j)$, $1 \leq \quad t \quad \leq T$, $1 \leq \quad j \quad \leq N$ se requieren $N(N+1)(T-1)+N$ multiplicaciones y $N(N+1)(T-1)$ sumas, por lo tanto, el orden del algoritmo de avance es N^2T en vez de TN^T que tomaría el calculo directo. Para N=5 y T=100, se requieren $\approx$ 3000 calculos, significativamente menos a los $7.8E71$ necesarios con la forma directa.

Algoritmo de retroceso (Backward) Para el problema de evaluación solo necesita la parte de anvace del algoritmo forware-backware. Se introduce el algoritmo de retroceso ya que es útil para resolver el problema de aprendizaje.

De igual manera, se introduce la variable de retroceso:

$$\beta_t(i) = P(O_{t+1}\ O_{t+2}\ ...\ O_T, q_t = S_i|\lambda) \tag{47}$$

que representa la probabilidad de la secuencia de observación parcial desde $t+1$ hasta T dado un estado j en el tiempo t y el modelo λ. Resolviendo de forma inductiva, el

algoritmo es el siguiente:

1. Inicialización:
   $\beta_T(i) = 1, \quad 1 \leq i \leq N$

2. Inducción:
   $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad t = T-1, T-2, ..., 1, \quad 1 \leq i \leq N$

La cantidad de operaciones necesarias para $\beta_t(i)$ $1 \leq t \leq T$, $1 \leq i \leq N$ es del orden de $N^2 T$.

Solución al problema de decodificación Se desea obtener la ruta óptima de secuencias de estado $q = (q_1 q_2 ... q_T)$ asociada con una secuencia de observaciones $O = O_1\ O_2\ ...\ O_T$ dado un modelo λ. La ruta óptima se puede obtener usando el algoritmo de Viterbi. Para esto, se define la siguiente cantidad:

$$\delta_t(i) = \max_{q_1, q_2, ..., q_{t-1}} P[q_1, q_2, ..., q_{t-1}, q_t = i, O_1 O_2 ... O_t | \lambda] \tag{48}$$

que se refiere a la probabilidad más alta a lo largo de un camino simple que toma en cuenta las primeras t observaciones y finaliza en el estado S_j. Por inducción se tiene

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij} \right] \cdot b_j(O_{t+1}) \tag{49}$$

Para obtener la secuencia de estados, es necesario conservar el argumento que maximiza la ecuación 49 para cada t y j; esto se logra con el arreglo $\psi_t(j)$. El procedimiento para encontrar la mejor secuencia de estados se establece de la siguiente manera:

1. Inicialización:
   $\delta_1(i) = \pi_i b_i(O_i), \quad 1 \leq i \leq N \qquad \psi_1(i) = 0$

2. Recursion:

$$\delta_t(j) = \max_{1\leq i\leq N} \left[\delta_{t-1}(i)a_{ij}\right] \cdot b_j(O_t), \quad 2 \leq t \leq T \quad 1 \leq j \leq N$$
$$\psi_t(j) = \operatorname*{argmax}_{1\leq i\leq N} \left[\delta_{t-1}(i)a_{ij}\right], \quad 2 \leq t \leq T \quad 1 \leq j \leq N$$

3. Terminación: $p^* = \max_{1\leq i\leq N} [\delta_T(i)]$
$q_T^* = \operatorname*{argmax}_{1\leq i\leq N} [\delta_T(i)]$

4. Ruta inversa: $q_t^* = \psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-1, ..., 1$

El algoritmo de Viterbi esta basado en métodos de programación dinámica. La complejidad de este algoritmo es de orden $O(N^2T)$.

Solución al problema de aprendizaje La solución de este problema consiste en encontrar el conjunto de parámetros $\lambda(A, B, \pi)$ que describen un modelo a partir de secuencias de observación que provienen de una misma clase. No hay una forma analítica de solucionar este problema, sino que se pueden estimar parámetros y encontrar aquellos que maximicen localmente la probabilidad de la secuencia de observación dado el modelo, es decir $P(O|\lambda)$. Esto se puede lograr mediante un algoritmo iterativo llamado Expectation-Maximization (EM) o su equivalente, el algoritmo Baum-Welch. La técnica consiste en inicializar los parámetros con un estimado y después ir actualizando los valores de forma iterativa hasta que se satisfaga una condición de terminación. En la sección anterior, se definieron las variables de avance y retroceso $\alpha_t(i)$ y $\beta_t(i)$. Para este problema, se define además las variable $\xi_t(i, j)$ y $\gamma_t(i)$ las cuales quedan expresadas en términos de las variables de avance y retroceso:

$$\begin{aligned} \xi_t(i,j) &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \end{aligned} \tag{50}$$

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)} \tag{51}$$

donde esta última representa la probabilidad de estar en el estado S_i en el tiempo t, dada una secuencia de observación O y un modelo λ.

Es posible establecer una relación entre $\gamma_t(i)$ y $\xi_t(i,j)$ mediante sumas en j, dando lo siguiente:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \tag{52}$$

A continuación se describen las fórmulas para re-estimar los parámetros π, A y B de un HMM.

$$\begin{aligned} \bar{\pi} &= \gamma_1(i) \\ \bar{a}_{ij} &= \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\ b_j\bar{(k)} &= \frac{\sum_{\substack{t=1 \\ O_t=vk}}^{T} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)} \end{aligned} \tag{53}$$

El algoritmo para el aprendizaje queda de la siguiente manera, donde ϵ es una condición de convergencia para terminar las iteraciones.

1. Inicializar a_{ij} y b_{jk}
2. t=t+1
3. Calcular $\bar{\pi}$, $\bar{a}_{ij}$ y $b_j\bar{(k)}$ como se muestra en la ecuación 53
4. $a_{ij} = \bar{a}_{ij}$ y $b_{jk} = b_j\bar{(k)}$
5. Realizar los pasos 2-4 hasta que $\text{argmax}_{i,j,k} \left[a_{ij}(t) - a_{ij}(t-1), b_{jk}(t) - b_{jk}(t-1)\right] < \epsilon$
6. Regresar $a_{ij}(t)$ y $b_{jk}(t)$

Los valores iniciales de a_{ij} y b_{jk} son críticos para la re-estimación; diferentes valores iniciales pueden llevar diferentes resultados ya que el algoritmo no garantiza converger a un mínimo global.

8.2.0.2. Modelos ocultos de markov continuos

Hasta este punto, se ha considerado el caso cuando las observaciones son símbolos discretos provenientes de un alfabeto finito de forma que es posible usar densidades de probabilidad discretas. Sin embargo, para muchas de las aplicaciones, las observaciones son continuas, por ejemplo, señales o vectores. Una forma de resolver esto, es mediante codebooks, pero tienen la desventaja que puede introducir degradaciones importantes asociadas a la cuantización. Otra forma de tratar este problema es reemplazar las probabilidades de observación discreta $b_j(k)$ por funciones de densidad de probabilidad continua $b_j(O)$, donde O es el vector a ser modelado que posee dimension d. Una práctica común es representar $b_j(O)$ como una mezcla de gaussianas.

$$b_j(o) = \sum_{m=1}^{M} c_{jm} \mathfrak{N}(o, \mu_{jm}, \Sigma_{jm}) \qquad 1 \leq j \leq S \tag{54}$$

donde $\mathfrak{N}$ denota una función de densidad de probabilidad Gaussiana, con vector de medias μ_{jm} de dimension d y matriz de covarianza Σ de dimensión dXd. Aquí, c_{jm} son los coeficientes de la m_{th} mezcla en el estado j. La distribución Gaussiana multi-variada $\mathfrak{N}(o, \mu_{jm}, \Sigma_{jm})$ de cada componente esta definida de la siguiente manera:

$$\mathfrak{N}(o, \mu_{jm}, \Sigma_{jm}) = \frac{1}{\sqrt{(2\pi)^K \, |\Sigma_{jm}|}} exp\left[-\frac{1}{2}(o - \mu_{jm})^T {\Sigma_{jm}}^{-1} (o - \mu_{jm}) \right] \tag{55}$$

Glosario de términos

Amplitud La amplitud se refiere al tamaño de las ondas acústicas y determina que tan fuerte es el sonido. La percepción subjetiva de la amplitud del sonido se conoce como volumen y debido a que el rango auditivo del ser humano es muy grande se utiliza una escala logarítmica en decibeles para hacer las mediciones de volumen.

Burn out El síndrome burnout se describe como una respuesta prolongada al estrés crónico emocional e interpersonal debido a el trabajo, determinada por dimensiones de cansancio, cinismo e ineficacia (Maslach *et al.*, 2001).

Clasificación supervisada En este tipo de clasificación se tiene un conocimiento a priori acerca de las class que se utilizan para el entrenamiento.

Clasificación no supervisada En este tipo de clasificación no se tiene un conocimiento a priori acerca de las class que se utilizan para el entrenamiento.

Decibel El decibel es una unidad logarítmica sin dimension que se usa para comparar la relación entre dos cantidades, por ejemplo la relación entre dos potencias, dos intensidades o entre dos voltajes. Matemáticamente, es 10 veces el logaritmo de relación entre las potencias de dos señales. $dB = 10log(P_1/P_0)$

donde P_0 se refiere a un valor de potencia de referencia que es comparado contra otro valor P_1.

Difuso (Fuzzy) Un conjunto se puede definir mediante la función indicadora. Si la función indicadora vale 1, el objeto está en el conjunto; si vale 0, no está en el conjunto. El conjunto de técnicas borrosas (fuzzy) están basados en la generalización de la función indicadora; un objeto pertenece en principio a todos los conjuntos, con un grado mayor o menor de pertenencia.

Digitalizacion de la señal El sonido que encontramos en la vida real es continuo en tiempo y en amplitud. Para poder procesar el sonido en microprocesadores, es necesario convertir la señal analógica a un formato digital, lo cual se logra mediante un procedimiento llamado digitalizacion el cual se compone de dos pasos: muestreo y cuantización (libro L Tan).

Dominio del tiempo y frecuencia Las señales sonoras en el dominio del tiempo contienen la amplitud de la señal en cada instante de tiempo muestreado. Sin embargo, existen aplicaciones que requieren descomponer la señal en tonos puros u ondas sinoidales. Este tipo de información se encuentra en una representación de la señal conocida como el dominio de la frecuencia. En la figura se ilustra una señal en el dominio del tiempo con sus componentes sinoidales, así como una ilustración de la señal en el dominio de la frecuencia que representa con que cantidad ciertos componentes de frecuencia aportan a la formación de la señal completa.

Energía La energía para una señal de duración finita x(n) esta definida como

$$E = \sum_{n=1}^{N} |x(n)|^2 \tag{56}$$

Entropía En el área de teoría de la información, la entropía de una señal mide que tan impredecible es dicha señal (Shannon, 2001). La formula para obtener la entropía de una variable aleatoria x dada una distribución de probabilidades $P = p_1, p_2, ..., p_n$ es la siguiente:

$$H(X) = -\sum_{i}^{n} P(x_i) log_b P(x_i) \tag{57}$$

donde b es la base del logaritmo.

La entropía mínima se obtiene cuando una señal es un valor constante k, en este caso su función de densidad de probabilidad es un impulso unitario en k, es decir; $pi = k$, lo cual produce una entropía de cero según la formula 58. En el caso opuesto, si la señal tiene una distribución de probabilidad uniforme, entonces su entropía es máxima, es decir,

si $pi = 1/n$ para todos los valores de n posibles, la entropía resultante se muestra en la fórmula 59.

$$H_{min} = -\sum_i \delta(k) log_b[\delta(k)] = -log_b(1) = 0 \tag{58}$$

$$H_{max} = -\sum_i \frac{1}{n} log_b \frac{1}{n} = -log_b(\frac{1}{n}) = log_b(n) \tag{59}$$

Escala MEL En 1937, Stevens, Volkman y Newman obtuvieron una escala perceptual para el tono llamada escala MEL (Stevens *et al.*, 1937). Para ello realizaron un trabajo de investigación en donde cinco participantes fraccionaron subjetivamente diversos tonos a varias frecuencias. La relación de esta escala subjetiva con la frecuencia se define asignando al tono de 1000Hz un tono de 1000 unidades mels. En la actualidad existen varias fórmulas para la escala MEL que han surgido de distintas tablas y curvas obtenidas en varios experimentos (Lindsay and Norman, 1977) (Beranek, 1949), aunque la más popular corresponde a la ecuación 60 (O'Shaughnessy, 1999).

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right) \tag{60}$$

donde f corresponde a la frecuencia en Hertz y m a las unidades mel.

Escala Bark La escala de frecuencia Bark, propuesta por Eberhard Zwicker en 1961, es una escala psicoacústica que esta definida de forma que las bandas críticas de la audición humana corresponden a un bark de ancho (Zwicker, 1961). Para convertir frecuencia en Hertz a unidades en Bark se utiliza la ecuación 61 (Traunmüller, 1990).

$$z = 13\, arctan(0.00076 f) + 3.5\, arctan((\frac{f}{7500})^2) \tag{61}$$

donde f es la frecuencia en Hertz y z es la frecuencia en Barks.

Espectrograma Una forma de mejorar la resolución temporal consiste en lo que se conoce como Transformada de Fourier en Tiempo Corto (STFT). La STFT consiste en dividir la señal en pedazos o tramas, que además se pueden traslapar unos con otros con el objetivo de reducir efectos de frontera. A cada pedazo se le calcula la FFT, obteniendo como resultado una matriz compleja que puede expresarse de la siguiente manera:

$$STFT\{x(n)\} = X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \tag{62}$$

Donde, x[n] es la señal y w[n] es una **función ventana**. Comúnmente se usa una ventana de Hamming, sin embargo existen diferentes tipos de ventanas,

A la magnitud cuadrada de la STFT se le denomina Spectrograma y es una de las formas originales para representar la información auditiva

$$spectrograma\{x(n))\} = |X(m,w))|^2 \tag{63}$$

Fase La fase se refiere a una relación en el tiempo entre dos o mas ondas acústicas en un punto especifico en sus ciclos. Si dos ondas periódicas poseen la misma frecuencia y empiezan al mismo tiempo, entonces ambas ondas se encuentra en fase. Si dos ondas idénticas empiezan en tiempos distintos entonces se encuentran fuera de fase con cierto grado. La fase se mide en grados desde 0 hasta 360, donde 0 grados significa sincronía y 180 grados implica que ambas señales son exactamente opuestas. Cuando dos sonidos que están en fase se mezclan, la combinación produce un sonido mas fuerte. Cuando dos sonidos fuera de fase se mezclan entonces se produce un sonido pequeño o ausencia de sonido. A esto se le conoce como cancelación de fase.

Frecuencia Es la tasa o numero de veces por segundo que una onda periódica completa un ciclo. La frecuencia se mide en ciclos por segundo o en Hertz. Para que el sonido sea audible por el ser humano, la frecuencia de las vibraciones debe estar dentro del espectro audible que corresponde a las ondas comprendidas entre los 20Hz y 20Khz.

Función ventana Es una función matemática que vale cero fuera de un intervalo dado. Un ejemplo de función ventana rectangular consiste en un valor constate dentro de un intervalo y cero fuera de ese intervalo. Al multiplicar una función ventana con una señal permite observar a la señal por un tiempo finito. Esta es la primer subsección

Ground truth Conjunto etiquetado de clasificaciones confiables para entrenamiento.

Leakage El Leakage consiste en la formación de nuevos componentes de frecuencia en el espectro.

Muestreo Debido a que es imposible digitalizar una infinita cantidad de puntos, se recurre al muestreo que consiste en tomar mediciones o muestras de la señal auditiva en un intervalo fijo de tiempo T. Al número de muestras tomadas por segundo se le conoce como frecuencia de muestreo. $f_s = \frac{1}{T}$ muestras por segundo (HZ). El teorema de Shannon-Nyquist indica que la frecuencia de muestreo debe tener al menos el doble de la frecuencia de la señal original para evitar un fenómeno conocido como Aliasing el cual produce datos incorrectos. Por ejemplo, las señales de voz humana pueden alcanzar frecuencias de hasta 4kHz, por lo que la frecuencia de muestreo mínima debe ser de 8kHz; para muestrear señales de audio con frecuencias hasta 20kHz, se necesita muestrear al menos 40000 veces por segundo.

NMF- Non Negative Matrix Factorization Es un grupo de algoritmos que se utilizan para descomponer datos multivariados donde una matriz no negativa V se factoriza en las matrices no negativas W y H. Estos algoritmos

Pitch De acuerdo con el Instituto Nacional de Estándares Americanos (ANSI) (Institute, 1999), ”el tono o pitch es un atributo auditivo del sonido que permite ordenarlos de más bajos a más altos en una escala. El tono depende principalmente del contenido frecuencial del estimulo auditivo, pero también depende de la presión del sonido y de la forma de onda del estimulo”. Aunque el tono puede cuantificarse en función a la frecuencia, éste no

es puramente una propiedad física sino un atributo psicológico del sonido. Normalmente, el tono de los sonidos complejos se cuantifica como frecuencia en ciclos por segundo o Hertz, mediante una comparación del sonido con formas de onda sinoidales periódicas (Hartmann, 1998).

Potencia La potencia para una señal de duración finita x(n) esta definida como

$$P = \frac{1}{N} \sum_{n=1}^{N} |x(n)|^2 \tag{64}$$

Relación señal ruido o Signal to Noise Ratio (SNR) La relación ruido señal es una especificación que mide el nivel de sonido deseado presente en una señal auditiva comparado contra el nivel de ruido que también esta presenta en la señal.

$$SNR_{dB} = 10log\frac{P_{senal}}{P_{ruido}} \tag{65}$$

donde P es la potencia de la señal.

$$P = \frac{1}{N} \sum_{i=1}^{n} (x_i)^2 \tag{66}$$

Resolución de bits La cuantización convierte a la variable dependiente de su forma continua a discreta. Esto se logra representando la amplitud de las muestras individuales como enteros expresados en forma binaria, asi las muestras se pueden medir en un número finito de niveles discretos. El rango de posibles enteros esta determinado por la profundidad de bits, es decir, el número de bits usados por muestra. Una señal muestreada con 8 bits de profundidad tiene un rango de $2^8 = 256$ enteros posibles, mientras que una muestreada con 16 bits de profundidad tiene $2^{16} = 65536$ lo cual brinda una mejor resolución que se traduce a una mejor calidad de sonido.

Ruido En el dominio de la comunicación, ”el ruido es cualquier cosa que provoca una pérdida de información mientras la información fluye desde la fuente hacia su destino”(Wood, 2012). Puede ser cualquier estimulo externo o ambiental que distrae la recepción de un mensaje enviado por un comunicador (Rothwell, 2009). Por ejemplo, en los trabajos de reconocimiento del habla, el ruido son todos aquellos sonidos externos al parlante de interés, por ejemplo la música de fondo o incluso otras personas hablando. El audio ambiental normalmente era considerado como ruido, sin embargo con el area de estudio Computationally Auditory Scene Analysis (CASA) ha surgido importancia en el ámbito científico y en vez de considerarse como ruido es considerado como información.

Sonido El termino sonido se refiere al fenómeno en el cual ondas acústicas producidas por vibraciones mecánicas se propagan a través de un medio elástico. Cuando la presión de las ondas acústicas produce vibración en el timpano, se envían señales al cerebro que pueden interpretarse como sonido.

La mas elemental de todas las señales acústicas es conocida como tono puro, que se refiere a una onda senoidal, la cual es de suma importancia en la ciencia de la acustica. Aunque es raro encontrar tonos puros en el mundo real, otros sonidos se pueden representar como una combinación de ondas senoidales. La formula de una onda senoidal es:

$$y(x) = A \times seno(2\pi fx - \varphi) \tag{67}$$

donde A es la amplitud de oscilación, f es la frecuencia y φ es la fase inicial.

Transformada del coseno La transformada discreta de coseno (DCT) es similar a la DFT, pero expresando a la señal solo en términos de funciones coseno de forma que solo se tienen valores reales. La fórmula para obtener la DCT es:

$$X_k = \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \; k = 0, ..., N-1. \tag{68}$$

Transformada de Fourier Las señales sonoras en el dominio del tiempo contienen la amplitud de la señal en cada instante de tiempo muestreado. Sin embargo, existen aplicaciones que requieren descomponer la señal en tonos puros u ondas sinoidales. Este tipo de información se encuentra en una representación de la señal conocida como el dominio de la frecuencia. En la figura se ilustra una señal en el dominio del tiempo con sus componentes sinusoidales, así como una ilustración de la señal en el dominio de la frecuencia que representa con que cantidad ciertos componentes de frecuencia aportan a la formación de la señal completa.

El análisis de Fourier es una familia de técnicas matemáticas que se basan en descomponer las señales en ondas sinoidales. Para el caso de señales digitales se utiliza la Transformada Discreta de Fourier cuya formula es:

En el dominio del tiempo, se aplica la DFT a una señal x que consiste de N puntos. Esto produce dos señales en el dominio de la frecuencia, una parte real Re y una parte imaginaria Im. Para obtener la magnitud en el dominio de la frecuencia, se utiliza la fórmula , mientras que para obtener la fase se usa

Existe un método para calcular la DFT conocido como Fast Fourier Transform el cual es sumamente eficiente reduciendo el tiempo computacional, el orden es O(ln(n) (Tan and Jiang, 2013). Este algoritmo ha permitido que muchas aplicaciones sean posibles de realizar en tiempo real.

Mientras que la DFT produce buena resolución en frecuencia, se pierde en su totalidad la información de la señal en el espacio del tiempo. En la siguiente figura se muestra un ejemplo, nótese como ambas señales constan de una onda sinoidal a 60Hz y otra onda de 20Hz. La DFT brinda la información en frecuencia pero no indica en que tiempo existen estos componentes de frecuencia.

Una forma de mejorar la resolución temporal consiste en lo que se conoce como Transformada de Fourier en Tiempo Corto (STFT). La STFT consiste en dividir la señal en pedazos o tramas, que además se pueden traslapar unos con otros con el objetivo de reducir efectos de frontera. A cada pedazo se le calcula la FFT, obteniendo como resultado una matriz compleja que puede expresarse de la siguiente manera:

$$STFT\{x(n)\} = X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \quad (69)$$

Donde, x[n] es la señal y w[n] es una **función ventana**. Comúnmente se usa una ventana de Hamming, sin embargo existen diferentes tipos de ventanas,

A la magnitud cuadrada de la STFT se le denomina espectrograma y es una de las formas originales para representar la información auditiva

$$espectrograma\{x(n))\} = |X(m,w))|^2 \quad (70)$$

Glosario de Acrónimos

Acrónimo	Significado
AAL	Ambient Assisted Living
AiP	Aging in Place
CDA	Comportamientos Disruptivos Audibles
CPU	Central Processor Unit
DAV	Detector de Actividad de Voz
dB	Decibel
FFT	Fast Fourier Transform
FN	Falsos Negativos
FP	Falsos Positivos
GT	Ground Truth
HMM	Hidden Markov Models
Hz	Hertz
IACM	Intervenciones de Agitaciones Cohen-Mansfield
INF	Intervenciones No Farmacológicas
MBSE	Multi Band Spectral Entropy
MFCC	Mel Frequency Cepstral Coefficients
NMF	Non Negative Matrix Factorization
PC	Personal Computer
PcD	Paciente con Demencia
RAM	Random Access Memory
RFID	Radio Frequency IDentification
RWCP	Real World Computing Partnership
SDPC	Síntomas de Demencia Psicológicos y de Comportamiento
SIA	Sistema de Intervenciones Asistidas
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
TES	Time Entropy Signature
VN	Verdaderos Negativos
VP	Verdaderos Positivos