DETECCIÓN E IDENTIFICACIÓN DE SEÑALES SONORAS EN ENTORNOS ASISTIVOS

Héctor Lozano Peiteado

Directora: Inmaculada Hernáez Rioja
Departamento de Ingeniería de Comunicaciones
Universidad del País Vasco / Euskal Herriko Unibertsitatea

Bilbao, 2015

Resumen

El trabajo desarrollado en este documento de Tesis Doctoral tiene como principal objetivo el estudio y aplicabilidad de técnicas de reconocimiento de sonidos no relacionados con el habla, denominados sonidos *no-habla* tales como timbres de puerta, grifos abiertos, despertadores, etc., que ayuden a mejorar la independencia y calidad de vida de las personas con discapacidad auditiva.

Para conocer el interés de estas técnicas, como punto de partida, se presenta un análisis detallado del colectivo de personas con discapacidad, centrándose más concretamente en aquellas con problemas auditivos. En este análisis se aportan datos demográficos de los años pasados y estimaciones futuras de su evolución. Además, se analizan los productos de apoyo existentes en el mercado, su consumo y las necesidades no cubiertas todavía en el sector de las tecnologías de apoyo, demostrando la importancia que supone la investigación en técnicas de reconocimiento de sonidos no-habla.

Esta tesis utiliza Modelos de Mezclas Gaussianas (GMM) junto con parámetros extraídos de la señal (Mel Frequency Cepstral Coefficients, Zero Crossing Rate, Roll-Off Point y Spectral Centroid) y los valida frente a varias bases de datos, algunas más frecuentes en la literatura científica (RWCP y CHIL) y otras propias, creadas específicamente para la problemática encontrada asociada al colectivo de personas con discapacidad auditiva. Estas bases de datos propias han sido creadas en base al análisis de necesidades del colectivo objeto, realizado a través de encuestas que se han llevado a cabo en esta tesis para establecer los principales sonidos de interés así como las pautas de diseño que un sistema de estas características debe tener.

En esta investigación se han desarrollado sistemas de reconocimiento capaces de trabajar en tiempo real utilizando micrófonos profesionales con una localización fija. Estos sistemas han sido diseñados tanto para avisar a las personas con problemas auditivos de sonidos de interés como para su uso en sistemas inteligentes que utilicen esta información para el reconocimiento de actividades de la vida diaria de la persona.

No obstante, la principal contribución de esta tesis reside en la investigación de este tipo de sistemas en teléfonos móviles donde las prestaciones hardware están más limitadas y las condiciones de entrenamiento de los sonidos y las de validación o testeo varían. Se ha demostrado cómo optimizando los algoritmos de detección y clasificación, estos sistemas pueden ser funcionales en dispositivos móviles en tiempo real. El trabajo en este campo ha derivado en el desarrollo de una aplicación funcional para teléfonos móviles, capaz de funcionar en tiempo real y diseñada en base a pautas de accesibilidad para el apoyo de personas con discapacidad auditiva. En esta

aplicación se ha evitado la creación de una base de datos universal (por la imposibilidad de recoger todos los diferentes timbres, alarmas,… existentes en el mercado), permitiendo al usuario que sea él mismo quien grabe sus propios sonidos que el sistema utilizará para crear los modelos necesarios para el reconocimiento de eventos acústicos del hogar.

Agradecimientos

Una de mis grandes amigas me contó una vez que el apartado de agradecimientos es, sin duda, uno de los más importantes de la tesis. "*Seguramente*", me dijo, "*sea el único que la gran mayoría de personas lea con gran detenimiento*". Pensar en esta frase hace que el contenido aquí escrito tome un sentido mágico porque, por un momento, puedes dar a conocer a la gente que lo lea las grandes personas con las que te has rodeado y de las que has crecido y aprendido durante estos años de trabajo.

Es por ello que no puedo por menos empezar de otra manera que no sea dando las gracias a mi tutora Inma Hernáez, la persona que me ha guiado, me ha empujado en los momentos difíciles y ha tenido tantísima paciencia conmigo. No sólo me ha apoyado con su incalculable experiencia sino que siempre me ha mostrado su faceta más humana. Me siento muy afortunado de haberla tenido conmigo todo este tiempo ofreciéndome no sólo sus conocimientos sino también su amistad. En especial a ti: ¡gracias!

También me gustaría tener unas palabras para la que ha sido mi supervisora todos estos años en Tecnalia, Igone Idigoras. Ha sido un placer tenerla como jefa y agradezco enormemente sus esfuerzos para que esta investigación haya salido adelante. Sé que estos últimos años han sido difíciles para ti y que posiblemente cuando defienda esta tesis estarás emprendiendo una nueva aventura fuera de Tecnalia. Por eso quiero aprovechar estas líneas para agradecerte todo tu trabajo, no sólo conmigo sino con todas las personas a las que nos has tenido a tu cargo o con las que has trabajado o compartido tantas experiencias. De verdad, eres todo un ejemplo de profesionalidad y de humanidad: No cambies.

En tercer lugar no me puedo olvidar de mi querida y gran amiga Andrea. Cuántas veces habré oído esa frase de "*venga Héctor, tienes que acabar la tesis ya…*" Sabes que tus sermones, aunque muchas veces no lo pareciera, han sido un empuje fundamental para mí. Incluso ahora que estás en Australia siempre me mandas algún mensaje de vez en cuando de "*¿qué tal la tesis?...*" Cuando regreses vas a tener que ser tú la que reciba sermones por no practicar con la guitarra... ¿o además de gurú en fotónica volverás convertida en una auténtica Paco de Lucía?

A Tecnalia y a su gente, gracias por permitirme crecer como persona y como profesional. En especial a la gente de la división de Salud como Lara, Alfonso, Javi C., Irati, Elisa, Javi A., Gabi, Sergio, Arantxa, Carmen, Alberto, Leire M., Javi G., Ricardo, Elías, Ander, Leire Z., Manu, Jesús,…, y a la gente no nombrada aquí que con sus aportes han hecho que esta tesis avance: Artzai, Itsaso, Aida, Arantxa, Patricia, Borja, Diego, Aida.

Gracias también al grupo de procesamiento de señal Aholab de la UPV: Eva, Dani, David, Luis, Agustín, Igor, Iñaki, Igor, Jon por su gran amabilidad y aporte de conocimientos científicos. Da gusto conocer a tanta gente con tan grandes capacidades técnicas que no dudan ni un momento en ayudar cuando alguien lo necesita.

Finalmente, gracias a mis padres, hermano y demás amigos por su continuo apoyo. Sé que sin todos ellos esta tesis no existiría.

Gracias a todos.

Contenido

Lista de Figuras

Lista de Tablas

Acrónimos

A/D	Analógico/Digital
AAATE	Association of Advancement of Assistive Technology in Europe
AAL	Ambient Assisted Living
ADL	Actividades de la Vida Diaria
AEED	Error de Detección
AEER	Error de Reconocimiento
AMS	Amplitude Modulation Spectrograms
AUD	Acoustic Unit Descriptor
BIC	Bayesian Information Criteria
CHIL	Computer in the Human Interaction Loop
CMN	Cepstral Mean Normalization
COORVISOR	Coordinadora Vizcaína de Sordos
DWT	Discrete Wavelet Transform
EDAD	Encuesta sobre Discapacidad, Autonomía personal y situaciones de Dependencia
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
HPSS	Harmonic Percussive Sound Separation
INE	Instituto Nacional de Estadística
K-NN	K-Nearest Neighbors
LFCC	Linear Frequency Cepstral Coefficients
LPC	Linear Prediction Coefficients
LVQ	Learning Vector Quantization
MFCC	Mel Frequency Cepstral Coefficients

MP	Matching Pursuit
MRMR	Minimum Redundancy Maximum Relevance
NN	Neural Networks
OMS	Organización Mundial de la Salud
PCA	Principal Component Analysis
RF	Roll-Off Point
RMS	Root Mean Square
RTP	Real Time Protocol
RUBICON	Robotic UBIquitous COgnitive Network
RWCP	Real World Computer Partnership
SNR	Signal Noise Ratio
STE	Short Time Energy
SVM	Support Vector Machines
TIC	Tecnologías de la Información y la Comunicación
UTC	Umbral de Tramas Consecutivas
UTT	Umbral de Tramas Totales
ZCR	Zero Crossing Rate

1. Introducción

1.1 Motivación y Antecedentes de la Tesis

El oído es uno de los sentidos más importantes del ser humano. Después de la vista, que es el sentido que más información aporta sobre el medio en el que nos movemos, podemos considerar el oído como la herramienta perceptiva humana con más trascendencia en cualquier actividad diaria. El oído humano es capaz de distinguir un gran número de sonidos ambientales con tan sólo prestar un poco de atención. Estos sonidos pueden ser entendidos como una gran cantidad de datos de información que, transmitidos por un medio, forman parte del proceso de comunicación entre el entorno y las personas que reciben los mensajes.

Pero… ¿qué sucede cuando esta capacidad de comunicación no existe o está muy limitada? Las personas con discapacidad auditiva son un claro ejemplo de ello. Las personas con discapacidad auditiva sufren cada día los inconvenientes que la imposibilidad de detectar e identificar sonidos conlleva. Diferenciar sonidos del entorno es una capacidad inherente de las personas, y carecer de ella puede suponer muchos problemas. Una alarma de incendios o una llamada telefónica a horas intempestivas avisando de un peligro pueden ser sonidos concebidos como prioritarios a detectar. El no hacerlo, no sólo puede poner a la persona en peligro sino también puede conllevar a una permanente preocupación e inseguridad, perjudicial para la salud del individuo.

Construir un sistema capaz de actuar con la suficiente inteligencia que le permita percibir y tomar las decisiones adecuadas como si de un ser humano se tratase, ha sido y sigue siendo en la actualidad una meta para investigadores de todo el mundo. En un sistema de este tipo, la tecnología de base son las técnicas de reconocimiento de eventos acústicos. Sin embargo, los esfuerzos actuales de investigación en detección e identificación de señales sonoras han estado centrados mayoritariamente en el amplio campo del reconocimiento del habla, dejando a un lado todo lo referente a otros sonidos producidos en el entorno. Esto se refleja en el número de escasas publicaciones si comparamos la gran fuente de información que produce una búsqueda basada en reconocimiento de voz. No obstante, la concienciación social y la búsqueda de la aplicabilidad, cada vez más exigida en la investigación, hacen que esta tendencia esté cambiando, surgiendo nuevos proyectos orientados al reconocimiento de este tipo de sonidos, pero siempre alejados del mercado y de un producto final robusto y consolidado orientado a la comunidad de personas con discapacidad auditiva.

Es por todo esto que el desarrollo de un sistema capaz de detectar y clasificar de forma automática los diversos sonidos de interés que puedan surgir en una vivienda, en la oficina, en el centro de estudio, etc, se concibe como una necesidad primordial para el colectivo de personas con disfunciones sensoriales auditivas que contribuirá a la mejora de su calidad de vida. La finalidad de esta tesis es contribuir a la mejora de la calidad de vida del colectivo de personas con discapacidad auditiva mediante el estudio de las técnicas de reconocimiento de sonidos no-habla desde un punto de vista práctico y funcional.

1.2 Objetivos de la Tesis

El principal objetivo de esta tesis es el estudio y aplicabilidad de técnicas de reconocimiento de sonidos no-habla orientadas al desarrollo de sistemas accesibles para personas con limitaciones auditivas.

Este objetivo marca el carácter aplicado de esta investigación. El alcance de esta tesis fija las pautas en la evaluación y desarrollo de técnicas y metodologías que permitan la construcción de sistemas capaces de trabajar en tiempo real, tanto en plataformas estándar como en dispositivos móviles, acercando la investigación en ella realizada a las necesidades del colectivo de personas con discapacidad auditiva.

Además, se quiere hacer partícipe de esta tesis tanto a empresas como usuarios finales. De esta forma, para las empresas, el desarrollo de esta tesis supondrá un análisis exhaustivo y elaborado del potencial impacto que este tipo de soluciones puede suponer en el mercado, aportándoles un estudio que les permita tomar decisiones sobre posibles inversiones en estas tecnologías. Para las personas con problemas auditivos esta tesis supondrá poner de manifiesto y en conocimiento de la comunidad científica una problemática con la que conviven día a día. De esta forma, a través de la colaboración con asociaciones y fundaciones de personas con discapacidad auditiva tanto a nivel provincial como estatal, se analizarán las preferencias y necesidades de este colectivo, estableciéndose las pautas finales de diseño y accesibilidad de este tipo de sistemas.

Debido a que el reconocimiento de sonidos no-habla es todavía una línea de investigación en desarrollo, parte de esta tesis se centra en comparar los algoritmos implementados con las bases de datos utilizadas en la bibliografía y disponibles para su evaluación. Este estudio permitirá la selección de los algoritmos y técnicas a emplear en el escenario de interés para el colectivo de personas con discapacidades auditivas: el entorno del hogar. A través del análisis de necesidades de este colectivo, se diseñará el sistema que integre estos algoritmos y técnicas de forma práctica y realista. De esta forma, la investigación se orientará al estudio de la optimización del sistema en plataformas móviles con condiciones de ruido y movilidad variables. En esta

investigación se diseñará la metodología de trabajo en base a la simulación de situaciones reales donde el propio usuario será el encargado de grabar y entrenar sus propios sonidos, diseñando una metodología de grabación que no implique la adquisición de varias muestras del mismo sonido y reduciendo el tiempo de entrenamiento.

Aunque el enfoque principal de esta tesis se centra principalmente en el reconocimiento de sonidos no-habla dentro del entorno del hogar, esta tesis analiza también otros entornos tales como la oficina o centro de estudios, calle o vehículo y ocio.

1.3 Estructura de la Tesis

La memoria de esta Tesis Doctoral está compuesta de 8 capítulos, el primero de los cuales corresponde con esta introducción.

El capítulo 2 hace una exposición del contexto al que va orientada esta investigación. En primer lugar se realiza un análisis del colectivo de personas con discapacidad, y más específicamente del colectivo de personas con discapacidad auditiva, analizando este grupo como potencial beneficiario y usuario de este tipo de soluciones. En segundo lugar se analizan los productos de apoyo disponibles en el mercado para problemas de audición y el consumo de los mismos, así como el impacto que sistemas de reconocimiento de sonidos no-habla pueden tener en dicho mercado.

El capítulo 3 define el conjunto de sonidos no-habla y expone las características que éstos poseen frente a otros tipos de sonidos. Se estudian las diferentes taxonomías que los caracterizan y se presentan los sistemas asistivos que hacen uso de ellos.

El capítulo 4 realiza un análisis del estado del arte de las técnicas de reconocimiento de sonidos actuales. Se describen las bases de datos más utilizadas y se analizan las diferentes etapas que componen la arquitectura de estos sistemas: extracción de características, algoritmos de detección y algoritmos de clasificación.

El capítulo 5 muestra los resultados del estudio realizado para analizar las necesidades del colectivo de personas con discapacidad auditiva en el área del reconocimiento de sonidos. Se analizan los entornos de mayor interés y, para cada uno de ellos, se analizan los sonidos más relevantes y los medios más adecuados que utilizar para su correcta indicación y visualización.

El capítulo 6 ofrece los resultados del estudio experimental obtenidos de la aplicación de las técnicas desarrolladas de reconocimiento de sonidos no-habla utilizando las bases de datos más habituales en la literatura científica, tanto con sonidos aislados como con sonidos en audio continuo.

En el capítulo 7, estas técnicas son aplicadas al entorno del hogar, cubriendo las necesidades establecidas por el colectivo de personas con discapacidad auditiva. En este capítulo las técnicas son analizadas no sólo con micrófonos convencionales en plataforma de PC sino también con micrófonos móviles, mostrando finalmente la implementación de las mismas sobre estos dispositivos en tiempo real.

Finalmente, en el capítulo 8 se resumen las conclusiones y aportaciones de esta tesis, así como las líneas de trabajo futuras que deja abiertas.

2. Tecnología y Discapacidad

El avance de la tecnología brinda una nueva herramienta de apoyo a la sociedad y, más concretamente, a las personas con discapacidad auditiva. Invertir en ofrecer sistemas accesibles y robustos a este colectivo debe ser contemplado como una apuesta a la excelencia y al éxito para las grandes empresas tecnológicas. La investigación en tecnologías y productos de nueva generación puede marcar la diferencia en el mercado y posicionar a una compañía a un nivel superior, sin embargo, las empresas necesitan datos objetivos que les permitan cuantificar económicamente la inversión y el reembolso que esto supone.

En este capítulo se analizan los datos demográficos de la población española con discapacidad. En el primer apartado se analiza el incremento del número de personas con discapacidad con respecto a la edad, mostrando cómo el envejecimiento de la población y la relación existente entre discapacidad y edad hacen evidente la necesidad de invertir en productos de apoyo. En posteriores apartados se presentan las clasificaciones actuales de deficiencias y discapacidades, acotando la población a las personas con problemas auditivos. Además, se analizan los productos de apoyo existentes y el consumo de los mismos por las personas con discapacidad auditiva. En este sentido, se presentan cifras de ventas de los productos más relevantes (audífonos e implantes cocleares) así como información y comparativa de los sistemas de señalización de eventos actuales en el mercado. Este capítulo pretende ofrecer un estudio de mercado objetivo y exhaustivo sobre tecnologías de reconocimiento de sonidos no-habla.

2.1 Análisis del Colectivo de Personas con Discapacidad en España

Al tratar de analizar los tipos de deficiencias y discapacidades, y el tamaño de la población objetivo para el sector de las tecnologías de apoyo, existe una importante divergencia entre datos según la fuente consultada cuando son estudios con un número reducido de participantes. Para este fin, cabe destacar la importancia del Instituto Nacional de Estadística (INE) el cuál ha realizado durante estos últimos años diferentes macro-encuestas sobre discapacidad. Estas encuestas son operaciones estadísticas que cubren buena parte de las necesidades de información sobre los fenómenos de la discapacidad, la dependencia, el envejecimiento de la población y el estado de salud de la población residente en España. Se han realizado tres macro-encuestas en 1986, 1999 y 2008: la *Encuesta sobre Discapacidades, Deficiencias y Minusvalías* [1], la *Encuesta sobre Discapacidades, Deficiencias y Estado de Salud* [2] y

la *Encuesta de Discapacidad, Autonomía personal y situaciones de Dependencia* [3]. Las metodologías siguen las recomendaciones de la *Organización Mundial de la Salud* (OMS), y en particular las clasificaciones internacionales vigentes en el año de realización de cada encuesta. El estudio realizado en este capítulo se basa en los datos obtenidos de la *Encuesta sobre Discapacidad, Autonomía personal y situaciones de Dependencia* (EDAD) de 2008 [3]. En ella se investiga la percepción subjetiva de las personas acerca de su discapacidad, entendida como limitación en la realización de alguna actividad. Aunque ya han pasado varios años desde su publicación, se trata de la mayor operación estadística realizada hasta la fecha (88.725 viviendas distribuidas en 3550 secciones censales, siendo 25 el número de viviendas entrevistadas en cada sección). Es importante reseñar que la amplia mayoría de informes encontrados sobre discapacidad y TICs se realizaron entre los años 2006 y 2010. La situación de crisis en la que vive el país ha limitado los recursos de asociaciones, fundaciones y entidades gubernamentales que antes se encargaban de recopilar, analizar y publicar estos datos.

Los cambios demográficos experimentados en las últimas décadas en España han traído consigo profundas transformaciones en la pirámide poblacional, entre ellas un proceso de envejecimiento notable. Uno de los posibles efectos es el aumento de las personas con discapacidad, ya que la edad es un factor determinante en la aparición de este fenómeno. Además, el aumento de la longevidad ha coincidido con importantes cambios sociales que han llevado a que instituciones sociales y políticas deban ajustar sus objetivos a la nueva realidad, que demanda más protección social y apoyo a las personas que se encuentran en situación de dependencia [4].

Según la encuesta del INE, en 2008, el número total de personas con discapacidad mayores de 6 años en España era de 3.787.500 (8,9% de la población mayor de 6 años). El 58,8% de esta población con discapacidad corresponde a personas mayores de 65 años, lo que muestra la incidencia de la edad en la tasa de discapacidad. La Tabla 1 ofrece los datos correspondientes tanto en España como en las diferentes Comunidades Autónomas del país.

	Total	De 6 a 64 años	>= 65 años
España	3787,5 (8,9%)	1560,4 (3,7%)	2227,1 (5,2%)
Andalucía	716,2	325,3	390,9
Aragón	111,6	41,5	70,1
Asturias	104,6	37,7	66,9
Baleares (Illes)	68,8	34,4	34,4
Canarias	135,7	68,6	67,1
Cantabria	37,6	14,8	22,8
Castilla y León	255,8	81,3	174,5
Castilla-La Mancha	182,9	63,7	119,2
Cataluña	511,6	212,1	299,5
Comunitat Valenciana	452,8	198,2	254,6
Extremadura	110,9	41,7	69,2
Galicia	292,9	102,5	190,4
Madrid (Comunidad de)	434,9	189	245,9
Murcia	127,5	56,8	70,7
Navarra	41,5	13,1	28,4
País Vasco	169,3	65,1	104,2
La Rioja	18	6,8	11,2
Ceuta	7,5	3,8	3,7
Melilla	7,3	4	3,3

Tabla 1 Número total de personas mayores de 6 años con discapacidad en España (2008) (Unidades: miles de personas)

Analizando la encuesta a un nivel más detallado, en la Figura 1 se ilustra la evolución en España de la discapacidad en función de la edad. Se observa la curva exponencial que indica la mayor incidencia de las limitaciones sensoriales, físicas y cognitivas cuando la edad aumenta.

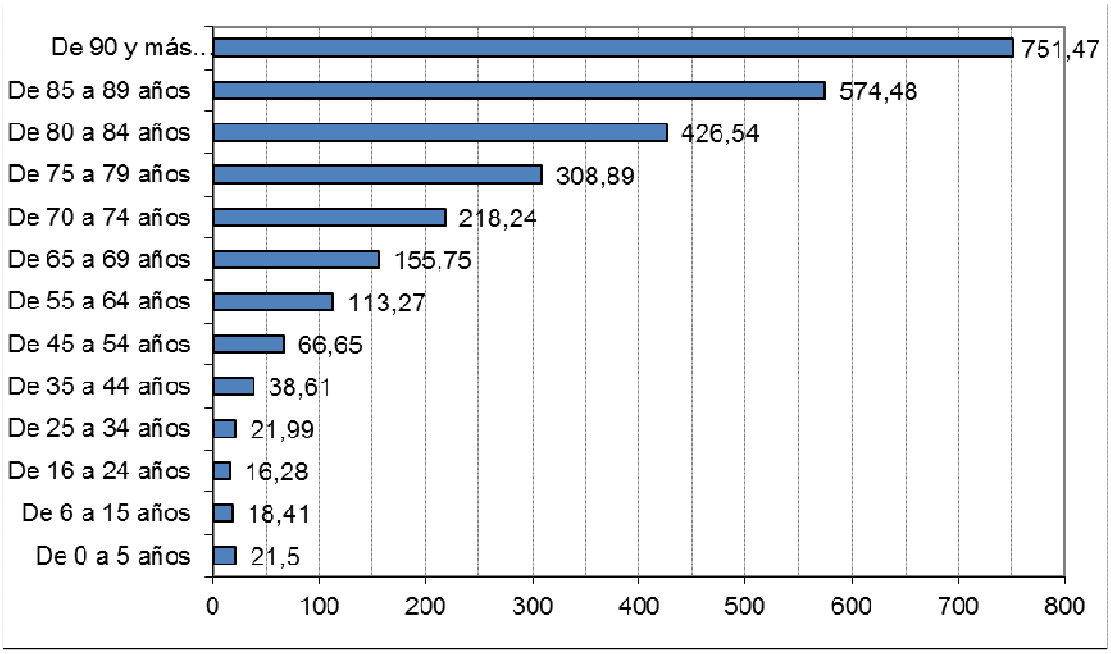


Figura 1 Evolución de la discapacidad en España (2008) (Unidades: miles de personas)

Los datos referentes a España muestran que la tasa de prevalencia de personas con discapacidad sobre el total de la población en cada una de las franjas de edad, pasa de un 6,6% en el tramo de 45 a 54 años a un 30,8% en el tramo de 75 a 79 años, hasta alcanzar un 57,4% en las personas mayores de 85 años. Se concluye por tanto que la evolución futura de la población dependiente estará determinada por el proceso de envejecimiento demográfico.

En la gráfica de la Figura 2 se muestra la evolución de los datos de prevalencia de la población española mayor de 65 años según estimaciones del INE. Si verificamos esta estimación hecha en 2008 con el último dato real aportado por el INE a fecha del **1 de enero del 2015** donde se indica que el porcentaje de personas mayores de 65 años es del 19,7% vemos cómo la estimación era incluso más moderada (siendo 17,5% la estimación realizada para esta fecha).

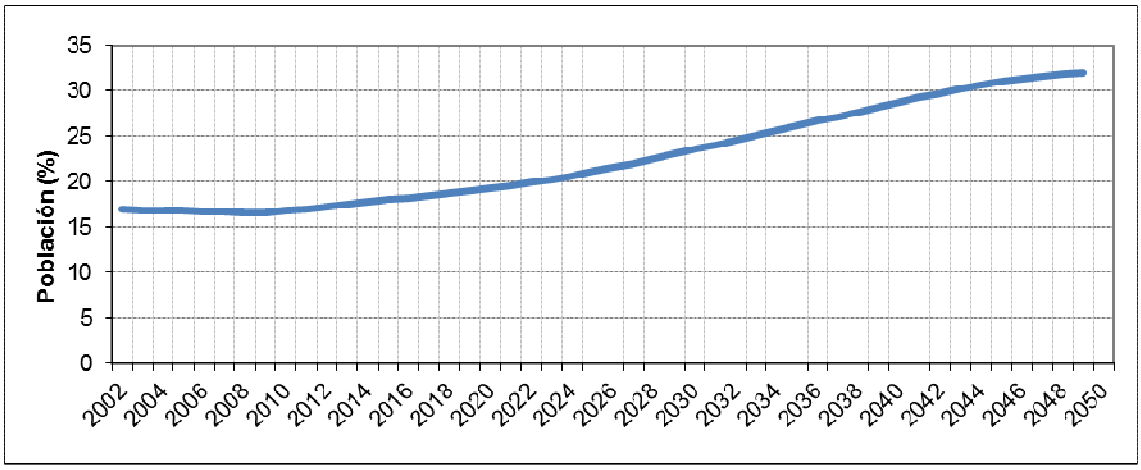


Figura 2 Evolución de los datos de prevalencia en España de personas mayores de 65 años (2008)

Si comparamos los datos procedentes de las Naciones Unidas [5] entre España y la Unión Europea (ver Tabla 2) podemos observar que, en España la evolución demográfica de personas mayores de 65 años se espera superior, pasando de porcentajes inferiores a la media Europea a cifras por encima de la media Europea en el 2050. Los escenarios de población mayor indican que entre los años 2010 y 2020 se incorporarán a la vejez aquellas generaciones que nacieron durante la Guerra Civil y la posguerra. En consecuencia, se producirá un vacío relativo de la población mayor en los primeros tramos de edad frente a un crecimiento de la población que supera los 75 y 85 años. Posteriormente, entre los años 2020 y 2040 se incorporarán a la vejez las generaciones más numerosas de la historia de España. Durante este período, la población mayor aumentará y lo hará en todos los tramos de edad [6].

Año	% España	Total España	% Europa	Total Europa
1950	7,3	2044	8,2	449080
2000	16,8	6772	14,8	107305
2050	31,8	16298	27,4	189118

Tabla 2 Evolución de los datos de prevalencia España - Europa de personas mayores de 65 años (Unidades: % y miles de personas)

2.2 Tipos de Deficiencias y Discapacidades

Dentro del ámbito de la salud, se entiende como ***deficiencia*** toda pérdida o anormalidad de una estructura o función psicológica, fisiológica o anatómica que puede ser temporal o permanente, entre las que se incluye la existencia o aparición de una anomalía, defecto o pérdida producida en un miembro, órgano, tejido y otra estructura del cuerpo, incluidos los sistemas propios de la función mental.

Por otro lado una ***discapacidad*** es toda restricción o ausencia (debido a una deficiencia) de la capacidad de realizar una actividad en la forma o dentro del margen que se considera normal para un ser humano. Pueden ser temporales o permanentes, reversibles o irreversibles. Pueden surgir como consecuencia directa de la deficiencia o como una respuesta del propio individuo. La discapacidad representa la objetivación de una deficiencia y en cuanto tal, refleja alteraciones a nivel de la persona.

De los datos del INE se extrae que en España las deficiencias osteoarticulares son las que presentan mayor incidencia en la población, representando un 27% del total (ver Figura 3). Cabe destacar que las deficiencias de audición, que son las más relacionadas con el núcleo principal de esta tesis, ocupan el segundo lugar con un 17%. Esta categoría abarca las personas con deficiencias de funciones y estructuras asociadas al aparato de la audición:

- **Sordera prelocutiva:** Se refiere a personas con sordera, previa a la adquisición del lenguaje (niños). Incluye la sordomudez cuya mudez se ha presentado como consecuencia de una sordera prelocutiva.
- **Sordera postlocutiva:** Se refiere a personas con sordera que se presenta después de la adquisición del lenguaje (adultos) con pérdida total de audición y que no pueden beneficiarse del uso de prótesis auditivas.
- **Mala audición:** Se refiere a personas con diferentes niveles de pérdida auditiva: moderada (45-50 dB), grave (71-91 dB), profunda (>91 dB). Pueden beneficiarse del uso de prótesis auditivas.
- **Trastornos del equilibrio:** Se refiere a personas que padecen vértigos laberínticos (el más frecuente es el vértigo *Ménière*), mareos y defectos de locomoción por trastornos vestibulares.

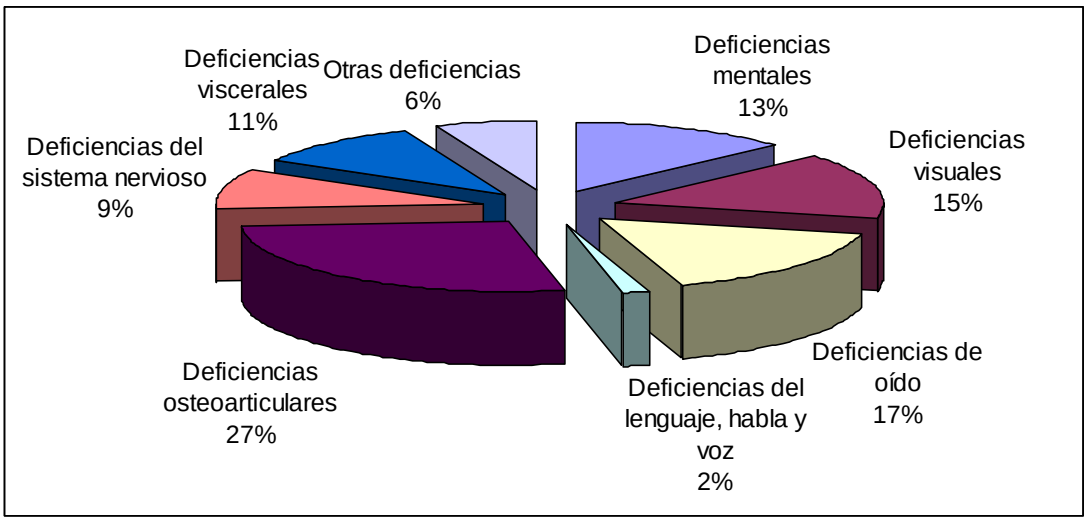


Figura 3 Incidencia de deficiencias en España (2008)

Al analizar las discapacidades que estas deficiencias provocan en la vida diaria de una persona (ver Figura 4), los datos del INE reflejan que las discapacidades más frecuentes son las encuadradas en las categorías de "*Movilidad*", "*Vida doméstica*" y "*Autocuidado*". Se aprecia una relación clara con la distribución de las deficiencias debido a que estas tres discapacidades, en alto grado, suelen tener origen en deficiencias osteoarticulares que, en el gráfico de la Figura 3, ocupaban el primer lugar de deficiencias más comunes. Ya en un segundo grupo en cuanto a importancia cuantitativa se encuentran las discapacidades de audición, seguidas de las relacionadas con la vista y la comunicación.

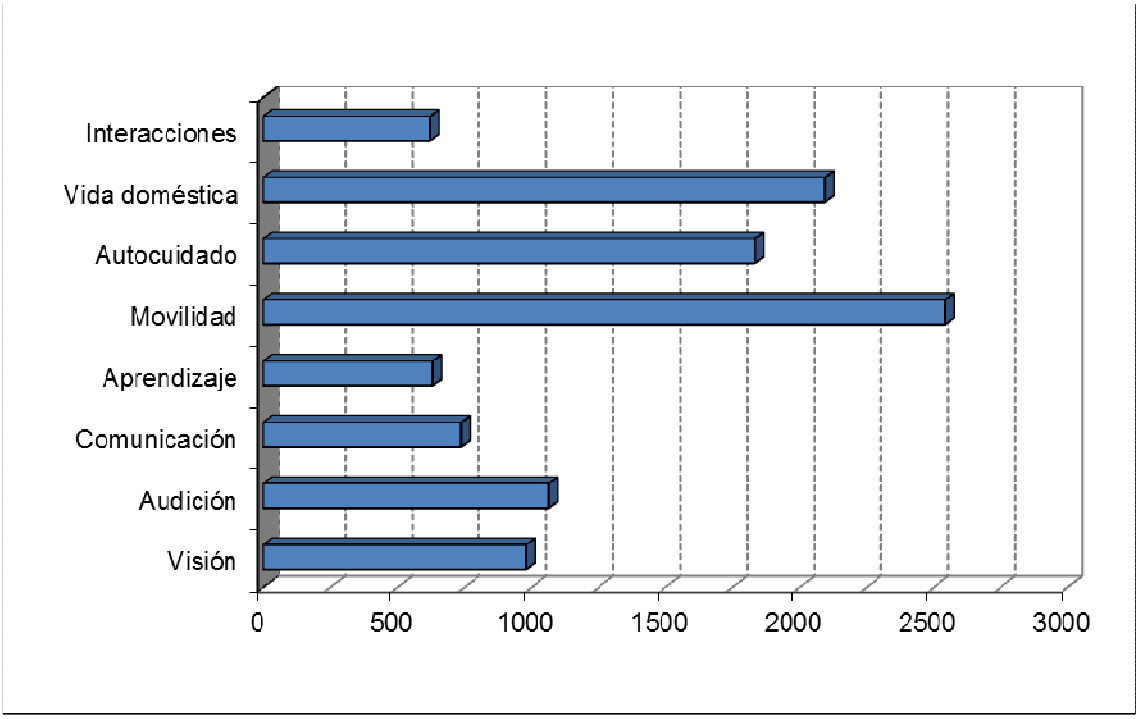


Figura 4 Discapacidades más frecuentes en España (2008) (Unidades: miles de personas)

Conviene indicar en este punto que una misma persona puede presentar más de un tipo de discapacidad. En la encuesta del INE se recoge este hecho, observándose cómo a nivel nacional más del 85% de las personas con deficiencias manifiestan tener dos o más tipos de discapacidad tal y como refleja la Tabla 3.

Sólo una discapacidad	Dos discapacidades	De tres a diez discapacidades	De once a más discapacidades
548,4 (14,5%)	545,5 (14,4%)	1689,6 (44,6%)	1004 (26,5%

Tabla 3 Número de personas en España con una o más discapacidades (2008) (Unidades: miles de personas)

Centrándonos en el colectivo de personas con deficiencias auditivas, la Tabla 4 muestra el número de discapacidades que tienen como origen las diferentes disfunciones del oído. Los resultados evidencian el hecho de que una persona con problemas auditivos puede estar limitada en varias actividades diferentes.

	Sólo una discapacidad	Dos discapacidades	De tres a diez discapacidades	De once a más discapacidades
Sordera prelocutiva	1,3	1	11,3	4,2
Sordera postlocutiva	7,9	3	8	12,1
Mala audición	183,6	176,8	287,7	192,4
Trastornos de equilibrio	3	3,6	13,5	5,3

Tabla 4 Número de personas en España con una o más discapacidades con deficiencias auditivas (2008) (Unidades: miles de personas)

A pesar de esta panorámica, los avances tecnológicos ofrecen soluciones con las cuales muchas discapacidades son compensadas con productos de apoyo que permiten a la persona con problemas poder llevar una vida más independiente. Desde lectores de pantallas a sistemas más avanzados son varias las opciones disponibles en el mercado de las que los usuarios se pueden beneficiar. En el siguiente apartado se hará una descripción de los productos de apoyo actuales.

2.3 Productos de Apoyo. Definición y Clasificación

Hablamos de "*Ayudas Técnicas*", "*Dispositivos o Productos de Asistencia*", "*Tecnologías y Productos de Apoyo*", etc. En ocasiones se utiliza también el término "*Tecnología Asistiva*"; término que proviene de una traducción, no totalmente aceptada y/o correcta, del inglés "*Assistive Technology*". En su lugar, en castellano, nos referiremos a "*Tecnología de Apoyo*" o "*Tecnología de Asistencia*".

Tomando como referencia la norma internacional ISO 9999 en su última edición publicada en 2007, el término "*Productos de Apoyo*" ("*Assistive products*") reemplaza al término "*Ayudas Técnicas*" utilizado en las anteriores ediciones de dicha norma. Las

definiciones que de ambos términos establece dicha norma internacional son las siguientes:

Ayuda Técnica ***UNE-EN ISO 9999: 2002***

Cualquier producto, instrumento, equipo o sistema técnico usado por una persona con discapacidad, fabricado especialmente o disponible en el mercado, para prevenir, compensar, mitigar o neutralizar la deficiencia, limitación para la actividad o dificultades para la participación.

Nota: Las ayudas técnicas son nombradas frecuentemente como "dispositivos de asistencia" o "tecnología de apoyo".

Producto de Apoyo ("Assistive product") ***ISO 9999: 2007***

Cualquier producto (incluidos dispositivos, equipamiento, instrumentos, tecnología y software), fabricado especialmente o disponible en el mercado, para prevenir, compensar, supervisar, mitigar o neutralizar discapacidades, limitaciones para la actividad y la participación.

Nota: En esta edición el término " Productos de Apoyo" ("Assistive products") reemplaza al término "Ayudas Técnicas" utilizado en ediciones previas.

Esta definición abarca una gama muy amplia y variada de productos que van desde dispositivos, como las prótesis, concebidos específicamente para compensar en lo posible una deficiencia funcional concreta, hasta productos dirigidos más a facilitar y mejorar la vida diaria de las personas con discapacidad y/o personas mayores. La propia norma ISO 9999 recoge una detallada clasificación de las ayudas técnicas o productos de apoyo. En la Tabla 5 se muestra el primer nivel de la clasificación definida según la norma ISO 9999 tanto en la versión de 2002 como en la última versión de 2007.

	UNE-EN ISO 9999:2002		ISO 9999:2007
No	**Concepto**	**No**	**Concepto**
04	Ayudas para el tratamiento médico personalizado	04	Productos de apoyo para tratamiento médico personalizado
05	Ayudas para el entrenamiento/ aprendizaje de capacidades	05	Productos de apoyo para el entrenamiento en habilidades
06	Ortesis y prótesis	06	Ortesis y prótesis
09	Ayudas para el cuidado y la protección personales	09	Productos de apoyo para el cuidado y la protección personales
12	Ayudas para la movilidad personal	12	Productos de apoyo para la movilidad personal
15	Ayudas para actividades domésticas	15	Productos de apoyo para tareas domésticas
18	Mobiliario y adaptaciones para viviendas y otros inmuebles	18	Mobiliario, adaptaciones para viviendas y otros inmuebles
21	**Ayudas para la comunicación, la información y la señalización**	22	**Productos de apoyo para la comunicación e información**
24	Ayudas para la manipulación de productos y bienes	24	Productos de apoyo para manejar productos y bienes
27	Ayudas y equipo para mejorar el ambiente, maquinaria, herramientas	27	Productos de apoyo para mejorar el entorno, maquinaria, herramientas
30	Ayudas para el esparcimiento	30	Productos de apoyo para el esparcimiento

Tabla 5 Primer nivel – Norma ISO 9999

Las diferencias entre ambas versiones de la norma, en cuanto a la clasificación, son pocas, aunque cabe mencionar cómo el grupo 21 "*Ayudas para la comunicación, la información y la señalización*" ha pasado a ser identificado con el código 22, modificándose significativamente la denominación de los subgrupos incluidos en el mismo (ver Tabla 6).

	UNE-EN ISO 9999:2002		ISO 9999:2007
Cód.	**Concepto**	**Cód.**	**Concepto**
21.03	Ayudas ópticas	22.03	Productos de apoyo para ver
21.06	Ayudas electro-ópticas	22.06	Productos de apoyo para oír
21.09	Unidades de salida de ordenadores, máquinas de escribir y equipos electrónicos	22.09	Productos de apoyo para producción de voz
21.12	Ordenadores	22.12	Productos de apoyo para dibujo y escritura
21.15	Máquinas de escribir y procesadores de texto	22.15	Productos de apoyo para cálculo
21.18	Calculadoras	22.18	Productos de apoyo para recepción de información audio, visual y video
21.24	Ayudas para dibujo y escritura manual	22.21	Productos de apoyo para comunicación cara a cara
21.27	Ayudas no ópticas para la lectura	22.24	Productos de apoyo para telefonear (y para mensajería telemática)
21.30	Grabadoras y receptores de audio	22.27	Productos de apoyo para alarma, indicación y señalización
21.33	Equipo de televisión y video	22.30	Productos de apoyo para lectura
21.36	Teléfonos y ayudas para telefonear	22.33	Ordenadores y terminales
21.39	Sistemas de transmisión de sonido	22.36	Dispositivos de entrada para ordenador
21.42	Ayudas para comunicación cara a cara	22.39	Dispositivos de salida para ordenador
21.45	Ayudas para audición		
21.48	Ayudas para señalar e indicar		

Tabla 6 Grupo 21/22 – Norma ISO 9999

Dentro de estos grupos se pueden encontrar todos los productos y/o dispositivos de apoyo para la audición, ya sea habla como alarmas o indicaciones varias.

Se destaca cómo el sector de productos de apoyo engloba una muy amplia gama de productos. Estos tienen un carácter muy heterogéneo en cuanto a proceso, sector productivo y contenido tecnológico, sin embargo, su desconocimiento o precio, en muchas ocasiones muy elevado, limita su consumo.

2.4 Consumo Actual de Productos de Apoyo

Una vez caracterizado el colectivo de potenciales usuarios de los Productos de Apoyo, el siguiente paso en el estudio es el conocimiento del nivel de demanda o de consumo que en la actualidad existe de los productos de apoyo a nivel nacional.

Durante mucho tiempo, los diversos colectivos de personas con discapacidad en España han gozado de los beneficios del estado de bienestar otorgado por las políticas instauradas por los gobiernos en los pasados años. Sin embargo, en los últimos años, como consecuencia de la situación de crisis económica, estas políticas se han visto seriamente afectadas. Tal y como se cita en el informe "*Acceso y uso de las TIC por las personas con discapacidad*" [7], un ejemplo claro de estas políticas lo constituye la "*Estrategia Española sobre Discapacidad 2012-2020*", que se redactó por el *Ministerio de Sanidad, Política Social e Igualdad* con la idea de establecer un espacio general de acción político-social para mejorar la calidad de vida de las personas con discapacidad. Todo ello, mediante el diseño de unas líneas básicas para trazar las políticas públicas que habrían de desarrollarse en los años de su aplicación, y que se dibujaron alineándolas con las enunciadas a nivel europeo en la "*Estrategia Europea sobre Discapacidad 2010-2020*" y la "*Estrategia Europea 2020*". El objetivo último de estas intervenciones es avanzar hacia la plena integración del colectivo de personas con discapacidad en la sociedad, en condiciones reales de igualdad, tanto de derechos como de oportunidades. En este informe se vio manifestado el hecho de que las personas con discapacidad en España tienen tasas significativamente menores del uso de productos de apoyo que las personas sin discapacidad. Al tener menores ingresos económicos, algunos de ellos ni siquiera pueden acceder a servicios básicos como la telefonía, el ordenador o Internet.

Analizando los productos de las categorías descritas en la norma ISO 9999, al hablar de productos de apoyo para personas sordas o con limitaciones auditivas se diferencian dos grandes grupos o tipos de productos:

- *Audífonos e implantes cocleares*.
- *Dispositivos y Sistemas de Asistencia a la Escucha*: Cualquier dispositivo o sistema, excluidos los audífonos, diseñado para mejorar la habilidad de una persona con discapacidad auditiva para comunicarse y para funcionar de forma más independiente a pesar de su pérdida auditiva. Estos dispositivos actúan bien amplificando el sonido y transmitiéndolo de forma más directa desde su fuente al que escucha, o transformando el sonido en señales visuales o vibratorias.

2.4.1 Audífonos e Implantes Cocleares

La gran mayoría de los informes sobre consumo se han realizado enfocándose en el primer grupo: los audífonos y los implantes cocleares, siendo más difícil encontrar datos de consumo y uso en el segundo grupo.

Según informes del *Portal de Profesionales de la Audiología*[*], los datos de consumo de productos del primer grupo muestran que existe un importante mercado de potenciales usuarios. El mercado mundial de **audífonos** está por encima de los 10 millones de unidades. Está dividido de manera uniforme entre los Estados Unidos y Canadá (alrededor del 31% del mercado de audífonos del mundo por unidades), Europa (38%), y Asia y el Pacífico (22%), así como el resto del mundo (9%). Los países con mayores ventas de audífonos son los que tienen economías más establecidas, como las que pertenecen a la *Organización para la Cooperación y el Desarrollo Económico*, y los que tienen una gran población de personas mayores. Los Estados Unidos, con ventas de más de 2,8 millones de audífonos en 2012 (28% del mercado mundial), está muy por encima de los dos siguientes países del ranking, Alemania y Japón – dos de los países que cuentan con un nivel más rápido de envejecimiento de la población- que representan por encima del 8% y 7% de las ventas de audífonos en todo el mundo, respectivamente.

Hearing Review, revista líder en audiología, estima que China es ahora el cuarto país más grande en términos de volumen de unidades de audífonos, pero admite que es muy difícil conseguir cifras de mercado sólidas sobre las ventas de audífonos en este país. Al igual que Alemania y Japón, China también cuenta con una creciente población de la tercera edad y cabe predecir que se convertirá en el mayor consumidor de aparatos auditivos en las próximas décadas.

Según la consultora *Grand View Research* se espera que la tasa de crecimiento anual compuesta, que mide la tasa de crecimiento del mercado año a año, sea del 3,2% desde 2014 a 2020. Los principales factores que impulsarían este aumento serían el tamaño cada vez mayor de la población que envejece y una mayor prevalencia de las pérdidas auditivas y de la sordera. Otros factores también podrían contribuir a este crecimiento. El informe cita el rápido incremento en las tasas del uso de audífonos digitales, así como los niveles de conocimiento de los pacientes, que se espera que aumenten durante el período pronosticado. El estudio también indica que Europa representó la mayor parte del mercado de audífonos en 2013 (más del 40,0%). Sin embargo, se espera que la región de Asia Pacífico obtenga una mayor tasa de crecimiento para el período 2014-2020 (más de 4,5%), ya que los sistemas de salud de la región están en constante proceso de mejora y el nivel adquisitivo continúa

[*] *www.audifono.net*

aumentando. El informe cita a los principales fabricantes del mercado como *Sonova Holding AG, William Demant Holding, GN Resound, Widex* y *Siemens AG*.

Estas cifras muestran un mercado cuyo volumen es creciente y con perspectivas positivas en el futuro. Representa un mercado de productos de alto nivel tecnológico que está liderado por multinacionales extranjeras. España desempeña en este sector el papel de distribuidor y de adaptación del producto, ya que el nivel de fabricación propia es muy bajo en comparación con otros países de la Unión Europea.

Datos de la compañía fabricante de **implantes cocleares** *MED-EL* indican que, durante los últimos doce meses anteriores a Julio de 2013, aproximadamente se vendieron 50.000 implantes cocleares en todo el mundo. En este mercado son dos las empresas principales, siendo la distribución de ventas como se indica a continuación:

- *MED-EL*: 14.027
- *COCHLEAR*: 26.674
- Resto: 9.000

Aproximadamente, 30.000 de todos ellos fueros usados con niños. El número de niños que nacen en todo el mundo se estima en 134 millones por año y se predice que continúe estable.

2.4.2 Dispositivos y Sistemas de Asistencia a la Escucha

Dentro de la categoría "*Dispositivos y Sistemas de Asistencia a la Escucha*" no se han encontrado datos sobre ventas, sin embargo, según la AAATE (*Association of Advancement of Assistive Technology in Europe*) se espera que a nivel europeo el volumen de mercado de la tecnología de apoyo alcance en 2015 la cifra de los 60.000 millones de euros y una de las áreas con mayor crecimiento será la relacionada con las TICs en Productos de Apoyo.

Estudios realizados desde la *Comisión Europea* auguran que el sector de los productos de apoyo va a ser muy dinamizado por la introducción masiva de las TIC. El subsector de las TIC aglutina una gran variedad de productos de apoyo y en la actualidad resulta el de mayor crecimiento potencial. Se está comprobando cómo las TIC son un importante apoyo tanto para las personas con discapacidad y mayores, como para sus cuidadores.

A nivel de España, de las 1.824 solicitudes de información recibidas en el área de información y asesoramiento en ayudas técnicas de *CEAPAT-IMSERSO* en 2001, el 64,47% fueron sobre productos de apoyo para la información y comunicación, y de ellas un 33% se referían a dispositivos de acceso al ordenador. Esto supone que el 85,75% de las consultas estuvieron relacionadas con las TIC. No obstante no existen

datos suficientes que puedan evaluar la adquisición y compra de estos productos por la dificultad que implica hacer un seguimiento de estos.

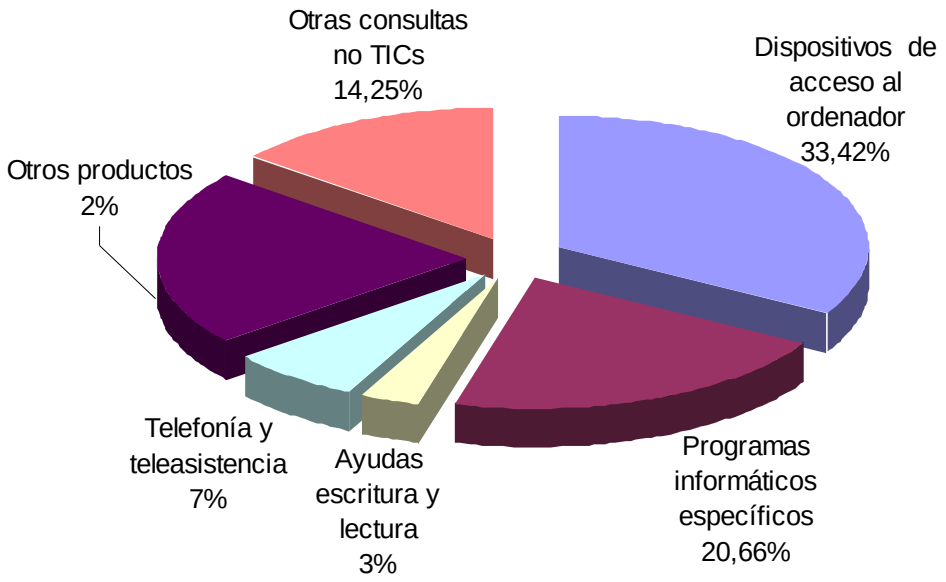


Figura 5 Solicitudes de información sobre Ayudas Técnicas recibidas por CEAPAT-IMSERSO en 2001

Aunque estos datos son muy antiguos, estas conclusiones se refrendan a nivel estatal por los resultados del informe prospectivo "*TICs y Dependencia*" [8] y "*Acceso y uso de las TIC por las personas con discapacidad*" [7] realizados por Fundación Vodafone España en 2007 y 2013. El siguiente apartado ofrece un resumen de los resultados de estos informes.

2.5 Discapacidad Auditiva y TICs como Producto de Apoyo

El estudio llevado a cabo por *Fundación Vodafone España* para la elaboración del informe "*TICs y Dependencia*", que contó con la participación de más de 450 personas pertenecientes a distintos colectivos de personas con dependencia a través de entrevistas personales y encuesta telefónica, revela el amplio calado que las TICs han comenzado a tener en las personas con discapacidad, debido a los beneficios que su uso les reporta.

El análisis inicial se realizó en base a la utilización de dos tecnologías: los teléfonos móviles e Internet; y su grado de uso por personas con distinto tipo de discapacidad o dependencia.

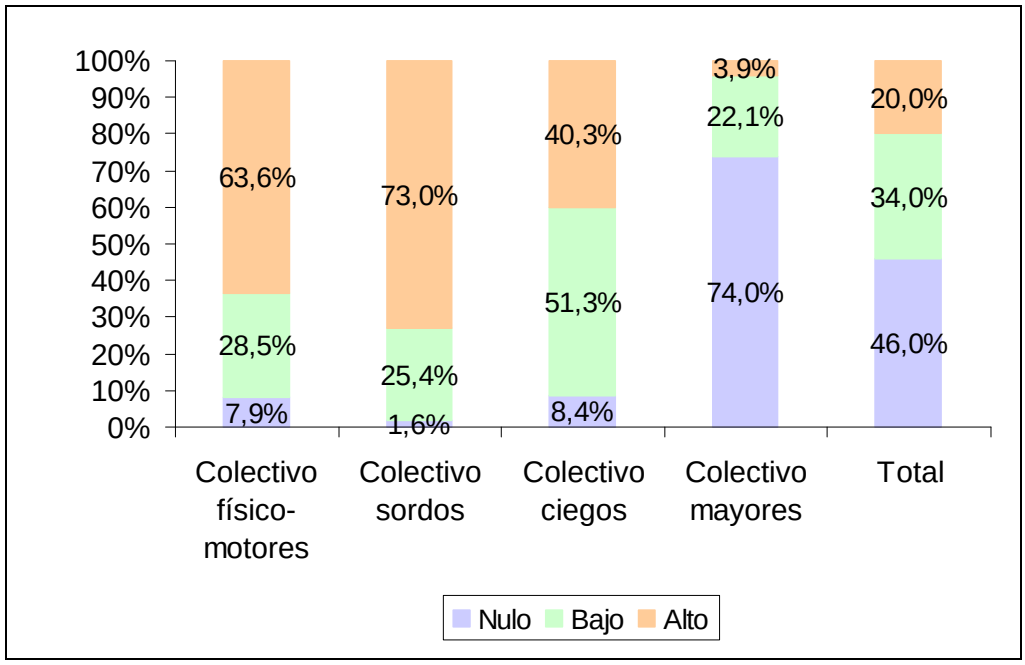


Figura 6 Uso de las TICs en base a colectivos de distinto tipo de discapacidad o dependencia (F. Vodafone)

Tal y como se muestra en la gráfica de la Figura 6, es el colectivo de personas sordas el más proclive a la utilización de las TICs, seguido del colectivo de personas con discapacidad físico-motora. En el polo opuesto se encuentra el colectivo de personas mayores con un 74% de encuestados que no utilizan estas tecnologías.

En el Informe de 2013, Vodafone destaca como nota importante el uso de las TICs en el colectivo de personas con discapacidad auditiva teniendo especial importancia los siguientes puntos.

- El uso del teléfono móvil es muy habitual entre el colectivo de personas con discapacidad auditiva (90,6%), siendo especialmente importante para ellos el envío y recepción de mensajes (45,5%).

- El colectivo de personas con discapacidad auditiva es el que más utiliza el ordenador (61,8%), respondiendo a las pautas socio-demográficas observadas por la población general, ya que lo utilizan menos las mujeres (61,0%), las personas mayores de 45 años (12,1%) y las que tienen un nivel de estudios bajo (52,4%), excepto en lo referente a la situación laboral, puesto que no suelen usar el ordenador si trabajan (34,4%) debido a la baja cualificación de sus desempeños. El colectivo de personas con discapacidad auditiva que no utiliza el ordenador alude sobre todo a su elevado precio: el 28,7% lo considera un gasto innecesario y el 29,9% piensa que podría hacerle gastar más de lo deseado.

- Prácticamente la mitad del colectivo de personas con discapacidad auditiva utiliza Internet (47,5%).

Entre las principales dificultades con las que se encuentran en su vida cotidiana las personas con discapacidad auditiva consultadas son:

- *Las dificultades para la comunicación y las relaciones sociales*: En las dificultades para la comunicación y las relaciones sociales las tecnologías de reconocimiento del habla pretenden ser una tecnología clave. Actualmente estas tecnologías están lejos de alcanzar un reconocimiento con gran precisión del habla libre, sin embargo, la tecnología actual sí que permite implementar sistemas que pueden trabajar con un error aceptable en entornos semánticos restringidos. Hay que destacar que el número de estudios e investigaciones sobre voz es cada vez mayor y los resultados obtenidos hacen albergar la esperanza de que en un futuro cercano la comunidad de personas con discapacidad auditiva pueda beneficiarse de una comunicación sin limitaciones ni barreras.

- *Los problemas para la identificación y reconocimiento de señales acústicas*: Dentro de esta categoría se encuentran los sistemas de señalización. Estos sistemas han sido creados para notificar a las personas sordas de señales que pueden producirse en el entorno. La mayoría de estos sistemas están centrados en el hogar pudiendo alertar a través de luces parpadeantes, alarmas de gran potencia sonora y/o a través de vibración. Sin embargo, la tecnología para la identificación y reconocimiento de señales sonoras no está tan avanzada como en el caso anterior. Son estos sistemas en los que se centra el contenido de la tesis y es por ello que en el siguiente apartado se hace una descripción más detallada de los mismos.

2.6 Sistemas de Señalización en el Mercado

Los sistemas de señalización son productos diseñados para notificar al usuario de diferentes eventos. Estos eventos pueden ser sonoros (timbre de la puerta, llanto de bebé,…) o salidas de sensores que llevan un aviso de emergencia (alarma de gas, inundación…) que no tienen por qué implicar una salida de audio. A continuación se muestra más en detalle las diferencias en cuanto a la arquitectura de estos sistemas.

- *Transmisores*: El transmisor es el encargado de una vez detectado el evento enviarlo al sistema para su posterior notificación al usuario. Generalmente es necesario comprar un transmisor por cada evento que se quiera detector. Cada transmisor está diseñado para alertar de un evento específico; sin embargo, algunos transmisores funcionan con más de uno.

- *Receptores*: El receptor es el encargado de recibir las señales de los diferentes transmisores y notificar al usuario. Existen receptores tanto fijos como

portables. En el caso de receptores fijos sería necesaria su instalación en tantas habitaciones como uno quiera ser alertado. Sin embargo, en el caso de receptores portables, algunos fabricantes han diseñado receptores tipo "*beeper*" o relojes que el usuario puede llevar consigo.

- *Tecnología de envío y recepción de la señal*: Hay dos tipos de tecnología disponibles para los sistemas de señalización. Éstas pueden ser:
    - *Radio Frecuencia*: Esta tecnología envía una señal desde el transmisor al receptor a través de ondas de radio. Es apropiada tanto para el uso en casas propias como para apartamentos o casas de alquiler debido a que no necesita ser añadido dentro de la instalación eléctrica de la vivienda.
    - *Línea portadora*: Esta tecnología envía la señal desde el transmisor al receptor a través del circuito eléctrico de la vivienda. El principal inconveniente es que precisa de una instalación previa.

En este nicho de mercado son varios los fabricantes que ofrecen soluciones propietarias. Los principales se listan en la Tabla 7.

Fabricante	Logotipo
Sonic Alert	Sonic Alert Specialized Hearing Solutions
Serene Innovations	SERENE innovations
Clarity	CLARITY Amplifying Your Life
Bellman & Symfon	Bellman & Symfon®
Silent Call	SilentCall Communications
Gentex	GENTEX CORPORATION

Tabla 7 Principales fabricantes de sistemas de señalización

En la Tabla 8 se muestran los eventos para los que cada fabricante tiene producto.

Tipo	Sonic Alert	Serene Innovations	Clarity	Bellman Symfon	Silent Call	Gentex	Otros
Alarmas Audio		x	x		x		
Bebé	x	x	x	x			x
Monóxido Carbono					x		x
Timbre Puerta	x	x	x	x	x		x
Golpes Puerta		x					x
Apertura Puerta				x	x		x
Alarma Incendios		x			x		x
Portero	x	x					
Sensor Movimiento		x	x				
Beeper		x	x		x		x
Detector Humo				x	x	x	x
Teléfono	x	x	x	x	x		x
Alerta Meteorológica		x			x		x

Tabla 8 Eventos detectados por los fabricantes de sistemas de señalización

Algunos ejemplos de los productos existentes se muestran en la Tabla 9. Tanto el *"Central Alert AX Audio Alarm Sensor"* como el "*Visit Baby Cry Transmitter*" basan su detección en un umbral de energía. El fabricante exige que los productos deban estar situados justo al lado de la fuente de audio, sin embargo, aunque el primero lo catalogan como detector de audio en general el segundo está tan sólo catalogado como detector de llanto de bebé. Como ejemplos de productos que ofrecen una detección a través de cable se muestran el "*TR-50 Telephone Ring Signaler*" y el "*Signature Series Weather Alert Transmitter*". El primero se conecta al cable de teléfono para poder detectar cuándo una llamada es recibida. El segundo ofrece la posibilidad de conectarse a radios americanas especiales que envían alertas del tiempo cuando se producen emergencias meteorológicas.

Central Alert AX Audio Alarm Sensor	Visit Baby Cry Transmitter	TR-50 Telephone Ring Signaler	Signature Series Weather Alert Transmitter
			
Serene Innovations	Bellman & Symfon	Sonic Alert	Silent Call

Tabla 9 Productos comerciales de señalización

De este análisis se observa que la clasificación de sonidos no relacionados con el habla es todavía una línea de investigación en auge de expansión. Los productos existentes en el mercado no están orientados a un reconocimiento de múltiples fuentes sino que es necesario el disponer de un transmisor por cada evento a detectar. En el caso de productos catalogados como válidos para la detección de diferentes fuentes estos funcionan estableciendo umbrales de intensidad que producen falsos positivos cuando otra fuente de sonido no controlada, cercana a la deseada, supera también este límite. Todos los productos mencionados, no basados en umbral de energía, necesitan tener el elemento transmisor cableado a la fuente de sonido, por lo que eventos acústicos no artificiales (gritos de auxilio, rotura de cristales,…) no pueden ser detectados.

Al contrario que en el reconocimiento de voz, en el reconocimiento de sonidos no-habla no existen amplios estudios entre diferentes algoritmos y parámetros del audio que comparen cuál de ellos da mejores resultados. Este hecho puede deberse a la idea equivocada de que esta línea carece de interés o que, simplemente, se desconoce su verdadera necesidad. No obstante, la concienciación con el colectivo de personas con discapacidad, en este caso personas sordas, hace que la tendencia en I+D+i esté cambiando. Cada vez surgen más proyectos europeos e internacionales que promueven la investigación en este campo y demuestran cómo no todas las técnicas empleadas en el reconocimiento de la voz obtienen iguales niveles de precisión en sonidos no-habla. Se entiende por tanto de interés el desarrollo de sistemas no-habla cuya tecnología permita clasificar un evento acústico en base a su naturaleza cubriendo un amplio rango de sonidos, además de poder ofrecer la información en diferentes interfaces accesibles (visuales o vibratorias) y en diferentes medios (teléfono móvil, televisión, PDA, ordenador,…).

2.7 Conclusiones del Capítulo

Los apartados del presente capítulo aportan información objetiva y contrastada del panorama actual y futuro, tanto a nivel demográfico como de mercado. Los datos extraídos demuestran el latente envejecimiento de la población y, con ello, el aumento del número de personas con discapacidad. El cuidado de personas dependientes y la necesidad de ayudas técnicas que aporten seguridad y permitan realizar actividades cotidianas de la vida diaria prevén un futuro próspero a las empresas que inviertan en nuevas tecnologías de apoyo.

Según los datos del INE, dentro del tipo de deficiencias, la auditiva tiene un peso muy relevante frente al resto, encontrándose en el número dos del ranking de deficiencias presentes en la población española. Cruzando estos datos con los datos de las actividades donde surgen las mayores limitaciones se observa un amplio mercado a explorar en el colectivo de personas con discapacidad auditiva y la vida doméstica.

Hasta la fecha, las mayores inversiones económicas para este colectivo se han centrado en los audífonos donde se generan grandes cifras de facturación y de ventas en todo el mundo. Sin embargo, excluyendo los audífonos, los productos de apoyo basados en las TIC son todavía escasos y poco maduros. A pesar de que los informes elaborados por Fundación Vodafone demuestran que el público objeto es un público con altos conocimientos de informática y habituado a las nuevas tecnologías, no se ha invertido suficiente en investigación capaz de llevar productos tecnológicos e innovadores para este colectivo al mercado.

La mayoría de trabajos de investigación orientados al colectivo de personas con discapacidad auditiva se han centrado en el reconocimiento de voz, dejando a un lado el reconocimiento de sonidos no-habla presentes en el entorno. En la actualidad, son los sistemas de señalización presentados en el apartado anterior los que intentan cubrir este hueco. Sin embargo, estos sistemas poseen varias carencias. En primer lugar, la mayoría de productos necesitan estar cableados. De esta forma, sonidos como, por ejemplo, el timbre de la puerta son reconocidos a través de la señal eléctrica y no a través del análisis de su sonido. Esto implica la necesidad de una instalación previa en la vivienda y la imposibilidad de reconocer eventos no eléctricos (gritos de auxilio, un grifo abierto,…) que puedan surgir en una vivienda. Por otra parte, cada sonido a reconocer posee su propio dispositivo. Aunque existen productos catalogados como válidos para varios sonidos, éstos requieren que el dispositivo esté colocado cerca de la fuente de sonido y se basan en el establecimiento de un umbral de energía que puede provocar falsos positivos en la detección. Ser capaces de reconocer diferentes fuentes de audio con un solo dispositivo, sin necesidad de cableado es un reto todavía lejano que mejoraría la calidad de vida de las personas con discapacidad auditiva.

3. Sonidos No-Habla

El término sonido no-habla puede resultar en muchas ocasiones muy confuso. La falta de una definición concreta hace que su significado posea un carácter abstracto que implique una necesidad de acotación donde se establezcan las pautas que una persona puede aplicar para identificar y referirse a este tipo de eventos acústicos.

Dentro de los diferentes artículos encontrados en la literatura científica, en el área de reconocimiento de sonidos no-habla, la diversidad de definiciones es muy amplia, tanto como el subconjunto de sonidos utilizados.

En este capítulo se asientan las bases que se utilizarán a lo largo de la tesis a la hora de referirse a este tipo de sonidos. En primer lugar, se establecerá la terminología de lo que, en esta investigación, se entiende como sonido no-habla, basándose en las características acústicas y físicas que los definen. En segundo lugar se hace un análisis del estado del arte referente a las taxonomías utilizadas en las diferentes investigaciones. Posteriormente se estudian los diferentes sistemas de apoyo y métodos de visualización encontrados desarrollados para favorecer la independencia de personas con discapacidad auditiva.

3.1 Definición de Sonidos No-Habla

Desde un punto de vista físico, el sonido es un fenómeno vibratorio transmitido en forma de ondas. Las vibraciones pueden ser transmitidas a través de diversos medios elásticos. Entre los más comunes se encuentran el aire y el agua.

Cuando hacemos mención a sonidos no-habla nos estamos refiriendo a aquellos sonidos audibles no generados por el aparato fonador de un ser humano que no tengan una construcción lógica (palabras, frases,…). Uno de los mayores problemas de este tipo de sonidos es que no están acotados frecuencialmente como lo está la voz o la música en general. Si se observa el gráfico de la Figura 7 el área en la que se encuentran este tipo de sonidos abarcaría no sólo el espacio disponible de la voz y de la música sino también el restante hasta alcanzar el umbral de audibilidad.

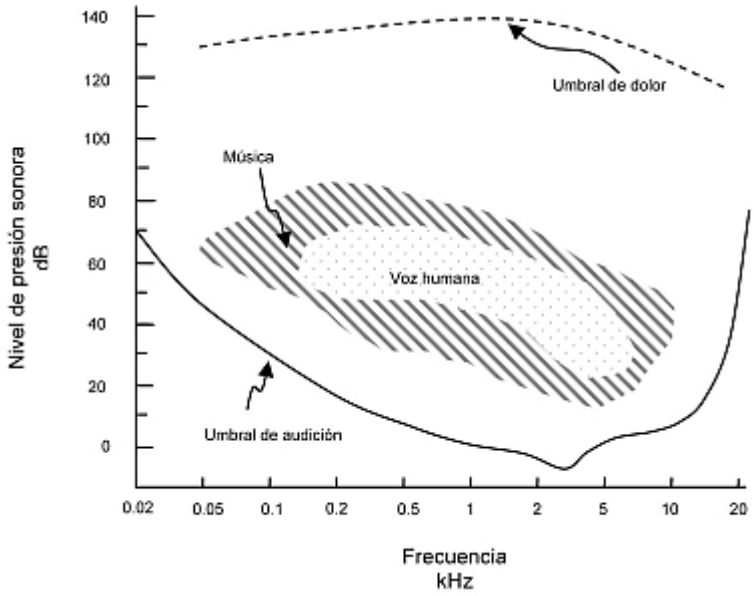


Figura 7 Umbral de audibilidad del ser humano

Existen pocos intentos para definir el término sonido no-habla. Vaderveer [9] define cuatro puntos generales que pueden ayudar a identificarlos:

1. Son producidos por eventos reales.
2. Tienen un significado en virtud de los hechos causales.
3. Son más complejos en su composición que los sonidos generados en laboratorios tales como tonos puros.
4. No son parte de un sistema de comunicación como el habla.

Sin embargo, estos puntos excluyen sonidos como timbres de puerta, teléfonos,… y se centra más en sonidos puramente *ambientales* (en esta tesis utilizaremos el término *ambiental* como sinónimo al término *no-habla*).

Diferentes estudios se contradicen en la forma en la que el ser humano es capaz de percibir los sonidos no-habla. Ballas y Howard [10] indican cómo los seres humanos perciben los sonidos no-habla de forma equivalente al habla. Estudios posteriores en psicoacústica [11] afirman que mientras el hemisferio izquierdo (oído derecho) es el encargado del procesamiento de sonidos del habla, es el hemisferio derecho (oído izquierdo) el que controla el procesamiento de sonidos no-habla. Esto es perfectamente consistente con el talento del lado derecho del cerebro en el reconocimiento de patrones, en asociar inmediatamente un nuevo estímulo a una experiencia previa almacenada en memoria, sin primero tener que descifrar el contexto asociado a ese sonido. Un sonido no-habla es un evento / entidad y el lado derecho del cerebro lo reconoce como tal, muy diferente a una sílaba al azar en una

secuencia determinada de sonidos del habla, que es un mero significante en relación a otros significantes.

Además, si el conjunto a reconocer no está bien acotado, el reconocimiento de estos sonidos puede ser en muchos casos muy complicado de realizar. Investigaciones sobre percepción sonora apuntan que al igual que con el reconocimiento automático del habla, existen sonidos homónimos ("*knight*" y "*night*" para el caso del habla) que no pueden ser determinados sin el contexto de otros sonidos [12]. Si por ejemplo, se produce un sonido metálico precedido de un chirrido, la interpretación semántica podría ser la de un choque de un coche, entendiéndose este chirrido como el sonido de los neumáticos al derrapar por la carretera. Por otra parte, si el mismo sonido metálico se combina con goteo de agua y ráfagas de aire, la interpretación semántica podría ser la del ruido de una máquina en una fábrica.

A pesar de no existir ningún tipo de semántica que ayude a la clasificación de estos sonidos, estudios realizados en el ámbito del reconocimiento demuestran que, trabajando con un conjunto limitado de clases, ciertos parámetros o características de una señal dan información suficiente para que ésta pueda ser clasificada de una forma correcta, incluso cuando la fuente del sonido que se produzca pueda estar en movimiento [13].

3.2 Taxonomías de Sonidos No-Habla

Ballas y Howard [10] hacen hincapié en la falta de un alfabeto fonético de sonidos no-habla. Concluyen que esto es debido a que el habla es producida por un conjunto limitado de diferentes acciones por los mecanismos de fonación de los seres humanos, y en cambio, los sonidos no-habla pueden ser producidos desde un amplio rango de fuentes. Un sonido no-habla puede ser producido por el agua de la lluvia al caer sobre un tejado o por el sonido producido por un taladro mecánico en contacto con el asfalto de la carretera. Es claro por tanto el gran abanico de posibilidades que esto otorga y la complejidad que implica para el área de reconocimiento.

La gran cantidad de sonidos no-habla hace necesario acotar el problema a un conjunto más limitado. Este es el motivo por el cuál usualmente los autores desarrollan diferentes taxonomías de sonidos, estructurando los sonidos en niveles semánticos. El desarrollo de una taxonomía de sonidos ayuda a entender mejor el dominio de los datos [14] e incrementa la precisión y la velocidad de clasificación [15]. Diferentes artículos aplican una jerarquía básica de clasificación entre sonidos habla y no-habla [16], o entre sonidos habla, no-habla y música [17] pero la aplicación de una jerarquía dentro del conjunto no-habla no está tan extendida por la dificultad que esto entraña.

Gaver hace una primera aproximación con una descripción jerárquica de tres niveles [18]. En el primer nivel se encuentran las clases de materiales y las interacciones que pueden causar al sonar. En un segundo nivel los eventos se dividen en tres categorías generales: aquellas que involucran sólidos vibrátiles, sonidos aerodinámicos y sonidos líquidos. Finalmente se muestran los eventos producidos a bajo nivel, definidos por la simple interacción que pueden causar los sólidos, gases y líquidos al sonar.

Cowling sigue la misma idea de Gaver para la implementación de su propia taxonomía [12]. Con una jerarquía de niveles, en el primer nivel, el sistema identifica los sonidos pertenecientes a una de seis clases. Cada clase engloba la interacción de dos objetos, cada uno relacionado con los tres estados del entorno (sólido, líquido y gaseoso): sólido-sólido, sólido-líquido, sólido-gas, líquido-líquido, líquido-gas, gas-gas. El siguiente nivel del sistema clasifica los sonidos basándose en las características de los materiales. Por ejemplo, un sólido puede ser madera, cristal, metal, etc.

Gerhard, por su parte, hace una clasificación basada en contexto [14], sin tener en cuenta el material con el que se produce el sonido. Distingue entre sonidos naturales, artificiales y ruido. Es una clasificación más cercana y fácil de entender para el ser humano, sin embargo, pierden importancia las características físicas de cada clase para el reconocimiento.

Utilizando también un enfoque basado en la acústica de la señal Gygi y Shafiro proponen su propia taxonomía en [19] siguiendo los resultados obtenidos de estudios previos en búsqueda de similitudes entre sonidos [20]. En un primer nivel separan las señales armónicas y no armónicas. En un segundo nivel se refleja la continuidad o discontinuidad. A pesar de que se indica que la estructuración planteada no cubre todo el rango posible de sonidos ambientales afirman que es una estructuración bastante simple que tiene en cuenta un porcentaje muy elevado de eventos sonoros necesarios para un sistema de clasificación efectivo.

Tal como afirma Temko [21], la concepción de una taxonomía de sonidos es subjetiva y depende en gran medida del dominio de clasificación escogido. Para una buena clasificación es necesario conocer el conjunto de sonidos que pueden ser encontrados en las pruebas, las interacciones que estos pueden producir y las características más descriptivas de los mismos. En sus estudios, Temko combina dos tipos de taxonomías: una taxonomía de sonidos acústica para descripciones generales de sonidos y una taxonomía semántica para una tarea específica.

Centrarse en el colectivo de personas con discapacidad auditiva hace que el abanico de sonidos a identificar se reduzca sustancialmente. Es comprensible la necesidad que una persona sorda puede tener de reconocer el timbre de la puerta cuando alguien está llamando, sin embargo, reconocer el sonido que produce un papel al ser manipulado no parece tan evidente que sea de necesidad primordial para ella.

3.3 Sistemas Asistivos de Reconocimiento de Sonidos No-Habla

Aunque no existen productos comerciales, sí que podemos encontrar en la literatura científica descripción de prototipos y proyectos de investigación que ahondan en el empleo de sonidos no-habla en aplicaciones asistivas. A continuación se hace un breve resumen de los más destacados.

En [22] se desarrolla un sistema de teleasistencia móvil Scribe4Me para personas sordas con una arquitectura cliente-servidor basada en teleoperadores. La herramienta fue diseñada para mejorar el conocimiento acústico que rodea a una persona en cualquier lugar. Cuando un botón de la aplicación es pulsado la información grabada de los últimos treinta segundos es enviada a un centro de teleasistencia donde un operador lo recibe y transcribe el contenido para devolver la información a la persona con problemas auditivos en un mensaje de texto. Las ventajas de este sistema son notables. Al depender de una persona humana es capaz de transcribir todo tipo de señales acústicas, ya sean voz como sonidos ambientales. Además puede ser utilizado en cualquier entorno y su precisión depende de la capacidad auditiva humana, en general, superior a la de una máquina. Las desventajas son dos principalmente. En primer lugar, el depender de personas externas implica un coste elevado y un tiempo de respuesta de entre tres y cinco minutos. Además, el sistema sólo transcribe bajo petición de la persona con problemas. En el caso de que un usuario de la aplicación no se diera cuenta de una situación importante y no diese al botón de "*transcribir*", esa información se perdería en el tiempo. Este último punto es fundamental y de suficiente peso para valorar el uso de sistemas automáticos capaces de ofrecer una funcionalidad aceptable y útil para las personas con discapacidad auditiva.

Haciendo uso de técnicas de procesamiento de señal, en [23] se presenta un sistema de reconocimiento de sonidos optimizado para aquellos sonidos de carácter mecánico para asistir a personas con discapacidad auditiva. La arquitectura que detallan consta de un micrófono, un ordenador encargado de procesar la señal de audio dada y un reloj de pulsera con *display* para mostrar la información. Las pruebas fueron realizadas con sonidos electrónicos: timbres de puerta, teteras y teléfonos. Un estudio más amplio fue dado en [24], [25] y en [26] donde los autores amplían el reconocimiento a siete tipos de sonidos: golpes de puerta, teléfonos, pasos, platos, cerraduras, cristales rompiéndose y gritos. El sistema consta de una interfaz de audio multicanal con cinco micrófonos distribuidos por las diferentes habitaciones de la casa-laboratorio y un PC encargado de todo el procesamiento. Aunque los autores orientan la aplicación al seguimiento de pacientes en hospitales también indican cómo es posible su aplicación en diferentes áreas. Estudios posteriores de los mismos investigadores ampliaron el número de clases a nueve [27]. En [28], motivados por la misma idea, desarrollaron un sistema del mismo tipo ampliando el número de clases a reconocer a diecinueve,

utilizando diferentes algoritmos en el reconocimiento. Experimentos similares se encuentran en [29] donde clasifican once clases diferentes de sonidos mezclándolos con diferentes niveles de ruido para comprobar la robustez de las técnicas utilizadas en el procesado. [30] incorpora el reconocimiento en un robot humanoide para acompañar a las personas dependientes con 12 clases de eventos diferentes.

Aunque más orientado a la monitorización de personas dependientes que a las personas con problemas auditivos, en [31] y [32] buscan la identificación de los sonidos producidos en el baño. Tal y como exponen, un elemento clave en la prevención de enfermedades cognitivas es el entendimiento de los patrones de conducta de las personas en sus *Actividades de la Vida Diaria* (ADL). Un tipo de actividad de la vida diaria importante tanto para cuidadores como médicos es la concerniente a la higiene personal. Los objetivos del sistema son la detección, e identificación para la generación de informes personales de higiene. Como salida presenta una lista de actividades que han ocurrido en el baño con detalles asociados incluyendo el tiempo de inicio de las ocurrencias y su duración, y en un informe resumido, indica si la persona siguió un patrón de comportamiento normal (ej. si se ducha una vez al día, si usa el baño cierto número de veces al día, etc). Como sistema de monitorización de dietas, en [33] se desarrolla un sistema que detecta el tipo de alimento que una persona está comiendo entre los siguientes cuatro alimentos: patatas fritas, manzana, pasta y lechuga. Este sistema se percibe importante especialmente para personas mayores con deterioros cognitivos con las que haya que tener especial cuidado con su alimentación. En el artículo se muestra cómo sonidos producidos al masticar pueden ser obtenidos desde un micrófono situado en el oído con buena calidad. Debido a que la mayoría de las señales acústicas generadas por interacción mecánica de dientes y comida durante la oclusión son transmitidas por la conducción de los huesos, estos sonidos son más fuertes que las señales de voz, pudiéndose discriminar fácilmente.

Fuera del entorno del hogar se encuentran varios estudios enfocados en el área de la conducción con el fin de garantizar una mayor seguridad a los conductores con problemas auditivos trabajando en la detección de sirenas de vehículos de emergencia [34], [35], [36]. La problemática en este caso viene dada por la complejidad de una detección en movimiento así como el aislamiento de todos los ruidos de fondo, ya sean motor de coche o sonidos externos de la ciudad o autopista, que reducen la fiabilidad y exactitud de las técnicas de reconocimiento. A su vez, en [37] se analizan diferentes tipos de sonidos de alerta de la calle: claxon, campanas, timbres de tranvía y teléfonos.

3.4 Sistemas Asistivos de Visualización de Sonidos No-Habla

Con un enfoque más centrado en el usuario, fuera de la etapa más técnica de reconocimiento, diferentes estudios analizan la forman en la que la información debe mostrarse a las personas con discapacidad auditiva. Primeras aproximaciones fueron dadas en [38] donde se evaluaron dos prototipos con los que mostrar información de sonidos ambientales a personas sordas mediante el uso de espectrogramas y de anillos dentro de un mapa indicando posición e intensidad. Para el estudio se contó con la colaboración de ocho participantes que testearon el sistema. De la investigación destaca la mayor aceptación del *display* en forma de mapa y anillos que el del espectrograma. En el experimento no se tuvo en cuenta la posibilidad de dar la información del sonido, sino que era el usuario el que tenía que hacerlo, basándose en los prototipos propuestos.

Más prototipos fueron desarrollados en [39] y en [40], añadiendo la opción de indicar la clase de sonido en el *display*. Aunque, entre uno de los prototipos se encontraba un boceto similar al mapa con anillos del estudio anterior, los usuarios se decantaron por aquellos que contenían iconos representando el evento. Los participantes indicaron que esta forma les ofrecía una forma más rápida de reconocer el sonido. En [41] las conclusiones fueron similares, siendo la opción de visualización mediante icono una de las más votadas.

El interés de poder disponer de un sistema de reconocimiento de sonidos ambientales también se vio confirmado en [42], [43]. Pensando en *displays* posicionados en el techo, mediante la utilización de encuestas, cuestionarios online y workshops, se diseñaron distintos prototipos. Del estudio derivan diferentes pautas de diseño para este tipo de sistemas que complementan las obtenidas en [39] y [40].

3.5 Conclusiones del Capítulo

El amplio abanico de eventos acústicos que se enmarcan dentro de la definición de sonidos no-habla hace difícil establecer una metodología común a la hora de crear taxonomías o acotar características acústicas y físicas específicas que puedan ser aplicadas a todos los sonidos.

Esta gran diversidad hace que surjan muy variados proyectos y aplicaciones de interés dentro del área de las tecnologías asistivas y de la salud en general. Así como las tecnologías del reconocimiento del habla han ido incrementando sus capacidades y su aplicabilidad, las tecnologías de reconocimiento de sonidos no-habla parece que pueden seguir su camino. De momento la mayoría de los esfuerzos se centran en la etapa de investigación donde las pruebas y experimentos están muy acotados a un subconjunto reducido de eventos y unas determinadas condiciones acústicas de

laboratorio. Dentro de los sistemas específicos para personas con discapacidad auditiva todavía es necesario profundizar en mayor medida. A nivel técnico es necesario trabajar con eventos de mayor interés para este colectivo y analizar cómo estos se comportan con los algoritmos de reconocimiento actuales y cómo de lejos se encuentran de la etapa de mercado. Del mismo modo, es necesario realizar análisis más profundos y con un mayor número de participantes que faciliten al investigador la labor de diseño de los sistemas adaptados a los usuarios finales.

4. Técnicas de Reconocimiento de Sonidos

En este capítulo se presenta el estado del arte de las técnicas de reconocimiento de sonidos no-habla actuales.

En el primer apartado se describe brevemente el funcionamiento general de estos sistemas y las etapas que constituyen su arquitectura. En el segundo apartado se presentan las bases de datos de eventos acústicos existentes dentro del campo de la investigación. Posteriormente se hace un análisis de la etapa de extracción de características acústicas de estos sistemas, aportando información sobre el tipo de enventanado, los parámetros acústicos más utilizados y los mecanismos de reducción de variables más frecuentes. Posteriormente se presentan los algoritmos de clasificación más habituales y las técnicas de detección utilizadas en sistemas de audio continuo. Se concluye el capítulo con un resumen de todo lo mencionado en el capítulo.

4.1 Esquema General de un Sistema de Reconocimiento de Sonidos

Según [24], [26], la arquitectura general de un sistema de reconocimiento de sonidos no-habla se muestra en la Figura 8. En un primer paso la señal es captada y enviada al módulo de detección. El módulo de detección es el encargado de analizar la señal y buscar segmentos de audio que puedan contener eventos que estén dentro del conjunto de señales acústicas a identificar. Será el módulo de clasificación en un siguiente paso el que decida a qué clase de sonido pertenecen los eventos (timbre de la puerta, llanto de bebé,…), para notificárselos al usuario. Sin embargo, como se verá más adelante, existen técnicas en las que la detección y la clasificación se realizan al mismo tiempo.

En la actualidad, en la mayoría de investigaciones, los módulos de detección y clasificación están basados en modelos probabilísticos que reciben como entrada características acústicas extraídas de la señal y, en base a patrones creados con muestras de entrenamiento previo, deciden la clase más probable.

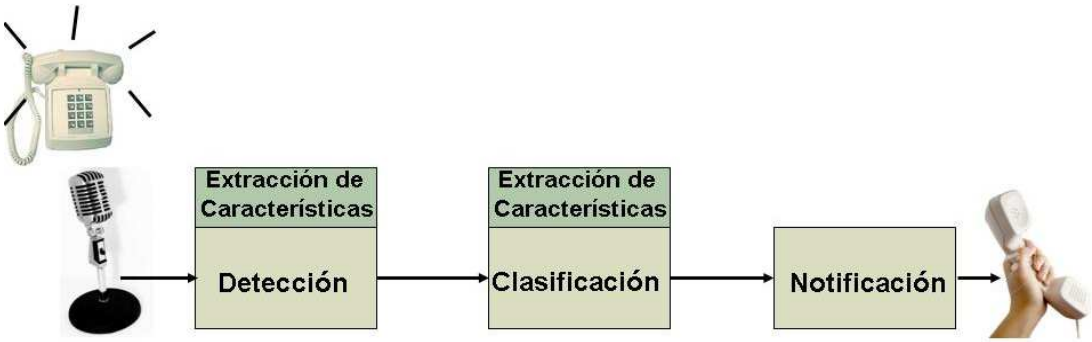


Figura 8 Arquitectura general de un sistema de reconocimiento de sonidos

Si se conoce el tipo de sonidos que se quieren detectar y estos están acotados en frecuencias se puede introducir al inicio un nuevo módulo conocido habitualmente como *pre-procesado*. En este módulo se aplica un filtrado de la señal para eliminar ruidos que puedan provocar confusiones al clasificador. Al contrario que la voz, la cual está acotada entre aproximadamente los 200 y los 8000 Hz, los sonidos no-habla pueden ser de muy distinta índole y éste es el motivo por el cual tampoco sea clara la aplicabilidad de un filtrado idóneo.

Como complemento a esta arquitectura, existen estudios que hacen uso de varios micrófonos en el sistema con el objetivo primordial de localizar la fuente de audio [44], [45], [46], [47]. Esto implica incorporar un módulo adicional que sea el encargado de separar el audio en base a técnicas de triangulación. Sin embargo, en esta tesis nos centraremos sólo en el reconocimiento. A pesar de la importancia que la localización puede suponer en entornos desconocidos para el usuario, en el entorno doméstico la persona con discapacidad auditiva conoce el lugar donde están los elementos que generan el sonido. Con la indicación de que han llamado al timbre de la puerta o que ha terminado el microondas la persona es capaz de asociar esa información a una ubicación específica.

4.2 Bases de Datos

En las etapas de detección y clasificación, como se ha mencionado, existe una tendencia hacia el uso de técnicas probabilísticas. Estas técnicas se basan en el entrenamiento del sistema mediante bases de datos de sonidos. Se requiere por tanto el uso de bases de datos que contengan suficiente cantidad de muestras de sonidos para que los modelos que se obtengan tengan información suficiente para discernir entre todas las clases del conjunto de sonidos a evaluar.

Los avances tecnológicos en equipos y métodos de grabación han conducido a un aumento en el número de grabaciones de sonidos ambientales que se han producido para fines comerciales, científicos o artísticos [48]. El avance de Internet y el

incremento en velocidad de la banda ancha ha hecho posible la simplificación del proceso de compartición y distribución de los sonidos grabados. Consecuentemente, un gran número de grabaciones de sonidos ambiente están disponibles para el público en forma de CDs comerciales y bases de datos electrónicas accesibles a través de la red. Sin embargo, la dispersión de éstas, el escaso número de muestras de una misma clase y el desconocimiento del método de grabación utilizado no las hacen válidas para su uso en la investigación.

En el campo del reconocimiento de sonidos no relacionados con el habla, la base de datos comercial de investigación más frecuentemente encontrada en los diferentes artículos estudiados, es la "*Sound Scene Database in Real Acoustical Environments*" (comúnmente referida como RWCP "*Real World Computer Partnership*") cuyos datos fueron grabados de 1998 a 2000 [49]. Usando un único micrófono, las señales acústicas de aproximadamente 100 tipos de sonidos fuente fueron grabadas en una habitación insonorizada. Los eventos acústicos son clasificados en tres categorías: sonidos de colisión, sonidos de acción y sonidos característicos.

Por otro lado, en 2004 tuvo lugar la primera evaluación CHIL "*Computer in the Human Interaction Loop*". Durante este evento uno de los aspectos tecnológicos evaluados fue el correspondiente a la clasificación de eventos acústicos. El objetivo de la clasificación de eventos acústicos en el proyecto CHIL fue adquirir conocimiento del entorno monitorizando el sonido ambiente. En esta evaluación, y en sus posteriores (CHIL 2005, CLEAR 2006, CLEAR 2007), varias bases de datos de sonidos no-habla fueron creadas y puestas a distribución. El entorno de las grabaciones consistió en una sala de reuniones donde fueron grabados sonidos tales como apertura y cierre de puertas, teléfonos, aplausos, movimiento de sillas,….

Posteriormente, centrada en sonidos de la oficina, en 2013 tuvo lugar la evaluación IEEE AASP [50] donde se incluyó también reconocimiento de escena. Sin embargo, posiblemente por su más reciente creación, no ha tenido muchas referencias y comparaciones en artículos publicados. Es por eso que en esta tesis las dos bases de datos que se evalúan son las dos anteriores: RWCP y CHIL.

Tanto el corpus de sonidos como la metodología de grabación son totalmente distintos entre las bases de datos de RWCP y CHIL. Cabe destacar que en RWCP los sonidos fueron grabados independientemente en una cámara insonorizada mientras que en CHIL los sonidos están grabados dentro de sesiones continuas (con la posibilidad de analizar su detección) y en un entorno real como es una sala de reuniones. Una descripción más detallada de estas bases de datos se encuentra en el capítulo 6.

Aunque el uso de bases de datos es de gran ayuda a la hora de analizar y mejorar algoritmos, éstas sólo expresan un subconjunto muy pequeño del posible. Esto implica que, en pruebas finales, cuando se trabaja en entornos no controlados se producen

caídas considerables en el rendimiento del sistema. El funcionamiento de los sistemas de reconocimiento sufre degradaciones importantes cuando se incorpora ruido de fondo a las grabaciones o cuando las condiciones acústicas de los datos de entrenamiento y los datos de test son muy diferentes. Esta situación es habitual en los sistemas que funcionan en aplicaciones de tiempo real en las que, por una parte, la fuente de sonido suele estar contaminada por la presencia de diversos ruidos de fondo, separación de los micrófonos, reverberaciones, movimiento, etc., y, por otra parte, resulta imposible disponer de datos de entrenamiento grabados en todas las condiciones acústicas posibles. Estas circunstancias han dado lugar, en los últimos años, a una extensa investigación (mayoritariamente en el ámbito de la voz) sobre diversas técnicas cuyo objetivo primordial es el de proporcionar una gran robustez a los sistemas de reconocimiento inmersos en un entorno acústico diferente al existente en la fase de entrenamiento. Para tal fin, se pueden encontrar artículos que comparan la robustez de sus algoritmos mezclando los sonidos a detectar con ruido de fondo de diferentes entornos (como por ejemplo el aire acondicionado [51]) o ruido artificial generado (por ejemplo ruido HIS [52]). Otras técnicas existentes se basan en adaptar el sistema "online" al ruido de fondo que se esté detectando, reentrenando los modelos dinámicamente.

4.3 Extracción de Características Acústicas

La extracción de características acústicas que sean discriminantes requiere un intenso trabajo de análisis de la problemática y la señal. Una sola característica difícilmente será capaz de dar información para clasificar todas las clases de sonidos, y cuando el conjunto obtenido es amplio, se hace necesaria la aplicación de métodos de validación basados en selección y búsqueda de características que requieren la realización de varias pruebas con los datos [53] [54].

En los siguientes subapartados se explican los puntos más críticos de este módulo, empezando con la segmentación de la señal en ventanas, continuando con una revisión de las características acústicas más importantes y finalizando con las técnicas utilizadas para la reducción de variables.

4.3.1 Enventanado

La división de la señal en pequeños trozos (conocido como "*enventanado*" o "*segmentación*") es un proceso necesario siempre que se trabaje en tiempo real. Además, este proceso también es necesario debido a que, en general, un sonido no-habla (al igual que ocurre con una señal de habla) no es una señal estocástica estacionaria. Sin embargo, si ésta se divide en porciones pequeñas, puede ser considerada como tal. El enventanado persigue precisamente esto al dividir la señal en

partes más pequeñas. Habitualmente su tamaño es fijo, de entre 10 a 100 milisegundos, aunque existen estudios como [55] que trabajan con ventanas variables estimadas mediantes técnicas de lógica difusa. Existen muchas y diferentes funciones de enventanado pero en el campo del procesamiento de audio la ventana de *Hamming* es la más utilizada. Estas ventanas suelen extraerse aplicando un solapamiento entre ellas que permite a los investigadores 1) tener más muestras con las que entrenar y testear y 2) que no se produzca una pérdida de información en los bordes de los segmentos.

Artículos como [56] aplican un enventanado de forma jerárquica. Por una parte, la señal es dividida en ventanas grandes (*superframes*). Seguidamente cada *superframe* es dividido en ventanas más pequeñas (*frames*) de las que se extraen las características acústicas. Del *superframe* se extraen cálculos estadísticos como la media y desviación típica de los *frames* contenidos en él. Igualmente, [57] analiza la señal de audio a dos niveles de granularidad para la detección de eventos anómalos como disparos, gritos o rotura de cristales. Los autores explican cómo esta arquitectura en el diseño de sistemas de clasificación viene motivada por la simple observación de que la mayoría de descriptores de audio disponibles en la literatura están definidos a nivel de *frame* mientras que los eventos de interés están típicamente caracterizados por una mayor duración. [58] introduce la idea de los *Descriptores de Unidad Acústica* (AUD) donde la división de la señal se hace en base a unidades atómicas extraídas de forma no supervisada. Los autores asumen la posibilidad de poder definir la mayoría de sonidos con un conjunto limitado de estas unidades capaz de representarlos a todos ellos de una manera precisa.

4.3.2 Características Acústicas

Una vez se tiene dividida la señal en ventanas, será de estas pequeñas porciones de audio de las que se obtenga la información a enviar al modelo de clasificación. Esta información son los parámetros o características acústicas descriptivas del audio. La pregunta que surge a los investigadores es: ¿qué característica hace diferente un sonido, como por ejemplo sonidos de pasos, del resto de sonidos que se puedan encontrar en un entorno dado?

Según Cowling [12], la extracción de características puede ser dividida en dos tipos: *estacionarias* (basadas en frecuencia) y *no estacionarias* (basadas en tiempo-frecuencia). Las características estacionarias producen un resultado general detallando las frecuencias contenidas en la señal completa. Con la extracción de características estacionarias no hay distinción de dónde estas frecuencias ocurren dentro de la señal. Por el contrario, las características no estacionarias dividen la señal en unidades discretas de tiempo. Esto permite identificar las frecuencias en áreas particulares de la señal. Sin embargo, el mismo autor indica cómo la gran mayoría de características

estacionarias (como es el caso de MFCC o LPC) pueden ser consideradas como *pseudo-estacionarias* porque ellas dividen también la señal en segmentos de tiempo. En esta clasificación no se incluyen características que se encuentren tan sólo en el dominio del tiempo. Características como Zero Crossing Rate o Short Time Energy quedan fuera de esta categorización. Es por esto que otros autores como Mitrovic [59] hacen mención a taxonomías más simples basadas simplemente en dos divisiones: *características basadas en tiempo* (utilizando los valores de intensidad) o *características basadas en frecuencia* (cuando a la señal se le aplica una transformada de Fourier, Wavelet, etc que la transforma al espacio frecuencial).

En [14], Gerhard divide las características en dos categorías: *perceptuales* y *físicas*. Las características perceptuales están basadas en la forma en la que los humanos escuchan un sonido. Ejemplos de características perceptuales son "*pitch*", "*timbre*" y "*ritmo*". Las características físicas están basadas en propiedades estadísticas y matemáticas de las señales. Ejemplos de características físicas son la frecuencia fundamental, Zero Crossing Rate y Energía. Se indica cómo algunas características perceptuales están relacionadas con características físicas, tal y como la frecuencia fundamental está relacionada con el pitch, y el timbre con el contenido espectral.

Al margen de clasificaciones, a continuación se presentan los parámetros acústicos más utilizados en el reconocimiento de sonidos ambientales, algunos de los cuales son comúnmente utilizados en el reconocimiento de instrumentos musicales y la gran mayoría de ellos son utilizados en el reconocimiento del habla.

- *Short Time Energy (STE):* Es la energía total comprendida en una ventana de audio [60].
- *Zero Crossing Rate (ZCR):* Calcula las veces que pasa por cero la señal en un cierto intervalo de tiempo [60], [61], [62].
- *Root Mean Square (RMS) Level:* Medida que cuantifica la intensidad de la señal. Su valor es calculado obteniendo la raíz cuadrada de la media de las muestras al cuadrado [61], lo que viene a ser la raíz cuadrada de la Short Time Energy.
- *Band Energy ratio:* Establece la relación entre la energía de las bajas y de las altas frecuencias de la señal [61].
- *Fundamental Frequency:* Es la frecuencia fundamental de la señal. Aplicable cuando los sonidos a reconocer son periódicos [63].
- *Spectral Centroid:* Calcula la frecuencia en la que reside el centro de la energía espectral de la señal analizada. Perceptualmente mide cómo de brillante es el sonido que se escucha [61].

- *Roll-Off Point (RF):* Calcula el ancho de banda de frecuencias en la que se concentra el 85% o 95% (este valor cambia dependiendo de la implementación) de la energía del espectro [61].

- *Spectral Bandwidth:* Una medida de la extensión del espectro alrededor del Spectral Centroid. Se calcula a través de la media ponderada de las distancias entre los componentes espectrales y el Spectral Centroid [61].

- *Mel Frequency Cepstral Coefficients (MFCCs):* Se computan los coeficientes cepstrales utilizando una escala logarítmica tipo Mel [60], [61], [64]. Algunos autores analizan diferentes modificaciones de estos parámetros teniendo en cuenta el espectro de los sonidos a clasificar. Como ejemplo, en [65] modifican el banco de filtros para no considerar las bajas frecuencias.

- *Linear Frequency Cepstral Coefficients (LFCCs):* A diferencia que los MFCCs, estos coeficientes utilizan una escala lineal para crear su banco de filtros [60], [64].

- *Linear Prediction Coefficients (LPCs):* Utilizados frecuentemente en sistemas de comunicación para codificar y decodificar señales de voz. El tracto vocal es modelado mediante un filtro *todo polos* de respuesta infinita impulsiva. Los coeficientes LPC son los coeficientes de la combinación lineal de este filtro, siendo coeficientes que representan la envolvente espectral de la señal [64].

- *Wavelet Coefficients:* Obtenidos de la transformada Wavelet capaz de trabajar en el dominio frecuencial con varias resoluciones. Indican qué tan parecida es la señal a analizar de una función madre, siendo los portadores del detalle de la señal [60], [66].

Al margen de las citadas características, también encontramos trabajos que utilizan características acústicas más sofisticadas, como el descrito en [67]. En él los autores introducen el uso de características de fluctuación espectro-temporales basadas en el algoritmo *Harmonic Percussive Sound Separation* (HPSS) donde se extraen tres espectrogramas que contienen los componentes estacionarios, transitorios e intermedios de la señal. Otros estudios aplican conocimientos del campo de la percepción auditiva humana usando características acústicas extraídas de un análisis del espectro denominado "*cocleograma*" que imita la respuesta de la membrana basilar [68], [69], [70], [71]. Guiados también por estudios de psicoacústica, en [72] se usan los llamados *Amplitude Modulation Spectrograms* (AMS) motivados por la importancia de modulaciones temporales para el reconocimiento de objetos acústicos. AMS representa una descomposición de la señal a través de las dimensiones de frecuencia, modulación y tiempo, y son computadas por una descomposición espectral de subbandas. Esta técnica es aplicada para detectar sonidos que no han sido

previamente recogidos en la base de datos, aunque los experimentos presentados tan sólo han tratado con un corpus de 4 clases de sonidos. [73] utiliza características para medir el rango de variación de la frecuencia fundamental postulando la hipótesis de que en muchos sonidos no-habla éste cambia abruptamente sin permanecer estacionario. En este caso, estos parámetros obtienen resultados muy similares a los MFCCs añadiendo información adicional si se complementan. Otros trabajos [74], [75] utilizan características basadas en la técnica *Matching Pursuit* (MP) aplicando funciones de *Gabor*, pero en este caso, la aplicación es más dirigida al reconocimiento de entornos que al de sonidos aislados. Sin embargo, el cálculo computacional que estas técnicas requieren es elevado. La implementación de estos algoritmos en dispositivos móviles en tiempo real requeriría unas prestaciones hardware alejadas de las actuales disponibles en el mercado. El procesamiento de todos los filtros y técnicas aplicadas incrementan considerablemente el tiempo de respuesta del sistema, haciendo inviable su uso.

En los últimos años, técnicas aplicadas en el reconocimiento de canciones [76] han tenido gran impacto en la literatura científica. Estas técnicas buscan encontrar puntos clave dentro del espectrograma que definan un patrón en dos dimensiones del sonido (tiempo – frecuencia) [77]. No obstante, la base de datos con la que se trabaja contiene un número muy limitado de clases. Además, estas técnicas están basadas en la realización de conexiones a lo largo del tiempo y esto implica que la duración del sonido no puede ser muy corta para su correcto reconocimiento.

4.3.3 Reducción de Variables

Cuando el número de parámetros utilizados es muy elevado, los clasificadores sufren problemas de sobredimensionamiento. Para mitigar este efecto contraproducente es importante hacer una reducción previa de variables para evitar parámetros redundantes y una alta complejidad en el clasificador.

En los sistemas de clasificación se utilizan mecanismos de reducción de parámetros que permiten separar aquellas características que realmente aportan información descriptiva, de las que sólo añaden ruido a la parametrización [78]. Con ello también se consigue reducir la dimensión de la parametrización, lo que reduce la carga computacional. El detectar los parámetros verdaderamente discriminantes puede permitir además comprender mejor la relación entre los sonidos ambientales y las características acústicas de la señal de audio. Existen dos aproximaciones: *Construcción de Características* y *Selección de Conjuntos de Características* [79]. A continuación se hace un breve resumen de ellas.

4.3.3.1 Construcción de Características

Las técnicas de Construcción de Características buscan las posibles relaciones entre las variables y devuelven un nuevo conjunto de variables, combinando las originales.

Las técnicas de análisis de componentes, por ejemplo *Principal Component Analysis* (PCA), son un caso de amplio uso, en que el espacio trasformado se forma de elementos ortogonales y aunque no se considera ningún modelo de aleatoriedad, en la estimación de las respectivas matrices de covarianza, es conocido que la efectividad de la técnica es más alta entre más se ajuste la condición de Gaussividad para las distribuciones de las características [80].

4.3.3.2 Selección de Conjuntos de Características

La Selección de Conjuntos de Características busca las variables más significativas, devolviendo un subconjunto del conjunto de las variables originales.

Los algoritmos de selección de parámetros se suelen dividir en dos grupos, los *wrappers* y los *filtros* [81]. A continuación se hace una breve descripción de ellos.

Los ***métodos wrapper*** utilizan un clasificador como mecanismo para estimar la precisión que puede alcanzar un cierto conjunto de parámetros. Esta estimación se obtiene entrenando y evaluando el clasificador con cada conjunto de parámetros considerado, generalmente mediante algún tipo de validación cruzada. Una vez estimado el error de clasificación cometido por cada conjunto de parámetros, se selecciona aquél que proporciona mayor tasa de aciertos. En este tipo de métodos existen dos aproximaciones básicas:

- *Selección Forward*: Se empieza sin ninguna variable y se van añadiendo una a una. En cada paso se añade la que más hace decrecer el error hasta que no se encuentren mejoras significativas.

- *Selección Backward*: Se empieza con todas las variables y se van eliminando una a una. En cada paso se elimina la que más hace decrecer el error hasta que no se encuentren mejoras significativas.

Este tipo de métodos proporcionan conjuntos de parámetros que alcanzan una gran precisión para el clasificador considerado. Sin embargo, los resultados no siempre son generalizables a otros clasificadores. Además, son algoritmos de gran carga computacional, debido a la necesidad de entrenar el clasificador por cada conjunto de parámetros analizado.

Los ***métodos filtro*** seleccionan los parámetros en función de su distribución y su relación con el resto de parámetros y con la clase objetivo. Como resultado, el conjunto seleccionado no está optimizado para ningún clasificador, y generalmente

proporciona una precisión comparable con diferentes clasificadores. También requieren una carga computacional significativamente menor que el método *wrapper*, al no tener que entrenar un clasificador en cada etapa. Estos métodos emplean medidas de ganancia de información, distancia o consistencia, entre el parámetro y la clase [82]; sin embargo, debido a que miden la importancia de cada parámetro en forma aislada, no pueden detectar si existen parámetros redundantes, y tampoco son capaces de determinar si la combinación de dos o más parámetros, aparentemente irrelevantes en forma aislada, se pueden transformar en relevantes [83]. Algunas aproximaciones a este problema son planteadas en [84] en donde utilizan el algoritmo denominado *Minimum Redundancy Maximum Relevance* (MRMR).

4.4 Clasificación de Sonidos

Las técnicas de clasificación de audio permiten decidir a qué clase de entre un conjunto de clases previamente definido pertenece cada observación (conjunto de características acústicas). Si en el proceso de reducción de variables se ha utilizado un método *wrapper* la técnica de clasificación está ligada a dicho proceso. Aunque existen estudios que aplican algoritmos semi-supervisados [85] o técnicas deterministas [76], la gran mayoría de ellos aplican algoritmos supervisados en los que es necesario que los sonidos sean etiquetados previamente.

4.4.1 Técnicas de Clasificación de Sonidos

Clarkson [86] diseñó un sistema *wearable* para reconocer diferentes tipos de sonidos tales como habla, sonidos del coche, etc. usando *Modelos Ocultos de Markov* (HMM) como método de clasificación. Similarmente, Ma [87], [88] usó HMM con topología de izquierda a derecha. Su base de datos estaba compuesta de diez tipos diferentes de sonidos de 3 segundos de duración. La precisión global obtenida fue aproximadamente del 92%. Eronen aplicó *Modelos de Mezclas Gaussianas* (GMM) y *Redes Neuronales* (NN) usando diferentes parámetros para clasificar 26 eventos de audio [89]. Selina [90], [91] clasificó varios tipos de sonidos de cinco diferentes entornos usando características en el dominio temporal y frecuencial con *K-Vecinos Más Cercanos* (K-NN), GMM y *Máquinas de Vectores Soporte* (SVM).

Ellis utilizó una red neuronal orientada a la detección de alarmas [92]. En este trabajo se hace uso del *perceptrón multicapa* con una única capa oculta con 100 neuronas ocultas y 2 neuronas de salida que corresponden a las clases "*alarma*" y "*no alarma*". Cowling y Sitte [93] aplicaron una red *Learning Vector Quantization* (LVQ) para el reconocimiento de pasos. Dufaux y Besacier [94], [95] usaron modelos HMM para reconocer sonidos impulsivos (rotura de cristales, gritos, disparos y explosiones).

Valenzise y Gerosa clasificaron disparos y gritos usando GMM [96], [97]. Claver [98] también intentó detectar explosiones de diferentes armas de fuego utilizando GMM. Usando HMM de dos estados, Ntalampiras [99] clasificó gritos, disparos y explosiones en el metro. Regunathan y Kim [100], [101] clasificaron varios sonidos para vigilancia en el ascensor usando nuevamente GMM.

A partir del año 2007 las Máquinas de Vectores Soporte (SVM) han tomado mucha fuerza en el área de reconocimiento de sonidos no-habla. Aunque Temko no fue el primero en utilizarlas [102] [103], tras sus trabajos [21], [104] el número de investigaciones encontradas con este algoritmo de clasificación han tenido un mayor impacto. Junto a los trabajos con HMM [105], [106], [107] y GMM [108], [109], [110], [91] estos son los principales clasificadores utilizados en la literatura científica.

En la Tabla 10 se listan los artículos más relevantes desde 2010 indicando el algoritmo/s de clasificación utilizado/s.

Artículo	Clasificador
[108], [109], [110], [111]	GMM
[67], [112], [113], [114], [46], [115], [84], [116], [45], [30]	HMM
[117], [72], [69], [70], [65], [73]	SVM
[118], [119], [120]	SVM-GMM Híbrido
[56]	Random forest
[68]	NN
[36]	Part-based Models
[77]	Codebook

Tabla 10 Artículos y clasificadores más relevantes a partir de 2010

4.4.2 Modelos de Mezclas Gaussianas (GMM)

Los modelos de mezclas gaussianas (GMM) fueron los elegidos para esta tesis por su bajo coste computacional y su facilidad de entrenamiento necesario para aplicaciones en tiempo real como se demostrará en los capítulos 6 y 7.

Los GMM son combinaciones de distribuciones normales o funciones de Gauss. La fórmula que describe la función de densidad de probabilidad del tipo gaussiana multivariable se describe en la fórmula 1:

$$N(x;\mu,\Sigma)=\frac{1}{\sqrt{2\pi^D\,|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \quad (1)$$

Donde x es el vector de datos de entrada de dimensión D, μ es el vector de medias y Σ la matriz de covarianzas.

Un modelo de mezcla de K gaussianas (GMM), denotado por λ, es una suma de densidades de gaussianas con ciertos pesos para cada gaussiana tal y como se enuncia en la fórmula 2:

$$P(x\mid\lambda)=\sum_{k=1}^{K}w_k\cdot N(x;\mu_k,\Sigma_k) \quad (2)$$

siendo los pesos todos positivos y la suma de todos ellos igual a 1:

$$\sum_{k=1}^{K}w_k=1 \qquad w_k\geq 0 \quad \forall k:1\ldots K \quad (3)$$

K es el número de gaussianas u orden del modelo. La Figura 9, representa la función de distribución de una mezcla de 3 gaussianas unidimensional.

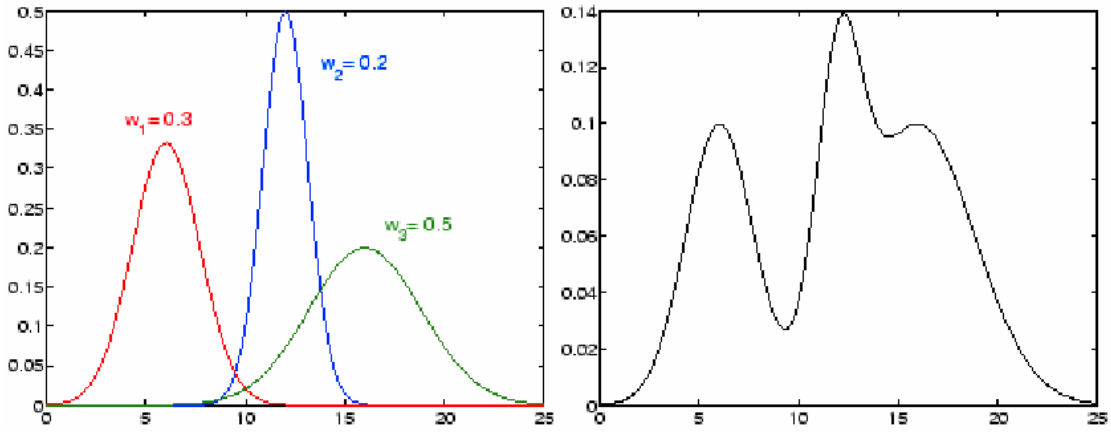


Figura 9 Función de distribución GMM

El proceso de entrenamiento del modelo GMM parte de una colección de vectores de entrenamiento de un sonido, con la que se estiman los parámetros del modelo (μ, Σ) para el número de gaussianas elegido K usando el algoritmo conocido como *Expectation Maximization* EM [121].

El número de gaussianas K necesario generalmente es establecido a través de prueba y error, aunque existen estudios que utilizan otros criterios, como el *Criterio de Información Bayesiano* (BIC) [26]. Si se elige un valor demasiado grande, puede darse el caso de que el modelo hallado sobreajuste demasiado a los datos extraídos

(*overfitting* o *sobreentrenamiento*). En el caso contrario, elegir un K demasiado pequeño puede llevar a que un determinado modelo no sea lo suficientemente diferente a los demás modelos.

Partiendo de un modelo inicial, el algoritmo EM refina iterativamente el modelo GMM incrementando su verosimilitud. Así, en la i-ésima iteración se encuentra el modelo $\lambda^{(i)}$ y se cumple que:

$$P(X \mid \lambda^{(i)}) > P(X \mid \lambda^{(i-1)}) \quad (4)$$

En donde X es el conjunto de vectores de entrenamiento correspondientes a una clase o sonido y $\lambda^{(i)}$ es el modelo obtenido en la iteración i-ésima. Éste es el nuevo modelo inicial para repetir el proceso hasta llegar a un nivel de convergencia predeterminado.

Debido a que en general el conjunto de vectores de características es muy grande, y por tanto, los valores de probabilidades son a menudo muy pequeños, es común utilizar el logaritmo de verosimilitud que viene dado por:

$$LogP(X \mid \lambda) = \frac{1}{N}\sum_{i=1}^{N} \log P(x_i \mid \lambda) \quad (5)$$

en donde los x_i son los vectores de entrenamiento.

Este valor, denotado también como Logl (*log – likelihood*), es la medida que nos dice cuánto de probable es que los vectores X pertenezcan al modelo λ

Para interrumpir la iteración, además de poder establecerse un número máximo de iteraciones, podemos establecer un umbral:

$$LogP(X \mid \lambda^{(i)}) - LogP(X \mid \lambda^{(i-1)}) < \varepsilon \quad (6)$$

En el proceso de testeo se deben evaluar los logl para cada modelo. Aquel con mayor valor para logl es el que tiene mayor probabilidad de que los vectores de entrada pertenezcan a ese modelo.

4.5 Técnicas de Detección de Sonidos

Cuando el sistema a desarrollar requiere funcionar en tiempo real, el audio con el que se trabaja es un audio continuo. Esto significa que, como a priori se desconoce cuándo se va a producir un evento, es necesario un proceso previo de búsqueda para saber dónde éstos se encuentran para clasificarlos. Esta búsqueda es conocida como *detección*.

La etapa de detección tiene un papel clave en la precisión y fiabilidad global del sistema. De nada sirve tener un clasificador capaz de identificar eventos con un 100% de acierto si la etapa de detección pierde eventos, genera falsos positivos o no indica de forma precisa el inicio y el final de los mismos.

En las técnicas de detección existen dos enfoques bien diferenciados: "*Detección por clasificación*" y "*Detección y clasificación*" [21]. El primero de ellos detecta en base a la clasificación de tramas consecutivas. El segundo busca el inicio y el fin sin clasificar de antemano. A continuación se detallan brevemente los dos enfoques.

4.5.1 Detección por Clasificación

Este tipo de enfoque se basa en detectar los eventos frente a una serie continua de clasificaciones. La mayoría de trabajos utilizan esta aproximación por ser de naturaleza muy simple al ya disponer de los modelos clasificatorios. Se puede decir que, en este caso, la tarea de detección se convierte en una tarea de clasificación.

En esta aproximación es necesario contar con una clase "*no evento*" entre el conjunto de las clases utilizadas para entrenar los diferentes modelos (u obtener dicha clase en base a un umbral de paso de probabilidades establecido en el resto de modelos). De alguna forma es necesario un mecanismo capaz de dar como opción de salida la posibilidad de que una trama no pertenezca a los eventos seleccionados a reconocer.

Un aspecto importante de esta estrategia es la aplicación de técnicas de post-procesado. Debido a que en la etapa de clasificación se trabaja con porciones muy pequeñas de audio (del orden del milisegundo) un suavizado de los resultados obtenidos de la clasificación es requerido mediante filtrados paso bajo o similares. De tal modo, si hemos clasificado las cinco primeras tramas como "*no evento*", una sexta trama como "*teléfono*" y cinco tramas más como "*no evento*", la aplicación de un suavizado haría que todos ellos perteneciesen a la clase "*no evento*" reasignando a la sexta trama "*teléfono*" la etiqueta de "*no evento*".

4.5.2 Detección y Clasificación

Otro enfoque distinto es el planteado con la estrategia "*Detección y clasificación*". En esta estrategia primero se busca el evento y luego se clasifica. Una vez encontrado el inicio y el final del evento, todo el trozo será enviado al clasificador para que decida la clase a la que pertenece en un segundo paso.

Utilizando esta técnica, la detección no está obligada a la utilización de una ventana del mismo tamaño que la clasificación. Además, se podrán utilizar distintos modelos probabilísticos o diferentes características acústicas que definan mejor el conjunto de sonidos en general. Aunque la aproximación más habitual hace uso de modelos

probabilísticos (habitualmente con la creación de dos modelos: "*evento*" y "*no evento*" en su primer nivel), existen otras aproximaciones que serán las que analicemos en este apartado por ser diferentes a los algoritmos probabilísticos ya comentados previamente.

El ejemplo más básico que no hace uso de modelos probabilísticos es aquel que utiliza un umbral de energía para separar los segmentos del audio capturado en eventos a clasificar. A priori no se sabe a qué clase pertenecen las tramas analizadas, sólo se puede decir que en las zonas donde la energía es alta hay "algo". Una vez se dispone del inicio y el final del evento sin clasificar, todas las tramas serán enviadas al clasificador para que sea él quien decida a qué clase pertenece. El principal inconveniente de este ejemplo es que es poco robusto frente a ruidos de fondo, sin embargo, dependiendo de la situación, puede ser un método ideal.

En [24] se proponen mejoras a la aplicación única de un umbral de energía basadas en correlación cruzada. Dado que la función de correlación cruzada es la medida de similitud entre dos señales, se aplica esta función entre dos ventanas sucesivas de la señal para detectar cambios bruscos. Obteniendo el valor máximo de la función de correlación, se aplica un umbral y si la resultante está por debajo del umbral se genera una detección de evento. Como segunda propuesta en [24], se utiliza la técnica de detección basada en predicción de energía. Este algoritmo está basado en calcular la diferencia entre el valor real de energía de una trama, y el valor predicho a través de una interpolación *SPLINE* utilizando tramas anteriores (10 tramas en el trabajo descrito). Para decidir si hay o no un evento, se establece un umbral autoajustable (que depende de la media y la desviación estándar de la señal).

Una técnica muy similar a las anteriores consiste en la disposición de un módulo donde la energía de la señal es estimada para cada sucesivo bloque de 100 milisegundos [94]. La secuencia de energía obtenida es filtrada y la salida del filtro substraída de la energía. Esto resulta en una nueva secuencia que es normalizada, enfatizando los pulsos de energía relevantes. Un umbral adaptativo, dependiendo de la desviación estándar de una secuencia de energía de ventana pasada, es aplicado a continuación. Este método provee un esquema de detección modificable y muy sensible para señales impulsivas, donde los pulsos pueden ser detectados bajo condiciones de bastante ruido adverso de fondo, con una relación señal a ruido *Signal Noise Ratio* (SNR) tan baja como -10 dB. Los valores de SNR son medidos sobre una ventana que incluye la parte decreciente de la señal. Igualmente, [47] aplica para la detección un umbral adaptativo utilizando ventanas de N=4096 muestras a 48Khz.

En [26] se utilizan *wavelets de Daubechies* con 6 momentos de desvanecimiento para la detección de sonidos impulsivos. Se utiliza para ello una ventana de 32 milisegundos. La *Transformada Wavelet Discreta* (DWT) es aplicada a los datos que se han muestreado y su salida forma un vector de la misma longitud que la señal. Este

vector tiene una estructura piramidal y está compuesto por 10 coeficientes wavelet. El algoritmo aplicado calcula la energía del octavo, noveno y décimo coeficiente (los tres de orden más alto) ya que los coeficientes más significativos de los sonidos a ser detectados se encuentran en las frecuencias altas. La detección es conseguida aplicando un umbral en la suma de energías de los tres coeficientes wavelets de más alto orden. El umbral es autoajustable y depende de los últimos 10 valores.

4.6 Conclusiones del Capítulo

En este capítulo se han analizado las diferentes etapas de los sistemas de reconocimiento de sonidos no-habla dentro del ámbito de la investigación. Se observa la tendencia a la división de los mismos en dos grandes bloques: detección y clasificación, dentro de los cuales se encuentra implícita la extracción de características acústicas de la señal para el entrenamiento y testeo de los modelos usados para el reconocimiento.

Pese a la existencia de bases de datos comerciales de sonidos no-habla, la gran mayoría de artículos analizados tienden a la creación de su propio corpus en base a los requerimientos u objetivos de los proyectos en los que se enmarcan. Esto es debido principalmente a la imposibilidad de poder reflejar todos los tipos de sonidos existentes en una única base de datos y, si esto fuera factible, la imposibilidad de manejo de un número tan elevado de clases resultantes. Debido a ello, existe una gran dificultad de realizar una comparativa clara sobre los trabajos estudiados en este campo. Las técnicas y algoritmos utilizados por los investigadores son aplicados a conjuntos muy diversos de sonidos, de número variable y grabaciones hechas en condiciones muy dispares, sin mucha relación entre trabajos.

Dentro de los parámetros acústicos más frecuentes podemos destacar los coeficientes cepstrales MFCCs. Estos parámetros se encuentran en la mayoría de artículos estudiados y ya habían demostrado previamente su fuerza dentro del campo del reconocimiento del habla. A su vez existe también tendencia al uso de otros parámetros más vinculados al reconocimiento de instrumentos musicales como ZCR, Roll-Off Point o Spectral Centroid. El uso de estos dos grupos de sonidos puede ser debido a la necesidad inherente de determinar el sonido no sólo por sus componentes tonales sino también por componentes más del lado de la percusión y del timbre que los mismos producen.

En las etapas de detección y clasificación las técnicas más utilizadas están basadas en modelos probabilísticos. Aunque dentro de la detección se encuentran algunos estudios donde la señal es procesada sin la aplicación de este tipo de algoritmos, la gran mayoría de artículos presentan para el reconocimiento modelos estadísticos tales como HMM, GMM o SVM que necesitan de un entrenamiento previo para su afinación

y funcionamiento correcto. Para esta tesis se han elegido los modelos GMM debido a su bajo coste computacional, pero también por presentar un modelado más genérico, capaz de representar un mayor número de tipos de sonidos y un entrenamiento fácil de adaptar a aplicaciones funcionales como se verá en posteriores capítulos.

5. Necesidades del Colectivo de Personas con Discapacidades Auditivas

Como se ha expuesto en capítulos anteriores, el conjunto de sonidos no-habla es muy extenso. Aunque existen bases de datos para la investigación, las primeras preguntas que surgen son si las muestras de sonidos que ellas contienen son de verdadero interés para el colectivo de personas con discapacidad auditiva. ¿Son todos los sonidos importantes? ¿Faltan de incluir sonidos considerados de mayor interés para las personas con problemas auditivos? Contestar a estas preguntas permitirá el diseño e implementación de soluciones orientadas a un usuario final, acercando la investigación a una realidad más cercana.

Para dar respuesta a todas estas preguntas, en este capítulo se ha realizado un estudio estadístico a nivel nacional, con personas con discapacidad auditiva. El estudio analiza sus necesidades, preferencias y puntos de vista acerca de cómo debería ser un sistema final de reconocimiento de sonidos no-habla en cuatro entornos: hogar, lugar de trabajo / estudios, calle y vehículo.

5.1 Metodología Aplicada

El estudio estadístico fue realizado mediante un cuestionario electrónico que se dejó disponible a través de Internet. De esta manera se suprime la necesidad de que los encuestados tengan que desplazarse, repercutiendo en poder llegar a un mayor número de usuarios en un menor tiempo. A través de asociaciones, fundaciones y entes públicos y privados de personas sordas se difundió el cuestionario por correo electrónico para que estos pudieran difundirlo a las personas objeto.

La encuesta fue dividida en cuatro bloques que se detallan a continuación.

1. *Datos Personales*: serie de preguntas personales a los encuestados sobre edad, sexo, formación y su discapacidad.

2. *Uso de Tecnologías*: serie de preguntas para conocer el uso que hace la persona de las nuevas tecnologías como el móvil o los ordenadores.

3. *Reconocimiento de Sonidos*: incluye 4 secciones, una por cada uno de los entornos que hemos considerado (hogar, oficina o centro de estudios, calle y vehículo) y una general para todos los entornos. En cada entorno interesa conocer:

    a. si se dispone de ayudas técnicas para la detección de sonidos

b. cuáles son los sonidos más importante a identificar en cada entorno

c. cómo le gustaría recibir a la persona el aviso y la información cuando uno de esos sonidos se detecte.

4. *Otras Sugerencias*: bloque dedicado a recabar información de cualquier otra problemática o necesidad que se encuentre en la vida diaria del encuestado (fuera del análisis de este capítulo).

Las cifras de participación del estudio se muestran en la Tabla 11:

Nº de e-mail enviados	92
Nº de cuestionarios respondidos	46
Nº de cuestionarios completos	37

Tabla 11 Cifras de participación de la encuesta realizada

Los datos que se presentan en los apartados siguientes derivan de los resultados obtenidos de los cuestionarios completos. Los cuestionarios respondidos parcialmente se desecharon de la muestra para no crear confusión en las estadísticas.

5.2 Datos Personales de los Encuestados

De los 37 encuestados (entre los que el 30% son hombres y el 70% restante mujeres), en la gráfica de la Figura 10 se puede apreciar que las franjas de edades sigue una distribución gaussiana. El grupo más activo fue el comprendido entre aquellos con edades comprendidas entre 25 y 35 años, siendo el grupo de personas mayores de 65 años el que tiene menos presencia en la encuesta. La falta de datos en la franja de más de 65 años puede venir derivado del hecho de que los cuestionarios fueron electrónicos, siendo el grupo de personas mayores el menos habituado a los mismos. A la hora de diseñar un sistema de reconocimiento de sonidos habrá que tener en cuenta este dato ya que será un sector cuya curva de aprendizaje será más alta y será necesario un diseño accesible y amigable para que este tipo de soluciones sean aceptadas.

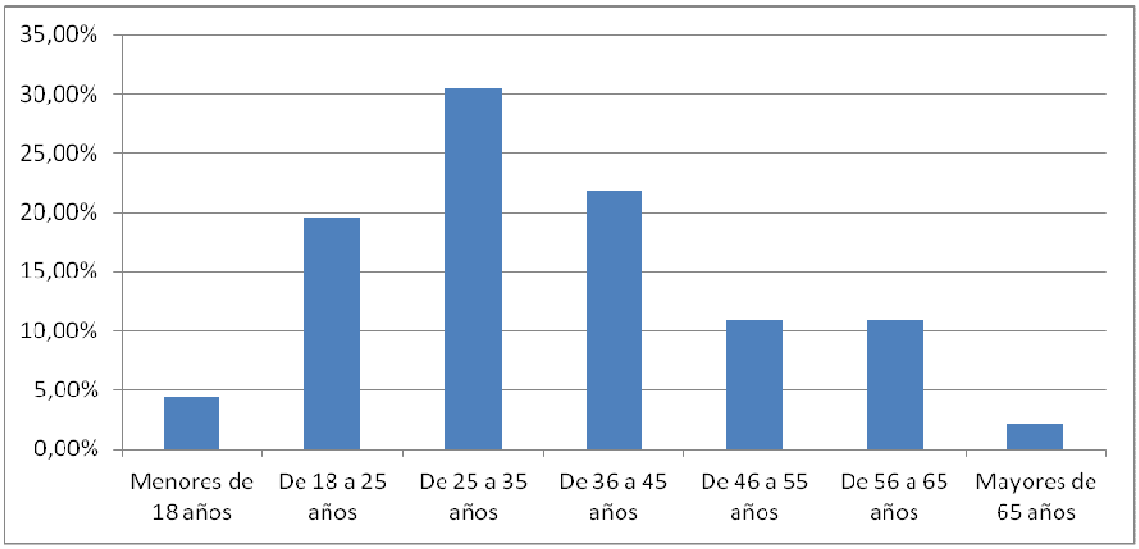


Figura 10 Encuesta realizada – Edad de los participantes

Al analizar la formación académica de los participantes, destacamos de la encuesta el alto nivel formativo de las personas que la realizaron. Tal y como se muestra en la gráfica de la Figura 11, más de la mitad de las personas con discapacidad auditiva que la contestaron poseen estudios universitarios o superiores. Estamos hablando por tanto de un conjunto de personas formadas y con una alta capacidad analítica que pueden ayudar a definir y acotar de una forma mejor la problemática planteada.

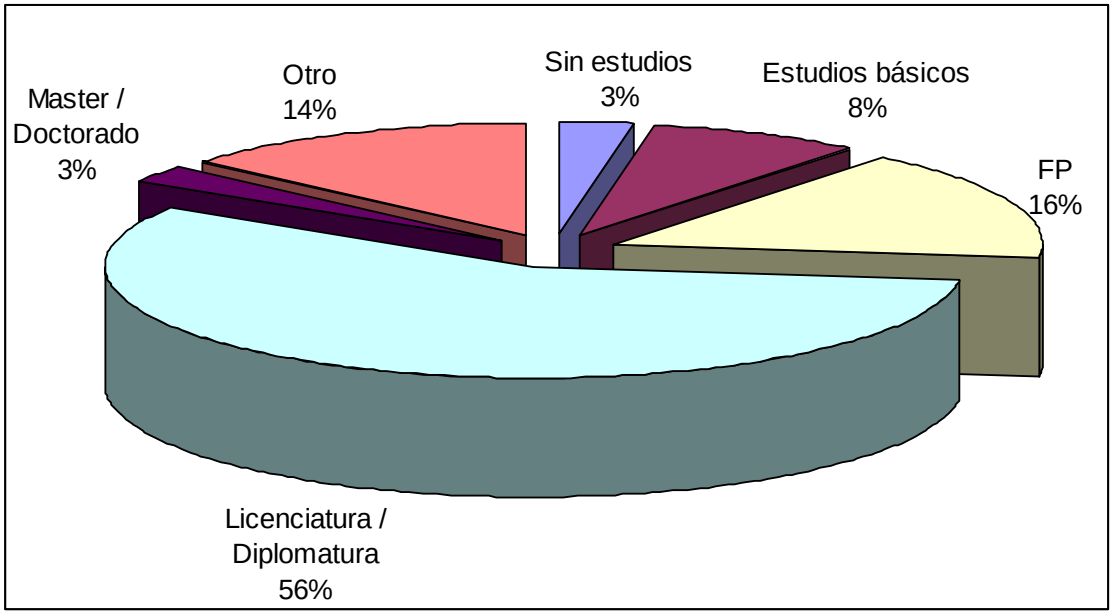


Figura 11 Encuesta realizada – Nivel de estudios de los participantes

Los datos del estudio estadístico definen también un sector en el que el 67% de sus individuos se encuentran en situación laboral de trabajo. Pese a que 37 cuestionarios es una muestra pequeña de la población, este dato es importante para conocer la solvencia económica del colectivo de personas sordas como futuro adquisidor de este tipo de soluciones.

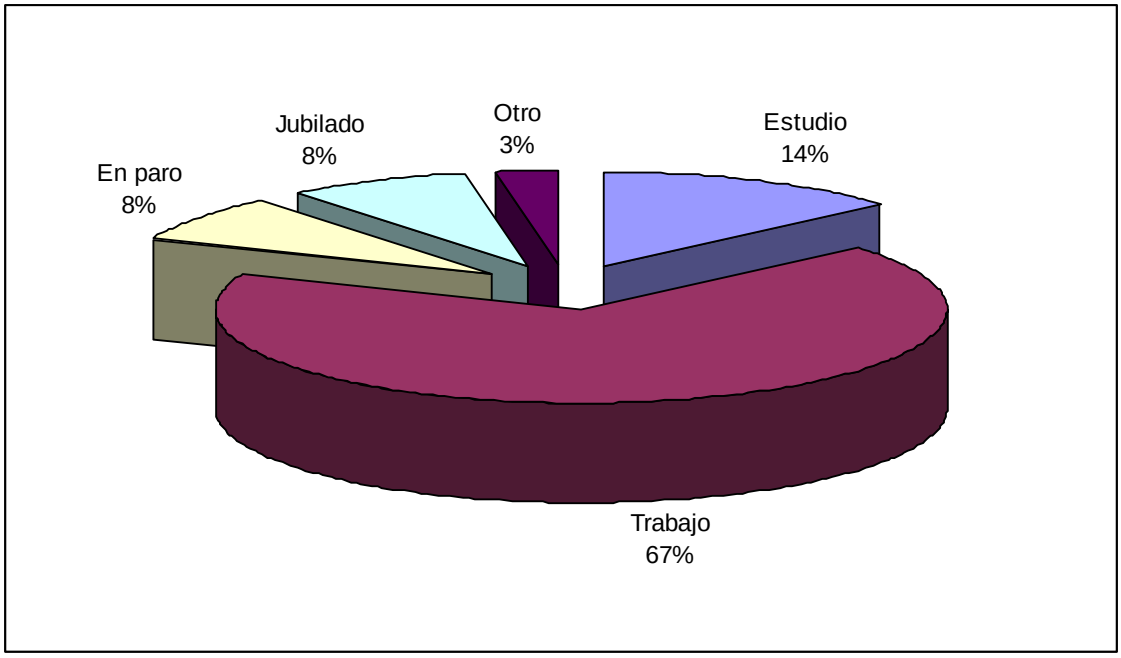


Figura 12 Encuesta realizada – Situación laboral/formativa de los participantes

En lo que se refiere al grado de sordera de los encuestados, las gráficas de la Figura 13 y Figura 14 indican cómo sólo un 5% no alcanza el 33% de discapacidad y el 97% de los encuestados posee una pérdida auditiva entre *Moderada* y *Total*. Este hecho es beneficioso para el análisis de los resultados ya que se trata de un nicho de personas cuya limitación auditiva es alta y sufren claramente en su vida diaria los problemas que supone el no ser capaces de reconocer los sonidos del entorno.

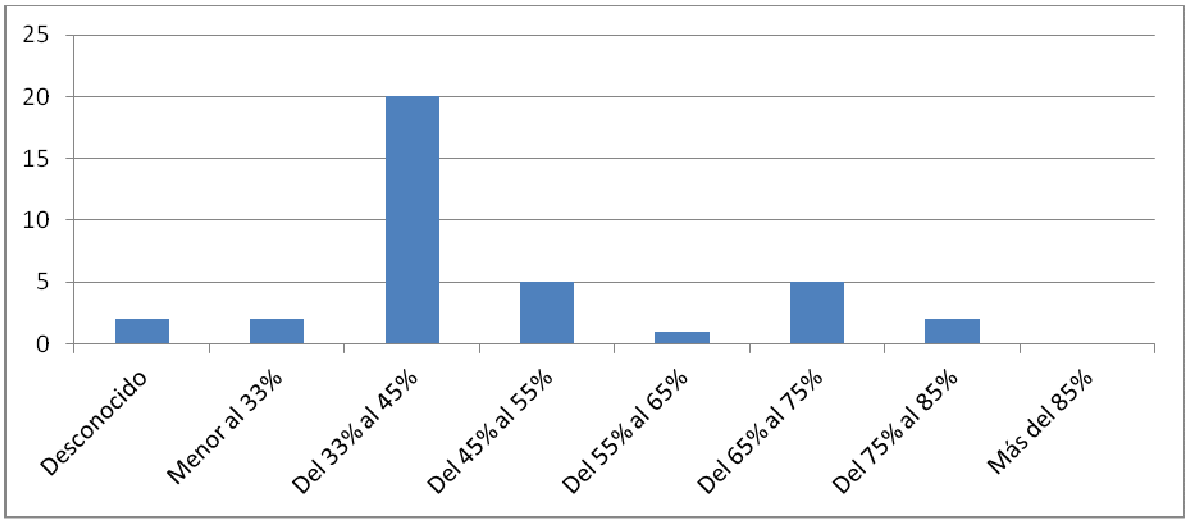


Figura 13 Encuesta realizada – Porcentaje de discapacidad de los participantes

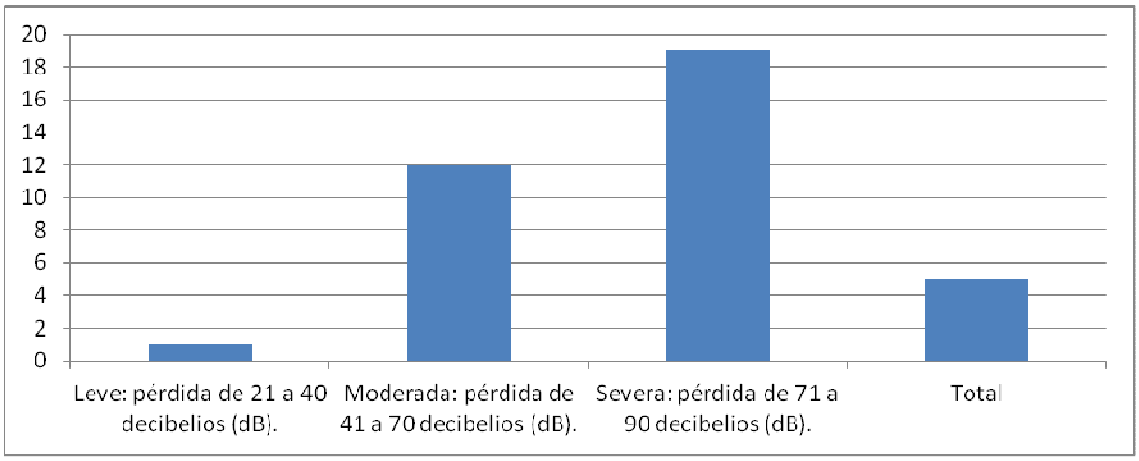


Figura 14 Encuesta realizada – Pérdida auditiva de los participantes

Además, tal como se ilustra en la gráfica de la Figura 15, reseñar que más del 27% de los encuestados nacieron con los problemas auditivos, siendo personas acostumbradas a convivir con la falta de audición. Es necesario aclarar que en el apartado "*Otro*" (pregunta abierta) se encuentran las personas que por culpa de medicamentos perdieron parte o toda su capacidad auditiva.

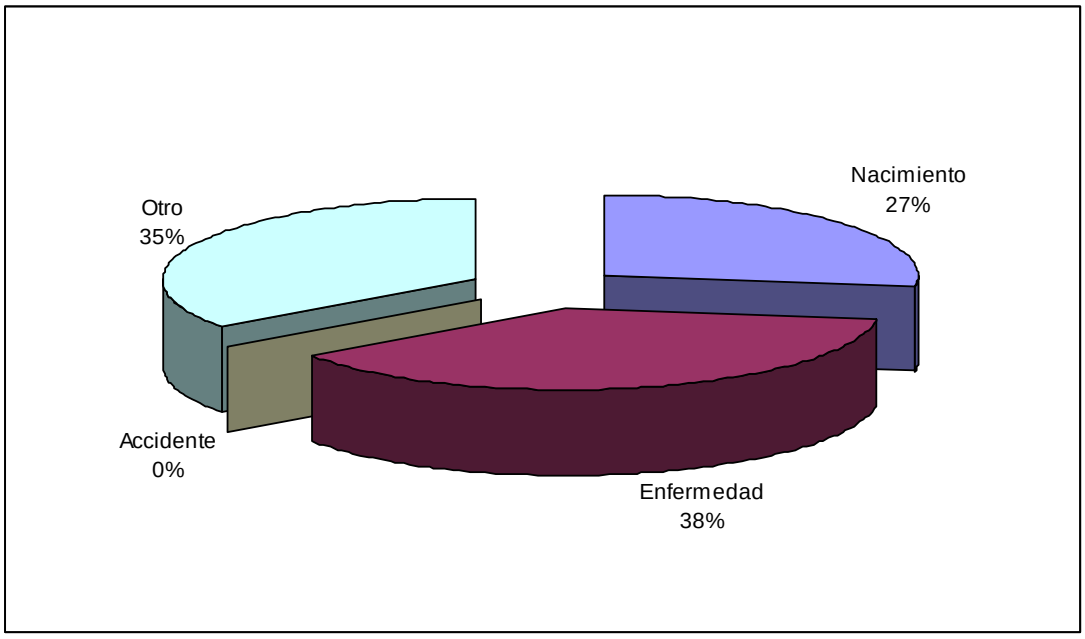


Figura 15 Encuesta realizada – Origen de la discapacidad de los participantes

Para conocer preferencias sobre posibles interfaces de salida en el diseño final de la solución, se analizó el uso de la Lengua de Signos entre los encuestados. La muestra escogida, tal y como se muestra en la gráfica de la Figura 16, está formada por personas donde aproximadamente la mitad de ellas puede comunicarse a través de la Lengua de Signos. Aunque más tarde, tal y como se verá, los usuarios son consultados acerca del interés de avisarles de sucesos acústicos mediante una interfaz que muestre

el mensaje a través de Lengua de Signos, los resultados de esta encuesta indican que este mecanismo no podría ser único, y debería ser acompañado de texto o imagen.

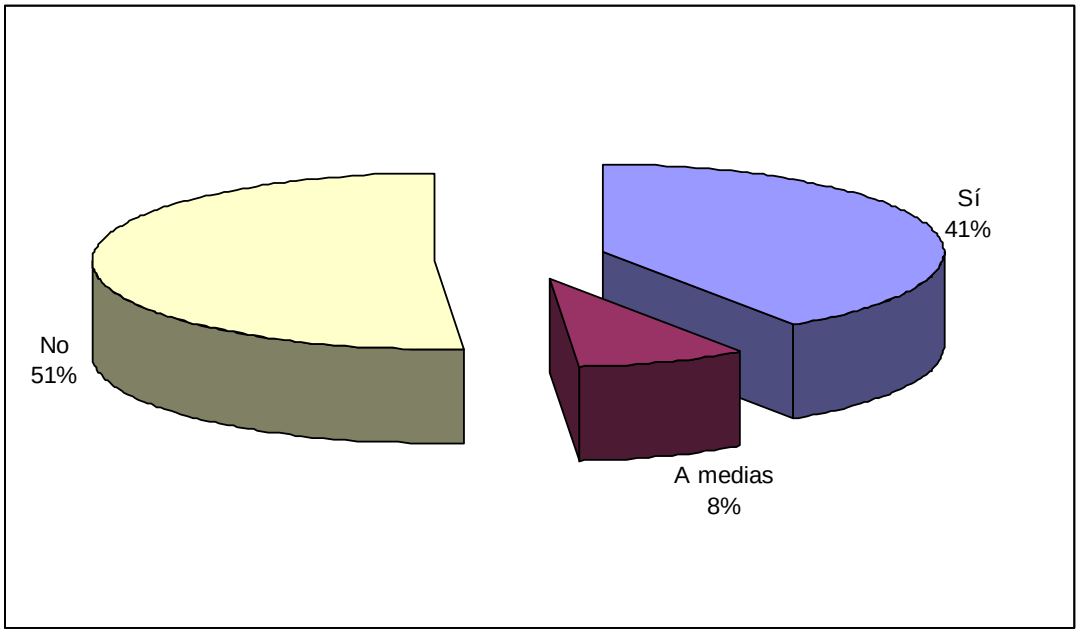


Figura 16 Encuesta realizada – Uso de la lengua de signos de los participantes

5.3 Uso de Tecnología de los Encuestados

Los datos analizados de los cuestionarios muestran cómo el 95% de los participantes posee un ordenador personal. Así mismo, su percepción de nivel de manejo es "*alta*" para un 41% y "*media*" para un 49% de las personas, tal y como se ilustra en la gráfica de la Figura 17. Estos datos corroboran que las nuevas tecnologías tienen gran penetración en la comunidad sorda, tal y como indicaban los datos presentados en el capítulo 2. No obstante, cabe destacar que el hecho de realizar los cuestionarios online, mediante un ordenador y a través de Internet, ya denota un mayor conocimiento tecnológico de los encuestados, por lo que esta información debe manejarse teniendo en cuenta este factor para no dar pie a posibles equivocaciones.

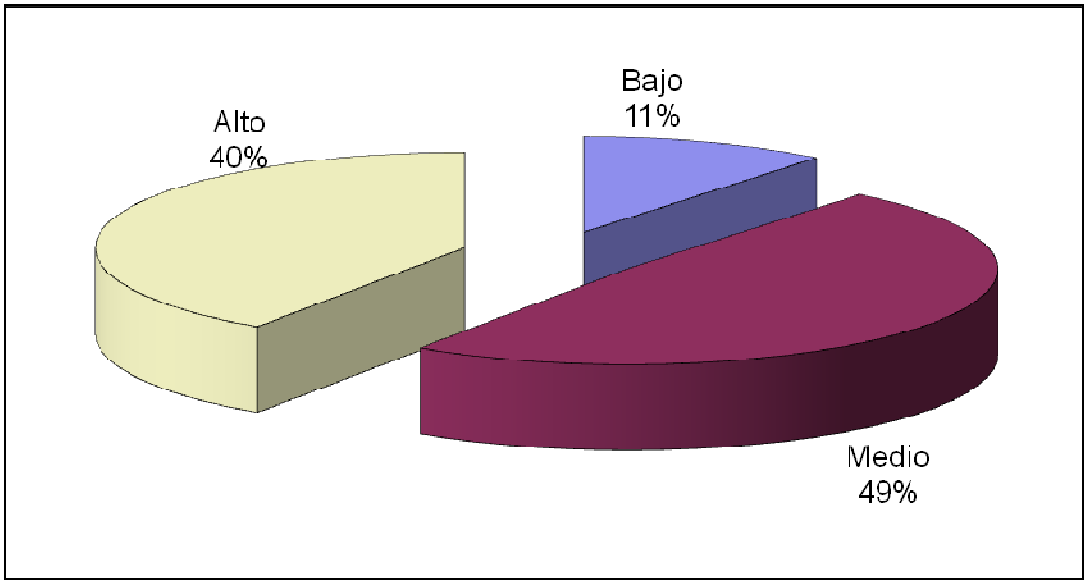


Figura 17 Encuesta realizada – Manejo del ordenador de los participantes

El 97% de los encuestados afirman tener acceso a Internet y, tal y como se muestra en la gráfica de la Figura 18, el 46% de la muestra tiene un nivel de manejo "*alto*" y otro 46% nivel "*medio*". Dentro del uso de Internet, el 40% del tiempo está dedicado a la comunicación (27% *e-mail* y 13% *chat*) y un 30% a la búsqueda de información como se muestra en la gráfica de la Figura 19. Esta información, igual que la interior, también viene relacionada con la metodología de realización de los cuestionarios.

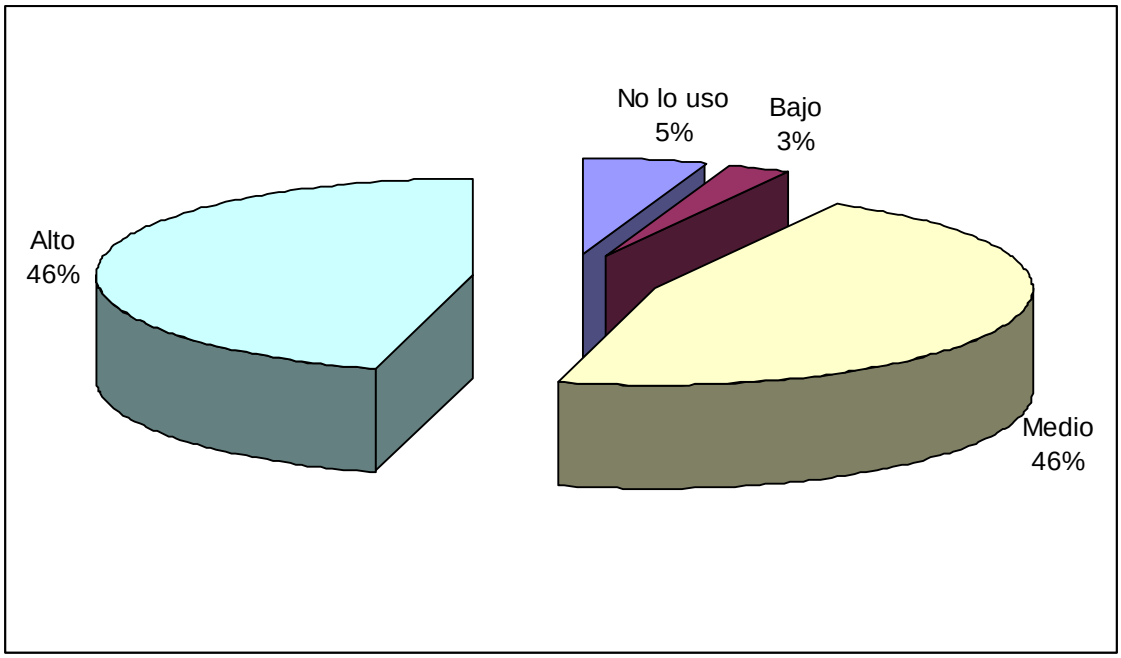


Figura 18 Encuesta realizada – Manejo de Internet de los participantes

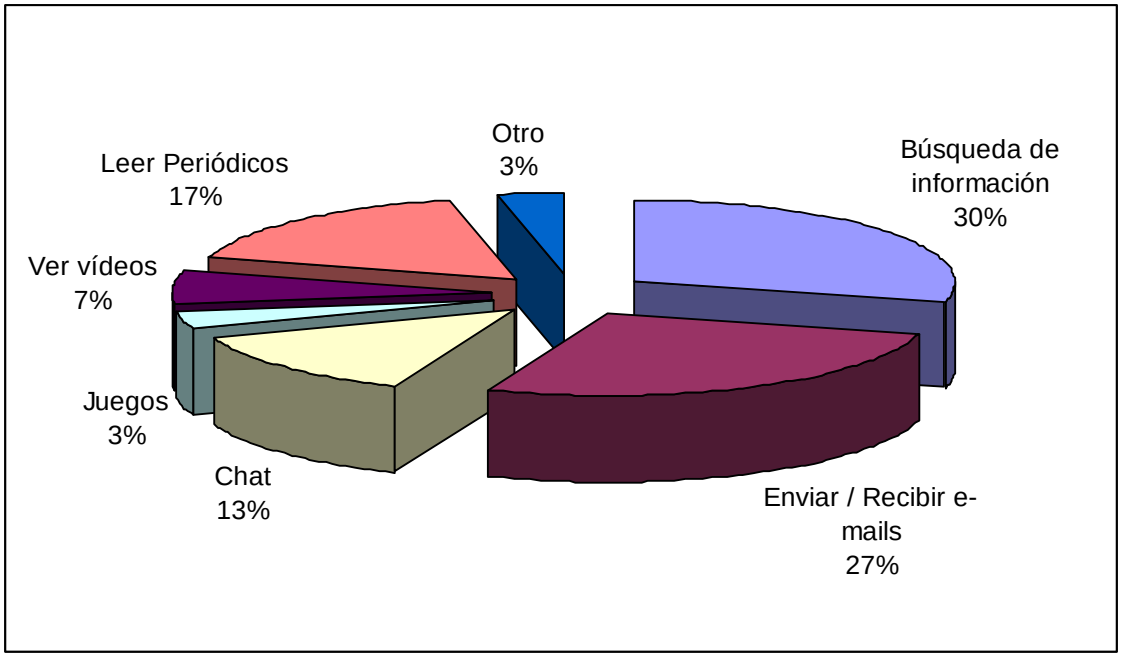


Figura 19 Encuesta realizada – Uso de Internet de los participantes

El uso del teléfono móvil va también acorde con el del uso del ordenador y el de Internet. El 97% de los encuestados aseguran disponer de un teléfono móvil utilizándolo mayoritariamente para el envío o recepción de mensajes de texto SMS, como ilustra la gráfica de la Figura 20.

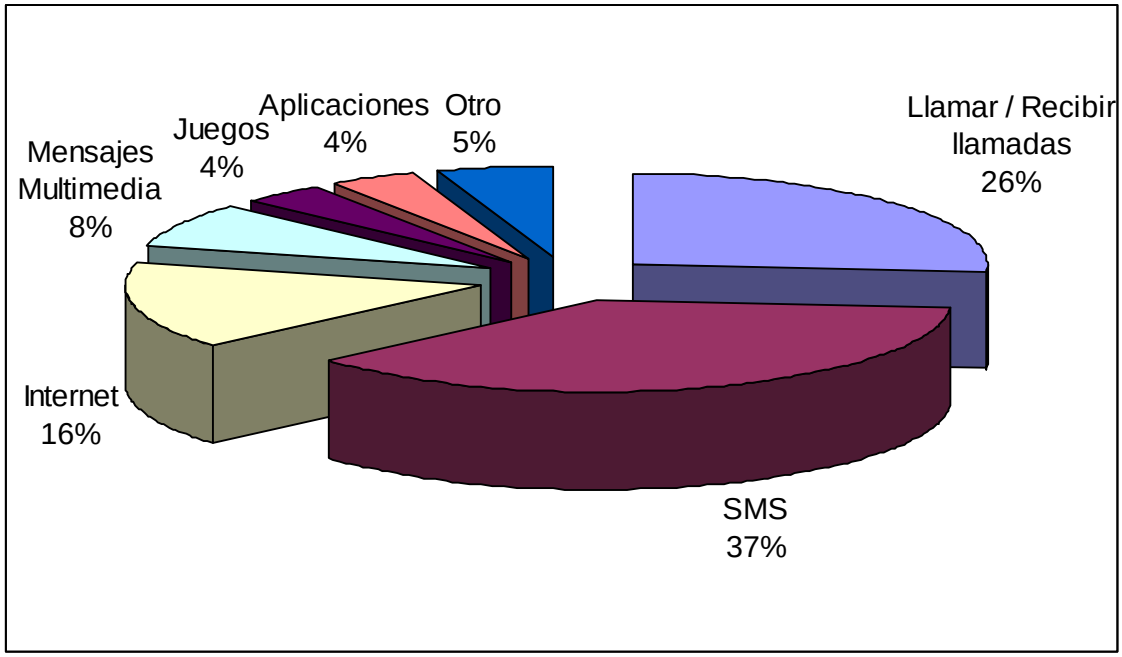


Figura 20 Encuesta realizada – Uso del móvil de los participantes

5.4 Reconocimiento de Sonidos No-Habla en Diferentes Entornos

En este apartado se presentan los resultados del bloque central de la encuesta referente al ámbito del reconocimiento de sonidos no-habla en diferentes entornos. Los entornos analizados son "*Casa*", "*Trabajo / Estudios*", "*Calle*" y "*Vehículo*". En todos ellos se analiza cuáles son los sonidos no-habla de mayor interés y el medio preferido donde recibir la información.

En primer lugar se quiso conocer qué entorno resulta de mayor interés para el reconocimiento de sonidos. La pregunta formulada para ello requería la elección de los dos entornos preferidos. Se añadió un entorno "*Otro*" con la idea de permitir que el encuestado o encuestada pudiera incluir algún otro entorno libremente. Tal y como muestra la gráfica de la Figura 21, el 34% de la suma de todos los votos fueron para el hogar ("*Casa*"), seguido muy de cerca del entorno "*Trabajo / "Estudios*" que obtuvo el 30%. Un 20% y un 10% recayó en los entornos "*Calle*" y "*Vehículo*" respectivamente y sólo un 6% en el grupo "*Otro*" donde, mediante respuesta abierta, los encuestados hacían alusión al teatro, cine y lugares de ocio en general. Se recuerda al lector que estos porcentajes están calculados sumando todos los votos recibidos sobre las dos opciones que debía marcar el encuestado. Por lo tanto, en este caso, el máximo porcentaje posible a alcanzar por un entorno sería el 50%.

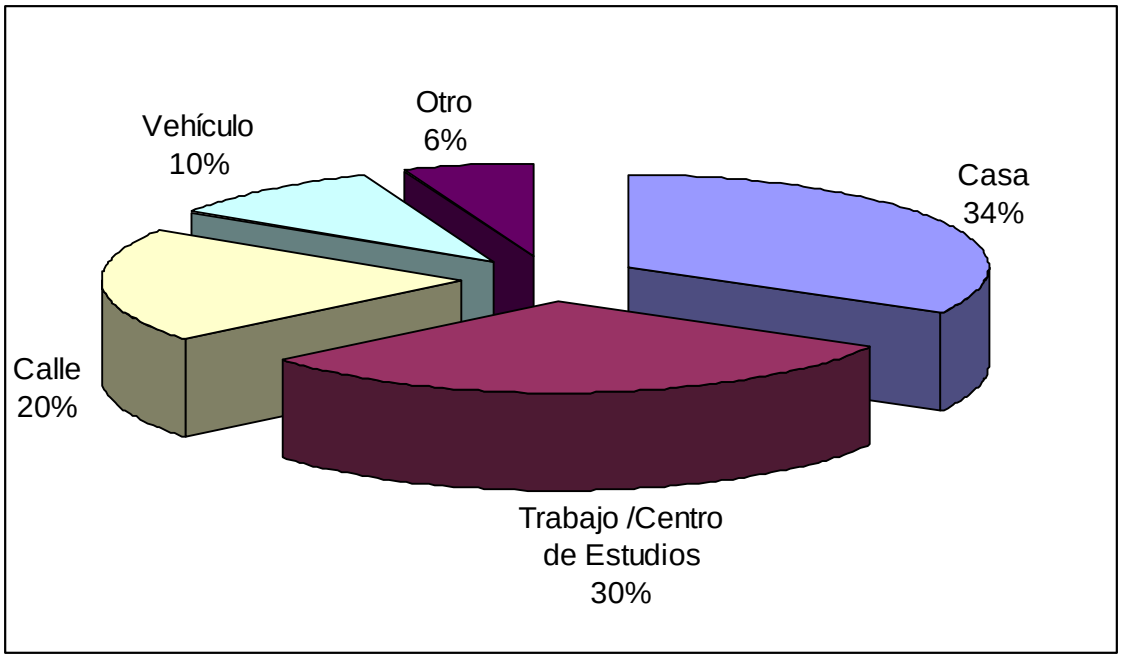


Figura 21 Encuesta realizada – Entornos de interés para los participantes

5.4.1 Reconocimiento de Sonidos No-Habla en el Hogar

En una siguiente pregunta, a los encuestados se les pidió marcar los tres sonidos de mayor interés sobre un listado elaborado de 22 tipos de sonidos que pueden producirse en una vivienda. Los resultados a la cuestión se muestran en la gráfica de la Figura 22. Se puede apreciar cómo el sonido más importante a reconocer es el timbre de la puerta con un 13% de la suma de todos los votos seguido muy de cerca del despertador con un porcentaje del 12%.

El resultado elevado del interés mostrado por la TV y la radio se debe (deducido a partir de consultas informales realizadas a algunos encuestados con posterioridad) a que los encuestados tienen interés en entender los comunicados de la radio y de la TV, y no el interés de saber si la radio o la TV están encendidos. Es decir, se trata de un fallo en la redacción de la pregunta que por tanto no tendremos en cuenta.

Dejando a un lado la TV y la radio, Los tres siguientes sonidos más importantes son el portero automático, el teléfono fijo y los golpes en la puerta con porcentajes del 9%, 7% y 7% de los votos respectivamente. De los 6 sonidos que ocupan el ranking de sonidos no-habla más importantes cabe destacar que tres cuartas partes (timbre de la puerta, despertador, portero automático y teléfono fijo) son sonidos de carácter determinista. Sus características dependen del fabricante, modelo o preferencias del usuario y, por lo general, se encuentran bien acotados en frecuencia.

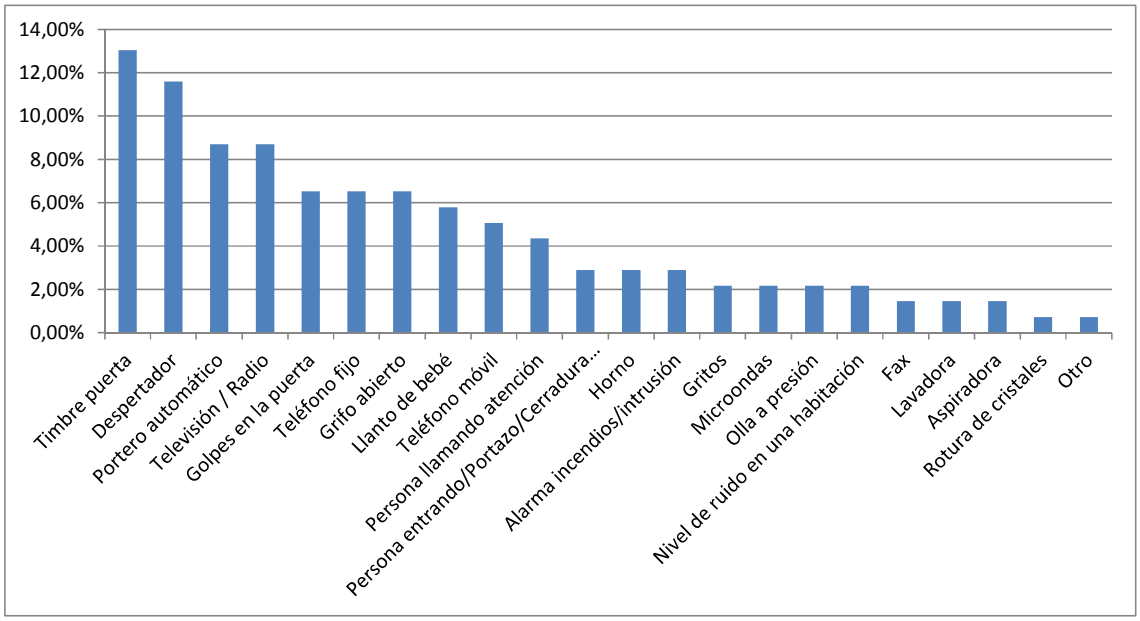


Figura 22 Encuesta realizada – Sonidos de Interés en el hogar para los participantes

En base a una lista de 8 opciones, se les pidió que marcaran los 2 medios de visualización que considerasen más adecuados donde pudiera mostrarse la información del sonido reconocido. La opción "*en mi móvil*" fue la más valorada con un 32% de la puntuación, superando incluso a la opción "*en el dispositivo que estuviese*

más cerca" que alcanzó un 28% de los votos. El tercer puesto recae en la opción "*en un reloj de pulsera*" con un 16% como se muestra en la gráfica de la Figura 23.

Esto demuestra la gran utilización del teléfono móvil por el colectivo de personas con discapacidad auditiva. Este dispositivo está totalmente asentado y aceptado por esta comunidad y, junto a los avances tecnológicos que día a día van surgiendo sobre él, puede ser entendido como el medio óptimo a utilizar en cualquier tipo de solución.

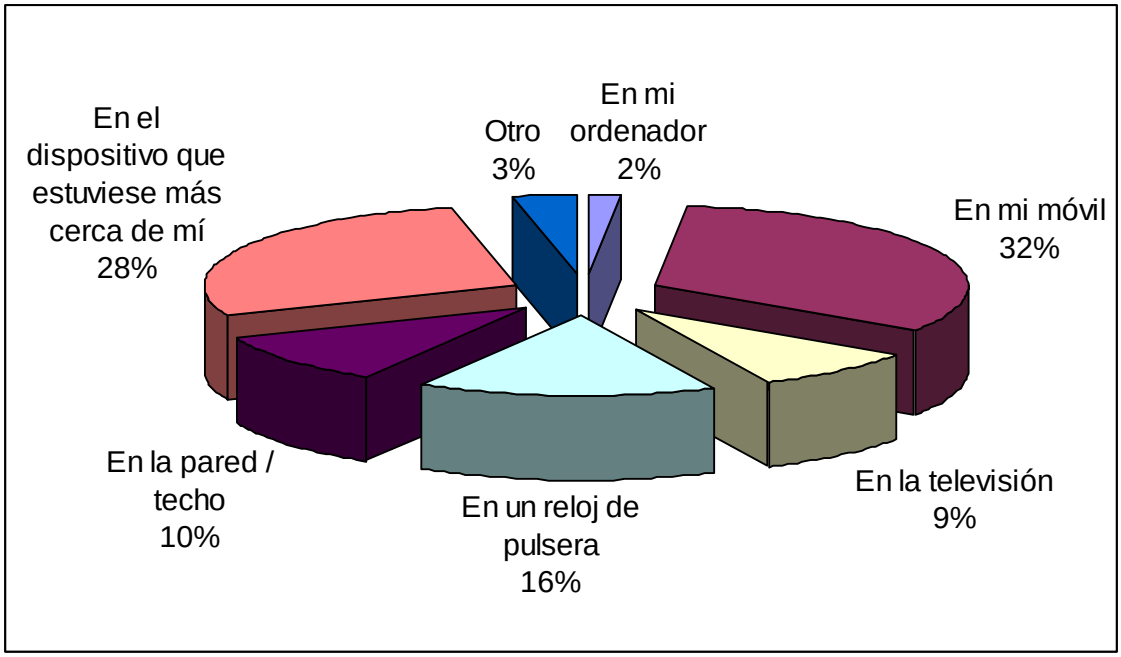


Figura 23 Encuesta realizada – Medios de visualización en el hogar para los participantes

5.4.2 Reconocimiento de Sonidos No-Habla en el Lugar de Trabajo y/o Estudio

Con la misma metodología que en el caso del Hogar, se elaboró un listado con 11 tipos de sonidos y se les pidió a los encuestados que marcasen los 3 que para ellos fuesen más importantes para reconocer. En este entorno el sonido con mayor valoración fue el etiquetado como "*persona llamando tu atención*" con un porcentaje del 21% como se indica en la gráfica de la Figura 24. En segundo lugar, como sonido más valorado, están las "*alarmas de incendios / emergencia*" con un 17% de los votos, seguido del teléfono móvil con una valoración del 15%. El hecho de que el mayor interés recaiga en saber cuándo una persona está llamando su atención refuerza la impresión obtenida de los comentarios de los encuestados en relación a los problemas generados por su sordera. Uno de ellos relata cómo su discapacidad le ha producido aislamiento: "*indicar que mi sordera me ha producido aislamiento, y a veces cuando alguien quiere decirme algo no presto atención, hasta que me chilla, pero eso me pone muy nerviosa, a la defensiva.....*".

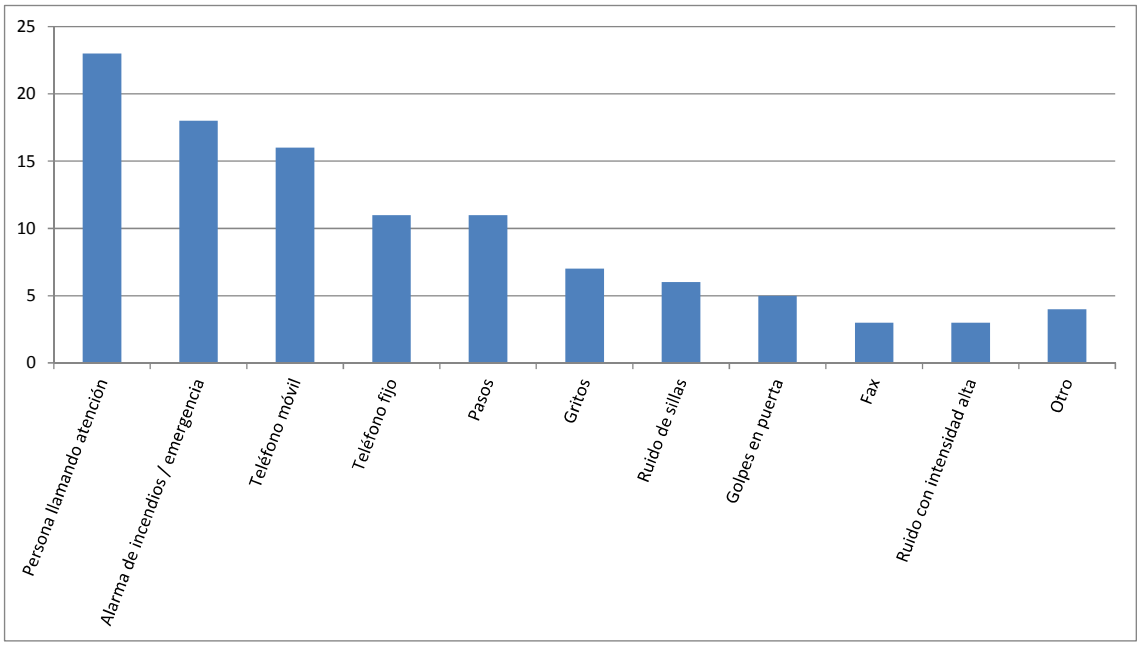


Figura 24 Encuesta realizada – Sonidos de Interés en el trabajo/estudio para los participantes

Igualmente, en base a una lista de 5 opciones, se les pidió que marcaran los 2 medios de visualización que considerasen los más adecuados donde pudiera mostrarse la información del sonido reconocido. Nuevamente el medio predominante fue el teléfono móvil con un 29% de los votos, dejando otra vez vigente la importancia de este dispositivo para el colectivo de personas con discapacidad auditiva.

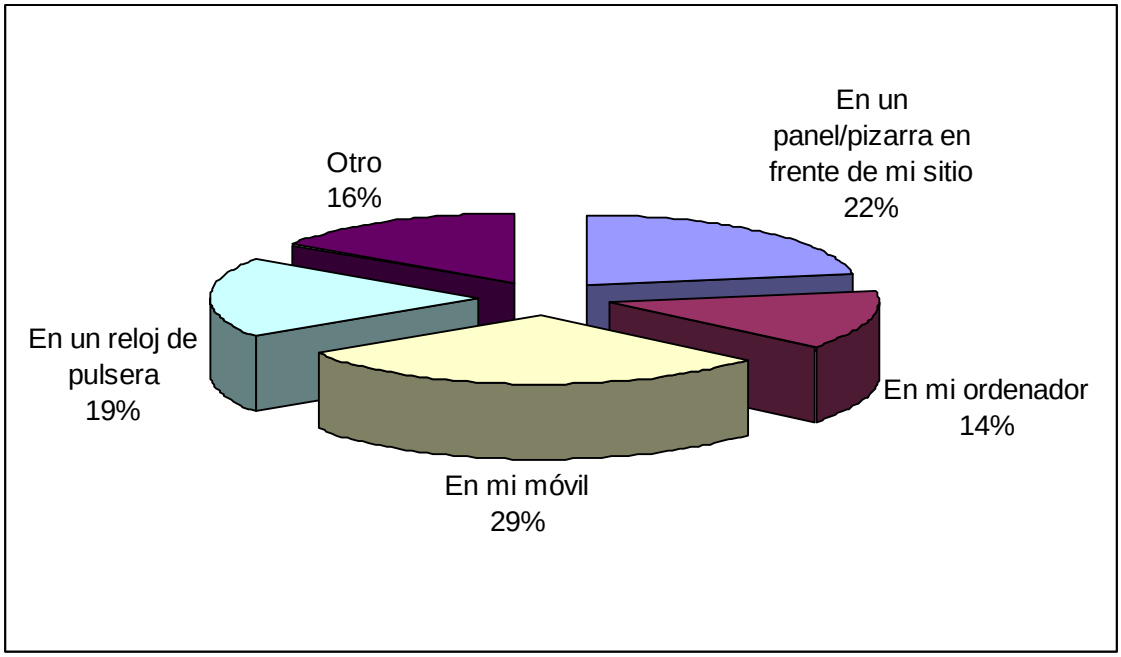


Figura 25 Encuesta realizada – Medios de visualización en el trabajo/estudios para los participantes

5.4.3 Reconocimiento de Sonidos No-Habla en la Calle

Para el entorno de la calle fueron 9 los tipos de sonidos presentados para la evaluación. De éstos, se pidió que se marcaran los tres considerados más interesantes a reconocer. Al igual que en el lugar de trabajo o estudio, fue el saber cuándo una persona está intentando llamar su atención lo más valorado con un 20% de los votos. El claxon de los vehículos, las sirenas de vehículos de emergencia y el teléfono móvil ocuparon consecutivamente los siguientes puestos con unos valores de 19%, 18% y 15% respectivamente tal y como indica la gráfica de la Figura 26.

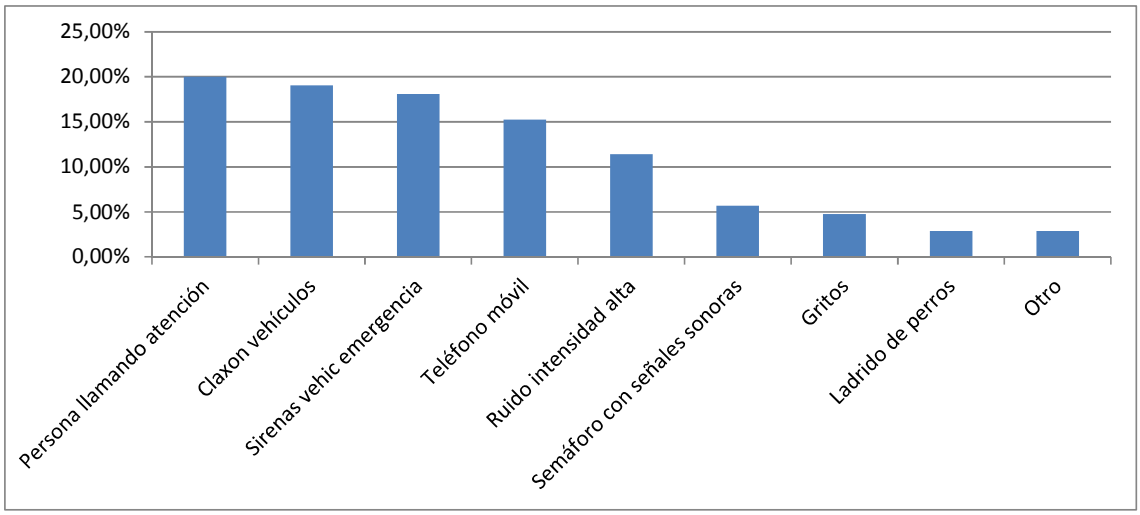


Figura 26 Encuesta realizada – Sonidos de Interés en la calle para los participantes

En cuanto al medio de visualización de los mensajes, el teléfono móvil sigue siendo el favorito con un 48% de los votos. La opción "*reloj de pulsera*" también fue una opción demandada con un 38% de porcentaje. Dentro del apartado "*Otro*", con un 14% de los votos, a parte de las respuestas de personas que no lo tenían claro, hay propuestas que se decantan por gafas subtituladas. Esta tecnología ya se puede encontrar en el mercado. De momento se trata de gafas prototipo cuya portabilidad se debe mejorar, pero no hay que descartar esta opción en un futuro como medio a ser utilizado.

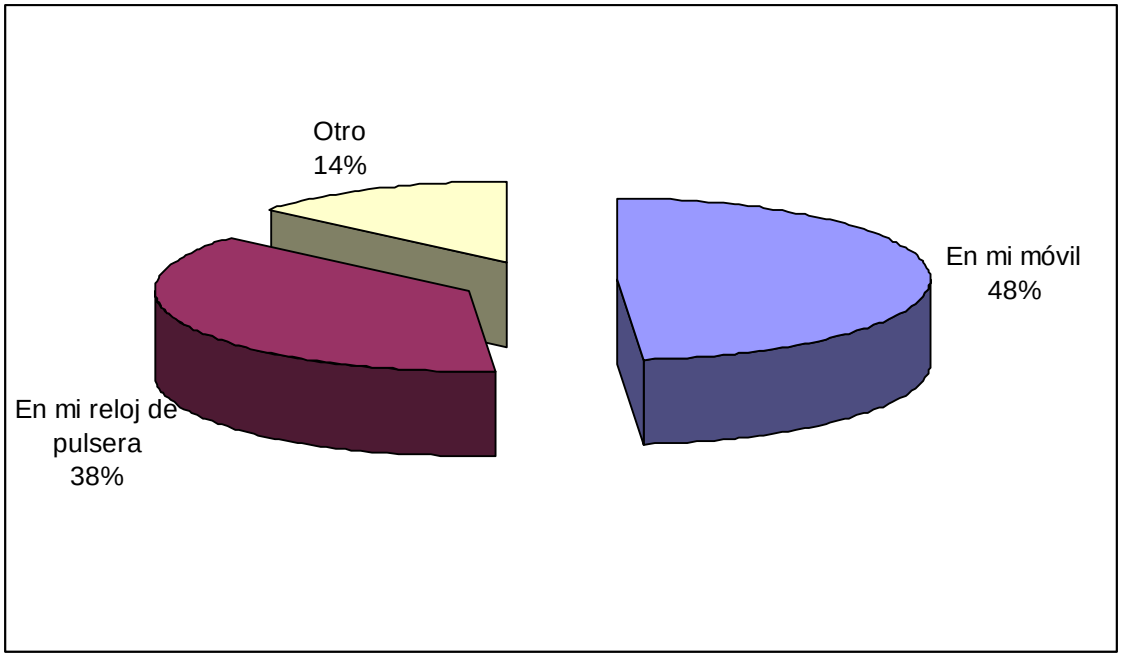


Figura 27 Encuesta realizada – Medios de visualización en la calle para los participantes

5.4.4 Reconocimiento de Sonidos No-Habla en el Vehículo

En esta ocasión se dieron a elegir 6 opciones para que los encuestados eligieran las 3 que para ellos describían a los sonidos más importantes a reconocer cuando se encontraban conduciendo su vehículo. Recalcar que el 76% de los participantes conducían habitualmente y que el 58% eran capaces de escuchar la radio o música cuando estaban en un vehículo.

Los encuestados en esta evaluación dieron a las sirenas de ambulancias / bomberos / policías la mayor puntuación con un 29% de los votos. Seguidamente, el claxon acaparó el 25% de la puntuación, seguido muy de cerca de las alarmas interiores del vehículo con un 21% de porcentaje (ver gráfica en Figura 28).

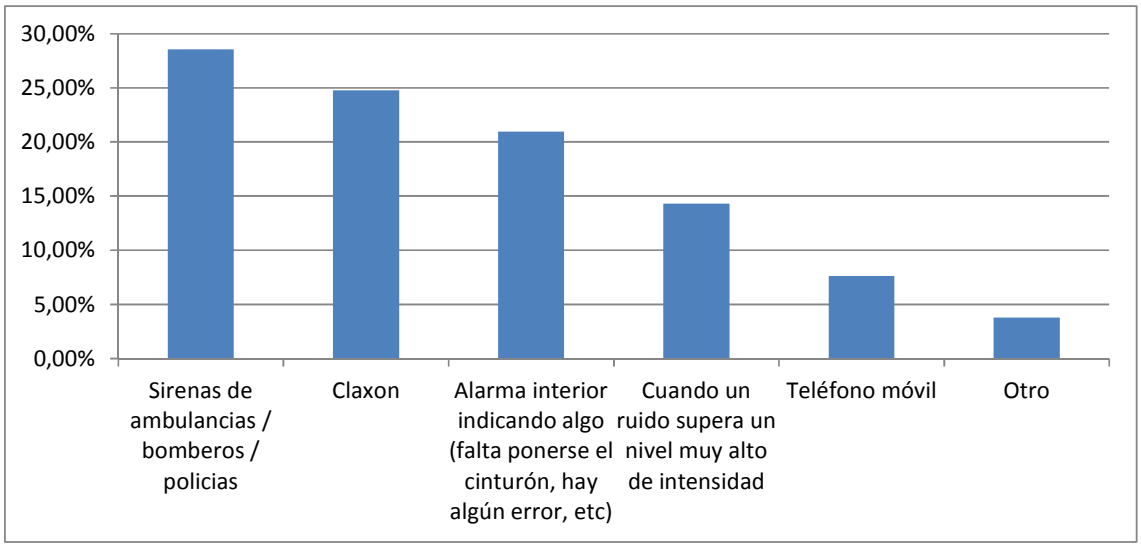


Figura 28 Encuesta realizada – Sonidos de Interés en el vehículo para los participantes

Debido a que dentro del vehículo los dispositivos que pueden utilizarse para visualizar sonidos están más limitados, la pregunta de dónde visualizar los mensajes se centró principalmente en la posición más adecuada de colocación del dispositivo. Con un 36% de los votos el salpicadero fue la opción más marcada seguida con un 29% por la opción del espejo retrovisor. La elección del salpicadero como primera opción puede estar influenciada por el uso del GPS. Además, el teléfono móvil, tan solicitado en los anteriores entornos, podría ser utilizado sin problema alguno en esta ubicación. Espejos retrovisores con información también empiezan a aparecer. Sobre todo, se empieza a ver en taxis que los utilizan para indicar el precio que el cliente debe pagar a medida que transcurren los kilómetros y el tiempo.

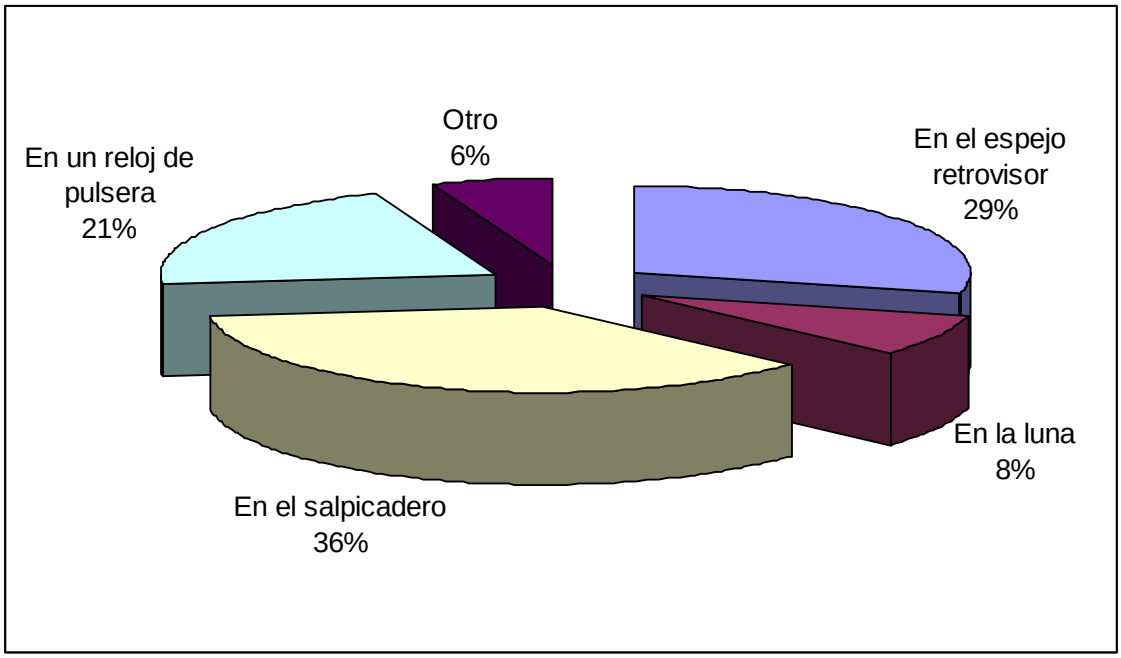


Figura 29 Encuesta realizada Medios de visualización en el vehículo para los participantes

5.4.5 Reconocimiento de Sonidos en General

Referentes a todos los entornos se plantearon dos preguntas para conocer la forma en la que la persona quiere ser avisada de que un evento acústico es detectado y, por otro lado, cómo se le tiene que mostrar la información. En el primer caso nos referimos tan sólo a la forma en la que un sistema debe llamar la atención para que se sepa que ha sucedido un evento sin la necesidad de mostrar todavía la información. Se consideraron como opciones: "*vibración*", "*luces*" y "*sonido amplificado*". El segundo caso se centra en conocer cómo deben ser las interfaces de usuario más adecuadas a utilizar para este tipo de sistemas. En este caso las opciones fueron: "*texto*", "*imagen*", "*lengua de signos*", "*código de luces*", "*sonido amplificado*".

Salvo en la calle, en el resto de entornos la elección principal para avisar de que algo sucede fue a través de luces. La excepción de la calle se debe claramente a que la calle es un entorno en el que es más complicado disponer de sistemas luminosos visibles todo el tiempo. En la Tabla 12 se muestra el resultado de esta evaluación.

	Vibración	Luces	Sonido Amplificado
Hogar	35%	**57%**	8%
Trabajo / Estudio	38%	**46%**	16%
Calle	**65%**	24%	11%
Vehículo	35%	**51%**	14%

Tabla 12 Preferencias de personas con discapacidad auditiva en modos de aviso

En cuanto a la forma de indicación del tipo de sonido (ver Tabla 13), la imagen se percibe en todos los entornos como elemento preferente a utilizar. El segundo puesto depende del entorno, aunque sí hay que denotar cómo dentro del vehículo, y posiblemente al haber un conjunto más reducido de sonidos, el código de luces tiene un gran peso con un 28.1% del porcentaje.

	Texto	Imagen	Lengua Signos	Código de Luces	Sonido Amplificado
Hogar	19.7%	**29.5%**	9.8%	21.3%	19.7%
Trabajo/Estudio	27.4%	**30.6%**	8.1%	21%	13%
Calle	21%	**29%**	9.7%	19.4%	21%
Vehículo	15.6%	**31.2%**	7.8%	28.1%	17.2%

Tabla 13 Preferencias de personas con discapacidad auditiva en señalización

Finalmente, se ofreció la posibilidad a los participantes de describir libremente cómo debería ser su sistema ideal de reconocimiento. De entre las respuestas recibidas, los encuestados hablan de la necesidad de un sistema que no moleste a los demás y que su forma de aviso no asuste a la persona utilitaria del mismo. Además, éste debe ser rápido y claro.

5.5 Conclusiones del Capítulo

En este capítulo se ha hecho un análisis que servirá para tener información acerca de cómo diseñar e implementar sistemas de reconocimiento de sonidos no-habla aceptables por el usuario. La realización de cuestionarios online y su difusión a varias decenas de organizaciones relacionadas con la comunidad de personas con discapacidad auditiva ha permitido tener datos de 37 encuestados que, considerando otros estudios, resulta un número elevado (en [38] se realizaron encuestas a 8 personas con problemas auditivos, en [40] fueron 10 los encuestados y en [122] fueron 24).

Las respuestas a las preguntas planteadas permiten conocer cuáles son los sonidos más importantes para la persona con problema de audición en los diferentes entornos analizados (Hogar, Trabajo/Centro de estudios, Calle y Vehículo) que podrán ayudar a la elaboración de bases de datos para la investigación más específicas y sistemas más robustos.

Se ha demostrado cómo el teléfono móvil puede ser un elemento fundamental en los futuros sistemas de reconocimiento de sonidos que se desarrollen. Su amplia utilización entre el colectivo de personas sordas y las posibilidades de desarrollo que éste posee hacen de él un elemento clave. En todos los entornos vistos ha sido el medio más demandado y sus amplias características para el diseño de interfaces multimodales reafirman técnicamente los resultados.

La forma de aviso más aceptada es la luminosa, siempre y cuando estemos en un entorno cerrado. Sin embargo, la opción de vibración también ha tenido altas valoraciones, pudiendo ser interesante la combinación de ambas opciones en un sistema híbrido. En cuanto a lo referente a la visualización de la información del sonido detectado, el uso de imágenes representativas de los mismos es la opción mejor valorada en todos los entornos. Esta opción es rápida e intuitiva, y además, puede ser combinada con otras opciones (como por ejemplo texto) para una mayor usabilidad del sistema.

Este capítulo, por tanto, aporta conocimientos clave para el diseño de sistemas de reconocimiento de sonidos no-habla para el apoyo a personas con discapacidad auditiva. En el ámbito de la investigación ayuda a centrar la investigación del área de

procesado de señal para que ésta pueda tener mayor aplicabilidad. En el ámbito industrial y/o de accesibilidad ofrece pautas de diseño a tener en cuenta para el desarrollo de futuras aplicaciones que puedan salir al mercado y sean bien aceptadas por los usuarios.

6. Estudio Experimental de Técnicas de Reconocimiento de Sonidos No-Habla

Analizados los trabajos y técnicas de mayor predominio en el área del reconocimiento de sonidos ambientales, el presente capítulo ofrece una perspectiva más técnica con la implementación y evaluación de diferentes algoritmos utilizando bases de datos comerciales de investigación. El estudio experimental se divide en dos puntos principales: por un lado el reconocimiento de sonidos aislados y, por otro lado, el reconocimiento de sonidos sobre audio continuo.

Dentro del apartado de reconocimiento de sonidos aislados se realiza una comparativa para conocer los parámetros de configuración óptimos del sistema de reconocimiento basado en Modelos de Mezclas Gaussianas con las bases de datos más utilizadas en la literatura científica: RWCP [49] y CHIL [123]. Se analizará si estas configuraciones pueden ser generalizadas a las dos bases de datos sin la pérdida sustancial de rendimiento del sistema.

En el apartado de reconocimiento de sonidos sobre audio continuo se refinan los algoritmos para adaptarlos a un reconocimiento online de los eventos, analizando diferentes estrategias de detección.

6.1 Configuración de los Experimentos

En este subapartado se describen las bases de datos de sonidos que serán utilizadas para evaluar el sistema. Posteriormente se establecen los algoritmos (características acústicas y algoritmo de clasificación) que definirán los experimentos de este capítulo.

6.1.1 Bases de Datos

Los trabajos descritos dentro de la literatura científica en el área del reconocimiento de sonidos no-habla aplican las técnicas desarrolladas a un único corpus de sonidos. Este corpus además varía de un trabajo a otro y no existe una base de datos de trabajo común, debido principalmente a que el rango de sonidos no-habla es muy amplio y dependiente de la aplicación. Ello impide conocer a ciencia cierta el alcance de los algoritmos. El presente estudio busca realizar una contribución en este área evaluando los algoritmos implementados con las dos bases de datos comerciales más extendidas dentro del campo del reconocimiento de sonidos ambientales: RWCP y CHIL.

En los experimentos presentes, los dos corpus de sonidos fueron divididos en tres subconjuntos: entrenamiento, validación y testeo.

El conjunto de clases de sonido provisto en RWCP fue reducido haciendo una selección previa de sonidos de interés por la gran cantidad de clases existentes en él. Esta selección consta de 11 clases las cuales fueron elegidas por encontrarse dentro del ámbito de la vivienda (ver Tabla 14). Una vez hecha esta selección se permutaron todos los sonidos de cada clase dividiendo el 60% de ellos para el entrenamiento, el 20% para la validación y el 20% restante para el testeo. Además, debido a que se encuentran muchos silencios al comienzo y al final de los sonidos de esta base de datos, se realizó un preprocesado en el que se eliminó el comienzo y final de las señales utilizando una detección de nivel (ya que las grabaciones son limpias y no ruidosas).

Clases	Denominación RWCP	Nº Total de Ejemplos
Botella (Bottle)	Bottle1 + Bottle2	200
Porcelana (China)	China1 + China2 + China3 + China4	400
Reloj (Clock)	Clock1 + Clock2	200
Vaso (Cup)	Cup1 + Cup2	200
Cerradura (Doorlock)	Doorlock	100
Secador (Dryer)	Dryer	100
Sonajero (Kara)	Kara	100
Sartén (Pan)	Pan	100
Teléfono (Phone)	Phone1 + Phone2 + Phone3 + Phone4	305
Afeitado (Shaver)	Shaver	100
Spray (Spray)	Spray	100

Tabla 14: Sonidos Base de Datos RWCP

Los sonidos de CHIL fueron todos seleccionados. Utilizando el etiquetado disponible, los sonidos fueron extraídos del audio en continuo para trabajar con ellos aisladamente. En la Tabla 15 se indica el número de sonidos por cada una de las 14 clases utilizadas. La base de datos se dividió en entrenamiento (DVD1), validación (DVD2) y testeo (DVD3). Esta división coincide con la planteada y utilizada en la evaluación CLEAR [123].

Clases	DVD1	DVD2	DVD3	Nº Total Ejemplos
Aplausos (Ap)	20	20	20	60
Cucharas (Cl)	23	21	20	64
Sillas (Cm)	23	28	24	75
Tos (Co)	22	22	21	65
Puerta abierta (Do)	20	20	21	61
Puerta cerrada (Ds)	20	21	19	60
Llaves (Kj)	21	21	23	65
Golpes puerta (Kn)	15	16	17	48
Teclas (Kt)	21	25	20	66
Risas (La)	22	21	21	64
Teléfono (Pr)	52	36	43	131
Papel (Pw)	32	29	24	85
Pasos (St)	28	24	22	74
Desconocidos (Un)	38	46	42	126

Tabla 15: Sonidos Base de Datos CHIL

6.1.2 Modelo de Clasificación

Tal y como se concluyó en el capítulo 3, el abanico de sonidos no-habla es muy amplio y variado. Unos son impulsivos y cortos (portazo o golpes en la puerta), otros siguen secuencias de tonos (teléfono, despertador), otros son largos y estacionarios (agua del grifo),… Esto implica que el clasificador a elegir tiene que tener la capacidad de comportarse de forma generalista y poder modelar la distribución de todos ellos.

Los HMM son capaces de establecer diferentes estados, sin embargo, el clasificador GMM fue el elegido debido a que es un modelo más simple y generalista. Eventos diferentes en estructura pueden ser clasificados observando la distribución que siguen los valores extraídos de sus tramas. Es una opción menos costosa de implementar y, una vez creado el modelo, éste no necesita de grandes cálculos computacionales a la hora de clasificar, lo que supone una ventaja si se desea hacer el algoritmo funcional en tiempo real. Además, en una aplicación funcional donde se pida al usuario que entrene sus propios sonidos, los HMM necesitan conocer el inicio y final de los eventos entrenados. En este caso el usuario debería grabar varios sonidos del mismo tipo independientemente (por ejemplo, grabando 50 veces el grifo del agua). Sin embargo, con los modelos GMM el usuario puede hacer una única grabación por cada sonido al no necesitar conocer el inicio y final del evento (por ejemplo, dejando el grifo de agua abierto y grabando su audio durante un tiempo estimado).

6.1.3 Características Acústicas

Los parámetros acústicos implementados para nuestros estudios fueron elegidos basándose en el estado del arte realizado en el capítulo 4. Las características acústicas elegidas fueron por una parte los MFCCs (habituales en el reconocimiento del habla) y

por otra parte Zero Crossing Rate, Spectral Centroid y Roll-Off Point (habituales en el reconocimiento de instrumentos y sonidos ambientales). A continuación se da una breve descripción de ellos.

6.1.3.1 Zero Crossing Rate (ZCR)

La Tasa de Cruces por Cero (Zero Crossing Rate, ZCR) es el número de veces que la señal pasa por cero por unidad de tiempo. En una señal discreta, el paso por cero se da cuando muestras consecutivas tienen signos diferentes, por lo que el cálculo de este parámetro se realiza contando el número de cambios de signo por unidad de tiempo. Un valor alto del ZCR está relacionado con una mayor cantidad de componentes de alta frecuencia en la señal.

La expresión 7 nos permite calcular el valor de ZCR:

$$Z_t = \frac{\sum_{n=1}^{N-1} \left| sign(x[n+1]) - sign(x[n]) \right|}{N} \qquad (7)$$

Siendo $x[n]$ el valor de la muestra y N el número de muestras totales en la trama t.

Este parámetro ayuda a distinguir entre eventos cuyo espectro cambia bruscamente (un portazo) y eventos con frecuencias definidas (una alarma), capturando el ZCR por cada trama de audio [124].

ZCR, al igual que Roll-off Point y Centroid, no es un parámetro habitualmente utilizado en el reconocimiento del habla, sino que su uso se aplica con mayor intensidad en el ámbito de la identificación de música y sonidos impulsivos no-habla, así como en segmentación de voz/música/ruido [125].

6.1.3.2 Roll Off Point (RF)

El punto RF (Roll-Off Point) es una frecuencia de corte tal que por debajo de ella reside el 85% (o 95%) de la energía total de la señal. En música este valor permite distinguir sonidos percusivos (como golpes de batería y ataques de notas) de sonidos con dinámicas más suaves (por ejemplo notas mantenidas por un violín).

El valor del Roll Off Point RF es la solución a la ecuación 8:

$$\sum_{k<RF} \left| X[k] \right|^2 = 0.85 \cdot \sum_{k} \left| X[k] \right|^2 \qquad (8)$$

Siendo $X[k]$ la Transformada Discreta de Fourier de la trama, k es el *bin* (unidad mínima en el dominio espectral) del eje de frecuencias discreto.

6.1.3.3 Spectral Centroid

Este parámetro calcula el centro de gravedad espectral de una trama de la señal. Se obtiene por tanto a partir de la transformada de Fourier mediante la expresión:

$$C_t = \frac{\sum_{k=1}^{N} |X[k]| \cdot k}{\sum_{k=1}^{N} |X[k]|} \quad (9)$$

donde $X[k]$ representa la muestra k-ésima de la Transformada Discreta de Fourier correspondiente a la trama y N es el número de muestras de la ventana.

Este parámetro está relacionado con lo que se suele llamar la 'brillantez' del sonido. Es por esto que muchas veces se cataloga dentro del conjunto de *características perceptuales* del audio (véase Capítulo 4).

6.1.3.4 Mel Frequency Cepstral Coefficients (MFCCs)

Los parámetros MFCC son muy populares en tratamiento del habla y se han empleado con éxito en prácticamente todas las áreas relacionadas (reconocimiento de voz y de locutores, reconocimiento de emociones y otros).

MFCC es un parámetro basado en la Transformada de Fourier de la señal. Tras calcular el logaritmo de la magnitud de la Transformada de Fourier, los *bins* se agrupan y suavizan según la escala frecuencial de *Mel* definida matemáticamente en la fórmula 10:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (10)$$

Existen muchas implementaciones de los parámetros MFCCs. La implementación elegida en esta tesis hace uso tanto de filtros lineales como logarítmicos. Para frecuencias bajas utiliza 13 filtros lineales y para las frecuencias altas utiliza 27 filtros con espaciado logarítmico. El peso asignado a cada filtro será proporcional al ancho de banda que estos tengan.

Esta implementación divide la ventana en segmentos de 10 ms, y es en cada uno de los segmentos resultantes donde se aplican los pasos finales de MFCC:

- Aplicación de un pre-énfasis a la señal
- Enventanado del segmento con ventana *Hamming*
- Obtención de la magnitud en el espacio frecuencial tras aplicar la *Transformada Discreta de Fourier*

- Aplicación del banco de filtros a los datos obtenidos de la *Transformada Discreta de Fourier* y posterior transformación a base logarítmica
- Multiplicación del resultado por la matriz de la *Transformada Discreta Coseno* para reducir dimensión y obtener los coeficientes *cepstrales*.

Los valores de los coeficientes MFCC serán calculados tras aplicar la media a los coeficientes de todos los segmentos de 10 ms. En total se obtuvieron los 13 primeros coeficientes.

6.1.3.5 Primera y Segunda Derivada

Para todas las características acústicas anteriores se calcularon la primera y segunda derivada (delta y delta-delta) que modelarán las características dinámicas de los parámetros.

El vector resultante está formado por 46 características acústicas como se indica en la Tabla 16.

Característica	Cantidad
MFCCs + delta + delta-delta	39
ZCR + delta + delta-delta	3
Roll-Off Point + delta + delta-delta	3
Spectral Centroid + delta + delta-delta	3
TOTAL	**46**

Tabla 16: Conjunto de características acústicas de los experimentos

6.2 Reconocimiento de Sonidos Aislados

Seleccionados los algoritmos a evaluar en los experimentos, a continuación analizamos la etapa de clasificación donde los sonidos están aislados, conociéndose el instante donde empieza y finaliza el evento.

6.2.1 Estudio Experimental

En este estudio se analiza el efecto que tiene la variación del número de gaussianas de los *Modelos de Mezclas Gaussianas* en la clasificación de sonidos procedentes de dos bases de datos distintas. Aunque algunos estudios que utilizan GMM como modelo de clasificación establecen este número en base a evaluaciones con sus corpus de sonidos o utilizando métodos probabilísticos como BIC, la mayoría de artículos no detallan la procedencia del mismo. Además, todos los artículos analizados evalúan sus experimentos con una única base de datos. No hay comparativa que demuestre la dependencia del tipo de sonidos con la variación de este parámetro.

A continuación se presenta el análisis realizado para la búsqueda del número óptimo de gaussianas sobre dos bases de datos comerciales: RWCP y CHIL. El experimento quiere ofrecer información acerca de la independencia de un corpus de sonidos frente a elementos de la clasificación.

6.2.1.1 Tamaño de Ventana

Como es conocido, el tamaño de la ventana definirá la resolución frecuencial del análisis, de forma que obtendremos mayores resoluciones frecuenciales con ventanas más largas y a la inversa. Por otra parte, una variación en el tamaño de ventana implica que el número de muestras con las que los modelos son entrenados, validados y testeados varíe si no hay solapamiento o éste no es constante. Si aplicamos tamaños de ventana pequeños, por cada sonido tendremos más elementos con los que alimentar el modelo probabilístico que si utilizamos ventanas de gran tamaño. Generalmente, trabajar con conjuntos grandes de datos es lo más deseado, y por ese motivo se incorpora un solapamiento de ventanas.

Comparar diferentes bases de datos entre sí requiere hacerlo de forma que la configuración de los experimentos sea la misma para las dos. Sin embargo, los datos pertenecientes al corpus de RWCP pueden ser confusos si no analizamos las características de sus sonidos. Así como la duración mínima de los eventos grabados en CHIL es aceptable para el uso de varios tamaños de ventana, en RWCP los eventos son más cortos, encontrándose sonidos que, eliminando los silencios del comienzo y del final, no pasan siquiera de los 30 milisegundos. En las gráficas de la Figura 30 y Figura 31 podemos observar los tamaños en una gráfica tipo *boxplot* de cada clase de sonido en las dos bases de datos. Se puede apreciar cómo, así como en CHIL todos los eventos tienen una mediana superior a los 1000 milisegundos, en RWCP sólo son dos los que superan esta cifra.

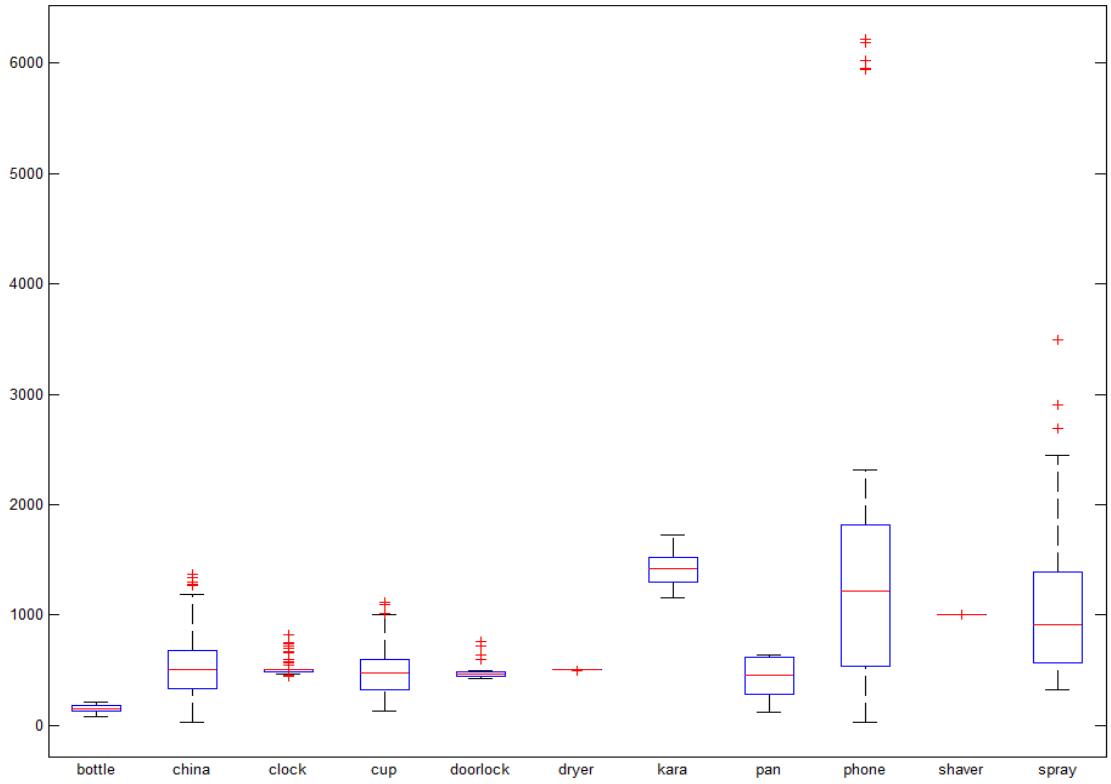


Figura 30 Duración de las muestras de RWCP en milisegundos

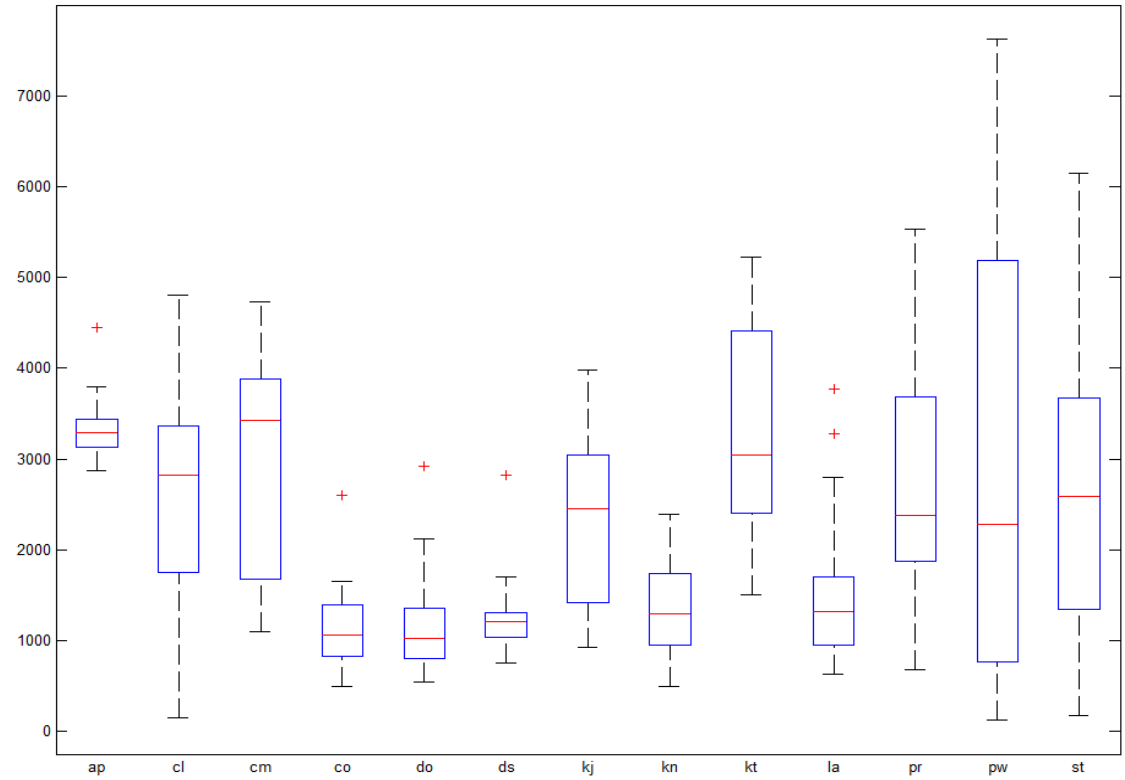


Figura 31 Duración de las muestras de CHIL en milisegundos

Del resultado de este análisis se optó por establecer el tamaño de ventana a 20 milisegundos con un desplazamiento constante de 10 milisegundos para las pruebas con eventos aislados que se presentan a continuación. Esto permite que todo sonido a clasificar ocupe más de una ventana de análisis (el sonido con menor duración dura apenas 26,9 milisegundos), no teniendo que rellenar la ventana con ceros o información repetida de la muestra. La relación de número de tramas por cada base de datos se muestra en la Tabla 17.

BD	Entrenamiento (Nº Tramas)	Validación (Nº Tramas)	Test (Nº Tramas)
RWCP	76946	25600	25607
CHIL	70295	75105	76853

Tabla 17: Relación del número de tramas con ventanas de 20 ms.

6.2.1.2 Número de Gaussianas

El parámetro más significativo de los GMM es el número de gaussianas que se utiliza para modelar las clases de sonidos seleccionadas. Un número reducido de gaussianas caracteriza las muestras de una forma muy general, sin encontrar separaciones entre los diferentes vectores de cada clase. Un número muy elevado de gaussianas puede sobreentrenar el modelo.

Los valores de número de gaussianas analizados en los experimentos fueron de 3 a 90, siendo éstos: 3, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85 y 90.

6.2.1.3 Métrica de Evaluación

En el área del reconocimiento de patrones son dos las medidas más utilizadas para ofrecer resultados: "*precisión*" y "*cobertura*" (es más frecuente encontrarlas con su terminología anglosajona: "*precision*" y "*recall*"). Sin embargo, trabajar con ellas por separado puede conllevar una mejora relativa en una pero un empeoramiento sustancial en la otra. La fórmula 11 define estas dos medidas.

$$precision = \frac{(verdaderos\_positivos)}{(verdaderos\_positivos + falsos\_positivos)}$$

$$recall = \frac{(verdaderos\_positivos)}{(verdaderos\_positivos + falsos\_negativos)} \quad (11)$$

Con el objetivo de ofrecer una medida capaz de aunar *precision* y *recall* los resultados de los experimentos planteados se darán en base a la fórmula del *F1-score*. Se trata de una medida clásica en la comunidad de la clasificación estadística. Los fundamentos se basan en la idea de calcular la media armónica entre estos dos valores. La fórmula 12 define esta medida.

$$F1-Score = 2*\frac{(\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})} \quad (12)$$

Los rangos de valores resultantes del *F1-score* se encuentran acotados entre 0 y 1. Cuanto más cerca esté el valor de 1, mejor será el sistema evaluado.

En los experimentos, la métrica del *F1-score* se aplicará tanto a nivel de trama como a nivel de evento.

6.2.2 Resultados

En las gráficas de la Figura 32 y Figura 33 se muestran los resultados obtenidos en la fase de validación para las dos bases de datos en función del número de gaussianas seleccionado..

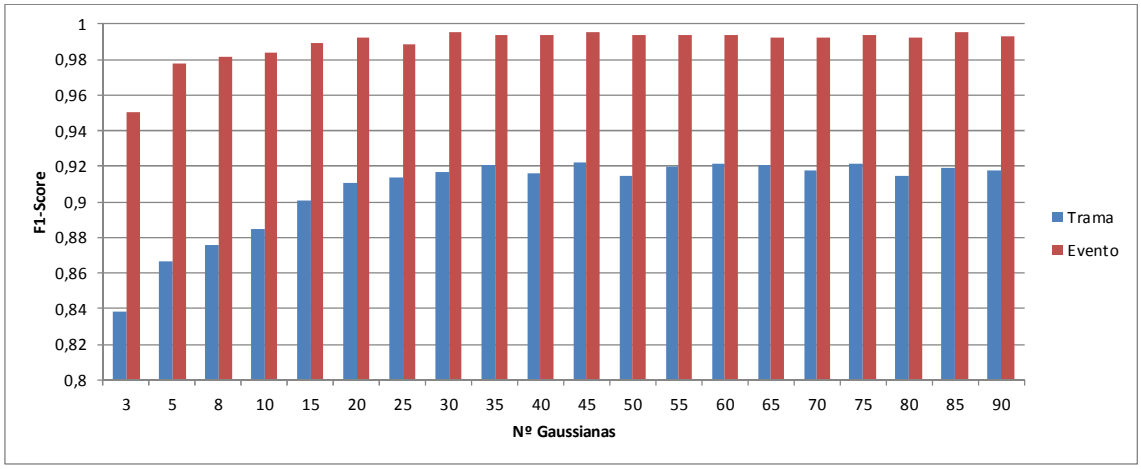


Figura 32 Clasificación BD Validación de eventos aislados con RWCP

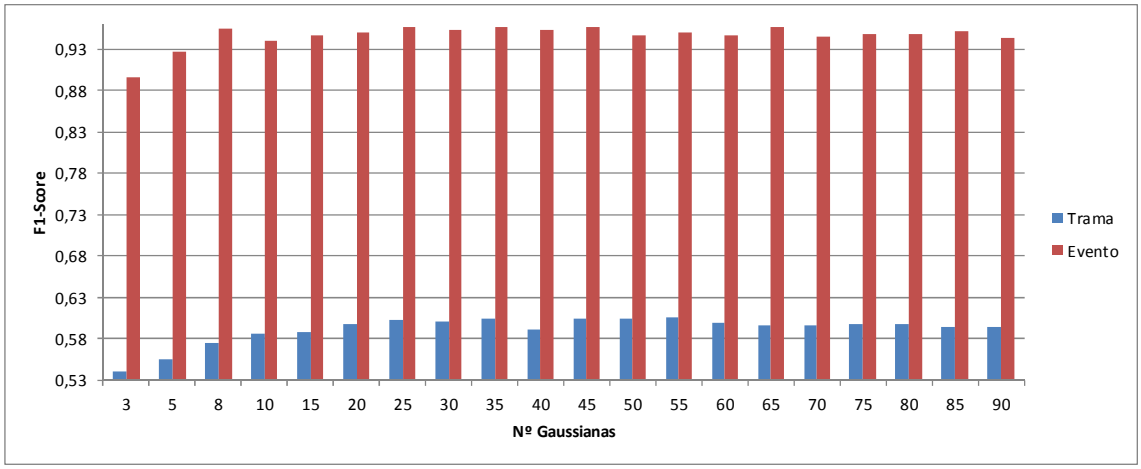


Figura 33 Clasificación BD Validación de eventos aislados con CHIL

Los mejores valores de *F1-score* a nivel de trama con la modificación del algoritmo EM son obtenidos con 45 gaussianas para ambas bases de datos. A nivel de evento RWCP

obtiene un *F1-score* de 1 con 80 gaussianas, mientras que el mejor valor para CHIL sigue obteniéndose con 45 gaussianas.

Los modelos creados con 45 gaussianas en RWCP y CHIL, fueros evaluados con el conjunto de datos de la base de datos de test obteniendo los valores que se muestran en la Tabla 18.

BD	Num. Gaussianas	*F1-score* Trama	*F1-score* Evento
RWCP	45	0,92	0,99
CHIL	45	0,60	0,92

Tabla 18: Clasificación con eventos aislados en RWCP y CHIL con 20 y 45 gaussianas sobre la BD de Test.

Un análisis más detallado del comportamiento de los algoritmos sobre la base de datos de test se ilustra en la gráfica de la Figura 34 y Figura 35. Los resultados son muy similares a los mostrados con la base de datos de validación.

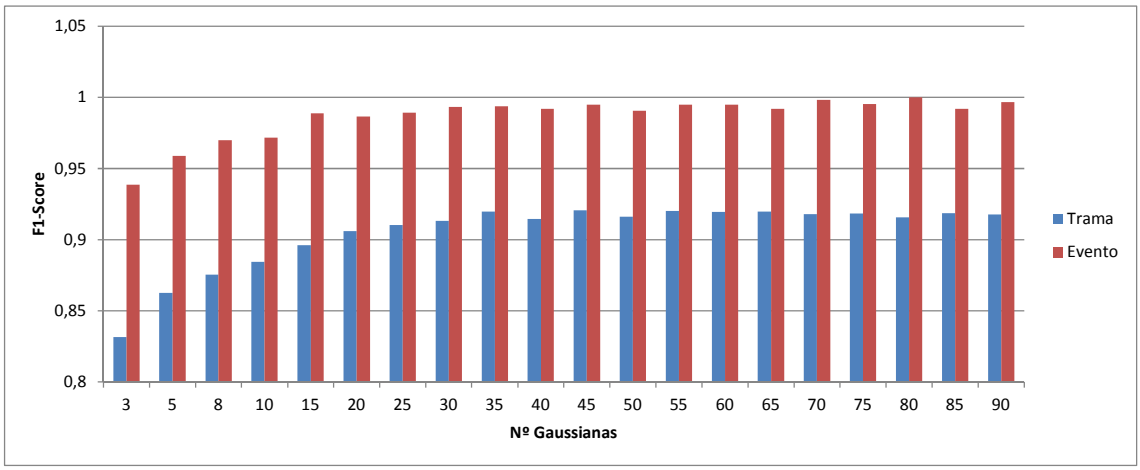


Figura 34 Clasificación BD Test de eventos aislados con RWCP

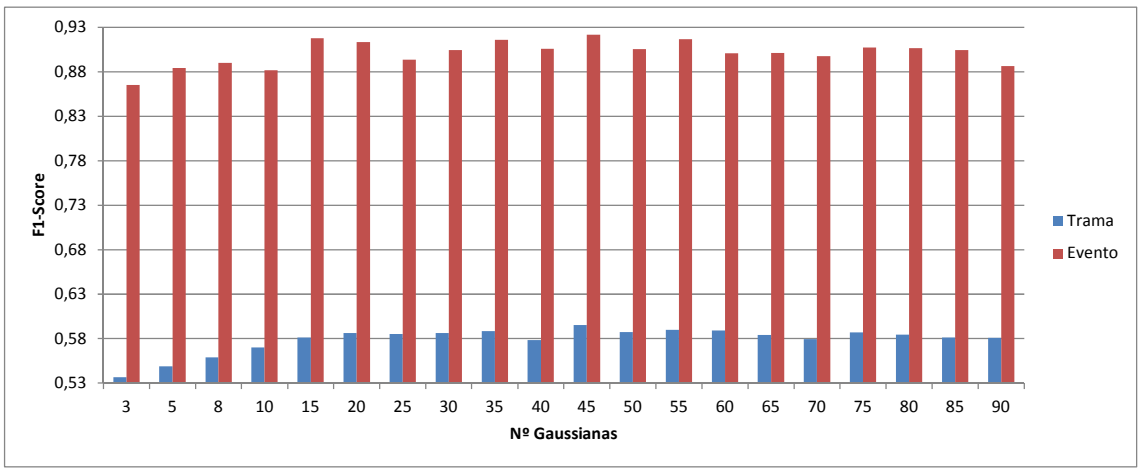


Figura 35 Clasificación BD Test de eventos aislados con RWCP

De los resultados se confirma también lo extraído de otros estudios o de la propia evaluación CLEAR donde se deja en evidencia que, cuando se trabaja con sonidos aislados conociendo el instante de tiempo de inicio y fin del evento, los ratios de aciertos con estas técnicas son muy altos. Además, también queda evidente que el número de gaussianas no es excesivamente importante en la clasificación. **Con unas pocas gaussianas se obtienen resultados muy cercanos a los óptimos**. Tanto con RWCP como con CHIL para todos los números de gaussianas evaluados el *F1-score* por eventos es cercano o superior a 0,9

6.3 Reconocimiento de Sonidos sobre Audio Continuo

En el apartado anterior se ha podido observar cómo el campo de la clasificación sobre sonidos aislados ofrece muy altos resultados de *precisión* y *recall*. Sin embargo, cuando se trabaja con audio continuo, que es lo que corresponde a una situación realista, es necesaria una etapa de detección previa.

Ya en las investigaciones previas en CHIL [21] se demuestra cómo, aunque los resultados obtienen altos ratios de precisión en la clasificación, cuando se trabaja con el flujo continuo de audio, los resultados obtenidos son muy inferiores. Si la detección no es muy robusta se pueden perder sonidos, identificar trozos de audio como evento que no son, pasar al clasificador un evento cortado o pasarle un trozo demasiado largo cuyo contenido no pertenezca en su totalidad a un evento.

En este capítulo se utilizarán dos técnicas de detección y se compararán entre sí para analizar cuál obtiene mejores resultados.

6.3.1 Estudio Experimental

Debido a que RWCP sólo posee eventos aislados, en este capítulo sólo podemos evaluar las grabaciones realizadas en CHIL. En esta base de datos, la relación entre *ruido ambiente* (o segmentos de audio no etiquetados como evento) y *eventos* se muestra en la Tabla 19.

	Tramas con Ruido ambiente	Tramas con Eventos a reconocer
DVD1	62,6%	37,4%
DVD2	61,1%	38,9%
DVD3	58,9%	41,1%

Tabla 19: Proporción de eventos en Base de Datos CHIL

Tal y como hemos visto en el capítulo 4, en las investigaciones analizadas existen dos estrategias de detección que se aplican a las bases de datos de eventos acústicos ("*Detección por clasificación*" y "*Detección y clasificación*") pero en ningún caso existe una comparativa entre ambas.

6.3.1.1 Estrategias de Detección

En la estrategia "*Detección por clasificación*" la detección se realiza en base a la clasificación de todas las tramas del audio continuo. En la estrategia "*Detección y clasificación*" el problema se aborda desde una perspectiva jerárquica, de forma que en el primer nivel se detectan eventos (sin asignarles etiqueta de a qué clase pertenecen) y, en el segundo nivel, los segmentos clasificados como eventos son enviados al clasificador para su etiquetado. Como se ha descrito en el estado del arte del capítulo 4, la estrategia *Detección y clasificación* puede realizarse tanto con modelos probabilísticos como con diferentes métodos basados en umbrales de intensidad, correlación, wavelets,… Para que la comparativa sea justa en los experimentos, estas estrategias serán aplicadas con los mismos modelos probabilísticos GMM. En los experimentos que se presentan a continuación se implementan las dos estrategias y se comparan entre ellas.

Utilizando el primer DVD de la base de datos CHIL se crearon los modelos GMM, incluyendo el modelo "*evento*" (que engloba a todos), el modelo "*no evento*" y el de la clase "*desconocidos*" (etiquetado en la base de datos CHIL). Una vez hecho esto, los pasos para las diferentes estrategias se indican en la Tabla 20.

Detección por Clasificación	Detección y Clasificación
Clasificar las tramas entre las 14 clases + la clase *No evento*.	Clasificar las tramas entre *Evento* y *No Evento*.
Aplicar suavizado. El resultado del suavizado dará el reconocimiento final del sistema.	Aplicar suavizado.
	Por cada bloque resultante clasificar las tramas entre las 14 clases y asignar al bloque general la etiqueta que se haya dado más número de veces.

Tabla 20: Proceso de las estrategias de detección

En ambas estrategias de detección se ha incluido una etapa de suavizado, cuyo fin es prevenir la falsa detección de eventos demasiado cortos (falsos positivos). Para ello se establecen dos parámetros: el número mínimo de **tramas consecutivas** (UTC, *Umbral Tramas Consecutivas*) con la misma etiqueta que es necesario disponer para detectar un evento con dicha etiqueta; y por otro lado el número mínimo de **tramas totales** (UTT) de un cierto evento necesarias para detectar el evento.

El algoritmo que aplica el umbral de *número mínimo de tramas consecutivas* UTC en el suavizado, recorre el vector de etiquetas predichas por el modelo GMM. Por cada trama i si ésta es distinta de la etiqueta de la trama anterior $i-1$, se observa si las (UTC-1) tramas siguientes a la actual ($i+1$, $i+2$,..,$i+UTC-1$) tienen todas la misma etiqueta que la trama i. En caso negativo, la etiqueta de la trama i es reasignada con el valor de la etiqueta de la trama $i-1$.

El *número mínimo de tramas totales* UTT se establece para eliminar los eventos resultantes del procesado previo cuya duración no sea igual o superior a UTT tramas. Un ejemplo gráfico del efecto de estos parámetros se muestra en la Figura 36.

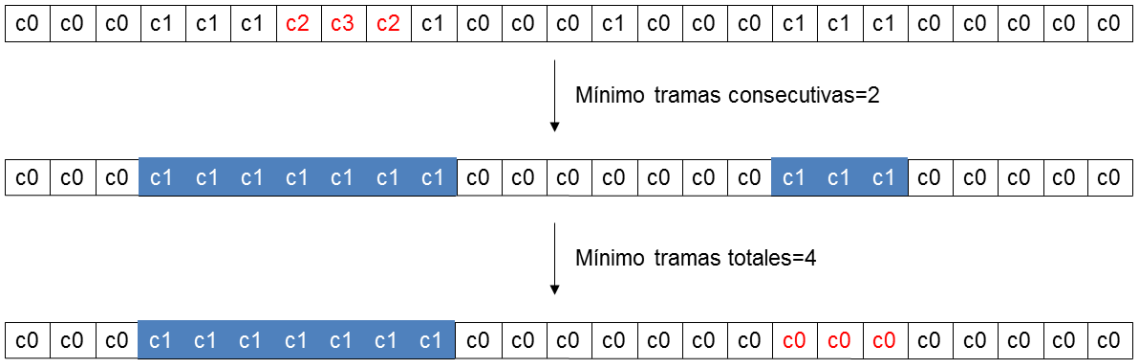


Figura 36 Parámetros del suavizado. ci indica trama de la clase i, siendo c0 la clase "no evento"

6.3.1.2 Métrica de Evaluación

En la identificación de eventos en audio continuo, la medida de *F1-score* sólo puede ser utilizada por tramas, ya que se desconoce el número de eventos a reconocer. Por ello, los parámetros que se utilizan en la evaluación de los resultados de reconocimiento en audio continuo son los siguientes:

- *F1-score a nivel de trama*

$$F1-Score = 2*\frac{(\text{Precision}*\text{Recall})}{(\text{Precision}+\text{Recall})} \quad (13)$$

En donde *Precision* y *Recall* están definidas en la ecuación 11.

- *Error de Detección (AEED) y Error de Reconocimiento (AEER)*

Las métricas AEED y AEER están basadas en la evaluación aplicada en los seminarios CLEAR 2006. A diferencia con CLEAR 2006, en nuestro caso tendremos en cuenta los eventos de la clase "*desconocidos*". Según esta métrica:

Un evento es correctamente detectado cuando el hipotético centro temporal está situado dentro del intervalo que corresponde al evento real sea cual sea su etiqueta.

Un error de inserción ocurre cuando el centro temporal del hipotético evento está situado en un intervalo correspondiente al silencio.

Un error de eliminación ocurre cuando hay un evento en la lista de eventos a detectar que no ha sido marcado como detectado.

Un error de sustitución ocurre cuando se detecta correctamente un evento pero su etiqueta no corresponde con la de referencia.

El Error de Detección AEED evalúa únicamente la etapa de detección (al no añadir las sustituciones) y se computa como:

$$AEED= (D+I)/N*100 \quad (14)$$

Donde N es el número de eventos a detectar, D el número de errores de eliminación e I los errores de inserción.

El Error de Reconocimiento (AEER) calcula el error global del sistema combinando las etapas de detección y de clasificación. Se diferencia de AEED en que se suman en la ecuación las sustituciones (S). De esta forma el error global (AEER) se computa como:

$$AEER= (D+I+S)/N*100 \quad (15)$$

Tanto el número de eventos eliminados (D) como el de nuevas inserciones (I) son variables correspondientes a la etapa de detección, ya que no consideran la identidad del evento. Las sustituciones (S) dependen evidentemente de la etapa de clasificación pero también de la de detección si se han suprimido tramas o insertado silencios.

6.3.2 Resultados

En este apartado se muestran los resultados obtenidos para las diferentes pruebas.

6.3.2.1 Mejor Combinación de Tamaño de Ventana y Número de Gaussianas

Antes de comparar las dos estrategias de detección se procedió a realizar una configuración previa de los parámetros del experimento (tamaño de ventana y número de gaussianas). En el apartado 6.2 se estableció un tamaño de ventana de 20 milisegundos para comparar las bases de datos RWCP y CHIL con el mismo parámetro, debido a la poca duración de los sonidos de RWCP. En este apartado, al trabajar tan sólo con la base de datos de CHIL, para evaluar las dos estrategias de detección se procedió previamente a ampliar las posibilidades de tamaños y buscar el número óptimo de gaussianas. De esta forma, los experimentos realizados en el apartado 6.2

para eventos aislados, con ventanas de 20 milisegundos se repitieron con tamaños de ventana mayores (hasta 100 ms).

La gráfica de la Figura 37 ofrece la media de *F1-score* a nivel de evento para los diferentes tamaños de ventana sobre la base de datos de validación.

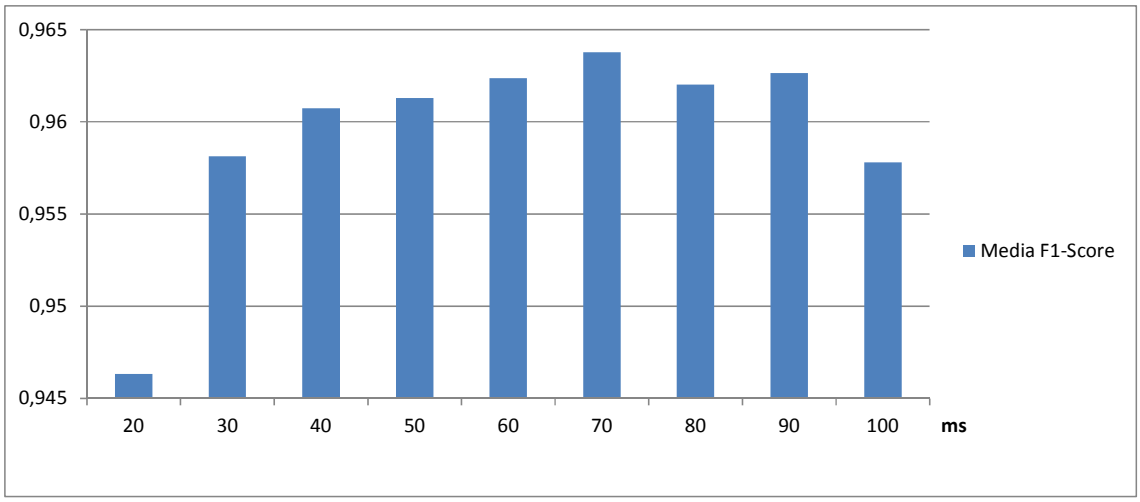


Figura 37 Media *F1-score* a nivel de evento por tamaño de ventana en CHIL

Tal y como se observa, los valores *F1-score* medios son muy similares. Si bien es cierto que el tamaño de 20 milisegundos aplicado anteriormente es el que peores resultados da, la diferencia entre todos ellos es muy pequeña. La Figura 38 muestra el número óptimo de gaussianas para cada tamaño de ventana. Este valor varía significativamente en función del tamaño de la ventana utilizada.

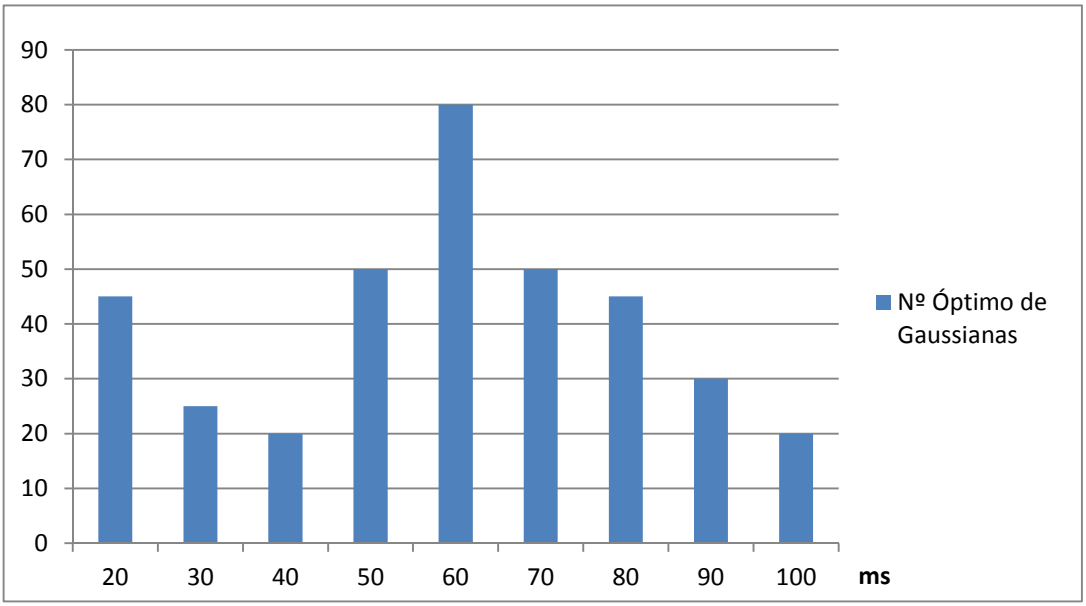


Figura 38 Número óptimo de gaussianas por tamaño de ventana en CHIL

Para los siguientes experimentos se estableció el tamaño de ventana en **70 milisegundos,** por ser éste el que obtuvo el valor medio de *F1-score* más alto (0,964).

El **número de gaussianas se fijó en 50** debido a que, para ventanas de 70 milisegundos fue el que dio mejores resultados.

6.3.2.2 Detección por Clasificación

Haciendo uso de la base de datos de testeo, se buscó la configuración óptima del suavizado. Para obtener el valor óptimo para el número mínimo de tramas consecutivas UTC se hizo una evaluación sobre un rango de valores UTC comprendidos entre 2 y 60. Los resultados obtenidos del experimento se muestran en la Figura 39 (valor AEED y AEER) y la Figura 40 (valor de *F1-score* a nivel de trama).

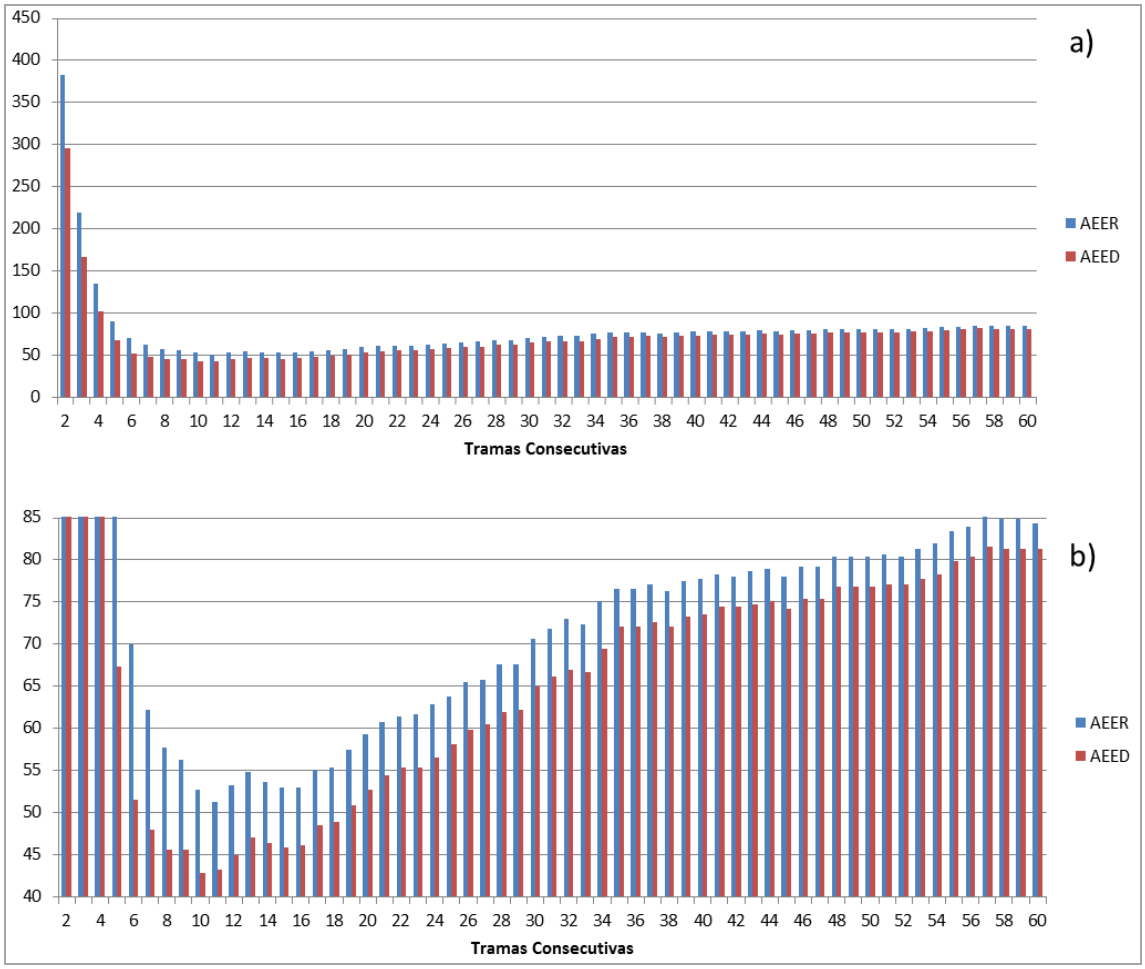


Figura 39 Detección por clasificación – a) AEER y AEED en función del valor UTC. b) Zoom de la gráfica

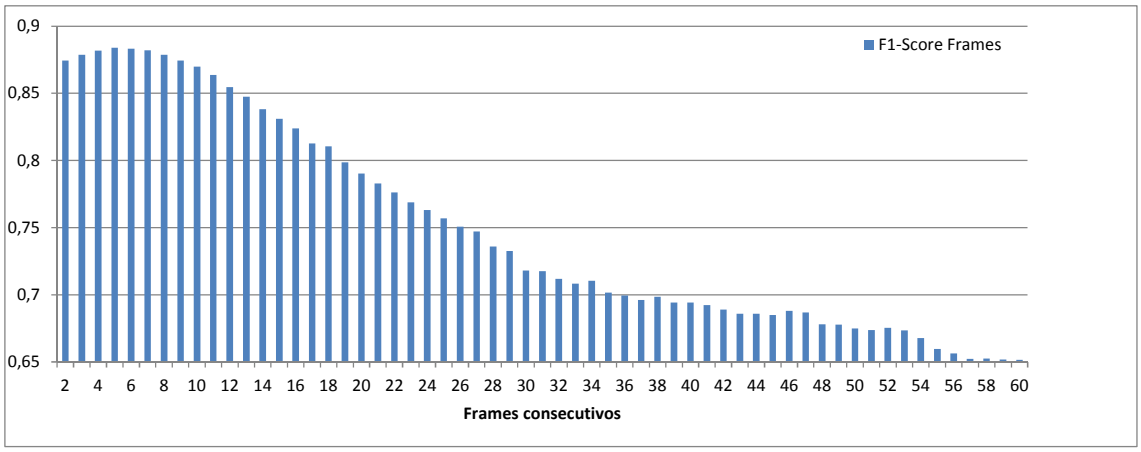


Figura 40 Detección por clasificación - *F1-score* en función del valor UTC

En la gráfica de la Figura 39 se observa cómo a medida que se incrementa el valor de UTC los errores de detección (AEED) y reconocimiento (AEER) disminuyen rápidamente. Al llegar a 10 (AEED) o 11 (AEER) tramas consecutivas el error vuelve a subir ligeramente. También *F1-score* disminuye a medida que aumentamos UTC (Figura 40) aunque en este caso, las pendientes de subida y de bajada no son tan pronunciadas y el número de tramas consecutivas que dio mejor resultado fue de UTC=5 (*F1-score*=0,884).

Como segundo procesado se añadió el *número mínimo de tramas totales* UTT que un sonido debe tener para ser clasificado como tal. Se consideró únicamente el rango 10<UTT<60 teniendo en cuenta las duraciones de los eventos (ver Figura 31). Para cada valor de UTC se calculó el valor de UTT óptimo. Los resultados de aplicar este procesado se muestran en la gráfica de la Figura 41.

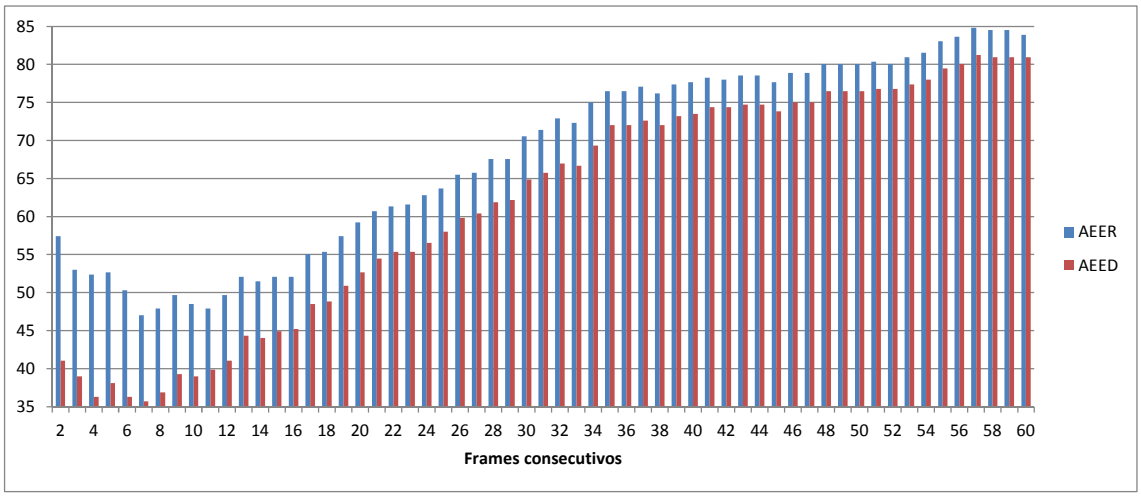


Figura 41 Detección por clasificación - AEER y AEED en función de UTC y UTT

Los beneficios de introducir este parámetro en los algoritmos se evidencian en la comparación de la Figura 41 con la Figura 39. Este suavizado hace que el error de detección (AEED) descienda considerablemente evitando falsos positivos omitidos en

la fase anterior. Así como en la gráfica de la Figura 39 los valores AEED y AEER son muy similares, al introducir el umbral de número mínimo de tramas totales UTT la diferencia entre esos errores crece. Adicionalmente, el valor óptimo para UTC desciende. Los sistemas con errores más bajos de AEED y AEER hicieron uso de un valor UTC=7, obteniendo valores de AEED=35,714 y AEER=47,024.

En la Tabla 21 se indica la combinación de parámetros que proporciona el mejor resultado al sistema y, consecutivamente, en la gráfica de la Figura 42 se muestran los errores de inserción, eliminación y sustitución obtenidos por el mismo para cada tipo de sonido.

UTC	UTT	*F1-score*	AEED	AEER
7	28	0,882	35,714	47,024

Tabla 21: Detección por clasificación – Mejor combinación de parámetros

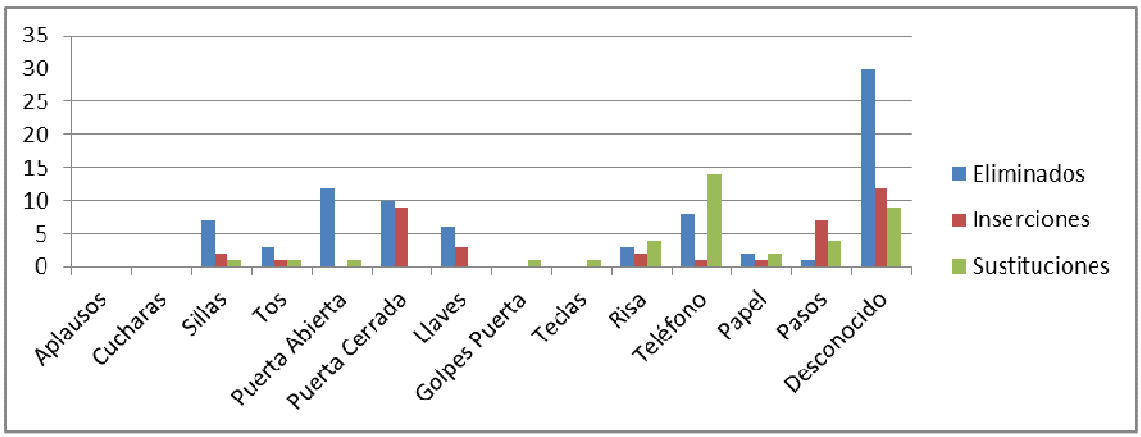


Figura 42 Detección por clasificación – Errores de Inserción, Eliminación y Sustitución para la mejor combinación

Como puede observarse, el mayor número de fallos recae en los eventos eliminados. Trabajar con eventos tan cortos hace que sea difícil detectar todos los eventos sin cometer fallos. Además, dentro de la categoría "*desconocidos*" el número de eventos eliminados es aún mayor que en el resto de categorías debido a que se trabaja con conjuntos de sonidos diferentes y el sistema no puede disponer de un modelo que aglutine a todos ellos.

Si comparamos los resultados obtenidos con los aportados en CLEAR 2006 [123] la mejora es significativa con respecto a los sistema propuestos UPC-D (AEER=58,9) y CMU-D2 (AEER=52,5). El sistema ITC-D2 obtiene un error menor (AEER=33,7), sin embargo, en la métrica AEER de estos tres sistemas no tenían en consideración la clase "*desconocido*", siendo esta clase la que produce los mayores errores en nuestro sistema (36,58% de los eventos eliminados pertenecen a esta clase). Además, el sistema ITC-D2 utiliza modelos HMM y estos requieren conocer el inicio y final de todos los eventos a entrenar, no práctico para un sistema funcional final (véase apartado 6.1.2).

6.3.2.3 Detección y Clasificación

La estrategia "*Detección y Clasificación*" proporciona los resultados de la Figura 43 y Figura 44 al aplicar el procesado correspondiente al *número de tramas consecutivas* UTC en la base de datos de testeo.

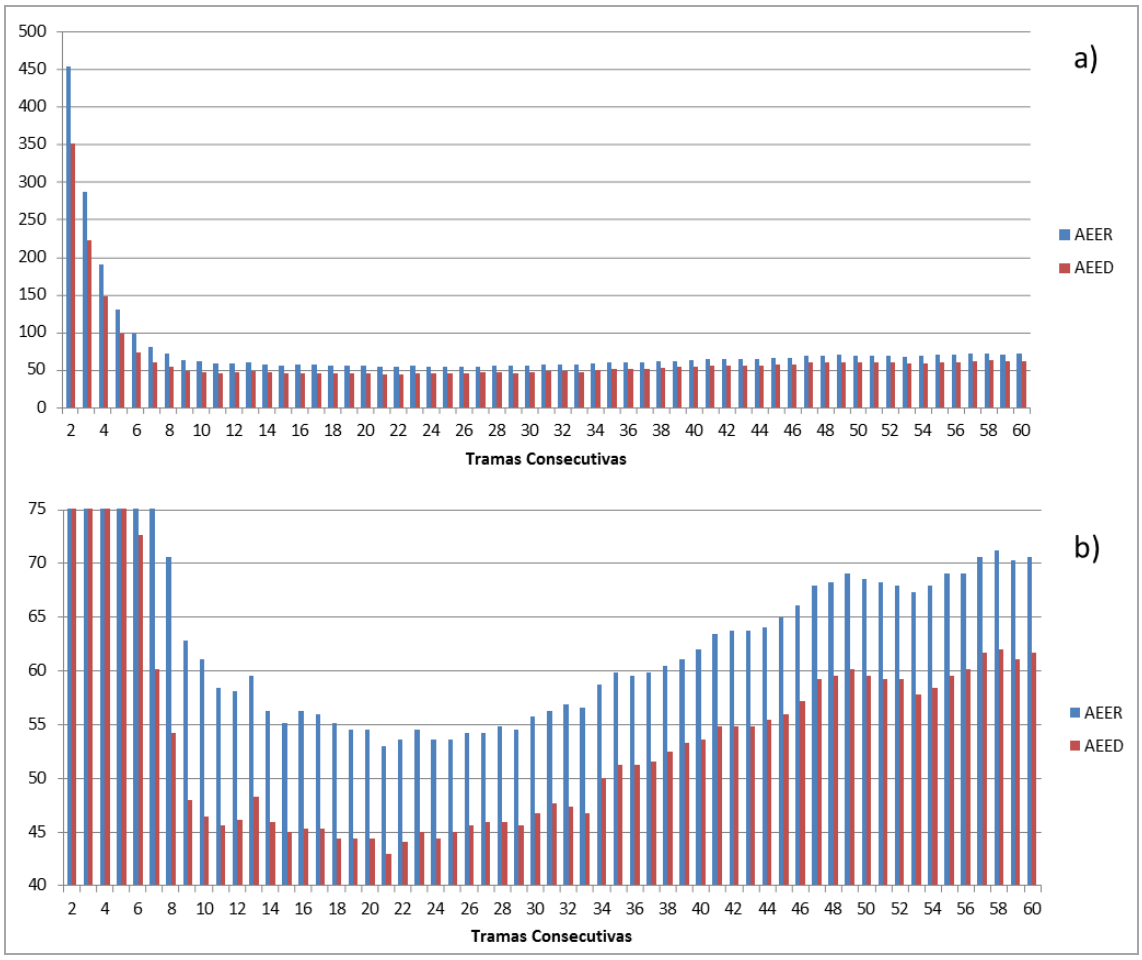


Figura 43 Detección y clasificación – a) AEER y AEED en función del valor UTC. b) Zoom de la gráfica

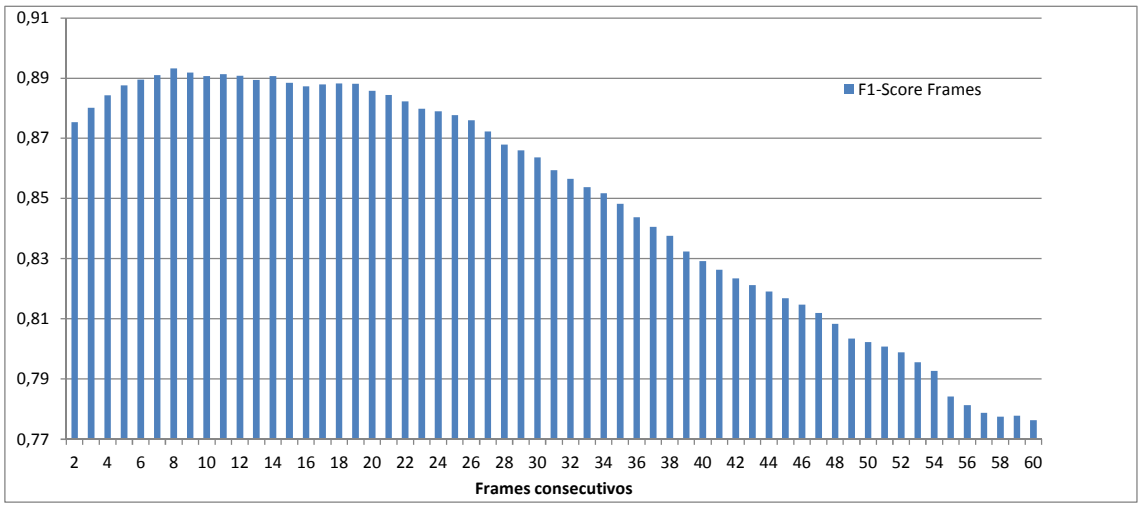


Figura 44 Detección y clasificación - *F1-score* en función de UTC

La evolución de las gráficas sigue la misma tendencia que en la estrategia *Detección por Clasificación*, (Figura 39 y Figura 40) obteniéndose resultados muy similares. Los mejores resultados de AEED y AEER obtienen un valor de 42,857 y 52,976 respectivamente. Los valores mínimos de AEED y AEER se obtienen para UTC=21 en este caso, muy superior UTC=10 (para AEED) o UTC=11 (para AEER) en el caso de *Detección por clasificación.* Esto implicaría un retardo de valor doble en la detección. Sin embargo la diferencia en AEER y AEED para UTC=10 y UTC=21 es muy pequeña.

Los resultados de aplicar el umbral de *mínimo número de tramas totales* UTT se muestran en la gráfica de la Figura 45.

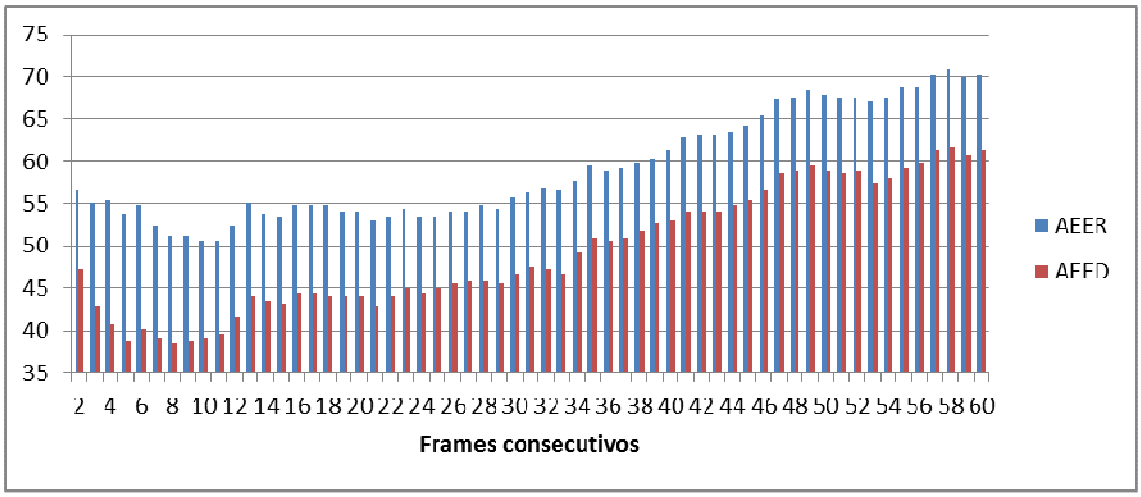


Figura 45 Detección y clasificación - AEER y AEED en función de UTC y UTT

Al igual que en el caso anterior (*Detección por Clasificación*), el valor UTT proporciona una reducción significativa de los errores, aunque estos son más altos que en la estrategia *Detección por Clasificación*. En la gráfica de la Figura 45 los valores mínimos de AEED y AEER alcanzan valores de 38,690 y 50,595 respectivamente mientras que previamente los valores obtenidos eran de 35,714 y 47,024.

A continuación en la Tabla 22 se indica la combinación de parámetros que proporciona el mejor resultado para la estrategia *Detección y Clasificación* y, consecutivamente, en la gráfica de la Figura 46 se muestran los errores de inserción, eliminación y sustitución obtenidos.

UTC	UTT	*F1-score*	AEED	AEER
10	59	0.8907	38.9880	50.5952

Tabla 22: Detección y clasificación – Mejor combinación de parámetros

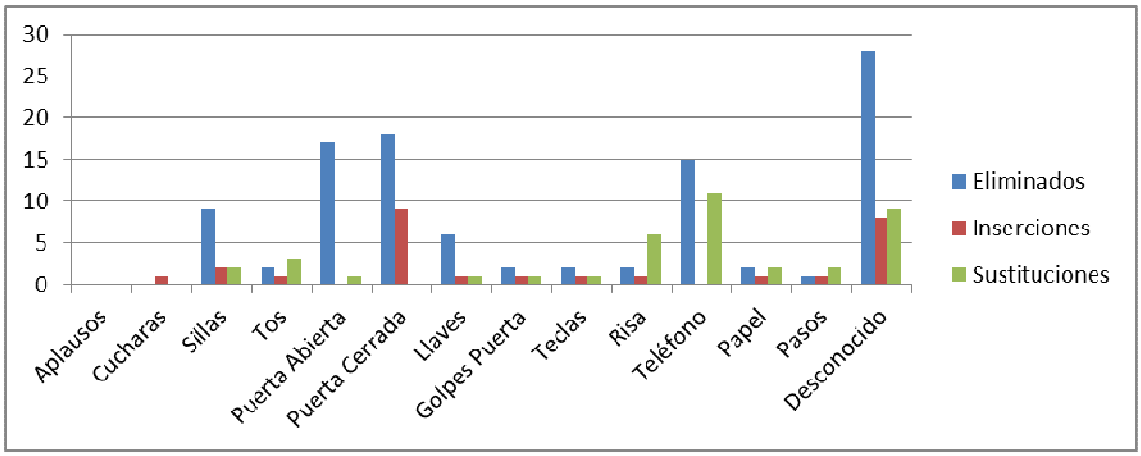


Figura 46 Detección y clasificación – Errores de Inserción, Eliminación y Sustitución para la mejor combinación

También en esta estrategia existe un problema con el elevado número de eventos eliminados que se producen (30,95%), más elevado aún que en *Detección por Clasificación* (24,4%). Esto posiblemente sea debido a la utilización de UTT más altos. Al aumentar este número hace que, eventos cortos, sean rechazados con mayor facilidad.

6.4 Conclusiones del Capítulo

En este capítulo se ha realizado una búsqueda de combinaciones óptimas de parámetros utilizando dos bases de datos de sonidos no-habla habituales en la literatura científica: RWCP y CHIL. Mediante métodos de clasificación probabilísticos GMM se ha comprobado cómo la gran variedad de sonidos existentes, su distinta procedencia y el entorno en el que se producen hace difícil encontrar un número de gaussianas óptimo para tamaños de ventana diferentes. Aun así, los resultados obtenidos en la etapa de clasificación son muy prometedores. Incluso alejándose de la parametrización óptima del clasificador, las características acústicas estudiadas demuestran comportarse de forma muy aceptable con este tipo de sonidos. Cuando se trabaja con sonidos aislados los resultados de clasificación son cercanos o superiores al 0,9 de *F1-score* a nivel de evento, aunque el número de gaussianas se aleje del óptimo. Esto ayuda a la generalización de los algoritmos, pudiéndose encontrar valores muy cercanos a los óptimos reduciendo el número de gaussianas aplicable a conjuntos muy variados de sonidos. Esto también demuestra que la mayor dificultad en estos sistemas no radica en la etapa de clasificación.

La comparativa realizada entre las dos estrategias de detección da como ganadora a la denominada como "*Detección por Clasificación*". Esta estrategia demuestra obtener mejores resultados (AEER=47,024) para la medida de AEER sobre todo por reducir el número de eventos eliminados mejor que la técnica "*Detección y Clasificación*". Si comparamos los resultados obtenidos con los aportados en CLEAR 2006 [123] la

mejora es significativa con respecto a los sistema propuestos UPC-D (AEER=58,9) y CMU-D2 (AEER=52,5). El sistema ITC-D2 obtiene un error menor (AEER=33,7), sin embargo, en la métrica AEER de estos tres sistemas no tenían en consideración la clase "*desconocido*", siendo esta clase la que produce los mayores errores en nuestro sistema (36,58% de los eventos eliminados pertenecen a esta clase). Además, el sistema ITC-D2 utiliza modelos HMM y estos requieren conocer el inicio y final de todos los eventos a entrenar, no práctico para un sistema funcional final (véase apartado 6.1.2).

En la etapa de detección, el mayor problema se encuentra en los eventos desconocidos. Ambas estrategias presentan carencias al detectar y clasificar eventos fuera del conjunto fijado. En entornos no controlados es muy probable que se produzcan sonidos nuevos con los que los algoritmos no hayan sido entrenados. Es necesario fortalecer esta problemática si se desea tener un sistema robusto y funcional para las personas con limitaciones auditivas.

7. Aplicación de las Técnicas de Reconocimiento de Sonidos No-Habla en el Hogar

La vida independiente en el hogar conlleva poder valerse por uno mismo y ser capaz de actuar frente a circunstancias adversas que puedan suceder. Sin embargo, las limitaciones sensoriales auditivas pueden implicar una pérdida de información que recaiga en la no actuación frente a una situación de peligro.

En este capítulo se presenta el diseño y la evaluación de un sistema que realiza la detección y la clasificación de los eventos acústicos en el entorno más demandado por los potenciales usuarios de estas aplicaciones: el hogar. Las técnicas utilizadas en el capítulo 6 son aplicadas ahora sobre una base de datos de sonidos ajustada a condiciones realistas, tanto en la tipología de sonidos que se utilizan como en el entorno acústico en el que se van a obtener. Inicialmente, se utilizan micrófonos profesionales para la obtención del sistema de referencia. Este sistema ha sido utilizado tanto para el aviso a personas con problemas auditivos como para su uso en sistemas inteligentes capaces de usar la información para el reconocimiento de actividades. Por otra parte, se implementa y se evalúa el sistema diseñado sobre dispositivos móviles, en diferentes condiciones de ruido y movilidad.

Como resultado de esta investigación, el capítulo concluye con la descripción del desarrollo de una aplicación funcional de reconocimiento automático de sonidos del hogar para móvil capaz de trabajar en tiempo real sobre sistema operativo Android. Su diseño, además de estar influenciado por varios estándares y recomendaciones de accesibilidad, está guiado por los resultados de la encuesta elaborada en el capítulo 5..

7.1 Reconocimiento de Sonidos No-Habla en el Hogar

En este apartado se muestra la metodología y técnicas utilizadas para el desarrollo de un sistema de reconocimientos de sonidos del hogar sobre plataforma PC. Aunque el objetivo último es el desarrollo de una aplicación para dispositivos móviles, se hace necesario investigar el comportamiento de los algoritmos en un entorno menos dificultoso en el que el hardware utilizado reúne unas características de calidad mínimas y nos permitirá obtener resultados que podrán ser utilizados como referencia. Además este sistema fue integrado en una aplicación de reconocimiento de actividades de la vida diaria, en el contexto del proyecto europeo RUBICON. Esta integración se describe en el último apartado de esta sección.

7.1.1 Corpus y Análisis de sonidos

Como ya se ha comentado en puntos anteriores, aunque existen bases de datos para la investigación que contienen sonidos no-habla que pueden ser encontrados en el hogar, estos sonidos son muy limitados y la mayoría de ellos no coinciden con los sonidos más relevantes indicados por las personas con problemas auditivos. En el capítulo 6 se trabajó con un conjunto reducido de la base de datos de RWCP donde se encontraban sonidos tales como sartenes, vasos, platos, secadores de pelo,… siendo el reconocimiento de estos sonidos de gran utilidad en sistemas AAL con el objetivo de reconocer actividades de una persona. Sin embargo, el número de sonidos de aviso para personas con problemas auditivos que contiene esta base de datos es muy escaso. Los únicos sonidos de aviso de interés que en ella podemos encontrar son el teléfono, el despertador y la cerradura de la puerta. Además, existen dos problemas aún mayores que hacen necesaria la creación de otro corpus diferente. Por una parte los sonidos fueron grabados de forma aislada, por lo que no se puede evaluar su rendimiento en la etapa de detección y, por otro lado, los sonidos carecen de ruido de fondo ya que éstos fueron grabados dentro de una cámara insonorizada.

Por estos motivos, se decidió llevar a cabo la evaluación de las técnicas anteriores en un entorno más realista. Con este fin, el estudio que se presenta a continuación fue realizado con sonidos producidos en un *Homelab* totalmente funcional de 45m^2. La vivienda cuenta con una zona de entrada, salón-comedor, cocina, dormitorio y baño (véase Figura 47). El *Homelab* está completamente amueblado y equipado con diversas tecnologías para el control inteligente del entorno y monitorización de actividades en la vivienda. Entre las diferentes tecnologías que se encuentran cabe destacar una red de micrófonos distribuidos en los distintos espacios de la vivienda que fue la que se utilizó para la grabación, entrenamiento y evaluación del sistema que se presenta en el presente documento.

Figura 47 Homelab utilizado para la grabación de las muestras de sonidos del hogar

Para el estudio experimental se realizó una selección de 12 clases de sonidos teniendo en cuenta los sonidos mejor valorados en la encuesta del capítulo 5. Además, en este

conjunto se incorporaron otros sonidos que, aunque no poseen un carácter de aviso, se entendieron como eventos acústicos importantes para poder utilizar en sistemas *AAL* (sillas, vajilla, armario, microondas). La tabla de todos los sonidos junto a la ubicación de las fuentes de sonidos evaluados se muestran en la Tabla 23 y en la Figura 48 respectivamente.

Personas con problemas auditivos	Reconocimiento de actividades
Timbre Puerta	Sillas
Portero automático	Vajilla
Golpes Puerta	Apertura de armarios
Teléfono	Microondas Funcionando
Grifo	Microondas Apertura
Microondas Fin	
Persona hablando	

Tabla 23: Sonidos del hogar por tipo de aplicación

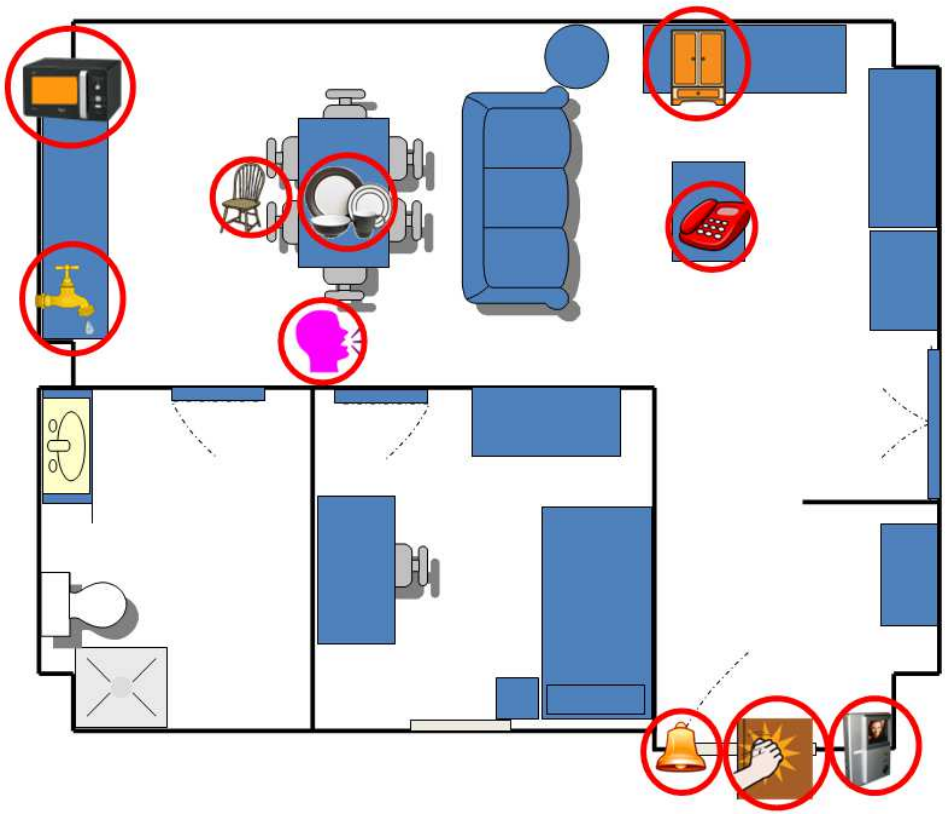


Figura 48 Localización de los sonidos del hogar en el homelab de pruebas

7.1.2 Metodología Aplicada

Del grupo de sonidos seleccionados muchos de ellos son sonidos sintéticos (timbre de la puerta, portero automático, teléfono y pitido del microondas) cuyas características dependen del fabricante y pueden ser totalmente distintos en cada casa. En una casa un timbre podrá tener un sonido cuyas frecuencias se mantengan en el tiempo y en otra, sin embargo, el sonido del timbre podría ser más melódico haciendo que se encuentre mayor variabilidad en las frecuencias. La robustez actual de los sistemas

comerciales de reconocimiento del habla viene precedida de un gran trabajo metodológico de creación de enormes bases de datos por grandes empresas. Los corpus con los que se cuenta son trabajos de varios meses / años de grabación con un gran número de personas de diferentes lugares y con amplios textos con los que entrenar. Llevar esto al terreno de los sonidos relacionados con el habla es totalmente inviable debido a su gran variabilidad por lo que, debido al interés práctico de este capítulo, parece clara por tanto plantear la necesidad de entrenar el sistema en cada vivienda con los sonidos que uno tenga y desee reconocer, y descartar la idea de tener una base de datos universal que pueda ser utilizada en todas los hogares.

Partiendo de la idea de construir un sistema capaz de ser entrenado inicialmente por el usuario, la sistemática utilizada debe ser rápida y sencilla de aplicar. En los sistemas de reconocimiento, el número de muestras con las que se entrena un sistema es determinante en su precisión, y contar con un número muy reducido de las mismas puede ocasionar problemas de clasificación. Sin embargo, no es viable pedir a la persona que va a entrenar el sistema que grabe, por ejemplo, 100 muestras de cada sonido.

La base de datos que se ha creado se ha grabado siguiendo dos metodologías diferentes, la primera para obtener las muestras de entrenamiento y la segunda para obtener las muestras de testeo.

Para crear el subconjunto de datos de entrenamiento, las fuentes acústicas emisoras de los sonidos se accionaron durante un tiempo aproximado de 2 minutos ininterrumpidamente. Como ejemplos, el grifo se dejó abierto durante este periodo de tiempo y el portero automático fue pulsado de seguido hasta que el tiempo estipulado finalizó. Para poder separar los eventos acústicos del ruido de fondo, a la lista de sonidos se añadió la clase "*no evento*", grabándose para el entrenamiento dos minutos de audio continuo del ruido del *Homelab*.

Para crear el subconjunto de datos de testeo, se grabó el audio en el *Homelab* de forma continuada durante un tiempo de una hora. Durante este tiempo, aproximadamente 25 sonidos fueron producidos por cada clase (301 muestras de sonidos en total).

Tanto durante la grabación de la parte de entrenamiento como para la de test, no se realizó un control del ruido ambiental o externo procedente de ambientes cercanos. Así por ejemplo, el aire acondicionado se mantuvo en modo automático, se produjeron ruidos de máquinas procedentes de laboratorios contiguos etc. Es decir, podemos decir que se trata de un entorno de grabación semi-controlado.

7.1.3 Experimentos y Resultados

Una vez creada la base de datos, se realizó el cálculo de parámetros utilizando en ventanas de 60 milisegundos. Al igual que en los experimentos previos, se calcularon los trece primeros coeficientes MFCCs, el parámetro ZCR, el parámetro Spectral Centroid y el parámetro Roll-Off Point, además de sus primeras y segundas derivas (delta y delta-delta). El vector resultante está formado de 48 parámetros. Las GMMs se entrenaron con N=40 gaussianas.

Los experimentos descritos en las siguientes líneas fueron realizados usando tan sólo el micrófono localizado en la cocina del *Homelab*. En la base de datos de entrenamiento, mediante la asignación de un umbral de intensidad, los segmentos de señales con amplitudes inferiores al umbral fueron eliminados automáticamente.

Para determinar la *precisión* y *recall* del sistema se hizo una primera evaluación de sólo clasificación (con los sonidos aislados) para observar el porcentaje de error que el sistema poseía en esta etapa donde, con otras bases de datos, se habían alcanzado muy buenos resultados. El *F1-score* obtenido fue de 0,97. En la Figura 49 se observa la matriz de confusión resultante tanto para la *precisión* como para la *recall*, calculadas de acuerdo a la ecuación 11.

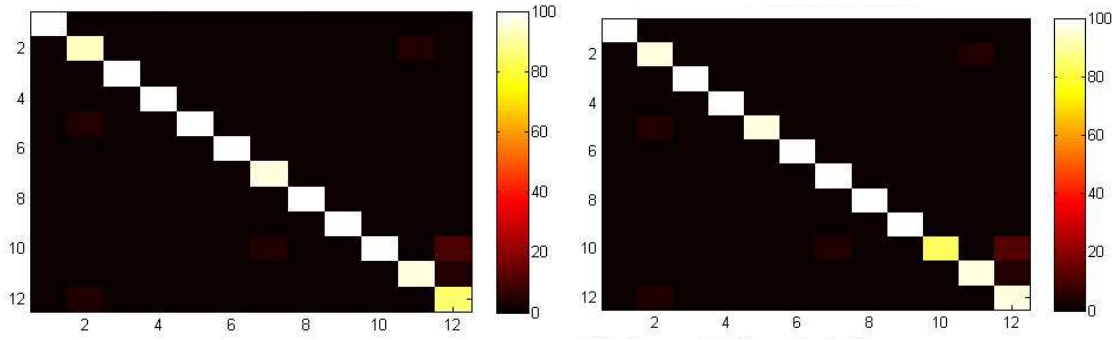


Figura 49 Matriz de confusión para la precisión (Izq) y recall (Dcha) de la clasificación de sonidos del hogar

Consecutivamente se procedió a evaluar el sistema sobre el audio en continuo. El valor de *AEER* (obtenido con la expresión 15) fue de 17.6. La mayoría de los errores se produjeron con sonidos impulsivos y de duración corta. Aunque durante las grabaciones los sonidos se activaron durante breves segundos de tiempo, los sonidos más impulsivos como apertura de microondas, pitido de microondas, habla (tan sólo se grababa la palabra "*hola*"), golpes en la puerta,… son los que obtuvieron más cantidad de errores. La Figura 50 muestra el número de eventos eliminados, sustituidos e insertados.

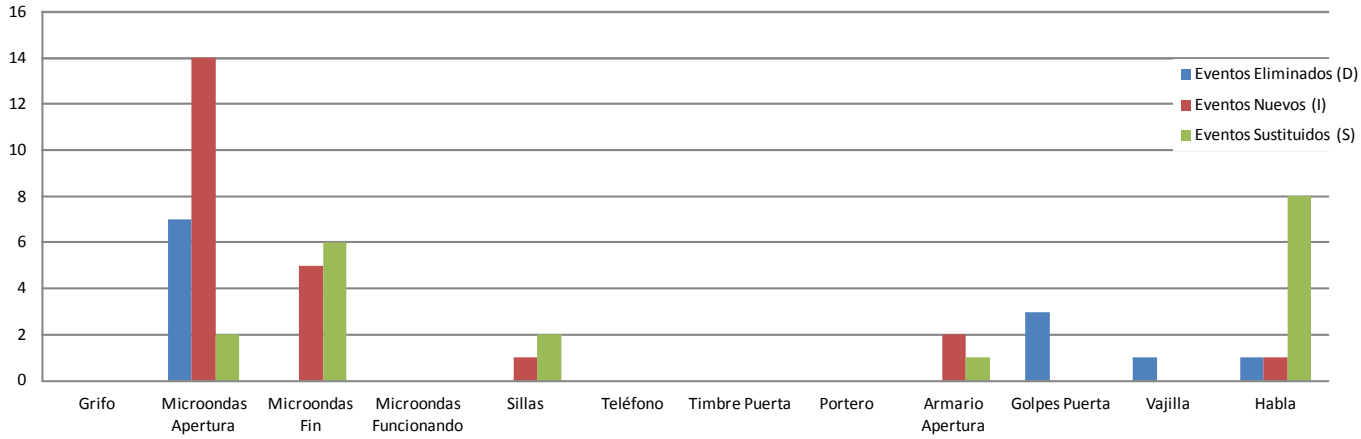


Figura 50 Detección y clasificación – Errores de Inserción, Eliminación y Sustitución en el hogar

7.1.4 Información de Contexto

Como se expuso en el capítulo 3, cuando se trata de un entorno no controlado, incluso para el oído humano, la única forma que una persona tiene de reconocer los sonidos cortos e impulsivos es combinando información acústica con información del ambiente. Gracias a las nuevas tecnologías, existen en la actualidad diversos tipos de sensores que pueden aportar información adicional de qué está sucediendo en un entorno. Aparte de la información sonora que se puede recoger desde un micrófono, en una vivienda puede haber sensores que aportan posibilidades adicionales. Si tenemos en cuenta que la mayoría de eventos acústicos son producidos por acciones de un usuario (golpes en la puerta, apertura de armarios, apertura de grifo,…) no es exagerado pensar la importancia que tiene realizar un seguimiento de la persona para inferir accionamientos de fuentes acústicas.

Para comprobar esta hipótesis se hizo uso de las facilidades del *Homelab* donde se realizaron las grabaciones. En cada habitación del *Homelab* se encuentran instalados sensores de presencia que se activan cuando detectan una persona en el área. Por esta razón, la información de estos sensores fue combinada con las grabaciones de audio usadas para el entrenamiento y testeo.

La posibilidad de combinar información de presencia del usuario permite reducir el conjunto de sonidos a reconocer en un momento dado. Si, por ejemplo, sabemos que la persona no se encuentra en la cocina podemos descartar de la lista de eventos sonidos como el grifo o la apertura del microondas. Para ello se realizó un conjunto de reglas simples como se muestra a continuación:

```
if (not presencia_en_entrada and (frame_rec==('portero' or 'timbrePuerta' or
  'golpesPuerta')))
        frame_rec='no_evento';
elseif (not presence_en_salón and frame_rec=='armario')
        frame_rec='no_evento';
elseif (not presence_en_cocina and (frame_rec==('grifo' or 'microondas' or 'sillas' or
  'vajilla')))
        frame_rec='no_evento';
end
```

Siguiendo estas reglas, las tramas clasificadas por el GMM como sonidos no apropiados a la localización del usuario fueron re-asignadas con la clase "*no evento*". Una vez realizado este procedimiento el error del sistema se volvió a calcular con la fórmula del *AEER* utilizada anteriormente. Aplicando la combinación de información acústica con la información de los sensores de presencia el error del sistema se redujo del 17.6 al 11.6. La gráfica de la Figura 51 demuestra la hipótesis planteada. El añadir la información de los sensores de presencia consigue eliminar falsos positivos que, sólo utilizando la información del audio, no habían sido eliminados. Incluso, en algunos casos como en el de la vajilla, soluciona problemas de sustitución.

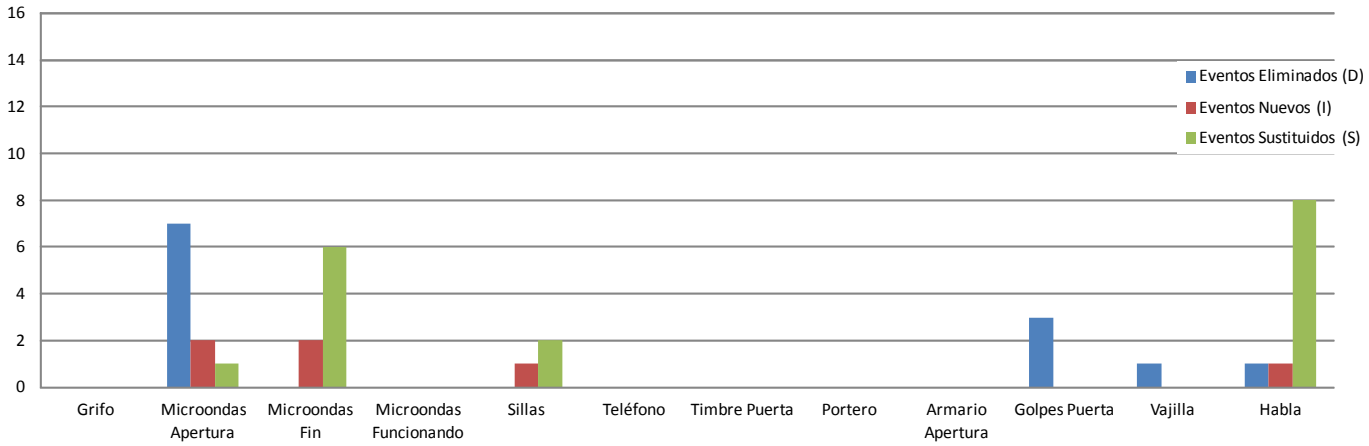


Figura 51 Detección y clasificación – Errores de Inserción, Eliminación y Sustitución en el hogar con contexto

Aplicando las técnicas mencionadas se implementó el software de reconocimiento en tiempo real cuyos resultados fueron publicados en [126] y en [127]. Buscando eficiencia y velocidad en el procesado, el lenguaje de programación elegido fue C/C++. En este apartado cabe destacar la utilización de las librerías GMM del grupo de procesado del habla de la UPV/EHU Signal Processing Laboratoy (*Aholab*) que fueron utilizadas para crear y entrenar los Modelos de Mezclas Gaussianas arriba descritos. El software se instaló en un PC servidor capaz de trabajar en tiempo real monitorizando y procesando el audio del entorno.

Siguiendo las indicaciones de las encuestas del capítulo 5, como medio de interfaz se utilizó un teléfono móvil con sistema operativo *Android* capaz de vibrar cuando un evento es detectado y mostrar su información gráfica en la pantalla. Así mismo también se implementó la interfaz en un *tablet PC* con el mismo sistema operativo que muestra el icono del sonido detectado sobre el mapa del *Homelab*. Cuando el PC servidor detecta un sonido éste es enviado al *Tablet* o al teléfono móvil para su visualización.

Figura 52 Interfaz de salida del sistema desarrollado sobre PC

7.1.5 Aplicación en el Reconocimiento de Actividades de la Vida Diaria

Parte del sistema de reconocimiento de sonidos fue desarrollado a través del proyecto Europeo FP7 *Robotic UBIquitous COgnitive Network* (RUBICON). En este proyecto, el sistema de reconocimiento de sonidos fue utilizado como un sensor más de la red de sensores planteada para inferir información sobre un escenario *Ambient Assisted Living* (AAL) planteado.

El proyecto RUBICON se fundamenta en una arquitectura de ecología robótica [128]. Las ecologías robóticas son sistemas conformados de varios dispositivos robóticos, incluyendo robots móviles, sensores inalámbricos o dispositivos embebidos en entornos del día a día. El proyecto RUBICON demuestra cómo, dotando a estas ecologías robóticas de algoritmos de procesamiento de la información, tales como percepción, aprendizaje, planificación y detección de nuevos eventos, se puede hacer que estos sistemas sean capaces de entregar soluciones *AAL* modulares, flexibles y seguras.

Los retos clave que persigue este proyecto son:

1. La extracción de significado de datos recogidos que puedan contener ruido y ser imprecisos.

2. El aprendizaje de qué servicio invocar

3. La manera de invocar estos servicios, desde la experiencia más que apoyándose en estrategias de metas predefinidas y estrategias de selección de planes.

El primero de estos retos, y el único que contemplaremos en este apartado, es resuelto mediante la *Capa de Aprendizaje* cuyas bases fueron publicadas en el artículo "*Self-Sustaining Learning for Robotic Ecologies*" presentado en la conferencia *1st International Conference on Sensor Networks* [129]. La capa de aprendizaje procesa flujos de datos obtenidos de todos los sensores para clasificar eventos y hacer predicciones sobre el estado de la ecología y de sus usuarios. La capa de aprendizaje puede ser entrenada, por ejemplo, para predecir actividades que realiza un usuario, tales como saber si está cocinando, mediante el análisis de la señal y el patrón temporal recibido de diversos sensores instalados en la vivienda.

En el escenario AAL se definieron un conjunto de actividades a predecir, todas ellas relacionadas con *Actividades de la Vida Diaria* de la persona, tal y como muestra la Tabla 24, con las que entrenar, validar y testear el sistema.

Actividad	Inputs (Sensores)	Script a desarrollar
Usuario al teléfono	- Sist. Rec. Sonidos - Acelerómetros	El teléfono suena y vibra, el usuario lo coge y habla.
Preparar la comida	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento - Sensore magnéticos	El usuario está en la cocina. Abre cajones para buscar comida, la prepara, la calienta en el microondas, coge platos, vasos y cubiertos, usa el grifo para coger agua,…
Poner la mesa	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento - Sensores magnéticos	El usuario está en la cocina. Pone la mesa, mueve sillas, abre cajones,…
Fregar los platos	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento - Sensores magnéticos	El usuario recoge las cosas de la mesa, las lleva al fregadero y friega.
Salir de casa	- RSSI - Sensores de movimiento - Sensores magnéticos	El usuario sale de la casa.
Dormir	- RSSI - Sensores de movimiento - Sensores de presión - Sensores de luz	El usuario va al dormitorio, baja la persiana, se mete a la cama y apaga las luces.
Relajarse en el sofá	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento - Sensores de presión	El usuario pone música y se sienta en el sofa a leer una revista.
Limpiar	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento - Sensores magenéticos	El usuario mueve sillas para limpiar el suelo, limpia la mesa. Además abre cajones para coger trapos y detergente.
Hacer ejercicio	- Sist. Rec. Sonidos - Acelerómetros - RSSI - Sensores de movimiento	El usuario pone música y realiza ejercicios con la WII (ejem: jugando al tenis).

Tabla 24 Actividades a reconocer en RUBICON

Como se indica en la Tabla 24, el sistema de reconocimiento de sonidos fue utilizado por la capa de aprendizaje como una entrada de datos para alimentar sus modelos en casi todas las actividades a predecir. Esto demuestra la gran información que el medio acústico ofrece de lo que está sucediendo en el entorno. El sonido prolongado del agua

cayendo del grifo aumenta la probabilidad de que el usuario esté fregando. El sonido de platos y vasos aumenta la probabilidad de que el usuario esté preparando la comida y poniendo la mesa, siendo el sonido del microondas funcionando lo que hace que la actividad de preparar comida tomara mayor peso. El ruido de sillas es característica clave de la actividad de limpieza y de preparar la mesa, y el sonido de la música hace que el sistema dé más peso a la posibilidad de que el usuario esté relajándose o haciendo ejercicio, etc.

Una descripción más detallada de este trabajo puede encontrarse en los artículos publicados en las revistas *Engineering Applications of Artificial Intelligence* [130] y *Journal of Intelligent & Robotic Systems* [131].

7.2 Reconocimiento de Sonidos No-Habla en el Hogar sobre Teléfonos Móviles

Una vez implementada la solución bajo plataforma PC utilizando micrófonos profesionales, los algoritmos y técnicas de reconocimiento fueron analizados sobre dispositivos móviles.

Los problemas que introduce el uso de dispositivos móviles son fundamentalmente dos. Por un lado, la velocidad de procesamiento, ya que ésta es en general mucho menor que la disponible en un ordenador. Por otro lado, la calidad de la señal adquirida. En los experimentos descritos en capítulos anteriores los micrófonos utilizados han sido micrófonos profesionales con alimentación *Phantom* utilizados en el campo de la medición acústica con una respuesta en frecuencia muy plana. Los micrófonos embebidos en los dispositivos móviles son, en general, micrófonos de peor calidad, no alimentados a 48 voltios y con una peor sensibilidad.

En este capítulo se analiza el rendimiento del sistema cuando los sonidos son grabados con micrófonos embebidos de teléfonos móviles, añadiendo la complejidad de que estos pueden estar en diferentes ubicaciones. Este experimento demostrará la robustez de los algoritmos, fortaleciendo el estado del arte en este área concreta del procesamiento de señal.

A su vez, al final del capítulo se describe el diseño y desarrollo de una aplicación final de reconocimiento de sonidos para plataforma móvil. Esta aplicación, como se detallará más adelante, ha sido diseñada siguiendo criterios de estándares de accesibilidad y en base a los resultados obtenidos de la encuesta del capítulo 5. La aplicación es capaz de funcionar en tiempo real sobre móviles con sistema operativo Android y ha sido validada con procesadores tanto de gama alta como de gama baja.

7.2.1 Corpus y Análisis de Sonidos

Dentro de la comparativa planteada cinco fueron los micrófonos testeados:

- 4 micrófonos embebidos en móviles
- 1 micrófono profesional con alimentación *Phantom* que servirá como referencia.

En la Tabla 25 se lista todo el equipo incluido en el experimento.

Identificador	Modelo	Tipo
M0	Behringer ECM 8000	Micrófono Profesional
M1	HTC Wildfire	Móvil
M2	LG Nexus 4	Móvil
M3	Samsung Google Nexus S	Móvil
M4	Samsung Galaxy Mini	Móvil

Tabla 25 Micrófonos / dispositivos móviles utilizados en los experimentos

El corpus de sonidos, aunque muy similar al anterior, fue cambiado ligeramente centrándonos más en eventos de interés para el colectivo de personas con discapacidad auditiva y no tanto a su utilización en entornos de reconocimiento de actividades *AAL*. Es por ello que eventos como el ruido de platos o la apertura de la puerta del microondas fueron eliminados y se añadieron otros como el despertador, el llanto de bebé (generado artificialmente) y dos alarmas (de inundación e incendios) que habían tenido un alto grado de aceptación en los resultados de la encuesta de interés (ver Figura 53).

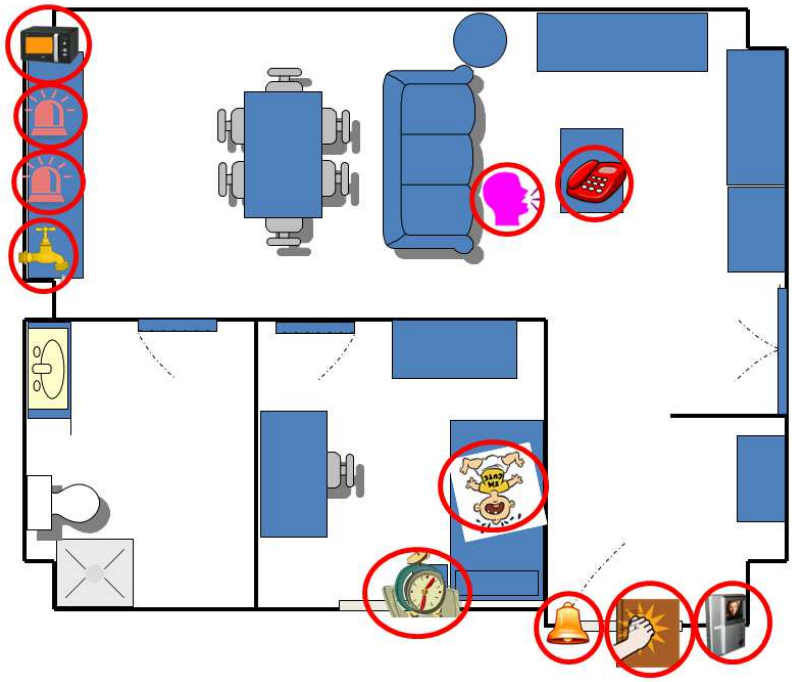


Figura 53 Ubicación de las fuentes de sonido seleccionadas para el análisis con móviles.

Todas las muestras se grabaron utilizando una frecuencia de muestreo de 44.100 Hz con 16 bits de resolución. Estas muestras fueron utilizadas para crear 3 bases de datos: entrenamiento, validación y testeo (*DB_Clean_Train*, *DB_Clean_Val*, *DB_Clean_Test* respectivamente). La diferencia entre estos experimentos y los realizados en apartados anteriores viene marcada principalmente por la libertad de movilidad que ofrecen los teléfonos móviles a los usuarios. En las pruebas anteriores los micrófonos se situaban en el techo en una ubicación fija, y tanto los sonidos que se utilizaban para el entrenamiento como para el testeo eran captados desde esa misma posición. En este experimento la posición de los micrófonos varía entre el entrenamiento y la validación/testeo, queriendo plasmar en la medida de lo posible la situación que se puede dar en la vida real.

Las muestras de la base de datos de entrenamiento *DB_Clean_Train* fueros adquiridas colocando todos los micrófonos sobre una superficie común con una mínima distancia entre ellos (aproximadamente 3 centímetros). Para ganar variabilidad en las muestras, las grabaciones de los sonidos fueron hechas i) manteniendo la superficie común quieta (manteniendo siempre la misma distancia) y ii) moviendo la superficie común más cerca y más lejos de la fuente de audio de forma aleatoria. Con el fin de eliminar diferencias entre tiempos de inicio entre las señales, todas las grabaciones fueron posteriormente manualmente sincronizadas. Para las muestras de entrenamiento, igual que se hizo en experimentos anteriores, cada sonido fue grabado durante un periodo de tiempo de aproximadamente 3 minutos (continuamente presionando el timbre, golpeando la puerta, etc.).

Las muestras de validación y testeo (*DB_Clean_Val* y *DB_Clean_Test*) se grabaron en 55 sesiones (27 y 28 respectivamente). Cada sesión consta de la reproducción de todos los sonidos a reconocer uno detrás del otro con una separación entre sí aproximada de 15 segundos. Por cada entorno del *homelab* (entrada, salón, dormitorio, cocina y baño) se establecieron 3 puntos donde se colocaron los micrófonos. Por cada punto se grabaron aproximadamente 4 sesiones. En este caso, durante todas las grabaciones los micrófonos estuvieron fijos en el punto establecido. La ubicación de los micrófonos en las diferentes sesiones de validación/testeo se muestra en la Figura 54.

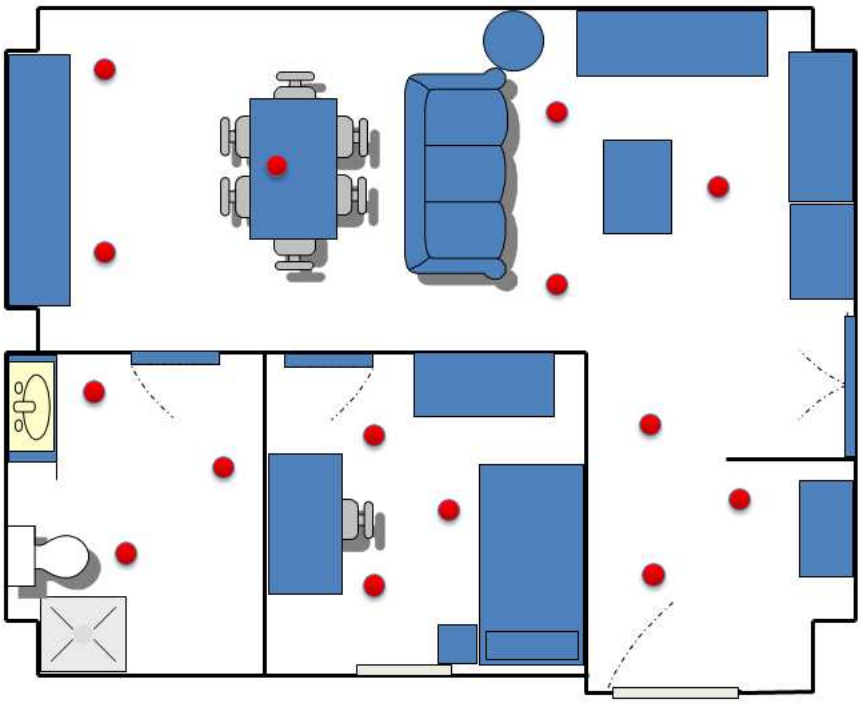


Figura 54 Ubicación de los micrófonos para las BDs de Validación y Testeo.

En la Tabla 26 se muestra la duración aproximada de cada sonido en las sesiones grabadas de validación / testeo.

Sonido	Duración aproximada
Alarma1	3 segundos
Alarma2	3 segundos
Despertador	5 segundos
Golpes en la puerta	2 segundos
Grifo	5 segundos
Habla	5 segundos
Llanto de bebé	5 segundos
Microondas fin	1 pitido (<1 segundo)
Portero automático	2 tonos (3 segundos)
Teléfono	3 tonos (6 segundos)
Timbre Puerta	2 tonos (5 segundos)

Tabla 26 Duración aproximada por cada tipo de sonido en BD de Validación y Testeo

7.2.2 Metodología Aplicada

En el capítulo 6 y sección 7.1 se obtuvieron diferentes combinaciones de parámetros que otorgaron altos ratios de precisión *F1-score* y bajos errores AEER, sin embargo, la inclusión de dispositivos móviles implica la necesidad de una mayor investigación en estas técnicas.

Las limitaciones de procesamiento y capacidad de memoria de los dispositivos móviles hace necesario reducir el número de cálculos computacionales posibles sin que esto

suponga un alto decaimiento de rendimiento en el sistema. En este sentido, algunos parámetros de diseño pueden ser modificados:

- *Número de gaussianas del modelo GMM:* Cuanto más grande sea este valor más cálculos se deberán emplear para calcular el log-likelihood de las tramas analizadas. En estudios iniciales [132] se demostró que el decrecimiento de la precisión del sistema causado por el número de gaussianas en sonidos no-habla es relativamente bajo. Basados en estos estudios un valor final de 3 gaussianas para los modelos GMM fue seleccionado con el que optimizar la relación precisión / cálculo computacional.

- *Solapamiento o desplazamiento entre ventanas:* Es habitual en este tipo de sistemas trabajar con un solapamiento entre ventanas, el cual incrementa el número de muestras para el entrenamiento, validación y testeo, y por tanto incrementa la confianza cuando se decide la clase de un evento acústico. No obstante, esto conlleva un alto coste computacional y, en dispositivos no muy potentes, puede implicar un funcionamiento no deseado de la aplicación final. Es por ello que en los experimentos siguientes no se aplicó solapamiento entre ventanas.

- *Algoritmo de detección:* En experimentos anteriores se ha demostrado cómo la técnica de *detección por clasificación* es la que mejores resultados obtiene. Con combinaciones óptimas en el suavizado de *mínimo número de tramas consecutivas* y *mínimo número de tramas totales* los errores AEER eran reducidos. No obstante, al no aplicar solapamiento entre ventanas estos valores cambian y no hay tanto margen debido a la corta duración de algunos sonidos. Es por esto que en los experimentos siguientes, aunque se utilizó la técnica de *detección por clasificación*, el suavizado fue reducido, utilizándose un valor igual tanto para el *mínimo número de tramas consecutivas* como para el *mínimo número de tramas totales.* De tal forma, un evento es detectado cuando un número mínimo *m* de ventanas consecutivas que pertenecen a la misma clase es encontrado. Este valor se analiza con la base de datos de Validación.

El resto de parámetros no varía, siendo el tamaño de ventana de 70 milisegundos y las características acústicas utilizadas 12 MFCCs, Zero Crossing Rate, Roll-Of Point y Spectral Centroid (con su primera y segunda derivada).

7.2.3 Experimentos y Resultados

7.2.3.1 Evaluación de la Base de Datos de Validación

En primer lugar se crearon los modelos GMM a partir de la base de datos de entrenamiento *DB_Clean_Train*. Por cada micrófono se creó un modelo por cada clase de sonido. Para todas las clases se extrajeron las características acústicas de exactamente 3 minutos de audio, utilizando tal y como se ha indicado anteriormente 3 gaussianas para cada modelo.

En segundo lugar, para la evaluación se hizo uso de la base de datos de validación *DB_Clean_Val*. Cada micrófono fue evaluado con sus propios datos (utilizando los modelos creados con sus datos a partir de la base de datos de entrenamiento y validándolos con sus propios datos a partir de la base de datos de validación). Esto simula el comportamiento esperado de la aplicación final donde el usuario entrenará sus propios sonidos con su móvil que será el que más tarde también utilice para reconocerlos. Se realizó una evaluación variando el valor del *número mínimo de tramas m*. Los valores de AEER se muestran en la gráfica de la Figura 55.

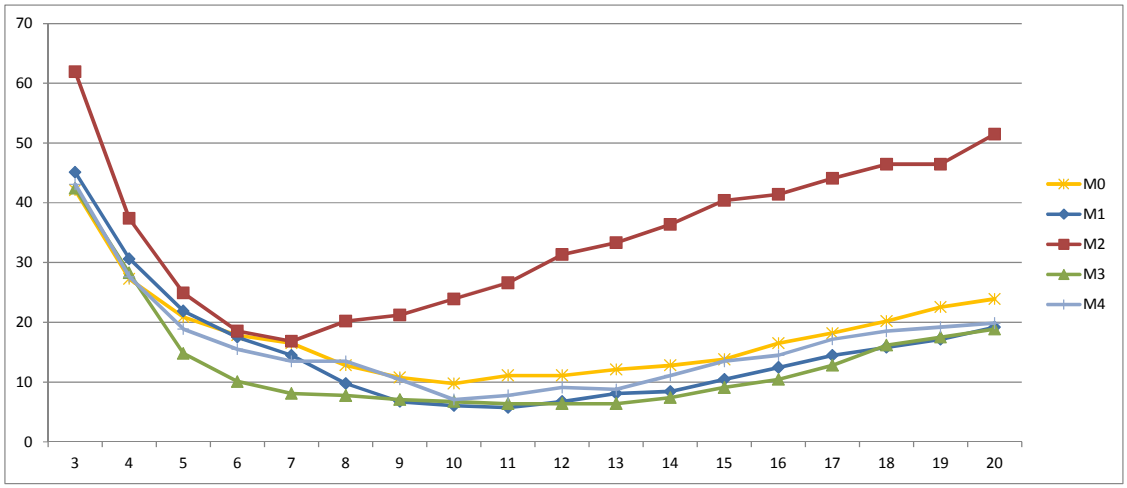


Figura 55 Valor AEER por número mínimo de tramas *m* por cada micrófono.

Los resultados demuestran que, a excepción del micrófono del teléfono móvil M2, el resto de móviles obtienen valores muy similares a los del micrófono profesional (M0). Aunque los micrófonos de los teléfonos móviles están construidos para comunicaciones de voz, primando el rango de frecuencias del habla, su respuesta en frecuencias no es condicionante para los algoritmos elegidos de reconocimiento de sonidos no-habla.

La Figura 56 representa separadamente los diferentes tipos de errores, promediados entre los micrófonos M1 y M4. Se muestra la media de Eventos Eliminados (D), Eventos Nuevos (I) y Eventos Sustituidos (S) extraídos de la fórmula AEER para todos los móviles. Lo más significativo de este análisis es que las inserciones de nuevos

eventos permanecen estables en relación al incremento del mínimo número de tramas consecutivas *m* mientras que los eventos sustituidos decrecen y los eventos eliminados incrementan. Esto significa que la mayoría de los errores generados cuando el número mínimo de tramas *m* es bajo, son generados dentro del propio evento acústico, siendo el evento parcialmente sustituido (Evento Nuevo) o totalmente sustituido (Evento Sustituido) por otro (ver Figura 57). Esto puede ser explicado ateniéndonos a la naturaleza de los sonidos, sus fases de ataque, mantenimiento y decaimiento, y a la forma de entrenamiento de las GMMs. Es en las fases de ataque y decaimiento cuando el "núcleo" del sonido no está enteramente definido y las frecuencias son más fluctuantes. Esto implica que en estas dos fases la clasificación pueda provocar errores graves de sustitución.

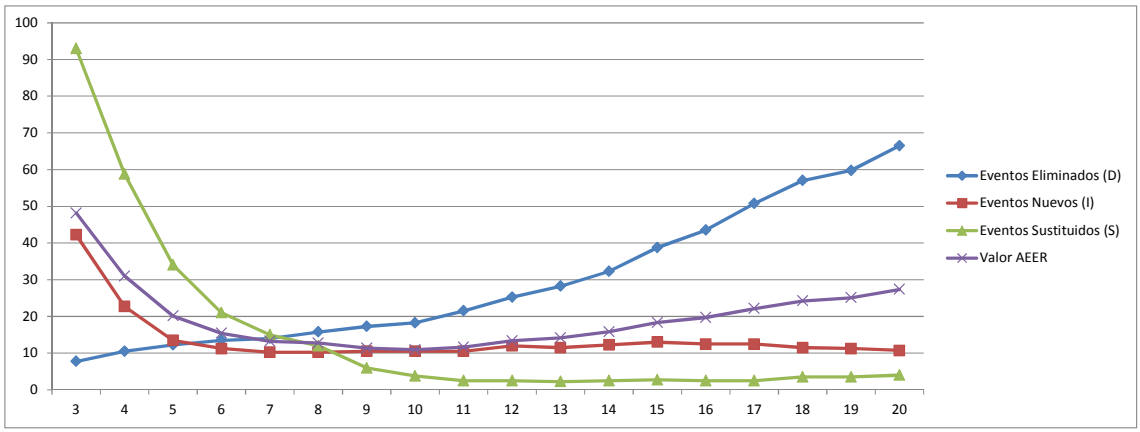


Figura 56 Media ponderada de los teléfonos móviles sobre eventos eliminados, nuevos, sustituidos y valor de AEER por número mínimo de tramas *m*.

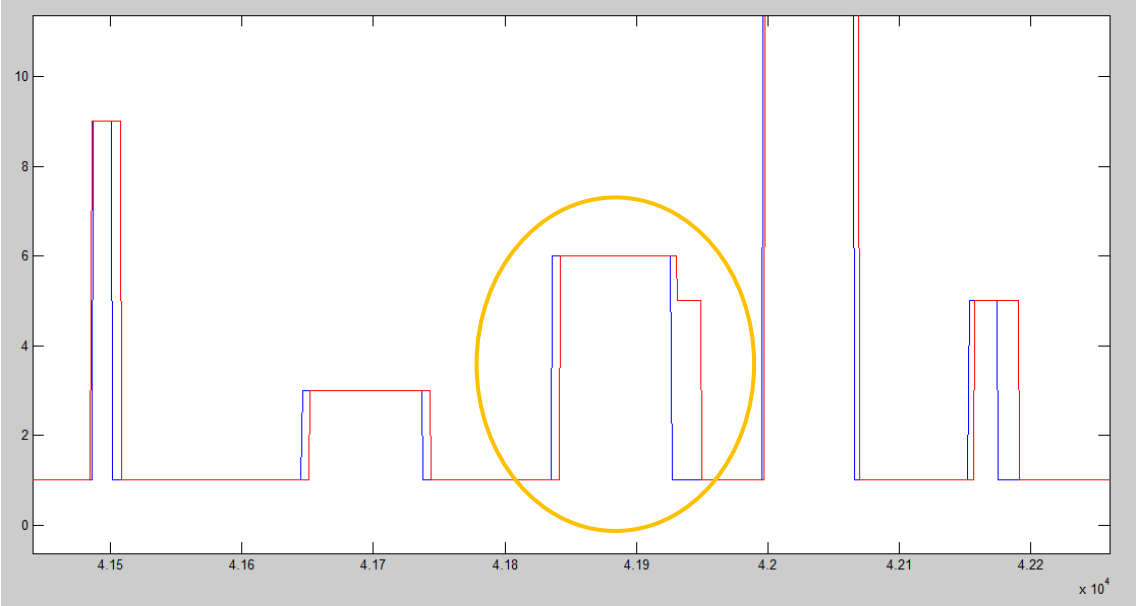


Figura 57 Ejemplo de fallo por sustitución al final de un evento.

Para intentar paliar esta problemática los modelos de entrenamiento fueron modificados. De los 3 minutos antes utilizados para entrenar los eventos acústicos, sólo se tomaron aquellas tramas cuyo valor medio de energía superase un umbral del 110% por encima del valor medio de 5 segundos de ruido ambiente. La ecuación 16 expresa el cálculo de este umbral.

$$Theshold = \frac{\sum_{t=1}^{T} STE(t)}{T} * \frac{110}{100} \quad (16)$$

siendo t el índice de trama, T el número de tramas correspondientes a 5 segundos y STE el valor del parámetro *Short Time Energy* definido como:

$$STE = \frac{\sum_{n=1}^{N} x(n)x(n)}{N} \quad (17)$$

donde N corresponde al número de muestras de una trama.

El resultado de aplicar esta medida en el entrenamiento se muestra en la Figura 58, donde se indica la mejora del valor de AEER con respecto a la evaluación anterior para la media de todos los móviles. El mejor valor pasa de un error de 10,9 a 6,5, siendo para ambos el mejor *número mínimo de tramas m* igual a 10.

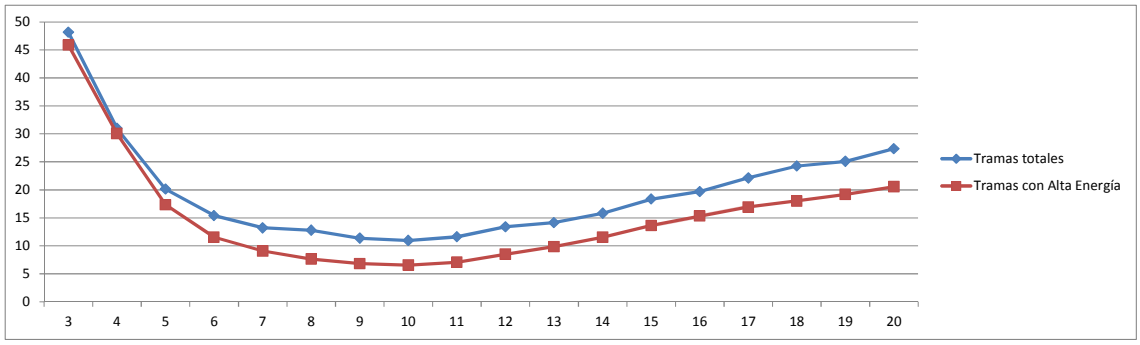


Figura 58 Comparativa de valor AEER sobre entrenamiento con todas las tramas y tramas de alta energía.

En la gráfica de la Figura 59 se muestran los eventos eliminados, insertados y sustituidos con esta mejora para un *número mínimo de m* igual a 10, para todas las clases.

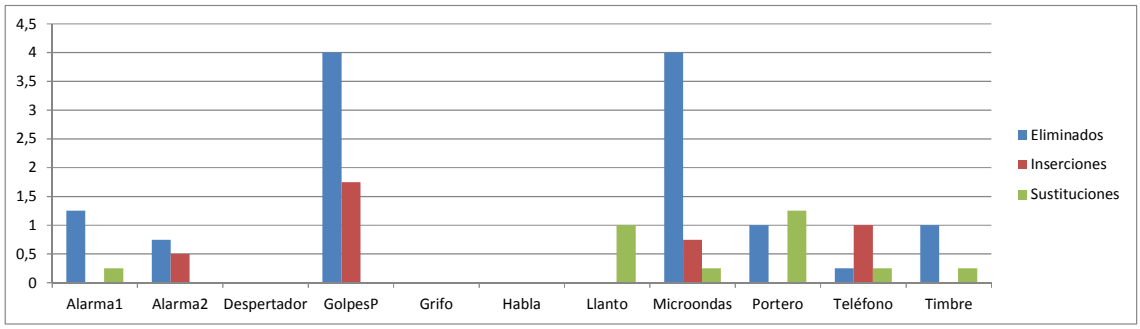


Figura 59 Media de eventos eliminados, insertados y sustituidos por todos los teléfonos móviles para cada clase.

7.2.3.2 Evaluación de la Base de datos frente al Ruido

Hasta ahora, las bases de datos evaluadas no han tenido alteraciones entre los datos de entrenamiento y los datos de validación y/o testeo. Las condiciones de grabación han sido las mismas, no llegándose a estudiar la problemática que pueda surgir cuando los datos de validación y/o testeo se ven modificados por condiciones de ruido. En el experimento siguiente se quiso analizar este condicionante, variando los datos de las bases de datos de validación y testeo.

Se extrajeron 5 archivos de audio de la base de datos PacDV disponible a través de Internet (*www.pacdv.com*). Estos archivos son ficheros de ruido ambiente de diferentes situaciones. En la Tabla 27 se muestra la descripción y el espectrograma de los mismos.

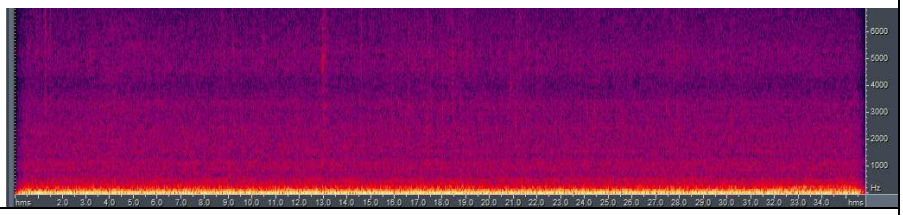
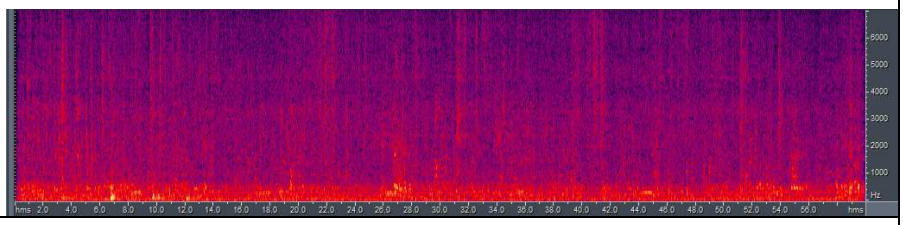
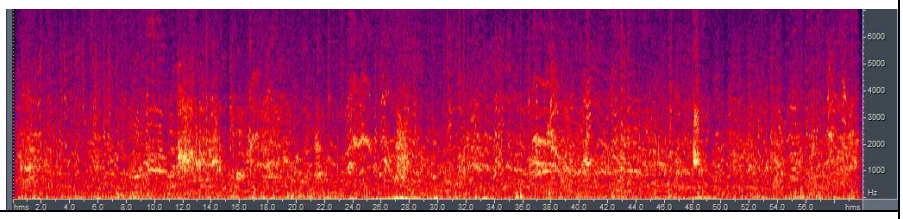
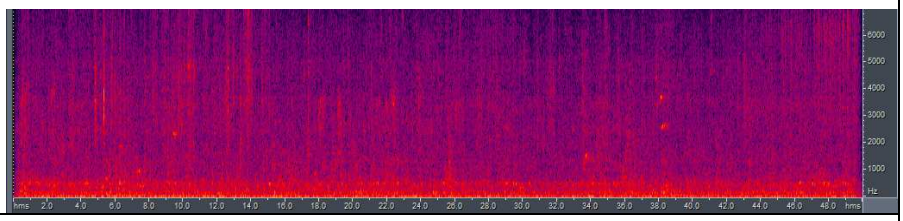
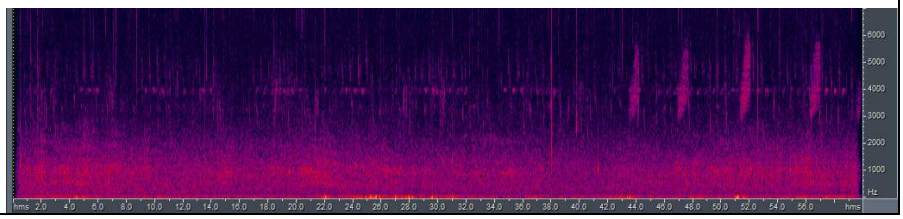
Id	Descripción	Espectrograma
R1	Interior autobús (35”)	
R2	Comedor (59”)	
R3	Niños jugando (59”)	
R4	Gente reunida (50”)	
R5	Naturaleza (59”)	

Tabla 27 Descripción y espectrograma de las señales de ruido utilizadas para mezclar

Para todas las sesiones de validación y testeo, estos ficheros fueron mezclados a un nivel *Signal Noise Ratio* (SNR) de valor 15 dB. Así, se creó una nueva base de datos de validación y testeo con diferentes ruidos de fondo denominadas *DB_Noise_Val* y *DB_Noise_Test*.

Un nuevo experimento de validación fue realizado aplicando el umbral de intensidad en el entrenamiento, usando los datos originales para el entrenamiento (*DB_Clean_Train*) y la base de datos de validación mezclada con ruidos de fondo (*DB_Noise_Val*) para su evaluación. Los resultados de los nuevos valores de AEER son mostrados en la Figura 60.

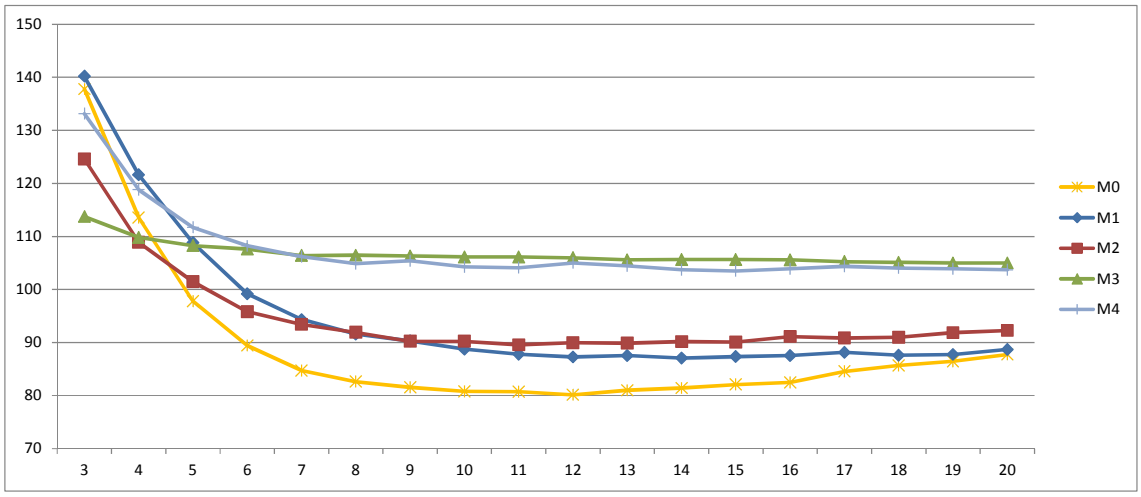


Figura 60 Valor AEER por número mínimo de tramas *m* por cada micrófono sobre *DB_Noise_Val*.

El empeoramiento de los resultados es considerablemente alto. Aplicando la media AEER de todos los móviles, el mejor valor medio obtiene un AEER de 96,9 cuando previamente estaba situado en 6,5. Si analizamos los resultados por tipos de ruidos mezclados (ver gráfica de Figura 61) observamos cómo es R1 (interior bus) el que da un menor error, estando el resto de mezclas muy equiparadas. Si analizamos los espectrogramas de la Tabla 27, el correspondiente de R1 no posee tanta variabilidad en frecuencias como el resto, encontrándose en los espectrogramas de R2, R3, R4 y R5 ruidos de golpes en mesas, sillas, cubiertos, gritos, habla, sonido de pájaros,… Esto puede explicar las diferencias de resultados entre los datos mezclados con R1 y con el resto de ruidos.

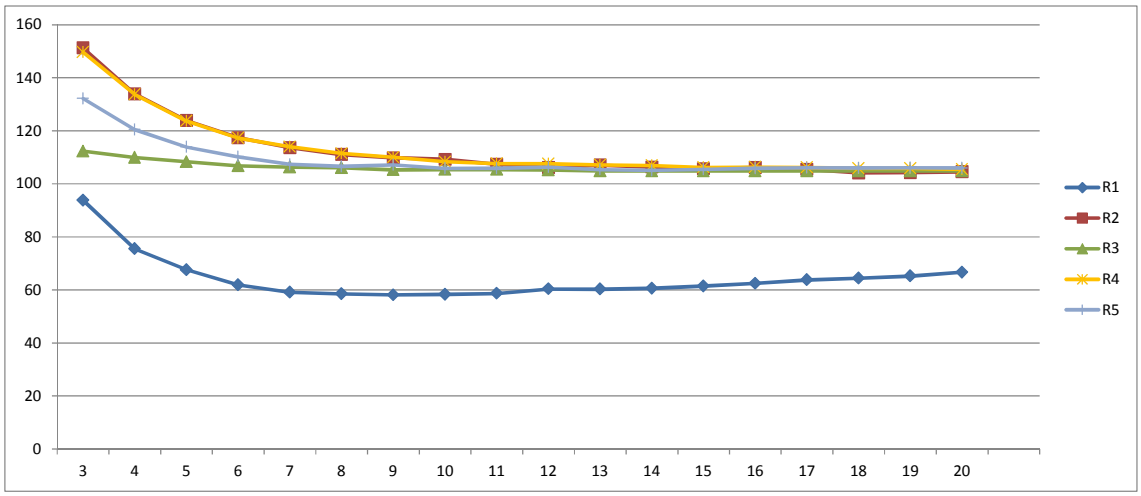


Figura 61 Valor AEER por tipos de ruidos (Valor medio entre teléfonos móviles) sobre *DB_Noise_Val*.

7.2.3.3 Análisis de la Robustez de los Parámetros Acústicos

En esta sección se analiza el decaimiento de la precisión del sistema debido al ruido de fondo en la base de datos de validación. Para ello se compararon los valores de todos

los parámetros acústicos entre las dos bases de datos de validación (*DB_Clean_Val* y *DB_Noise_Val*) y se observaron diferencias entre ellos.

Aunque los valores de los parámetros MFCC fluctúan entre las bases de datos *DB_Clean_Val* y *DB_Noise_Val*, (ver como ejemplo la distribución de las muestras de MFCC1 en la Figura 62 para la clase no-evento y el ruido R1), esta fluctuación no es tan alta como la observable en los parámetros ZCR, Roll-Off Point y Centroid (ver como ejemplo la distribución de las muestras de Roll-Off Point en la Figura 63 para la clase no-evento y el ruido R1).

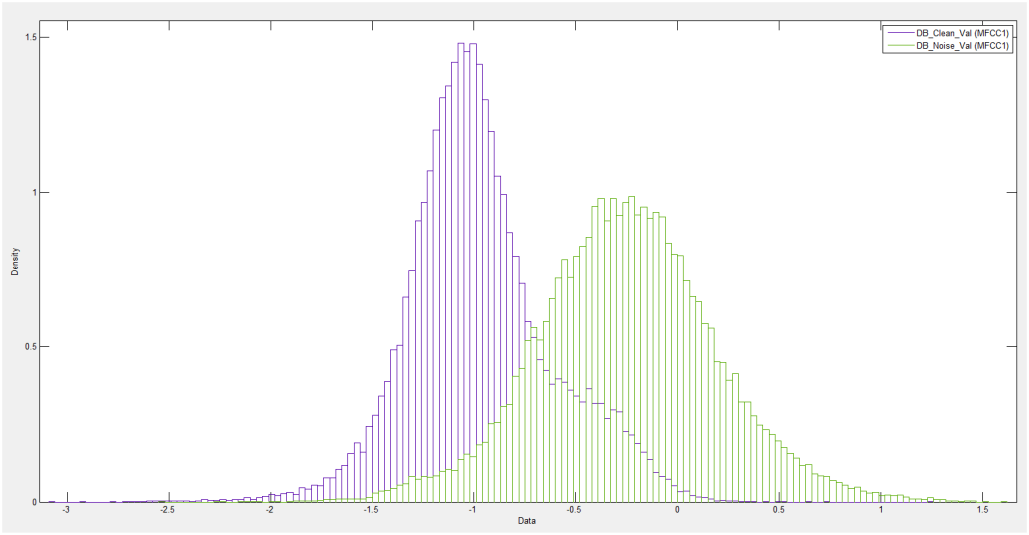


Figura 62 Distribución de MFCC1 para las bases de datos *DB_Clean_Val* y *DB_Noise_Val*

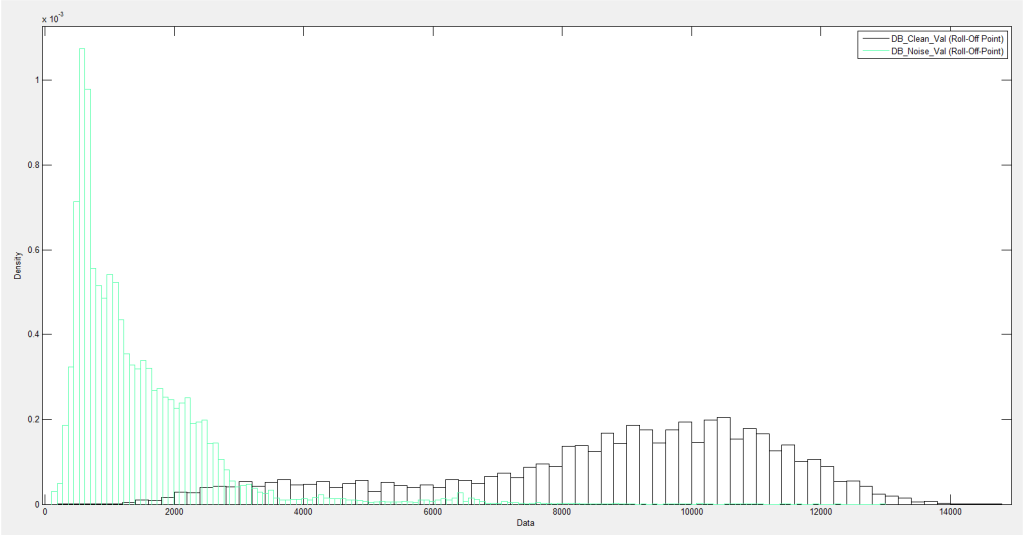


Figura 63 Distribución de Roll-Off Point para las bases de datos *DB_Clean_Val* y *DB_Noise_Val*

Los tres parámetros con la mayor fluctuación (ZCR, Roll-Off Point y Centroid) fueron eliminados del vector de parámetros utilizado y se procedió a evaluar de nuevo los experimentos. El resultado se muestra en las gráficas de la Figura 64 y Figura 65. La eliminación de los parámetros ZCR, Roll-Of Point y Centroid hace que el valor AEER se

reduzca, siendo anteriormente el mejor error medio 96,6 y después de la eliminación de estos parámetros bajó a 71,9. No obstante, este error todavía queda lejos de ser aceptable.

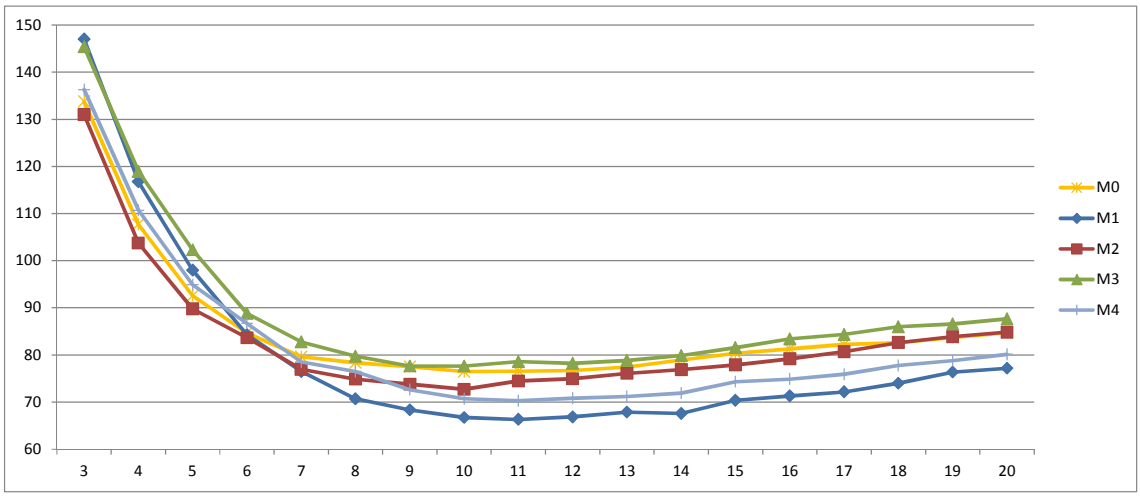


Figura 64 Valor AEER por número mínimo de tramas *m* por cada micrófono sobre *DB_Noise_Val*. Eliminados los parámetros ZCR, RF y Centroid.

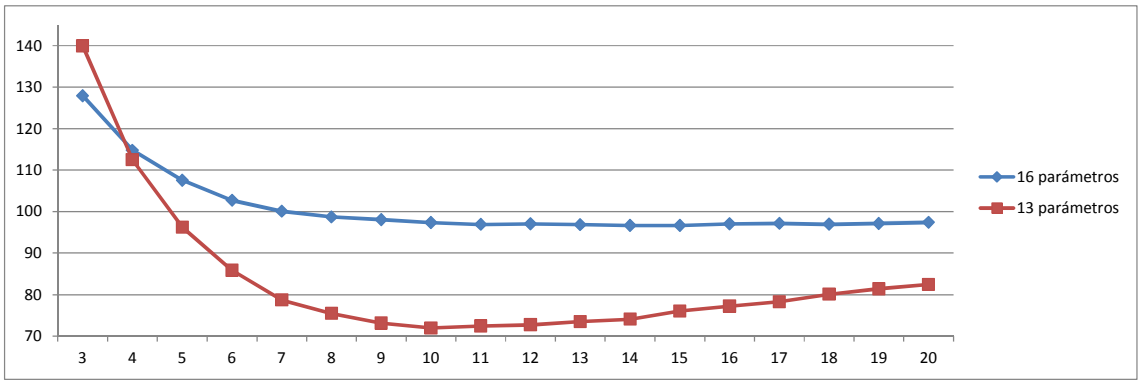


Figura 65 Comparativa de Valor AEER sobre *DB_Noise_Val* con y sin parámetros ZCR, RF y Centroid. Valor AEER calculado con la media de los teléfonos móviles.

7.2.3.4 Normalización de Canal

En el reconocimiento del habla es común encontrarse con grabaciones producidas en diferentes medios o canales. Para trabajar con todas ellas diferentes trabajos [133] utilizan el *Cepstral Mean Normalization / Substraction*.

Esta técnica normaliza los coeficientes MFCC de forma que sus valores no sean afectados por el cambio de canal.

En el siguiente experimento se aplicó esta técnica a las bases de datos de la siguiente manera:

1) *En el entrenamiento:* (*DB_Clean_Train*)

    a. Se calculó la media de cada coeficiente de la clase “no evento”.

b. Para todas las clases, por cada coeficiente, se resta la media obtenida a su valor.

c. Los modelos GMM son creados con estos nuevos datos

2) *En la validación:* (*DB_Noise_Val*)

a. Por cada experimento se obtienen los primeros 10 segundos de tramas "no-evento" en base a un nivel de intensidad y se calcula la media de cada coeficiente.

b. Para toda trama, por cada coeficiente analizado se resta la media obtenida a su valor.

c. El vector resultante de coeficientes fue el que se utilizó para clasificar la trama.

Los resultados de aplicar esta técnica se muestran en las gráficas de la Figura 66 y de la Figura 67.

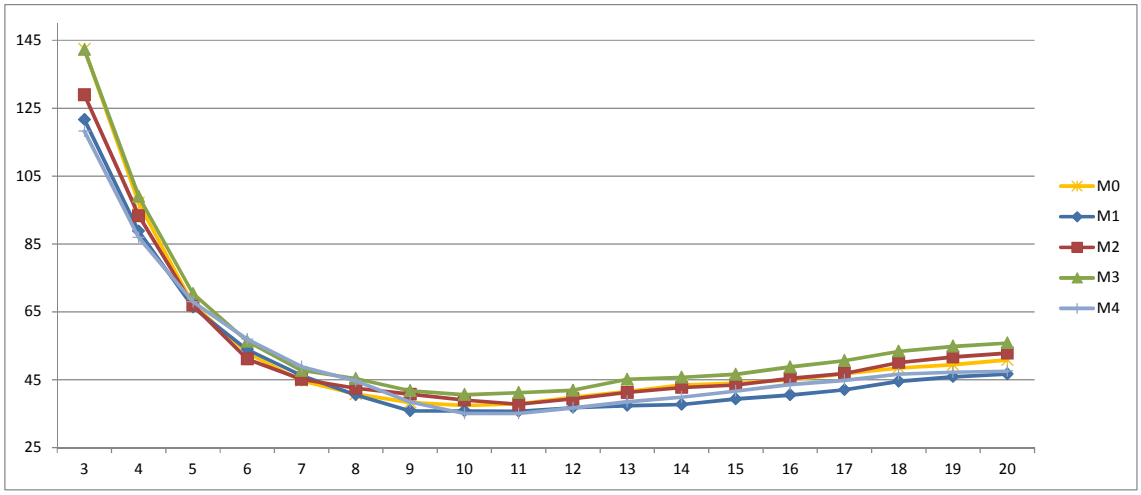


Figura 66 Valor AEER por número mínimo de tramas *m* por cada micrófono sobre *DB_Noise_Val*. Aplicando Cepstral Mean Normalization.

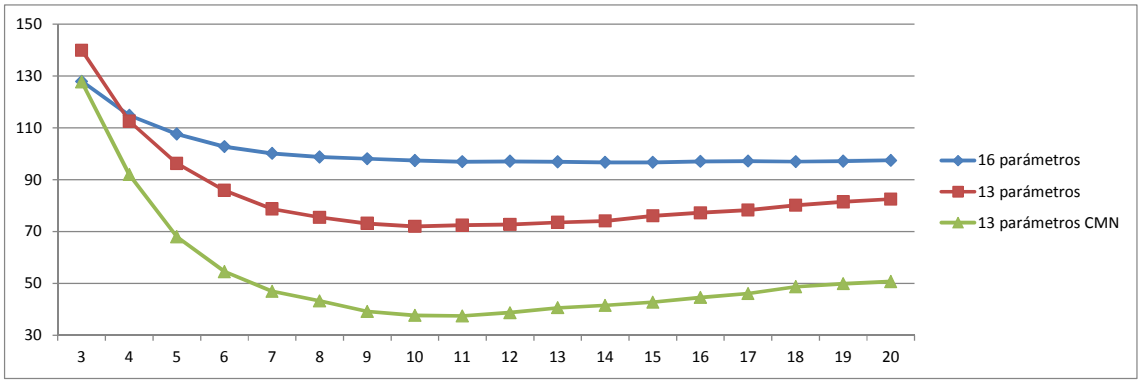


Figura 67 Comparativa de Valor AEER sobre *DB_Noise_Val* añadiendo CMN. Valor AEER calculado con la media de los teléfonos móviles.

Las gráficas demuestran la mejora que se obtiene con la técnica *Cepstral Mean Normalization / Substraction*. El error más bajo aplicando la media sobre todos los teléfonos móviles fue de 37,4. Aunque esta cifra no es comparable a los valores obtenidos con las señales originales mejora más del doble al valor inicial obtenido (96,6).

La técnica CMN fue a su vez evaluada con las bases de datos originales (*DB_Clean_Val*). Como se muestra en la Figura 68, con los datos sin modificar, la combinación de 16 parámetros (MFCC, ZCR, Roll-Off Point, Centroid) sin CMN obtiene peores resultados que la combinación de 13 parámetros MFCC normalizados con CMN (6,5 frente a 4,5 en el menor error AEER). Esto demuestra la mejora que esta técnica implica en los sistemas de reconocimiento no-habla.

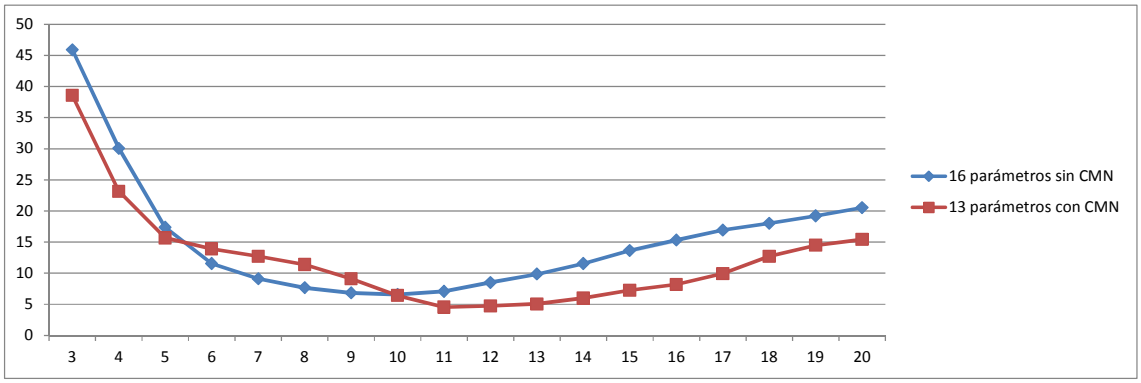


Figura 68 Comparativa de Valor AEER sobre *DB_Clean_Val* con y sin CMN. Valor AEER calculado con la media de los teléfonos móviles.

7.2.3.5 Reducción del Tiempo de Entrenamiento

El número de tramas utilizadas en la etapa de entrenamiento de los sistemas de reconocimiento del habla determinará su rendimiento. En una aplicación final, si el usuario tiene que grabar su propia base de datos, reducir el tiempo de entrenamiento implica una mejora en funcionalidad. Grabar un timbre de la puerta, un portero automático o una alarma de incendios no es sólo molesto para el usuario y para aquellos a su alrededor, sino que es también una tarea complicada en muchos casos. La aceptación de esta tecnología varía dependiendo de factores determinantes como el tiempo de entrenamiento. En los siguientes experimentos se evaluó la variación del error AEER con respecto al tiempo de entrenamiento (o número de tramas utilizadas para la creación del modelo).

Los experimentos fueron realizados reduciendo el tiempo de entrenamiento por intervalos de 10 segundos. El mínimo número de tramas consecutivas *m* fue fijado al óptimo valor encontrado previamente (10 sin CMN y 11 con CMN). Los resultados para la validación con la base de datos original *DB_Clean_Val* se muestran en la gráfica de la Figura 69. En la gráfica de la Figura 70 se muestra el resultado de la validación con la base de datos mezclada *DB_Noise_Val*.

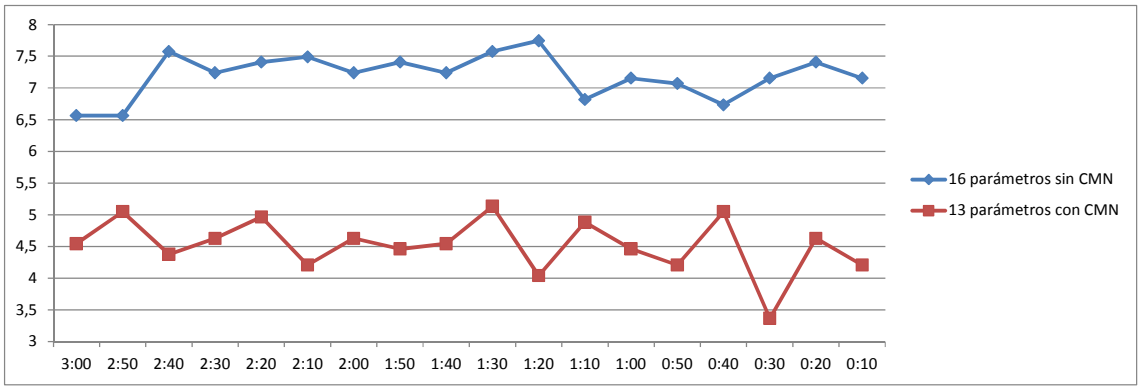


Figura 69 Comparativa de Valor AEER sobre *DB_Clean_Val* por tiempo de entrenamiento. Valor AEER calculado con la media de los teléfonos móviles.

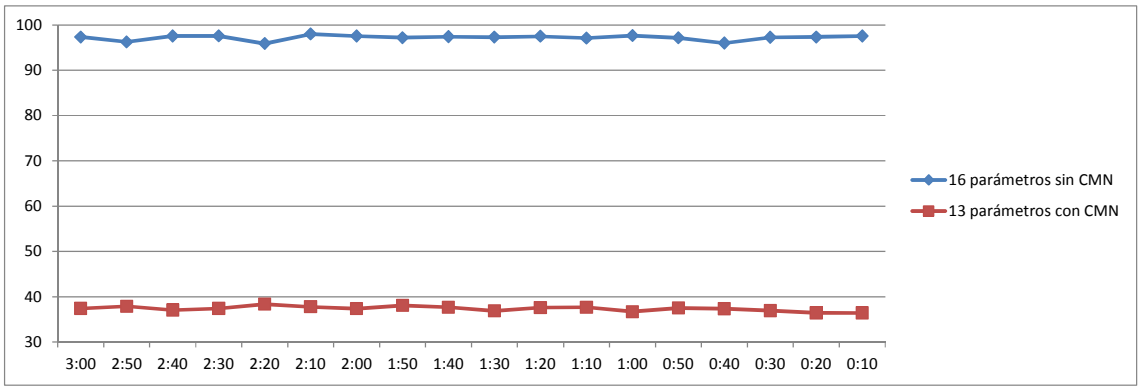


Figura 70 Comparativa de Valor AEER sobre *DB_Noise_Val* por tiempo de entrenamiento. Valor AEER calculado con la media de los teléfonos móviles.

En ambas gráficas la disminución del tiempo de entrenamiento no implica un aumento del error AEER. Esto significa que no es necesario grabar los eventos acústicos a reconocer durante un tiempo de entrenamiento prolongado. Con tan solo unos pocos segundos los modelos GMM pueden ser creados definiendo la distribución de las características acústicas al mismo nivel que si éste fuese aumentado.

7.2.3.6 Evaluación Final del Sistema

La Tabla 28 muestra los resultados de los experimentos realizados usando la base de datos de testeo con y sin ruido añadido, así como aplicando y no aplicando la normalización de parámetros. Con el objetivo de comparar con los experimentos previos, Tabla 28 muestra a su vez los resultados sobre la base de datos de validación. El número mínimo de tramas consecutivas *m* fue fijado a 10 y 11 (el mejor valor observado sin y con CMN respectivamente). El tiempo de entrenamiento se estableció en 10 segundos.

Como se esperaba, el rendimiento de ambas bases de datos es muy similar ya que fueron grabadas en condiciones muy similares.

	16 parámetros sin CMN	13 parámetros con CMN
DB_Clean_Val	7,15	4,20
DB_Clean_Test	4,87	5,92
DB_Noise_Val	97,52	36,41
DB_Noise_Test	97,38	35,61

Tabla 28 Resultados finales con la base de datos de Validación y Testeo. Valor AEER calculado con la media de los teléfonos móviles.

7.2.4 Sistema Desarrollado: myEardroid

El principal objetivo de este trabajo ha sido proveer a la comunidad de personas con problemas auditivos de una aplicación útil que puedan usar en su vida diaria. Como se demostró en el capítulo 5, la opción más aceptada es el uso de teléfonos móviles. Las personas con problemas auditivos hacen un uso extenso de estos dispositivos y los escogieron en la encuesta como plataforma preferida en la que implementar estos avances tecnológicos.

Se eligió *Android* como sistema operativo debido a ser el sistema con más usuarios del momento tanto a nivel nacional como internacional. Además, *Android* da la posibilidad a los desarrolladores de programar el código utilizando tanto Java como sus librerías nativas escritas en C++. Esto permite el desarrollo de aplicaciones más eficientes y con un mayor rendimiento en el procesado.

La aplicación detecta y clasifica sonidos del hogar seleccionados por el usuario (timbre, teléfono, alarma de incendios, etc.). El uso de una base de datos genérica o estándar que contenga los modelos de referencia fue descartado debido a la naturaleza de los sonidos a clasificar. Por ejemplo, el número de timbres diferentes es inmenso, y no sería práctico crear una base de datos que contenga todos ellos. Este el motivo por el cual, el usuario entrenará sus propios modelos con los sonidos seleccionados.

La consecución de este desarrollo dio lugar al lanzamiento de una primera versión *Beta* de la aplicación en *GooglePlay*. Además el proyecto fue presentado a la **VIII Edición del Premio Vodafone a la Innovación en Telecomunicaciones**, alzándose éste vencedor en la categoría "***Premio al desarrollo de Aplicación Mobile for Good***". Cabe mencionar que este premio se concede a una aplicación que contribuya a la mejora de la calidad de vida, la autonomía personal y la promoción del envejecimiento activo y la solidaridad intergeneracional.

En los siguientes párrafos se describe la arquitectura y funcionalidad de esta primera versión "Beta" de la aplicación.

7.2.4.1 Diseño Orientado a las Personas

Dentro de la etapa de diseño se aplicó la metodología o *técnica de Personas*. La *técnica de Personas* sintetiza los datos disponibles sobre los usuarios previstos del sistema en unos usuarios arquetipos. Es una técnica de uso habitual por expertos en usabilidad y se aplica cuando el número de usuarios es demasiado extenso. Los usuarios ficticios resultado de aplicar esta técnica son representativos de cómo es un usuario. En este sentido hubo ciertas premisas que se contemplaron a la hora de definir las pantallas y funcionalidad de la aplicación:

Perfil del usuario

Tal y como se expone en el libro 'The essentials of interaction design' [134], la mayoría de las personas no son ni principiantes ni expertos cuando se enfrentan a una nueva herramienta. La razón de este hecho es que en la mayoría de los casos disponen de modelos mentales para enfrentarse a ella, lo que les convierte en usuarios intermedios. Al mismo tiempo, no pueden considerarse usuarios expertos. Hay que tener en cuenta que en la mayoría de las ocasiones alcanzar el nivel de experto conlleva la necesidad de ir más allá del uso estándar, por lo que no son muchos los usuarios que dan el salto.

Por norma, haciendo una lectura generalista, el usuario es un ***intermedio perpetuo***. Ésta es la razón por la que se definió la aplicación desde el punto de vista de un usuario intermedio, huyendo de opciones avanzadas que complicasen el desarrollo y la interacción con la aplicación. No obstante, los usuarios principiantes también han sido considerados a través de la implementación de tutoriales de navegación al ejecutarse por primera vez la aplicación, y de pantallas simplificadas e intuitivas.

Internacionalización

Aunque en la versión inicial de la aplicación sólo se contemplaron el castellano y el inglés como idiomas, posteriormente se añadió el italiano por petición de varias asociaciones. Además, gracias a la simplicidad que *Android* provee a la hora de desarrollar aplicaciones multilenguaje, la arquitectura está preparada para una rápida inclusión de más idiomas.

Accesibilidad cognitiva de las personas con deficiencias auditivas

Este aspecto parte de las conclusiones del estudio realizado por el Departamento de Psicología Experimental de la Facultad de Psicología de la Universidad de Granada [135], centrado en la comunicación de personas sordas pre-locutivas. La conclusión más relevante de este estudio se cita a continuación:

"el déficit auditivo de los sordos no sólo les dificulta o impide la compresión del habla en la comunicación oral, sino que les conduce a un notable déficit en el procesamiento

de cualquier información verbal, sea oral o escrita. Este hecho les supone un handicap a la hora de interpretar el lenguaje escrito que […] está fuertemente implicado en la comunicación vía Internet."

A continuación se extraen del estudio mencionado los puntos más relevantes a la hora de diseñar la aplicación:

- *Tarea de búsqueda visual:* se recomienda el uso de iconografía para la identificación de las tareas, agrupaciones lógicas, menús, etc. y acompañarlas de texto. Este aspecto también quedó destacado en el capítulo 5 de esta tesis.
- *Tarea de atención dividida:* se recomienda reducir la cantidad de información o el número de tareas *visioespaciales* por página.
- *Tarea de navegación:* las personas con problemas auditivos son mejores a la hora de navegar gracias a que sus habilidades *visioespaciales* están más desarrolladas. El estudio propone reducir la información en pantalla aumentando el número de pantallas y de interacciones.
- *Tareas de lectura y comprensión de textos:* se debe utilizar lenguaje directo, expresiones cortas, lenguaje familiar, resaltar aspectos claves y evitar frases negativas.

Estos puntos guiaron el diseño de la aplicación aun siendo conscientes de que no todas las personas que se definan corresponderán al perfil de sordos pre-locutivos. Se entiende por tanto que ningún usuario con sordera post-locutiva tendrá dificultad añadida por la toma de esta decisión.

Edad de los usuarios

El público adolescente se identifica como usuario mayoritario de la aplicación, por ser éste el que está más familiarizado con *smartphones* y más interesado en el ámbito de las nuevas tecnologías. No obstante, la aplicación puede ser utilizada por cualquier persona que disponga de un dispositivo con sistema operativo *Android*. Por este motivo y por los aspectos expuestos en el apartado anterior como: uso de iconografía, reducción de la información por pantalla, etc, se optó por una navegación lineal acompañada siempre de mensajes intuitivos durante el proceso de configuración y entrenamiento. De esta forma se facilita la cumplimentación de las tareas y la navegación para usuarios infantiles o, sobre todo, de avanzada edad. La gente mayor empieza a tener más contacto con estas tecnologías y, sus problemas de sordera derivados de la edad hacen de ellos un colectivo con un potencial enorme en el uso de esta aplicación.

7.2.4.2 Navegación y Diseño de Pantallas

El diseño final de las pantallas viene marcado por las pautas de accesibilidad antes mencionadas y los resultados de la encuesta del capítulo 5. Además, éstas siguen también las recomendaciones de diseño de Google para sistemas *Android* orientadas a una mayor usabilidad y accesibilidad de las personas.

Diseño de marca

Para el diseño de la marca se recopilaron imágenes relacionadas con las deficiencias auditivas, seleccionándose como la más representativa el audífono. A partir de diferentes fuentes y referencias fotográficas de audífonos, se desarrolló el diseño del logotipo como se ilustra en la Figura 71.

Figura 71 Propuesta de logotipo de la aplicación myEardroid.

Pantalla de primeros pasos

La primera vez que se inicia la aplicación *myEardroid*, se muestra una guía de uso constituida de 5 pasos. Esta ayuda se podrá seguir consultando más adelante tocando sobre el botón de ayuda que contiene como icono descriptivo un símbolo de interrogante.

Pantalla principal

Esta pantalla es la interfaz de inicio de la aplicación (excepto la primera vez de ejecución que será la pantalla de "primeros pasos" la que se muestre). Los elementos específicos de esta pantalla se muestran a continuación (ver ejemplo en Figura 72):

Sonido ambiente. Éste es un elemento fundamental para el buen funcionamiento de la aplicación. Dentro de la etapa de detección es importante contar con el modelo de "*ruido ambiente*" más característico de cada entorno. Este modelo cambiará si estamos reconociendo sonidos en una casa en el campo o si la aplicación es usada en un piso céntrico de la ciudad pegado a la autopista. Para obtener una mayor precisión en los algoritmos de reconocimiento el usuario debe obligatoriamente grabar el sonido ambiente de su casa. Éste es el único sonido que debe ser entrenado obligatoriamente. Los elementos que describen su diseño se muestran a continuación (comenzando desde la izquierda):

- *Icono representativo*.
- *Nombre del sonido*: "*Ambiente*".
- *Icono de Entrenamiento*: da acceso al entrenamiento del ruido ambiente. Puede mostrarse en dos estados diferentes que indican si el sonido ha sido entrenado o si el sonido está pendiente de entrenar (aparecerá el icono con una exclamación).

Listado de sonidos del hogar. Enumeración de los sonidos definidos para el entorno del hogar. En la versión *Beta* desarrollada los sonidos son fijos por defecto. A futuro se valorará tanto la posibilidad de personalizar los alias de los sonidos, como dar de alta nuevos sonidos. En este caso la navegación y capacidades serán una extensión de lo presentado en este apartado. Los elementos que componen cada registro del listado de sonidos son los siguientes (comenzando desde la izquierda):

- *CheckBox*: para identificar los sonidos activos dentro del entorno. Los sonidos marcados con el *check* serán los que actualmente se estén "escuchando" cuando la aplicación empiece a reconocer.
- *Nombre del sonido*: en la versión beta se han asignado por defecto 5 sonidos: teléfono, timbre, alarma, despertador y portero automático.
- *Icono de Entrenamiento*: da acceso al entrenamiento del ruido ambiente. Puede mostrarse en dos estados diferentes que indican si el sonido ha sido entrenado o si el sonido está pendiente de entrenar (aparecerá el icono con una exclamación).

Botón de Empezar. Al pulsar este botón la aplicación comenzará a analizar el audio proveniente del micrófono para el reconocimiento de los eventos acústicos a través de los algoritmos anteriormente comentados.

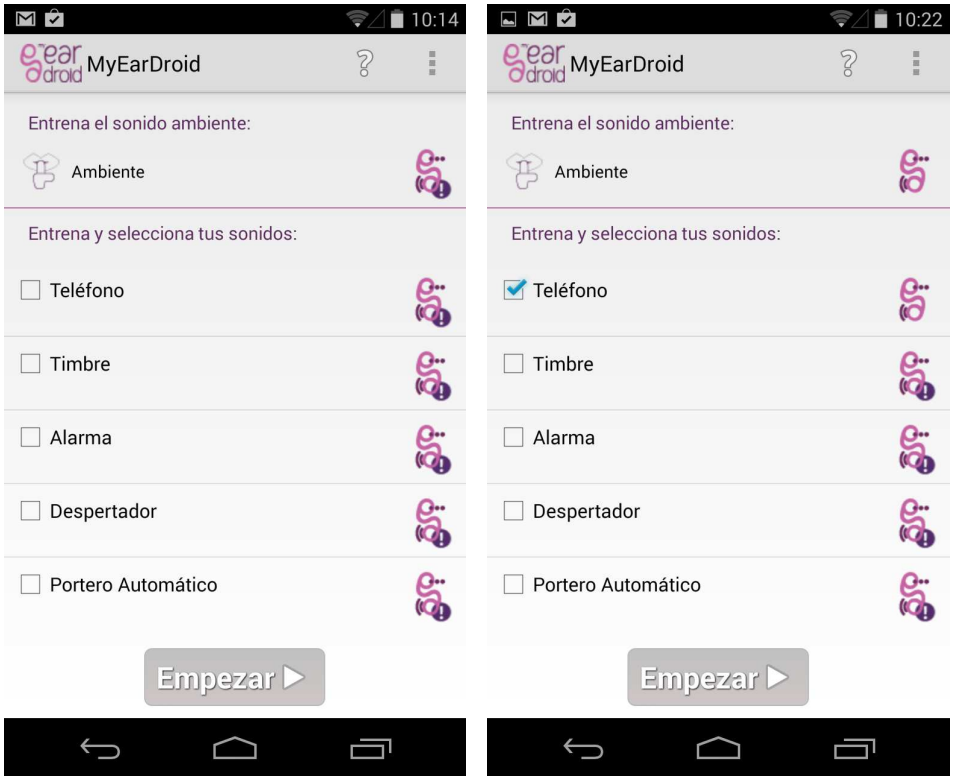


Figura 72 Pantalla principal de la aplicación myEardroid.

Pantalla de entrenamiento

Si el usuario accede a esta pantalla a través del *Icono de Entrenamiento*, éste podrá grabar sus sonidos. La aplicación utilizará el audio grabado para crear los modelos GMM que definan cada clase de sonido. Los elementos principales que definen esta pantalla son los siguientes (ver Figura 73):

Barra de progreso circular: representación gráfica del porcentaje de entrenamiento activo transcurrido y restante. Además dispondrá de una serie de elementos auxiliares de apoyo a la tarea a realizar:

- Indicador de los primeros cinco segundos: Este tiempo se marca para que el usuario se mantenga en silencio. El motivo de esta acción viene dado para establecer el valor de intensidad actual. De esta forma, el cronómetro sólo avanzará cuando se graben muestras de audio cuya intensidad supere este umbral con cierta diferencia (para evitar que el usuario, a mitad de grabación, deje de reproducir un sonido durante un tiempo y esas muestras sean contempladas como pertenecientes al sonido en sí). Al ser un periodo durante el cual se espera un comportamiento concreto y diferenciado del resto, esto aparecerá indicado.

- Zona central con pictograma del sonido esperado: Como asistente para el usuario, que de esta forma sabrá en todo momento que espera de él la aplicación. Este pictograma variará al paso de los cinco segundos a menos que

se trate del entrenamiento del sonido ambiente. El diseño de los pictogramas se ha orientado a reflejar de forma directa y clara el significado de la notificación, minimizando así la carga cognitiva, en especial de las personas con sordera pre-locutiva.

Botón 'Empezar a escuchar' / 'Parar de escuchar': este botón presentará los dos estados indicados. Será el botón con el cual el usuario tendrá el control para comenzar o detener el entrenamiento.

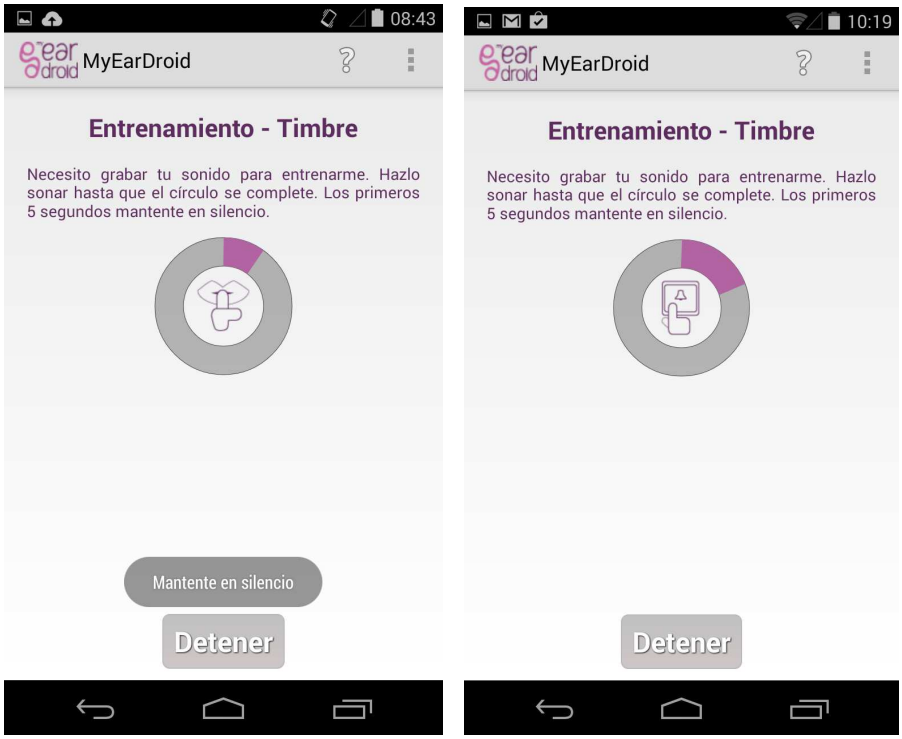


Figura 73 Pantalla de entrenamiento de la aplicación myEardroid.

Una vez finalizado el entrenamiento, la aplicación retornará a la pantalla principal desde la cual se inició el entrenamiento. El icono de entrenamiento se actualizará indicando que ese sonido ha sido entrenado y ya puede seleccionarse como sonido a reconocer.

Pantalla de análisis

Una vez se inicia el proceso de escucha a través del *Botón de Empezar* de la pantalla principal, el algoritmo de reconocimiento empieza a funcionar mostrándose la pantalla de análisis. Los elementos que componen esta pantalla son los siguientes (ver Figura 74):

Último sonido detectado: Muestra en la parte superior el último evento que ha sido detectado (icono y texto) junto con su hora de reconocimiento.

Estado del procesamiento: Mientras la aplicación está en escucha se muestra un círculo progresivo que gira constantemente. Esto da *feedback* al usuario de que la aplicación sigue activa y no se ha quedado parada o ha tenido fallo. En cuanto un sonido es detectado, el dispositivo móvil vibra y el círculo cambiará por el icono en grande del sonido reconocido. Esta imagen permanecerá visible durante unos segundos, yendo más tarde a la parte superior como "*Último sonido detectado*" y volviendo a mostrarse el círculo progresivo.

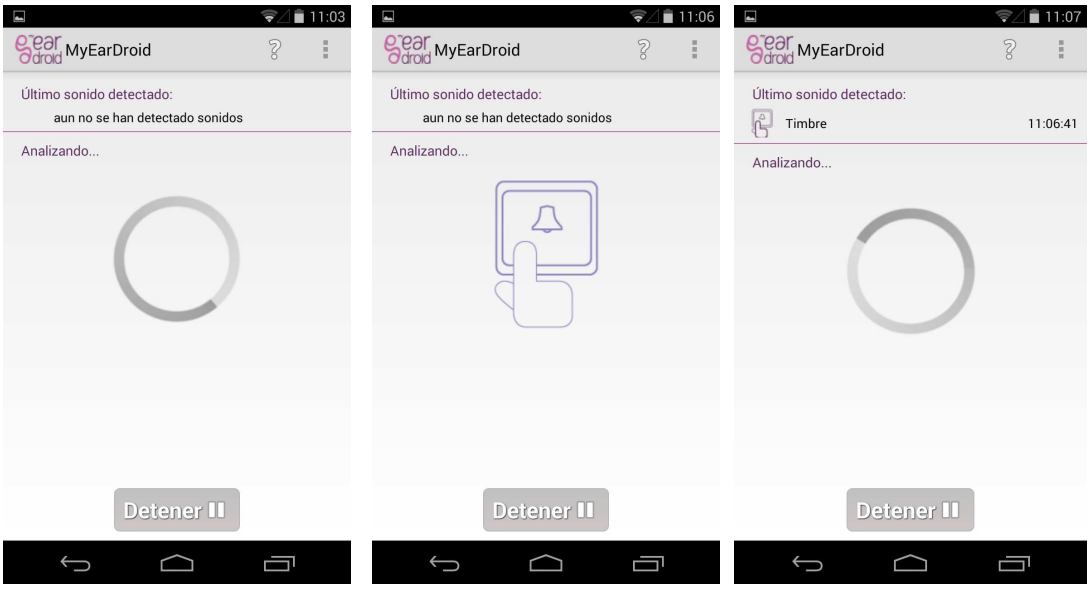


Figura 74 Pantalla de análisis de la aplicación myEardroid.

Mensajes de notificación

Cuando la aplicación *myEardroid* no se encuentra corriendo en primer plano o el móvil tiene la pantalla apagada, si el sistema detecta un evento acústico de los seleccionados deberá notificárselo al usuario. Siguiendo las pautas recogidas en el capítulo 5 estos mensajes son mostrados a través de vibración y a través de patrones luminosos desde el led situado generalmente en la parte posterior de los móviles. El aviso también aparecerá en la barra superior, mostrando un icono de notificación perteneciente al sonido detectado. En la Figura 75 se muestran los pictogramas utilizados para los sonidos en la aplicación junto con su correspondiente icono de notificación.

Figura 75 Pictogramas e iconos de notificación de la aplicación myEardroid.

Historial de sonidos detectados

La pantalla del historial muestra la lista de sonidos detectados desde el inicio de la misma. De esta forma, si el usuario se ha marchado de casa y ha dejado el móvil con la aplicación funcionando, podrá ver a su vuelta los eventos detectados (ver Figura 76).

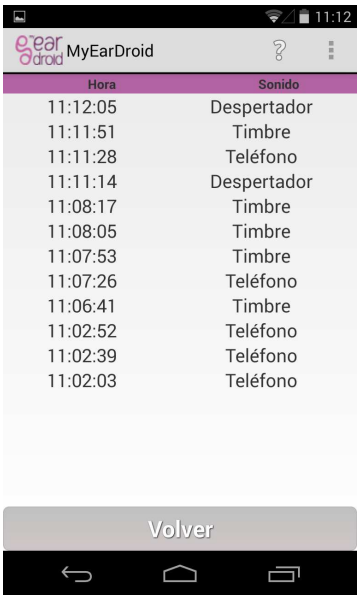


Figura 76 Pantalla del historial de la aplicación myEardroid.

7.2.4.3 Funcionamiento en Tiempo Real

El funcionamiento de la aplicación, antes de ser publicada, fue evaluado con 5 dispositivos móviles cubriendo gamas de procesadores de baja, media y alta prestación (ver Tabla 29).

Modelo	Procesador	Memoria	Versión Android
LG Nexus 4	Gama Alta (Quadcore – 1,5 GHz)	2GB RAM	4.2
Samsung Galaxy S3	Gama Alta (Quadcore – 1,4 GHz)	1GB RAM	4.0.4
Tablet Motorola Xoom	Gama Media (Dualcore – 1 GHz)	1GB RAM	3
Samsung Google Nexus S	Gama Media (1 GHz)	512 MB RAM	2.3
Samsung Galaxy Mini	Gama Baja (600 MHz)	384MB RAM	2.3.3

Tabla 29: Dispositivos móviles con Sistema Operativo Android testeados inicialmente.

Los móviles de gama media y alta no presentaron problema alguno en el funcionamiento en tiempo real. Sin embargo, el móvil de gama baja con procesador a 600MHz y memoria de 512MB, no fue capaz de procesar a la misma velocidad las tramas que tenía en la cola de audio que las que iban llegando. Esto suponía que la cola interna de paquetes se hacía cada vez más grande y se añadía un retardo en el reconocimiento. Aunque este efecto no era muy pronunciado, teniendo la aplicación horas funcionando el usuario podría recibir los mensajes de aviso varios segundos (o incluso minutos) más tarde.

Para solventar esta problemática se decidió implementar un protocolo basado en el estándar RTP (*Real Time Protocol*). Así, cuando el buffer de reconocimiento tiene demasiadas tramas sin procesar, las más antiguas se desechan. Aunque esto puede implicar perder algunos sonidos por detectar es una solución para móviles de gama muy baja más funcional.

Este trabajo de evaluación permitió que, al publicar la versión *Beta* en *GooglePlay* ésta fuera válida para la mayoría de dispositivos. Aun así, se hizo un seguimiento de los errores reportados de aquellos usuarios que mostraron problemas, mejorando fallos derivados de las diferentes versiones de *Android* de las diferentes marcas de fabricantes y modelos.

7.3 Conclusiones del Capítulo

En este capítulo se ha demostrado la aplicabilidad de las técnicas de reconocimiento de sonidos no-habla en el entorno del hogar, siendo este entorno valorado por los usuarios como el más importante.

En el primer apartado, en los experimentos sobre plataforma PC, se demuestra cómo el análisis en tiempo real con audio continuo produce mayores errores en el reconocedor al no tener éste conocimiento de dónde empieza y acaba el evento. Al

igual que en el capítulo 6 con la base de datos CHIL, en sonidos impulsivos este factor se agrava. Los sonidos más cortos como apertura de microondas, pitido de microondas, habla (tan sólo se grabaron palabras sueltas), golpes en la puerta,… son los que obtuvieron más errores. Detectar un sonido que posee una duración corta, si éste no está muy acotado en frecuencias tiene el problema de producir muchos falsos positivos o pérdidas del mismo ya que no hay suficientes tramas con las que poder asegurar un reconocimiento correcto. Cuando se trata de un entorno no controlado, donde pueden aparecer sonidos no registrados previamente en la base de datos, las probabilidades de fallo aumentan afectando al reconocimiento.

El oído humano funciona de igual manera y, muchas veces, la única forma que una persona tiene de reconocer este tipo de sonidos cortos e impulsivos es combinando información acústica con información del contexto. Este trabajo demuestra cómo al añadir información de sensores de presencia se consiguen eliminar falsos positivos y solucionar errores en la clasificación que, sólo utilizando la información del audio no había sido posible solventar. Esta estrategia fue la utilizada en el proyecto europeo RUBICON, cuyos resultados se publicaron en las revistas *Engineering Applications of Artificial Intelligence* [130] y *Journal of Intelligent & Robotic Systems* [131]

En el segundo apartado, se describe la investigación llevada a cabo para integrar los desarrollos previos en teléfonos móviles. La investigación demuestra cómo los micrófonos convencionales integrados en los móviles de hoy en día ofrecen un rendimiento similar a los micrófonos profesionales en esta área particular del reconocimiento de sonidos no-habla. Se demuestra la importancia de usar muestras de entrenamiento de calidad, utilizando sólo las tramas centrales del sonido y eliminando las tramas iniciales y finales con bajo nivel de intensidad. Adicionalmente, el tiempo de entrenamiento ha sido reducido a unos pocos segundos, lo cual es de gran relevancia para este tipo de aplicación donde el usuario graba sus propios sonidos.

El rendimiento final del sistema en condiciones sin ruido obtiene errores muy reducidos. Sin embargo, el mayor problema encontrado en los experimentos viene dado por la aplicación de estas técnicas en condiciones de ruido. Se ha demostrado cómo los parámetros más habituales en los sistemas de reconocimiento de sonidos no-habla (MFCC, y principalmente ZCR, Roll-Off Point y Centroid) no son robustos frente al ruido. La aplicación de la técnica Cepstral Mean Normalization mejora notablemente los resultados en presencia de ruido, sin embargo, esfuerzos más extensos en esta área son necesarios.

El software de reconocimiento fue finalmente diseñado e implementado para su funcionamiento en tiempo real sobre teléfonos móviles con sistema operativo Android. Siguiendo los resultados de la encuesta del capítulo 5, la aplicación se programó para vibrar cuando un evento es detectado y para mostrar con texto e iconos la información sobre el evento detectado. Este software ha sido validado en

varios teléfonos móviles con diferentes procesadores y ha sido premiado en la **VIII Edición del Premio Vodafone a la Innovación en Telecomunicaciones**, en la categoría "***Premio al desarrollo de Aplicación Mobile for Good***". Una descripción de esta aplicación ha dado lugar al artículo "*Household Sound Recognition System for the Hearing Impaired based on Mobile Phone Platforms*" enviado para su publicación a la revista *Universal Access in the Information Society*.

8. Conclusiones y Líneas Futuras

8.1 Aportaciones de la Tesis

Esta tesis presenta un trabajo pionero en el área del reconocimiento de eventos acústicos de interés para el colectivo de personas con discapacidad auditiva y en el diseño e implementación de soluciones reales donde las técnicas utilizadas puedan ser integradas de una manera práctica y sencilla para el usuario.

Mediante un **análisis exhaustivo del mercado** se ha **aportado información objetiva y contrastada del panorama actual y futuro de las tecnologías de reconocimiento de sonidos no-habla**. Este análisis muestra el actual envejecimiento de la población y, con ello, el aumento del número de personas con discapacidad. El cuidado de personas dependientes y la necesidad de ayudas técnicas que aporten seguridad y permitan realizar actividades cotidianas de la vida diaria prevén un futuro próspero a las empresas que inviertan en estas tecnologías. La madurez de los productos existentes en el mercado de aviso de eventos acústicos se encuentra en un nivel aún muy bajo. La implementación de algoritmos robustos puede aportar una gran diferenciación a empresas fabricantes de este tipo de tecnología y grandes beneficios económicos.

Sin embargo, hasta hoy en día, los estudios en sistemas de reconocimiento de sonidos no-habla se han limitado a aplicaciones en general muy alejadas de las orientadas al colectivo de personas con discapacidad auditiva. Además, el amplio abanico de eventos acústicos que se engloban dentro de la definición de sonidos no-habla incluye trabajos muy variados (desde el reconocimiento del canto de los pájaros hasta la identificación de tipos de disparos por arma). Pese a la existencia de algunas bases de datos comerciales, la gran mayoría de artículos analizados tienden a la creación de su propio corpus en base a los requerimientos u objetivos de los proyectos en los que se enmarcan. Esto es debido principalmente a la imposibilidad de poder reflejar todos los tipos de sonidos existentes en una única base de datos y, si esto fuera factible, la imposibilidad de manejo de un número tan elevado de clases resultantes. Debido a ello, existe una gran dificultad en realizar una comparativa clara sobre los trabajos estudiados en este campo. Las técnicas y algoritmos utilizados por los investigadores son aplicados a conjuntos muy diversos de sonidos, de número variable y grabaciones hechas en condiciones muy dispares, y en los que el foco de interés principal es más aumentar el número de clases o la precisión del sistema que la practicidad que estos sistemas deben tener para ser llevados al mercado.

Es por todo esto que la investigación presente en este documento se ha abordado desde dos enfoques diferentes. Por un lado se han validado los algoritmos

implementados con las **bases de datos más frecuentes en la literatura científica** y, por otro lado, estos resultados han servido para el diseño de sistemas que se han evaluado con **bases de datos creadas específicamente para cubrir las necesidades asociadas al colectivo de personas con discapacidad auditiva**. En esta investigación, como algoritmo de clasificación, se ha optado por la utilización de *Modelos de Mezclas Gaussianas* (GMM) por su facilidad en el entrenamiento, por su bajo consumo computacional y por presentar un modelado más genérico, capaz de representar un mayor número de tipos de sonidos. A su vez, las características acústicas elegidas fueron *Mel Frequency Cepstral Coefficients*, *Zero Crossing Rate*, *Roll-Off Point* y *Spectral Centroid*, junto con sus primeras y segundas derivadas (*delta* y *delta-delta*) muy habituales en la literatura científica.

Al validar estas **técnicas frente a las bases de datos más frecuentes** se demuestra cómo en distintas condiciones de grabación y para diferentes tipos de sonidos **el sistema se comporta de forma muy aceptable con números reducidos de gaussianas** con resultados de clasificación en sonidos aislados cercanos o superiores al 0,9 de *F1-score*. La mayor dificultad en estos sistemas radica en la etapa de detección de los eventos sobre audio continuo. En esta tesis **se han comparado dos estrategias de detección**: "*Detección y Clasificación*" y "*Detección por Clasificación*". Esta última demuestra obtener mejores resultados (AEER=47,024) bajo la medida AEER sobre todo en reducir el número de eventos eliminados que la técnica "*Detección y Clasificación*" produce. Si comparamos los resultados obtenidos con los aportados en CLEAR 2006 [123] la mejora es significativa con respecto a los sistema propuestos UPC-D (AEER=58,9) y CMU-D2 (AEER=52,5). El sistema ITC-D2 obtiene un error menor (AEER=33,7), sin embargo, en la métrica AEER de estos tres sistemas no tenían en consideración la clase "*desconocido*", siendo esta clase la que produce los mayores errores en nuestro sistema (36,58% de los eventos eliminados pertenecen a esta clase). Además, el sistema ITC-D2 utiliza modelos HMM y estos requieren conocer el inicio y final de todos los eventos a entrenar, no práctico para un sistema funcional final.

Para validar los **algoritmos con una base de datos de interés para el colectivo de las personas con discapacidad auditiva**, en esta tesis se ha llevado a cabo una **encuesta a través de cuestionarios online** difundidos gracias a la colaboración de varias organizaciones y fundaciones de personas con problemas auditivos. Con los datos extraídos de las respuestas de los encuestados ha quedado evidente cómo es el hogar el entorno más demandado para el reconocimiento de sonidos no-habla para este colectivo y cuáles son los sonidos de mayor interés. Con esta información, se crearon varias bases de datos donde los algoritmos se volvieron a evaluar sobre sonidos de interés y específicos del hogar, tanto para el aviso de sonidos de interés como para su uso por sistemas inteligentes en el reconocimiento de actividades de la vida diaria de la persona. En los experimentos se reafirma cómo es la etapa de detección sobre audio

continuo la que produce mayores errores en el reconocedor. Además, en sonidos impulsivos este factor se agrava. Detectar un sonido que posee una duración corta, si éste no está muy acotado en frecuencias tiene el problema de producir muchos falsos positivos o pérdidas del mismo ya que no hay suficientes tramas con las que poder asegurar un reconocimiento correcto. Es por ello que para dar mayor fiabilidad a los sistemas puede ser importante la fusión de datos con otros sensores que hagan que las probabilidades de reconocimiento aumenten o disminuyan en base a los datos recogidos. En esta tesis, el sistema implementado fue provisto de información de sensores de presencia localizados en las diferentes habitaciones donde se realizaron las grabaciones de los eventos acústicos. Esta información extra, integrada a través de simples reglas de decisión, permite crear subconjuntos más limitados de posibles eventos a reconocer, donde se demostró cómo el número de falsos positivos era reducido. Este mismo **sistema fue integrado en el proyecto europeo RUBICON, aportando información sonora del entorno para el reconocimiento de actividades de la persona**.

A través de la encuesta elaborada se ha constatado cómo las personas con discapacidad auditiva consideran **el teléfono móvil el medio ideal donde implementar soluciones de reconocimiento de sonidos no-habla**. No obstante, la integración de las técnicas de reconocimiento en teléfonos móviles conlleva dos problemas: por una parte, las limitaciones de cálculo computacional de estos dispositivos, y por otra parte, la movilidad del micrófono frente a las fuentes de audio. En esta tesis se demuestra cómo **las técnicas utilizas previamente pueden ser optimizadas para poder funcionar sobre teléfonos móviles en tiempo real**. Se ha demostrado cómo los micrófonos convencionales integrados en los móviles de hoy en día ofrecen un rendimiento similar a los micrófonos profesionales en esta área particular del reconocimiento, aunque la posición del teléfono móvil varíe e incluso pudiendo reducir el tiempo de entrenamiento a unos pocos segundos. Además, el sistema fue evaluado utilizando una base de entrenamiento fija pero modificando las bases de datos de validación y testeo con ruido adicional. El rendimiento del sistema en condiciones sin ruido es muy alto, sin embargo, cuando las bases de datos de validación y testeo son mezcladas con ruido adicional el rendimiento del sistema decae notablemente. Se demuestra cómo los parámetros más habituales en los sistemas de reconocimiento de sonidos no-habla que habíamos utilizado previamente, y que se encuentran muy presentes en la literatura científica, (MFCC, y principalmente ZCR, Roll-Off Point y Spectral Centroid) no son robustos frente al ruido. La normalización de canal ofrece grandes beneficios en este sentido. A través de la aplicación de la técnica Cepstral Mean Normalization los errores en el reconocimiento frente al ruido fueron disminuidos notablemente con un error casi tres veces inferior al error obtenido sin la normalización, además de ofrecer resultados muy similares en condiciones sin ruido.

Derivado de todos estos experimentos, el trabajo de esta tesis presenta el **diseño e implementación de una aplicación funcional de reconocimiento automático de sonidos del hogar para móvil capaz de trabajar en tiempo real sobre sistema operativo Android**. Su diseño, además de estar influenciado por varios estándares y recomendaciones de accesibilidad, está guiado por los resultados de la encuesta elaborada en el capítulo 5 que aporta información acerca de la forma óptima de aviso y visualización de mensajes. La gran variedad de timbres, despertadores y demás sonidos producidos en el hogar ha hecho que se descarte la idea de contar con una base de datos genérica de sonidos. Es por ello que en esta aplicación se ha primado la practicidad, permitiendo al usuario que sea él mismo el que grabe sus propios sonidos con los que se creen los modelos que más tarde se usarán en el reconocimiento de eventos acústicos del hogar. Este software ha sido validado en varios teléfonos móviles con diferentes procesadores y ha sido premiado en la **VIII Edición del Premio Vodafone a la Innovación en Telecomunicaciones**, en la categoría "***Premio al desarrollo de Aplicación Mobile for Good***".

8.2 Líneas Futuras

Del trabajo realizado en esta tesis se han detectado un conjunto de problemas y posibles mejoras que deben ser abordados en futuras investigaciones. En este apartado se resumirán críticamente las desventajas de las técnicas y desarrollos implementados y se propondrán líneas de investigación que puedan ser prometedoras.

8.2.1 Detección de Eventos

Como se ha expuesto, uno de los principales problemas que se deben afrontar en este tipo de sistemas es la detección de sonidos de duración muy corta. El timbre de la puerta o el portero automático, por ejemplo, pueden durar escasos segundos. Además, si utilizamos el teléfono móvil como medio para la detección, una ubicación de éste muy alejada de la fuente de sonido puede dificultar aún más la tarea. En la presente tesis se ha expuesto la fusión de esta información de audio con la información procedente de sensores de presencia. Sin embargo, es necesario profundizar más en esta alternativa. A través de técnicas de fusión de expertos diferentes sensores pueden proveer información del entorno que permita el reconocimiento más robusto de los sonidos no-habla. Incluso centrándonos sólo en los teléfonos móviles, la gran mayoría de éstos llevan integrados muy variados sensores que pueden proveer información de contexto adicional como la hora actual o información de posición para calcular la distancia a la fuente si este teléfono está provisto de dos o más micrófonos. Además, la detección también puede estar adaptada al tipo de sonido. Como se ha visto, hay sonidos muy cortos y sonidos más largos. Para evitar falsos positivos el sistema puede utilizar un suavizado de mínimo

número de tramas adaptado a la clase de evento a detectar. De esta forma, para la detección de un teléfono móvil podrán ser requeridas un número mayor de tramas consecutivas ya que este sonido, generalmente, dura más en el tiempo.

8.2.2 Mejora de la Robustez del sistema

La técnica Cepstral Mean Normalization otorga grandes mejoras al sistema cuando las bases de datos de validación y testeo son modificadas con la adición de ruido respecto a los habituales parámetros MFCC sin normalizar y los parámetros Zero Crossing Rate, Roll-Off Point y Spectral Centroid. Sin embargo, esta mejora queda lejos de ser suficiente. Es necesario profundizar más en esta problemática e investigar en algoritmos y técnicas que puedan reducir aún más este error.

Una propuesta a analizar, complementaria a la búsqueda de parámetros más robustos, es la utilización de una clase de evento tipo "*desconocidos*" que esté entrenada con grandes cantidades de diversos sonidos. En los experimentos realizados se ha demostrado cómo los mayores errores con ruido de fondo se encontraban cuando el ruido contenía eventos impulsivos fuera del conjunto de eventos seleccionados (golpes en mesas, sillas, cubiertos, gritos, habla, sonido de pájaros,…). Poseer una clase de eventos "*desconocidos*" de miles de sonidos variados puede hacer que la confusión del sistema sea menor. El problema a afrontar sería el número de gaussianas que este modelo requeriría o si sería más conveniente disponer de varias clases de eventos tipo "*desconocidos*" en base a jerarquías analizadas en el estado del arte que tengan en cuenta el tipo de material o la tipología del sonido.

Otra alternativa es la posibilidad de incorporar funcionalidades de re-entrenamiento al software desarrollado. Cuando el sistema detecta un evento, ese sonido puede ser guardado para posteriormente ser reproducido por una persona con capacidades auditivas. Si el sistema ha producido un falso positivo, la persona podrá saberlo reproduciendo el sonido, pudiendo indicar al sistema que el evento detectado es otro para que los modelos GMM se puedan re-entrenar y mejorar en su clasificación.

8.2.3 Ampliación de los Entornos de Interés

La presente tesis se ha centrado en el colectivo de personas con discapacidad auditiva pero, principalmente, en el reconocimiento de sonidos en el entorno del hogar. No obstante los trabajos realizados han dejada abiertos otros entornos de interés a nuevas investigaciones, fijando pautas de diseño y aportando posibles estrategias en lo que a algoritmos posibles se refiere. El capítulo 5 ha plasmado las necesidades de las personas con discapacidad auditiva en entorno como la oficina, centro de trabajo, calle o vehículo. Se ha aportado información de los sonidos de mayor interés así como de los medios más aceptados y la forma de aviso y visualización de los mensajes en cada medio. Queda a disposición de otros investigadores la labor de profundizar y aplicar

esta información en el desarrollo de sistemas reales que satisfagan las necesidades de este colectivo.

8.2.4 Reconocimiento de Actividades

Finalmente, es importante mencionar que el área de reconocimiento de sonidos puede también ser útil en otros campos complementarios al de la discapacidad auditiva. La información acústica puede ayudar a entender que está ocurriendo en el entorno [87], [136]. En sistemas *Ambient Assisted Living* [137], [138] la capa de aprendizaje y razonamiento puede usar el reconocimiento de eventos acústicos para responder de forma más precisa a las necesidades de los usuarios. Es por ello que, una línea futura de trabajo es el reconocimiento de una secuencia de eventos de sonido (abrir la nevera, ruido de platos, ruido de sillas, pitido de microondas,…) como una manera de inferir actividades (ej. la persona está preparando la comida). Esto puede conllevar a la detección de cambios en los patrones de comportamiento [139], [140], [141], [142] que alerten de potenciales problemas de deterioro cognitivo.

Aunque los desarrollos presentes han sido aplicados con este fin en el proyecto Europeo FP7 RUBICON (*Robotic UBIquitous Cognitive Network*), la dificultad que supone este reto hace necesario profundizar aún más en esta línea de investigación y contrastar los resultados en estudios reales más extensos.

Publicaciones

- **H. Lozano**, I. Odriozola, I. Hernáez, J. Camarena, E. Navas, and I. Gutierrez, "Household Sound Recognition System for the Hearing Impaired based on Mobile Phone Platforms," *Universal Access in the Information Society*, 2015 [**en revision**]. (Impact Factor 0.475, Q4 in Computer Science, Cybernetics)

- Dragone, G. Amato, D. Bacciu, S. Chessa, S. Coleman, M. Di Rocco, C. Gallicchio, C. Gennaro, **H. Lozano**, L. Maguire, M. McGinnity, A. Micheli, G. M. P. O'Hare, A. Renteria, A. Saffiotti, C. Vairo, and P. Vance, "A cognitive robotic ecology approach to self-configuring and evolving AAL systems," *Eng. Appl. Artif. Intell.*, vol. 45, pp. 269–280, Oct. 2015. (Impact Factor 2.207, Q1 in Computer Sicence, Artificial Intelligence)
- G. Amato, D. Bacciu, M. Broxvall, S. Chessa, S. Coleman, M. Di Rocco, M. Dragone, C. Gallicchio, C. Gennaro, **H. Lozano**, T. M. McGinnity, A. Micheli, A. K. Ray, A. Renteria, A. Saffiotti, D. Swords, C. Vairo, and P. Vance, "Robotic Ubiquitous Cognitive Ecology for Smart Homes," *Journal of Intelligent & Robotic Systems*, pp. 1–25, Feb. 2015. (Impact Factor 1.178, Q3 in Computer Science, Artificial Intelligence and Robotics).
- D. Bacciu, M. Broxvall, S. Coleman, M. Dragone, C. Gallicchio, G. Gennaro, R. Guzman, R. Lopez, **H. Lozano**, A. Ray, A. Renteria, A. Saffiotti, and C. Vairo, "Self-Sustaining Learning for Robotic Ecologies," *1st Int. Conf. on Sensor Networks*, pp. 99–103, 2012.
- **H. Lozano**, I. Hernáez, J. Camarena, I. Díez, and E. Navas, "Identification of sounds and sensing technology for home-care applications," in: *Ambient Assisted Living Home Care,* LNCS vol. 7657, pp. 74–81, 2012.
- **H. Lozano**, I. Hernáez, J. Camarena, I. Díez, and E. Navas, "A real-time sound recognition system in an assisted environment," in *Computers Helping People with Special Needs*, LNCS Vol. 7383, pp. 385–391, 2012.
- **H. Lozano**, I. Hernáez, A. Picón, J. Camarena, and E. Navas, "Audio classification techniques in home environments for elderly/dependant people," *Computers Helping People with Special Needs.* LNCS, Vol.6179 pp. 320–323, 2010.
- **H. Lozano**, I. Hernáez, E. Navas, F. González, and I. Idigoras, "Non-Speech Sounds Classification for People with Hearing Disabilities," *Assitive Technology Research Series, , Vol. 20, Challenges for Assitive Technology , IOPress,* pp. 276–280, 2007.
- **H. Lozano,** I. Hernáez, E. Navas, F. González, and I. Idigoras, "Household Sound Identification System for People with Hearing Disabilities", in *Conference and Workshop on Assistive Technology for People with Vision and Hearing Impairments* (CVHI), 2007
- **H. Lozano**, I. Hernáez, E. Navas, F. González, and I. Idigoras, "Técnicas de Clasificacion de Sonidos del Hogar," in *URSI*, 2007.

Referencias

[1] INE, “Encuesta sobre Discapacidades, Deficiencias y Minusvalías (EDDM1986),” *Instituto Nacional de Estadística*, 1986. [Online]. Available: http://www.ine.es. [Accessed: 24-Nov-2015].

[2] INE, “Encuesta sobre Discapacidades, Deficiencias y Estado de Salud (EDDS1999),” *Instituto Nacional de Estadística*, 1999. [Online]. Available: http://www.ine.es. [Accessed: 24-Nov-2015].

[3] INE, “Encuesta de Discapacidad, Autonomía personal y situaciones de Dependencia (EDAD2008),” *Instituto Nacional de Estadística*, 2008. [Online]. Available: http://www.ine.es. [Accessed: 24-Nov-2015].

[4] INE, “Panorámica de la discapacidad en España,” *Cifras INE - Bol. Inf. del Inst. Nac. Estadística*, pp. 1–12, 2009.

[5] UNDESA, “World Population Prospects,” *Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat*, 2008. [Online]. Available: http://www.un.org/. [Accessed: 24-Nov-2015].

[6] A. Salvá, A. Rivero, and M. Roqué, *Evolución del proceso de envejecimiento de la población española y análisis de sus determinantes*. Fundación Pfizer, 2007.

[7] FVA, *Acceso y uso de las TIC por las personas con discapacidad*. Madrid: Fundación Vodafone, 2013.

[8] FVA, *Tics y Dependencia*. Madrid: Fundación Vodafone, 2007.

[9] N. VanDerveer, “Ecological acoustics: Human Perception of Environmental Sounds,” Cornell University, 1979.

[10] J. Ballas and J. Howard, “Interpreting the language of environmental sounds,” *Environ. Behav.*, vol. 19(1), pp. 91–114, 1987.

[11] J. Kane, “Poetry as right-hemispheric Language,” *J. Conscious. Stud.*, vol. 11(5–6), pp. 21–59, 2004.

[12] M. Cowling, “Non-speech environmental sound classification system for autonomous surveillance,” *PhD thesis, Griffith Univ.*, 2004.

[13] C. Couvreur, “Environmental sound recognition: a statistical approach,” *PhD thesis, Fac. Polytech. Mons*, 1997.

[14] D. Gerhard, *Audio signal classification: History and current techniques*. Technical Report TR-CS 2003-07 November, 2003.

[15] S. Cheng and H. Wang, "METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation," in *8th International Conference on Spoken Language Processing*, 2004, pp. 1617–1620.

[16] Y. SHI, W. XUE, and S. BIN, "Several features for discrimination between vocal sounds and other environmental sounds," *EUSIPCO Conf.*, pp. 2099–2102, 2004.

[17] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proceedings of the ninth ACM international conference on Multimedia - MULTIMEDIA '01*, 2001, p. 203.

[18] W. Gaber, "What in the World Do We Hear-An Ecological Approach to Auditory Event Perception," *Ecol. Phychology*, vol. 5(1), pp. 1–29, 1993.

[19] B. Gygi and V. Shafiro, "Development of the Database for Environmental Sound Research and Application (DESRA): Design, Functionality, and Retrieval Considerations," *EURASIP J. Audio, Speech, Music Process.*, vol. 2010, pp. 1–12, 2010.

[20] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds.," *Percept. Psychophys.*, vol. 69, no. 6, pp. 839–55, Aug. 2007.

[21] A. Temko, "Acoustic event detection and classification," *PhD Thesis. Univ. Politècnica Catalunya*, 2007.

[22] T. Matthews, S. Carter, and C. Pai, "Scribe4Me: Evaluating a Mobile Sound Transcription Tool for the Deaf," *UbiComp 2006 Ubiquitous Comput.*, vol. 4206, pp. 159–176, 2006.

[23] I.-C. Yoo and D. Yook, "Automatic sound recognition for the hearing impaired," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 2029–2036, Nov. 2008.

[24] M. Vacher, D. Istrate, L. Besacier, and J. Serignat, "Life Sounds Extraction and Classification in Noisy Environment.," *SIP. Fifth IASTED Int. Assoc. Sci. Technol. Dev. Int. Conf. Signal Image Process.*, pp. 13–15, 2003.

[25] M. Vacher, D. Istrate, L. Besacier, J. Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," *IASTED Biomed. Conf.*, pp. 395–399, 2004.

[26] D. Istrate, M. Vacher, J. Serignat, and E. Castelli, "Multichannel smart sound sensor for perceptive spaces," *Complex Syst. Intell. Mod. Technol. Appl.*, pp. 691–696, 2004.

[27] E. Castelli, M. Vacher, D. Istrate, L. Besacier, and J. Serignat, “Habitat telemonitoring system based on the sound surveillance,” *ICICTH (International Conf. Inf. Commun. Technol. Heal.*, pp. 141–146, 2003.

[28] N. Laydrus, E. Ambikairajah, and B. Celler, “Automated sound analysis system for home telemonitoring using shifted delta cepstral features,” *15th Int. Conf. Digit. Signal Process.*, pp. 35–38, 2007.

[29] J. Wang, H. Lee, J. Wang, and C. Lin, “Robust environmental sound recognition for home automation,” *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 25–31, 2008.

[30] M. Janvier, X. Alameda-pineda, L. Girin, and R. Horaud, “Sound-Event Recognition with a Companion Humanoid,” *12th IEEE-RAS Int. Conf. Humanoid Robot.*, pp. 104–111, 2012.

[31] a. H. Kam and L. Shue, “An Automatic Acoustic Bathroom Monitoring System,” *2005 IEEE Int. Symp. Circuits Syst.*, pp. 1750–1753, 2005.

[32] J. Chen, A. Kam, J. Zhang, N. Liu, and L. Shue, “Bathroom activity monitoring based on sound,” *Pervasive Comput.*, pp. 47–61, 2005.

[33] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, “Analysis of chewing sounds for dietary monitoring,” *UbiComp 2005 Ubiquitous Comput.*, vol. 3660, pp. 56–72, 2005.

[34] D. K. Fragoulis and J. N. Avaritsiotis, “A Siren Detection System based on Mechanical Resonant Filters,” *Sensors*, vol. 1, no. 4, pp. 121–137, Sep. 2001.

[35] F. Beritelli, S. Casale, A. Russo, and S. Serrano, “An Automatic Emergency Signal Recognition System for the Hearing Impaired,” in *2006 IEEE 12th Digital Signal Processing Workshop & 4th IEEE Signal Processing Education Workshop*, 2006, pp. 179–182.

[36] J. Schroder, S. Goetze, V. Grutzmacher, and J. Anemuller, “Automatic acoustic siren detection in traffic noise by part-based models,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 493–497.

[37] S. Oberle and A. Kaelin, “Recognition of acoustical alarm signals for the profoundly deaf using hidden Markov models,” in *Proceedings of ISCAS’95 - International Symposium on Circuits and Systems*, 1995, vol. 3, pp. 2285–2288.

[38] F. W. Ho-Ching, J. Mankoff, and J. A. Landay, “Can you see what i hear?,” in *Proceedings of the conference on Human factors in computing systems - CHI ’03*, 2003, p. 161.

[39] T. Matthews, J. Fong, and J. Mankoff, “Visualizing non-speech sounds for the deaf,” in *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility - Assets ’05*, 2005, p. 52.

[40] T. Matthews, J. Fong, F. W.-L. Ho-Ching, and J. Mankoff, “Evaluating non-speech sound visualizations for the deaf,” *Behav. Inf. Technol.*, vol. 25, no. 4, pp. 333–351, Jul. 2006.

[41] J. Azar, H. Saleh, and M. Al-Alaoui, “Sound visualization for the hearing impaired,” *Int. J. Emerg. Technol. Learn.*, vol. 2(1), pp. 1–7, 2007.

[42] M. Tomitsch and T. Grechenig, “Design Implications for a Ubiquitous Ambient Sound Display for the Deaf.,” *Conf. Work. Assist. Technol. People with Vis. Hear. Impair.*, 2007.

[43] M. Tomitsch, T. Grechenig, A. Vande Moere, and S. Renan, “Information Sky: Exploring the Visualization of Information on Architectural Ceilings,” *2008 12th Int. Conf. Inf. Vis.*, pp. 100–105, Jul. 2008.

[44] J. Nam, G. J. Mysore, and P. Smaragdis, “Sound Recognition in Mixtures,” *Latent Var. Anal. Signal Sep.*, vol. 7191, pp. 405–413, 2012.

[45] T. Heittola and A. Mesaros, “Sound event detection in multisource environments using source separation,” *CHiME*, pp. 36–40, 2011.

[46] T. Heittola, A. Mesaros, T. Virtanen, and M. Gabbouj, “Supervised model training for overlapping sound events based on unsupervised source separation.,” *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013.

[47] J. Kotus, K. Lopatka, and A. Czyzewski, “Detection and localization of selected acoustic events in acoustic field for smart surveillance applications,” *Multimed. Tools Appl.*, vol. 68, no. 1, pp. 5–21, Jul. 2012.

[48] V. Shafiro and B. Gygi, “How to select stimuli for environmental sound research and where to find them,” *Behav. Res. Methods, Instruments, Comput.*, vol. 36, no. 4, pp. 590–598, Nov. 2004.

[49] RWCP, “The RWCP Sound Scene Database in Real Acoustical Environments,” *Real Acoustic Environment Subgroup of RWCP Intellectual Resource for Research and Development Working Group in Real World Computing Partnership*, 2000. [Online]. Available: http://tosa.mri.co.jp/sounddb/indexe.htm. [Accessed: 10-Oct-2015].

[50] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, “Detection and Classification of Acoustic Scenes and Events,” *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[51] A. F. Smeaton and M. McHugh, “Towards event detection in an audio-based sensor network,” *Proc. third ACM Int. Work. Video Surveill. Sens. networks - VSSN ’05*, p. 87, 2005.

[52] M. Vacher, D. Istrate, and J. Serignat, “Sound detection and classification through transient models using wavelet coefficient trees,” *12th Eurasip Eur. Signal Process. Conf.*, pp. 1171–1174, 2004.

[53] D. Mitrovic, M. Zeppelzauer, and H. Eidenberger, *Towards an optimal feature set for environmental sound recognition*. Technical Report TR-188-2-2006-03, 2006.

[54] R. K. Reddy, V. Ramachandra, N. Kumar, and N. C. Singh, “Categorization of environmental sounds.,” *Biol. Cybern.*, vol. 100, no. 4, pp. 299–306, Apr. 2009.

[55] I. Ding, “Events Detection for Audio Based Surveillance by Variable-Sized Decision Windows Using Fuzzy Logic Control,” *Tamkang J. Sci. Eng.*, vol. 12, no. 3, pp. 299–308, 2009.

[56] H. Phan and A. Mertin, “A voting-based technique for acoustic event-specific detection,” *40th Annu. Ger. Congr. Acoust. (DAGA 2014)*, 2014.

[57] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, “An Ensemble of Rejecting Classifiers for Anomaly Detection of Audio Events,” *2012 IEEE Ninth Int. Conf. Adv. Video Signal-Based Surveill.*, pp. 76–81, Sep. 2012.

[58] A. Kumar, P. Dighe, R. Singh, S. Chaudhuri, and B. Raj, “Audio event detection from acoustic unit occurrence patterns,” *2012 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 489–492, Mar. 2012.

[59] D. Mitrovic, “Discrimination and Retrieval of Environmental sounds,” *PhD Thesis. Vienna Univ. Technol. Vienna*, 2005.

[60] S. Essid, G. Richard, and B. David, “Hierarchical classification of musical instruments on solo recordings,” *ICASSP’06*, vol. 5, pp. 817–820, 2006.

[61] M. McKinney and J. Breebaart, “Features for audio and music classification.,” *ISMIR*, vol. 4, pp. 151–158, 2003.

[62] S. Al-Zhrani and M. Alqahtani, “Audio Environment Recognition using Zero Crossing Features and MPEG-7 Descriptors,” *J. Comput. Sci.*, vol. 6, no. 11, pp. 1262–1266, 2010.

[63] J. Eggink and G. Brown, “Application of missing feature theory to the recognition of musical instruments in polyphonic audio.,” *ISMIR*, no. 1999, 2003.

[64] J. Marques and P. Moreno, “A study of musical instrument classification using Gaussian mixture models and support vector machines,” *Cambridge Res. Lab. Tech. Rep.*, 1999.

[65] J. Ludeña and A. Gallardo, “NMF-Based Spectral Analysis for Acoustic Event Classification Tasks,” *Adv. Nonlinear Speech Process.*, vol. 7911, pp. 9–16, 2013.

[66] M. Espi, M. Fujimoto, D. Saito, N. Ono, and S. Sagayama, “A tandem connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, vol. 7683, pp. 4293–4296.

[67] M. Espi, M. Fujimoto, D. Saito, N. Ono, and S. Sagayama, “A TANDEM CONNECTIONIST MODEL USING COMBINATION OF MULTI-SCALE SPECTRO-TEMPORAL FEATURES FOR ACOUSTIC EVENT DETECTION Graduate School of Informations Science and Technology , University of Tokyo , Japan NTT Communication Science Laboratories , NTT Corporat,” pp. 4293–4296, 2012.

[68] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, “Acoustic Monitoring and Localization for Social Care,” *J. Comput. Sci. Eng.*, vol. 6, no. 1, pp. 40–50, Mar. 2012.

[69] A. G. F. Properties, “Gammatone Cepstral Coef fi cients: Biologically Inspired Features for Non-Speech Audio Classi fi cation,” vol. 14, no. 6, pp. 1684–1689, 2012.

[70] X. Valero and F. Alías, “Gammatone wavelet features for sound classification in surveillance applications,” *Signal Process. Conf. (EUSIPCO), 2012 Proc. 20th Eur.*, pp. 1658–1662, 2012.

[71] C. Kim and R. M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4101–4104.

[72] H. Kayser, “Audio Classification and Localization for Incongruent Event Detection,” *Detect. Identif. Rare Audiov. Cues*, vol. 384, pp. 39–46, 2012.

[73] B. Uzkent, B. Barkana, and H. Cevikalp, “Non-speech environmental sound classification using SVMs with a new set of features,” *Int. J. Innov. Comput. Inf. Control ICIC Int.*, vol. 8, no. 5, pp. 3511–3524, 2012.

[74] S. Chu, S. Narayanan, and C.-C. J. Kuo, “Environmental sound recognition using MP-based features,” *2008 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1–4, Mar. 2008.

[75] S. Chu, “Unstructured Audio Classification for Environment Recognition.,” *Twenty-Third AAAI Conf. Artif. Intell.*, pp. 1845–1846, 2008.

[76] A. Wang, “An Industrial Strength Audio Search Algorithm.,” *Int. Conf. Music Inf. Retr.*, pp. 713–718, 2003.

[77] J. Dennis, H. D. Tran, and E. S. Chng, “Overlapping sound event recognition using local spectrogram features and the generalised hough transform,” *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 1085–1093, Jul. 2013.

[78] I. Luengo, “Análisis y Evaluación de Parámetros para Identificación Automática de Emociones en el Habla,” *PhD Thesis, Univ. del País Vasco*, 2010.

[79] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers, 1998.

[80] I. Jolliffe, *Principal component analysis*. Springer series in statistics, 2005.

[81] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[82] R. Leite and P. Brazdil, “Decision tree-based attribute selection via sub sampling,” *Work. minería datos y aprendizaje, Herrera, F., Riquelme, J. (eds), VIII Iberamia, Sevilla, Spain*, pp. 77–83, 2002.

[83] S. Piramuthu, “Evaluating feature selection methods for learning in data mining applications,” in *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, 2004, vol. 5, pp. 294–301.

[84] E. Vozarikova, M. Lojka, J. Juhar, and A. Cizmar, “Performance of Basic Spectral Descriptors and MRMR Algorithm to the Detection of Acoustic Events,” *Multimed. Commun. Serv. Secur.*, vol. 287, pp. 350–359, 2012.

[85] Z. Zhang and B. Schuller, “Semi-supervised learning helps in sound event classification,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 333–336.

[86] B. Clarkson, N. Sawhney, and A. Pentland, “Auditory context awareness via wearable computing,” *Work. Percept. User Interfaces*, 1998.

[87] L. Ma, D. Smith, and B. Milner, “Context awareness using environmental noise classification.,” *Eighth Eur. Conf. Speech Commun. Technol. ISCA*, pp. 2237–2240, 2003.

[88] L. Ma, B. Milner, and D. Smith, “Acoustic environment classification,” *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–22, Jul. 2006.

[89] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, “Audio-based context recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[90] S. Chu, S. Narayanan, and C. Kuo, “Content analysis for acoustic environment classification in mobile robots,” *AAAI Fall Symp. Aurally Inf. Perform. Integr. Mach. List. Audit. Present. Robot. Syst. Arlington, Va, USA*, 2006.

[91] S. Chu, S. Narayanan, C. -c. Kuo, and M. Mataric, “Where am I? Scene Recognition for Mobile Robots using Audio Features,” in *2006 IEEE International Conference on Multimedia and Expo*, 2006, pp. 885–888.

[92] D. Ellis, “Detecting alarm sounds,” *CRAC Work. Aalborg, Denmark*, pp. 3–6, 2001.

[93] M. Cowling and R. Sitte, “Analysis of speech recognition techniques for use in a non-speech sound recognition system,” *6th Int. Symp. Digit. Signal Process. Commun. Syst.*, pp. 16–20, 2002.

[94] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, “Automatic sound detection and recognition for noisy environment,” *Signal Process. Conf. 2000 10th Eur.*, pp. 1–4, 2000.

[95] L. Besacier, A. Dufaux, M. Ansorge, and F. Pellandini, “Automatic sound recognition relying on statistical methods, with application to telesurveillance,” *Proc. COST 254, Int. Work. Intell. Commun. Technol. Appl. with Emphas. Mob. Commun.*, pp. 116–120, 1999.

[96] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, “Scream and gunshot detection and localization for audio-surveillance systems,” in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 2007, pp. 21–26.

[97] W. Huang, S. Lau, T. Tau, L. Li, and L. Wyse, “Audio events classification using hierarchical structure,” in *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, 2007, vol. 3, pp. 1299–1303.

[98] C. Clavel, T. Ehrette, and G. Richard, “Events Detection for an Audio-Based Surveillance System,” in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 1306–1309.

[99] S. Ntalampiras, I. Potamitis, and N. Fakotakis, “On acoustic surveillance of hazardous situations,” *2009 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 165–168, Apr. 2009.

[100] R. Radhakrishnan, a. Divakaran, and P. Smaragdis, “Audio analysis for surveillance applications,” *IEEE Work. Appl. Signal Process. to Audio Acoust. 2005.*, pp. 158–161, 2005.

[101] K. Kim and H. Ko, “Hierarchical approach for abnormal acoustic event classification in an elevator,” *2011 8th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, pp. 89–94, Aug. 2011.

[102] W. Huang, S. Lau, and T. Tan, “Audio events classification using hierarchical structure,” *… Signal Process. …*, pp. 1299–1303, 2003.

[103] G. Guo and S. Z. Li, “Content-based audio classification and retrieval by support vector machines.,” *IEEE Trans. Neural Netw.*, vol. 14, no. 1, pp. 209–15, Jan. 2003.

[104] A. Temko, C. Nadeu, and D. Macho, “Acoustic Event Detection,” *Comput. Hum. Interact. Loop*, vol. 4625, pp. 61–73, 2008.

[105] C. Boukis and L. Polymenakos, “The Acoustic Event Detector of AIT,” *Multimodal Technol. Percept. Humans*, vol. 4625, pp. 328–337, 2008.

[106] C. Zieger, “An HMM Based System for Acoustic Event Detection,” *Multimodal Technol. Percept. Humans*, vol. 4625, pp. 338–344, 2008.

[107] X. Zhou, X. Zhuang, and M. Liu, “HMM-Based Acoustic Event Detection with AdaBoost Feature Selection,” *Multimodal Technol. Percept. Humans*, vol. 4625, pp. 345–353, 2008.

[108] M. Vacher, A. Fleury, F. Portet, J. Serignat, and N. Noury, “Complete sound and speech recognition system for health smart homes: application to the recognition of activities of daily living,” *Intech B.*, pp. 645–673, 2010.

[109] M. Vacher, F. Portet, A. Fleury, and N. Noury, “Development of audio sensing technology for ambient assisted living: Applications and challenges,” *Int. J. E-Health Med. Commun.*, vol. 2(1), pp. 35–54, 2011.

[110] W. Choi, J. Rho, D. K. Han, and H. Ko, “Selective Background Adaptation Based Abnormal Acoustic Event Recognition for Audio Surveillance,” *2012 IEEE Ninth Int. Conf. Adv. Video Signal-Based Surveill.*, pp. 118–123, Sep. 2012.

[111] A. Alkilani, “Automatic Acoustic Events Detection, Classification, and Semantic Annotation for Persistent Surveillance Applications,” *PhD Thesis, Tennessee State Univ.*, 2014.

[112] M. Rossi, S. Feese, O. Amft, N. Braune, S. Martis, and G. Troster, “AmbientSense: A real-time ambient sound recognition system for smartphones,” in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, 2013, pp. 230–235.

[113] E. Kiktova, M. Lojka, M. Pleva, J. Juhar, and A. Cizmar, “Comparison of Different Feature Types for Acoustic Event Detection System,” *Multimed. Commun. Serv. Secur.*, vol. 368, pp. 288–297, 2013.

[114] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, p. 1, 2013.

[115] M. Lojka, M. Pleva, J. Juhar, and E. Kiktova, “Modification of Widely Used Feature Vectors for Real-time Acoustic Events Detection,” *ELMAR, 2013 55th Int. Symp.*, pp. 25–27, 2013.

[116] G. Wichern, J. Xue, H. Thornburg, B. Mechtley, and A. Spanias, “Segmentation, Indexing, and Retrieval for Environmental and Natural Sounds,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 18, no. 3, pp. 688–707, Mar. 2010.

[117] M. Rossi, S. Feese, O. Amft, N. Braune, S. Martis, and G. Tr, “AmbientSense : A Real-Time Ambient Sound Recognition System for Smartphones,” no. March, pp. 230–235, 2013.

[118] M. A. Sehili, D. Istrate, B. Dorizzi, and J. Boudy, “Daily sound recognition using a combination of GMM and SVM for home automation,” *Signal Process. Conf. (EUSIPCO), 2012 Proc. 20th Eur.*, pp. 1673–1677, 2012.

[119] I. Ding, “Fuzzy Rule-Based System for Decision Making Support of Hybrid SVM-GMM Acoustic Event Detection,” *Int. J. Fuzzy Syst.*, vol. 14, no. 1, pp. 118–130, 2012.

[120] X. Zhuang, X. Zhou, M. a. Hasegawa-Johnson, and T. S. Huang, “Real-world acoustic event detection,” *Pattern Recognit. Lett.*, vol. 31, no. 12, pp. 1543–1551, Sep. 2010.

[121] T. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[122] J. Dhruv, L. Findlater, J. Gilkeson, B. Holland, R. Duraiswami, D. Zotkin, C. Vogler, and J. Froehlich, “Head-Mounted Display Visualizations to Support Sound Awareness for the Deaf and Hard of Hearing,” *ACM Conf. Hum. Factors Comput. Syst.*, 2015.

[123] A. Temko, R. Malkin, and C. Zieger, “CLEAR Evaluation of Acoustic Event Detection and Classification Systems,” *Multimodal Technol. Percept. Humans*, vol. 4122, pp. 311–322, 2007.

[124] P. Atrey, C. Maddage, and S. Kankanhalli, “Audio based event detection for multimedia surveillance,” *ICCASP. IEEE Intern. Conf.*, pp. 813–816, 2006.

[125] M. Vacher, D. Istrate, and J. Serignat, “Detection and speech/sound segmentation in a smart room environment,” *3rd Int. Conf. Speech Technol. Human-Computer Dialogue*, pp. 37–48, 2005.

[126] H. Lozano, I. Hernáez, J. Camarena, I. Díez, and E. Navas, “A real-time sound recognition system in an assisted environment,” in *Computers Helping People with Special Needs*, 2012, pp. 385–391.

[127] H. Lozano, I. Hernáez, J. Camarena, I. Díez, and E. Navas, “Identification of sounds and sensing technology for home-care applications,” *Ambient Assist. Living Home Care*, vol. 7657, pp. 74–81, 2012.

[128] G. Amato, M. Broxvall, S. Chessa, M. Dragone, G. Gennaro, R. Lopez, L. Maguire, T. Mcginnity, A. Micheli, A. Renteria, G. O’Hare, and F. Pecora, “Robotic UBIquitous COgnitive Network,” *Ambient Intell. - Softw. Appl. Adv. Intell. Soft Comput.*, vol. 153, pp. 191–195, 2012.

[129] D. Bacciu, M. Broxvall, S. Coleman, M. Dragone, C. Gallicchio, G. Gennaro, R. Guzman, R. Lopez, H. Lozano, A. Ray, A. Renteria, A. Saffiotti, and C. Vairo, “Self-Sustaining Learning for Robotic Ecologies,” *2012 Int. Conf. Sens. Networks*, pp. 99–103, 2012.

[130] M. Dragone, G. Amato, D. Bacciu, S. Chessa, S. Coleman, M. Di Rocco, C. Gallicchio, C. Gennaro, H. Lozano, L. Maguire, M. McGinnity, A. Micheli, G. M. P. O’Hare, A. Renteria, A. Saffiotti, C. Vairo, and P. Vance, “A cognitive robotic ecology approach to self-configuring and evolving AAL systems,” *Eng. Appl. Artif. Intell.*, vol. 45, pp. 269–280, Oct. 2015.

[131] G. Amato, D. Bacciu, M. Broxvall, S. Chessa, S. Coleman, M. Di Rocco, M. Dragone, C. Gallicchio, C. Gennaro, H. Lozano, T. M. McGinnity, A. Micheli, A. K. Ray, A. Renteria, A. Saffiotti, D. Swords, C. Vairo, and P. Vance, “Robotic Ubiquitous Cognitive Ecology for Smart Homes,” *J. Intell. Robot. Syst.*, pp. 1–25, Feb. 2015.

[132] H. Lozano, I. Hernáez, E. Navas, F. González, and I. Idigoras, “Non-Speech Sounds Classification for People with Hearing Disabilities,” *Conf. Assoc. Adv. Assist. Technol. Eur. AAATE*, pp. 276–280, 2007.

[133] P. Pujol, D. Macho, and C. Nadeu, “On real-time mean-and-variance normalization of speech recognition features,” *IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1, pp. 773–776, 2006.

[134] A. Cooper, R. Reimann, and D. Cronin, “About face 3: the essentials of interaction design,” *Indianapolis. Wiley*, 2007.

[135] I. Fajardo, J. Cañas, A. Antolí, and L. Salmerón, “Accesibilidad Cognitiva de los Sordos a la Web,” 2002. [Online]. Available: http://www.anobium.es/docs/gc_fichas/doc/CJMUafknpu.pdf. [Accessed: 14-Nov-2015].

[136] D. Smith, L. Ma, and N. Ryan, “Acoustic environment as an indicator of social and physical context,” *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 241–254, Oct. 2005.

[137] W. Wu, S. Dasgupta, E. E. Ramirez, C. Peterson, and G. J. Norman, “Classification accuracies of physical activities using smartphone motion sensors.,” *J. Med. Internet Res.*, vol. 14, no. 5, p. e130, Jan. 2012.

[138] J. Kropf, L. Roedl, and A. Hochgatterer, “A modular and flexible system for activity recognition and smart home control based on nonobtrusive sensors,” in *Proceedings of the 6th International Conference on Pervasive Computing Technologies for Healthcare*, 2012, pp. 245–251.

[139] L. Chen, C. D. Nugent, and H. Wang, “A Knowledge-Driven Approach to Activity Recognition in Smart Homes,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 961–974, Jun. 2012.

[140] M. Louter, W. C. C. A. Aarden, J. Lion, B. R. Bloem, and S. Overeem, “Recognition and diagnosis of sleep disorders in Parkinson’s disease.,” *J. Neurol.*, vol. 259, no. 10, pp. 2031–40, Oct. 2012.

[141] M. Borazio and K. Van Laerhoven, “Combining wearable and environmental sensing into an unobtrusive tool for long-term sleep studies,” in *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics - IHI ’12*, 2012, p. 71.

[142] E. Visch-Brink, “Improvement of spontaneous speech in early stage Alzheimer’s disease with rivastigmine,” *J Nutr Heal. Aging*, vol. 13, pp. 34–38, 2009.