UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

Departamento de Electrónica

Máster Oficial en Sistemas Electrónicos Avanzados.
Sistemas Inteligentes

Tesis de Máster

Estudio, implementación y evaluación de un sistema de seguimiento de múltiples locutores usando fusión audiovisual

Frank Sanabria Macías

2010

UNIVERSIDAD DE ALCALÁ

Escuela Politécnica Superior

Departamento de Electrónica

Máster Oficial en Sistemas Electrónicos Avanzados.
Sistemas Inteligentes

Tesis de Máster

Estudio, implementación y evaluación de un sistema de seguimiento de múltiples locutores usando fusión audiovisual

Autor: Frank Sanabria Macías

Director/es: Marta Marrón Romera, Javier Macías Guarasa

Tribunal:

Presidente: D. Manuel Mazo Quintas.

Vocal 1º: D. Pablo Ramos Sainz.

Vocal 2º: D. Javier Macías Guarasa.

Calificación: .....

Fecha: .....

A mi familia.

Agradecimientos

A los tutores, Marta y Javi por la ayuda constante y el esfuerzo grande para que llegara hasta aquí,... no ha sido fácil, yo sé.

A mis compañeros Pacheco y Legrá por la amistad sincera y la compañía en las buenas y malas,..., mejores no quiero.

Al piquete del almuerzo + cafe + charla, Cano, Fernando, Isabel, Carlos y Susel por los momentos de relax al mediodía

A los que hicieron posibles mis viajes , profes y no profes de allá y de aquí,... ustedes saben

A la pandilla del ISPACE, Raquel, David, Alvaro, Jose y Jose Luis por la acogida, ..., que se repita

A Dani por estar siempre involucrado en el proyecto

A Yainet por los café,...sigue tan oportuna

En especial a Daileny, mi pareja, a mi mamá, abuela, Alfonso y el Charle, por aguantarme y apoyarme siempre,... no se cansen.

A mi Papá, mi Tía , Tíos y resto de la familia por estar lejos y cerca a la vez, siempre.

Índice general

Índice de figuras

Índice de tablas

Parte I

Resumen

Resumen

En este trabajo se ha diseñado, implementado y evaluado un sistema de seguimiento de locutor usando fusión audiovisual. La información de audio y vídeo es obtenida a partir de *arrays* de micrófonos y cámaras de vídeo situados en el entorno. El sistema está compuesto por dos bloques que extraen información de audio y vídeo y esta información es mezclada en un plano de ocupación, paralelo y a una altura "h" del suelo. Un filtro de partículas modela la dinámica de dicha mezcla, obteniendose finalmente la localización estimada del locutor en cada instante de tiempo. Como bloque de audio se implementa un algoritmo de detección de actividad acústica por sectores (volúmenes cónicos alrededor de cada *array*). Posteriormente, se busca en el interior de las intersecciones de los sectores activos de todos los *arrays* y el plano de ocupación, la región de máxima potencia acústica, usando el algoritmo *Steered Response Power*. El bloque de visión detecta rostros humanos en las imágenes de las cámaras de video, con una versión multi-pose del algoritmo *Viola and Jones*. Las proyecciones de dichas detecciones sobre el plano de ocupación, son combinadas con operadociones lógicas. El sistema fue evaluado usando la base de datos AV16.3.

Palabras Claves: Fusión Audiovisual, Seguimineto de Personas, Detección de Actividad Acústica, Detección de Rostros, Filtrado de Partículas.

Parte II

Abstract

In this work, we present a multimodal system for tracking a single user in a smart room environment. The information is extracted from several arrays of microphones and video cameras. The system is based on three blocks: two of them process audio and video information over a 2D ocupation grid with height "h" while a particle filter do the tracking task with audiovisual measures. The audio block discretize the physical space around each microphone array into a few sectors, and for each time frame, it determines which sectors contain active acoustic sources. Intersections betwen all active sectors and the ocupation plane define active areas for an Steered Response Power algorithm to search for maxima. The video block runs a multipose face detector based on Viola and Jones algorithm in every camera images. Faces detected are projected to the ocupation plane and mixed on it. The system has been evaluated with the IDIAP AV16.2 database, showing promising preliminary results.

Key Words: Audiovisual Fusion, People Tracking, Acustic Source Detection, Face Detection, Particle Filtering

Parte III

Memoria

Capítulo 1

Introdución

En este capítulo de introducción se explicará de forma general los objetivos que se pretenden con la realización de la presente Tesis de Máster y la justificación del mismo tanto en el entorno investigador en el que ha sido realizado como a nivel internacional.

1.1. Presentación

Esta Tesis de Máster se enmarca dentro del campo de los "Espacios Inteligentes", entornos dotados de un conjunto de sistemas sensoriales, de comunicación, y de cómputo inteligente, transparentes e imperceptibles para el usuario, pero que están continuamente percibiendo el entorno y cooperando entre ellos para proporcionar la ayuda necesaria a cada persona [1]. En este contexto, en que se persigue una interacción natural entre los sistemas y los usuarios, es imprescindible disponer de información precisa acerca de la existencia, posición y orientación (pose) de dichos usuarios, así como el estado habla/silencio de los mismos. Para ello es necesario realizar tareas de detección, localización y seguimiento automático de los mismos en la escena en cuestión [2].

Esta propuesta se inserta en una línea de investigación que viene desarrollando el Grupo de Ingeniería Electrónica aplicada a Espacios Inteligentes y Transporte del Departamento de Electrónica (GEINTRA) de la Universidad de Alcalá, orientada a la explotación de la información audiovisual generada por usuarios (locutores) en un entorno inteligente y captada por un sistema de sensores que incluyen arrays de micrófonos y de cámaras, con el fin de mejorar la capacidad de interacción entre el entorno y sus usuarios.

En este trabajo se pretende avanzar en las líneas tecnológicas principales que explotan la extracción y fusión de información multi-sensorial para mejorar las prestaciones de los actuales sistemas de detección, localización, seguimiento y estimación de pose de múltiples locutores, ubicados dentro de un espacio inteligente, dinámico y poblado por un número variable de objetivos [3] [2]. Como aplicaciones objetivo iniciales pueden citarse las referidas a la localización y seguimiento de hablantes en un espacio inteligente [4].

Esta Tesis de Máster tiene como antecedentes varios trabajos realizados en GEINTRA como son el Trabajo de Investigación Tutelado de Diego Alonso Jiménez [1] ("Localización, seguimiento y pose de múltiples interlocutores utilizando fusión audiovisual"), la Tesis Doctoral de Marta Marrón Romera [4] ("Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos") y los Proyectos Fin de Carrera de Carlos Castro González [5] ("Speaker Localization Techniques in Reverberant Acoustic Environments"), María Cabello Aguilar [6] ("Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos") [7] ("Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual"), y Eva Muñoz Herraiz [8] ("Diseño, implementación y evaluación de técnicas de localización de fuente y de mejora de la señal de habla en entornos acústicos reverberantes: aplicación a sistemas de reconocimiento automático de habla").

1.2. Motivación y Objetivos

El objetivo general de la Tesis de Máster es :

Diseñar e implementar un sistema de seguimiento de múltiples locutores utilizando fusión audiovisual.

Los objetivos específicos de la Tesis de Máster son:

- Implementar un bloque de detección y localización de rostros en vídeo basado en el algoritmo Viola and Jones [9].
- Implementar un sistema de localización de fuentes sonoras basado en una medida de actividad acústica por sectores [10].
- Diseñar e implementar el sistema de seguimiento basado en fusión audiovisual usando filtros bayesianos y combinando los bloques de detección de audio y vídeo [3].
- Evaluar los algoritmos implementados, usando las bases de datos multimodales disponibles en GEINTRA. Esta evaluación debe realizarse para diferentes condiciones reales e implica el análisis de la relevancia de los parámetros de control y la interpretación de los resultados obtenidos para llegar a conclusiones sobre la validez de los algoritmos en comparación con los métodos de partida.

Los requisitos que debe cumplir el trabajo propuesto son los siguientes:

- Incorporar los procesos necesarios a un sistema de audio y vídeo existente o en desarrollo dentro de GEINTRA, con vistas a la mejora de rendimiento.
- Ser flexible en el sentido de permitir modificar con facilidad los parámetros de control disponibles en los algoritmos utilizados.
- Ser flexible en el sentido de permitir la cómoda incorporación y control de nuevos algoritmos de localización y seguimiento.
- Estar bien documentado para facilitar su utilización en futuros proyectos.
- Disponer de un software eficiente y robusto.

1.3. Estructura del Documento

Esta Tesis de Máster está formado por un total de cinco capítulos, cuyos contenidos se detallan a continuación:

- **Capítulo 1** *Introducción.* Se introduce el tema de seguimiento de locutores basado en fusión audiovisual en el contexto de los espacios inteligentes, los objetivos principales y secundarios de la tesis, se explican los puntos de partida en los que se ha basado todo el trabajo para terminar con explicación sobre la estructura del documento.
- **Capítulo 2** *Estudio teórico.* Se muestran los trabajos fundamentales del estado del arte sobre el tema en cuestión, y se presenta de forma general la propuesta desarrollada.
- **Capítulo 3** *Desarrollo algorítmico y herramientas.* Se tratan los aspectos teóricos y prácticos de los algoritmos desarrollados en la propuesta.
- **Capítulo 4** *Resultados experimentales.* Se muestran los resultados obtenidos mediante la aplicación de los desarrollos implementados, tanto en forma de tablas como de forma gráfica. Se analizan los factores que inciden en cada uno de los resultados, realizándose propuestas de mejora en cada caso.
- **Capítulo 5** *Conclusiones y trabajos futuros.* Se plantean las conclusiones obtenidas tras la finalización del trabajo, así como propuestas de continuación de la investigación en esta temática.

Capítulo 2

Estudio Teórico

2.1. Introducción

En este capítulo se introducen los aspectos teóricos fundamentales del desarrollo de esta Tesis de Máster con el objetivo de mostrar una panorámica del estado del arte de la tecnología de seguimiento del locutor usando fusión audiovisual y proponer una solución basada en trabajos previos.

2.2. Estado del arte

En la literatura científica existe una gran cantidad de trabajos que emplean la información procedente de un único tipo de sensores con el objetivo de localizar determinados objetivos dentro de la escena bajo análisis, como por ejemplo cámaras de vídeo [11] [12], agrupaciones (*arrays*) de micrófonos [13] [14], balizas infrarrojo, etc. Entre todos ellos, los dos primeros han sido los más empleados. En ciertas aplicaciones como videoconferencias "inteligentes", interfaz hombre-maquina, análisis automático de escenas, vigilancia, encuadre automático de cámaras, etc, en las que se ven implicadas fuentes sonoras, la utilización de esta información puede ser muy interesante a la hora de realizar la detección, localización y seguimiento de determinados objetivos dentro de la misma. Además, el habla es la forma natural del ser humano de comunicarse, por lo que resulta inmediato su utilización a la hora de facilitar la interacción hombre-maquina en el contexto de los espacios inteligentes.

La localización y seguimiento de locutores a partir de señales de audio únicamente, aún resulta poco precisa, especialmente cuando los *arreglos* de micrófonos que capturan dichas señales no cumplen ciertas características [1]. Además los espacios inteligentes son por lo general entornos cerrados donde está siempre presente el fenómeno de reverberación, que complica aún más la tarea, por demás en muchas ocasiones las fuentes sonoras se presentan con una relación señal ruido muy baja, imposibilitando completamente la tarea de localización. Por todo esto es común en este tipo de aplicaciones incluir cámaras de vídeo que aporten información adicional para resolver la tarea.

Las principales dificultades con que cuentan los sistemas basados en cámaras de vídeo son las oclusiones totales o parciales de los objetivos a seguir, producto del entorno o de otros objetivos, debido al carácter direccional de la luz. Por el contrario, las señales de audio, son poco direccionales, especialmente a baja frecuencia, por lo que puede servir de complemento cuando el sistema de visión no puede seguir al objetivo. Otro problema de los sistemas basados en visión son las incertidumbres propias del sistema de adquisición, cambios de iluminación (brillos, sombras, contraste),el ruido (característicos del sensor y la óptica) o la aparición de *outliers*, la ambigüedad de las imágenes obtenidas producto de la proyección de un espacio de tres a otro de dos dimensiones, las variaciones naturales de los objetivos (color, forma, tamaño, textura, relaciones con el resto, etc.), y la asociación o identificación de cada dato de posición obtenido con el objetivo del que ha sido extraído, además del gran volumen de información que generan y por ende su difícil manejo.

Dado que cada tipo de sensores posee sus propias fortalezas y debilidades, la comunidad científica se ha planteado recientemente la utilización combinada de varios de ellos como una solución que aumente la robustez de los algoritmos de detección, localización y seguimiento implementados, denominando al resultado "seguimiento multimodal". Dentro de este tipo de seguimiento el más común es combinar audio y vídeo.

Estos tipos de aplicaciones tienen su desarrollo fundamentalmente en dos áreas. La primera es el procesamiento de la información audiovisual, por ejemplo, localización del locutor o locutores

para mejorar la relación señal a ruido de las señales de voz a través de la orientación adecuada de los *arrays* de micrófonos, selección o encuadre de las cámaras hacia el sujeto que se encuentre hablando en un momento dado [15], lo que además de permite identificar las personas en la escena empleando las dos fuentes de información. La segunda está en el campo del análisis de la interacción humana, en el cual es importante analizar patrones de comunicación verbal y no verbal (gestos, miradas, expresiones faciales entre otros), además de correlaciones entre acciones del locutor y conductas del grupo, para los cuales el seguimiento resulta necesario. Además se prevé un incremento de estas aplicaciones en la medida en que aumente la precisión de los algoritmos.

2.2.1. Técnicas de localización basada en audio

Para la detección, localización y seguimiento basada en audio se utilizan *arrays* de micrófonos, situados en unas posiciones espaciales conocidas dentro del entorno, de tal manera que tengan la capacidad de filtrado espacio-temporal, de las fuentes sonoras presentes en el mismo.

2.2.1.1. Detección

Los métodos de detección se utilizar para segmentar tramos de señal donde se encuentra la parte de interés, en este caso la voz humana. Los detectores de voz tradicionales o *Voice Activity Detectors* (VAD) [16] emplean características individuales del canal para calcular las métricas, como por ejemplo niveles de energía, cruces por cero, etc. A partir de éstas se aplican reglas de clasificación basadas en umbrales fijos o recalculados en los períodos de silencio. Esta forma de actuar hace que estos algoritmos presenten problemas en entornos con baja relación señal a ruido, especialmente con ruidos no estacionarios

2.2.1.2. Localización

El objetivo de la localización es ubicar en el espacio de coordenadas tridimensional del respectivo las fuentes de actividad (voz) en intervalos de tiempo pequeños. Aunque dentro de las técnicas de localización existes varias estrategias, las más usadas en aplicaciones prácticas son aquellas que se apoyan en las diferencias entre los tiempos de llegada de las ondas acústicas producidas por una fuente a los distintos micrófonos.

Los métodos que se basan en este concepto, requieren normalmente un conocimiento muy preciso de la geometría del *array* de micrófonos, pero no necesitan ningún conocimiento particular del entorno. Entre los algoritmos más importantes están los que emplean *Time Difference Of Arrival* (TDOA), y los que utilizan *Steered Response Power* (SRP). Los que usan TDOA tratan de estimar en un primer paso las diferencias de estos tiempos de llegada, para luego calcular la posición de la fuente sonora a partir de relaciones geométricas. Los que se basan en SRP hacen uso de técnicas de orientación del patrón de radiación del *arrays* (*beamforming*), para "escanear" una serie de localizaciones espaciales conocidas, en cuyo caso a la salida del algoritmo se conoce como *Steered Response* [17].

La desventajas de los métodos basados en SRP, es que su precisión depende del número de localizaciones elevadas. Esto implica un alto costo computacional cuando el espacio de búsquedas crece y limita la posibilidad de ejecución en tiempo real. Una de las alternativas para solucionar esta dificultad consiste en dividir el espacio de búsqueda en sectores [18], [19], [10], estimar en cuales de estos hay actividad sonora y luego restringir la búsqueda a los sectores activos.

2.2.1.3. Pose

En el contexto de los *arrays* de micrófonos, la estimación de la pose se refiere a la obtención de la dirección hacia la que esta enfocada la fuente sonora, en este caso la orientación de la cabeza del locutor. La mayoría de los algoritmos de estimación de pose trabajan de manera cooperativa con robustos algoritmos de localización, lo cual se debe a que la estimación de la orientación de la cabeza es un factor que puede degradar considerablemente las características de dichos algoritmos de localización. Por ello estos algoritmos suelen trabajar en dos etapas: En la primera de ellas se trata de obtener la posición de la fuente y en la segunda la orientación de la misma. Entre las variantes presentes en la literatura existen varias basadas en algoritmos SRP [20], [21] y [22].

2.2.1.4. Seguimiento

Los métodos de seguimiento pueden ser vistos como una forma de filtrado de la posición instantánea estimada por el algoritmo de localización, de tal manera que generan una trayectoria espacial suavizada. Además sirve para resolver ocluciones y problemas de asociación. De esta manera si el algoritmo de localización genera una estimación de posición que representa un cambio abrupto en la trayectoria del objetivo, esta será automáticamente corregida por comparación con la trayectoria anterior y filtrada de modo que la nueva posición estimada esté de acuerdo con el comportamiento del objetivo en el pasado.

El proceso de seguimiento implica la definición de un espacio de estados [4], y de unos modelos de actuación y observación. En el vector de estados se identifican las variables estimadas a la salida del algoritmo de seguimiento (por ejemplo la posición espacial, el estado de habla y silencio y la pose del locutor), este se implementa a partir de las características de observación definidas (información seleccionada de la fuente) en el modelo de observación y con una evolución temporal caracterizada por el modelo de actuación.

La diferencia fundamental entre el seguimiento y la localización de objetivos, consiste en que el primero de ellos explota la dinámica de los objetivos en busca de una mayor eficiencia, exactitud y robustez. El modelado de esta dinámica suele realizarse a través de un proceso autorregresivo (ARP) [1], pudiendo ser diferente en función del número de observaciones del espacio de estados que se utilicen para estimar el vector de estados.

Del estudio de la literatura científica relacionada, se desprende que el problema de seguimiento de múltiples objetivos abordado en el presente trabajo es un problema complejo no resuelto aún de forma general por la comunidad científica. La principal razón de esta complejidad radica en las difíciles características del entorno, que implica la división del problema en dos procesos interrelacionados [23]: La estimación del vector de estados de cada objetivo y la asociación de los mismos con alguno de los objetivos presentes en el entorno. Estos procesos se pueden llevar a cabo mediante algoritmos determinísticos como los Filtros de Kalman (KF) [24], probabilísticos como los Filtros de Partículas (PF) [25] [11] o una combinación de ambos [4].

2.2.2. Técnicas de localización basada en vídeo

Para la detección, localización y seguimiento basada en vídeo se utilizan *arrays* de cámaras de vídeo. Las cámaras son situadas en posiciones espaciales conocidas dentro del entorno, de tal manera que quede cubierto en el campo de visión de las mismas todo el espacio de interés(espacio inteligente).

2.2.2.1. Detección

Para detectar la presencia de locutores en cada imagen, se pueden utilizar algoritmos de detección de rostros dado la riqueza de características del rostro humano. Estos se basan fundamentalmente en reconocimiento de patrones de iluminación [9] ó en histogramas de iluminación, color y textura de la piel (*blobs*) [26]. Los primeros se caracterizan por ser más robustos frente a cambios de iluminación del entorno y son más complejos computacionalmente mientras que los segundos son menos sensibles a cambios de pose, lo cual puede ser conveniente para la detección, sin embargo se perdería información de pose.

2.2.2.2. Localización

Una de las alternativas en el proceso de localización con técnicas de visión es utilizar un algoritmo de *visual hull* [27], [28], en el cual se proyectan las zonas de interés detectadas en la imagen de cada una de las cámaras, a través del foco, como un volumen cónico de tres dimensiones. La intersección de los volúmenes generados por todas cámaras contiene el objeto de interés (Figura 2.1).

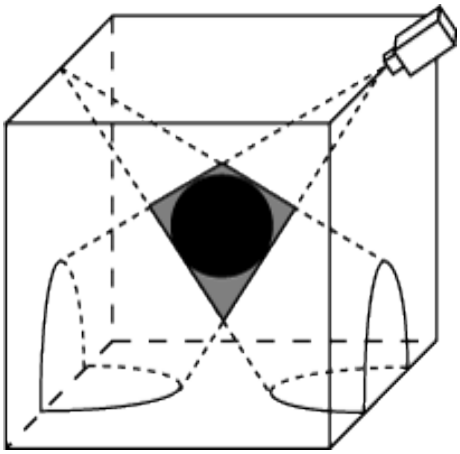


Figura 2.1: Esquema de *visual hull*.

La obtención del volumen completo del objetivo *visual hull* es computacionalmente costoso y no es necesario para tareas de localización y seguimiento. Por esta razón son comunes alternativas que solo obtienen la intersección de dicho volumen con un plano horizontal ubicado a determinada altura [29], [30], reduciendo de esta forma el tiempo de computo del algoritmo.

2.2.2.3. Pose

La estimación de la pose en el contexto de *arrays* de cámaras de vídeo [31], persigue el mismo objetivo que en caso de señales de voz: estimar la orientación de la cabeza de las personas presentes en el entorno. En este caso sin embargo el proceso se apoya con los algoritmos de detección, debido a que dicha orientación afecta el perfil del rostro visto por cada cámara, de modo que es necesario en muchos casos entrenar y ejecutar varias versiones del algoritmo para distintas orientaciones de rostros, de forma que una medida de verosimilitud permita la seleccionar la pose ganadora.

2.2.2.4. Seguimiento

El proceso de seguimiento de los objetivos con información de vídeo se realiza con las mismas herramientas que para el caso de audio [28] [32]. En este caso, también es posible el seguimiento de los múltiples objetivos en el plano imagen de cada una de las cámaras para luego estimar la localización.

2.2.3. Técnicas de localización basada en fusión audiovisual

Aunque desde hace algunos años se han venido presentado trabajos sobre localización y seguimiento de múltiples locutores con sistemas basados en audio o en visión en áreas de procesamiento de señales [33], [34], [13] y visión por computador [35], [36], solo en los últimos años se ha incrementado la atención a los sistemas que combinan estas dos fuentes [37], [38], [39], [40], [41], [42].

Estos trabajos se diferencian en tres aspectos fundamentales:

- El número de objetivos que persiguen, ya sea un único locutor, único locutor en presencia de varios o seguimiento de múltiples locutores,
- El marco de trabajo específico de detección y seguimiento usado.
- La configuración de los sensores usados: por ejemplo una única cámara con un par de micrófonos [37], [40], con más micrófonos [43] ó múltiples cámaras y micrófonos, calibradas [38] o no [44].

De todas las técnicas existentes los modelos probabilísticos generativos, aquellos basados en métodos de inferencias, exactas ó aproximadas [40], o basados en muestreo [38], [43], [44], parecen ser los más prometedores, dado sus principios de formulación y su rendimiento demostrado.

La principal clasificación de los algoritmos de fusión de audio y vídeo se refiere a cómo integran la información obtenida de las dos fuentes, existiendo dos clases de métodos, conocidos como “Orientados a Sistema” y “Orientados a Modelo” (Figura 2.2).

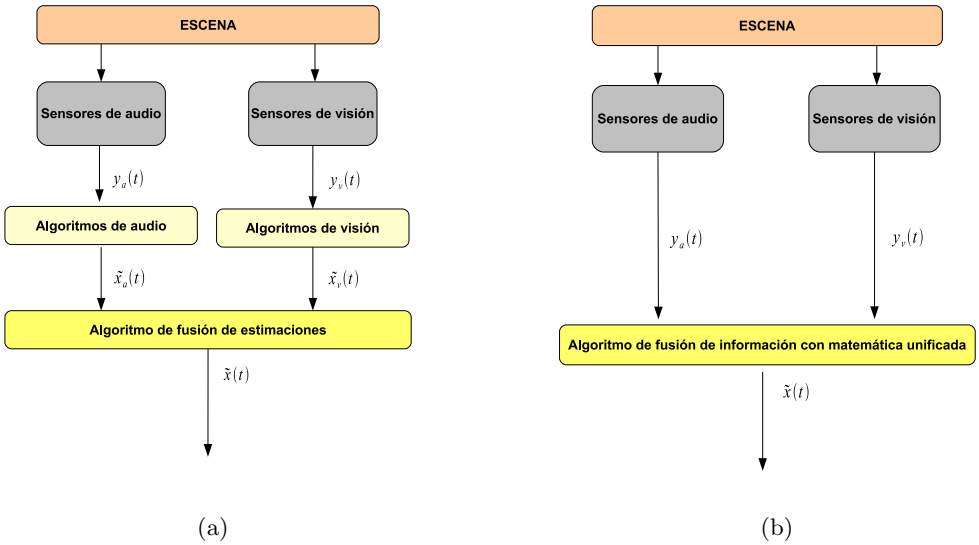


Figura 2.2: Clasificación de algoritmos de fusión audiovisual (a) Orientados a sistema, (b) Orientados a Modelo.

Los orientados a sistemas realizan énfasis en la integración de módulos de seguimiento con visión y audio independientes, aprovechando el desarrollo existente en algoritmos de seguimiento uni-modal (audio y video). A partir de la estimación arrojada por cada uno de ellos, se combinan dichas estimaciones para obtener la solución definitiva. Los orientados a modelo por el contrario, se basan en obtener una formulación matemática conjunta que aproveche al máximo las fortalezas de cada tipo de sensor en las diferentes etapas del algoritmo, de tal manera que se genere directamente una estimación óptima conjunta.

Entre los modelos orientados a sistemas el trabajo presentado en [45], describe un sistema basado en un dispositivo que integra un *array* de micrófonos circular pequeño y varias cámaras calibradas, cuyas vistas son combinadas en una imagen panorámica. El sistema, en el que cada persona es tratada independientemente, consta de tres módulos, uno para inicialización, usando o algoritmos estándares de localización de fuentes sonoras o pistas visuales, un segundo bloque de seguimiento basado en Modelos Ocultos de Markov, *Hidden Markov Model*, (HMM), y un último para verificar el seguimiento.

En [46] se describe un algoritmo de detección de múltiples locutores no probabilístico usando cámaras omni-direccionales (con limitaciones de resolución) y *array* de micrófonos, calibrados entre sí. En cada imagen del vídeo, el método extrae *blobs* de piel, y luego detecta la fuente sonora usando técnicas de *beamforming* en el conjunto de direcciones indicado por los *blobs* de piel.

En [47] se describe un sistema audio visual para múltiples locutores, basado en una cámara estereo y un *array* de micrófonos lineal, consistentes en tres módulos separados. Un módulo realiza el seguimiento de la cabeza de cada persona independientemente, basado en información visual. Otro módulo realiza la estimación de las direcciones de llegada, *Direction Of Arrival* (DOA) de las señales de audio. Y un módulo estima las de actividades síncronas de audio y vídeo. Para determinar el estado habla o silencio de los locutores se realiza una prueba de hipótesis basado en modelos estadísticos sencillos definidos a partir de las observaciones provenientes de los módulos anteriores.

En [48] se usan un conjunto de técnicas estándar en módulos separados, cuyos resultados son luego integrados en un sistema que estima la localización e identidad de cada uno de los participantes en entorno y detecta la persona que se encuentra hablando en cada momento usan cuatro cámaras calibradas, una de las cuales es omni-direccional (situada en el centro de una mesa) y un *arrays* de 16 micrófonos ubicado en un extremo de la mesa.

Como parte de los métodos de fusión "orientados a modelo", se han propuesto un conjunto de modelos generativos probabilísticos para inferir las localizaciones y el estado habla o silencio de múltiples interlocutores. Todos ellos están basados en filtros de partículas (PF) [42], [41], [49], [50] pero difieren en cuanto al espacio de estados, el modelos dinámico, el de observación y la técnica de muestreo. En [42] se usan dos cámaras calibradas y cuatro *arrays* lineales de micrófonos en la pared, y se basó en el modelo inicialmente propuesto en [35], el cual define un único espacio de estados que modela múltiples locutores, donde el número de estos puede variar en el tiempo. Un modelo de observación de cuerpo entero definido por dos términos, uno para vídeo, obtenido a partir de los *píxels* resultantes de aplicar un esquema de substracción del fondo, y uno para audio, obtenido del conjunto de transformadas de Fourier de corta duración del señales de cada micrófono. El PF se sustenta en un muestreo de importancia básico o ponderación de las partículas(IS de sus siglas en inglés), que resultan ineficiente a medida que se incrementa el número de personas en la escena [2].

En [41] se usa la misma configuración de sensores que en [45], y se plantea el seguimiento de múltiples interlocutores con un conjunto de PF independientes(uno para cada persona). Como

muestreo, los PF usan una mezcla de distribuciones específicas, que se obtienen de la salida de un seguidor de pistas simple (basado en audio, color, o información de forma). Esta propuesta incrementa la robustez en caso de fallo del seguidor más simples. Además, se asume que cada modelo de observación de un único objeto está factorizado sobre todas las pistas.

En [49] se usa una cámara estéreo y un *array* circular de 8 micrófonos, usa plantea el uso de un PF básico para realizar inferencias sobre un espacio estado de múltiples personas, asumiendo que el modelo de observación puede ser factorizado por participantes. El método sólo fue aplicado a escenas con dos personas, al parecer debido a las conocidas limitaciones del filtro de partículas básico.

El trabajo desarrollado en [50] presenta un método específico para salas de reuniones consistente en tres cámaras no calibradas cubriendo todo el espacio físico con un solapamiento mínimo en sus campos de visión, y un *array* circular de 8 micrófonos ubicado en el centro de la mesa. El modelo de fusión define un espacio de estados para múltiples locutores, que integra un modelo de oclusión de pares de personas a través de un MRF previo al modelo dinámico multi-objeto. Para resolver los problemas de los PFs tradicionales en el manejo de los espacios estado de muchas dimensiones, las inferencias en este modelo son realizadas por un PF de tipo Cadenas de Markov de Monte Carlo,*Markov Chain Monte Carlo*(MCMC),que tienen alta eficiencia de muestreo [51]. El espacio de estados integra datos de audio y vídeo a través de un modelo de observación donde las observaciones de audio son obtenidas con algoritmos de localización de audio y de observaciones visuales basadas en modelos de la forma y estructura espacial de la cabeza humana.

Este último modelo tiene dos ventajas sobre los presentados en [42], [41]. Primero incorpora de manera explícita un término relacionado con la interacción de pares de personas, que es especialmente útil para el manejo de oclusiones. Segundo, usa la técnica de muestreo MCMC, que permite seguir varios objetos de manera eficiente mientras preserva la formulación rigurosa del espacio de estados conjunto.

2.3. Propuesta Desarrollada

En el presente trabajo se propone una idea similar a la utilizada por [29], en la que se construye un mapa de ocupación en base a la información proporcionada por un conjunto de cámaras con coberturas solapadas. La construcción de un *grid* o mapa de ocupación resulta una manera eficiente de combinar la información de cámaras que se encuentran alejadas entre sí, no dependiendo de métodos que requieran complejas correspondencias entre imágenes o de modelos tridimensionales de objetos . Además, este tipo de técnicas admite la utilización de filtros multi-modales para la detección de múltiples objetos, como pueden ser PF, aseguran la coherencia espacial y temporal [30].

Como en [52] se combinan el *grid* de visión con otro mapa de "actividad sonora" generado a partir de un conjunto de sensores acústicos. El mapa de ocupación en este trabajo consiste en un plano paralelo al suelo (x, y) a una altura constante $z = h$, que se extiende sobre todo el espacio de busqueda (Figura 2.3).

La altura "h" del plano se ha de seleccionar de modo que coincida aproximadamente con la fuente de actividad, en este caso la boca de los locutores, con el fin de que los *grid* de audio y vídeo generen información coherente. Por ello su utilización se orienta a aplicaciones como reuniones (*meetings*) o conferencias en las cuales se espera que la altura del locutor permanezca dentro de un rango de variación pequeño.

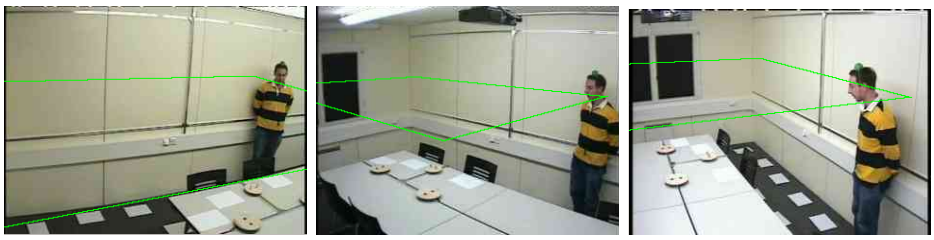


Figura 2.3: Plano de Ocupación sobre tres cámaras

El *grid* de actividad sonora se genera aplicando el algoritmo SRP, precedido de un bloque de detección basado en sectores (conjunto de regiones definidas en el espacio de busqueda) activos, propuesto en [10] con la finalidad de reducir el espacio de busqueda a solo a las regiones donde se estime existe una fuente de actividad sonora. Además este algoritmo permitirá no realizar el computo del *grid* en las tramas (*frames*) de silencio (ningún sector activo), con la consecuente reducción de carga computacional. En este sentido la detección basada en sectores puede considerarse un detector de actividad de voz(*Voice Active Detector* VAD).

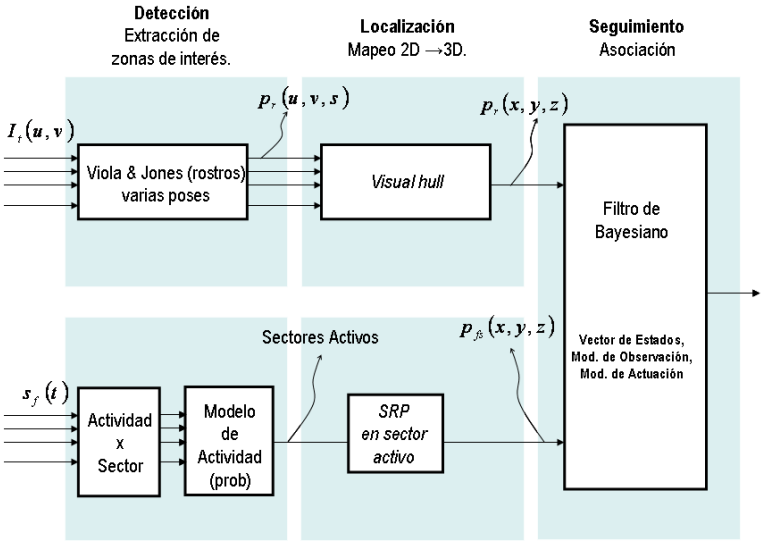


Figura 2.4: Esquema de fusión audiovisual propuesto

El modelo de actividad consiste en un umbral constante, seleccionado a partir de datos de entrenamiento. Esta es una primera aproximación al problema de detección y localización conjunta de actividad sonora, en el cual existen alternativas mas complejas como la propuestas por [10] donde se estima el umbral de manera adaptativa, ajustando un modelo de cinco variables, y las medidas de actividad de su métrica. Los sectores se distribuyen de forma esférica alrededor de cada uno de los *arrays* de micrófonos. y se definen de manera uniforme acuerdo a un valor de ancho en azimut, elevación y radio determinado. el modelo de actividad estimará un índice de la presencia de actividad sonora (SSM) para cada unos de los sectores. En el caso de existencia de varios *array* de micrófonos. se considerarán zonas activas los volúmenes de intersección (en adelante intersecciones) entre sectores activos de distintos *arrays*, como ha sido sugerido en Lathoud [10]. El mapa de actividad se obtendrá en los puntos del plano que pertenecen al volumen de intersección determinado como activo.

Para la obtención del mapa de ocupación de vídeo es preciso detectar la presencia del locutor en el plano imagen de cada una de las cámaras. Para esto se utiliza una busqueda de rostros con

una variante multi-poses del algoritmo *Viola and Jones* [9]. Una vez detectados los rostros se realizara una proyección homográfica de la detección de cada cámara hacia el plano de ocupación. En este se realiza una combinación "*booleana*", con las detecciones proyectadas siguiendo el principio de *visual hull*. obteniéndose la región de ocupación en el plano.

La fusión del *grid* de ocupación y el *grid* de actividad sonora se realiza mediante la operación de unión de los dos conjuntos (OR lógico). Finalmente el resultado de la fusión se toma como entrada al algoritmo del filtro de partículas. que realizará el seguimiento de las medidas, de ocupación o actividad a lo largo del tiempo, entregando en cada instante la localización mas probable del locutor (Figura 2.4).

El sistema se ha insertado dentro de una arquitectura cliente servidor que se ha venido desarrollado en trabajos anteriores ([4], [30], [6]) por GEINTRA. Cada cámara de vídeo se encuentra asociada a un servidor que realiza la detección de rostros 2D y la proyección al plano de ocupación. El detector de audio se asocia a otro servidor en el que se construye el mapa de actividad sonora sobre el plano "h". El cliente solicita a los servidores de vídeo el mapa de ocupación y al de audio el mapa de actividad, realiza la combinación entre las proyecciones de todas las camaras, la fusión de ambos mapas y ejecuta el PF sobre los datos fusionados.

Capítulo 3

Desarrollo Algorítmico y Herramientas

3.1. Introducción

En este capítulo se introducen los aspectos teóricos fundamentales del desarrollo de esta Tesis de Máster con el objetivo de llevar a cabo una adecuada tecnología de seguimiento de un locutor usando fusión audiovisual.

3.2. Localización de fuentes sonoras basada en audio

En un contexto de Espacio Inteligente la localización de fuentes sonoras se realiza mediante el análisis de las señales de audio almacenadas por un conjunto de micrófonos ubicados en el entorno (*array* de micrófonos). La idea consiste en extraer características de las señales, que brinden información sobre la localización de la fuente de emisión. Estas características están relacionadas con las deformaciones que sufre la onda acústica en su propagación por el aire.

El aire se considera un medio no dispersivo, de modo que la velocidad de propagación (c) de una onda mecánica en dicho medio, no depende de la frecuencia de la misma, y sí de la temperatura del medio. Considerando un temperatura ambiente constante puede asumirse que la velocidad de propagación de la voz en el aire es constante (Ejemplo. a 25 grados, c= 343.13m/s).

Las fuentes de voz por su parte, pueden ser modeladas como fuentes puntuales, y debido a su velocidad de propagación constante, generan una onda esférica centrada en el punto de emisión. Dicha onda, pasado un tiempo alcanzará a los micrófonos ubicados en las distintas localizaciones del entorno, produciendo una recepción, cuya amplitud y fase dependerán de la ubicación relativa entre la fuente y cada sensor. En la figura 3.1 se muestra el esquema de una onda sonora generada en un punto l_{ps}, y su recorrido hasta un micrófono situado en un posición l_m.

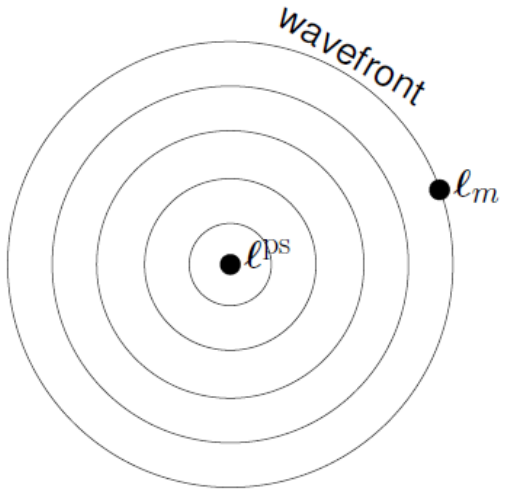


Figura 3.1: Modelo de onda acústica generada por una fuente puntual

Debido al modelo de onda esférica y despreciando la absorción de la onda por el medio, la variación de amplitud se describe de acuerdo a la ley del inverso del cuadrado de la distancia [1].

$$A\left(l_{ps}, l_m\right) = \frac{A_1}{\|l_{ps} - l_m\|} \tag{3.1}$$

Donde $A(l_{ps}, l_m)$ se refiere a la amplitud de la onda en la localización l_m generada con una amplitud inicial A_1.

Además, el modelo de onda esférica implica que el tiempo de vuelo (TOF: *time of flight*) entre la fuente y el sensor será proporcional a la distancia entre los mismos, como se muestra en la expresión 3.2.

$$TOF(l_{ps}, l_m) = \frac{\|l_{ps} - l_m\|}{c} \tag{3.2}$$

Donde $TOF(l_{ps}, l_m)$ se refiere al tiempo de vuelo de la onda desde la localización en que se generó l_{ps} hasta la localización del micrófono l_m.

En este modelo se ha asumido la condición de campo libre ó *free field*, que implica la ausencia de obstaculos, ó reflexiones. Esta condición no se cumple completamente en ambientes cerrados como los espacios inteligentes, dado que las paredes u otros obstáculos presentes en los mismos pueden producir múltiples reflexiones, absorción y dispersión en la onda acústica, siendo el modelado de la reverberación es aún un problema abierto en la literatura [53]. Por tanto es común asumir el modelo simple de campo libre y utilizar algoritmos que sean robustos ante este fenómeno.

Dado que la velocidad de propagación del sonido en el aire es relativamente baja, la mayoría de las aplicaciones prácticas de localización de fuentes sonoras aprovechan las pequeñas diferencias de las señales recibidas por los *array* de micrófonos.

Dichas diferencias se deben a las deformaciones de la señal de acuerdo al modelo anteriormente descrito y tendrán relación directa con la ubicación relativa entre los sensores(micrófonos) entre sí y con la fuente de voz. Aunque existen varias características de las señales que reflejan estas diferencias y que pueden ser usadas para estimar la ubicación de la fuente como la diferencias de nivel de las señales y la respuesta a impulso del local, los métodos más usados son aquellos que se basan en la diferencia de tiempos con que llega la señal a los distintos sensores.

3.2.1. Localización de fuentes sonoras a partir de agrupaciones de *array* de micrófonos

De estos métodos coexisten dos estrategias en general, la primera conocida como metodos de Diferencia de Tiempo de Llegada ó *Time difference Of Arrived*(TDOA), consisten en tratar de estimar las diferencias de tiempo de llegada entre cada par de sensores y luego a partir de relaciones geométricas, obtener la posición de la fuente, mientras que la segunda, conocidos como métodos directos, intentan estimar la localización directamente a partir de las señales, haciendo un barrido de todas las localizaciones donde pudiera estar la fuente. La principal problemática del primer grupo de métodos está en que el fallo de una estimación puede dar al traste con el resultado final,

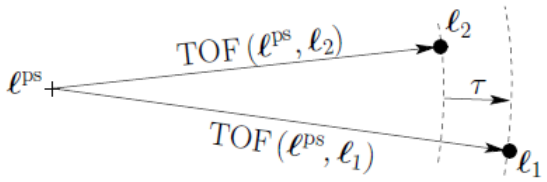


Figura 3.2: Diferencia de tiempo de arribo de un señal a dos micrófonos (TDOA)

En la figura 3.2 se muestra el esquema de una onda esférica y los tiempos de vuelo de la misma hasta dos localizaciones l_1 y l_2. También se muestra la diferencia de tiempos de llegada a dichas localizaciones *tau*

Entre los algoritmos de localización de fuentes de voz con estimación directa más eficaces está el *beamforming* o *Steered Response Power* (SRP), el cual es menos sensible a los efectos de la reverberación en señal de banda ancha como la voz. La idea en SRP es estimar la actividad sonora en cada punto del espacio, "orientando" el patrón de radiación del *array* hacia el punto a analizar. Esto se logra compensando las diferencias de tiempos de vuelo teóricas del punto bajo análisis y cada par de micrófonos.

3.2.1.1. Stered Response Power

En la práctica SRP se obtiene a partir de la función de Correlación Cruzada Generalizada ó *Generalized Cross Correlation* (GCC) entre cada par de micrófonos. De la diferencia de tiempo de vuelo teórica entre la posición evaluada como posible fuente de señal y la posición de un par de micrófonos se obtiene la demora τ, valor con el cual se evalúa la función GCC. Sumando el aporte de cada par de micrófonos se genera un mapa de actividad del espacio de búsqueda, conformado por el conjunto de puntos evaluados. La detección de fuentes activas puede realizarse definiendo umbrales de actividad o buscando máximos locales entre los datos. Determinando así la(s) localización(es) donde se encuentra(n) la(s) posible(s) fuente(s). La ecuación 3.3 expresa la suma retardada de potencias SRP, de las señales registradas por un conjunto de N_m micrófonos.

$$P_{SRP-PHAT}\left(l, X_1(t), X_2(t), ..., X_M(t)\right) = \sum_{k=2}^{N_f+1} \left| \sum_{n=1}^{N_m} \frac{X_m^{(t)}}{|X_m^{(t)}|} e^{-j\pi \frac{k-1}{N_f} TOF(l,l_m)} \right|^2 \tag{3.3}$$

donde l es la localización evaluada, $X_1(t), X_2(t), \cdots$ y $X_M(t)$ son los espectros de las señales registradas por los micrófonos, ubicados en las localizaciones $l_1, l_2, \cdots, l_M$, n y k son los índices de micrófonos y frecuencias, N_f es el número de frecuencias.

Una de las principales desventajas de los métodos directos es que necesitan evaluar un número elevado de localizaciones con vistas a obtener mayor precisión en el resultado. Esto conlleva un alto costo computacional, que dificulta la posibilidad de efectuar la busqueda en tiempo real. Para solucionar esta problemática se ha planteado la necesidad de reducir el espacio de busqueda. Una de las alternativas sugeridas en [18] y [19] consiste en dividir el espacio de búsqueda global en sectores. Los sectores son un conjunto de volúmenes que forman una partición del espacio de búsqueda (Figura 3.3), de los cuales pueda determinarse si en los mismos existe la presencia de una fuente sonora (sector activo) o no (sector inactivo), para luego aplicar el método directo solamente en los sectores activos. Los métodos que aplican este tipo de estrategia se conocen como algoritmos basados en sectores [10]

Otra de las desventajas de los métodos directos esta en que presumen la presencia de una o varias fuentes de voz durante todo el tiempo de señal. Siendo la voz una señal intermitente, existirán muchos intervalos de tiempo donde no habrá presencia de voz. En estos intervalos no sería necesario efectuar la tarea de localización. Por esta razón es necesario combinar estos métodos con algoritmos de detección de actividad de voz (*Voice Active Detectors VAD*) los cuales tradicionalmente se basan en características de señales de un único canal como la energía o el número de cruces por cero y son sub-óptimos cuando se trata de tareas de localización. Los métodos basados en detección por sectores detectan implícitamente si existe presencia o no de actividad ("algún sector activo"/"ningún sector activo") con lo cual resuelven también esta problemática.

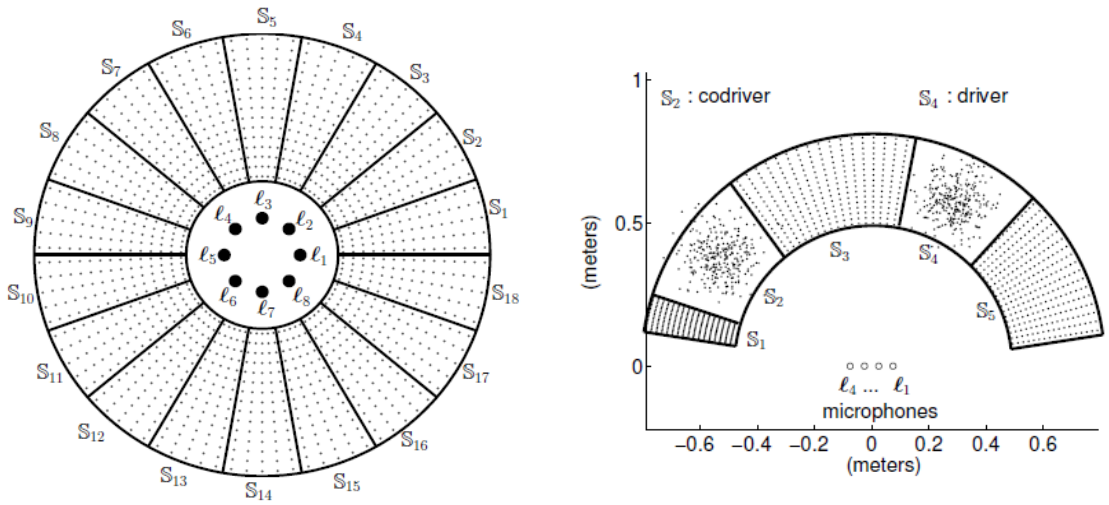


Figura 3.3: Sectores del espacio de búsqueda :(a) *Uniform Circular Array* (b) *Uniform Linear Array*

Entre los intentos que han seguido la filosofía de actividad por sectores están las propuestas por [54] y [55] que se basan esencialmente en calcular SRP en un punto central de cada sector. Esto puede causar problemas cuando la fuente se encuentra muy cerca de los bordes de dos o más sectores. En el trabajo de Lathoud [10] se propone un método en el cual se evalúa la actividad media en todo el volumen para caracterizar la actividad del sector, obteniéndose de esta forma una medida mas objetiva de la actividad acústica en los mismos. Las ideas esenciales propuestas en el trabajo de Lathoud son:

- La interpretación de SRP como una comparación entre las fases observadas y las fases teóricas correspondientes a una ubicación en particular Métrica en el Domino de fase ó *Phase Domain Metric* (PDM). Teniendo la particularidad de que la evaluación de la misma en un número elevado de localizaciones no influye en su tiempo de computo, permitiendo de esta manera promediar todo el volumen del sector sin dificulta contra la necesidad de tiempo real.
- Se presenta un modelo probabilístico no supervisado de la potencia acústica en un sector, basado en la métrica propuesta y con selección automática del umbral, permitiendo de esta manera realizar la tarea de detección adaptándose a diferentes condiciones del entorno.
- La métrica PDM se utiliza luego para realizar la búsqueda de la localización puntual de la fuente mediante un algoritmo de gradiente descendente.

En esta tesis de máster se ha planteado el objetivo de implementar inicialmente el módulo de detección por sectores, conjuntamente con la detección de las fuentes puntuales por el método del gradiente conjugado escalado (*Scaled Conjugated Gradient* SCG). A continuación se desarrolla su basamento teórico conjuntamente con aspectos de la implementación.

3.2.2. Detección basada en sectores

En los apartados 3.2.2.1,3.2.2.2 y 3.2.2.3 se exponen la fórmulas básicas del desarrollo teórico del método de detección basado en sectores propuesto por Lathoud, para más detalles ver [10]

3.2.2.1. Métrica en el dominio de fase *Phase Domain Metric*(PDM)

Dado un entorno con un conjunto de N_m micrófonos, sea $q \in \mathbb{N}$ el índice de cada par de micrófonos : $1 \leq q \leq N_q$, donde $N_q = N_m \frac{(N_m-1)}{2}$ es el número de pares de micrófonos. Sea $a_q \in \mathbb{N}$ y $b_q \in \mathbb{N}$ el índice de los dos micrófonos en el par $q : 1 \leq a_q < b_q \leq N_m$

Para una ventana de tiempo $t(frame)$ la señal recibida en los micrófonos a_q y b_q será $x_{a_q}(t)$ y $x_{b_q}(t)$. A un frecuencia discreta k, la suma retardada de potencias en frecuencias se puede definir alineando las señales con respecto a la fase teórica $u_q^{th}(k,l)$

$$E_q^{(t)}(k,l) = \left| X_{a_q}^{(t)}(k) + X_{b_q}^{(t)}(k) \cdot e^{j \cdot u_q^{th}(k,l)} \right|^2 \tag{3.4}$$

$$= \left| X_{a_q}^{(t)}(k) \right|^2 \cdot \left| 1 + \left| \frac{X_{b_q}^{(t)}(k)}{X_{a_q}^{(t)}(k)} \right| \cdot e^{j\left(-u_q^{(t)}(k) + u_q^{th}(k,l)\right)} \right|^2 \tag{3.5}$$

donde $X_{a_q}^{(t)}(k)$ y $X_{b_q}^{(t)}(k)$ son los espectros de las señales de los micrófonos a y b, l es una localización dada, $k \in \{1, \ldots, N_F\}$ es el vector de frecuencias discretas, la fase observada se define como:

$$u_q^{(t)}(k) = \angle X_{a_q}^{(t)}(k) - \angle X_{b_q}^{(t)}(k) \tag{3.6}$$

y la fase teórica $u_q^{th}(k,l)$ como:

$$au_q^{th}(k,l) = -\pi \cdot \frac{k-1}{N_F} \cdot \tau_q^{th}(l) \tag{3.7}$$

siendo $\tau_q^{th}(l)$ es el TDOA teórico entre una localizació l y el par de micrófonos q.

Asumiendo que las magnitudes de las señales son similares $|X_{a_q}^{(t)}(k)| \approx |X_{b_q}^{(t)}(k)|$, se puede transformar la expresión 3.5 en:

$$E_q^{(t)}(k,l) \propto 1 - \sin^2 \left[\frac{-u_q^{(t)}(k) + u_q^{th}(k,l)}{2} \right] \tag{3.8}$$

Teniendo en cuenta que la maximización de la expresión 3.8 es equivalente a minimizar el termino $\sin^2 \left[\frac{-u_q^{(t)}(k) + u_q^{th}(k,l)}{2} \right]$, Lathoud [10] propone una métrica en el dominio de fase para evaluar los vectores de fase observada y teórica de un conjunto de micrófonos.

$$d(\mathbf{u}, \mathbf{u'}) = \sqrt{\frac{1}{N_q} \sum_{q=1}^{N_q} \sin^2 \left(\frac{u_q - u'_q}{2} \right)} \tag{3.9}$$

donde $u^{(t)}(k) = \left[u_1^{(t)}(k), \cdots, u_q^{(t)}(k), \cdots, u_{N_q}^{(t)}(k) \right]$ y $u^{th}(k,l) = \left[u_1^{th}(k,l), \cdots, u_q^{th}(k,l), \cdots, u_{N_q}^{th}(k,l) \right]$ son los vectores de fases observadas y teóricas para cada par de micrófonos.

Lathoud [10] demuestra también que existe una equivalencia exacta entre maximizar SRP-PHAT y minimizar PDM $d^2(\cdot,\cdot)$ en un punto del espacio $\Re^3$. Con el objetivo de tener un índice que evalúe todo el sector Lathoud [10] propone calcular el valor RMS de la métrica PDM sobre

todo el sector. Esta operación resulta equivalente a calcular el promedio SRP-PHAT sobre el sector por lo cual se le denomina *Average Delay Sum Power* (ADSP) . Como la evaluación de todo el espacio continuo $\Re^3$ no es posible, se define una aproximación discreta $\hat{\overline{D}}_{\check{s}}^{(k)}$, en la cual se evalúan un número finito de puntos del volumen del sector $\check{s}$, lógicamente la aproximación será más precisa cuantos mas puntos se evalúen.

Para un sector S_s, en un *frame* t, y a una frecuencia k, la expresión discreta de la media cuadrática se define como :

$$\hat{\overline{D}}_{\check{s}}^{(k)} = \sqrt{\frac{1}{N_v}\sum_{n=1}^{N_v}\left[d\left(u^{(t)}(k), u^{th}(k, v_{\check{s},n})\right)\right]^2} \tag{3.10}$$

donde $\{v\} = \{v_{\check{s},1}, \cdots, v_{\check{s},n}, \cdots, v_{\check{s},N_v}\}$ es el conjunto de todas las localizaciones. Realizando transformaciones algebraicas a la expresión anterior se obtiene:

$$\left(\hat{\overline{D}}_{\check{s}}^{(k)}\right)^2 = \frac{1}{2N_v}\sum_{q=1}^{N_q}\left\{1 - \Re\left[e^{ju_q^{(t)}} Z_{\check{s},p}^{*}(k, \{v\})\right]\right\} \tag{3.11}$$

donde el término $Z_{\check{s},p} \in C$ se define como

$$Z_{\check{s},p}(k, \{v\}) = \frac{1}{N_v}\sum_{n=1}^{N_v} e^{ju_q^{th}(k, v_{\check{s},n})} \tag{3.12}$$

La expresión 3.11 resulta más conveniente dado que la precisión de la aproximación depende del termino Z, el cual es independiente de la información extraida de las señales (fases observadas), por lo que es posible calcular a priori (*off line*) estos términos para cada sector. Debido a esto, se puede realizar una aproximación tan precisa como se desee, sin que esto implique un aumento del tiempo de computo en la fase en tiempo real*on-line*.

3.2.2.2. Comparativa entre Sectores ó *SAM-SPARSE-MEAN*

Uno de los problema que tiene la medida ADSP es que como otros métodos por *beamforming* su directividad es limitada, especialmente en bajas frecuencias. Por esta razón es común que los sectores adyacentes a un sector activo, tengan valores significativos en $\hat{\overline{D}}_{\check{s}}^{(k)}$, no siendo activos en si mismos. Además, observaciones estadísticas en estudios de habla humana *multi-party speech* [56] que muestran que a una frecuencia dada, una fuente de voz puede ser considerada dominante. Suponiendo que un sector con una fuente acústica de banda ancha como la voz sea dominante sobre los otros sectores en un número elevado de frecuencias, [10] propone una medida derivada de ADSP para evaluar la actividad de un sector, basada en un conteo de sectores dominantes por frecuencia, denominada *SAM SPARSE MEAN* (SSM).

A una frecuencia k, un sector se considera dominante si la medida $\hat{\overline{D}}_{\check{s}}^{(k)}$ es menor estricta que en el resto de las frecuencias de acuerdo a la siguiente expresión:

$$\check{s}_{min}(k) = argmin_{\check{s}}\hat{\overline{D}}_{\check{s}}^{(k)}(k) \tag{3.13}$$

El valor SSM ($\zeta_{\check{s},t}$) para un sector s consiste en el número de sectores dominantes que contenga.

$$\zeta_{\check{s},t} = \sum_{k=2}^{N_F+1} \delta\left(\check{s} - \check{s}_{min}(k)\right) \tag{3.14}$$

3.2.2.3. Extensión a múltiples *array* de micrófonos

Aunque en principio la ubicación de micrófonos en el espacio puede ser cualquiera, evaluaciones experimentales* han mostrado que las correlaciones entre pares de micrófonos muy distantes no aporta buenos resultados, por lo que es común la ubicación de estos en grupos cercanos llamados *arrays* de micrófonos. La utilización de un único array en muchos casos dificulta la tarea de localización. La combinación de varios *arrays* con distinta orientación es capaz de mejorar los resultados. El método de detección por sectores relativo a un *arrays* de micrófonos, es fácilmente extensible a un conjunto mayor de *array*.

En [10] se sugiere la designación de región activa a los volúmenes resultantes de la intersección entre sectores activos de los distintos *array*. De manera similar al principio de *Visual Hull* en el área de visión. 3.4

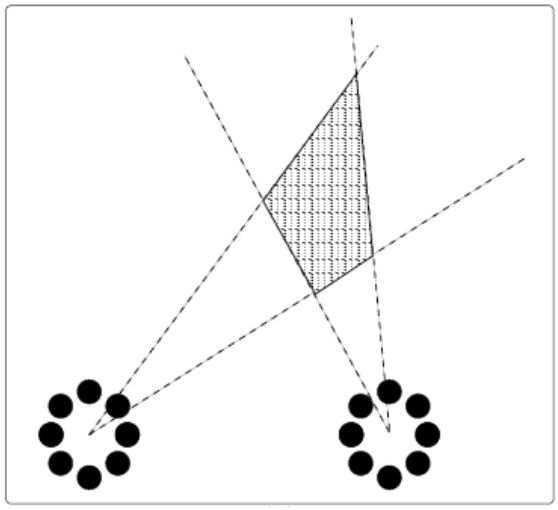


Figura 3.4: Intersección de Sectores

3.2.3. Esquema general del detector basado en sectores

Con lo visto, el sistema de detección de sectores queda dividido en dos fases, una realizada *off-line* que incluye el cálculo de los parámetros Z. En la fase *on-line* se procesan las señales hasta obtener la información de los sectores que se encuentran activos según cada *array*.

3.2.3.1. Fase *Off-Line*

En esta fase se generan los parámetros Z que serán usados en la fase *On-line* para obtener la medida de actividad SSM en cada sector. Para ello primeramente se definen la geometría de los sectores y el *grid* de localizaciones usado. Los "sectores" en este trabajo han sido definidos en coordenadas esféricas, considerando como origen el centro geométrico de los micrófonos del array. Cada sector ocupa un volumen caracterizados por los intervalos de azimut y elevación $[az_{min}\ az_{max}]$ y $[el_{min}\ el_{max}]$.

El *grid* se define también con distribución esférica y el mismo origen de coordenadas que el conjunto de sectores. Los parámetros que lo definen son $[r_{min}\ \Delta r\ r_{max}]$ $[az_{min}\ \Delta az\ az_{max}]$ y $[el_{min}\ \Delta el\ el_{max}]$. El conjunto de puntos es generados en coordenadas esféricas, luego transformados a coordenadas cartesianas y almacenados. A cada array de micrófonos se le asocia un *grid* esférico. Tanto los parámetros de configuración del *grid* como de los sectores se encuentran almacenados también en ficheros de configuración.

Para mantener compatibilidad con versiones anteriores en el sistema de localización basado en audio, la información de los puntos del *grid* (valor de x, y, z) se han almacenado de manera independiente a la información de pertenencia a los sectores. De esta forma se puede mantener el mismo *grid*, mientras se usan varias configuraciones de sectores, teniendo de esta manera mayor flexibilidad.

Dado un array de micrófonos, el cálculo de Z implica la obtención del tiempo de vuelo entre todos los micrófonos y todas las localizaciones $TOF(m_i, l_j)$, como indica el siguiente pseudocódigo.

Algorithm 1 Cálculo de los tiempos de vuelos

Require: conjunto de micrófonos $M = \{m_i/i \in N, 1 < j < N_m\}$, conjunto de localizaciones $L = \{l_j/j \in N, 1 < j < N_l\}$.

Ensure: Tiempo de vuelo entre todos los micrófonos y todas las localizaciones $TOF(m_i, l_j)$

```
1: for m_i ← 1 hasta número de micrófonos N_m do
2:    for l_j ← 1 hasta número de localizaciones N_l do
3:       TOF(m_i, l_j) ← dist(m_i, l_j)· f_s/c
4:    end for
5: end for
```

Luego por cada micrófono se promedian los desfases teóricos de todas las localizaciones pertenecientes al sector. El resultado es un valor dependiente del sector y la frecuencia $Z(q_i, f_i)$ que también es almacenado. .

Algorithm 2 Obtención de los parámetros Z de todos los sectores de un *array* de micrófonos

Require: Tiempo de vuelo entre todos los micrófonos y todas las localizaciones $TOF(m_i, l_j)$

Ensure: medida $Z(q_i, f_i)$ en todos los pares de micrófonos q_i y todas las frecuencias f_i

```
1: for todos los pares q_i = (m_{i_A}, m_{i_B}) de micrófonos do
2:    for f_i ← 0 hasta el número de puntos en frecuencia Nfft/2 − 1 do
3:       Z(q_i, f_i) ← 0
4:       for l_{i_s} ← 1 hasta número de localizaciones en el sector N_{l_s} do
5:          l_i ← locIndex(ilocSec)
6:          τ ← tof(m_{i_A}, l_{i_s}) − tof(m_{i_B}, l_{i_s})
7:          u_theo ← −π ∗ (f_i + 1) ∗ τ
8:          Z(q_i, f_i)+ = cos(u_theo) + isen(u_theo)
9:       end for
10:      Z(q_i, f_i) ← Z(q_i, f_i)/N_{l_s}
11:   end for
12: end for
```

Este procedimiento se repite para todos los *arrays*, generándose un fichero de parametros Z por cada *array*.

Algorithm 3 Cálculo de los parámetros Z de todos los *array*

Require: conjunto de micrófonos $M = \{m_i/i \in N, 1 < j < N_m\}$, conjunto de localizaciones $L = \{l_j/j \in N, 1 < j < N_l\}$

Ensure: parámetro $Z_{i,s}$ para todos los *arrays* i y sectores s

1: **for** $i \leftarrow 1$ hasta el número de *arrays* N_{arr} **do**
2: Calcular $TOF(m_i, \{l_j\})$ para todos los micrófonos del *arrays*
3: **for** $s = 1$ hasta el número de sectores N_s **do**
4: Calcular Z_s
5: **end for**
6: Copiar $Z_{i,s}$ a fichero
7: **end for**

3.2.3.2. Fase on-Line

En esta fase se procesan las señales de audio en bloque de muestras, llamados ventanas, tramas o *frames*. De cada *frames* se obtendrán las listas de las sectores activos de cada *array* y por último las intersecciones entre todos ellos.

3.2.3.3. Pre-procesamiento de las señales de audio

Todos los datos procedentes de cada micrófono pasan por una etapa de pre-procesamiento, asociada a la obtención del espectro de la señal. La señales son filtradas por un filtro de pre-énfasis $H(Z)$ con la propiedad de resaltar las componentes de alta frecuencia de la señal. Esto permite dar mayor importancia esta banda, la cual tiene mayor poder de localización dado que sus longitudes de ondas son más pequeñas.

$$H(Z) = 1 - 0{,}97Z^{-1} \tag{3.15}$$

Posteriormente la señal es segmentada en *frames* de 40 ms (640 muestras a $fs = 16kHz$). A cada segmento se le resta el valor medio y luego se le aplica una ventana de hamming para reducir los bordes del segmento de señal. Se añaden ceros al final de la señal para completar 1024 muestras, para finalmente calcular el espectro con una fft de 1024 muestras

3.2.3.4. Actividad por Sector

Para cada *array*, primeramente se calculan los valores de GCC-PHAT para todos los pares de micrófonos del *array* y para cada valor de frecuencia (f). Luego por cada sector s se obtiene la medida $ADSP(s, f)$ para luego realizar una búsqueda en cada frecuencia del sector con mayor valor ADSP (sector ganador) y contar en cada sector el número de frecuencias en que ha resultado ganador SSM.

Un sector es considerado activo si el número de frecuencias en que ha resultado ganador (SSM) supera un umbral dado. De esta forma por cada *array* se obtiene una lista de sectores activos L_{act}. La selección de este umbral será abordada en la sección 4.4.1.1 de este documento.

3.2.3.5. Intersecciones activas

Para obtener los volúmenes de intersecciones de sectores activos es necesario recorrer todas las combinaciones posibles con sectores de cada *array*, $N_{allcomb}$. N_{arr} es el número de *arrays* y

Algorithm 4 Detección de Sectores Activos

Require: señales de los m micrófonos $s_m(t)$, parámetro $Z_{i,s}$ para todos los *arrays* i y sectores s

Ensure: Lista de sectores activos por *array* L_{act}

1: **for** *frames* $t \leftarrow 1$ hasta el número de frames N_f **do**
2: **for** $m_i \leftarrow 1$ hasta número de micrófonos N_m **do**
3: Pre- procesamiento de los datos $s_m(t) \leftarrow w(H_z(s_m(t)))$
4: Cálculo del espectro $S_m(f) \leftarrow fft(s_m(t))$
5: **end for**
6: **for** todos los *array* i **do**
7: De todos los pares $q = (m_a, m_b)$ de micrófonos $GCC_{PHAT}(q, f) \leftarrow \frac{S_a(f)S_b(f)}{|S_a(f)S_b(f)|}$
8: $ADSP(s_i, f) =\leftarrow \sum_l^{N_{l_s}} \frac{1+\Re\{GCC_{PHAT}(q,f)\cdot\}\Re\{Z_s(f)\}+\Im\{GCC_{PHAT}(q,f)\}\cdot\Im\{Z_s(f)\}}{2N_q}$
9: $SSM(s_i) \leftarrow$ número de frecuencias f_i, con $Adsp(s_i, f_i) > Adsp(s_j, f_i)$ Donde $j \neq i$
10: **if** $SSM(s_i) > th_{actividad}$ **then**
11: $L_{act} \leftarrow s_i$
12: **end if**
13: **end for**
14: **end for**

Ns_j es el número de sectores del *array* j-ésimo.

$$N_{allcomb} = \prod_{j=1}^{N_{arr}} Ns_j \tag{3.16}$$

Cada intersección $I(s_1, \cdots, s_i, \cdot, s_{N_{arr}})$ está definida en un espacio N_{arr}-dimensional, donde s_i es un sector activo del i-ésimo *array*. Teniendo el cuenta que el número de *arrays* puede variar de un experimento a otro, surgió la necesidad de definir estructuras especiales para manejar esta dimensión variable. La estructura utilizada fue una lista unidimensional $L_{allcomb}$ que ordena todas las intersecciones, primero de acuerdo al orden definido de *arrays* y segundo por el orden definido de sectores en cada *array*, donde $s_{i,j}$ es el sector i-ésimo en el *array* j-ésimo. Esta estructura se genera en la fase *off-line*.

Lista de Intersecciones

$$\begin{aligned} L_{allcomb} = \{&(s_{1,1}, \cdots, s_{1,j}, \cdots, s_{1,N_{arr}}), \cdots, \\ &(s_{1,1}, \cdots, s_{i,j}, \cdots, s_{Ns_{N_{arr}},N_{arr}}), \cdots, \\ &(s_{Ns_1,1}, \cdots, s_{Ns_j,j}, \cdots, s_{Ns_{N_{arr}},N_{arr}})\} \end{aligned} \tag{3.17}$$

Una vez que se tiene las listas de sectores activos por cada *array*, se analizan todas las combinaciones posibles $N_{act-comb}$. A partir de los índices $\{s_1, \cdots, s_i, \cdot, s_{N_{arr}}\}$ de una combinación activa, se obtiene el índice de esta combinación en la lista de intersecciones.

$$N_{act-comb} = \prod_{array=1}^{N_{arr}} Nsa_j \tag{3.18}$$

3.2.4. Localización de las fuentes puntuales de actividad

Una vez detectadas las intersecciones de sectores activos, la siguiente tarea consiste en ubicar dentro de dichos volúmenes, el o los puntos más probables donde puede encontrarse la fuente de voz. Para ello se siguieron dos estrategias, una es el *beamforming* clásico (SRP-PHAT), evaluando solo en la región de las intersecciones activas, para luego buscar los máximos locales mientras la segunda fue utilizar un algoritmo de optimización según lo propuesto en [10].

SRP evalúa la actividad en cada punto del *grid* de actividad y obtiene un máximo mientras que en el SCG se parte de una posición inicial (centro geométrico de los puntos de la intersección) y se termina con una solución final correspondiente a un mínimo de la métrica PDM.

3.2.4.1. Búsqueda mediante SRP en la intersección de sectores

Al igual que en el caso de los parámetros Z, la evaluación de la potencia SRP-PHAT en el espacio de búsqueda, debe definirse en un conjunto discreto de puntos en dicho espacio. Como los *grid* generados para el cálculo de Z están asociados a cada *array* y tienen una distribución no regular (coordenadas esféricas), se generó un nuevo *grid* (*grid* de actividad), en este caso regular (coordenadas cartesianas) que cubre todo el espacio de búsqueda . Los parámetros que lo definen son $[x_{min}\ \Delta x\ x_{max}]$ $[y_{min}\ \Delta y\ y_{max}]$ $[z_{min}\ \Delta z\ z_{max}]$, al igual que los parámetros de configuración usados en el cálculo de Z estos parámetros se almacenan en ficheros de configuración.

Como primer paso se determinan cuales de los puntos del *grid* de actividad pertenecen a intersecciones activas. En la práctica los puntos del grid que pertenecen a cada intersección se obtienen a priori (*off line*) y son almacenados en memoria, a las que se accede una vez que se conocen las intersecciones activas.

El algoritmo de SRP usado corresponde a la implementación de [5], modificada para evaluar el subconjunto de puntos del *grid* de actividad que pertenecen a la intersección. Finalmente se realiza la búsqueda del máximo de actividad SRP en cada intersección y estas localizaciones conforman el resultado. En los casos en los que aparezca más de una intersección activa, los resultados son ordenados de mayor a menor potencia SRP-PHAT.

3.2.4.2. Búsqueda basada en la optimización de la métrica de fase y algoritmo de gradiente descendente

De acuerdo a la interpretación topológica de SRP que supone la métrica PDM, [10] propone una búsqueda de la ubicación final de la fuente mediante un algoritmo de optimización, el *Scaled Conjugate Gradient* (SCG) [57] usando como función de coste la métrica PDM. El SCG es más eficiente que otros en cuanto a velocidad debido a que sólo utiliza información de las derivadas de primer orden (gradiente).

La función de coste a minimizar, basada en PDM para un *array* de micrófonos se expresa:

$$\Delta\left(\left\{\mathbf{u}^{(t)}(k)\right\}, \Upsilon, \hat{l}_n^{(t)}\right) = \frac{1}{N_\Upsilon}\sum_{k\in\Upsilon} d^2\left[\mathbf{u}^{(t)}(k), \mathbf{u}^{th}\left(k, \hat{l}_n^{(t)}\right)\right] \tag{3.19}$$

Donde :
$\hat{l}_n^{(t)} = \left[\hat{x}_n^{(t)}, \hat{y}_n^{(t)}, \hat{z}_n^{(t)}\right]$, es la localización estimada $\left\{\mathbf{u}^{(t)}(k)\right\}$ y $\left\{\mathbf{u}^{th}\left(k, \hat{l}_n^{(t)}\right)\right\}$ son los vectores

de fases observadas y teóricas de todos los micrófonos y $\Upsilon \subset \{2, \cdots, N_F + 1\}$ el conjunto de frecuencias discretas estrictamente positivas

Desarrollando la ecuación 3.19 se obtiene:

$$\Delta = \frac{1}{2} - \frac{1}{2N_\Upsilon N_q} \sum_{k \in \Upsilon} \sum_{q=1}^{N_q} \Delta_{k,q} \tag{3.20}$$

$$\Delta_{k,q} = cos\left[u_q^{(t)}(k) - u_q^{th}(k, \hat{l}_n^{(t)})\right] \tag{3.21}$$

Así mismo, la componente del gradiente de la función de coste con respecto al eje x es:

$$\frac{\partial \Delta}{\partial \hat{x}_n^{(t)}} = \frac{1}{2N_\Upsilon N_q} \frac{\pi}{N_F} \frac{f_s}{c} \sum_{k \in \Upsilon} \sum_{q=1}^{N_q} \dot{\Delta}_{k,q} \tag{3.22}$$

Donde

$$\dot{\Delta}_{k,q} = \left\{ (k-1) sen\left[u_q^{(t)}(k) - u_q^{th}(k, \hat{l}_n^{(t)})\right] \left[\frac{\hat{x}_n^{(t)} - \hat{x}_{a_q}^{(t)}}{\left\| \hat{l}_n^{(t)} - \hat{l}_{a_q}^{(t)} \right\|} - \frac{\hat{x}_n^{(t)} - \hat{x}_{b_q}^{(t)}}{\left\| \hat{l}_n^{(t)} - \hat{l}_{b_q}^{(t)} \right\|} \right] \right\} \tag{3.23}$$

De la misma manera se pueden obtener las expresiones en las componente z e y. En [10] se sugiere en usar las mismas expresiones 3.20 y 3.22 algoritmo para el caso de múltiples *arrays*, considerando el conjunto de micrófonos como un único *array*. Sin embargo siguiendo resultados de experiencias previas en GEINTRA, la combinación de todos los pares de micrófonos aporta peores resultados que considerar solo las combinación de aquellos agrupados cerca (*array*) y luego combinar los resultados de cada unos de ellos. Es por ello que en este trabajo se propone un modificación a estas expresiones para el caso de múltiples *arrays*. La nueva función de coste se expresa como la suma del aporte de todos los *array*, $N_q(j)$ es el número de pares de micrófonos en el *array* j.

$$\Delta_m = \frac{1}{2} - \frac{1}{2N_\Upsilon N_q} \sum_{k \in \Upsilon} \sum_{j=1}^{N_j} \sum_{q=1}^{N_q(j)} \Delta_{k,q} \tag{3.24}$$

De la misma manera, el gradiente de la nueva función de coste:

$$\frac{\partial \Delta_m}{\partial \hat{x}_n^{(t)}} = \frac{1}{2N_\Upsilon N_q} \frac{\pi}{N_F} \frac{f_s}{c} \sum_{k \in \Upsilon} \sum_{j=1}^{N_j} \sum_{q=1}^{N_q(j)} \dot{\Delta}_{k,q} \tag{3.25}$$

El algoritmo SCG define constantes para finalizar la búsqueda por un número máximo de iteraciones o porque el error entre iteraciones disminuye por debajo de determinado umbral. El restado final en cada intersección, forma la salida del algoritmo.

3.2.5. Generación del *grid* de actividad

Para convertir la salida del bloque de audio en un mapa o *grid* de actividad se usó los valores de potencia SRP, calculada en las intersecciones activas de cada *frame*. Con estos se realiza el mismo procedimiento descrito en [7]:

- Normalización con el valor máximo de actividad
- Proyección de los valores normalizados [0,1] al intervalo [0,255]
- Umbralización los píxeles con valores $> th_1 = 250$.
- Dilatación de los regiones detectadas.
- Umbralización los píxeles con valores $> th_2 = 200$.
- Intersección AND de las imagen generadas.

Este proceso se muestra en las figuras 3.5 y 3.6:

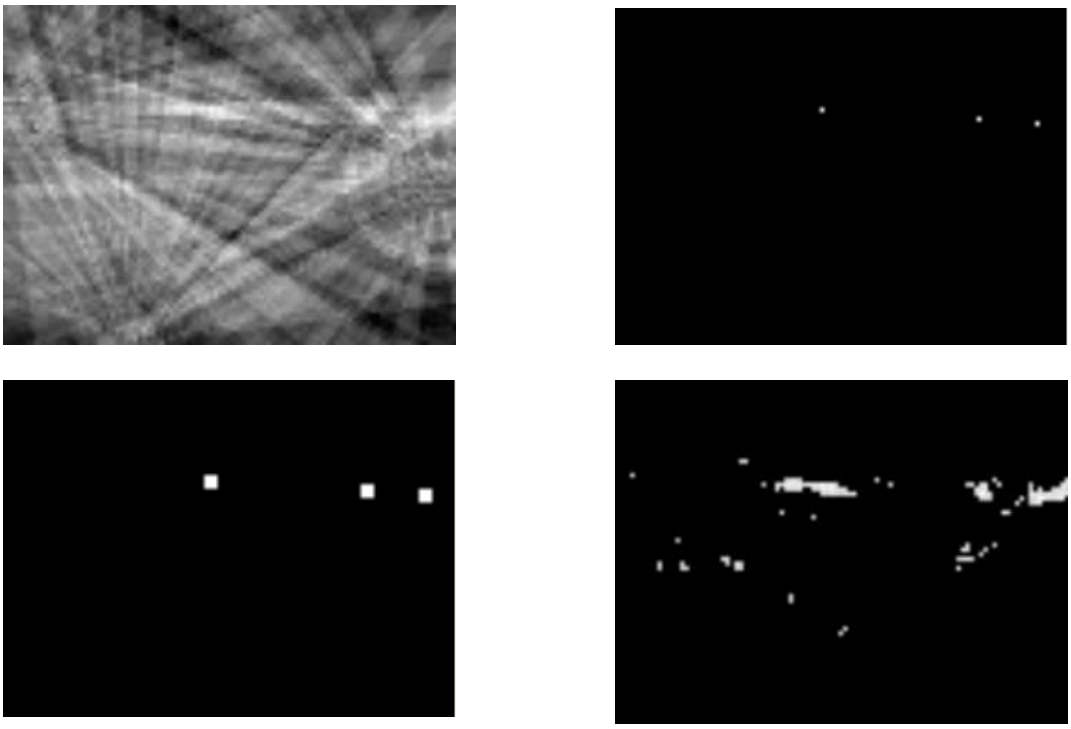


Figura 3.5: Arriba izquierda: imagen original. Arriba derecha: máximos de la imagen. Abajo izquierda: Puntos máximos dilatados. Abajo derecha: Imagen original filtrada a un nivel de gris de 200

Figura 3.6: Resultado de realizar la AND entre la imagen con los puntos dilatados y la filtrada a 200

3.3. Localización de personas basada en vídeo

Para localizar las posiciones de los locutores basados en la información que proporcionan las cámaras vídeo se realizará la detección de rostros en cada una de las cámaras. Luego se proyectan las zonas detectadas sobre el plano de ocupación a través de una homografía. Por último se combinarán las proyecciones de todas las cámaras para obtener las regiones de ocupación en el plano.

3.3.1. Detección de rostros, algoritmo Viola-and Jones

El algoritmo de Viola-Dones [9] es un algoritmo muy potente y popular en la detección de objetos en imágenes y especialmente en la detección de rostros. Es un método basado en apariencia, el cual se soporta en métodos de aprendizaje estadísticos que permiten construir un clasificador "rostro/no-rostro" (problema de reconocimiento de patrones con dos clases) en estos se busca patrones de luz y sombra que se asemejen a un rostro humano.

En esencia un detector basado en apariencia consta de tres bloques fundamentales. Una ventana de búsqueda o exploración (región rectangular de la imagen), esta determina la zona que posteriormente va a evaluar el clasificador, realizando en el caso más general una busqueda para diferentes escalas y posiciones. Si no existe ninguna restricción en cuanto a la zonas en las que deben aparecer los rostros en la imagen, lo común es incrementar la escala (tamaño de la ventana de evaluación) en un porciento (usualmente 10 ó 20 por ciento) y para cada escala desplazar la ventana tanto en horizontal como en vertical sobre la imagen analizada, en saltos proporcionales a la escala determinada. El clasificador evalúa las características de la imagen en cada ventana, determinando si coincide con la apariencia buscada. Por último un algoritmo combina soluciones que se solapen. Estas pueden aparecer producto de varios resultados positivos del clasificador en escalas y posiciones muy cercanas.

Algorithm 5 Algoritmo de detección basado en apariencia

Require: imagen I.

Ensure: Lista de rectángulos $\{(s_0, x_0, y_0), (s_1, x_1, y_1), \cdots, (s_N, x_N, y_N)\}$ clasificados como rostro

1: **for** s $\leftarrow$ Escala mínima (s_0), con salto de escala Δs hasta escala máxima(s_{max}) **do**
2: **for** $x \leftarrow x_{min} : \Delta x : x_{max}$ **do**
3: **for** $y \leftarrow y_{min} : \Delta y : y_{max}$ **do**
4: $D(s, x, y) \leftarrow$ clasificador($I[X : X + sW_w,\ Y : Y + sH_w\ ]$)
5: **end for**
6: **end for**
7: **end for**
8: Mezcla de soluciones solapadas

Las características que distinguen al *Viola and Jones* de otros detectores de rostros son los siguientes:

- El uso de un conjunto de características (*features*) simples de tipo Haar. Estas características sirven para detectar contrastes de iluminación (luz y sombra) y se definen como la suma de los *pixels* en las áreas en blanco y la resta en las áreas de negro (Figura 3.7). Su principal ventaja es que permiten ser evaluadas de manera eficiente.
- El uso de una cascada de clasificadores de complejidad incremental. Una arquitectura en cascada funciona como un árbol de decisión degenerado: En cada etapa el clasificador

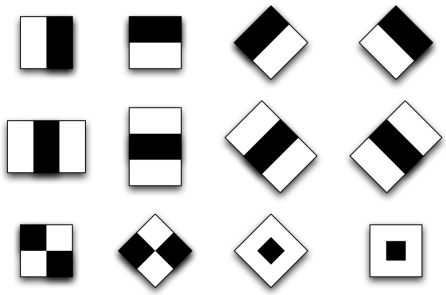


Figura 3.7: *features* Rectangulares de tipo Haar

rechaza la ventana y el proceso se detiene o la acepta y la ventana avanza a la siguiente etapa de la cascada. Cada etapa se diseña de manera que detecte todas los caras mientras rechace la mayor cantidad de "no-cara" posibles. En la Figura 3.8 se muestra el esquema de cascada de clasificadores usado por el algoritmo *Viola and Jones*.

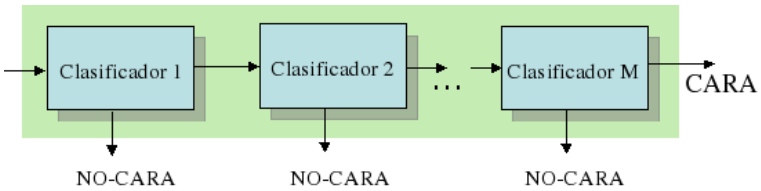


Figura 3.8: Cascada de clasificadores

La detección de un rostro en una imagen es un evento de baja probabilidad, considerando el numero de rectángulos evaluables en todo el espacio de escalas y posiciones (s, x, y) en la imagen. Además teniendo en cuenta que entre las ventanas que no contienen rostros, existen muchas que sean fácilmente detectables. Por ello se diseñan las primeras etapas del clasificador con pocos *features*,de modo que eliminar las ventanas de "no-rostro" mas evidentes, de esta manera se eliminan muchas ventanas con un costo computacional mínimo. Estas etapas se diseñan con razón de rechazo a falsos positivos baja (pasan muchos ventanas "no-caras") pero si una razón de detección alta (todas las caras deben pasar). Las ventanas de no-caras más difíciles de clasificar se eliminarán en etapas posteriores de la cascada.

- Cada etapa del clasificador se compone de una combinación de clasificadores "débiles" (clasificador simple, con bajo rendimiento), entrenados usando *boosting* [9]. La idea subyacente del *boosting* es combinar linealmente clasificadores "débiles" para construir un clasificador con mayor rendimiento (Figura 3.9). La selección de las *features*, así como la estimación de los pesos se obtienen en el proceso de entreamiento.

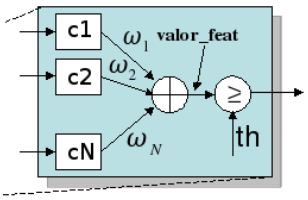


Figura 3.9: Etapa de clasificación. *Boosting* de clasificadores simples

En la Figura 3.9 se muestra un clasificador formado por una combinación de clasificadores

débiles $C1, C2, \cdots, CN$, usando en conjunto de pesos $\omega_1, \omega_2, \cdots, \omega_N$ y un umbral th.

Entre las dificultades de este algoritmo está que necesitan un número elevado de muestras para su entrenamiento, las cuales deben estar bien alineadas, para conseguir un buen resultado. Los clasificadores se entrenan para poses específicas ("frontal", "perfil", etc) teniendo poco margen de variabilidad de la pose para su detección.

Las diferentes poses que puede tener un rostro frente a una cámara en un contexto de seguimiento, como el planteado en este trabajo, generan la necesidad de tener algoritmos de detección robustos a las distintas poses. Para solucionar este problema existe un conjunto de alternativas en la literatura, entre las que están

- Esquema "en paralelo", este consiste en varios detectores, uno para cada pose y un mecanismo de votación que mezcle las salidas de los detectores. Esta estrategia tiene el problema de su alto coste computacional.
- "Estimador de pose + modelo simple", contiene un estimador de pose como etapa inicial, para luego aplicar un único detector entrenado para la pose estimada. Esta alternativa tiene la desventaja que una mala estimación de pose, de al traste con la detección.
- Los modelos piramidales van de lo general a lo particular, el nivel superior se entrena para detectar todas las vistas. En niveles inferiores se entrenan clasificadores de rangos de poses cada vez más pequeños.

Por las dificultades encontradas para el entreamiento de nuevas plantillas en este trabajo se optado por un modelo de detección en paralelo, usando dos plantillas, entrenadas para rostros frontales y otra de perfil respectivamente. Estas plantillas están disponibles en la librería de código libre OpenCV [58].

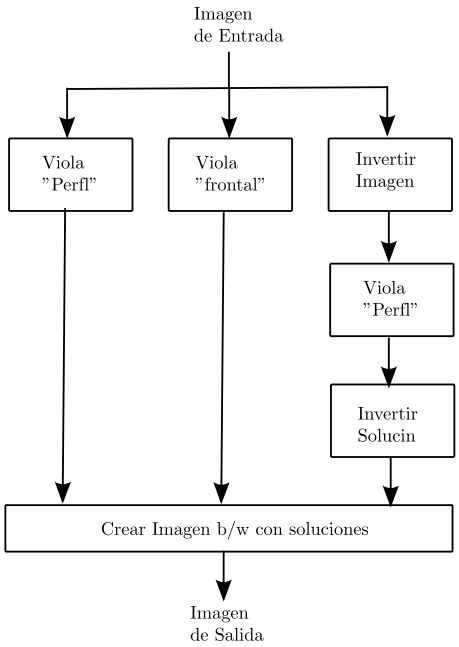


Figura 3.10: flujograma del algoritmo de detección de rostros

El algoritmo implementado (ver figura 3.10) realiza la llamada al método de *Viola and Jones* tres veces, uno con la plantilla frontal y dos con las plantillas de perfil, el primero de

ellos aplicado a la imagen original, y el segundo se aplica a la imagen despues de realizarle una reflexión con respecto su eje central vertical. De modo que se puede detectar rostros de perfil izquierdo y derecho. Despues, los resultados positivos generado por el algoritmo aplicado a la imagen invertida, se invierten para tenerlos en el mismo sistema de referencias .

Finalmente se construye una imagen en blanco y negro donde las ventanas, detectadas en cada pose se le asigna el color blanco, mientras que el resto se le asigna negro (ver figura 3.11).

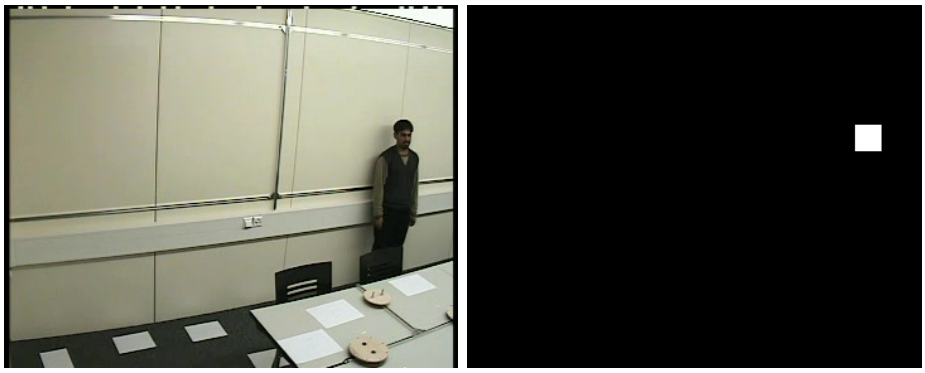


Figura 3.11: Resultados del algoritmo de detección de rostros (a) Imagen Original, (b) Resultado de la detección

3.3.2. Proyección de la localización por homografía al plano 2D (*Visual Hull*)

Las imágenes generadas por el algoritmo de detección se proyectan al plano de ocupación a partir de la relación de homografía entre todos los *pixels* de la imagen de detección y los *pixels* del *grid* de ocupación.

Dado un modelo de cámara *pin-hole* (ver figura 3.12) con sus correspondientes matrices de parametros intrínsecos K, rotación R, traslación t [59]; y dado el plano de ocupación π en el espacio 3D, la matriz de homografía, H, relaciona la posición de los puntos X_π del plano de ocupación π con los puntos X_I en el plano imagen Δ.

Partiendo de la relación entre los puntos del plano imagen X_I y los puntos de espacio X_w

$$X_I = K[R|T]X_w \tag{3.26}$$

$$\begin{bmatrix} cx_i \\ cy_i \\ c \end{bmatrix} = K\,[R|T] \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} \tag{3.27}$$

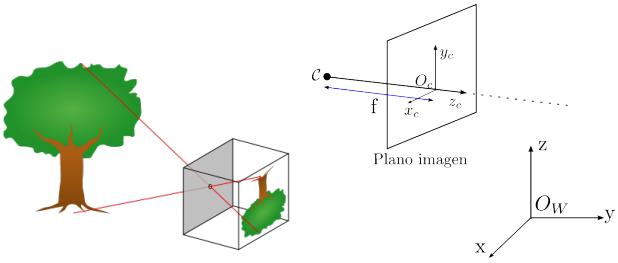


Figura 3.12: Modelo de Pin hole y ejes de coordenadas

donde c es un factor de normalización.

$$\begin{bmatrix} cx_i \\ cy_i \\ c \end{bmatrix} = KR \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + KT \tag{3.28}$$

Como se ha explicado anteriormente, en este trabajo se define el plano de ocupación π paralelo al suelo, a una altura "h". Debido a ello las coordenadas de los puntos del plano en el sistema de coordenadas 3D tendrán la forma $X_w = \begin{bmatrix} x_w & y_w & h \end{bmatrix}'$, sustituyendo en la ecuación 3.28 resulta:

$$X_I = KR \begin{bmatrix} x_w \\ y_w \\ h \end{bmatrix} + KT = KR \left\{ \begin{bmatrix} x_w \\ y_w \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ h \end{bmatrix} \right\} + KT \tag{3.29}$$

$$X_I = KR \begin{bmatrix} x_w \\ y_w \\ 0 \end{bmatrix} + KR \begin{bmatrix} 0 \\ 0 \\ h \end{bmatrix} + KT \tag{3.30}$$

Haciendo $C = KR$ y $T_p = KR \begin{bmatrix} 0 \\ 0 \\ h \end{bmatrix} + KT$ la expresión queda como sigue:

$$X_I = CX_\pi + T_p = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ 0 \end{bmatrix} + \begin{bmatrix} t_{p1} \\ t_{p2} \\ t_{p3} \end{bmatrix} \tag{3.31}$$

Esta expresión se puede representar de forma homogénea como sigue:

$$X_I = \begin{bmatrix} c_{11} & c_{12} & c_{13} & t_{p1} \\ c_{21} & c_{22} & c_{23} & t_{p2} \\ c_{31} & c_{32} & c_{33} & t_{p3} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ 0 \\ 1 \end{bmatrix} \tag{3.32}$$

que se puede reducir a:

$$X_I = C'X_\pi = \begin{bmatrix} c_{11} & c_{12} & t_{p1} \\ c_{21} & c_{22} & t_{p2} \\ c_{31} & c_{32} & t_{p3} \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \tag{3.33}$$

Una vez conocida la relación entre los puntos del plano de ocupación X_π en el espacio 3D y los puntos del plano imagen Δ, es necesario describir la relación entre los puntos del plano de ocupación π y el *grid* formado en ese plano X_G. En este caso con un origen de coordenadas distinto y con un factor que determina el tamaño del *pixel* en cada eje del grid. Dicha transformación se modela a través de la matriz M.

$$\begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} = \begin{bmatrix} m_x & 0 & D_{X_w} \\ 0 & m_y & D_{Y_w} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_g \\ y_g \\ 1 \end{bmatrix} \tag{3.34}$$

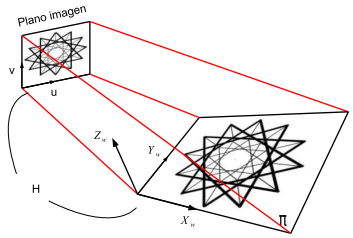


Figura 3.13: Relación de homografía

$$M = \begin{bmatrix} m_x & 0 & D_{X_w} \\ 0 & m_y & D_{Y_w} \\ 0 & 0 & 1 \end{bmatrix} \tag{3.35}$$

donde

m_{X_w} y m_{Y_w}: Equivalencia de milímetros a pixels

D_{X_w} y D_{Y_w}: Desplazamiento del origen de coordenadas X_w,Y_w, medido en milímetros.

El par de matrices de homografía H y H_{inv} relacionan las coordenadas de la imagen en la cámara X_I y las coordenadas de la imagen X_G formada en el plano π vendrá dada por las expresiones siguientes:

$$H = C'M = \begin{bmatrix} c_{11} & c_{12} & t_{p_1} \\ c_{21} & c_{22} & t_{p_2} \\ c_{31} & c_{32} & t_{p_3} \end{bmatrix} \begin{bmatrix} m_x & 0 & D_{X_w} \\ 0 & m_y & D_{Y_w} \\ 0 & 0 & 1 \end{bmatrix} \; y \; H_{inv} = H^{-1} \tag{3.36}$$

En la figura 3.13 se muestra la relación la proyección de puntos del plano imagen Δ a puntos de un plano de ocupación π, en este caso a una altura $h = 0$.

3.3.3. Combinación de las proyecciones homográficas de varias cámaras

Para combinar las proyecciones de las rostros detectados en cada imagen, se utilizó inicialmente el principio del *visual hull*, en el que la región de intersección de todas la proyecciones determina la localización del objeto. En este caso la intersección se logra mediante la operación AND lógico [7] [30] en el plano de ocupación.

En la figura 3.14 se muestra la imagen X_G el plano de ocupación con las proyecciones de rostros detectados en cada cámara, (a),(b),(c) y el AND lógico (d) entre las mismas. Esta operación tiene la ventaja de eliminar falsos positivos de la fase de detección de rostros, al ser menos probable que se produzca intersección de todas las cámaras en un área concreta. Sin embargo tiene el inconveniente de que un falso negativo en la detección de rostros en una cámara puede eliminar su detección en el plano de ocupación, teniendo como consecuencia que aparezcan muchos falsos negativos en los resultados (apartado 4.5).

Con vistas a obtener una combinación menos restrictiva, se planeó como alternativa realizar la unión(OR) de las intersecciones (AND) dos a dos de las proyecciones X_{IC}. De esta manera con dos cámaras cuyas proyecciones se intercepten habrá un resultado positivo en la detección. Presenta como inconveniente que puede ser más propensa a devolver zonas de falsos positivos.

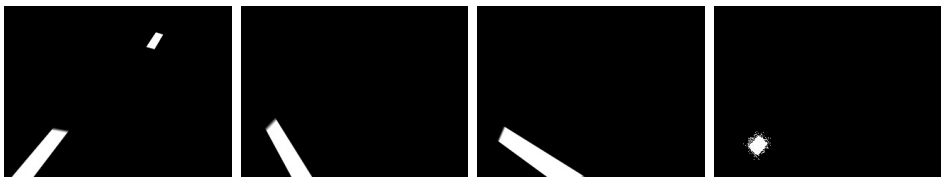


Figura 3.14: Combinación de las proyecciones en el plano de homografía: (a,b,c) cámara 1,2y 3 (d) AND Lógico

3.4. Localización de personas basada en fusión audio-visual

En esta sección se presenta el esquema propuesto para la combinació de los sistemas de extracción de información de audio y vídeo planteado sobre el plano de ocupación y el PF como algoritmo de seguimiento.

3.4.1. Modelo de integración de la información audiovisual

Para combinar las dos fuente de información de posición se procedió a realizar la operación de OR lógico entre las regiones detectadas con actividad por los bloques de audio y vídeo [7]. De esta manera se espera que ambas fuentes de información se complementen para lograr una mejor detección, en caso de que el vídeo proporcione información de localización en los intervalos donde falla la detección de audio y viceversa. Este método tiene el inconveniente es que no elimina los falsos positivos detectados por algoritmo de localización de audio, independientemente si coincida o no con las medidas de vídeo

Adicionalmente, este esquema permite que las medidas de vídeo proporcionen información de posición incluso cuando los locutores no estén hablando, lo cual permita al PF seguir una trayectoria continua, facilitando así en general su funcionamiento. Es necesario tener en cuenta en este trabajo se pretende evaluar localización solo en los intervalos en que los locutores están hablando.

3.4.2. Seguimiento usando Filtro de Partículas

Los Filtros Bayesianos (BF) son algoritmos de estimación que permiten modelar el comportamiento del sistema mediante funciones de densidad de probabilidades (pdf). La esencia de estos métodos consiste en estimar la pdf del estado del sistemas dado el conjunto de medidas realizadas, conocida como probabilidad *a posteriori* o creencia $p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t})$ donde $\overrightarrow{x}_t$ es el vector de estados y $y_{1:t}$ es el conjunto de medidas realizadas a lo largo de la evolución del sistema.

El valor de la creencia anteriormente definida se obtiene mediante la aplicación de la regla de Bayes de la siguiente manera:

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = \frac{p(\overrightarrow{y}_t|\overrightarrow{x}_t, \overrightarrow{y}_{1:t-1})p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t-1})}{p(\overrightarrow{y}_t|\overrightarrow{y}_{t-1})} \tag{3.37}$$

Donde:

- $p(\overrightarrow{y}_t|\overrightarrow{x}_t)$: Verosimilitud, Modelo de percepción u observación, este pondera la probabilidad de los estados actuales(instante t) de acuerdo a las medidas realizada en el instante t.

- $p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t-1})$: Probabilidad *a priori*, estimación de la probabilidad del estado en el instante t, dado las observaciones anteriores. Es una aproximación o predicción de la función de densidad a buscar.
- $p(\overrightarrow{y}_t|\overrightarrow{y}_{1:t-1})$: Probabilidad total del vector de medidas.

La ecuación anterior se puede simplificar aplicando la condición de Markov, en base a la cual la historia pasada de un sistema puede ser resumida en su estado actual, para cada instante de tiempo.

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = \frac{p(\overrightarrow{y}_t|\overrightarrow{x}_t)p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t-1})}{p(\overrightarrow{y}_t|\overrightarrow{y}_{t-1})} \tag{3.38}$$

Además si se tiene en cuenta que el conjunto completo de creencias ha de sumar la unidad, el valor del denominador se suele sustituir por un como factor de normalización η , obteniéndose:

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = \eta p(\overrightarrow{y}_t|\overrightarrow{x}_t)p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t-1}) \tag{3.39}$$

En este punto, desarrollando la ecuación anterior mediante la aplicación de la ley de probabilidad total se obtiene:

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = \eta p(\overrightarrow{y}_t|\overrightarrow{x}_t)\int p(\overrightarrow{x}_t|\overrightarrow{x}_{t-1},\, \overrightarrow{y}_{1:t-1})p(\overrightarrow{x}_{t-1}|\overrightarrow{y}_{t-1})dx \tag{3.40}$$

Aplicando de nuevo la condición de Markov se obtiene la formulación compacta del BF:

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = \eta p(\overrightarrow{y}_t|\overrightarrow{x}_t)\int p(\overrightarrow{x}_t|\overrightarrow{x}_{t-1})p(\overrightarrow{x}_{t-1}|\overrightarrow{y}_{t-1})d\overrightarrow{x}_t \tag{3.41}$$

Donde $p(\overrightarrow{x}_t|\overrightarrow{x}_{t-1})$:es un modelo de actuación o de estado, que brinda información de la dinámica del sistema(probabilidad de que un sistema esté en una estado $\overrightarrow{x}_t$ si su estado anterior fue $\overrightarrow{x}_{t-1}$)

La dificultad de trabajar con esta expresión consiste en que la integral no tiene solución analítica, excepto en casos particulares cuando las funciones de densidad, tienen un comportamiento lineal y gaussiano, obteniéndose para estos casos una solución analítica conocida como filtros de Kalman (KF). Una alternativa frente a casos más generales es consiste en utilizar representaciones discretas de dichas funciones.

Los Filtros de partículas(PF) discretizan la pdf *a posteriori* o creencia $p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t})$ a través de un conjunto de n partículas pesadas $(\overrightarrow{x}_t^{(i)}, \widetilde{w}_t^{(i)})$. Estas partículas contienen una representación del vector de estados $\overrightarrow{x}_t^{(i)}$, en el instante de tiempo t y un peso asociado $\widetilde{w}_t^{(i)}$, que determina la importancia de esa partícula, o sea,la probabilidad de que el objetivo seguido se encuentre en el estado $\overrightarrow{x}_t^{(i)}$.

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = S_t = \left\{(\overrightarrow{x}_t^{(i)}, \widetilde{w}_t^{(i)})\right\}_{i=1}^{n} \tag{3.42}$$

Con lo que la ecuación del BF se transforma en:

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = p(\overrightarrow{y}_t|\overrightarrow{x}_t^{(i)})\sum_{i=1}^{n} p(\overrightarrow{x}_t^{(i)}|\overrightarrow{x}_{t-1}^{(i)})p(\overrightarrow{x}_{t-1}^{(i)}|\overrightarrow{y}_{1:t-1}) \tag{3.43}$$

En la ecuación anterior se puede observar que el segundo término del sumatorio $p(\overrightarrow{x}_{t-1}^{(i)}|\overrightarrow{y}_{1:t-1})$, se corresponde con la creencia en el instante *t-1*. De esta forma la resolución de esta ecuación se vuelve un proceso iterativo. En la práctica, estos filtros trabajan en dos pasos, en una primera llamada Predicción, se obtiene la probabilidad *a priori* a partir de la creencia obtenida en el instante *t-1* y del modelo de actuación del sistema $p(\overrightarrow{x}_t^{(i)}|\overrightarrow{x}_{t-1}^{(i)})$.

$$p(\overrightarrow{x}_t^{(i)}|\overrightarrow{y}_{1:t-1}) = \sum p(\overrightarrow{x}_t^{(i)}|\overrightarrow{x}_{t-1}^{(i)})p(\overrightarrow{x}_{t-1}^{(i)}|\overrightarrow{y}_{1:t-1}) \tag{3.44}$$

visto de otro modo, en este paso las partículas se propagan a través del modelo de estado del sistema bajo estudio para obtener un nuevo conjunto $S_{t|t-1} = \{\overrightarrow{x}_{t|t-1}^{(i)}, \widetilde{w}_{t-1}^{(i)}\}_{i=1}^{n}$ que representa la distribución *a priori* del vector de estado en el instante t, $p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t-1})$.

En la segunda etapa, denominada Corrección, se multiplica la probabilidad *a priori* por la función de verosimilitud, obteniéndose la creencia del instante t.

$$p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t}) = p(\overrightarrow{y}_t|\overrightarrow{x}_t)p(\overrightarrow{x}_t|\overrightarrow{y}_{t-1:1}) \tag{3.45}$$

En el caso del PF el peso de cada partícula $\overrightarrow{w}_t = \{w_t^{(i)}\}_{i=1}^{n} \equiv (\overrightarrow{x}_{0:t})$ el que se obtiene comparando el vector de salida $\overrightarrow{y}_t$ y el valor predicho basado en la estimación $h(\overrightarrow{x}_{t|t-1})$, lo que implica que se obtiene directamente de la función de verosimilitud $p(\overrightarrow{y}_t|\overrightarrow{x}_t)$ del modelo elegido:

$$w(\overrightarrow{x}_{0:t}) = w(\overrightarrow{x}_{0:t-1}) \cdot \frac{p(\overrightarrow{y}_t|\overrightarrow{x}_t) \cdot p(\overrightarrow{x}_t|\overrightarrow{x}_{t-1})}{q(\overrightarrow{x}_t|\overrightarrow{x}_{0:t-1}, \overrightarrow{y}_{1:t})} \tag{3.46}$$

Además de estos dos pasos básicos es necesario incluir un mecanismo de eliminación de partículas con muy baja probabilidad e inclusión de nuevas partículas en las regiones de mayor probabilidad, para evitar problemas de degeneración del conjunto [60]. Este mecanismo se implementa en un nuevo paso, llamado de Selección.

el paso de Selección usa el vector de pesos $\overrightarrow{w}_t = \{w_t^{(i)}\}_{i=1}^{n}$ calculado en el paso anterior de muestreo para obtener un nuevo conjunto $S_t = \{\overrightarrow{x}_t^{(i)}, \widetilde{w}_t^{(i)}\}_{i=1}^{n}$ con las partículas más probables, que representarán la nueva creencia $p(\overrightarrow{x}_t|\overrightarrow{y}_{1:t})$ o probabilidad *a posteriori* del vector de estado.

Este conjunto de muestras es usado como entrada para la iteración del algoritmo en el instante *t+1*. De esta forma el diagrama de flujograma de un PF queda como se muestra en la figura 3.15.

En este trabajo se ha utilizado, una variante de PF, llamado "Filtro de Partículas Extendido con Proceso de Clasificación" o "eXtended Particle Filter with Clustering Process", XPFCP propuesto en [4]. Este algoritmo contiene modificaciones para poder seguir varios objetivos con un único filtro. Entre las mejoras significativas está la inclusión de un bloques de agrupamiento o clasificación, uno para el conjuntos de medidas y otro para el conjuntos de partículas a los distintos objetivos. El esquema del XPFCP utilizado se muestra en la siguiente figura 3.16

En este trabajo se define como medida $\overrightarrow{y}$ el conjunto de las posiciones en las cuales se ha detectado información de la unión de audio y vídeo. Se utiliza un algoritmo K-Medias para un número variable de objetivos [61], que agrupa el conjunto de medidas en clases. Como modelo de actuación se ha utilizado uno de velocidad constante, en el que se asume que las partículas en la etapa de predicción se desplazan en todas las direcciones del plano, a una velocidad constante. El peso de cada partículas es calculado a partir de su distancia al centro geométrico de la clase de medidas más cercano. Finalmente en el paso de selección se eliminan las partículas de más bajo

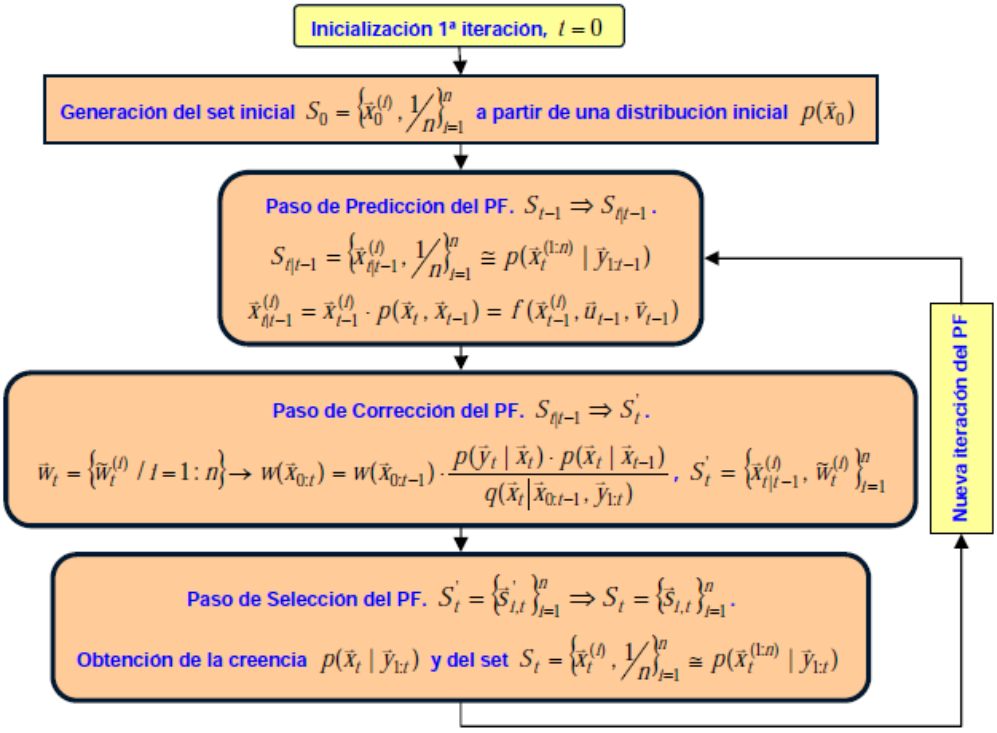


Figura 3.15: Flujograma de PF (*Sequential Importance Resampling*)

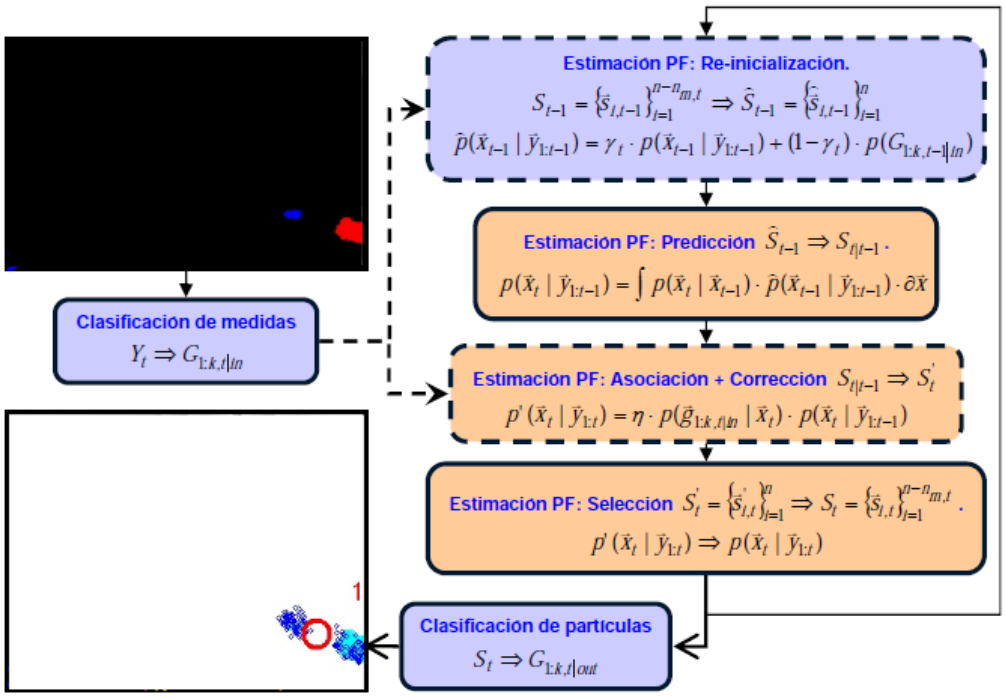


Figura 3.16: Flujograma de Filtro de Partículas Extendido con Proceso de Clasificación

peso y luego se agrupa el conjunto de partículas con el algoritmo k-Medias. El centro geométrico de las clases serán las salidas del algoritmo.

En la figura 3.16 se muestra un flujograma general del XPFCP usado. En la entrada del sistema aparecen las medidas de audio (rojo) y vídeo (azul) y a la salida se muestran las medidas en azul claro, las partículas (cian) y un círculo rojo centrado en la localización de obtenida por el filtro.

Capítulo 4

Resultados Experimentales

4.1. Introducción

A lo largo de este capítulo se exponen los resultados obtenidos en los diversos sistemas de experimentación utilizados en este trabajo. Para cada propuesta (sistema de localización acústica, visual y multimodal) se muestran los resultados obtenidos mediante una serie de tablas y algunos ejemplos gráficos de estos. Así mismo se analizan estos resultados.

4.2. Estrategia de evaluación y métricas

Para evaluar los resultados generados por cada uno de los distintos bloques desarrollados, así como los generados por el sistema completo se ha realizado la evaluación utilizando un conjunto de datos de audio y vídeo de la base de datos AV16.3 [62], [63] [64], y a los resultados parciales y totales obtenidos se le han aplicado un conjunto de métricas definidas en el proyecto *Computers in the Human Interaction Loop* (CHIL) [64].

El objetivo de este procedimiento es evaluar la localización de un único locutor, en los instantes en los que se encuentra hablando, por lo que las métricas usadas, permiten obtener información de cuan eficiente es el algoritmo detectando actividad de voz y cuan preciso es en la localización de la fuente sonora. También se muestran el número de *frames* detectados y el la duración en segundos de los segmentos anotados.

Las métricas utilizadas han sido:

- El *Pcor* o *Localization Rate* mide la fiabilidad del algoritmo como el porcentaje de estimaciones clasificadas como "error preciso" o "*fine error*" (F_{EF}) entre el total de *frames*, (TF) [65]. Un *frame* es etiquetado con "error preciso" si el error de localización del locutor no supera un umbral (E_r) de 50cm y como localización grosera o "*gross error*" en el caso contrario.

$$Pcor = \frac{F_{EF}}{TF} \cdot 100 \tag{4.1}$$

- Los parámetros *Bias fine*, *Bias fine+gross*, *Bias AEE fine* y *Bias AEE fine+gross* miden errores promedios de localización en milímetros de los errores etiquetados como "*fine error*" para los casos de *Bias fine* y *Bias AEE fine*, y de todos (*fine* y *gross*) para *Bias fine+gross*, *Bias AEE fine+gross*. El *Bias fine* y *Bias fine+gross* muestran los resultados por dimensión (x, y, z), mientras que el *Bias AEE fine* y *Bias AEE fine+gross* muestra el error global. El *Bias AEE fine* es equivalentes al *Múltiple Object Tracking Precision* (MOTP) definido a continuación.

$$MOTP = \frac{\sum_{i,t} d_{i,t}}{\sum_t c_t} \tag{4.2}$$

  donde c_t es el número de posiciones anotadas con una hipótesis asociada (generada por el algoritmo) en el *frame* t y $d^{i,t}$ es la distancia euclídea entre la posición anotada y la posición de la hipotesis correspondiente.

- El parámetro *deletion* o tasa de borrados se refiere a la fracción entre el número de *frames* anotados en el *groundtruth* en los que el algoritmo no devuelve un resultado de posición, falsos negativos (FN), y el total de *frames* anotados (falsos negativos (FN) y verdaderos

positivos (TP)), dado en por cientos, lo que es lo mismo la razón de falsos negativos o "False Negatives Rate"(FNR).

$$Deletion = \frac{FN}{TP + FN} 100 \tag{4.3}$$

Además para la evaluación del detector de actividad por sectores se utilizó la tasa de verdaderos positivos ó *True Positive Rate* (TPR). Esta se define a partir de la razón entre loas verdaderos positivos (TP) y el total de *frames* anotados o verdaderos

$$TPR = \frac{TP}{TP + FN} 100 \tag{4.4}$$

Para el cálculo de las métricas se utilizó un evaluador proporcionado por CHIL con este fin. Para el análisis de los bloques de detección de sectores basados en audio y detección de rostros y su proyección al plano de homografía se desarrollaron funciones en Matlab.

4.3. Bases de datos

En esta sección se describen las características de los datos usados para la evaluación del sistema desarrollado. Estos datos provienen de la base de datos AV16.3 [62] creada por investigadores de IDIAP [66]. Los datos de AV16.3 se han registrado en el contexto de una sala de reuniones. Una de las ventajas del uso de esta base de datos está que presenta una serie de fenómenos relevantes en las tareas de localización y seguimiento. Algunos de estos fenómenos son voces solapadas, localizaciones cercanas y lejanas, separaciones angulares grandes y pequeñas, inicialización de objetivos, número variable de objetivos, oclusiones parciales y totales y cambios "naturales" de iluminación.

En AV16.3 todos los datos fueron adquiridos de manera síncrona con 3 cámaras y 16 micrófonos lejanos. Estos datos son objeto de varias áreas de investigación como el seguimiento de locutores usando técnicas de audio, vídeo o multimodal (audiovisual).

A partir de este punto, el término de "secuencia" se refiere a los datos correspondientes a una única grabación, que contienen:

- 3 ficheros de vídeo DIVX AVI (resolución 288x360), uno por cada cámara, muestreados a 25 Hz. Cada fichero de vídeo además incluye un canal de audio.
- 16 ficheros de audio en formato WAV, grabados con dos *arrays* de 8 micrófono, muestreados a 16 kHz.
- En algunos casos, un fichero de audio WAV con la señal registrada del micrófono de solapa en el locutor, muestreado a 16 kHz.

Los ficheros de vídeo llevan asociada una marca de tiempo en cada imagen y los ficheros de audio siempre empiezan en la marca de tiempo 00:00:10.00 de vídeo. Los detalles de implementación del hardware para sincronizar todos los dispositivos de grabación puede encontrarse en [67]. La base de datos cuenta con un total de 42 secuencias, incluyendo 10 anotadas y 32 no anotadas, con una duración de estas que oscila entre 14 segundos y 9 minutos (total de 1h25 acumulado).

4.3.1. Configuración física

En la base de datos AV16.3 se han usado 3 cámaras de vídeo y dos *arrays* circulares de ocho micrófonos de 10 cm de radio cada uno [67]. Los dos *arrays* de micrófonos están separados 0.8 m entre sí.

Figura 4.1: Sala de grabación de AV16.3. Imagen de la secuencia 18.

En la figura 4.1 se observan las vistas de las tres cámaras de una secuencia de la base de datos. En la imagen del centro se han señalado los *arrays* circulares de micrófonos. También se puede observar en la cabeza de los locutores presentes una "esfera marcador", usada en algunas secuencias para facilitar la tarea de anotación.

4.3.2. Descripción de las secuencias usadas

Como se ha mencionado, la motivación inicial de este trabajo es evaluar el algoritmo desarrollado en condiciones donde esté presente un único locutor. Por ello se han seleccionado las cinco secuencias de la base de datos que presentan esta característica:

En la tabla 4.1 se muestra la lista de las secuencias (primera columna) usadas en esta evaluación. Así mismo se muestra la duración en segundos de la secuencia (columna 2), la modalidad de interés , [A]udio, [V]ideo o [AV]Audiovisual (columna 3), el número de locutores (columna 4) y el comportamiento del locutor, predominante voz solapada [ov], al menos una oclusión visual [occ], locutor estático [S], locutor [D]inámico, [U] movimiento no restringido (columna 5).

Nombre	Dur. (s)	Modalidad de interés	Num. de locutore(s)	comportamiento(s)
seq01-1p-0000	217	A	1	S
seq02-1p-0000	189	A	1	S
seq03-1p-0000	242	A	1	S
seq11-1p-0100	30	A, V, AV	1	D
seq15-1p-0100	35	AV	1	S,D(U)

Tabla 4.1: Lista de secuencias usadas.

- seq01-1p-0000, seq02-1p-0000, seq03-1p-0000: Un único locutor, que se encuentra estático mientras habla, y se ubica de frente a los *arrays* de micrófonos. El propósito de estas secuencias es evaluar los sistemas de localización con fuentes de audio y un único locutor..
- seq11-1p-0100: Un único locutor, que se mueve mientras habla, ubicándose siempre de frente a los *arrays* de micrófonos. El propósito de estas secuencia es evaluar algoritmos de localización de locutores basados en información de audio, vídeo o audio y vídeo en situaciones difíciles de movimiento. El locutor se encuentra hablando la mayor parte del tiempo.

- seq15-1p-0100: Un único locutor, que se mueve mientras habla, alternando voz con largos períodos de silencio. Los propósitos de esta secuencia son 1) demostrar que el seguimiento de fuentes sonoras, no puede recuperarse cuando las trayectorias son impredecibles en los períodos de silencio. 2) un caso de prueba para el seguimiento audiovisual.

4.3.3. Anotación

Para la obtención de la localización 3D de la boca del locutor se utilizan dos alternativas, la primera a través de una interfaz, *Mouth Annotation Interface* (MAI), de los propios creadores de AV16.3, que permite anotar en la imágenes 2D, la posición de la boca del locutor. Luego se reconstruye la posición 3D a través de la proyección de las 3 cámaras, usando los parámetros de calibración de las mismas. La otra alternativa consiste en usar los datos 3D de la "esfera marcador", partiendo de su anotación 2D y reconstrucción 3D luego, con otra interfase denominada *Ball Annotation Interface* (BAI).

En cualquiera de los dos casos el resultado del proceso de anotación está almacenado en ficheros texto con extensión .mouthgt. La información de segmentación de voz o silencio esta implícita en lo ficheros de anotación de la posición de la boca (.mouthgt), pues en los mismos solo está anotada la posición de la boca en los *frames* donde hay presencia de voz.

4.3.4. Preparación de los datos

Para poder procesar los datos de AV16.3 con la aplicación desarrollada fue necesario realizar una serie de modificaciones a los estos:

- Transformación del formato de vídeo (avi) a secuencia de imágenes.
- Sincronización (alineación) de las secuencias de imágenes de cada cámara con relación a las señales de audio.

La primera modificación fue necesaria debido a que la aplicación desarrollada está preparada para aceptar una secuencia de imágenes, y no un vídeo. La transformación del formato se llevó a cabo a través de la aplicación ffmpeg [68], creando ficheros con formato jpg y con nombre asociado al número del *frame* generado partiendo del *frame1*. Este procedimiento se realizó para todos los vídeos asociados a cada una de las cámaras, sin compresión adicional.

La segunda de las transformaciones fue debido a la falta de sincronismo entre los ficheros de vídeo de cada cámara entre sí y con los de las señales de audio. Como se menciona anteriormente todos los ficheros de vídeo contienen una marca de tiempo codificada en la fila superior del fichero, sin embargo ésto no significa que las secuencias obtenidas con cada cámara comiencen en el mismo instante.

En la tabla 4.2 se muestra el instante de comienzo de grabación de cada cámara en cada vídeo (el instante correspondiente al primer *frame* de la secuencia de imágenes). Para crear una correspondencia entre el mismo *frame* de todas las cámaras y los *frames* de audio, se adicionó en cada secuencia de vídeo, un número de *frames* de forma tal que el primero de cada cámara correspondiera con el inicio de la secuencia de audio. Estos *frames* adicionales son copias del primer *frame* de la secuencia de vídeo original.

Por ejemplo, en la secuencia de vídeo 01, las 3 cámaras tienen 5,21s 6,02s y 7,06s segundos respectivamente de retraso con relación al comienzo de los *frames* de audio. En este caso se

	cámara 1	cámara 2	cámara 3
seq01-1p-0000	15,21	16,02	17,06
seq02-1p-0000	17,22	16,07	19,11
seq03-1p-0000	15,24	14,24	17,00
seq11-1p-0100	16,16	16,18	15,12
seq15-1p-0100	08,16	09,18	11,06

Tabla 4.2: Instante de inicio de los vídeos, por secuencia y cámara.

añadieron, $\lfloor 5{,}21s \cdot 25Hz \rfloor - 1 = 129$, $\lfloor 6{,}02s \cdot 25Hz \rfloor - 1 = 149$ y $\lfloor 7{,}06s \cdot 25Hz \rfloor - 1 = 175$ *frames* (copias del *frame*1) al inicio de la secuencia de vídeo correspondiente. La nueva secuencia de *frames* comienza en cualquier caso en el *frame0*. De esta forma el *frame* i-ésimo de cada cada cámara corresponde al intervalo de tiempo del *frame* i-ésimo de audio.

4.4. Evaluación del sistema basado en audio

Para evaluar el sistema de localización basado en audio debemos evaluar sus dos bloques funcionales, el detector de sectores activos, y los algoritmos de localización de fuentes puntuales descritos en el capítulo anterior.

4.4.1. Evaluación del algoritmo de detección por sectores

El algoritmo de detección de sectores cumple dos objetivos fundamentales en el sistema desarrollado, el primero consiste en detectar la presencia de voz o no en un *frame* dado, o sea que en este caso se comportar como un detector de actividad de voz (VAD). El segundo objetivo es reducir el espacio de búsqueda de las fuentes de voz.

Como se ha sido mencionado en apartados anteriores se utilizó un umbral fijo sobre la actividad SSM de cada sector para determinar si un se encuentra activo o no.

Para evaluar el algoritmo en su función de VAD, se consideran el total de *frames* (tamaño del *frame* = 40 ms) de cada secuencia evaluada, un *frame* t se considera "positivo" si algún volumen de intersección de sectores I, se ha considerado activo en t y "negativo" en caso contrario. De esta forma se pretende analizar la detección de actividad del algoritmo sin tener en cuenta la precisión en la localización.

En la evaluación como localizador se consideraron todos los pares intersección-*frame* (I, t) evaluados. En este caso, todos los pares volumen de intersección-*frame*, (I, t) detectados por el algoritmo con actividad, son considerando como "positivos" y el resto como "negativos". Se considera que un par (I, t) es TP si en el *groundtruth* de ese *frame* existe una anotación y la localización de la misma se encuentra dentro del sector considerado activo por el algoritmo.

Las tablas 4.3 y 4.4 muestran el comportamiento del algoritmo sobre la secuencia 01, como VAD y como localizador respectivamente, para un valor de umbral de 63.

donde:

$$TPR = \frac{TP}{TP + FN} \tag{4.5}$$

y

	Frames Activos	*Frames* Inactivos	Total
Positivos	1365	232	1597
Negativos	3118	883	4001
Total	54483	1115	

TPR_F	60,7
FPR_F	6,9

Tabla 4.3: Matriz de confusión por *frame* y tasas de verdaderos positivos (TPR_F) y de falsos positivos (FPR_F)

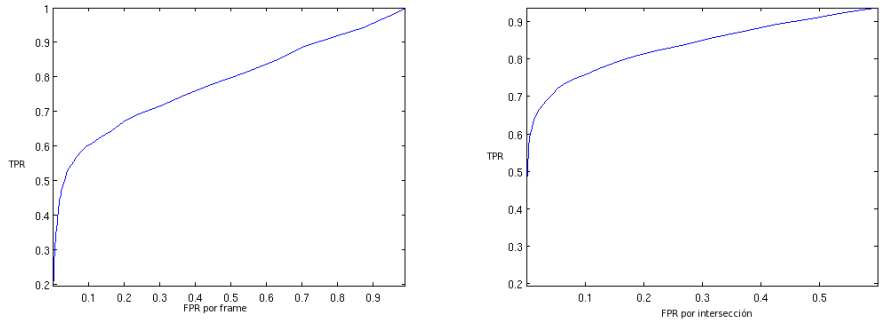
	Frames Activos	*Frames* Inactivos	Total
Positivos	1272	1253	2525
Negativos	505917	976	506893
Total	507189	2229	

TPR_I	56,52
FPR_I	0,24

Tabla 4.4: Matriz de confusión por par (Intersección de Sectores,*frame*) y tasas de verdaderos positivos (TPR_I) y de falsos positivos (FPR_I)

$$FPR = \frac{FP}{TN + FP} \tag{4.6}$$

Para una evaluación más completa se utilizó la gráfica *Receiver Operating Characteristic* (ROC) que representa el valor TPR vs *False Positive Rate* (FPR), para un conjunto de umbrales. En la figura 4.2 se muestran las curvas ROC del algoritmo en su tarea de detector de actividad de voz y de localizador en la secuencia 01, para las misma se utilizó como conjunto de umbrales $th = \{20, 21, \cdots, 100\}$.



(a) ROC por *frame* (b) ROC por *frame*-intersección

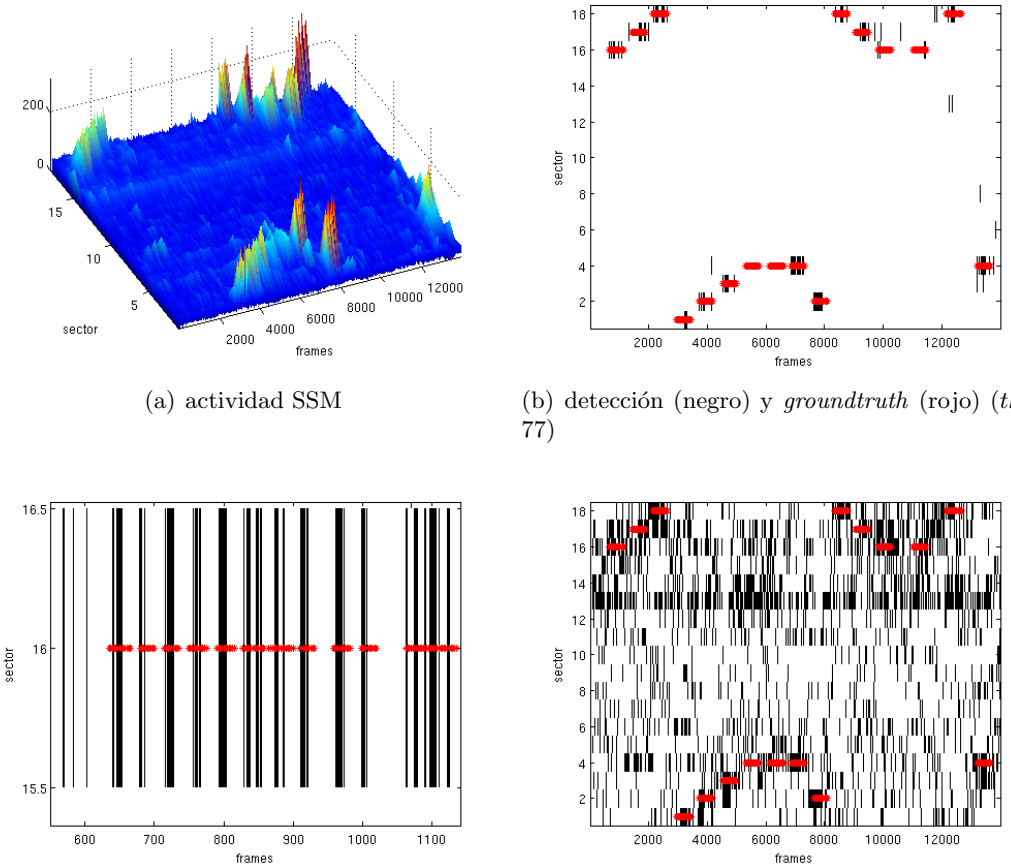
Figura 4.2: *Receiver Operating Characteristic* (a) como detector de voz y (b) como localizado

El algoritmo de detección basado en sectores no presenta buenas características en su función de detector de voz, en la curva ROC por *frame*. En figura 4.2(a) se puede observar que para

lograr un 90 % de TPR, es necesario permitir un elevado porcentaje de falsos positivos (FPR).

En la figura 4.3(a) se muestra la métrica SSM por sectores y por *frames* obtenida con el *array* 1 en la secuencia 01. La gráfica 4.3(b) muestra en color negro la detección de actividad seleccionando un umbral con valor elevado ($th = 77$) y en rojo el sector donde esta el locutor (*groundtruth*). Un análisis cualitativo de estas gráficas muestra que la medida SSM de actividad por sectores tiene un aspecto visual coherente con los sectores ocupados por el locutor en el transcurso de la secuencia.

Observando la gráfica 4.3(b) a menor escala (figura 4.3(c)), se observa que muchos de los errores de detección se presentan en el inicio y final de los intervalos de actividad. Esto sugiere que es necesario un mejor modelo para aproximar la actividad de sector en estas zonas de difícil detección. Una reducción en el valor del umbral no soluciona esta problemática dado que aumentan los sectores detectados incorrectamente. En la figura 4.3(d) se presenta la misma gráfica que en 4.3(b) pero con umbral $th = 33$.



(a) actividad SSM

(b) detección (negro) y *groundtruth* (rojo) ($th = 77$)

(c) *zoom* de la figura 4.3(b)

(d) detección (negro) y *groundtruth* (rojo) ($th = 33$)

Figura 4.3: Actividad SSM por sectores de la secuencia 01 (*array* 1), detección con umbrales altos y bajos

Es importante hacer notar que esta dificultad con el algoritmo de detección no es debido a la inclusión en nuestro esquema del concepto de intersección de sectores, que podría producir fallos para intersecciones vacías entre sectores, observando en la figura 4.4 el comportamiente del algoritmo por cada *array* (puramente por sectores) no es muy diferente a su rendimiento usando el esquema de intersección de sectores.

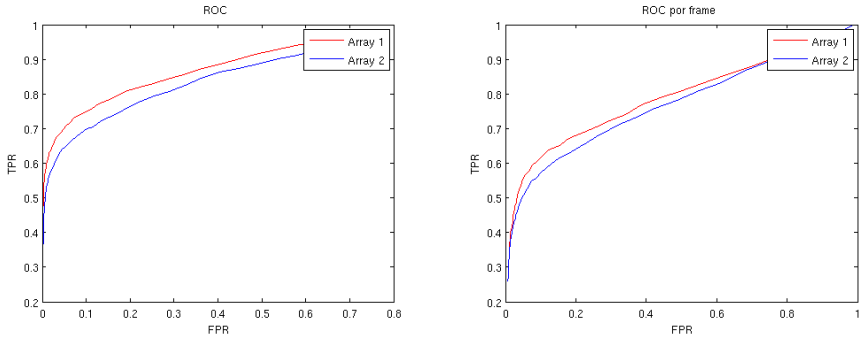


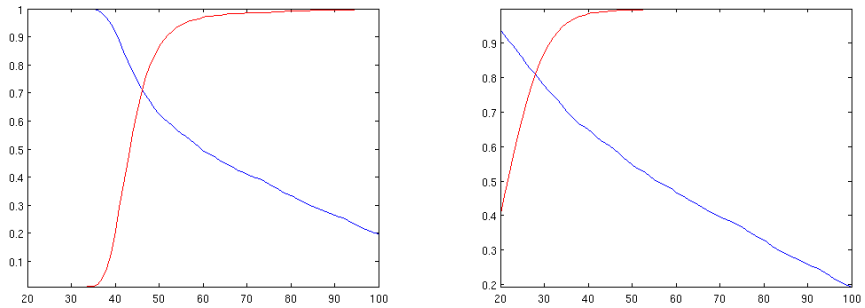
Figura 4.4: Receiver Operating Characteristic como detector de voz y como localizado

La explicación de este comportamiento radica en el modelo de umbral fijo que se ha implementado como primera aproximación. Este modelo resulta ser mucho más simple que el propuesto en [10] con varias parámetros que permite ajustarse en función de las condiciones del medio y seleccionar el umbral de manera automática, además de que en su esquema de trabajo aparecen otros bloques que discriminan basados en otras características espectrales de la voz que no han sido incluidas en esta propuesta.

Analizando el comportamiento en la función conjunta de detección-localización el algoritmo presenta mejores resultados que como VAD. Esto es posible dado que el número de pares *frame*-intersección que se evalúan en mucho mayor haciendo los porcentaje de falsos positivos menores

4.4.1.1. Selección del umbral

Dado que nuestra aplicación no cuenta aún con el bloque de detección automática de umbral ha sido necesario definir un valor constante, como primera solución al problema. Un criterio objetivo de selección de umbral que es usado comúnmente consiste en seleccionar de las curvas de comportamiento un valor de umbral en el cual las razones TPR y 1-FPR coinciden. Denominado umbral de "*Equal Error Rate*" (EER).



(a) TPR y 1-FPR por *frame* (b) TPR y 1-FPR por intersección

Figura 4.5: TPR y 1-FPR vs umbral, como detector de voz y como localizador en la secuencia 01

En las gráficas 4.5(a) y se muestran la curvas TRP y 1-FPR, en su tarea VAD (izquierda) y como localizador (derecha) en la secuencia 01, usando como umbrales el conjunto $th = \{20, 21, \cdots, 100\}$. En las mismas se observan los dos umbrales que cumplen la condición EER.

Un método objetivo para analizar el comportamiento del algoritmo, implicaría que los datos

	detector-localizador (intersección-*frame*)	detector (por *frame*)
seq01 (caso estático)	33	51
seq11 (caso dinámico)	36	63

Tabla 4.5: Umbrales correspondientes valores de equal error rate

	TPR/FPR por *frame*	TPR/FPR por intersección
tr-estat/ts-estat(33)	100/100	89/18
tr-estat/ts-dinam(33)	100/100	84/18
tr-dinam/ts-dinam(36)	100/99,6	84/10
tr-estat/ts-estat(51)	84/33	75/0,9
tr-estat/ts-dinam(51)	81/32	61/1
tr-dinam/ts-dinam(63)	67/8	61/0,3

Tabla 4.6: Resultados del algoritmo de detección de sectores para los distintos umbrales, análisis como detector de voz

de entranamiento, con las que se obtienen los umbrales, deben ser diferentes a las muestras con las cuales se procede a su evaluación, (por ejemplo *Leave-one-out*, *bootstraping*, entre otros). En este caso, dado que la solución de un umbral fijo es una solución intermedia, y que el conjunto de datos es solo de 5 secuencias, se ha optado por una solución más simple: seleccionar como muestras de entrenamiento una en la cual el locutor se encuentra estático mientras habla y otra en la cual el locutor se mueva mientras habla y analizar el comportamiento en tres situaciones:

- Selección de umbral en condiciones estáticas (seq01), evaluación con otras secuencias estáticas (seq02, seq03),"*tr-estat/ts-estat.*"
- Selección de umbral en condiciones estáticas (seq01), evaluación con secuencias dinámicas (seq11, seq15), "*tr-estat/ts-dinam*".
- Selección de umbral en condiciones dinámicas (seq11), evaluación con secuencias estáticas (seq01, seq02, seq03),"*tr-dinam/ts-estat*)".

En cada caso se tiene en cuenta el comportamiento del algoritmo como detector y como localizador. De este procedimiento se obtienen cuatro umbrales (tabla 4.5)

En la tabla 4.6 se muestran los resultados del algoritmo en los tres casos mencionados.

Los umbrales más bajos $(33, 36)$ presentan un porcentaje muy elevado de falsos positivos teniendo en cuenta su uso como detector de actividad. Los umbrales más elevados $(51, 63)$ eliminan un mayor número de *frames* inactivos a expensas de eliminar un mayor número de *frames* e intersecciones activas. Una decisión entre estos dos últimos depende de lo que se pretenda potenciar ante cada error, bien la robustez del algoritmo ante falsos positivos, la efectividad de detección de *frames* activos. Dado que este trabajo se pretende evaluar el efecto de la fusión audiovisual y sus mejora frente a la detección/localización con algoritmos uni-modales, un entorno en el que cada una de las fuentes no presenten razones de aciertos elevadas permitirá saber mejor cuánto se puede mejorar con la fusión. Además número elevados de intersecciones activas generaran muchas medidas de entrada al filtro de partículas en todo el espacio de búsqueda, y dado que la mezcla con las medidas de vídeo se realizan de una manera no ponderada, el filtro no

podrá como distinguir entre las medidas correctas y las erróneas, deteriorando en buena medida el resultado final. Con lo cual para las etapas anteriores se usará un umbral de 63 como decisión para determinar si una intersección se considera activa en un *frame* dado.

4.4.2. Evaluación del algoritmo de localización 3D

Una vez seleccionado el umbral se presenta la evaluación de los dos algoritmos de localización puntual desarrollados, SBD-SRP (Detección de Sectores más SRP) y SDB-SCG (Detección de Sectores más SCG). En la tabla 4.7 se muestran los resultados promedios de las cinco secuencias para estos dos algoritmos. En la tabla 4.8 se presenta la comparación entre el método usando sólo SRP y SBD-SRP. Todos los algoritmos fueron evaluados usando un *grid* regular en el volumen de (2.1m x 3.1m x 1m) con salto de 10cm en cada eje. En el caso de las métricas *Pcor*, se muestra el porcentaje de confianza.

	SBD+SCG	SBD+SRP
Pcor	$76{,}0 \pm 1{,}1\,\%$	$96{,}0 \pm 0{,}5\,\%$
Rel. error reduction		26,3 %
Bias fine (x:y:z) [mm]	$9 : 15 : 46$	$17 : -1 : 42$
Bias fine+gross (x,y,z) [mm]	$74 : 145 : 165$	$10 : -12 : 38$
Bias AEE fine [mm] = MOTP	155	130
Rel. AEE reduction		16,1 %
Bias fine+gross [mm]	524	161
Rel. BIAS f+g reduction		69,3 %
Deletion rate	33	33
Deletion rate reduction		−0,0 %
Loc. *frames*	5505	5505
Ref. duration (s)	641,0	641,0

Tabla 4.7: Comparación de algoritmos de localización con Audio (a) SBD+SRP (b) SBD+SCG

4.4.2.1. Comparación entre SBD+SRP y SBD+SCG

Los resultados indican que el método combinado de SBD-SRP presenta mejores índices de localización. El 96 % de *Pcor* (primera fila) del algoritmo SBD+SRP representa una reducción del error del 21.5 % con respecto a SBD+SCG. En el caso de los parámetros *Bias fine* y *Bias AEE fine* ambos algoritmos presentan resultados similares, sobre los 100 mm. Sin embargo en *Error fine+gross* SBD-SRP resulta mucho mejor con un resultado de 161 mm frente a 524 mm de SBD+SCG, mostrando una mejora relativa el error del 69.3 % . Estos resultados indican que el algoritmo SBD-SCG que en el 76 % de los casos en que localiza a la fuente con un error inferior al 0.5 m, el resultado de la localización resulta con un precisión similar a SRP. El hecho que en el 24 % restante de los casos el error supere los 0.5m hace que el resultado global se incremente significativamente. El el caso de SCG-SRP el elevado porcentaje de *Pcor* hace que ambos resultados (*Bias fine* y *Bias gross+fine*) tengan valores similares.

En la figura 4.6 se muestran las localizaciones detectadas por los algoritmos SRP (rojo) y SCG (verde) en la secuencia 01. También se muestran la ubicación del *groundtruth* (negro) y los micrófonos (azul). Haciendo un análisis cualitativo de los resultados de localización de ambos métodos, se puede notar que el SDB-SCG tiene dificultades para localización en el eje radial. O sea, si se observa distribución de dichas localizaciones desde una vista superior y lateral (plano

x-z), muchos de resultados presentan (al menos visualmente) una buena resolución de azimut (ver gráfico 4.6(a)) teniendo como origen de coordenadas el centro de uno de los *array*. Así mismo la localización en elevación también parece ser coherente con el plano formado por los centros de los *array* y la posición de la fuente, siendo su mayor dificultad definir la distancia radial a la que se encuentra la fuente.

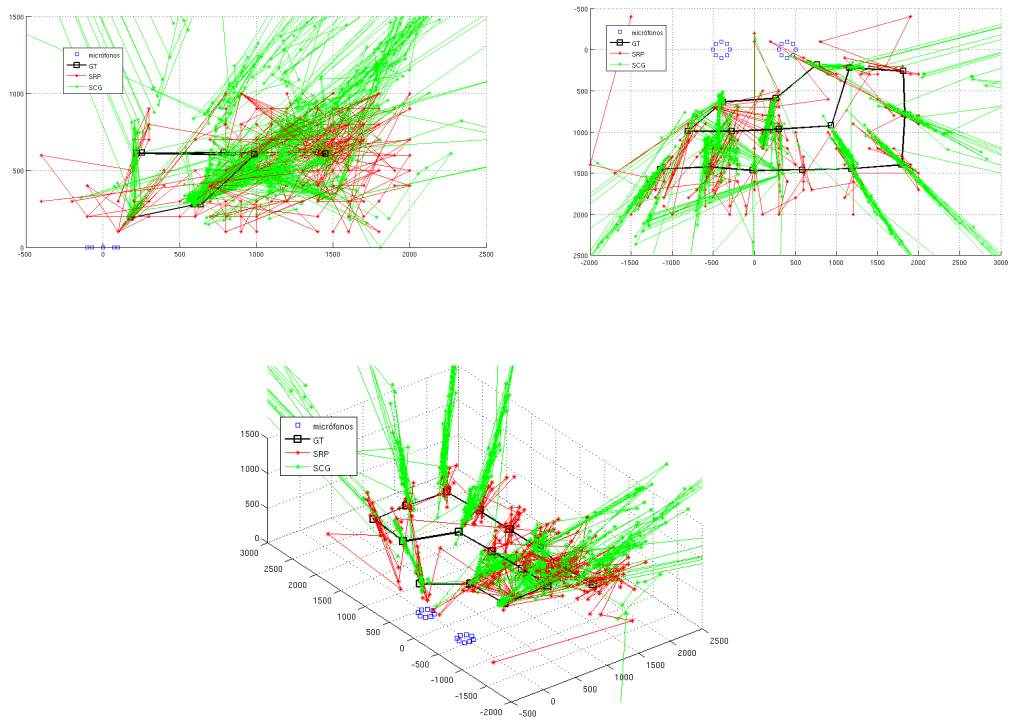


Figura 4.6: Resultados de Algortimos con audio

Una de las causas de esta dificultad radica en que SRP realiza una evaluación "exhaustiva" (con determinado salto entre puntos) en el volumen de la intersección, mientras que SDB-SCG es un método de optimización no exento del problemas de mínimos locales Por otro lado en la implementación realizada no se ha restringido las soluciones de SDB-SCG-PDM a la región definida por la intersección, presentando varias soluciones fuera del área de búsqueda. Otra de las posibles causas puede ser la distribución particular de los micrófonos en AV16.3, a un mismo lado de la fuente, lo que combinado con la métrica ADSP, hace que no generen en muchas *frames* un mínimo lo suficiente mente definido o en la zona correcta, como se muestra en la figura 4.7.

En la gráfica 4.7 se muestra el valor de la métrica PDM sobre los punto del plano de actividad. En dicha tabla se observa una vaguada muy definida a lo largo del eje que marca el centro del *arra* y la posición de la fuente. Esto se debe a que en este caso el efecto del otro *array* de micrófonos no esta presente con la misma "fortaleza" que el primero, creando una zona de mínimos bastante poco definida que hace fracasar al SCG. Una alternativa para salvar esta dificultad en trabajos futuros pude ser encontrar una manera de pesar el efecto de cada *array* en la métrica que disminuya este efecto. En vistas que el funcionamiento de que SBD-SRP ha presentado mejores resultados, se opta por mantener este método de localización puntual.

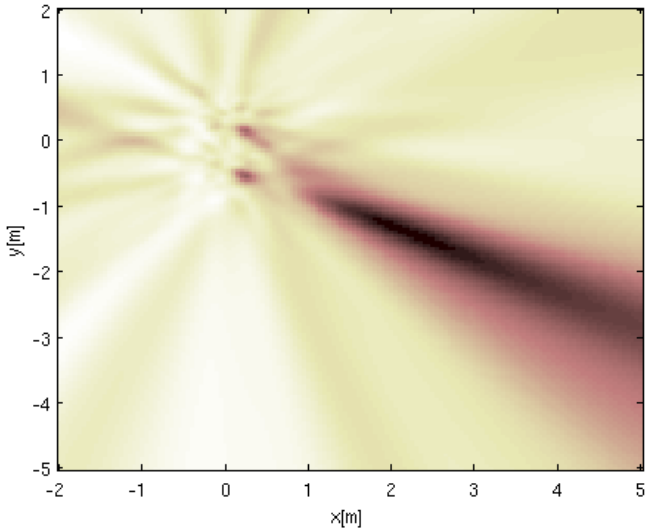


Figura 4.7: Métrica *Avarage Delay Sum Power* (ADSP) en el plano h = 0,7 m

	SRP	SBD+SRP
Pcor	$79{,}0 \pm 0{,}9\,\%$	$96{,}0 \pm 0{,}5\,\%$
Rel. error reduction		21,5 %
Bias fine (x:y:z) [mm]	$12 : -4 : 34$	$17 : -1 : 42$
Bias fine+gross (x,y,z) [mm]	$-169 : -240 : 7$	$10 : -12 : 38$
Bias AEE fine [mm] = MOTP	139	130
Rel. AEE reduction		6,5 %
Bias fine+gross [mm]	478	161
Rel. BIAS f+g reduction		66,3 %
Deletion rate	0	33
Deletion rate reduction		$-inf\,\%$
Loc. *frames*	8212	5505
Ref. duration (s)	641,0	641,0

Tabla 4.8: Comparación de algoritmos de localización con Audio (a) SRP (b) SBD+SRP

4.4.2.2. Comparación entre SBD+SRP y SRP

En tabla 4.8 se observa que en cuanto a localización, los resultados de combinando la detección por sectores y SRP resultan superiores a la utilizar solamente de SRP. El resultado de *Pcor* de 96 % representa un 21.5 % de reducción del error con respecto a SRP. El Error *Bias fine+gross* de SBD-SRP mejoran un 66.3 % a SRP, mientras que en *Bias fine* presentan resultados similares. Por otro lado SBD elimina *frames* en los que hay actividad resultando en un *Deletion* de 33 % como se ha analizado anteriormente.

Estos resultados se pueden interpretar de acuerdo a dos efectos posibles. El primero sería que SBD elimina regiones del espacio de búsqueda que representan posibles errores para SRP (zonas con picos superiores al pico verdadero). Con lo cual el uso de SBD presenta la ventaja adicional de "filtrar" falsos positivos en SRP, mejorando los resultados. El segundo sería que los *frames* que se han eliminado por SBD, coincidan con aquellos en los que SRP comete los errores. El primero de esto efectos tiene como es lógico tiene un factor beneficioso para el sistema. Un análisis mucho más exhaustivo que el realizado en este trabajo sería necesario para evaluar el peso de estos efectos en SBD+SRP.

4.4.3. Conclusiones sobre el sistema basado en audio

El sistema de localización basado en sectores desarrollados no presenta características muy buenas en su función conjunta detección localización, menos aún como VAD, presumiblemente por la simpleza del detector usado (umbral fijo), por lo que se puede plantear como trabajo futuro incluir el modelo presentado en [10] en este esquema. Como método de localización puntual la búsqueda con SBD-SRP ha presentado mejores resultados, por lo cual se ha optado por mantener este bloque en la propuesta. Sin embargo aún quedan aspectos a evaluar en un futuro como otras configuraciones de micrófonos y mejoras a la métrica presentada que pueden hacer posible la utilización de éste u otro método de optimización en el futuro, con la consecuentes mejoras en velocidad. El método SBD-SRP presenta mejoras en cuanto a la precisión de localización con respecto SRP, a espesa de un aumento al 33 % en *deletion*.

4.5. Evaluación del sistema basado en vídeo

En este epígrafe se analizaran los resultados de los bloque de detección y localización basados en información visual. El bloque de detección de rostros en 2D, y el bloque de proyección y combinación en plano de homografía.

4.5.1. Evaluación del algoritmo de detección de rostros y proyección

En este caso el algoritmo no tiene la posibilidad de entregar una localización, como el detector de audio (posición del máximo SRP) sino zonas del plano de homografía. Con la finalidad de evaluar el resultado en cada una de las secuencias y cámaras se consideró una detección positiva, si en un *frame* anotado como activo, el algoritmo devuelve alguna zona activa en la imagen proyectada en el plano de homografía y como error de localización la menor distancia de todos los puntos activos de la imagen a la proyección de la anotación en la imagen (valores x e y en el sistema de coordenadas de la imagen). Este resultado es convertido nuevamente a milímetros.

En las tablas 4.9,4.10 y 4.11 se muestran los resultados obtenidos por el algoritmo para las secuencias analizadas y cada una de las cámaras. Los parámetros presentados son *Pcor*, los

errores *Bias fine* y *Bias fine+gross* y la taza de *deletion*. Las columnas se refieren a los resultados con las cámaras 1, 2 y 3 en las columnas 2, 3 y 4 de las tablas respectivamente. En las columnas 4 y 5 aparecen los resultados de la combinación And (intersección) y Or-And (unión de las intersecciones dos a dos).

	cam1	cam2	cam3	And	Or-And
Pcor	92 %	99 %	99 %	100 %	100 %
Rel. error reduction		7,6 %	7,6 %	8,6 %	8,6 %
Bias AEE fine [mm] = MOTP	18,96	1,2	0,72	3,02	2,9
Rel. AEE reduction		94 %	97 %	84 %	85 %
Bias fine+gross [mm]	93,61	6,2	3,64	3,02	2,9
Rel. BIAS f+g reduction		93 %	96 %	97 %	97 %
Deletion rate	63,1	8,7	2,6	68,7	9,9
Deletion rate reduction		86 %	95 %	−8,9 %	84 %
Loc. *frames*	676	1672	1786	572	1651

Tabla 4.9: Resultados detección y proyección de homografía, de las tres cámaras y su combinación AND lógico, con la secuencia 01

	cam1	cam2	cam3	And	Or-And
Pcor	87 %	99 %	99 %	100 %	100 %
Rel. error reduction		14 %	14 %	15 %	15 %
Bias AEE fine [mm] = MOTP	9,04	0,68	1,39	6,75	1,9
Rel. AEE reduction		92 %	85 %	25 %	78 %
Bias fine+gross [mm]	139	13,4	14,8	6,75	1,9
Rel. BIAS f+g reduction		90 %	89 %	95 %	98 %
Deletion rate	60,2	6,2	10,7	70,2	10,1
Deletion rate reduction		90 %	82 %	−16 %	83 %
Loc. *frames*	774	1825	1739	580	1749

Tabla 4.10: Resultados detección y proyección de homografía, de las tres cámaras y su combinación AND lógico, con la secuencia 02

El resultado observado en el parámetro *Pcor* (primera fila) muestra valores elevados para la mayoría de las cámaras, la de menor valor resulta siempre la cámara 1, con resultados de 92, 87 y 94 % para las tres secuencias, el resto muestra valores de 99 y 100 %. Las combinación de AND y AND-OR lógico de las tres proyecciones muestra un valor superior, en todos los casos del 100 %. Los parámetros de Bias, tanto fine como fine+gross, muestran un comportamiento similar. En ambos casos los resultados de las cámaras 2 y 3, son superiores a la cámara 1. En ambos casos con porcentajes de reducción del error de 93, 90 y 86 % de la cámara 2 y 96, 83 y 76 % de la cámara 3 con respecto a la cámara 1, en las secuencias 01, 02 y 03 respectivamente. En cuanto a los indicadores de precisión, *Bias fine*+gross de la combinación OR-AND (columna 4) mejora en todos los casos al obtenido con la combinación AND (columna 3) con valores de 2.9 % 1.9 % y 14 % frente a 3 ,6 y 45 %.

En nuestro caso, se espera que las variaciones de altura de las fuentes con respecto al plano de homografía sean pequeñas. Por esta razón es poco probable que una detección de rostro correcta sea proyectada a una distancia superior 0.5 metros, con lo cual podemos asumir que estos errores, clasificados como “gross error” corresponden a falsos positivos del detector de rostro, coincidentemente con que el algoritmo no pudo detectar el rostro correctamente (falso

	cam1	cam2	cam3	And	Or-And
Pcor	94 %	99 %	99 %	100 %	100 %
Rel. error reduction		5,3 %	5,3 %	6,3 %	6,3 %
Bias AEE fine [mm] = MOTP	30,5	7,94	17,4	45,1	14
Rel. AEE reduction		74 %	43 %	−47 %	54 %
Bias fine+gross [mm]	91,2	12,0	24,1	45,1	14
Rel. BIAS f+g reduction		86 %	73 %	50 %	84 %
Deletion rate	51,6	3,4	6,4	64,9	3,3
Deletion rate reduction		93 %	87 %	−25 %	93 %
Loc. *frames*	986	1966	1905	715	1968

Tabla 4.11: Resultados detección y proyección de homografía, de las tres cámaras y su combinación AND lógico, con la secuencia 03

positivo), afectando en este caso también a los errores *Bias* de localización. El hecho de que la combinación por intersección de las tres proyecciones, supere en estos parametro al resultado de todas las cámaras muestra que se está logrando la deseada eliminación de falsos positivos, pero queda por hacer un estudio detallado de las tasas de *deletion*.

En el caso del parámetro *deletion*, igualmente muestra siempre peores resultado la cámara 1, con 86, 90 y 93 % en la cámara 2 y 95, 82 y 87 % de reducción de error para la cámara 3 (secuencias 01, 02 y 03). La razón por la cual la cámara 1 presenta los peores resultados no está clara, presumiblemente por la posición relativa del locutor con relación a las cámaras, y la dificultad del algoritmo para detectar determinadas poses. En este caso, la intersección actúa negativamente, con resultados inferiores a la peor de las cámaras, dado que la omisión de una única cámara anula la detección de ese *frame*.

La combinación de las proyecciones a travéz de unión (OR) de las intersecciones dos a dos (cuarta columna) resultó ser mucho más efectiva en cuanto a la tasa de *deletion*, con resultados inferiores al 10 % (9.9, 7.0 y 3.3 % en las secuencias 01, 02 y 03).

En las secuencias 01 y 02 el resultado es inferior a la cámara de mejor resultado (2.6 y 6.2 %) con valores cercanos a la que le sigue (10 y 8.7 %) lo cual resulta lógico, dado que la cámara 1 está detectando mucho menos *frames* y es la intersección de las cámara 2 y 3 la que esta produciendo la mayoría de las detecciones, con lo cual se comporta de forma similar a la peor de estas dos. En la secuencia 03 el rendimiento es incluso mejor que la mejor de las cámaras (3.4 %), lo cual coincide con una mejora en las detecciones de la cámara 01.

Un análisis cualitativo de los resultados de estas combinaciones muestra que en varios *frames* las zonas detectadas con la combinación OR-And resulta de mayor área que la AND 4.8, lo cual es coherente ya que la intersección es subconjunto de la combinación OR-And y que además se ve incrementado dado que las cámaras 2 y 3, (las que mejores detectan) son las que están mas cercanas, produciendo proyecciones alargadas en la dirección de estas cámaras. Este aumento del área puede tener efecto en el resultado dado que un mayor área con partículas en el filtro.

En la figura 4.8 se muestra un *frame* con las proyecciones de las cámaras 2 y 3 (a) (no se ha detectado en la cámara 1 b), y cómo la zona detectada con la intersección And (c) y la unión de las intersecciones dos a dos AND-OR(c).

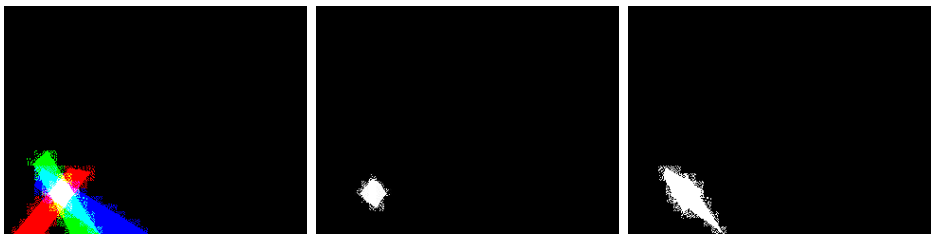


Figura 4.8: Plano de Homografía con zonas detectadas en las tres cámaras (R,G,B)-(cam1,cam2,cam3)(a), Región de intersección de las tres cámaras(b) y región de unión de las intersecciones dos a dos (c)

4.5.2. Conclusiones sobre el sistema basado en vídeo

El detector de rostro no presenta buenos índices de detección, especialmente en una de las cámaras. Esto se debe a que el algoritmo solo cuenta con plantillas entrenadas para poses frontales y de perfil, mientras que la posición relativa del locutor frente a la cámara no está restringida a estas poses. La combinación And-OR muestra mejores resultados que la intersección AND en cuanto a *deletion* y *Bias*. Las zonas detectadas con And presentan menor área que las de And-OR, este aspecto puede incidir en el resultado final en función de cómo se distribuyan las partículas dentro de las medidas, con lo cual este aspecto se evaluará con el filtro de partículas.

4.6. Evaluación del sistema basado en fusión audio-visual

En este epígrafe se analizaran los resultados de los bloque de detección y localización basados en información visual. El bloque de detección de rostros en 2D, y el bloque de proyección y combinación en plano de homografía.

4.6.1. Evaluación del seguimiento usando filtro de partículas

En las tablas 4.12, 4.13, 4.14, 4.15, 4.16 y 4.17 se muestran los resultados obtenidos del algoritmo de seguimiento, usando medidas de audio, vídeo y combinación de audio vídeo para cada una de las secuencias evaluadas. Las tablas 4.12, 4.14, y 4.16 comparan los resultados de audio, de vídeo con la combinación AND y audio visual igualmente con AND, mientras que en 4.13,4.15 y 4.17 se muestran los resultados con la alternativa de vídeo a partir de la unión de las combinaciones dos a dos (OR-AND). Además de las métricas usadas se muestran en la parte inferior de cada casilla la reducción del error, relativa a la variante de solo audio.

	A	V(And)	AV (And)
Pcor	100,0 ± 0,0 %	100,0 ± 0,0 %	100,0 ± 0,0 %
Rel. error reduction		0,0 %	0,0 %
Bias fine (x:y:z) [mm]	97 : −98 : 92	42 : 36 : 94	72 : −49 : 93
Bias fine+gross (x,y,z) [mm]	97 : −98 : 92	42 : 36 : 94	77 : −50 : 93
Bias AEE fine [mm] = MOTP	192	113	157
Rel. AEE reduction		41,1 %	18,2 %
Bias fine+gross [mm]	192	113	164
Rel. BIAS f+g reduction		41,1 %	14,6 %
Deletion rate	86	79	55
Deletion rate reduction		8,1 %	36,0 %
Loc. *frame*	259	379	822
Ref. duration (s)	154,2	154,2	154,2

Tabla 4.12: Comparación seguimiento seq01, vídeo And

	Audio	Vídeo (And-Or)	AV (And-Or)
Pcor	100,0 ± 0,0 %	100,0 ± 0,0 %	99,0 ± 0,5 %
Rel. error reduction		0,0 %	−1,0 %
Bias fine (x:y:z) [mm]	97 : −98 : 92	55 : −13 : 92	58 : −21 : 92
Bias fine+gross (x,y,z) [mm]	97 : −98 : 92	54 : −16 : 92	56 : −25 : 92
Bias AEE fine [mm] = MOTP	192	129	134
Rel. AEE reduction		32,8 %	30,2 %
Bias fine+gross [mm]	192	131	138
Rel. BIAS f+g reduction		31,8 %	28,1 %
Deletion rate	86	16	12
Deletion rate reduction		81,4 %	86,0 %
Loc. frames	259	1538	1617
Ref. duration (s)	154,2	154,2	154,2

Tabla 4.13: Comparación seguimiento seq01, vídeo And-Or

	Audio	Vídeo(And)	AV (And)
Pcor	82,0 ± 3,1 %	100,0 ± 0,0 %	98,0 ± 0,9 %
Rel. error reduction		22,0 %	19,5 %
Bias fine (x:y:z) [mm]	71 : −119 : 51	51 : 15 : 3	52 : −92 : 5
Bias fine+gross (x,y,z) [mm]	133 : −111 : 95	51 : 15 : 3	49 : −87 : 5
Bias AEE fine [mm] = MOTP	209	71	146
Rel. AEE reduction		66,0 %	30,1 %
Bias fine+gross [mm]	316	71	178
Rel. BIAS f+g reduction		77,5 %	43,7 %
Deletion rate	75	84	57
Deletion rate reduction		−12,0 %	24,0 %
Loc. *frames*	599	378	1040
Ref. duration (s)	171,4	171,4	171,4

Tabla 4.14: Comparación seguimiento seq02, vídeo And

	Audio	Vídeo(And-Or)	AV(And-Or)
Pcor	$82{,}0 \pm 3{,}1\,\%$	$99{,}0 \pm 0{,}5\,\%$	$99{,}0 \pm 0{,}5\,\%$
Rel. error reduction		20,7 %	20,7 %
Bias fine (x:y:z) [mm]	$71 : -119 : 51$	$75 : 31 : 6$	$75 : 12 : 6$
Bias fine+gross (x,y,z) [mm]	$133 : -111 : 95$	$79 : 35 : 6$	$77 : 17 : 6$
Bias AEE fine [mm] = MOTP	209	105	114
Rel. AEE reduction		49,8 %	45,5 %
Bias fine+gross [mm]	316	110	120
Rel. BIAS f+g reduction		65,2 %	62,0 %
Deletion rate	75	29	27
Deletion rate reduction		61,3 %	64,0 %
Loc. frames	599	1702	1762
Ref. duration (s)	171,4	171,4	171,4

Tabla 4.15: Comparación seguimiento seq02, vídeo And-Or

	Audio	Vídeo(And)	AV (And)
Pcor	$91{,}0 \pm 2{,}4\,\%$	$100{,}0 \pm 0{,}0\,\%$	$96{,}0 \pm 1{,}3\,\%$
Rel. error reduction		9,9 %	5,5 %
Bias fine (x:y:z) [mm]	$49 : -114 : -14$	$77 : 0 : -80$	$30 : -123 : -78$
Bias fine+gross (x,y,z) [mm]	$55 : -130 : -4$	$77 : 0 : -80$	$26 : -135 : -78$
Bias AEE fine [mm] = MOTP	219	127	188
Rel. AEE reduction		42,0 %	14,2 %
Bias fine+gross [mm]	275	127	214
Rel. BIAS f+g reduction		53,8 %	22,2 %
Deletion rate	79	93	66
Deletion rate reduction		−17,7 %	16,5 %
Loc. *frames*	551	193	900
Ref. duration (s)	220,0	220,0	220,0

Tabla 4.16: Comparación seguimiento seq03, vídeo And

	Audio	Vídeo(And-Or)	AV(And-Or)
Pcor	$91{,}0 \pm 2{,}4\,\%$	$100{,}0 \pm 0{,}0\,\%$	$100{,}0 \pm 0{,}0\,\%$
Rel. error reduction		9,9 %	9,9 %
Bias fine (x:y:z) [mm]	$49 : -114 : -14$	$164 : 149 : -83$	$148 : 114 : -83$
Bias fine+gross (x,y,z) [mm]	$55 : -130 : -4$	$164 : 150 : -83$	$148 : 114 : -83$
Bias AEE fine [mm] = MOTP	219	245	231
Rel. AEE reduction		−11,9 %	−5,5 %
Bias fine+gross [mm]	275	246	232
Rel. BIAS f+g reduction		10,5 %	15,6 %
Deletion rate	79	36	36
Deletion rate reduction		54,4 %	54,4 %
Loc. frames	551	1386	1383
Ref. duration (s)	220,0	174,5	174,5

Tabla 4.17: Comparación seguimiento seq03, vídeo And-Or

De los resultados obtenidos se puede extraer que en cuanto al número de *deletion* (fila 6 de las tablas) en las secuencias 02 y 03 (especialmente en ésta última) el seguimiento usando medidas de audio muestra mejores resultados que usando sólo medidas de vídeo "combinadas con AND", con un decremento del error relativo de 12 % y 17 % respectivamente, mientras que en la secuencia 01 el seguimiento con vídeo "combinadas con AND" supera en un 8.1 % al detector que incluye solamente información de audio. La alternativa de vídeo "combinadas con OR-AND" en todas las secuencias supera al seguimiento con audio, con incrementos del 81 %, 61 % y 54 % para la secuencia 01, 02 y 03 respectivamente.

En todos los casos, estos resultados de *deletion* son peores a los esperados debido a las limitaciones de los bloques de detección del sistema. A pesar de esto, los resultados obtenidos la fusión audiovísual "combinadas con AND" mejora en todos los casos a los obtenidos con una única fuente, con un incremento del 36 %, 24 % y 16.5 % en las secuencias 01, 02 y 03 respecto al seguimiento con solo audio, y en un 30 %, 32 % y 29 % (no aparece en tabla) respecto al seguimiento solo con vídeo "combinadas con AND". El seguimiento multimodal usando vídeo (Or-And) presenta mejores resultados que el seguimiento con audio, con 86 %, 64 % y 54 % y comparado con solo vídeo (And-Or) del 25 %, 6.8 % y 0 % (no aparece en tabla).

En cuanto a los resultados de detección, ninguno de los bloques de extracción de audio ó vídeo presenta buenos resultados, de acuerdo al por ciento de *deletion* que presentan. Sin embargo, la fusión audio-vídeo es capás de mejorar este parámetro superando los resultados relativos a cada fuente por separado. Este resulta lógico dado que la combinación a través de la suma OR permite que los *frames* que uno de los bloques no detecte, se compense con las detecciones del otro.

Con relación a la precisión de localización, observando el parámetro *Pcor* (fila 1 de las tablas), se ve que se obtienen valores elevados prácticamente en todas las secuencias. Sólamente en la seguimiento basados sólamente en audio y en las secuencias 02 y 03 se obtienen valores inferiores al 95 %, concretamente de 86 % y 91 % respectivamente. Esto demuestra que una vez detectadas las medidas, los tres algoritmos localizan el objetivo con un error inferior a 0.5m en todos los casos.

En cuanto a los errores de *Bias fine+gross* (fila 5 de las tablas), se observa que en todas las secuencias, la localización basado sólo en vídeo tanto combinadas con AND como con Or-And, son mejores que la basada sólo en audio, con reducciones del error relativo del 41 %, 77 %, 53 % y 81.4 %, 43.7 %, 15 % respectivamente. Los algoritmos de seguimiento con fusión de audiovisual combinados con AND superan a la variante basada sólo en audio en todos los casos, en 14 %, 43.7 % y 22.2 %, pero con resultados inferiores al seguimiento solo con vídeo "combinadas con AND". Finalmente el seguimiento con fusión de audiovisual combinación "Or-And" mejora al seguimiento solo con audio en 28.1 %, 62 % y 15,6 %, superando también las tasas obtenidas en el seguimiento con vídeo (Or-And) en las secuencias 01 y 03.

En cuanto a los parámetros a los parametros *Pcor* y *Bias*, el resultado intermedio de la fusión audio vídeo (menor precisión que vídeo y mas que audio) consideramos se debe a varias causas. La primera es que las medidas de audio y vídeo se mesclan de forma "plana", o sea la combinación OR, hace de operación de unión de conjuntos, y no se puede determinar cuál de las medidas es mas importante. Este hecho puede crear dos situaciones que influyan en el error: primero, si las medidas de audio y vídeo no se encuentran ceranas entre sí, el algoritmo debe crear dos clases de medidas diferentes, con un resultado final de localización para cada uno, pudiendo confundir al evaluador, si se le asigna como primer resultado el cluster erroneo. En segundo lugar, si ambas medidas tienen intersección o están muy cercanas de manera que se asuman como una única clase de medidas, si una de las medidas, tiene más dispersión o estas sesgada con relación a la ubicación correcta, debe incidir sobre el centro geométrico del cluster

deteriorando con esto el resultado.

Figura 4.9: Medidas de Audio y vídeo en dos *frames* de la secuencia 01, mostrando posibles errores provocados por la fusión "plana"

En la figura 4.9 se muestran dos *frames* en los que se muestran las medidas de audio y vídeo (cian), las partículas (rectángulos azules) y la localización resultante (circunferencia en rojo). En el caso mostrado en el primer *frame* el algoritmo genera dos clases de medidas y por tanto de partículas se distribuirán entre dos hipótesis de observación. El segundo *frame* muestra medidas audiovisuales conectadas, pero , lo que provocará la creación de una sola clase, pero con centro geométrico desplazado.

Otra de las problemáticas que aparecen en los experimentos realizados es que la detección de rostro no ubica sólo la posición de la boca en el plano de ocupación sino en todo el rostro, creando un centro geométrico de las medidas desplazado con relación a la ubicación real de la boca.

Una manera de solucionar estas dificultades sería ponderar la "importancia" de las medidas de cada una de las fuentes, así como de su combinación. Para llevar a cabo este procedimiento sería suficiente con utilizarcomo conjunto de medidas, no sólo el *grid* booleano de ocupación, sino uno probabilístico o continuo con el valor de potencia de audio, y algún valor indicativo de la probabilidad de encontrar rostro y de la posición de la boca dentro del mismo.

Es también importante notar que la designación de una altura constante para el *grid* de ocupación basado en vídeo en un seguimiento 3D de locutores es fuente de errores cuando la boca del locutor no se encuentra a la altura seleccionada. En los resultados obtenidos con las secuencias 01 y 03 han sido afectados en este sentido, incrementando el parámetro alrededor de 100 milímetros aproximadamentes.

Una propuesta interesante a incluir en el seguimiento audiovisual donde se estime la altura de la fuente inicialmente y se proyecte sobre un plano de homografía variable o una completamente en 3D, debe reducir el efecto de esta fuente de error.

4.6.2. Conclusiones sobre el sistema basado en fusión audio-visual

El algoritmo de seguimiento usando fusión audiovisual implementado, es sus dos versiones vídeo "combinadas con AND", vídeo (OR-And) ha mejorado los resultados de los algoritmos de seguimiento con ambas fuentes en cuanto al número de falsos negativos obtenidos, aunque aún los porcentajes alcanzados no son los esperados, debido a las limitaciones en la detección de información de audio y vídeo ya comentadas. En cuanto a la precisión, la versión que combina audio y vídeo (And) mejora los resultados de audio, mientras que resulta inferior a vídeo. La alternativa que combina audio y vídeo (Or-And) mejora al seguimiento uni-modal con audio y

con vídeo.

Capítulo 5

Conclusiones y Líneas Futuras

5.1. Conclusión

Tras la realización del sistema de localización de locutores usando fusión de audio vídeo propuesto en este trabajo se han podido obtener las siguientes conclusiones:

- Se ha implementado un bloque detección de actividad sonora por sectores (SBD) propuesto en [10], capáz de reducir el espacio de búsqueda de algoritmos de localización. Este bloque aún presenta resultados pobres como detector de actividad de voz debido a la simplificación del modelo de actividad. Un modelo con mayor complejidad [10], así como la utilización de características propias de la voz humana deben mejorar los resultados de este bloque.

- Se ha generalizado el concepto detección de "sectores activos", dado un *array* de micrófonos a detección por "intersección de sectores activos" de múltiples *arrays* de micrófonos. Esta variante presenta características similares a la detección para un solo *array*, con lo cual se logra una reducción mayor del espacio de búsqueda, mientras se mantienen resultados similares como detector de actividad de voz.

- Se han evaluado dos métodos de localización de la fuente de audio de manera puntual, en los volúmenes de intersección activos, SRP y minimización de la métrica en el dominio de fase (PDM) propuesta en Lathoud [10](SCG). Esta última con una modificación para el caso de múltiples *arrays*. El primero de estos métodos obtuvo mejores resultados de precisión.

- Se observó que la métrica basada en fase presentó poca resolución en el eje radial de los *arrays*, lo cual le impide tener mejores resultados de resolución en distancia. Este hecho puede deberse a la distribución particular que presentan los *arrays* de micrófonos en la base de datos Av16.3 con que se ha evaluado el sistema. Queda pendiente una evaluación de dicha métrica frente a otras distribuciones de *arrays* que permitan evaluar el efecto de este factor en el resultado.

- La combinación de detección por sectores y SRP mejora los resultados de precisión de SRP, "filtrando" resultados erróneos de SRP a costa de eliminar algunos *frames* en la detección, incrementando la tasa de borrados.

- Se ha implementado un bloque de detección de rostros multi-pose (paralelo), basado en el algoritmo *Viola and Jones* [9] sobre imágenes. Este tiene la capacidad de detectar rostros con poses frontales y de perfil. Este bloque ha presentado resultados pobres de detección, especialmente en una de las cámaras. Estos problemas se deben a que el sistema no es capáz de detectar poses muy diferentes a las dos entrendas. Un esquema capáz de detectar rostros en un número mayor de poses, permitiría reducir los falsos negativos en todas las cámaras y por tanto más información de vídeo serviría para conformar mejor la información de ocupación. Una alternativa que proporcione información más continua "no *booleana*" sobre la presencia de rostros en una posición concreta serviría como modelo de ponderación de las muestras de vídeo, con vistas a lograr mayor precisión de localización.

- Se han evaluado dos variantes para combinar las proyecciones de todas las cámaras en el plano de ocupación. La primera encontrando la intersección de todas las proyecciones, de acuerdo con el principio de *visual hull*. La segunda se forma a partir de la unión de todas las intersecciones dos a dos. La primera de las variantes resultó ser una condición muy propensa a falsos negativos cuando los índices de detección de alguna de la cámaras no son buenos. La segunda alternativa presentó mejores índices a costa de un aumento en el área de las medidas detectadas.

- Se implementó un modelo de fusión a partir de la unión de las regiones detectadas sobre el plano de ocupación/actividad generadas por audio y vídeo. Este sistema puede ser interpretado como un sensor generalizado que combina información de ambas fuentes. La información generada por el mismo sirvió de entrada a un filtro de partículas XPFCP que se encarga de modelar la dinámica de dichas medidas a través de funciones de densidad de probabilidades y entregar el resultado de localización.

- Se evaluó el modelo de fusión desarrollado frente a dos alternativas unimodales de seguimiento, usando solamente información de audio o de vídeo. El modelo de fusión fue capáz de mejorar el parámetro *deletion* de los seguidores unimodales, combinando detecciones de ambas fuentes. La reducción del error fue superior cuando la detección de vídeo fue inferior. En cuanto a precisión, el resultado de la fusión fue superior a la versión de audio y ligeramente inferior al de vídeo en la versión consecuente con *visual hull*, superando en el otro caso como promedio a las versiones unimodales.

- Estos resultados pueden ser mejorados con modelos de fusión y seguimiento que incluyan por un lado la debida ponderación de las muestras de cada fuente de información, así como en la mezcla de ambas fuentes de manera que se realice un pesado de las partículas del filtro que aproxime mejor las funciones de densidad de probabilidades.

- Una alternativa que permita realizar el seguimiento en 3D reduciría el error relacionado con las diferencias de altura entre el plano seleccionado y la altura real de la fuente.

5.2. Líneas de Trabajo Futuro

Teniendo en cuenta las conclusiones del trabajo, se plantean las siguientes líneas futuras:

- Hacer un estudio detallado de los resultados y sus errores, para afinar y validar algunas de las conclusiones obtenidas

- Proponer esquema de seguimiento que sea capáz de localización en altura variable o plenamente en 3D.

- Implementar el modelo probabilístico de actividad acústica en un sector propuesto por Lathoud [10] para estimar el umbral de actividad de un sectores de manera automática. Esto contribuiría a mejores índices de detección del sistema de audio y por tanto mas información de audio llegará al algoritmo de seguimiento.

- Evaluar la métrica PDM propuesta en este trabajo en nuevas bases de datos con otras configuraciones de *arrays* que permitan evaluar nuevamente su capacidad de localización.

- Desarrollar el bloque de detección de rostros con un número mayor de poses y proponer un índice continuo con sentido probabilístico como salida del detector de rostro. Estas capacidades permitirán mejorar los índices de falsos positivos y negativos del sistema y se proporcionará información para reducir errores de precisión.

- Desarrollar un modelos donde se utilice la información probabilística de los detectores de rostros para ponderar las muestras de vídeo en el *grid* de ocupación y la información de potencia SRP para el *grid* de actividad sonora. Igualmente combinar de manera probabilística los resultados de ambos bloques y que esta información sirva para incidir en el peso asociado a las partículas.

- Implementar el sistema de localización basado en fusión audiovisual en tiempo real en el En el Espacio Inteligente del Departamento de Electrónica de la Universidad de Alcalá de Henares.

Parte IV

Bibliografía

Bibliografía

[1] D. A. Jiménez, "Localización, seguimiento y pose de múltiples interlocutores utilizando fusión audiovisual," Tech. Rep., 2007.

[2] A. M. party Interaction (AMI) project, "State of the art overview: Localization and tracking of multiple intelocutors with multiple sensors," Tech. Rep., 2007.

[3] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual probabilistic tracking of multiple speakers in meetings," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 2, p. 601, 2007.

[4] M. Marrón-Romera, "Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos," Ph.D. dissertation, Escuela Politécnica Superior. Universidad de Alcalá. Spain, 2009.

[5] C. Castro, "Speaker localization techniques in reverberant acoustic environments," Master's thesis, School of Electrical Engineering. Royal Institute of Technology (KTH). Sweden, 2007.

[6] M. C. Aguilar, "Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos," 2007, bachelor Thesis.

[7] ——, "Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual," Master's thesis, Escuela Politécnica Superior. Universidad de Alcalá. Spain, 2010.

[8] E. M. noz Herraiz, "Diseño, implementación y evaluación de técnicas de localización de fuente y de mejora de la señal de habla en entornos acústicos reverberantes: aplicación a sistemas de reconocimiento automático de habla," Master's thesis, ETSI Telecomunicación. Universidad Politécnica de Madrid. Spain, 2005.

[9] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE CVPR 2001*. Citeseer.

[10] G. Lathoud, "Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays," Ph.D. dissertation, Lausanne, Switzerland, 2006, phD Thesis #3689 at the École Polytechnique Fédérale de Lausanne (IDIAP-RR 06-77).

[11] K. Okuma, A. Taleghani, N. D. Freitas, O. D. Freitas, J. J. Little, and D. G. Lowe, "A boosted particle filter: Multitarget detection and tracking," in *In ECCV*, 2004, pp. 28–39.

[12] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1436–1449, 2006.

[13] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 265–268, mar 2005.

[14] C. Castro García, "Speaker localization techniques in reverberant acoustic environments," Master's thesis, Royal Institute of Technology (KTH), Stockholm, 2007.

[15] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *ICASSP '97: Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97) - Volume 1.* Washington, DC, USA: IEEE Computer Society, 1997, p. 187.

[16] J. Ramirez, J. Gorriz, and J. Segura, "Voice Activity Detection. Fundamentals and speech recognition system robustness," *M. Grimm, and K. Kroschel, Robust Speech Recognition and Understanding*, pp. 1–22.

[17] J. H. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.

[18] B. Friedlander and A. Weiss, "Direction finding for wide-band signals using an interpolated array," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1618–1634, 1993.

[19] J. jacques Fuchs, "On the application of the global matched filter to doa estimation with uniform circular arrays," *IEEE Trans. Signal Process*, pp. 702–709, 2001.

[20] B. M. Parham, B. Mungamuru, and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, pp. 1526–1540, 2004.

[21] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," *INTERSPEECH 2005*, pp. 2337–2340, sep 2005.

[22] F. Abad, "A multi-microphone approach to speech processing in a smart-room environment," Ph.D. dissertation, Universidad Politécnica de Cataluña, feb 2007.

[23] C. Hue, P. Le Cadre, and P. Pérez, "A particle filter to track multiple objects," *IEEE Workshop on Multi-Object Tracking*, pp. 61–68, jul 2001.

[24] R. Kalman, *A new approach to linear filtering and prediction problems.* Transactions of the ASME-Journal of Basic Engineering, mar 1960, vol. 82.

[25] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Transactions on Speech and Audio Processing*, vol. SAP-11, no. 6, pp. 826–836, November 2003.

[26] R. Hsu, M. Abdel-Mottaleb, and A. Jain, "Face detection in color images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 696–706, 2002.

[27] R. J. Peñín, "Reconstrucción volumétrica exacta a partir de múltiples cámaras y su aplicación a los espacios inteligentes," Tesis de Master, Universidad de Alcalá, 2010.

[28] T. Starner, B. Leibe, D. Minnen, T. Westyn, A. Hurst, and J. Weeks, "The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and 3D reconstruction for augmented desks," *Machine Vision and Applications*, vol. 14, no. 1, pp. 59–71, 2003.

[29] A. Hoover and B. Olsen, "Sensor network perception for mobile robotics," in *IEEE International Conference on Robotics and Automation*, vol. 1. IEEE; 1999, 2000, pp. 342–347.

[30] V. E. Gómez, "Sistema de detección de obstáculos y robots mediante múltiples cámaras en espacios inteligentes," Master's thesis, Escuela Politécnica de Alacalá de Henares, dec 2008.

[31] J. Meynet, T. Arsan, J. Mota, J. Thiran, and E. de Lausanne, "Fast Multiview Face Tracking with Pose Estimation," 2007.

[32] P. F. Gabriel, J. G. Verly, J. H. Piater, and A. Genon, "The state of the art in multiple object tracking under occlusion in video sequences," in *In Advanced Concepts for Intelligent Vision Systems (ACIVS), 2003*, 2003, pp. 166–173.

[33] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 520–529, 2004.

[34] B. Vo, S. Singh, and W. Ma, "Tracking multiple speakers using random sets," in *IEEE Conference Proceedings on Acoustics, Speech, and Signal Processing*, vol. 2.

[35] M. Isard and J. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *ICCV*, 2001, pp. 34–41.

[36] T. Yu and Y. Wu, "Collaborative tracking of multiple targets," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, pp. 834–841, 2004.

[37] R. Cutler and L. Davis, "Look who is talking: Speaker detection using video and audio correlation," *Proc. IEEE Int. Conf. Multimedia (ICME)*, pp. 1589—-1592, jul 2000.

[38] D. Zotkin, R. Duraiswami, and L. S. Davis, "Active speech source localization by a dual coarse-to-fine search," *IEEE Internacional Conference on Acoustic, Speech and Singal Processing*, vol. 5, pp. 3309–3312, may 2001.

[39] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Inf. Fusion*, vol. 3, no. 2, pp. 209—-223, sep 2001.

[40] M. J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *in Proc. ECCV*, 2002.

[41] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. IEEE*, vol. 92, no. 3, pp. 485—-494, mar 2004.

[42] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.(ICASSP)*, pp. 881—-884, may 2004.

[43] D. Gatica-Perez, G. Lathoud, I. McCowan, J. Odobez, and D. Moore, "Audio-visual speaker tracking with importance particle filters," in *2003 International Conference on Image Processing, 2003. ICIP 2003. Proceedings*, 2003.

[44] D. Gatica-Perez, G. Lathoud, I. McCowan, and J. Odobez, "AMixed-STATE IParticle FILTER FOR MULTI-CAMERA SPEAKER TRACKING," 2003.

[45] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L. wei He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverberg, "Distributed meetings: A meeting capture and broadcasting system," in *In ACM Multimedia*, 2002, pp. 503–512.

[46] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," 2003.

[47] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell, "A multi-modal approach for determining speaker location and focus," *Proc. Int. Conf. Multimodal Interfaces (ICMI)*, pp. 77—-80, 2004.

[48] C. B. Sergi, S. Hernanz, C. wei Chu, S. il Kwon, S. Lee, P. G. Georgiou, I. Cohen, and S.Ñarayanan, "Smart room: Participant and speaker localization and identification," in *in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP*, 2005.

[49] H. Asoh, F. Asano, T. Yoshimura, K. Yamamoto, Y. Motomura, N. Ichimura, I. Hara, and J. Ogata, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion," in *in Proc. Int. Conf. on Information Fusion (IF*, 2004, pp. 805–812.

[50] D. Gatica-perez, G. Lathoud, J. marc Odobez, and I. Mccowan, "Multimodal multispeaker probabilistic tracking in meetings," in *in Proc. Int. Conf. on Multimodal Interfaces (ICMI*, 2005, pp. 183–190.

[51] J. S. Liu, *Monte Carlo Strategies in Scientific Computing.* Springer, 2001.

[52] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking," 2003.

[53] S. Gannot, J. Benesty, J. Bitzer, I. Cohen, S. Doclo, R. Martin, and S.Ñordholm, "Advances in multimicrophone speech processing," *EURASIP journal on applied signal processing*, vol. 2006, p. 12, 2006.

[54] W. Kellermann, "A self-steering digital microphone array."

[55] D. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 12, pp. 499–508, sep 2004.

[56] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *In EUROSPEECH*, 2003, pp. 1009–1012.

[57] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *NEURAL NETWORKS*, vol. 6, no. 4, pp. 525–533, 1993.

[58] "Página de la librería opencv," http://opencv.willowgarage.com/wiki/ [último acceso septiembre 2010].

[59] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, second edition ed. Cambridge University Press, Cambridge, UK, 2003.

[60] A. Doucet, Ph.D. dissertation.

[61] J. C. Jiménez, "Comparativa teórica y empírica de métodos de clasificación aplicados a medidas tridimensionales de visión," Master's thesis, Universidad de Alcalá, jul 2007.

[62] "Av16.3: an audio-visual corpus for speaker localization and tracking," http://www.idiap.ch/av16_3corpus [último acceso mayo 2010].

[63] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16.3: An audio-visual corpus for speaker localization and tracking," in *MLMI*, 2004, pp. 182–195.

[64] "D7.4 evaluation packages for the first chil evaluation campaign," http://chil.server.de/servlet/is/2712/ [último acceso mayo 2009].

[65] M. Omologo, A. Brutti, and P. Svaizer, "Speaker Localization and Tracking-Evaluation Criteria," 2005.

[66] "Página del instituto de investigación idiap," http://www.idiap.ch/ [último acceso septiembre 2010].

[67] D. Moore, "The IDIAP smart meeting room," *IDIAP Communication 02*, vol. 7, 2002.

[68] "Página de la aplicación ffmpeg," http://www.ffmpeg.org/ [último acceso septiembre 2010].