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 localización de locutores basado en el modelado de arrays de micrófonos como cámaras de perspectiva

Alejandro Legrá Rios

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 localización de locutores basado en el modelado de arrays de micrófonos como cámaras de perspectiva"

Autor: Alejandro Legrá Rios

Directores: Javier Macías Guarasa y Daniel Pizarro Pérez

Tribunal:

Presidente: Manuel Mazo Quintas

Vocal 1º: Marta Marrón Romera

Vocal 2º: Daniel Pizarro Pérez.

Calificación: .....

Fecha: .....

A mi hija y mis padres.

Agradecimientos

Quiero agradecer a mis tutores Javier Macías y Daniel Pizarro por brindarme todo su apoyo, tiempo y conocimientos. Gracias por sus oportunos consejos y sugerencias sin los cuales no hubiera podido llegar hasta aquí.

A mis compañeros del grupo, a Yainet, a mis padres y familia por darme las fuerzas y el ánimo cuando mas lo he necesitado.

Índice general

Índice de figuras

Índice de tablas

Parte I

Resumen

Resumen

Las tareas de localización y seguimiento de personas en el contexto de los espacios inteligentes, son fundamentales para mejorar la interacción con el entorno. En este trabajo se implementa un sistema de localización de locutores activos, basado en la información obtenida por varios *arrays* de micrófonos, ubicados en un espacio inteligente. Se implementa una nueva estrategia del uso de la información acústica a partir del modelado de los *arrays* de micrófonos como cámaras de perspectiva, que permite obtener imágenes con información de la potencia acústica en todas las direcciones que conforman el espacio de búsqueda y aplicar técnicas de visión por computador para realizar el posicionamiento. El sistema es evaluado con los datos proporcionados por las bases de datos de AIT e ITC del proyecto CHIL (Computer in the Human Interaction Loop) financiado por la Unión Europea, donde se definen un conjunto de métricas que son utilizadas para evaluar los resultados del sistema de localización.

Palabras clave: Localización de hablantes, *array* de micrófonos, cámaras generalizadas.

Parte II

Abstract

Abstract

The tasks of locating and tracking people in a smart room spaces are essential to improve the interaction between the system and the environment. In this work, we present a tracking system for active speakers, using the information provided by severals arrays of microphones placed in a smart space. We have implementing a new strategy modeling the microphone array as perspective cameras, that generates images with information related to the acoustic power. The system uses computer vision techniques to finally estimate the user position and it has been evaluated with data extracted from the 2007 CLEAR evaluation campaign, generated in the EU CHIL project (Computer in the Human Interaction Loop) financed by the European Union.

Keywords: Speaker location, microphone array, generalized cameras.

Parte III

Memoria

Capítulo 1

Introducción

1.1. Presentación

El análisis automático de espacios inteligentes a partir del procesamiento de múltiples sensores es un área de cada vez mayor actividad científica. En ese contexto, las tareas de detección, localización y seguimiento de personas son fundamentales para mejorar los procesos de interacción con el entorno, o con otras personas u objetos dentro del mismo [1].

Las áreas de explotación de dichas tareas abarcan tanto aspectos ligados al procesamiento de señal (por ejemplo técnicas de mejora de la señal de habla captada por micrófonos lejanos [2], [3], dada la fuerte sensibilidad de la misma a los problemas de reverberación, ruido aditivo y baja relación señal a ruido [4], [5] o técnicas de identificación de locutores y de detección de eventos acústicos localizados), como aquellos relacionados con el análisis de las interacciones humanas dentro del entorno, y de los humanos con otros elementos (por ejemplo robots móviles [6]).

El Grupo de Ingeniería Electrónica aplicada a Espacios Inteligentes y Transporte (GEINTRA) del Departamento de Electrónica de la Universidad de Alcalá ha arrancado una línea de actividad en la que se plantean trabajos orientados a la explotación conjunta (fusión) de la información acústica generada por locutores activos y captada por *arrays* de micrófonos, y la procedente de capturas de video del entorno, para mejorar la interacción de estos en espacios inteligentes.

El trabajo que aquí se propone está orientado a analizar la viabilidad de una nueva estrategia de uso de la información acústica, a través del modelado de los *arrays* de micrófonos como cámaras de perspectiva, y la explotación posterior de dicha información visual con técnicas de procesamiento de imágenes.

Se parte de trabajos iniciados por los Proyectos Fin de Carrera de Eva Muñoz Herraiz [7] ("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"), Carlos Castro González [8] ("Speaker Localization Techniques in Reverberant Acoustic Environments"), María Cabello Aguilar [9](*C*omparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos") [10] ("Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual").

1.2. Motivación y objetivos

Se dispone de un espacio dotado de agrupaciones *("arrays")* de micrófonos. Mediante la utilización de técnicas de localización de fuentes sonoras basadas en información de audio como SRP *(Steered Response Power)* se puede dirigir el patrón de recepción de los *arrays* de micrófonos a posiciones específicas del espacio permitiendo que estos puedan ser utilizados para medir la potencia acústica en distintas posiciones del espacio. Mediante esta técnica se plantea la posibilidad de generar imágenes a partir de información acústica, cuya representación se puede modelar mediante modelos matemáticos propios de proyecciones en cámaras de perspectivas [11]. La información visual puede ser utilizada por distintos algoritmos de posicionamiento de imágenes y de visión por computador para realizar la localización y seguimiento del locutor, tareas que son fundamentales para mejorar la interacción con el entorno. En esta Tesis de Máster se persiguen los siguientes objetivos:

- Diseñar e implementar un sistema de generación de imágenes a partir de la información acústica procedente de múltiples agrupaciones de micrófonos siguiendo el esquema de bloques mostrado en la Figura 1.1.

- Desarrollar algoritmos de tratamiento de las imágenes acústicas para la localización de fuentes acústicas.
- Aplicar técnicas de triangulación para la obtención de la posición tridimensional de las diferentes fuentes acústicas mediante múltiples *arrays* de micrófonos [11].
- Evaluar los algoritmos implementados, realizando experimentos con el software desarrollado y las bases de datos multimodales disponibles en GEINTRA. La evaluación cumpliría los siguientes requisitos:
  - Medir las prestaciones de la algorítmica desarrollada en las aplicaciones que se generen, en diferentes condiciones acústicas reales (en función de las bases de datos CLEAR 2007).
  - Buscar conclusiones razonadas sobre la validez de los resultados obtenidos con las técnicas implementadas. Además, se hará un estudio detallado que ofrezca información sobre la relevancia de los parámetros de control de la experimentación desde un punto de vista práctico.
  - Interpretar los resultados obtenidos a la vista de su fiabilidad estadística, considerando en su justa medida las mejoras o degradaciones observadas respecto a los sistemas de partida.

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 introducen los temas relacionados con la implementación de una nueva estrategia, para realizar la localización utilizando técnicas de imágenes a partir de señales acústica, mediante el modelado de *arrays* de micrófonos como cámaras de perspectivas. Se termina con la explicación de 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 la influencia de varios parámetros de configuración de los experimentos sobre los resultados. Por último se realiza la propuesta del sistema final.
- **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.

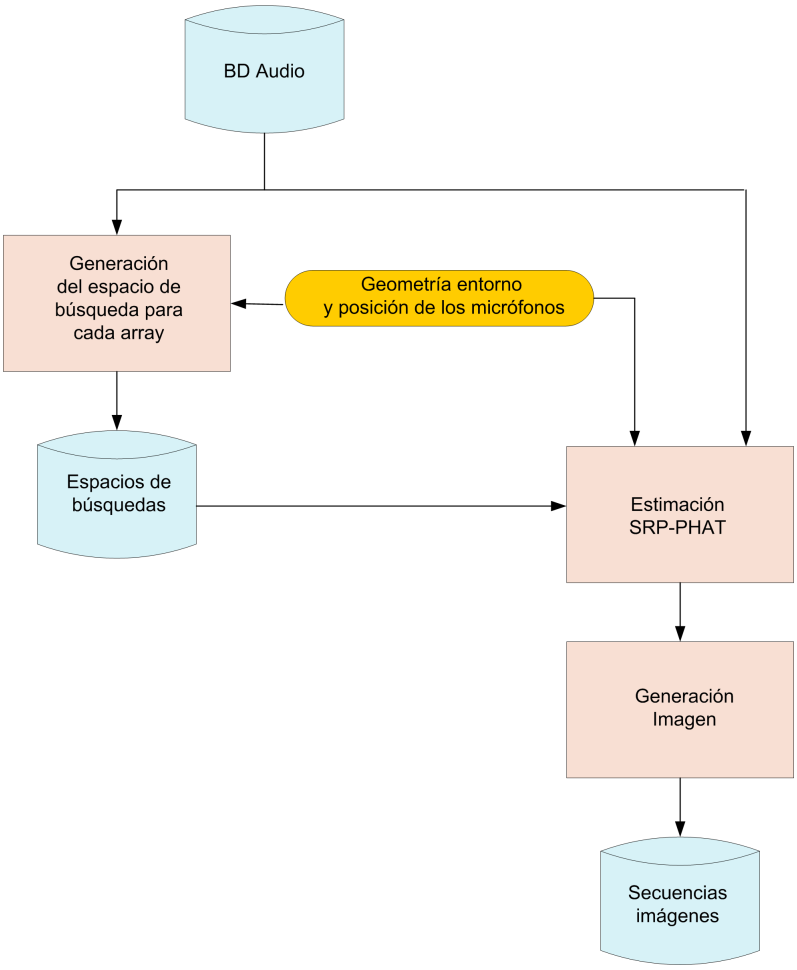


Figura 1.1: Arquitectura del sistema de generación de imágenes a partir de información acústica

Capítulo 2

Estudio teórico

2.1. Introducción

Los espacios inteligentes [12] permiten a los usuarios interactuar con el entorno de manera transparente a través del procesamiento y la interacción conjunta de múltiples señales. En este trabajo se pretende realizar el análisis de la información acústica proveniente de *arrays* de micrófonos para implementar un sistema de localización que permita estimar la posición de la persona que está hablando, tarea que es fundamental para mejorar los procesos de interacción con el entorno o con otras personas dentro del mismo. En esta tesis se propone enfoque novedoso, donde las señales de audio son interpretadas de manera "visual" mediante la generación de imágenes con información acústica. Dichas imágenes permiten el modelado de los *arrays* de micrófonos como si de cámaras de perspectivas se tratase. En este capítulo se hace un estudio de los principales trabajos que se encuentran en la literatura científica de las técnicas relacionadas en esta área que sirven de base para proponer la solución implementada en esta Tesis de Máster.

2.2. Estado del arte

En esta sección se realiza un estudio teórico de algunos de los trabajos disponibles en la literatura científica de los principales temas que se relacionan con el trabajo propuesto. Primero se hace un estudio de los aspectos que caracterizan a los espacios inteligentes, de las principales técnicas de localización basadas en señales de voz y de las técnicas de visión por computador que permiten la utilización y el procesamiento de las imágenes orientados a los sistemas de localización.

2.2.1. Espacios inteligentes

El espacio inteligente como concepto ha sido propuesto por diferentes autores y ha ido evolucionando separándose en diferentes implementaciones, donde se incluyen las aplicaciones destinadas a la localización. Este enfoque de Espacio Inteligente tiene origen en las áreas de investigación que guardan relación con el interfaz hombre-máquina (HMI); Un espacio inteligente es un entorno delimitado por elementos sensoriales no intrusivos que posibilitan la adaptación del sistema al mundo que perciben y además, permiten a los usuarios interactuar con dicho entorno. El objetivo final de este sistema debe ser prestar un servicio a los usuarios de un modo transparente y a través de una comunicación basada en el lenguaje natural entendido por ambos.

Los espacios inteligentes y la Computación Ubicua, en sus inicios, fueron conducidos por Weiser [12], el cual realiza una clasificación en función de cuatro elementos básicos:

- Ubicuidad. Característica de un sistema global formado por otros múltiples sistemas embebidos con total interconexión entre ellos.
- Conocimiento. Habilidad del sistema para localizar y reconocer lo que acontece en el entorno y su comportamiento.
- Inteligencia. Capacidad de adaptación al mundo que percibe.
- Interacción Natural. Capacidad de comunicación entre el entorno y los usuarios.

A finales de los 80, en el proyecto Xeroc PARC de Computación Ubicua (UbiComp) [13], se propuso uno de los primeros sistemas ubicuos implementados con esta filosofía. La red de sensores desplegada no estaba basada en información visual debido al coste prohibitivo para

la época. En la actualidad varios grupos de investigación están trabajando en el desarrollo y la expansión del concepto de espacio inteligente y la computación ubicua. Algunos de los más notorios se indican a continuación:

- *Intelligent Room* [14]. Desarrollado por el grupo de investigación del Artificial Intelligence Laboratory en el MIT (Masachussetts Institute of Technology). Es uno de los proyectos más evolucionados en la actualidad, bajo el nombre del proyecto "Oxygen" [15, 16]. Se trata de una habitación que posee cámaras, micrófonos y otros sensores que realizan una actividad de interpretación con el objetivo de averiguar las intenciones de los usuarios. Se realiza interacción mediante voz, gestos y contexto.

- *Smart Room* [17]. Desarrollado en el Media Lab del MIT bajo el grupo de investigación VisMod ("Vision and Modelling"). Mediante la utilización de cámaras, micrófonos y sensores se pretende analizar el comportamiento humano dentro del entorno. Se localiza al usuario, identificándolo mediante su voz y apariencia y se realiza un reconocimiento de gestos. La línea de investigación actual se centra en la utilización de redes de sensores para su aplicación a productividad en grandes empresas [18].

- Easy Living [19]. Se trata de un proyecto de investigación de la empresa Microsoft con el objetivo de desarrollar "entornos activos", que ayuden a los usuarios en tareas cotidianas. Al igual que otros proyectos comentados, se intenta establecer una comunicación entre el entorno y el usuario mediante lenguaje natural. Se hace uso de visión artificial para realizar identificación de personas e interpretación de comportamientos dentro del espacio inteligente. El proyecto no posee actividad investigadora relevante desde el año 2000.

Muchos de los proyectos mencionados no incluyen robots controlados por el entorno. El uso de agentes controlados por el espacio inteligente es estudiado por un grupo todavía reducido de laboratorios.

- T. Sogo [20] propone un sistema equipado con 16 cámaras fijas llamadas "Agentes de Visión", las cuales aportan la información necesaria para localizar a un grupo de robots a través de un espacio lleno de obstáculos. Los Agentes de Visión consisten en cámaras omnidireccionales no calibradas. La localización se realiza en dos pasos. En primer lugar un operador humano debe mostrar al espacio inteligente el camino que deberán seguir los robots. Una vez que el sistema ha aprendido en coordenadas de imagen el camino señalado, puede actuar sobre los robots con el objetivo de minimizar el camino que recorren en la imagen con respecto al que realizan durante el periodo supervisado.

  La técnica es similar a la utilizada en aplicaciones de servo óptico, por lo que resulta muy difícil estimar la precisión que se obtiene realmente con el sistema. Dependerá en gran medida de la técnica de comparación en el plano imagen y la posición relativa de las cámaras y los robots.

- Un trabajo más evolucionado fue propuesto en la universidad de Tokyo por Lee y Hashimoto [21]. En este caso se dispone de un Espacio Inteligente completo en el que se mueven robots y usuarios. Cada cámara, unida a un sistema de procesamiento es tratado como un dispositivo inteligente conectado a una red, denominado DIND (Distributed Intelligent Network Device). Cada uno de los DIND posee capacidad de procesamiento por si solo, por lo que el espacio inteligente posee flexibilidad a la hora de incluir nuevos nodos inteligentes. Puesto que se cuenta con dispositivos calibrados, la localización se realiza en espacio métrico. Los robots están dotados de balizamiento pasivo mediante un conjunto

de bandas de colores fácilmente detectables por cada DIND. En este espacio de trabajo (Ispace), se realizan experimentos de interacción entre humanos y robots y navegación con detección de obstáculos.

- Otra propuesta es el proyecto MEPHISTO *("Modular and Extensible Path Planning System using Observation")* [22,23] del Instituto para el Diseño de Computadores y Tolerancia a Fallos en Alemania. En este proyecto, el concepto de inteligencia distribuida se alcanza gracias a las denominadas Unidades de Procesamiento de Área Local (LAPU, "Local Area Processing Unit"), similares a los DIND de Lee y Hashimoto. Se define una unidad específica para el control del robot (RCU, "Robot Control Unit") que conecta y envía instrucciones a los diferentes robots. La localización se realiza sin marcas artificiales, usando una descripción poligonal de cada robot en coordenadas de la imagen, la cual es convertida a una posición tridimensional con un enfoque multicámara.
- En el Departamento de Electrónica de la Universidad de Alcalá existe un grupo de "Espacios Inteligentes" (grupo GEINTRA, "Grupo de Espacios Inteligentes y Transporte") [24–26] en el que, mediante un conjunto de sensores, se pretende realizar posicionamiento de robots móviles. Las alternativas incluyen visión artificial, ultrasonidos, infrarrojos y voz.

Estos proyectos hacen hincapié en el diseño de sistemas distribuidos y flexibles, donde la inclusión de un nuevo sensor al entorno se realiza de forma dinámica sin afectar al funcionamiento del sistema. Se han desarrollado a su vez trabajos, que si bien se diferencian en el planteamiento genérico, las técnicas usadas y el objetivo se relacionan directamente con la idea propuesta. Entre dichos trabajos, se pueden citar los siguientes:

- Destaca entre otros el trabajo realizado por Hoover y Olsen [27], en el que utilizando un conjunto de cámaras con coberturas solapadas, son capaces de detectar obstáculos estáticos y dinámicos. Mediante un mapa de ocupación [28] es posible guiar un robot dentro del espacio de cobertura conjunta.
- En [29], se presenta un proyecto similar al anterior denominado MONAMOVE *("Monitoring and Navigation for Mobile Vehicles")* y orientado a la navegación de un vehículo en un entorno industrial. En este trabajo se propone fundir la información de localización, obtenida del conjunto de cámaras distribuidas por el entorno, con sensores de ultrasonidos e infrarrojos ubicados en el robot. Al igual que el trabajo de Hoover y Olsen, se utilizan tablas de búsqueda para crear un mapa de ocupación del entorno.

2.2.2. Técnicas de localización basada en audio

Los sistemas de localización de fuentes sonoras deben de obtener la posición de varios targets activos en el espacio, durante el período de tiempo en que la señal de voz puede ser considerada estacionaria. Para ello desde hace más de 30 años [30–37] se han utilizados los *arrays* de micrófonos y un conjunto de sofisticados mecanismos de procesamiento de la señal que permiten la adquisición remota de las señales de audio con bastante calidad; aprovechando la capacidad del filtrado espacial que tienen los *arrays* para enfatizar la señal capturada en una dirección determinada, atenuando las señales sonoras producidas por personas y fuentes indeseadas. A este procedimiento se le conoce como *beamforming* y permite dirigir el *array* hacia un punto o dirección fija del espacio [8].

La fiabilidad y robustez de los algoritmos de localización se ven afectadas por diferentes problemáticas inherentes al entorno, a las características de la señal de audio y a los *arrays* de micrófonos [8], como se muestra a continuación:

- Los entornos reverberantes provocan que la señal viaje por múltiples trayectorias debido a las reflexiones y difracciones que ocurren cuando esta rebota con los objetos y paredes presentes, provocando que la señal recibida contenga componentes retardadas, atenuadas y distorsionadas [8] como se muestra en la Figura 2.1.

- La baja SNR de la señal recibida, debido a la atenuación que sufre la señal con la distancia, además del ruido que puede ser producido por el fondo, o por otras fuentes sonoras que se encuentren en el entorno.

- La voz humana es una señal de banda ancha. Además tiene un carácter intermitente, por lo que será necesario estimar de manera fiable su presencia en cada instante de tiempo.

- La naturaleza de las conversaciones humanas es muy dinámica y puede haber varias personas hablando simultáneamente (*multi-party speech*), cambios rápidos en el turno de palabra, intervenciones muy rápidas en cuanto a tiempo o movimientos no lineales de determinados participantes.

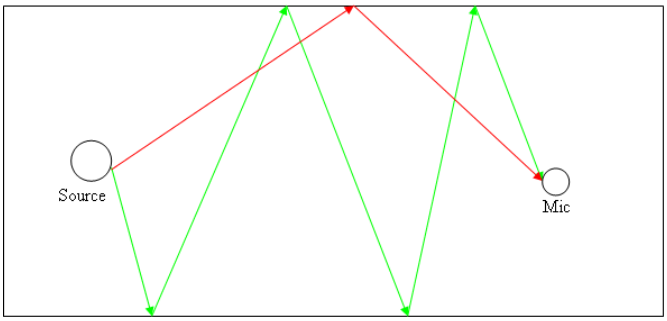


Figura 2.1: Propagación del sonido en el interior de una habitación

En la literatura se han descrito muchas técnicas de localización que utilizan los *arrays* de micrófonos los cuales pueden ser clasificados en tres grandes grupos [38]:

1. Los que utilizan las diferencias en los tiempos de llegada de la señal (*Time Difference of Arrival* (TDOA)) que están divididas en dos fases bien definidas:

   a) Estimación de los retardos en la señal acústica *Time Delay Estimation* (TDE) entre pares de micrófonos separados espacialmente. Según [39] hay dos formas principales para llevar a cabo esta primera etapa:

      - Utilizando la Correlación Cruzada Normalizada (CC). La cual analiza la similitud entre dos señales diferentes para distintos tiempos de retardo, por lo que maximizándola se puede obtener el retardo con que llega la señal a dos micrófonos diferentes [8]. Es una técnica adecuada para el cálculo del TDOA en entornos que se ven afectados por ruido incorrelado. Sin embargo es una técnica que se ve gravemente afectada por las reverberaciones, ya que cada reverberación generará un máximo relativo en la función de auto correlación, afectando negativamente la estimación de la posición.
      - Utilizando la Correlación Cruzada Generalizada (GCC). Esta mejora los resultados obtenidos por CC aplicando un filtro previo a cada una de las señales antes de calcular la correlación cruzada [8]. Para ello en [31] se proponen diferentes funciones de peso que son aplicados a GCC, con el objetivo de obtener los TDOA

reales que mejoran la respuesta la Función de Correlación Cruzada Generalizada en entornos afectados por el ruido y las reverberaciones. Algunas de las más importantes son:

- *Roth Processor:* Suprime las frecuencias donde el espectro de potencia de ruido es mayor.
- *Smoothed Coherente Transform (SCOT):* Mejora los resultados obtenidos con Roth ponderando la señal con pesos dependientes de su SNR.
- *Maximun Likelihood (ML) Estimator*: En esta función de pesos la señal y el ruido deben ser conocidos, por lo que en la práctica deberán ser estimados.
- *Phase Transform (PHAT):* Esta función de pesos normaliza la señal poniendo el mismo énfasis en todas las frecuencias y ha demostrado un mejor funcionamiento que el resto en entornos reales. A pesar de ello sigue siendo subóptima en entornos reverberantes, pues se ve afectada cuando la potencia de la señal es baja, para solucionar este problema en [40] se propone la utilización de un filtro paso banda $(300Hz, 6KHz)$ que deje pasar la banda de frecuencia en las que hay mayor potencia de señal acústica.

b) Una vez obtenidos los TDE y conociendo la posición espacial de los micrófonos, se obtienen curvas hiperbólicas que representan las posiciones en las que existe una mayor probabilidad de ubicación de los targets. Estas curvas se intersectan en puntos concretos del espacio permitiendo la estimación de la posición utilizando determinados criterios de optimización, [41] pags. 31-43. Su implementación es bastante simple y posee bajo coste computacional. Sin embargo se ve muy afectada por los problemas de reverberación del entorno y por la baja SNR. Además en lugar de la localización espacial proporcionan una estimación de la Direction Of Arrival (DOA), teniendo menos robustez que los sistemas basados en beamforming.

2. Los basados en *High-resolution Spectral Estimation*. En estos se estima la distribución de potencia de la señal, para poder detectar los picos de potencia presentes. Dichos métodos se basan principalmente en la explotación de las propiedades de la Cross-Sensor (spatial) Covariance Matriz (CSCM) del *array*, principalmente de sus autovalores, a partir de los cuales son capaces de dividirla en dos subespacios, uno que contiene la señal de habla y otro el ruido. Esta idea fue estudiada por [42] dando lugar a varios algoritmos de banda estrecha como:

   - MUSIC (Multiple Signal Classification): [43].
   - ESPRIT (Estimation of Signal Parameters via Rotacional Invariante Techniques): [44].
   - MIN-NORM (MINimun NORMs): [45].
   - WSF (Weighted Subspace Fitting): [46].

   Entorno a esto Benesty en [47] consigue estimar el retardo en los tiempos de vuelo entre un par de micrófonos basándose en la descomposición en autovalores. La idea fundamental consiste en que el autovector correspondiente al mínimo autovalor de la matriz de covarianza de las señales en los micrófonos, contiene la respuesta al impulso entre la fuente y los micrófonos y por tanto toda la información necesaria para la detección de las ondas directas y la estimación de los tiempos de vuelo.

   La Figura 2.2 muestra la configuración del sistema empleado en [47], el cual emplea altavoces ubicados en posiciones estáticas para generar las señales acústicas y un par de micrófonos separados a una distancia de 95 cm. La frecuencia de muestreo es de 48KHz, la cual es convertida mediante un proceso de submuestreo a 16KHz, y 5 segundos de duración.

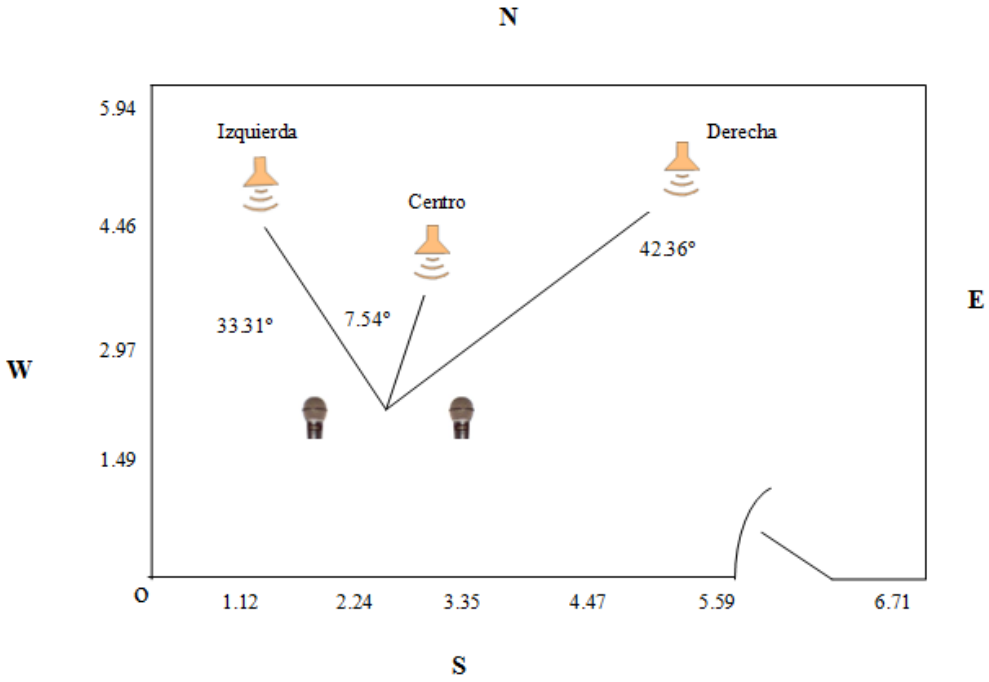


Figura 2.2: Experimento para el cálculo del tiempo de Vuelo con 2 micrófonos

Los resultados presentados por el autor demuestran que el algoritmo converge rápidamente (menos de 250 ms) a buenas estimaciones de TDOA, demostrando además su robustez en entornos con bajas SNR (10 dB) y en presencia de reverberaciones.

Sin embargo estos algoritmos presentan dos limitaciones fundamentales:

a) No son implementables cuando el número de llegadas incorreladas supera al número de sensores del *array*, siendo esta una condición bastante común en entornos reverberantes debido a la dificultad de distinguir entre señales reflejadas y fuentes sonoras adicionales. En el estudio presentado en [47] el número máximo de señales dentro de la ventana de trabajo es de dos (una directa mas una reflejada), por lo que no se reproduce esta problemática.

b) Este conjunto de técnicas funcionan razonablemente bien con señales de banda estrecha, no siendo este el caso de las señales de habla (20Hz, 20KHz). Su extensión a banda ancha genera problemas adicionales. Además la carga computacional es bastante alta.

Se han propuesto diferentes mecanismos para la extensión de este tipo de técnicas a banda ancha, siendo uno de los más destacables el *Coherent Signal Subspace Method (CSSM),* formulado por [34] y desarrollado en [48,49]. Estos métodos estiman la covarianza espacial de múltiples bandas de frecuencia, las cuales son alineadas *(focusing)* y a partir de las cuales se genera un conjunto reducido de estadísticos cuya media tiene la misma estructura de autovalores que los mecanismos de banda estrecha CSCM.

Aunque los métodos basados en CSSM aportan soluciones a algunos de los problemas de los métodos basados en subespacios, su uso en aplicaciones reales queda reducido debido a la limitada precisión del proceso de *focusing* y a la problemática no resuelta de tener que tratar con un mayor número de fuentes que de sensores.

3. Los que utilizan *Steered Response Power* (SRP). Estas técnicas hacen uso de determinadas técnicas de *beamforming*, bien con el objetivo de localizar a la fuente emisora de la señal o de aumentar la calidad de la señal recibida cuando la posición de la fuente emisora de la misma es conocida. En caso contrario el *array* puede ser utilizado para escanear una serie de localizaciones espaciales conocidas, en cuyo caso a la salida del beamformer se la conoce como *steered response* [40].

   La forma más simple de implementar un *steered response* consiste en la utilización de la salida de un *delay-and-sum beamformer*, en la que los diferentes retardos temporales (calculados en función de la posición del espacio a la que se desea apuntar) son aplicados a la señal de entrada de cada uno de los micrófonos, produciendo una alineación temporal de las mismas, que luego se suman para obtener la salida del algoritmo. Un procedimiento más general, denominado *filter-and-sum beamformer* aplica previamente un filtro a cada una de las señales de entrada a los micrófonos.

   Para mejorar las estimaciones del TDOA realizadas a partir de GCC en condiciones de baja SNR y entornos reveberantes se han propuesto varias soluciones como: aumentar el número de micrófonos [40]. utilizar la equivalencia entre SRP y GCC-PHAT de todas las combinaciones posibles de pares de micrófonos [50], esta técnica es conocida como SRP-PHAT y su robustez se basa en la explotación de la redundancia espacial de los micrófonos promediada por todas las posibles combinaciones de parejas GCC-PHAT.

   Los métodos basados en SRP brindan la posibilidad de aceptar un número variable de targets simultáneos [51], alta robustez y fiabilidad, siendo superior a la proporcionada por los métodos basados en TDOA [52], aunque tienen como principal desventaja la alta carga computacional que se ve incrementada con el número de micrófonos y de localizaciones espaciales a las que se debe orientar el *array*. Además el tamaño del *grid* de búsqueda utilizado limita la precisión del algoritmo [53].

También se han encontrados muchos trabajos entre los que se encuentran: el de Brandstein y Silverman que utilizaron la función de ponderación de Biwieght Tukey para contrarrestar el efecto de las reflexiones [50]. Ward and Williamson desarrollaron un filtro de partículas para resolver el problema de la reverberación [54] Potamitis propuso la técnica PDA *(probabilistic data association)* para mejorar los errores de estimación, Chen utilizó un estimador de máxima verosimilitud (ML) para detectar la localización del locutor bajo los efectos *near-far* y *far-field*. Para mejorar la carga computacional del algoritmo (ML), Chung en [55] propuso dos algoritmos recursivos de *Expectation Maximization* (EM) para localizar el locutor. Jwu-Sheng Hu utiliza un *Gaussian Mixture Model (GMM)* para modelar la distribución de las de las diferencias de fase de las distintas localizaciones. Recientemente se han encontrado algunos trabajos [56, 57] que utilizan técnicas de visión a imágenes creadas a partir de potencia acústica.

En el presente trabajo se modela cada *array* de micrófono como una cámara de perspectiva. Se utiliza el algoritmo de SRP-PHAT disponible en el grupo de investigación para integrar la potencia de la señal de audio medida en distintas direcciones por varios *arrays* de micrófonos y formar imágenes a partir de la información de la potencia acústica obtenida. El método propuesto permite utilizar técnicas clásicas del procesamiento de imágenes, cálculo de máximos y triangulación para la estimación de las posiciones.

2.2.3. Técnicas de estimación de máximos

La detección de máximos en los mapas de energía es una tarea fundamental en la estimación de la posición 3D del locutor, en el que cada máximo constituye una posible dirección en el

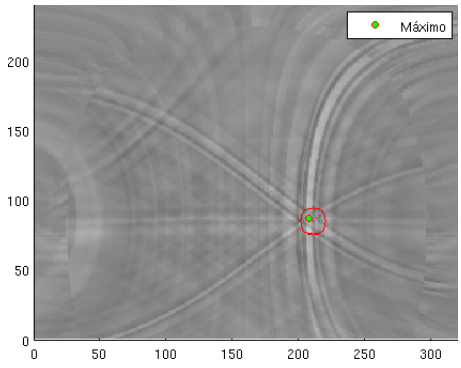


Figura 2.3: En la figura se representa una imagen con información acústica cuyo máximo se representa con un círculo de color verde

que la energía acústica es superior al resto de las direcciones vecinas. En dichos máximos se espera encontrar la actividad acústica generada por los locutores existentes en el espacio donde se encuentran los micrófonos (ver Figura 2.3). Esta fase se puede dividir en dos partes:

- Algoritmo de supresión no máxima para aislar máximos locales y umbralización. Este algoritmo tiene como objetivo ayudar a la búsqueda de máximos locales mediante la eliminación progresiva de la energía alrededor de puntos de acumulación. El resultado es un mapa de energía binario donde posiciones con valor unitario corresponden a máximos locales. Este algoritmo requiere definir un valor de umbral para determinar qué es considerado un máximo local y requiere a su vez una noción de radio de vecindad para suprimir valores de píxeles vecinos al máximo local [58]. Se debe notar que en este paso se obtiene la posición de cada máximo encontrado en posiciones enteras del vector de dirección.
- Detección de cada máximo con precisión subpixélica. Este algoritmo refina las posiciones de los máximos locales obtenidos en la fase anterior. Para ello es necesario hacer una suposición sobre la forma que tienen dichos máximos en su entorno [59]. En la literatura se exponen algunos tipos de funciones que son usadas para realizar este paso (e.g., funciones cuadráticas, gaussianas 2D etc.).

2.2.4. Técnicas de triangulación

En el contexto de visión por computador la triangulación se refiere al proceso de determinar un punto 3D en el espacio realizando la proyección de dos o más píxeles en imágenes diferentes [11]. En la cámara de perspectiva cada píxel dentro de una imagen corresponde a una línea 3D en el espacio, donde todos los puntos en la línea son proyectados a un píxel de la imagen. Si la proyección de un par de píxeles en dos o más imágenes diferentes se intersectan, se obtiene la proyección de un punto 3D [11]. El conjunto de líneas generadas por los puntos de la imagen deben intersectarse en X y la formulación algebraica de las coordenadas de X puede ser calculada de varias formas.

Sin embargo, en la práctica, las coordenadas de los píxeles de la imagen no se pueden medir con exactitud, debido a distintos tipos de ruido (e.g., ruido geométrico debido a la distorsión de la lente o error en el punto de interés detectado) dando lugar a imprecisiones en la medida de las coordenadas de la imagen [11]. Debido a esto, las líneas generadas por los píxeles de las imágenes no siempre se intersectarán en el espacio 3D, por lo que se busca encontrar un punto

3D óptimo. En la literatura se encuentran muchas propuestas para estimar de manera óptima el punto 3D cuando se está en presencia de ruido.

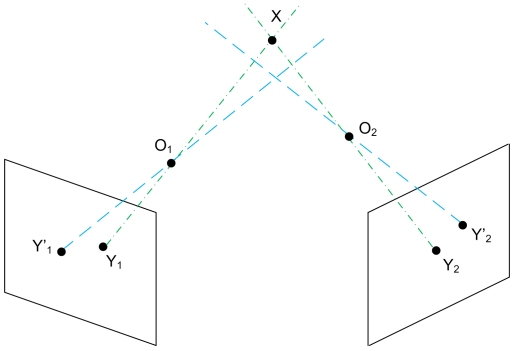


Figura 2.4: Triangulación

La Figura 2.4 muestra el caso ideal en el que la proyección de dos puntos Y_1 *y* Y_2 (representados con la línea verde) de dos imágenes diferentes se intersectan en un punto X en el espacio 3D. Mientras que con la línea azul se muestra el caso real en el que la posición del punto X no se puede encontrar con exactitud. Esto se debe a una serie de factores tales como:

- Distorsión geométrica, como por ejemplo, distorsión de la lente que provoca que proyección del punto 3D en el plano 2D de la cámara se desvíe debido al modelo de perspectiva de la cámara.
- El rayo de luz que viene desde el punto X es dispersado en el sistema de lente de la cámara debido a la respuesta de un sistema imagen proveniente de una fuente puntual.
- Los puntos de la imagen usados para la triangulación con frecuencia son obtenidos usandos algún extractor de características, como por ejemplo bordes o en general puntos de interés. Esto es inherente al error de localización que comenten los extractores de características.

Los algoritmos de triangulación deben suavizar estos problemas estimando el punto de intersección más cercano al real. A continuación se describen algunos de los métodos encontrados en la literatura:

- Mid-point: En este método se busca el punto medio de la perpendicular que une a los dos rayos [60]. Como mejor solución se propone dividir la perpendicular entre los dos rayos en proporción a la distancia desde los dos centros de la cámara. Sin embargo este método no da buenos resultados debido a la utilización de muchas aproximaciones.
- Basado en la matriz fundamental: este concepto fue introducido en 1981 por Longuet-Higgins y es una matriz de 3 x 3, que relaciona los puntos correspondientes en imágenes estéreo, asumiendo que las cámaras satisfacen el modelo de cámara de perspectiva. Mediante la matriz fundamental, se pueden aplicar una serie de métodos para determinar el punto 3D. En [61] se describe un algoritmo, que da una solución global óptima, la solución se basa en los conceptos de correspondencia epipolar y en la matriz fundamental. Es un algoritmo no iterativo que busca minimizar una función de costo determinada.
- Método de triangulación lineal (análogo al DLT). Un punto en el espacio X proyectado en dos puntos ($m(u, v)$ *y* $m'(u', v')$) en dos imágenes distintas de manera que: $m = PX$ *y* $m' = P'X$. Estas ecuaciones pueden ser combinadas en una ecuación lineal en X, $AX = 0$. Para

ello primero se elimina el factor de escala por un producto cruzado ($\times$) para obtener 3 ecuaciones por cada punto de las cuales dos son linealmente independientes [11], por ejemplo para la primera imagen, $m \times (PX) = 0$, esta puede ser escrita de la siguiente forma:

$$\begin{array}{rcl} u(p^{3T}X) - (p^{1T}X) & = & 0 \\ v(p^{3T}X) - (p^{2T}X) & = & 0 \\ u(p^{2T}X) - v(p^{1T}X) & = & 0 \end{array} \tag{2.1}$$

donde p^{iT} son las filas de P. Estas ecuaciones son lineales en los componentes X. Una ecuación en la forma $AX = 0$ puede ser escrita como:

$$A = \begin{bmatrix} up^{3T} - p^{1T} \\ vp^{3T} - p^{2T} \\ u'p'^{3T} - p'^{1T} \\ v'p'^{3T} - p'^{2T} \end{bmatrix} \tag{2.2}$$

Existen dos formas para resolver el sistema de ecuaciones:

- Método homogéneo: Resuelve el sistema de ecuaciones mediante la descomposición en valores singulares (SVD).
- Método no homogéneo: En este método se reducen el sistema de ecuaciones, $AX = 0$, que se encuentran en coordenadas homogéneas $X(x, y, z, 1)^T$ a un sistema de 4 ecuaciones no homogéneas de 3 incógnitas. La solución se obtiene mediante mínimos cuadrados.

2.2.5. Técnicas de estimación de coherencia

En este apartado se hace un estudio sobre algunas de las técnicas que se usan para estimar la coherencia de las proyecciones de los máximos de energía en las imágenes, las cuales definen como *outliers* los datos que son muy diferentes al resto.

Muchos de los algoritmos de detección de *outliers* requieren un conjunto de datos para entrenar el modelo, y estos implícitamente pueden asumir que los *outliers* pueden ser tratados como patrones de datos no observados anteriormente.

Los métodos de determinación de *outliers*, atendiendo al tipo de restricción que utilizan, se han dividido en diferentes categorías que se exponen a continuación:

- Métodos basados en la distribución de los datos [62–64]: Estos emplean algún modelo de distribución estándar, por ejemplo distribución normal, y se consideran *outliers* todos los datos que se desvíen del modelo. Yamanishi en [65] usa un modelo Gaussiano mezclado para presentar el comportamiento normal y a cada dato se le da una puntuación basada en cambios del modelo. Esta aproximación se puede combinar con algún método de aprendizaje supervisado para obtener el patrón general de *outliers* [66]. Sin embargo con el incremento en la dimensionalidad de los datos se incrementa la dificultad y la precisión para estimar las distribuciones de múltiples dimensiones de los datos.
- Basado en algún umbral de los datos [67, 68]: Se define un umbral y los datos que no cumplen el umbral son definidos como *outliers*.
- Basado en la desviación: Esta técnica identifica los *outliers* describiendo las características de los objetos y considerando como *outliers* los que no cumplen con estas características.

- Basados en distancia: Esta técnica fue originalmente propuesta por Knorr y Ng [69,70], se basa en la distancia entre un punto y el vecino más cercano [71]. De manera alternativa el factor de *outlier* de cada punto es calculado como la suma de las distancias desde el vecino más cercano [72]. Bay y Schwabacher [73] presentan un algoritmo que determina el *outlier* mediante una distancia basada en una línea de tiempo. y los K vecinos que superen un umbral de distancia determinado representan un *outlier*. La distancia de Mahalanobis es una de las más usada y existen trabajos que demuestran que es más robusta que la distancia euclidiana [74].
- Basadas en su densidad: Propuesto por Breunig en [75], se basa en un factor de *outlier* (LOF) local a cada punto, que depende de la densidad de vecinos cercanos. Los algoritmos que calculan los LOF de una muestra de datos realizan los siguientes pasos [74]:
  1. Para cada muestra de datos se calculan la distancia a los vecinos más cercanos y luego agrupan las muestras que están a una distancia esférica determinada.
  2. Calcula la distancia que se considera asequible para que una muestra O pueda pertenecer a un grupo determinado.
  3. Calcula la densidad de los datos que se encuentran en un grupo como la inversa del promedio de distancias. Basadas en el mínimo número de muestras que son vecinos de la muestra O.
  4. Calcula la LOF de la muestra de datos O como el promedio de los radios de las muestras que se encuentran agrupadas a una distancia asequible.
- Para mejorar las prestaciones del LOF, Tang en [76] introdujo un factor de *outliers* (COF) basado en la conectividad.
- Basadas en la clusterización de los datos: considerando pequeños cluster como *outliers* [77], que son removidos del conjunto de datos iniciales [78].
- Basadas en sub espacios: Aggarwal and Yu [79], realizan la detección de los *outliers* observando la distribución de densidad de las proyecciones de los datos.
- Basadas en un *Support Vector*: El *Support Vector Novelty Detector* lo propuso Tax and Duin [80].

En [81] se expone una variante del método *Support Vector Machines* (SVM) que no requiere ningún entrenamiento de los datos. Este algoritmo trata de separar los datos en dos clases de datos distintas espaciadas en un hiperplano.

Fischler Bolles en 1981 propuso el algoritmo RANSAC *(Random Sample Consensus)*, el cual es un método iterativo que realiza la estimación de parámetros desde un modelo matemático que contiene *outliers*. Es una técnica que es muy utilizada para la estimación de parámetros en un modelo, usando datos que pueden estar contaminados por *outliers*. RANSAC estima una relación global que se adapta a los datos, mientras de manera simultánea clasifica los datos y detecta los *outliers*.

2.3. Propuesta

En este trabajo, se propone implementar un sistema de localización basado en señales de voz, mediante el modelado de *arrays* de micrófonos como cámaras de perspectivas. Se utiliza la información visual obtenida, para realizar la estimación de la posición de la persona que está hablando. El sistema se divide en las siguientes etapas:

- Cálculo de la potencia acústica, en las distintas direcciones que cubren el espacio de búsqueda, donde se puede encontrar el locutor. Para realizar el cálculo de la potencia acústica, se plantea el uso de la técnica de SRP con filtrado PHAT (SRP-PHAT), que permite dirigir el patrón de recepción del *array* de micrófonos hacia las posiciones deseadas. Dicha técnica de localización es la que presenta mejores prestaciones en entornos reverberantes. Esta estrategia permite formar una imagen en la que cada píxel represente la potencia acústica medida en una dirección.
- Búsqueda de la dirección de máximos de potencia acústica, utilizando el algoritmo de *Non Maximun Supression* con aproximación subpixélica, a partir de una función cuadrática.
- En las salas donde hayan más de 4 *arrays* de micrófonos, utilizar una variante de la técnica RANSAC para eliminar las estimaciones de máximos de energía que no sean coherentes con el resto.
- Realizar el cálculo de la posición 3D a partir de la técnica de triangulación lineal DLT.

2.4. Conclusiones

En este capítulo se expusieron las características de los espacios inteligentes, entorno en el cual se implementará el sistema de localización propuesto. Se hizo un análisis de las técnicas más empleadas en la literatura por los sistemas de localización basados en audio, seleccionando la técnica SRP-PHAT como la más adecuada para medir la potencia acústica en los entornos reverberantes, se expusieron las principales técnicas de búsqueda de máximos de manera robusta. A su vez se plantearon las principales técnicas de visión por computador utilizadas para triangular y obtener el posicionamiento 3D de un objeto determinado, profundizando en el algoritmo DLT. Además se hizo, un estudio de los principales algoritmos que permiten estimar errores por coherencias *(outliers)* en los máximos de potencia detectados, dentro de las imágenes que pueden afectar el sistema de posicionamiento. Finalmente se ha descrito de manera resumida la propuesta que se aborda en esta Tesis de Máster.

Capítulo 3

Desarrollo algorítmico

3.1. Introducción

En este capítulo se profundiza en los aspectos teóricos fundamentales del desarrollo de la Tesis de Máster y se detalla, todo el desarrollo algorítmico seguido para la implementación del sistema de localización, analizando la información acústica brindada por *arrays* de micrófonos mediante técnicas de visión por computador.

3.2. Modelado de cámaras

Una cámara proporciona información del mundo 3D sobre una imagen en dos dimensiones, siendo necesario realizar una transformación de un espacio a otro [11]. La forma en la que se procesa dicha transformación, se explica mediante el modelo de una cámara de perspectiva, utilizado para generar las localizaciones donde se calcularán la potencia acústica a través de SRP_PHAT.

En el modelo de cámara de perspectiva un rayo que viene desde un punto en el espacio pasa a través de la lente de una cámara e incide en una película o dispositivo digital produciendo un punto en la imagen. Ignorando los efectos del foco y del espesor de la lente, se puede decir de manera aproximada que todos los rayos pasan por el centro de la lente, e intersectan el plano imagen donde los puntos en el espacio 3D son mapeados al plano imagen, [11]. Este proceso es denominado proyección central, considerando además, que todos los puntos de un rayo son iguales y que pasan a través del centro de proyección, formando un punto en la imagen. Se puede decir que el conjunto de puntos en una imagen, es el conjunto de rayos que pasan a través del centro de la cámara [11].

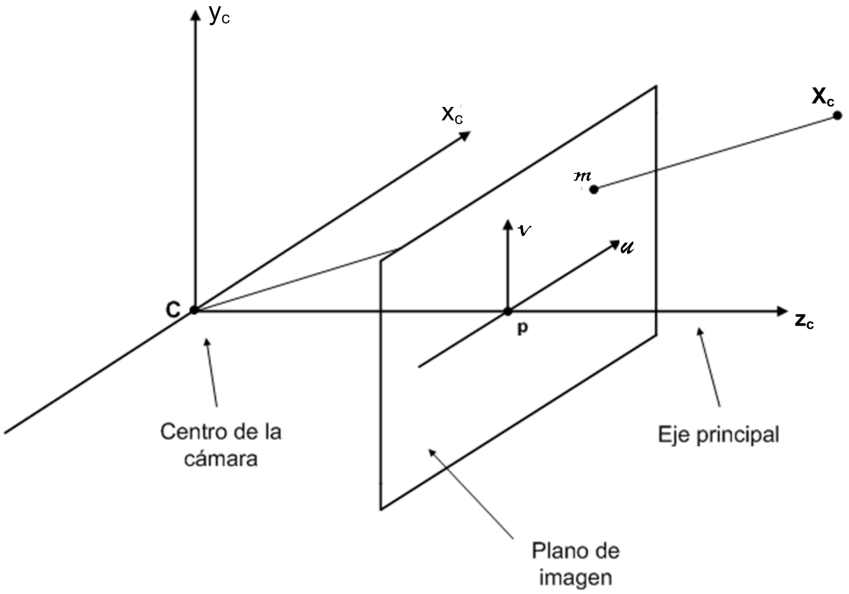
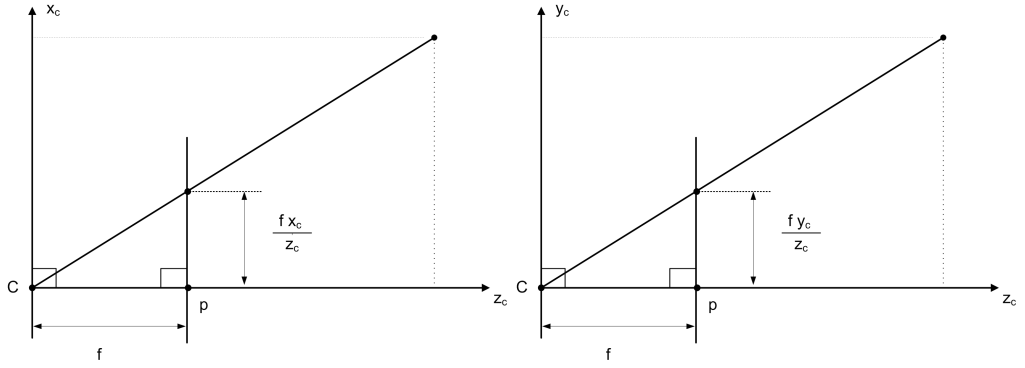


Figura 3.1: Geometría de la cámara de perspectiva

Considerando la proyección de los puntos del espacio sobre un plano, se tiene que el centro de proyección es el origen del sistemas de coordenadas euclidianas y que $z_c = f$ es el plano imagen o plano focal. En una cámara de perspectiva, un punto en el espacio con coordenadas $X_c = (x_c, y_c, z_c)^T$ es mapeado a un punto m en el plano imagen, donde se intercepta con la línea que une al punto X_c y el centro de proyección de la cámara [11], como se muestra en la Figura

3.1.



(a) Proyección en el eje x_c del plano imagen (b) Proyección en el eje y_c del plano imagen

Figura 3.2: Proyección de un punto 3D en la plano imagen

Aplicando semejanzas triangulares (ver Figura 3.2) se pueden proyectar las coordenadas $(x_c, y_c, z_c)^T$ del punto en el plano imagen como $(fx_c/z_c, fy_c/z_c, f)^T$ [11]. Sin tener en cuenta las coordenadas final de la imagen se obtiene la ecuación 3.1 que describe la proyección central de un punto del espacio 3D en plano de la imagen.

$$(x_c, y_c, z_c)^T \mapsto (f\frac{x_c}{z_c}, f\frac{y_c}{z_c})^T \tag{3.1}$$

Si el mundo y los puntos de la imagen son representados como vectores homogéneos, la proyección central se puede expresar como la proyección lineal entre sus coordenadas homogéneas [11]. La ecuación 3.1 se puede escribir en términos de la matriz de multiplicación [11] como:

$$\begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fx_c \\ fy_c \\ z_c \\ 1 \end{pmatrix} = \begin{bmatrix} f & & & 0 \\ & fy_c & & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} \tag{3.2}$$

Definiendo las coordenadas del punto (X_c) respecto al sistema de coordenadas de la cámara como el vector homogéneo $(x_c, y_c, z_c, 1)^T$, las coordenadas del plano imagen m mediante el vector homogéneo $(u, v, 1)^T$ y P_h como la matriz homogénea de proyección de la cámara [11], se puede calcular las coordenadas de la imagen de la siguiente forma:

$$m = P_h X_c \tag{3.3}$$

Hasta el momento se ha estado trabajando con el centro de la cámara como el origen del sistema de coordenadas. Sin embargo en la práctica el origen suele estar situado en un extremo del plano imagen, (ver Figura 3.2). La transformación se modela sumando a la coordenada (u, v) la posición del centro de la cámara (p_u, p_v) [11]. La expresión 3.1 se transforma en:

$$(x_c, y_c, z_c)^T \mapsto (f\frac{x_c}{z_c} + p_u, f\frac{y_c}{z_c} + p_v)^T \tag{3.4}$$

La ecuación puede ser expresada en coordenadas homogéneas como:

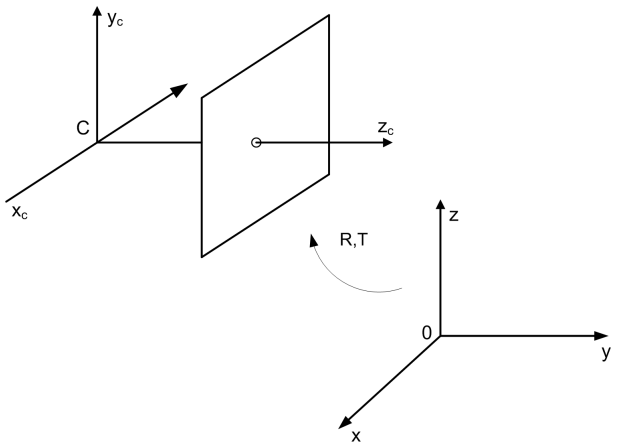


Figura 3.3: Transformación euclidiana entre los sistemas de coordenadas de la cámara y el mundo

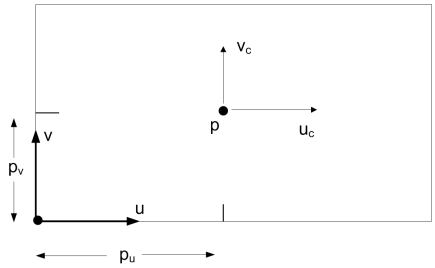


Figura 3.4: Coordenadas de la imagen (u_c, v_c) y de la cámara (u_c, v_c)

$$\begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fx_c + z_c p_u \\ fy_c + z_c p_v \\ z_c \\ 1 \end{pmatrix} = \begin{bmatrix} f & & p_u & 0 \\ & f & p_v & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} x_c \\ y_c \\ z_c \\ 1 \end{pmatrix} \tag{3.5}$$

A partir de la ecuación 3.5 se puede definir el parámetro de calibración K como:

$$K = \begin{bmatrix} f & & p_u \\ & f & p_v \\ & & 1 \end{bmatrix} \tag{3.6}$$

Para obtener las coordenadas de la imagen en píxel es necesario dividir cada coordenada por la anchura de cada píxel d_u y d_v de la siguiente forma:

$$u = \frac{1}{d_u}\left(f\frac{x_c}{z_c} + p_u\right) \tag{3.7}$$

$$v = \frac{1}{d_v}\left(f\frac{y_c}{z_c} + p_v\right) \tag{3.8}$$

Sustituyendo el parámetro *skew* S [82], que modela la no perfecta perpendicularidad de los ejes en la ecuación 3.6 la matriz de calibración se expresa como:

$$K = \begin{bmatrix} \alpha_u & S & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.9}$$

donde $\alpha_u = \frac{f}{d_u}$, $\alpha_v = \frac{f}{d_v}$, $u_0 = \frac{1}{f}(p_u)$ y $v_0 = \frac{1}{f}(p_v)$

La matriz de calibración K describe los parámetros intrínsecos de la cámara.

En general los puntos en el espacio deben ser expresados en términos de un sistema de coordenadas euclidiano, conocido como sistemas de coordenadas del mundo, (ver Figura 3.3). El sistema de coordenadas de la cámara (x_c, y_c, z_c) se relaciona con el sistema de coordenadas del mundo mediante las matrices de rotación R y de traslación T.

- La matriz de rotación R representa el cambio de dirección de un sistema respecto a otro, y dependerá de los ángulos de rotación (α, β, θ) de los ejes de coordenadas de la cámara respecto a los del mundo, (ver Figura). Las matrices de rotación se calculan como:

$$R_\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \tag{3.10}$$

$$R_\beta = \begin{bmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{bmatrix} \tag{3.11}$$

$$R_\alpha = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.12}$$

$$R = R_\theta R_\beta R_\alpha \tag{3.13}$$

La matriz T expresa el lugar donde se encuentra el centro óptico de la cámara (origen del sistema de coordenadas de la cámara), respecto al origen de coordenadas del mundo.

$$T = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} \tag{3.14}$$

Las matrices R y T describen los parámetros extrínsecos de la cámara. La matriz de proyección P permite pasar del sistema de coordenadas del mundo al sistema de coordenadas de la cámara y depende de los parámetros extrínsecos e intrínsecos de la cámara [82]. La matriz P se calcula como:

$$P = K(R;T) \tag{3.15}$$

3.3. Generación de imágenes a partir de información acústica

En la sección anterior se expusieron todos los elementos necesarios para definir una cámara acústica. Para cada *array* de micrófonos en el espacio 3D, se selecciona un punto C que será usado como el centro de la cámara, en este caso el centroide del *array* de micrófonos (ver Figura

3.5). A continuación se generan localizaciones que definen rayos que parten desde el centroide del *array* de micrófonos, haciendo un barrido en *azimuth* y elevación, de todo el espacio de búsqueda de la habitación donde se realiza la grabación. Para cada una de las localizaciones, se suma la potencia acústica medida por cada combinación de pares de micrófonos posible y por último, se suman las potencias acústicas calculadas en todas las localizaciones que forman un rayo o dirección determinada. Mediante esta técnica se obtiene una imagen en la que cada píxel representa la potencia acústica medida en esa dirección (*azimuth* y elevación). El ancho de la imagen representa el barrido en azimuth mientras el alto representa el barrido en elevación, ver Figura 3.5.

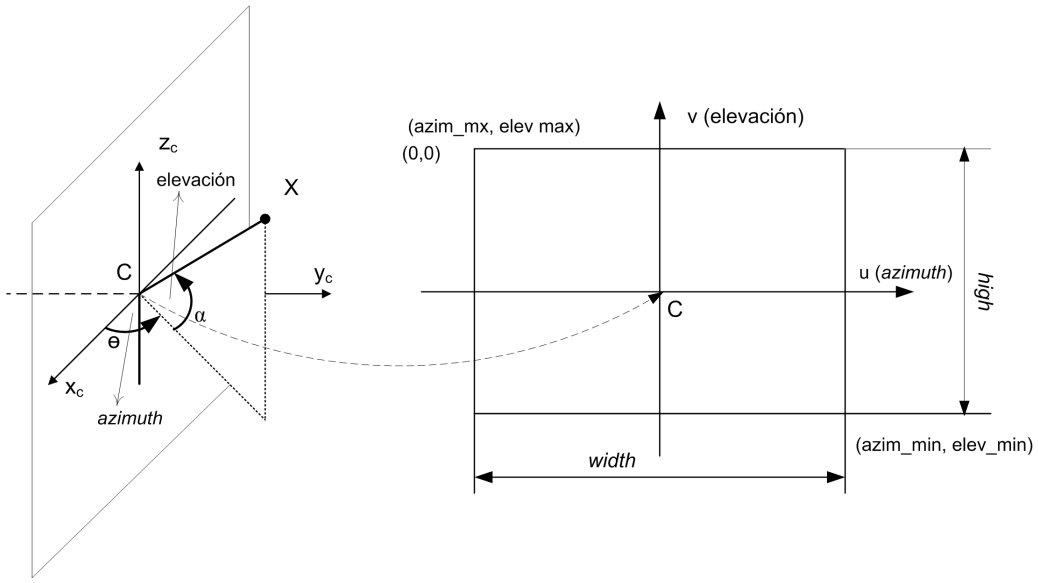


Figura 3.5: Formación de la imagen con información acústica.

3.3.1. Técnicas de Localización de fuentes sonoras a partir de arrays de micrófonos

Debido a las ventajas que presenta sobre las otras técnicas revisadas en la literatura se utilizó la variante de SRP-PHAT para calcular la potencia acústica en cada una de las direcciones. SRP permite direccionar el patrón de recepción del *array* de micrófonos hacia una posición determinada, esto permite calcular la potencia acústica en cada una de las direcciones del espacio de búsqueda. Para ello se utiliza la salida de un *delay-and-sum beamformer*, cuyo diagrama en bloque se muestra en la Figura 3.6. En este esquema, los retardos calculados en función del espacio al que se desea apuntar, son aplicados a la señal de entrada de cada uno de los micrófonos, produciendo la alineación temporal de las mismas. Dichas señales son sumadas para obtener la salida del algoritmo. Como paso previo al alineamiento, se puede filtrar la señal de entrada a los micrófonos. A este método se le conoce como *filter-and-sum beamforming* [8], ver Figura 3.7. En el dominio de la frecuencia la señal a la salida se calcula como:

$$Y(\omega, q) = \sum_{n=1}^{M} W_n(\omega) X_n(\omega) e^{j\omega\Delta_n} \tag{3.16}$$

donde $Y(\omega, q)$ es la señal en el dominio de la frecuencia que se obtiene a la salida del filtro, Δ_n es

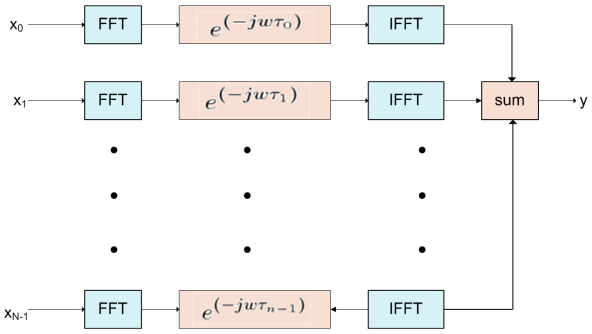


Figura 3.6: Diagrama en bloques de un *delay-and-sum beamformer* en el dominio de la frecuencia

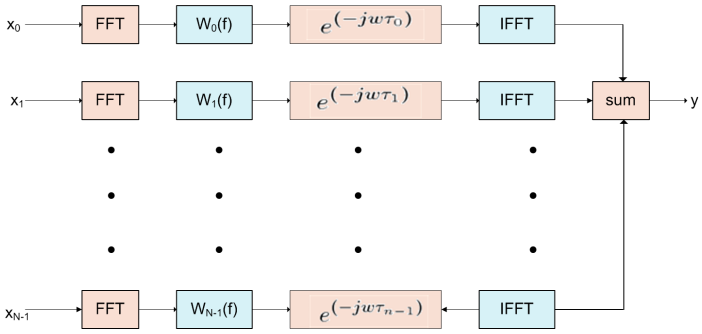


Figura 3.7: Diagrama en bloques de un *filter-and-sum beamformer* en el dominio de la frecuencia

el retardo aplicado al micrófono n, para dirigir el patrón de recepción del *array* a la localización espacial q, $X_n(\omega)$ es la Transformada de Fourier de la señal recibida por el micrófono n y $W_n(\omega)$ es la Transformada de Fourier del filtro aplicado W.

La potencia acústica medida en una posición q se puede expresar como la potencia de salida del filtro *filter-and-sum beamformer* mediante la ecuación 3.17.

$$P(q) = \int_{-\infty}^{\infty} |Y(\omega, q)|^2 d\omega = \int_{-\infty}^{\infty} Y(\omega, q) Y'(\omega, q) d\omega \tag{3.17}$$

Como la potencia calculada en la localización q, puede estar afectada por el ruido y la reverberación, se aplica la función de pesado PHAT que pone al mismo nivel todas las componentes espectrales de la señal. Esta estrategia es la que presenta mejores resultados en los entornos reales, aunque tiene como principal desventaja que el pesado se realiza con el inverso de su módulo, por lo que los errores se acentúan, cuando la potencia de la señal es pequeña [8], ver 3.18.

$$\Phi_{ij}(\omega) = \frac{1}{|X_i(\omega) X'_j(\omega)|} \tag{3.18}$$

La potencia en la localización q obtenida a la salida del algoritmo se calcula sustituyendo la ecuación 3.16 en 3.17 y aplicando la función de pesado PATH $\Phi_{ij}(\omega)$:

$$P(q) = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \int_{-\infty}^{\infty} \Phi_{ij}(\omega) X_i \omega X'_j(\omega) e^{j\omega(\Delta_j - \Delta_i)} \tag{3.19}$$

donde:

- $\Phi_{ij}(\omega) = W_i(\omega)W'_j(\omega) = \frac{1}{|X_i(\omega)X'_j(\omega)|} \Leftrightarrow W_n(\omega) = \frac{1}{|X_n(\omega)|}$ son los filtros PHAT.
- $\tau_{ij} = \Delta_j - \Delta_i$ es la TDOA entre el micrófono i y el j para la señal de audio en la localización q.

En [8] se demuestra que el SRP correspondiente a un *array* M de micrófonos es equivalente a la suma de las correlaciones cruzadas generalizadas (GCC) de todas las combinaciones de pares de micrófonos posibles $\left(\binom{M}{2} = \frac{M!}{2!(M-2)!} = \frac{M(M-1)}{2}\right)$ expresándose como:

$$P(q) = P(\Delta_1 ... \Delta_M) = 2\pi \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} c_{ij}(\Delta_j - \Delta_i) = 2\pi \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} c_{ij}(\tau_{ij}) \tag{3.20}$$

donde $\Delta_1 ... \Delta_M$ son los retardos en caso de dirigir el *array* a la localización q y $c_{ij}(\tau_{ij})$ es el GCC-PHAT de las señales de los micrófonos i y j evaluadas en τ_{ij}. La correlación cruzada de un par de micrófonos se calcula como:

$$c^{(g)}_{x_i x_j}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{x_i xj}(\omega) X_i(\omega) X'_j(\omega) e^{-j\omega\tau} d\omega \tag{3.21}$$

Las señales que se procesan son digitales obtenidas después de un proceso de muestreo y para analizarlas es necesario hacer un enventanado temporal que limite la cantidad de muestras en cada análisis, por lo que es necesario dividirlas en una serie de bloques antes de aplicarles la Transformada Discreta de Fourier (DFT). Con el objetivo de mejorar la representación espectral de la señal y eliminar los efectos causados al final de cada bloque por la DFT es necesario multiplicar cada bloque por una ventana, en este caso se utiliza la ventana de Hamming [8], ver ecuación 3.22. En la Figura 3.8 se muestra las respuestas en tiempo y frecuencia de la ventana de Hamming. Además los bloques de datos consecutivos se solapan en el tiempo para permitir que los datos del final estén centrados en el siguiente, dando así el mismo peso a todos ellos.

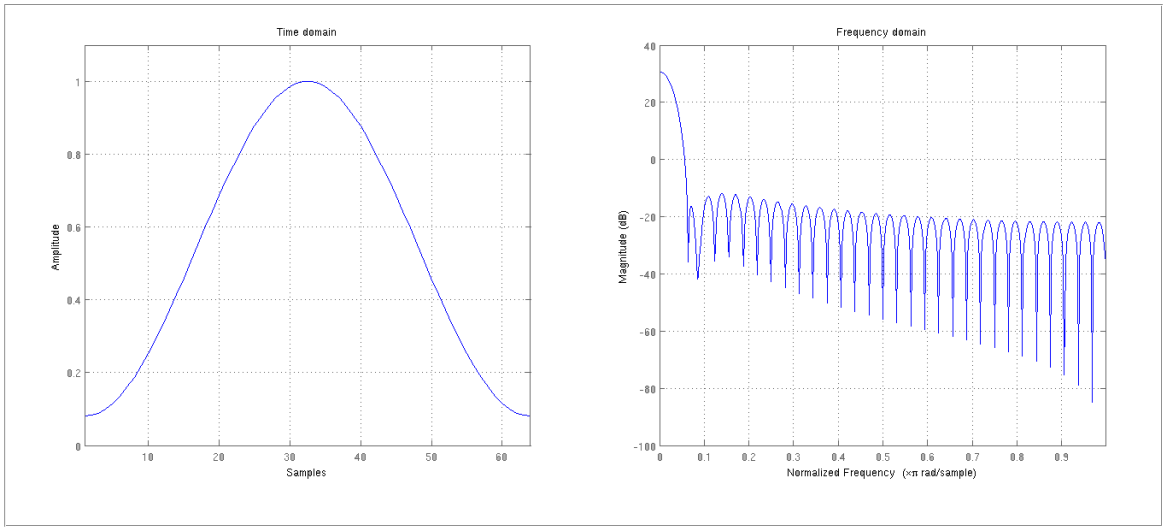


Figura 3.8: Ventana Hamming de 64 puntos en el dominio del tiempo y frecuencia

$$W(n) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{n}{N}\right), \quad 0 \leq n \geq N \tag{3.22}$$

El algoritmo de localización obtiene una estimación para la DFT de cada bloque de datos, asumiendo que el locutor no se moverá durante la duración en tiempo de ese bloque, llamado *frame size* y la tasa a la que se proporcionan las estimaciones se llama *frame-shift*.

La expresión para las señales discretas de los micrófonos $x_1[n]...x_M[n]$ y sus DFTs cuando se segmentan en bloques de longitud N es:

$$x_{m,b}[n] = \omega[n] \cdot x[bA + n] \text{ para n=0...N-1} \tag{3.23}$$

donde $x_{m,b}[n]$ son los datos dentro de la ventana del micrófono m y el bloque b. A es el $frameshift$. Cuando $A < N$ los bloques se solapan, $\omega[n]$ es la función de enventanado de Hamming [8].

La DFT del bloque b se puede expresar de la siguiente manera:

$$X_{m,b}[k] = \sum_{n=0}^{N-1} x_{m.b}[n]e^{-jk\frac{2\pi}{K}n} \text{ para k=0...K-1} \tag{3.24}$$

donde K es la longitud de la DFT y para hacer un cálculo eficiente de esta su longitud k debe de ser potencia de 2 y mayor que longitud N de cada bloque. Por lo tanto es necesario hacer un zero-padding al bloque para hacer $k = N$

La expresión de la función GCC-PHAT basada en la DFT entre los micrófonos i y j del bloque de datos b, $\hat{c}_{ij,b}$ se define sustituyendo en la ecuación 3.21 la transformada de Fourier de los bloques DFTs definidos anteriormente:

$$\widehat{c}_{ij,b}(\widehat{\tau}) = \frac{1}{K}\sum_{k=0}^{K-1} \Phi_{ij}[k]X_{i,b}[k]X'_{j,b}[k]e^{jk\frac{2\pi}{K}\widehat{\tau}} = \frac{1}{K}\sum_{k=0}^{K-1} \Phi_{ij}[k]C_{ij,b}[k]e^{jk\frac{2\pi}{K}\widehat{\tau}} \tag{3.25}$$

donde $\Phi_{ij}[k]$ es la versión discreta de la función de pesos $\Phi_{ij}[\omega]$ y $\omega_k = \frac{2\pi k}{K}$ es el índice de frecuencia de la DFT.

Empleando teoremas propios de la DFT [8], la función GCC-PHAT a partir de la FFT de las señales capturadas por un par de micrófonos se calcula como:

$$\widehat{c}_{ij,b}(\widehat{\tau}) = Re[IFFT(\Phi_{ij}[k]X_{i,b}[k]X'_{j,b}[k])](\widehat{\tau}) = Re\left[IFFT\frac{X_{i,b}[k]X'_{j,b}[k]}{|X_{i,b}[k]||X_{j,b}[k]|}\right](\widehat{\tau}) \tag{3.26}$$

Como SRP puede ser calculado como la suma de las correlaciones cruzadas (GCC) de todos los pares de micrófonos posibles; en el dominio de la frecuencia se puede plantear que el SRP en la posición q se calcula como:

$$\widehat{P}_q\left(\widehat{\Delta}_1 \ldots \widehat{\Delta}_M\right) = 2\pi\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} \widehat{c}_{ij,b}\left(\widehat{\tau}_{ij}\right) = \sum_{i=1}^{M-1}\sum_{j=i+1}^{M} Re\left[IFFT\frac{X_{i,b}[k]X'_{j,b}[k]}{|X_{i,b}[k]||X_{j,b}[k]|}\right](\widehat{\tau}) \tag{3.27}$$

Teniendo las localizaciones de todos los micrófonos del *array*, y las posiciones q que barren el espacio de búsqueda mediante la ecuación 3.27, se calcula la potencia acústica en cada una de las posiciones. En el sección 3.3.2 se explican los detalles de cómo se generan las imágenes a partir de la potencia acústica.

3.3.2. Estimación de mapas de energía acústica

En este apartado se explica en detalle el proceso de generación de mapas de energía acústica, empezando por el proceso de generación de los puntos de localización, que conforman los rayos del espacio de búsqueda, así como la utilización de la técnica SRP-PHAT para el cálculo de los mapas de energía acústica.

3.3.2.1. Generación de los rayos y el espacio de puntos de cálculo de potencia acústica

La generación de los rayos y el espacio de puntos es un paso imprescindible para poder calcular potencia acústica mediante la técnica de SRP-PHAT, debido a que, se necesita saber cada una de las localizaciones para poder redireccionar el patrón de recepción de los *arrays* de micrófonos. Como lo que se desea, es generar imágenes a partir de la potencia acústica, basándose en el modelo de cámara de perspectiva, las localizaciones en un paso inicial se generan en coordenadas polares a partir del centro del *array* de micrófonos, de forma tal que se cubra todo el espacio de búsqueda, haciendo un escaneo en *azimuth* y elevación, como se muestra en la Figura 4.18 de la página 77.

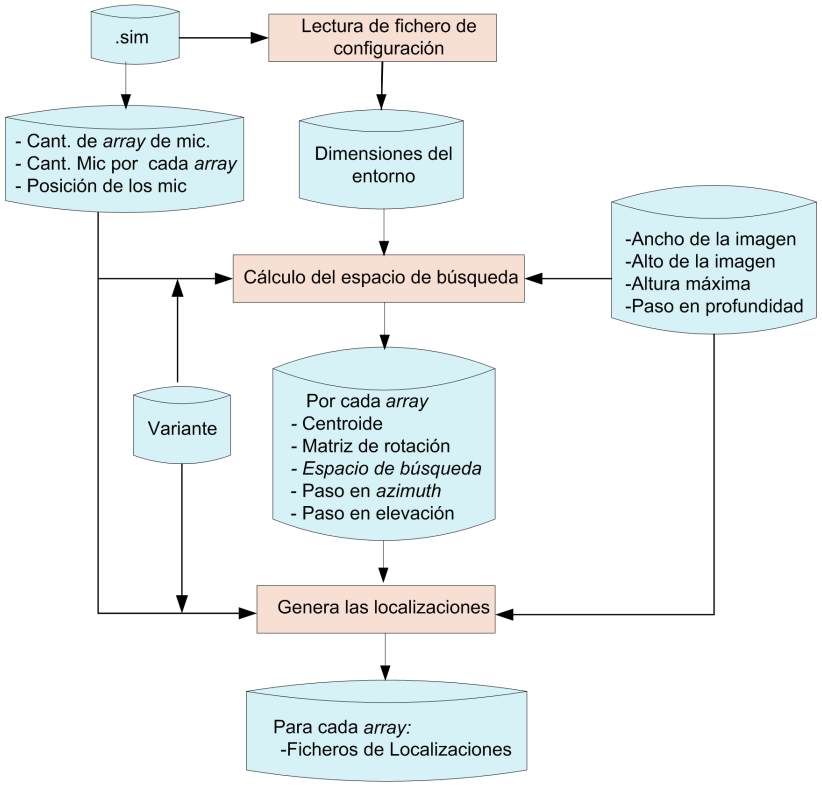


Figura 3.9: Diagrama de flujo del proceso de generación de las localizaciones

Para ello se creó la librería, "`generateSearchspace`" cuyo algoritmo se representa en el diagrama de flujo de la Figura 3.9. A continuación se explicarán los pasos realizados para la generación de las localizaciones:

1. En el primer paso se lee el fichero de simulación, que lleva el mismo nombre de la base de datos donde se tiene la grabación de la señal de audio proveniente del locutor. Para ello se utiliza la librería "`simulationconfiglib`" desarrollada en GEINTRA de la cual se obtiene los siguientes datos:

   - Para cada *array* la cantidad de micrófonos que poseen y la ubicación 3D, respecto al origen de coordenadas global de la habitación de cada uno de ellos.
   - Se obtienen todas las dimensiones de la habitación donde se realizó la grabación.

2. Una vez obtenida toda la información necesaria acerca de los micrófonos y de las dimensiones del entorno, se calcula para cada *array* de micrófonos los siguientes parámetros:

   - El centroide, que representará el centro de la cámara de perspectiva y a partir del cual se generarán todas las localizaciones. En la ecuación 3.28, n representa la cantidad de micrófonos que contiene el *array*, y $[x_i, y_i, z_i]$ las coordenadas 3D respecto al sistema de referencia global.
   $$Cent = \frac{\sum_{i=1}^{n} [x_i, y_i, z_i]}{n} \tag{3.28}$$
   - Se asume que el plano del *array* de micrófonos es perpendicular al suelo y por tanto la matriz de rotación Rot, que define la orientación del *array* respecto al sistema de coordenadas globales, se calcula como:
   $$Rot = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.29}$$
   donde α es el ángulo de rotación tomando como positivo el sentido anti horario.
   - La matriz de traslación T, define la traslación de la posición del *array* de micrófonos y que se calcula como:
   $$T = Cent \tag{3.30}$$
   - El espacio de búsqueda está definido por el rango en *azimuth*, elevación y distancia de las localizaciones que forman cada rayo. Con el objetivo de limitar el espacio de búsqueda, se fijó una distancia muerta ($DEAD\_DIST$) de 25 cm de separación respecto a las paredes, en la que se consideró que no iba a ver ningún locutor. Dicha distancia puede ser variable. A continuación se explica en detalle cómo se definió el espacio de búsqueda:
     - Distancia máxima por rayo: Define la distancia máxima que se va a recorrer en cada dirección y para la cual se implementaron dos variantes:
       a) La distancia máxima d_{max} de todos los rayos es la misma e igual a la cuerda más grande que pueda existir en el entorno. Por ejemplo, para una habitación que forme un prisma de base rectangular:
       $$d_{max} = \sqrt{x_{max}^2 + y_{max}^2 + z_{max}^2} \tag{3.31}$$
       b) Para cada rayo se determina el punto de intersección con cualquiera de las superficies que limita la habitación, para la cual se utiliza la librería "`geometrylib.h`" desarrollada en GEINTRA.

- Determinación del rango en que varía el *azimuth*. Define el barrido horizontal del espacio de búsqueda. La Figura 3.10 ayuda a entender las ecuaciones 3.32 y 3.33 que calcula el azimuth máximo, $Azim_{max}$, y el mínimo, $Azim_{min}$.

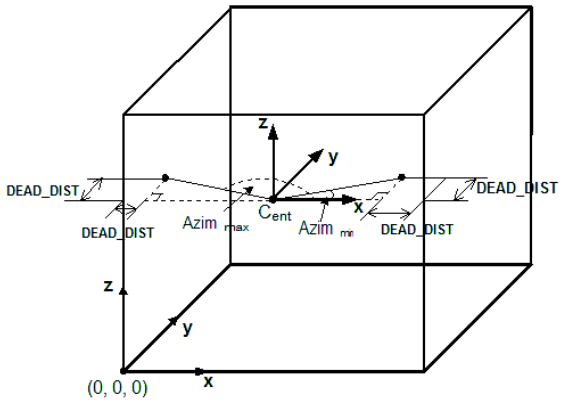


Figura 3.10: Cálculo del rango de exploración en *azimuth*

$$Azim_{max} = \Pi - \arctan\left(\frac{DEAD\_DIST}{(Cent - DEAD\_DIST)}\right) \tag{3.32}$$

$$Azim_{min} = \arctan\left(\frac{DEAD\_DIST}{(Cent - DEAD\_DIST)}\right) \tag{3.33}$$

- Determinación del barrido vertical. Calcula la elevación máxima $Elev_{max}$, definida por la ecuación 3.34, y la mínima, $Elev_{min}$ por 3.35, a la cual se van a generar el barrido vertical, ver Figura 3.11.

$$Elev_{max} = \left\{ \begin{array}{ll} \arctan\left(\frac{(H\_MAX - Cent_z)}{(y_{max} - DEAD\_DIST)}\right), & if \quad Cent_z > H\_MAX \\ \arctan\left(\frac{(H\_MAX - Cent_z)}{DEAD\_DIST}\right), & if \quad Cent_z < H\_MAX \\ 0, & if \quad Cent_z = H\_MAX \end{array} \right\} \tag{3.34}$$

$$Elev_{min} = \frac{(H\_MIN - Cent_z)}{DEAD\_DIST} \tag{3.35}$$

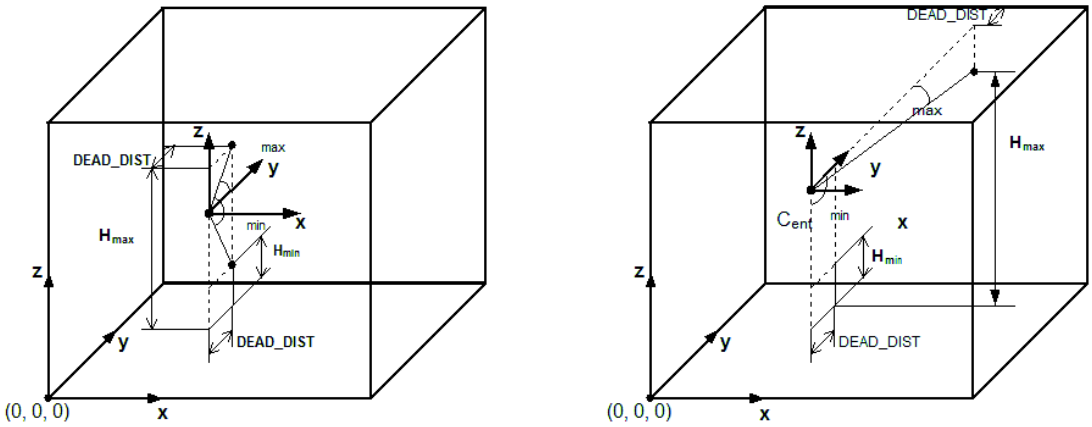
- El paso en ángulo en que varían el *azimuth* y elevación se muestra en las ecuaciones 3.36 y 3.37:

$$stepAzimuth = \frac{(Azim_{max} - Azim_{min})}{IMAGE\_WIDE} \tag{3.36}$$

$$stepElev = \frac{(Elev_{max} - Elev_{min})}{IMAGE\_HIGH} \tag{3.37}$$

3. En los pasos anteriores se obtuvieron todos los datos necesarios para la generación de las localizaciones de todos los *arrays* de micrófonos. En la Figura 3.12 se muestra el diagrama de flujo de la generación de las localizaciones, que se explicará a continuación.

Para cada *array* de micrófonos se genera un conjunto de localizaciones que barren todo el espacio de interés del entorno. En el paso (5) las localizaciones que inicialmente se tienen en coordenadas polares y en el sistema de coordenadas global se transforman a coordenadas cartesianas utilizando la función "`convertTPolarPointToTPoint`" de la librería "`geometrylib`"



(a) Centroide del *array* de micrófonos está debajo de la altura máxima de exploración

(b) Centroide del *array* de micrófonos por encima de la altura máxima de exploración

Figura 3.11: Cálculo del rango de exploración en elevación.

desarrollada en GEINTRA y luego son transformadas al sistema de coordenadas de cada *array* de micrófonos como se muestra en la ecuación 3.38 :

$$X = Rot * [x, y, z]^T + T \tag{3.38}$$

donde Rot es la matriz de rotación definida en la ecuación 3.49 y T la matriz de traslación definida en la ecuación 3.30. Una vez obtenida la localización se comprueba la variante utilizada para generar las localizaciones las cuales son:

- Variante 1. En cada rayo se generan la misma cantidad de localizaciones, barriendo desde la distancia mínima hasta la máxima respecto al centroide del *array* de micrófonos, sin importar que la localización quede fuera de los límites de la habitación, ver Figura 3.13(a).
- Variante 2. Las localizaciones que queden fuera de los límites de la habitación se desechan. Para ello se utilizó la función "_checkPathOk" de la librería "geometrylib.h" desarrollada en GEINTRA, que permite calcular la intersección de un rayo con un plano, ver Figura 3.13(b).

Para cada *array* se genera un fichero de localizaciones en el cual se almacena entre otros:

- Las localizaciones en el espacio 3D en el mismo orden en que se generan.
- La cantidad de localizaciones por rayo y la total.
- El espacio barrido en *azimuth* y elevación.

3.3.2.2. Cálculo de potencia acústica

Una vez generadas las localizaciones relativas a cada *array* de micrófonos y definida la técnica de localización utilizada en la sección 3.3.1 se pasaría al cálculo de la potencia acústica que se mide en el *array* de micrófonos en cada uno de los rayos que forman el espacio de búsqueda para cada *array* de micrófonos.

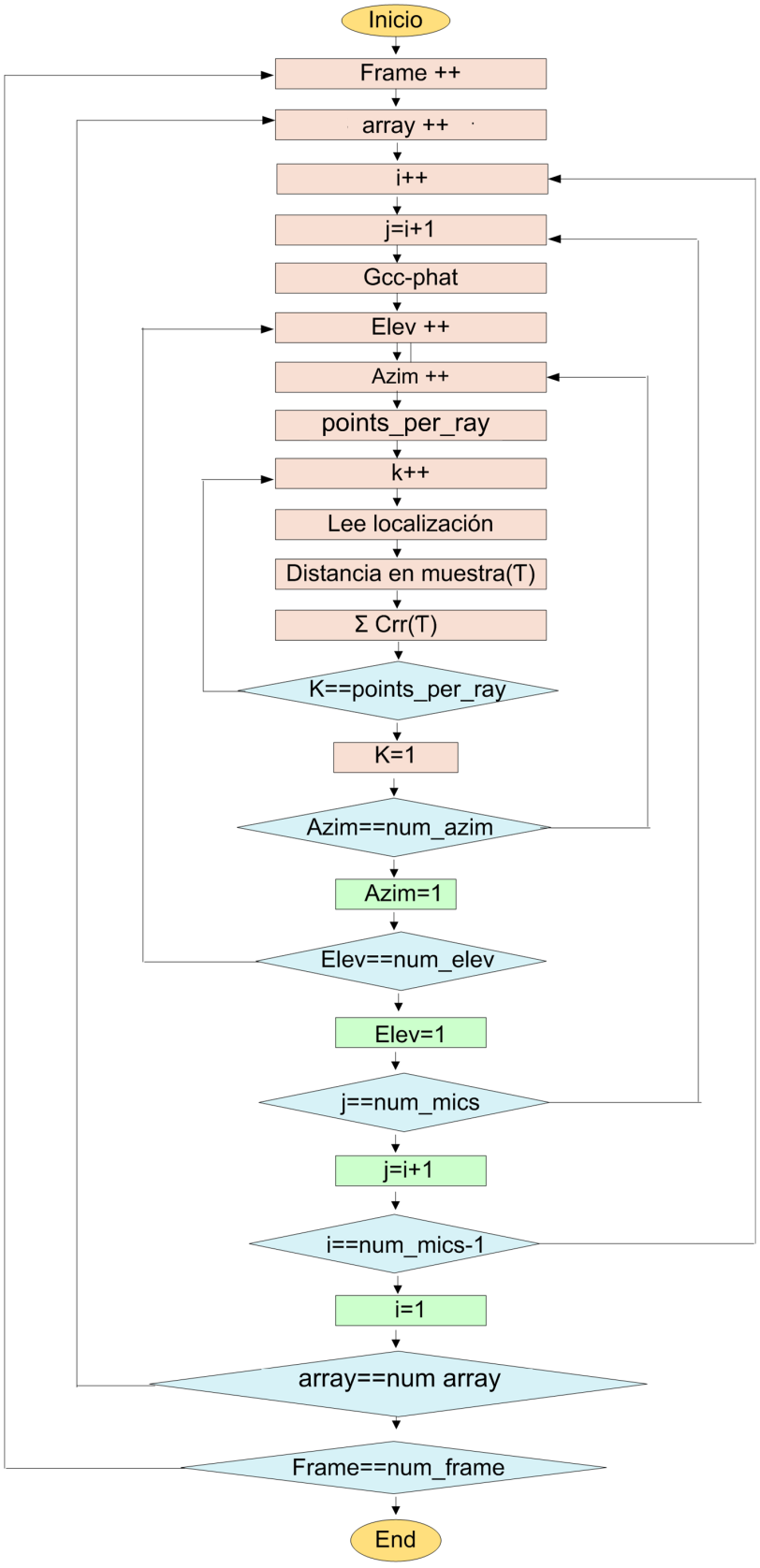
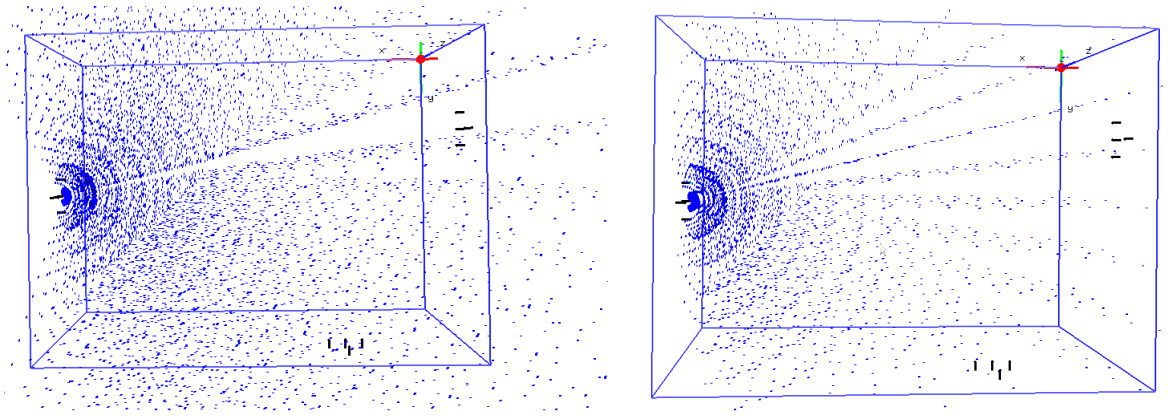


Figura 3.12: Algoritmo implementado en la generación del espacio de espacio de puntos para el cálculo de la potencia acústica



(a) Variante 1. Localizaciones generadas sin importar que estén fuera de los límites de la habitación

(b) Variante 2. Las localizaciones que están fuera de los límites de la habitación son desechadas

Figura 3.13: Representación gráfica de la habitación con las localizaciones generadas respecto a un *array* de micrófonos.

Como se comentó en la sección 3.3.1 la señal de audio recibida se divide en bloques de tiempos *(frames)* que limitan la cantidad de muestras que serán procesadas en cada estimación, tiempo en el cual se considera que el locutor no se mueve.

Para cada *frame*, para cada *array* de micrófonos y para todas las combinaciones de los pares de micrófonos que conforman el *array* se calcula:

- La correlación cruzada generalizada (GCC-PHAT), mediante la ecuación 3.26.
- Se hace un barrido en *azimuth* y elevación, obteniéndose las localizaciones correspondientes a cada rayo donde:
  - Se calcula la diferencia de distancia en muestra que existe entre cada localización y el par de micrófonos mediante la ecuación:

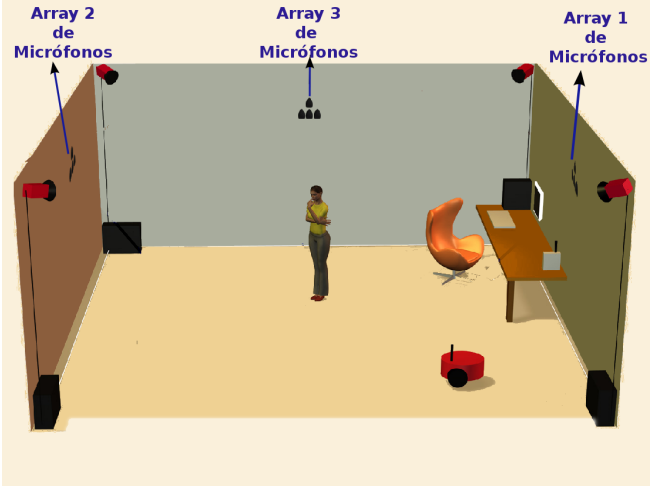
$$\tau(i,j,l) = \frac{dist(i,l) - dist(j,l)}{C_{son}} f_s \tag{3.39}$$

    siendo $dist(i,l) = \sqrt{(x_i - x_l)^2 + (y_i - y_l)^2 + (z_i - z_l)^2}$ la distancia entre el micrófono i y la localización j, se considera la velocidad del sonido C_{son} constante a 20º, pero puede adaptarse y f_s la frecuencia de muestreo.
  - Para cada *array* se obtiene la potencia medida en la localización l, sumando las correlaciones cruzadas $\widehat{c}_{ij}(\tau)$, correspondientes a todas las combinaciones posibles entre pares de micrófonos (*comb_par_minc*) desplazadas en τ, como:

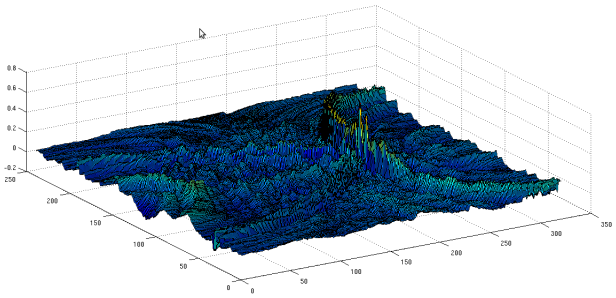
$$P_l = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{c}_{ij}(\tau) \tag{3.40}$$

    donde n es la cantidad de micrófonos que hay en el *array*.
- Como la cantidad de localizaciones que se generan en cada rayo no siempre es la misma (en la variante 2 expuesta en el apartado 3.3.2.1), se realiza un promediado de la potencia acústica obtenida para cada rayo quedando la potencia medida en un *array* como:

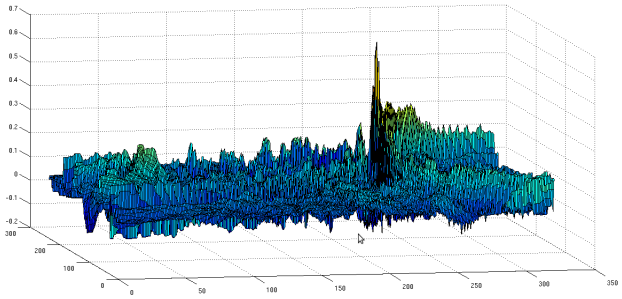
$$Pray_{array} = \frac{\sum_{k=1}^{k=points\_per\_ray} P_l}{points\_per\_ray} \tag{3.41}$$



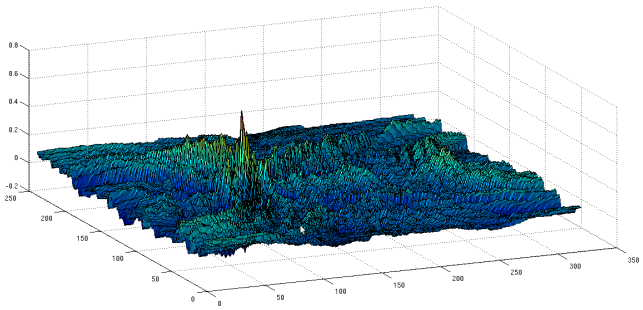
(a) Espacio Inteligente



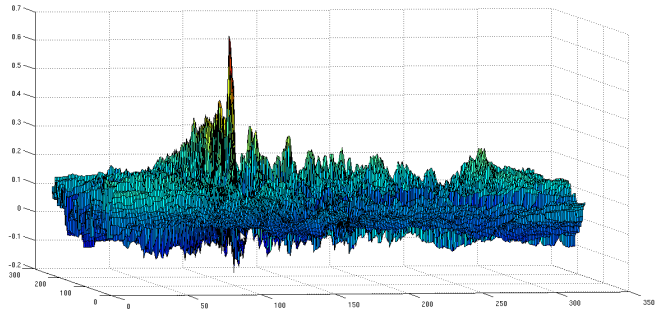
(b) Mapa de potencia que se obtiene desde el *Array 1*



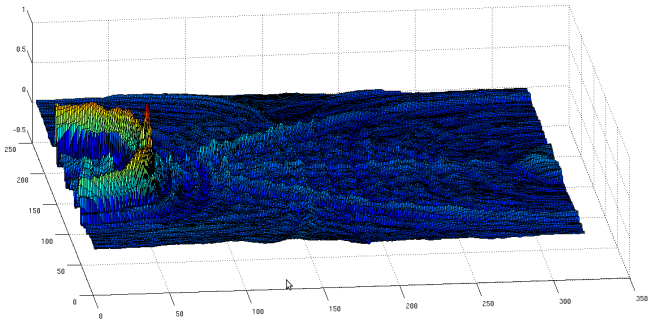
(c) Nivel del mapa de potencia



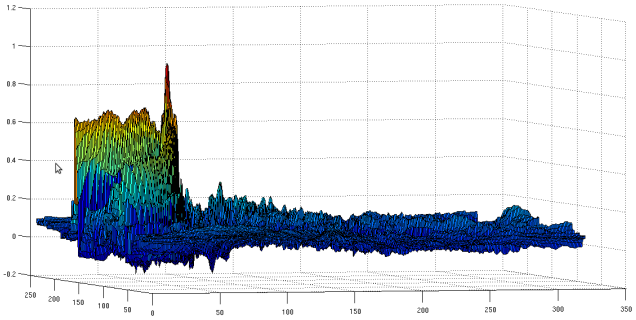
(d) Mapa de potencia que se obtiene desde el *Array 2*



(e) Nivel del mapa de potencia

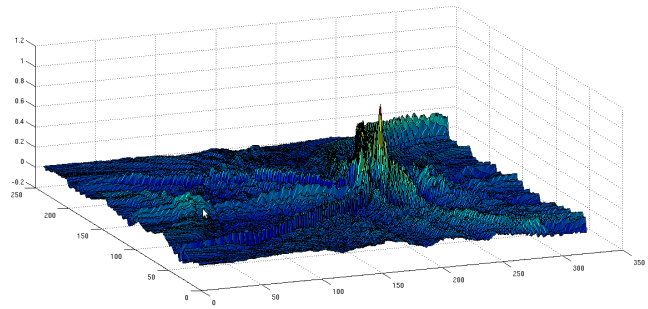


(f) Mapa de potencia que se obtiene desde el *Array 3*

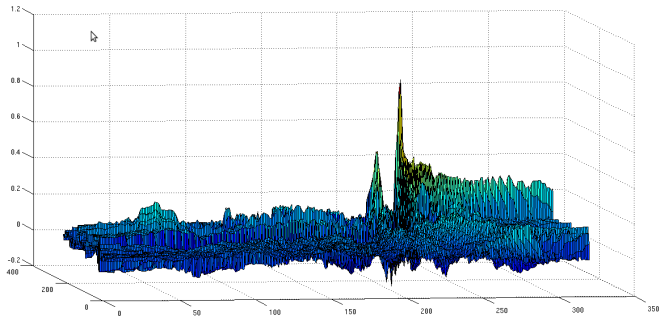


(g) Nivel del mapa de potencia

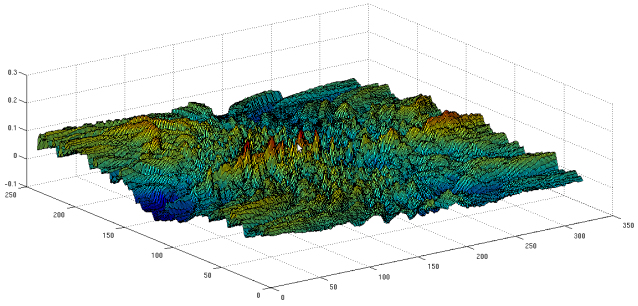
Figura 3.14: Mapas de potencia obtenidos en un *frame* con un locutor activo



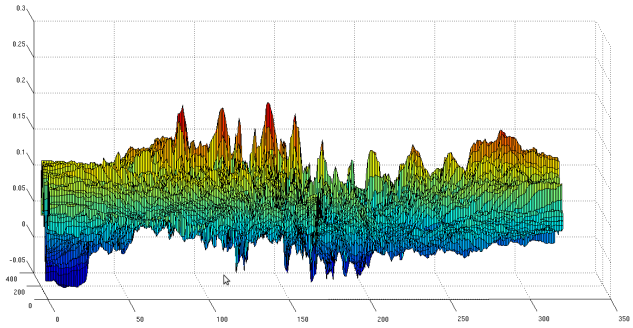
(a) Mapa de potencia que se obtiene desde el *Array 1*



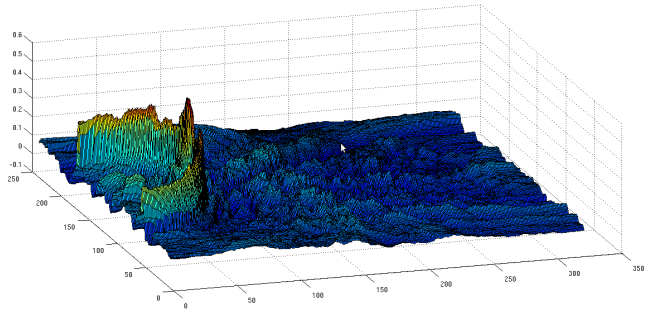
(b) Nivel del mapa de potencia



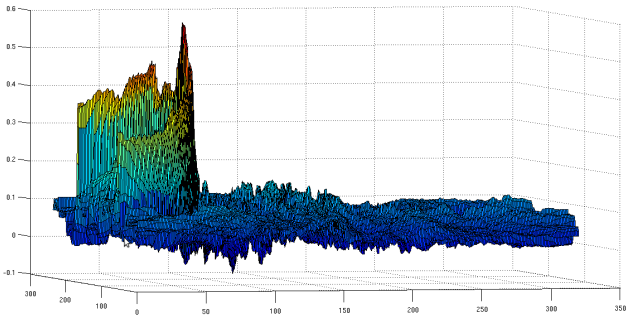
(c) Mapa de potencia que se obtiene desde el *Array 2*



(d) Nivel del mapa de potencia

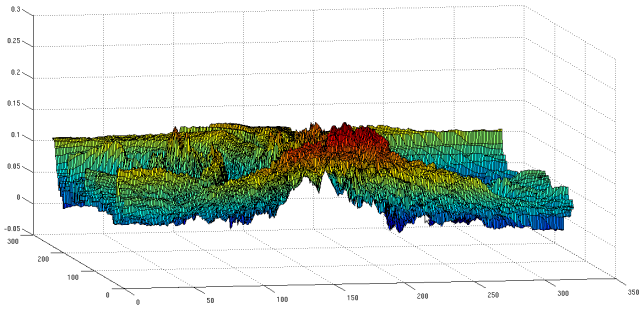


(e) Mapa de potencia que se obtiene desde el *Array 3*

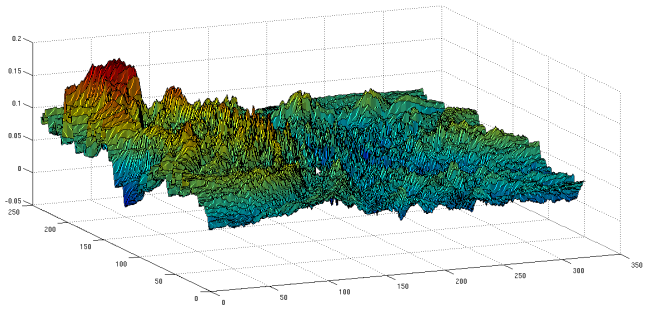


(f) Nivel del mapa de potencia

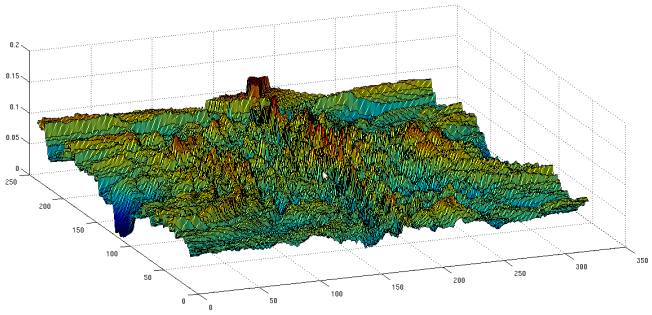
Figura 3.15: Mapas de potencia obtenidos en un *frame* con un locutor activo en el que ocurre un fallo en el mapa de energía obtenidos en el *Array* 2 de Micrófonos, (ver sub figuras c y d).



(a) Mapa de potencia que se obtiene desde el *Array 1*



(b) Mapa de potencia que se obtiene desde el *Array 2*



(c) Mapa de potencia que se obtiene desde el *Array 3*

Figura 3.16: Vista superior del mapa de potencia calculado en un *frame* donde no hay locutor activo

y la potencia medida en un rayo sería:

$$Pray = \sum_{k=1}^{k=num\_array} Pray_{array} \tag{3.42}$$

En la Figura 3.14(a) se representa un espacio inteligente donde están ubicados 3 *arrays* de micrófonos en el que se obtienen los mapas de potencias asociados a cada *array* en distintas circunstancia. En la Figura 3.14 se muestran los mapas de potencias generados para un *frame* con un locutor activo, el eje horizontal definen el barrido en azimuth desde cada *array* de micrófonos y el eje vertical el barrido en elevación. Como se puede ver en cada una de las imágenes existe un pico bastante bien definido, el cual representaría la dirección en la que se ha medido una mayor energía, en las otras direcciones también se pueden observar algunos picos generados por el ruido y las reverberaciones. En la Figura 3.15(d) se puede observar un *frame* en la que a pesar de haber un locutor hablando en el *arrays 2* no se obtiene una buena estimación encontrándose un fallo SRP, este tipo de error suele cometerse cuando la señal de audio llega con poca potencia al *array*, ya sea por la lejanía o porque el locutor estuviese hablando de espaldas al *array*. En la Figura 3.16 se representan los mapas obtenidos para un *frame* en el que no hay ningún locutor activo, viéndose cómo aparecen algunos niveles de energía, el cual es producido por el ruido de fondo que hay en la habitación.

3.3.3. Generación de imágenes

En esta sección se describe cómo se generan las imágenes de potencia acústica a partir de los mapas de potencias calculados en la sección anterior, para ello se utilizan algunas de las funciones

de la librería de tratamiento de imágenes OpenCV 2.0, en la que cada pixel equivale a un punto en el mapa de potencia. Las imágenes se generaron en escala de grises con una profundidad de 8 bits por píxeles. El ancho y el alto de la imagen representan el barrido en *azimuth* y en elevación respectivamente. Como los niveles de potencia son valores del tipo *float* y menores que 1 es necesario expandir el rango dinámico entre 0 y 255 para no perder información, por lo que se realiza un histograma con todos los mapas de potencias calculados en un experimento dado, para analizar la distribución de energía y poder seleccionar el rango dinámico con el que varían los píxeles de la imagen.

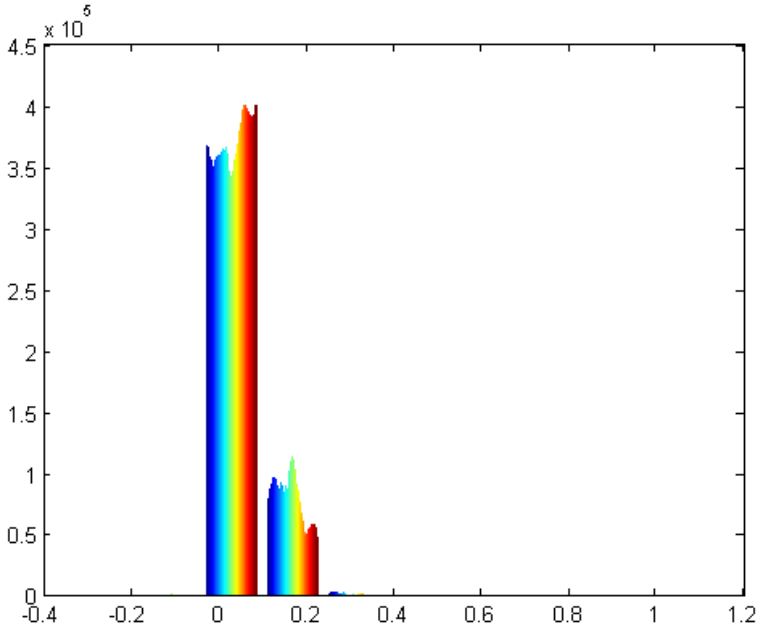


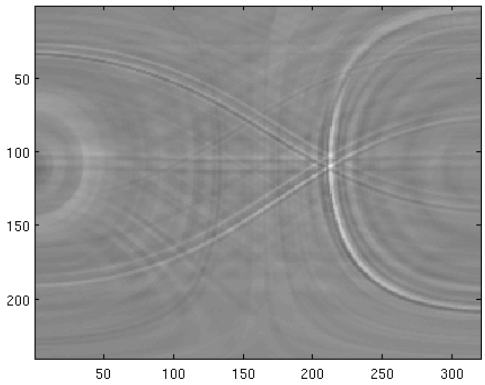
Figura 3.17: Histograma realizado con los valores de potencias obtenidos de los *frame* del experimento AIT de Chil

En el histograma que se muestra en la figura 3.17 correspondiente a un experimento de CHIL, se observan muy pocos puntos en los niveles de energía más alto, esto se debe a que solo existe un locutor y por tanto muy pocos píxeles tendrán valores de energía considerable. Para este caso en particular se escogió como valor mínimo, $min = -0{,}2$, y como máximo, $max = 0{,}6$. El valor de cada pixel, $Pixel$, se calcula como:

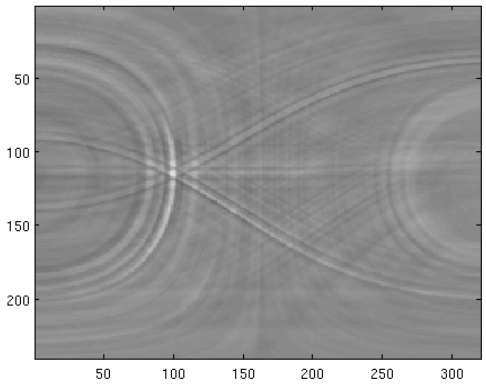
$$Pixel = \left\{ \begin{array}{ll} 0, & if \quad Pray < 0 \\ \frac{(Pray-min)*255}{(max-min)}, & if \quad Pray \geq 0 \end{array} \right\} \tag{3.43}$$

En la Figura 3.18 se muestran las imágenes obtenidas para un *frame*, con un locutor activo en el que se puede observar de manera bastante clara las direcciones donde se miden mayores niveles de potencia acústica. En la Figura 3.19(b) no se detecta con claridad la dirección de mayor energía a pesar de haber un locutor activo, lo cual representa un fallo en SRP. Este fallo puede ser causado por la poca energía acústica que se recibe, debido a que el locutor se encuentre lejos o hablando en dirección contraria. En la Figura 3.20 se representan las imágenes obtenidas para un *frame* donde no hay locutor activo.

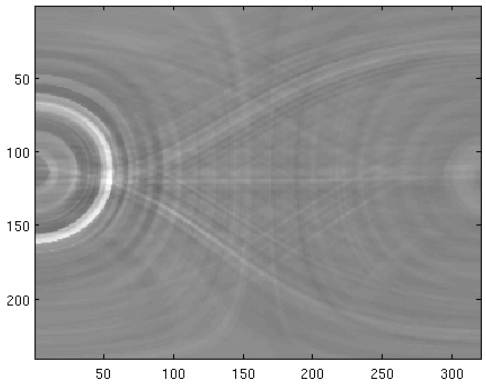
A pesar que en las imágenes se pueden observar de manera bastante clara las direcciones de máximos de potencia acústica, para buscar estos en un primer paso se utilizaron directamente



(a) El *array* de micrófono se encuentra en la pared derecha de la habitación

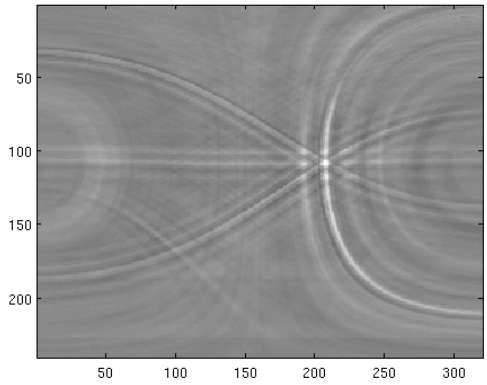


(b) El *array* de micrófono se encuentra en la pared izquierda de la habitación

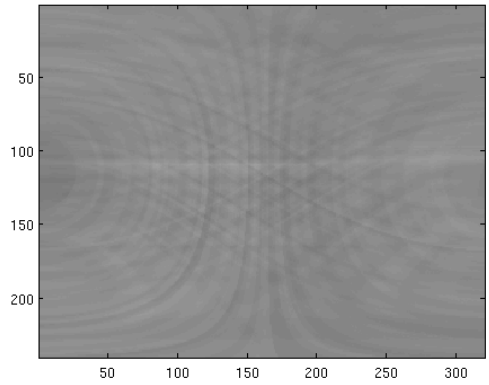


(c) El *array* de micrófono se encuentra en el fondo de la habitación

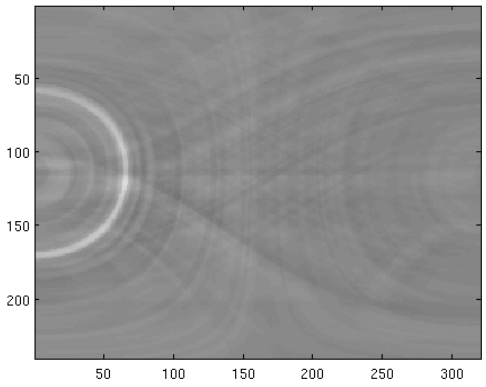
Figura 3.18: Imagen resultante de un *frame* con un locutor activo en tres *arrays* de micrófonos



(a) El *array* de micrófono se encuentra en la pared derecha de la habitación

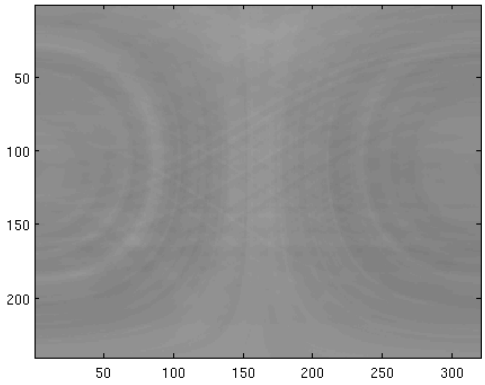


(b) El *array* de micrófono se encuentra en la pared izquierda de la habitación

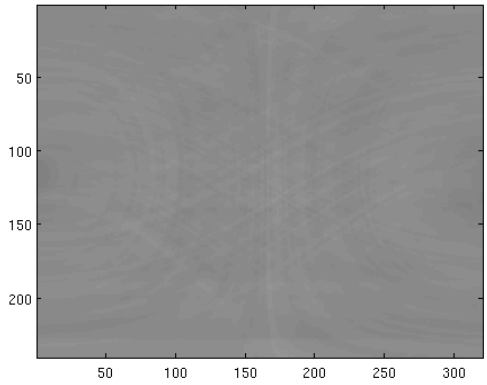


(c) El *array* de micrófono se encuentra en el fondo de la habitación

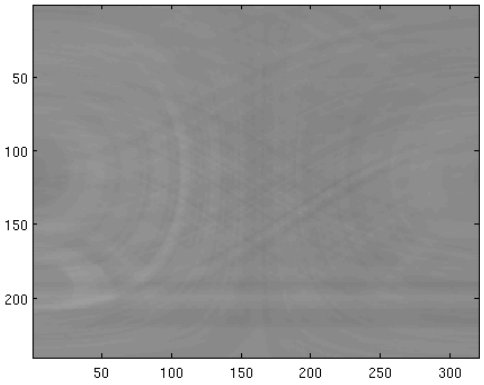
Figura 3.19: Imagen resultante de un *frame* con un locutor activo en tres *arrays* de micrófonos, donde en (b) se puede observar un fallo en SRP



(a) El *array* de micrófono se encuentra en la pared derecha de la habitación



(b) El *array* de micrófono se encuentra en la pared izquierda de la habitación



(c) El *array* de micrófono se encuentra en el fondo de la habitación

Figura 3.20: Imagen resultante de un *frame* en el que no hay locutor activo en ninguno de los tres *arrays* de micrófonos

los mapas 3D de potencia.

3.4. Técnicas de cálculo de máximos

Una vez generados los mapas y las imágenes de potencia acústica que se obtiene para cada *frame* por los *arrays* de micrófonos, se utiliza alguna de las técnicas existentes para buscar los máximos de energía que corresponden a una dirección o conjunto de las mismas donde la energía acústica es superior al resto de direcciones vecinas. En dichos máximos se espera encontrar la actividad acústica generada por los locutores existentes en el espacio donde se encuentran los micrófonos. Como se explicó en la sección 3.4, esta fase se divide en dos partes:

- Utilización del algoritmo *Nom Maximun Supression* para aislar máximos locales realizando una umbralización.
- Detección de cada máximo con precisión subpixélica mediante una función cuadrática.

A continuación se explica en detalles los pasos utilizados en la implementación del algoritmo *Nom Maximun Supression* con aproximación subpixélica.

1. Utilizando la función de matlab *ordfilt2*, cada pixel es sustituido por el valor máximo encontrado alrededor de un radio fijado. Por ejemplo suponiendo que el radio fijado es 1 la matriz:

$$cim = \begin{matrix} 92 & 99 & 1 & 8 & 15 & 67 & 74 & 51 & 58 & 40 \\ 98 & 80 & 7 & 14 & 16 & 73 & 55 & 57 & 64 & 41 \\ 4 & 81 & 88 & 20 & 22 & 54 & 56 & 63 & 70 & 47 \\ 85 & 87 & 19 & 21 & 3 & 60 & 62 & 69 & 71 & 28 \\ 86 & 93 & 25 & 2 & 9 & 61 & 68 & 75 & 52 & 34 \\ 17 & 24 & 76 & 83 & 90 & 42 & 49 & 26 & 33 & 65 \\ 23 & 5 & 82 & 89 & 91 & 48 & 30 & 32 & 39 & 66 \\ 79 & 6 & 13 & 95 & 97 & 29 & 31 & 38 & 45 & 72 \\ 10 & 12 & 94 & 96 & 78 & 35 & 37 & 44 & 46 & 53 \\ 11 & 18 & 100 & 77 & 84 & 36 & 43 & 50 & 27 & 59 \end{matrix}$$

   después de aplicarle *ordfilt2* queda:

$$mx = \begin{matrix} 99 & 99 & 99 & 16 & 73 & 74 & 74 & 74 & 64 & 64 \\ 99 & 99 & 99 & 88 & 73 & 74 & 74 & 74 & 70 & 70 \\ 98 & 98 & 88 & 88 & 73 & 73 & 73 & 71 & 71 & 71 \\ 93 & 93 & 93 & 88 & 61 & 68 & 75 & 75 & 75 & 71 \\ 93 & 93 & 93 & 90 & 90 & 90 & 75 & 75 & 75 & 71 \\ 93 & 93 & 93 & 91 & 91 & 91 & 75 & 75 & 75 & 66 \\ 79 & 82 & 95 & 97 & 97 & 97 & 49 & 49 & 72 & 72 \\ 79 & 94 & 96 & 97 & 97 & 97 & 48 & 46 & 72 & 72 \\ 79 & 100 & 100 & 100 & 97 & 97 & 50 & 50 & 72 & 72 \\ 18 & 100 & 100 & 100 & 96 & 84 & 50 & 50 & 59 & 59 \end{matrix}$$

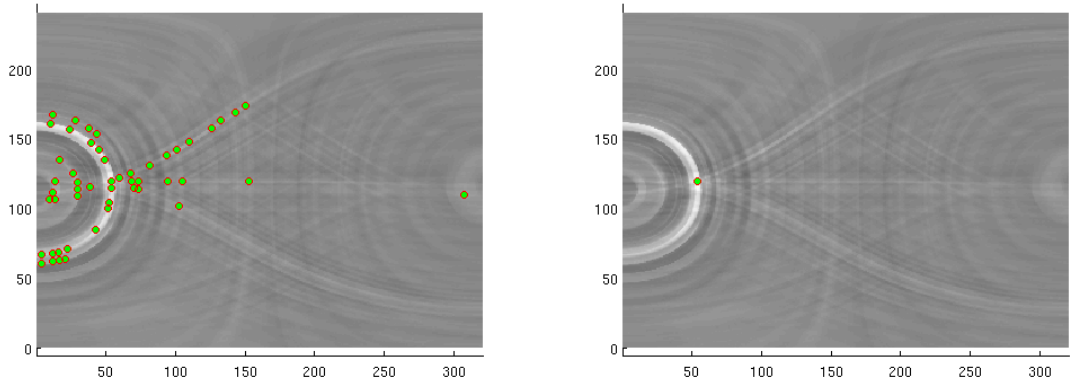
2. Se extraen las posiciones de todos los valores de mx que superen un umbral determinado eliminándose los que se encuentran a cada extremo una distancia menor o igual que el radio fijado. Luego se obtienen las posiciones de todos los máximos locales que superan el umbral.

3. Se realiza la aproximación subpixélica (sub_r, sub_c) de la posición de los máximos (r, c) mediante la ecuación 3.44:

$$\begin{aligned} sub\_r &= \frac{cim(r-1,c)-cim(r+1,c)}{2\cdot(cim(r-1,c)+cim(r+1,c)-2\cdot cim(r,c))} + r \\ sub\_c &= \frac{cim(r,c-1)-cim(r,c+1)}{2\cdot(cim(r,c-1)+cim(r,c+1)-2\cdot cim(r,c))} + c \end{aligned} \tag{3.44}$$

4. Para este caso, como solamente hay un locutor activo se escoge el máximo de mayor valor que constituirá el máximo global, ver Figura 3.21(b).

La Figura 3.21 muestra el resultado de la detección del máximo global para un mapa de energía obtenido desde un *array* de micrófonos.



(a) Máximos locales obtenidos en un mapa de energía (b) Máximo global obtenido en un mapa de energía

Figura 3.21: Máximos obtenidos en un mapa de energía por el algoritmo Nom Maximun Supression

La posición del máximo global obtenida, ver Figura 3.21(b), define la dirección de mayor potencia acústica medida en cada mapa de potencia, y es la utilizada en el algoritmo de triangulación DLT para estimar la posición 3D del locutor activo.

3.5. Técnicas de triangulación

En esta sección se describe como calcular la posición de un punto en el espacio 3D dado dos puntos en imágenes diferentes. Se asume que solamente hay errores en la estimación de las coordenadas de las imágenes y no en las matrices de proyección P, P'. Bajo estas circunstancia la triangulación clásica por reproyección de los rayos desde los puntos de la imagen fallará debido a que los rayos no se intersectarán y por tanto es necesario estimar la mejor solución para el punto 3D en el espacio. La mejor solución requerirá la definición y minimización de una función de coste. Se desea encontrar un método de triangulación que sea invariante a las transformaciones proyectivas del espacio. Partiendo de que se conoce la matriz fundamental de la cámara y que existe un error en la estimación de las coordenadas x y x' de las imágenes. Lo que significa que no existe un punto en el espacio X que satisfaga $x = PX$, $x' = P'X$; y que los puntos de las imágenes no satisfacen la restricción epipolar $x'^T Fx = 0$. Lo que es equivalente a decir que los dos rayos corresponden a un par de puntos $x \leftrightarrow x'$. ver la Figura 3.22 .

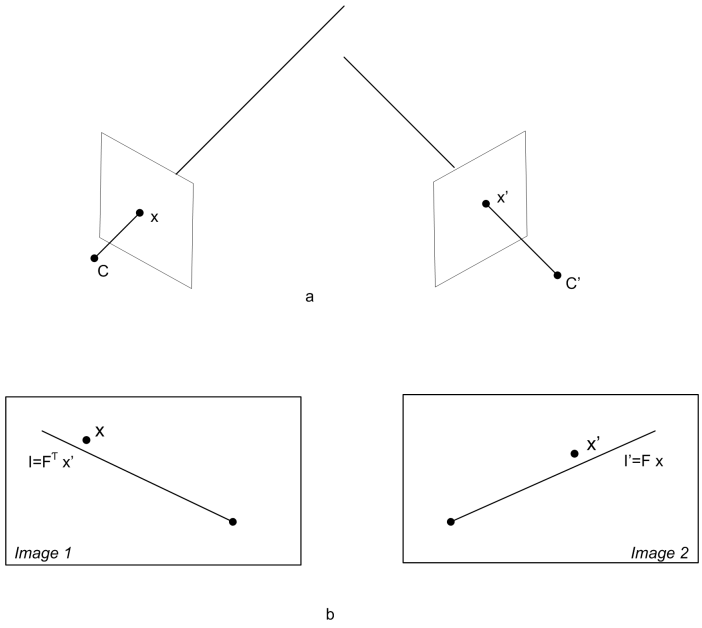


Figura 3.22: Reproyección de los rayos desde puntos de medida erróneos

Se denota por τ un método de triangulación usado para calcular un punto X en el espacio 3D desde la correspondencia $x \leftrightarrow x'$ y un par de matrices fundamentales de cámaras P y P' como:

$$X = \tau(x, x', P, P'). \tag{3.45}$$

La triangulación es invariante a la transformación H si:

$$\tau(x, x', P, P') = H^{-1}\tau(x, x', PH^{-1}, P'H^{-1}). \tag{3.46}$$

La reconstrucción proyectiva, es inapropiada para minimizar los errores en el espacio proyectivo 3D. Por ejemplo el método *midpoint* que busca el punto medio de la perpendicular entre los dos rayos en el espacio no es aplicable a la geometría proyectiva. Por lo que se necesita un método que sea invariante a la geometría proyectiva. La idea es estimar un punto $\hat{X}$ que satisfaga la ecuación:

$$\hat{x} = P\hat{X} \qquad \hat{x}' = P'\hat{X} \tag{3.47}$$

Partiendo de un punto X^{OW} en el espacio 3D respecto al sistema de coordenada global de la habitación, ver Figura 3.23, y conociendo las matrices de rotación y de traslación que relacionan el punto respecto al centro de las cámaras que forman las imágenes (origen de coordenadas local), por ejemplo para el sistema de referencia OA_1 en la Figura 3.23 se tiene que:

$$X^{OA_1} = R_1[X^{OW} - T^{OA_1}] \tag{3.48}$$

donde T^{OA_1} es un vector de 3 x 1 que representa la traslación en los ejes x, y, z del centro de la cámara OA_1 respecto al sistema de coordenadas global OW, que en este caso sería el centroide del *array* de micrófonos, R_1 representa la matriz de rotación 3 x 3 que relaciona la orientación de la cámara respecto al sistema global. Los *arrays* de micrófonos se encuentran ubicados en las paredes de la habitación por lo que todas las rotaciones se consideran respecto al eje Z del sistema de coordenadas global OW y se calcula como:

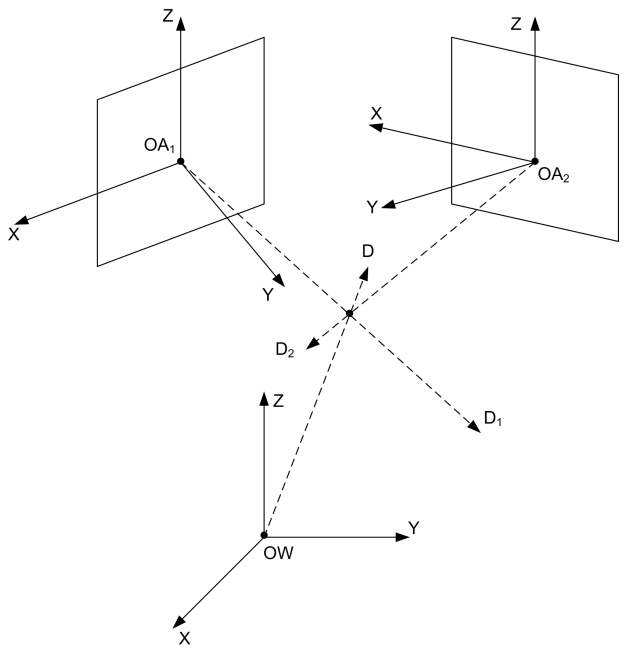


Figura 3.23: Triangulación lineal

$$R_1 = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.49}$$

Si se tiene el vector director D^{OA_1} que pasa por el vector X^{OA_1}. Entonces el producto cruzado de dos vectores colineales:

$$D^{OA_1} \times X^{OA_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.50}$$

Sustituyendo la ecuación 3.48 en 3.50:

$$D^{OA_1} \times (R_1 \cdot [X^{OW} - T^{OA_1}]) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.51}$$

Sustituyendo la ecuación 3.49 en 3.51 el vector X^{OW} por sus tres coordenadas $[x, y, z]'$ y el vector de traslación T^{OA_1} por $[T_{x_1}, T_{y_1}, T_{z_1}]'$se tiene:

$$D^{OA_1} \times \left(\begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \left([x, y, z]' - [T_{x_1}, T_{y_1}, T_{z_1}]' \right) \right) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.52}$$

Resolviendo la ecuación anterior se puede obtener la matriz A_1 de 3×3 ver (3.53) pero que tiene solo 2 filas que son linealmente independientes.

$$A_1 = \begin{bmatrix} D_z \sin\alpha & -D_z \cos\alpha & D_y \\ D_z \cos\alpha & D_z \sin\alpha & -D_x \\ -D_y \cos\alpha - D_x \sin\alpha & -D_y \sin\alpha + D_x \cos\alpha & 0 \end{bmatrix} \tag{3.53}$$

y la matriz de términos independientes B:

$$B_1 = \begin{bmatrix} D_{z_1} \cdot (\cos\alpha \cdot T_{y_1} - \sin\alpha \cdot T_{x_1}) - D_{y_1} T_{z1} \\ D_x T_{z_1} - D_{z_1} \cdot (\cos\alpha \cdot T_{x_1} + \sin\alpha \cdot T_{y_1}) \\ D_{y_1} \cdot (\cos\alpha \cdot T_{x_1} + \sin\alpha * T_{y_1}) - D_{x_1} \cdot (\cos\alpha \cdot T_y - \sin\alpha) \cdot T_{x_1}) \end{bmatrix} \tag{3.54}$$

El vector director $D_1^{OA} = [D_{x_1}, D_{y_1}, D_{z_1}]$ se obtiene a partir del resultado del algoritmo de estimación de máximos "Nom Maximun Supression" en la sección 3.4;

Se tiene entonces que:

$$A_1 X^{OW} = B_1 \tag{3.55}$$

Donde $A_1 X^{OW}$ representa un sistema de ecuaciones lineales que tiene 2 ecuaciones linealmente independientes y 3 incógnitas. De manera análoga para el segundo sistema de coordenadas se obtiene un sistema de ecuaciones lineales que tiene 2 ecuaciones linealmente independientes y 3 incógnitas. Combinándose los dos sistemas ecuaciones se obtiene un sistema con cuatro ecuaciones linealmente independientes y 3 incógnitas. Obteniendo X^{OW} como la solución de mínimo error cuadrático.

$$X^{OW} = \left(A^T A\right)^{-1} A^T B \tag{3.56}$$

En las Figuras 3.24 y 3.25 se representan los resultados obtenidos en la estimación de la posición en el espacio 3D por el proceso de triangulación, para el caso ideal en que no existen errores en las proyecciones y para el caso real en el que las proyecciones se ven afectadas por diferentes tipos de ruido respectivamente.

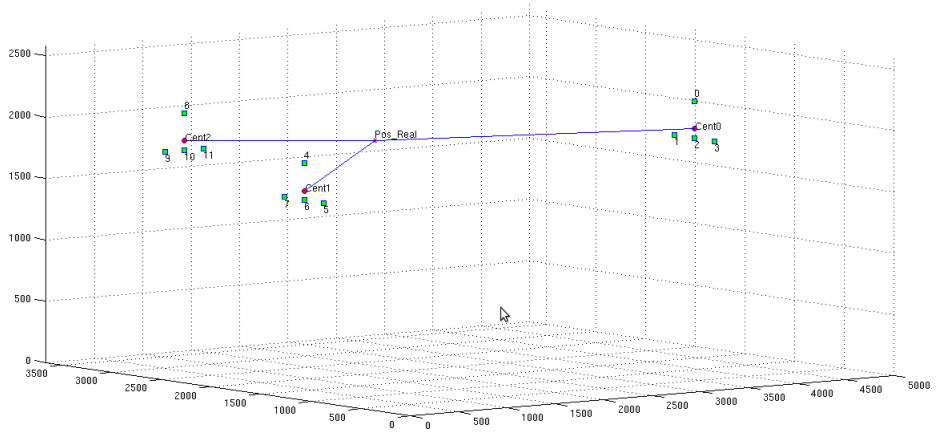


Figura 3.24: Triangulación en el caso ideal, las proyecciones no están afectadas por el ruido

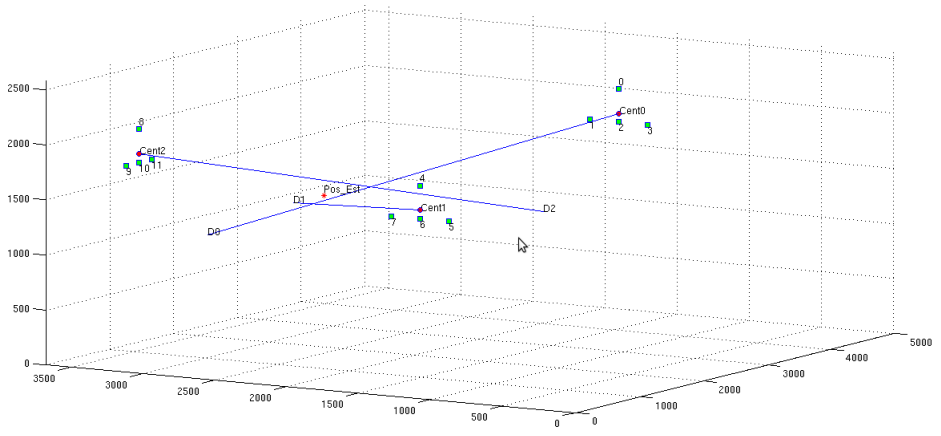


Figura 3.25: Triangulación en el caso real, las proyecciones están afectadas por el ruido

3.6. Técnicas de estimación de coherencia

La estimación de la posición del locutor mediante triangulación se ve muy afectado cuando ocurre algún error de consideración en el cálculo del mapa de potencia acústica en cualquiera de los *arrays* de micrófonos, por esta razón se consideró aplicar alguna técnica de eliminación de *outliers* que permita eliminar del proceso de triangulación los *arrays* de micrófonos que tengan errores muy grandes. Se implementó una variante del algoritmo RANSAC en la que a diferencia de este se usan las posiciones estimadas para todos los pares de *arrays* de micrófonos. La idea es estimar una posición para cada pareja de *array* de micrófono posible y el conjunto de posiciones que estén más lejanas se consideran *outliers* (a diferencia de RANSAC se utilizan todo el conjunto de soluciones). Luego se calcula la distancia euclidiana que existe entre cada una de las posiciones y los puntos que se encuentren a menos de una distancia definida como umbral se definen como *inliers* (conjunto de posiciones estimadas que son coherentes a la posición correspondiente), formándose un grupo de posiciones para cada una de las posibles combinaciones de *arrays* de micrófonos. Se seleccionan todos los *arrays* que hayan calculados las posiciones del grupo más poblado para realizar la triangulación y calcular la posición final. La distancia seleccionada como umbral, se debe de escoger de forma tal que se conozca y se fije la probabilidad de que un punto se considere como *inliers*, este cálculo requiere conocer la distribución de probabilidad del error en el posicionamiento. Como paso inicial en este trabajo el umbral es seleccionado empíricamente.

3.7. Conclusiones

En este capítulo se ha descrito todo el desarrollo teórico de las técnicas que intervienen en la implementación del sistema propuesto en la tesis de Máster, haciendo hincapié en la estrategia del modelado de los *arrays* de micrófonos como cámaras de perspectiva, para lo cual se partió del modelo de una cámara de perspectiva. Se abordó todo el proceso de generación de localizaciones formando rayos que parten del centroide de cada *array* y la utilización del la técnica SRP-PHAT para calcular la potencia acústica en cada una de las direcciones. Se explicó todo el proceso

de generación de la imágenes y de los mapas de energía acústica. Se expuso el algoritmo de triangulación DLT utilizado para estimar la posición 3D del locutor activo, se desarrolló una variante del algoritmo RANSAC para detectar los mapas en los que el máximo de potencia se detecta de manera errónea, para eliminarlos del proceso de localización final.

Capítulo 4

Resultados experimentales

4.1. Introducción

En este capítulo se exponen los diversos experimentos realizados sobre las bases de datos AIT e ITC del proyecto CHIL (Computer in the Human Interactional Loop). En cada experimento realizado se utilizan las métricas de evaluación definidas por el mismo proyecto CHIL bajo la campaña de evaluación CLEAR, que permiten evaluar las prestaciones del sistema de localización implementado en esta Tesis de Máster.

4.2. El proyecto CHIL

El proyecto CHIL [83] financiado por la Unión Europea comenzó el 1 de Enero del 2004 con una duración de 44 meses. En él se realizan intentos para desarrollar equipos que asistan a las actividades, interacciones e intenciones humanas. Para lograrlo se necesita que las máquinas se adapten y aprendan los intereses, actividades, metas y aspiraciones de las personas. Lo que requiere un entendimiento de todas las señales de comunicación como el habla, las expresiones faciales, la atención, la emoción, los gestos etc.. El sistema de evaluación de CHIL fue organizado conjuntamente con la campaña de evaluación de CLEAR (*Classification of Events, Activities and Relationships*) desarrollada desde el 2004 al 2007. Dicha campaña constituye un esfuerzo internacional para evaluar sistemas que son diseñados para analizar a las personas, sus identidades, actividades, interacciones y en escenarios con interacción hombre-hombre. CLEAR intenta unir a los proyectos e investigadores que trabajan en estas tecnologías con el fin de establecer una campaña de evaluación internacional común. La campaña oficial CLEAR 2007 [84] se realizó en los meses de Febrero a Abril del 2007 y fue conducida por el taller organizado en Mayo del 2007 en Baltimore, USA donde fueron evaluadas las nueves tecnologías siguientes:

- Dentro del área de visión:
  - Detección y seguimiento 2D de rostro. Este proceso evalúa la calidad y precisión del sistema de detección y seguimiento del rostro.
  - Seguimiento 2D de personas. Este proceso evalúa la calidad y la precisión del sistema de seguimiento, cuya posición se define como el centro de la cabeza de la persona proyectado al suelo.
  - Identificación visual de personas. El objetivo de este proceso es el reconocimiento de personas usando secuencias de imágenes de hasta 20 segundos. La evaluación se realizó en una base de datos de seminario y en otra base de datos de seminario iterativa. En la primera de ellas el objetivo es solamente la identificación del presentador, mientras en la otra además es necesario identificar a los participantes.
  - Estimación de pose. El objetivo es estimar la pose de personas. La pose es determinada por tres ángulos llamados: balance, inclinación y pan.
- Dentro del área de audio:
  - Seguimiento de la persona que está hablando. El objetivo principal es localizar las personas que están hablando dentro de la habitación. Conocer la posición del hablante es útil para realizar el *beamforming* muy usado en la transcripción automática del habla desde lejos.
  - Identificación de la persona que está hablando.
  - Detección de los hablantes activos.

- Dentro de las tecnologías multimodales
  - Seguimiento multimodal de las personas
  - Identificación multimodal de las personas. En la identificación de personas multimodal, se realiza una fusión de la señales de audio y video para identificar a la persona que está hablando.

CHIL está compuesta por 7 bases de datos, que fueron usadas por el sistema de evaluación, estas están cosntituidas por un conjunto de grabaciones de señales de audio y video en seminarios interactivos los cuales fueron:

- AIT Seminars, CLEAR 2007
- IBM Seminars, CLEAR 2007
- UKA-ISL Seminars, CLEAR 2007
- FBK Seminars, CLEAR 2007
- UPC Seminars, CLEAR 2007
- UKA-ISL, CLEAR 2007 Head Pose Estimation database
- ITC-irst, CLEAR 2007 Acoustic Event Detection database

Cada seminario consta de una presentación hecha en una sala de reuniones. Estas presentaciones se llevan a cabo por una o varias personas (3 a 7 por seminario) donde están presentes un grupo de oyentes. Los temas están relacionados con cuestiones técnicas del proyecto CHIL (en su mayoría Procesamiento del Lenguaje Natural). Durante y después de la presentación hay preguntas de los asistentes con las respuestas del presentador. También hay actividad en términos de personas que entran y/o salen de la habitación, abriendo y cerrando la puerta, que se levantan y van a la pantalla, con pausas para el café, y que hablan unas con otras, etc.

En algunos de los seminarios interactivos, se utilizaron unas secuencias de *scripts* para capturar determinadas actividades (e.g., las aperturas y cierras de las puertas). Estas actividades se detectan automáticamente en las tareas de evaluación (por ejemplo, la clasificación de eventos acústicos).

Los resultados del sistema de localización y seguimiento del locutor activo desarrollado en esta Tesis de Máster se realizan sobre los seminarios AIT e ITC. Partiendo de un experimento base utilizado durante todo el proceso de desarrollo se hace una evaluación de la influencia de las distintas etapas que componen el sistema implementado. En concreto se evaluaron los siguientes aspectos:

- Estudio del efecto de las localizaciones utilizadas para calcular la potencia acústica para los cuales se estudian:
  - Los efectos de barrido en ángulos.
  - Los efectos del barrido en profundidad.
  - Los efectos de la estrategia de generación de rayos.
- Estudio del efecto del proceso de triangulación.
- Estudio del efecto del proceso de estimación de coherencia.

4.3. Estrategia de evaluación y métricas

Para evaluar los resultados de los distintos bloques desarrollados, así como del sistema completo se ha evaluado el mismo utilizando un conjunto de datos de audio de las bases de datos de AIT e ITC. Los resultados son evaluados por el conjunto de métricas definidas en el proyecto Computers in the Human Interaction Loop (CHIL).

Se realiza la evaluación del sistema de localización de un locutor, en los instantes en que éste está activo. Con el objetivo de medir la precisión del sistema de localización se definen el siguiente conjunto de métricas:

- Pcor. Es la relación entre la cantidad de *frames* en que el error del posicionamiento medido respecto al etiquetado de *grountruth* es menor que 500 mm *(fine error)* y la cantidad de *frame* en los que hay algún locutor activo y se calcula como:

$$Pcor = \frac{\text{nº error fine}}{\text{nº frames con groundtruth y estimación}} \tag{4.1}$$

- *Bias fine*: La media del error con respecto a los *groundtruth* en las estimaciones donde el error es menor que 500 mm *(fine error)* y se calcula como:

$$Bias\ fine = \frac{\sum_{i=1}^{No.Fine\_error} dist\_err(i)}{No.Fine\_error} \tag{4.2}$$

- *Bias fine+gross*: El error medio en milímetros cometido en todas las estimaciones realizadas, se calcula como:

$$Bias\ fine + gross = \frac{\sum_{i=1}^{(No.Fine\_error+Gross\_error)} dist\_err(i)}{(No.Fine\_error + Gross\_error)} \tag{4.3}$$

- *MOTP (Multiple Object Tracking Precision)*: Indica la precisión del sistema como la suma de las distancias euclídeas en mm que existe entre cada estimación, cuyo error está definido como *fine error*, normalizada con la cantidad total de *fine error*. Este parámetro es igual al Bias fine.

- Deletions. se define como el por ciento (%) de la cantidad de *frames* en los que el sistema de localización no devuelve resultado respecto a la cantidad total de *groundtruths*.

También se muestran el número de *frames* detectados, la duración en segundos de los segmentos anotados y del experimento.

En el sistema de localización implementado siempre se estima la posición de un posible locutor y los *deletion* generalmente corresponden a estimaciones que se encuentran fuera del espacio de búsqueda, las cuales son eliminadas de manera automática. A lo que no se atiende es a la presencia de los falsos positivos, debido a la falta de un *Voice Active Detector* (VAD).

4.4. Bases de datos

Como se expresó anteriormente los seminarios utilizados para realizar la evaluación fueron:

1. "AIT-TEST Seminars 2006" que consta de cinco seminarios registrados en la sala AIT entre los meses de Julio y Octubre del 2006. Cada seminario tiene una duración de 30 minutos. El primero de ellos se utilizó para el desarrollo. En el resto de los seminarios se extrajeron dos segmentos de cinco minutos para realizar la evaluación, obteniéndose un total de 8 segmentos de evaluación.

   A continuación se hace una breve descripción de las características fundamentales de la habitación donde se realiza el seminario AIT mostrado en la Figura 4.1.

   - El extremo izquierdo inferior es seleccionado como el origen del sistema de coordenadas, donde el eje Z refleja la altura.
   - En la esquina superior derecha se muestra el extremo superior de la habitación que permite determinar sus límites.
   - Los *arrays* de micrófonos utilizados en el sistema de localización fueron los A, B y C cuya geometría se puede observar en la parte derecha de la figura.
   - En el plano se muestra en cada uno de los *arrays* de micrófonos las coordenadas 3D del micrófono 1.
   - En el centro se encuentra ubicada una mesa redonda donde se encuentran cada uno de los locutores que intervendrán en la habitación.
   - También se pueden observar las ubicaciones de las cámaras de video, la puerta de entrada y salida, la pizarra donde aproximadamente estará ubicado el locutor y el *array* de micrófonos D formado por 64 elementos pero que no es utilizado en el sistema implementado.

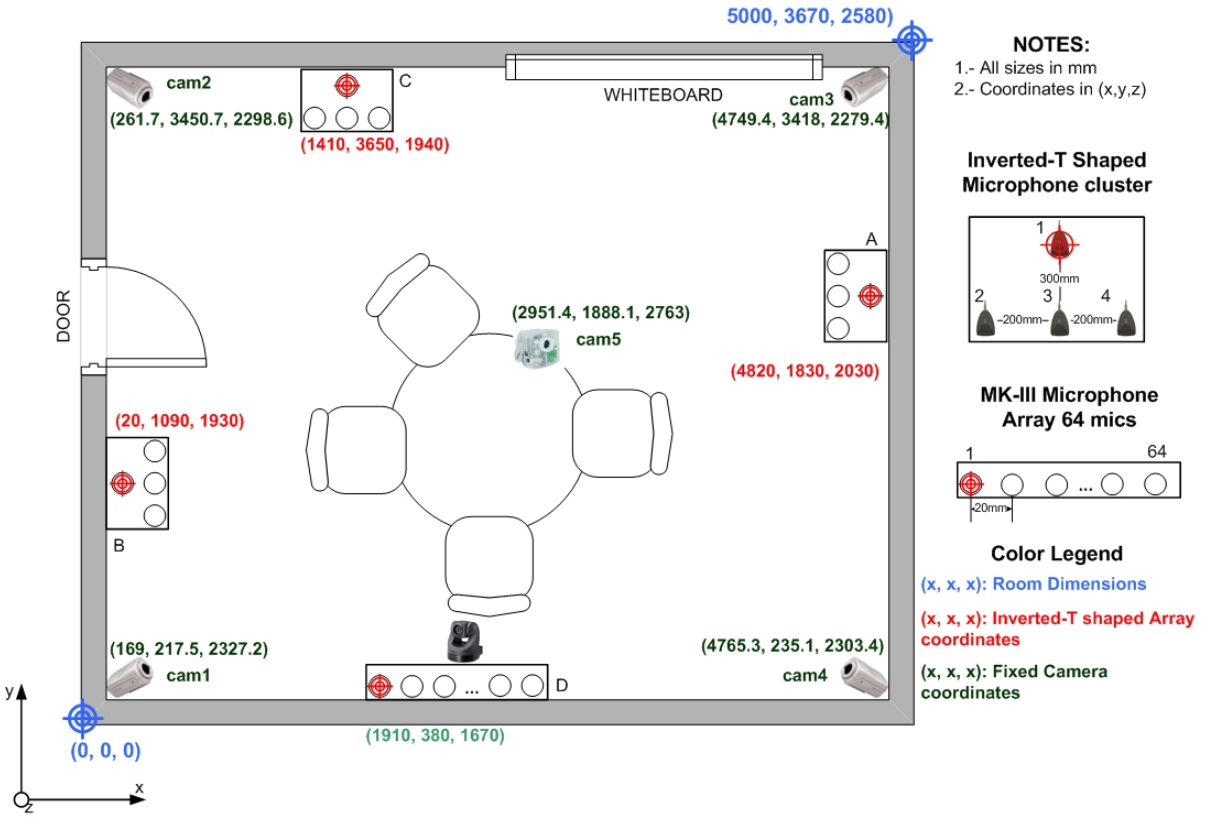


Figura 4.1: Plano del seminario AIT 2006

2. ITC-DEV-2007. En el que las grabaciones de audio se realizan con 32 micrófonos. De ellos en las paredes se ubican 7 *arrays* (array_a-array_g) formados por cuatros micrófonos con geometría en forma de T, que son los utilizados por el sistema de localización implementado, además encima de una mesa ubicada en el centro de la habitación hay un octavo *array* con cuatro micrófonos ubicados en forma circular (table_1-table_4). También se pueden observar las dimensiones de la habitación y la posición de las 4 cámaras de video. En la parte inferior izquierda se sitúa el origen del sistema de coordenadas de referencia tomándose la altura como el eje Z.

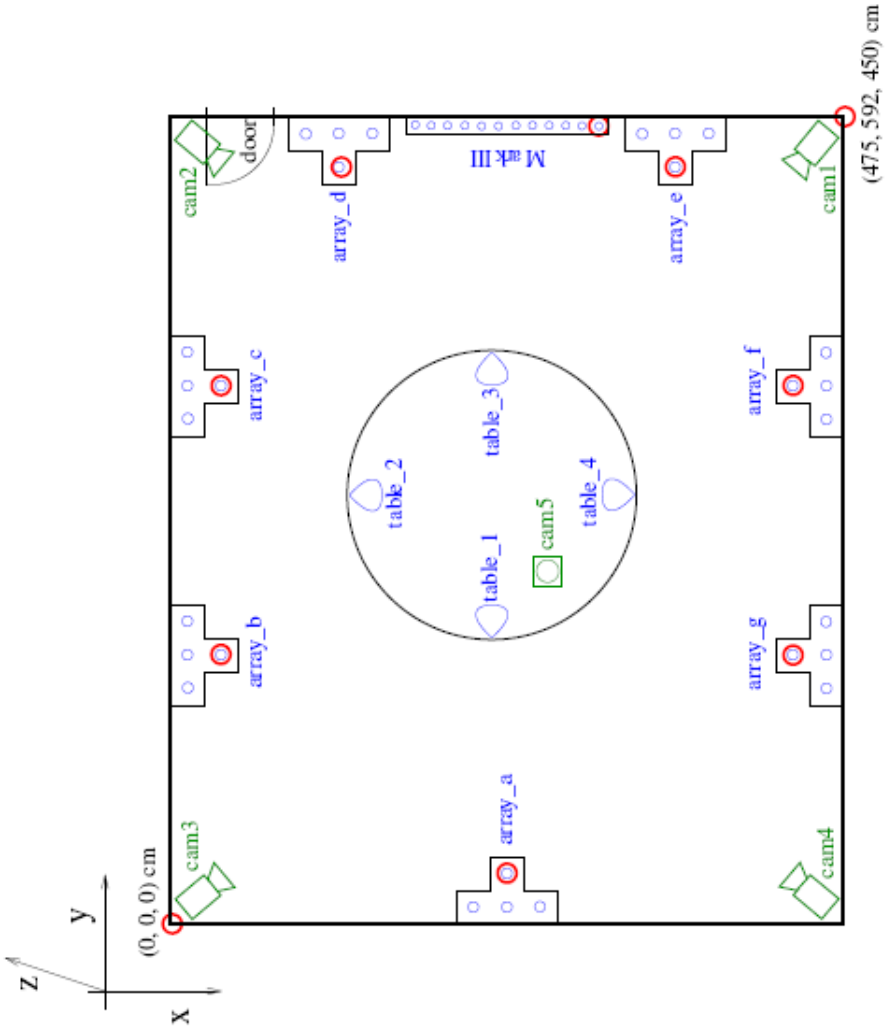


Figura 4.2: Plano del seminario ITC-DEV

4.5. Experimento base

En esta sección se describen los parámetros fundamentales utilizados para definir un experimento base *(baseline)*, que se utilizará como patrón de comparación con otros experimentos realizados en los entornos AIT e ITC de CHIL. Se definen dos grupos de parámetros, el primero de ellos está relacionado con el proceso del cálculo de SRP, los cuales se mantendrán invariantes a lo largo de toda la experimentación, el segundo está relacionado con todo el proceso de

generación de las localizaciones en las que se va a calcular la potencia acústica en las distintas direcciones para cada *subarray* de micrófonos alguno de los cuales serán variados a lo largo de todo el proceso de experimentación.

Descripción de los parámetros que definen el *baseline*.

1. Parámetros relacionados son el proceso de generación de localizaciones:
   - Tipo de localización no esférica. En el que cada rayo de localización se genera a partir del centroide del *array* hasta llegar a cualquiera de los límites de la habitación (ver Figura 4.3).
   - Cantidad de puntos utilizado en el barrido horizontal *(azimuth)*: 320 puntos.
   - Cantidad de puntos utilizados en el barrido vertical (elevación): 240 puntos.
   - Paso en profundidad: 100 mm.
   - Alturas máxima y mínima a la que se puede encontrar una persona hablando: 2000 mm *y* 750 mm.
2. Parámetros de configuración utilizados en el algoritmo de SRP-PHAT.
   - Frecuencia de muestreo con la que se realizaron las grabaciones en las bases de datos: 44,1kHz
   - Tamaño de la ventana de tiempo en la que se realizará la localización, "FRAME_SIZE_SECS", : 0,5seg.
   - Desplazamiento de la ventana de análisis "FRAME_SHIFT_SECS": 0,5seg. Se define del mismo tamaño que el "FRAME_SIZE_SECS" debido a que para realizar las estimaciones no se usa ningún algoritmo de seguimiento.
   - Cantidad de puntos de la FFT, "FFT_SIZE": 32768.
   - Tipo de ventana: Hamming
   - Correlación del tipo: GCC-PATH.
   - No se realiza filtrado de la señal de audio.

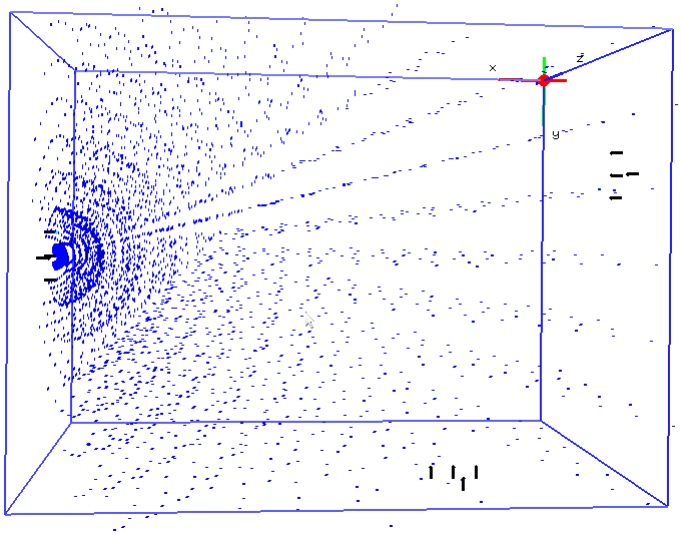


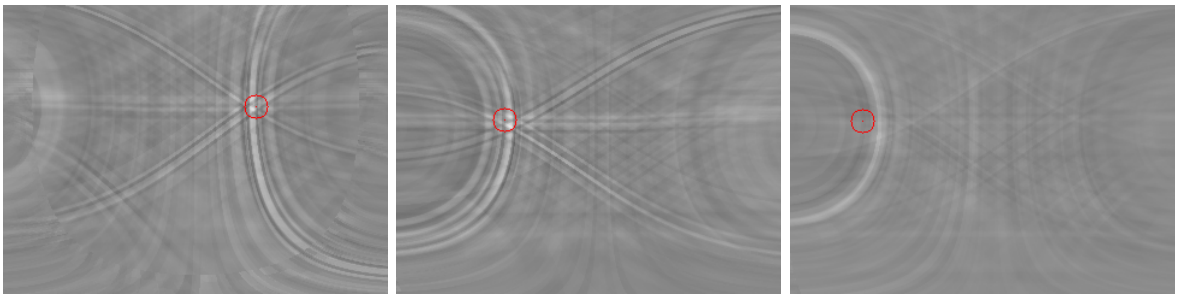
Figura 4.3: Localizaciones generadas del tipo no esférica

4.5.1. Resultados del proceso de generación de imágenes

Las imágenes son generadas en escala de grises y el valor de cada pixel corresponde a la potencia acústica calculada por SRP-PHAT en una dirección, partiendo del centroide del *array* de micrófonos. En las imágenes obtenidas a partir de los *frames* donde hay una persona hablando, se dibuja un círculo de color rojo en la proyección de la posición real, la cual se obtiene a partir del etiquetado de *groundtruth* proporcionada por las base de datos. Debido a que cada *array* está compuesto por 3 micrófonos colocados linealmente a la misma altura y un cuarto que se encuentra más alto formando una distribución en forma de T, (ver Figura 4.1) se observan arcos de color blanco más intensos que las líneas horizontales, esto sucede debido a que en SRP-PHAT la potencia en cada localización se calcula como la suma de las correlaciones cruzadas de todas las combinaciones de pares de micro posibles desplazadas por las diferencias de tiempos que se retrasa la señal de audio en llegar a cada micrófono del par correspondiente.

Como se explicó anteriormente la técnica de localización SRP-PHAT es la que mejor comportamiento tiene en entornos reverberantes pero se ve muy afectada cuando los niveles de potencia de la señal son bajos, debido entre otras cosas a que la distancia que existe entre el locutor y el *array* de micrófonos es grande o que éste no esté hablando en la dirección a donde se encuentra el *array*, estos factores provocan comportamientos erráticos en el cálculo de la potencia acústica en algunos *frames*. A continuación se muestran algunas de las imágenes generadas por cada *array* de micrófonos que definen los comportamientos de SRP-PHAT siguientes:

1. Un *frame* donde el punto de mayor intensidad es bastante congruente con el *groundtruth*, ver Figura 4.4.
2. Un *frame* donde el punto de mayor intensidad en el *array* 2 no coincide con el *groundtruth*, ver Figura 4.5.
3. Un *frame* donde el punto de mayor intensidad sólo coincide con el *groundtruth* en el *array* 1, ver Figura 4.6.
4. Un *frame* donde no existe ninguna persona hablando, ver Figura 4.7.

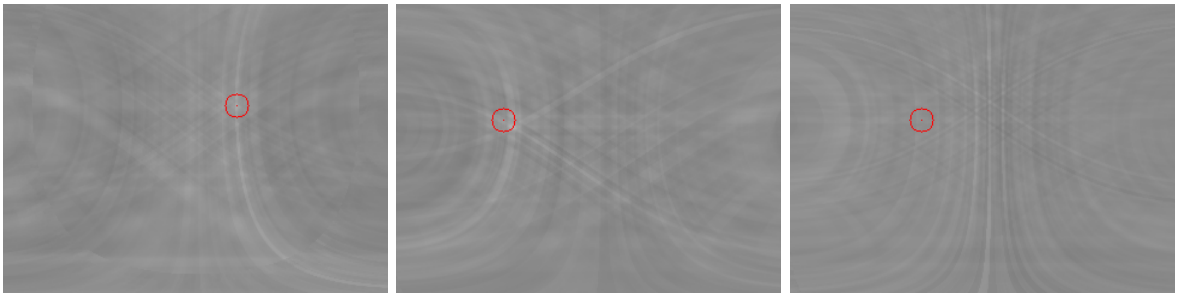


(a) Array 1, Frame 69 (b) Array 2, Frame 69 (c) Array 3, Frame 69

Figura 4.4: Imágenes de potencia acústica calculada en un *frame* de la base de datos AIT de CHIL, en la que los mapa de energía del algoritmo SRP es congruente con el *groundtruth* en los tres *arrays*

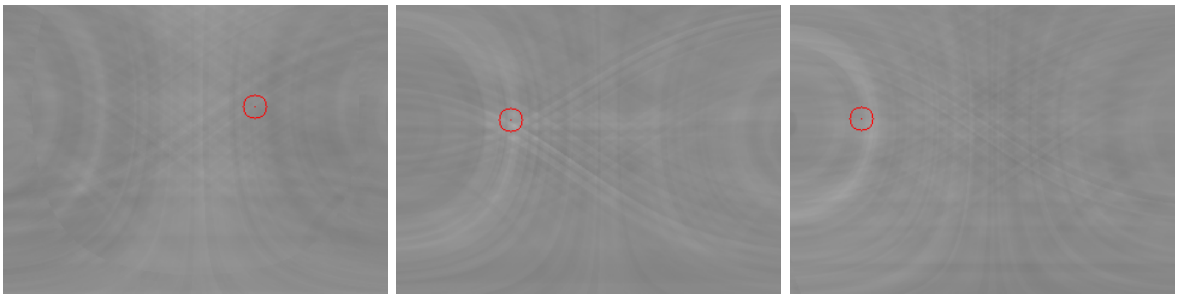
4.5.2. Resultados del proceso de cálculo de máximos

Una vez obtenidas las imágenes de potencia acústica para cada *array* de micrófonos se realiza el proceso de búsqueda de máximos con el objetivo de encontrar la dirección de máxima energía



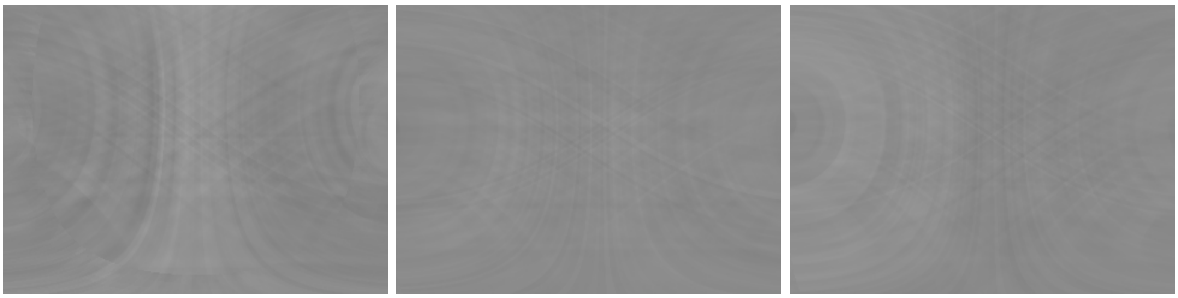
(a) Array 1, Frame 65 (b) Array 2, Frame 65 (c) Array 3, Frame 65

Figura 4.5: Imágenes de potencia acústica calculada en un *frame* de la base de datos AIT de CHIL, donde el mapa de energía del algoritmo SRP es congruente con el *groundtruth* en dos de los tres *arrays*



(a) Array 1, Frame 102 (b) Array 2, Frame 102 (c) Array 3, Frame 102

Figura 4.6: Imágenes de potencia acústica calculada en un *frame* de la base de datos AIT de CHIL, donde el mapa de energía del algoritmo SRP es congruente con el *groundtruth* en uno de los tres *arrays*

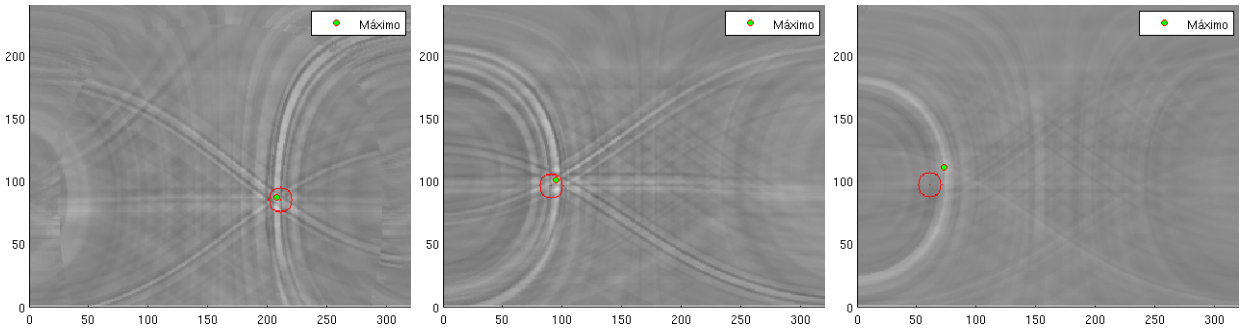


(a) Array 1, Frame 237 (b) Array 2, Frame 237 (c) Array 3, Frame 237

Figura 4.7: Imágenes de potencia acústica calculada en un *frame* de la base de datos AIT de CHIL, en la que no hay ningún locutor activo

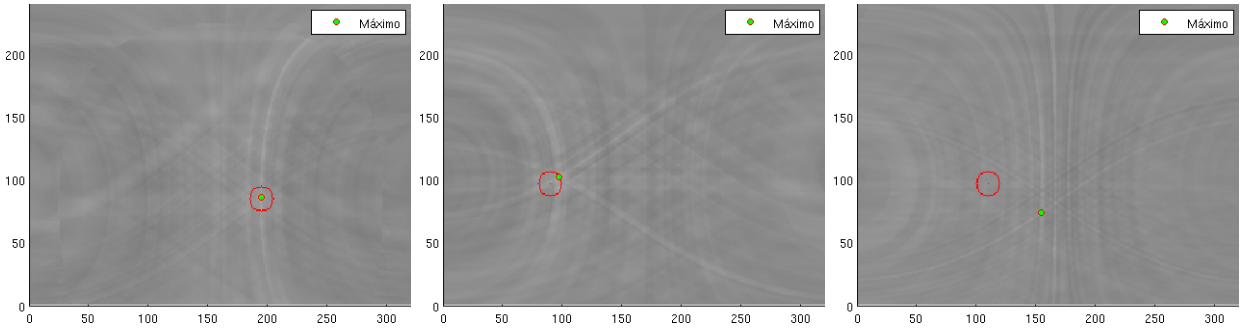
donde se espera encontrar la persona que está hablando.

A continuación se muestran unas serie de imágenes que describen el comportamiento del algoritmo *Non Maximun Supression* con aproximación subpixélica utilizado, ver Figuras 4.8, 4.9, 4.10 y 4.11, donde se representa con un círculo de color verde el máximo de energía encontrado.



(a) Array 1, Frame 69 (b) Array 2, Frame 69 (c) Array 3, Frame 69

Figura 4.8: Representación de los máximos encontrados en la que el error respecto al *groundtruth* en las tres imágenes es pequeño



(a) Array 1, Frame 65 (b) Array 2, Frame 65 (c) Array 3, Frame 65

Figura 4.9: Representación de los máximos encontrados en un *frame* donde el error respecto al *groundtruth* en una de las imágenes (c) es bastante grande.

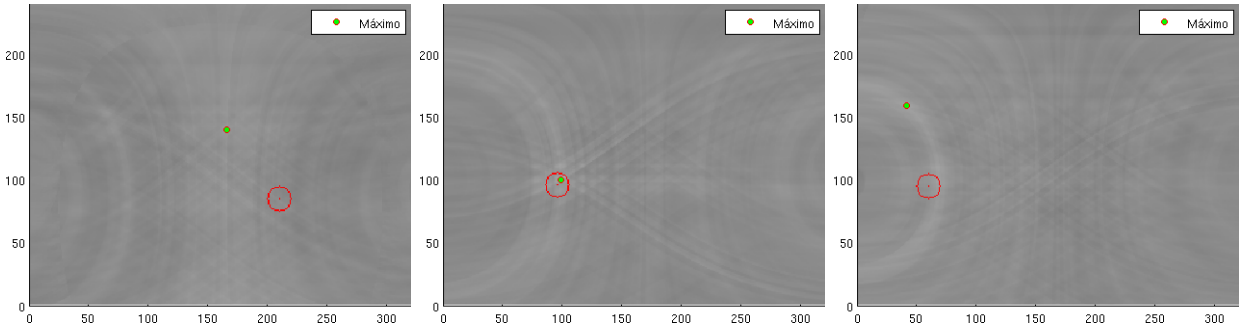
Como se puede ver en la tabla 4.1 los errores cometidos en la búsqueda de los máximos de potencia acústica en los tres *arrays* son parecidos, siendo considerablemente mayores en *azimuth* que en elevación.

	Azimuth			Elevación		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	11,70º	10,33º	11,82º	5,84º	4,63º	7,56º
error >20º	20,61 %	18,80 %	19,94 %	7,44 %	4,96 %	8,73 %

Tabla 4.1: Errores en *azimuth* y elevación cometidos en la estimación de los máximos en el experimento *baseline* de la base de datos de AIT

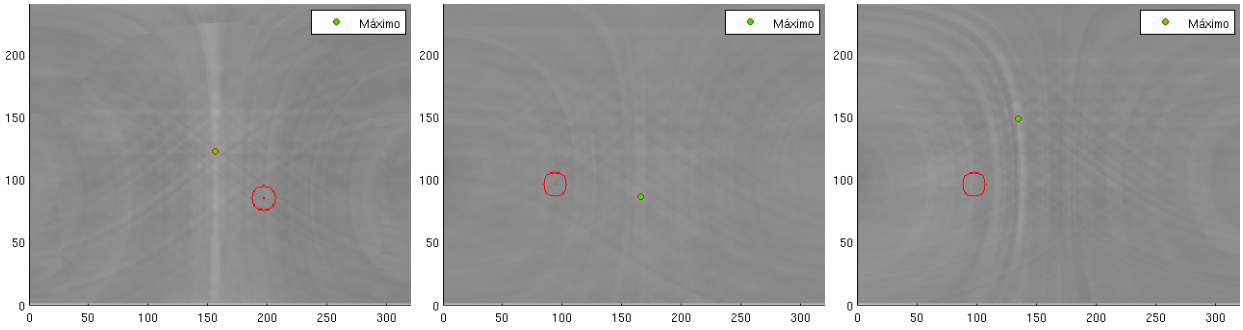
4.5.3. Resultados del proceso de triangulación

Las proyecciones de los máximos de energía en las imágenes correspondientes a cada *subarray* de micrófonos obtenidas en el paso anterior son utilizadas para realizar la triangulación como



(a) Array 1, Frame 102 (b) Array 2, Frame 102 (c) Array 3, Frame 102

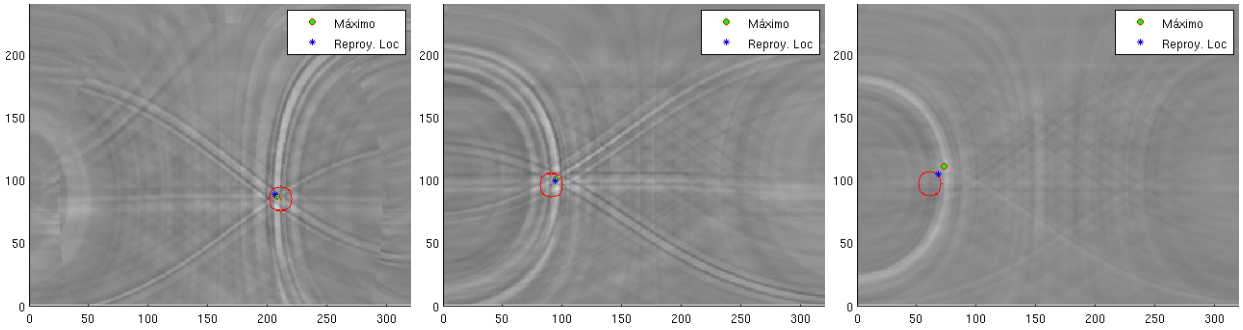
Figura 4.10: Representación de los máximos encontrados en un *frame* donde el error respecto al *groundtruth* en las tres imágenes es bastante grande



(a) Array 1, Frame 64 (b) Array 2, Frame 64 (c) Array 3, Frame 64

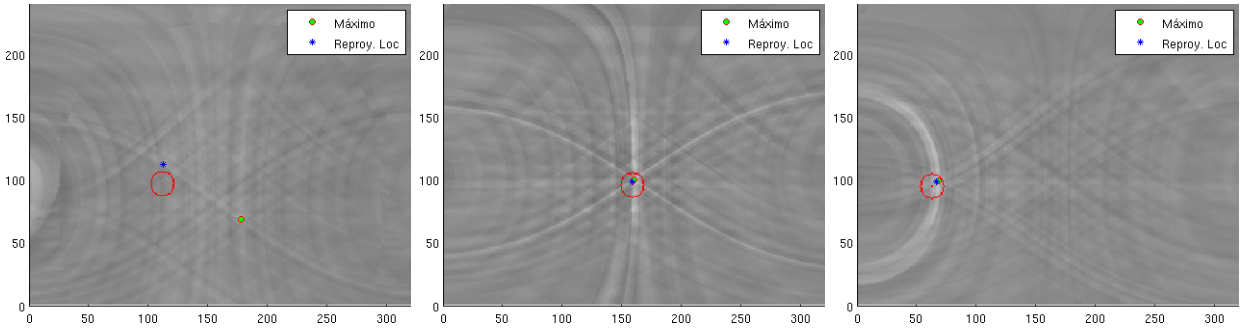
Figura 4.11: Representación de un *frame* donde los máximos encontrados no coinciden con el *groundtruth* en ninguno de los tres *arrays*

paso final del sistema de localización diseñado, donde se obtiene la estimación de la posición 3D del locutor activo. En las Figuras 4.12, 4.13, 4.14 y 4.15 se presentan distintos comportamientos del sistema de localización donde se representan las proyecciones de los máximos de energía utilizados en el proceso de triangulación con un círculo de color verde y las proyecciones de la estimación de la posición 3D con un punto de color azul.



(a) Array 1, Frame 69 (b) Array 2, Frame 69 (c) Array 3, Frame 69

Figura 4.12: Representación de los resultados del proceso de triangulación en un *frame* donde el error en la estimación es pequeño, las imágenes muestran con un asterisco en color azul la reproyección del posicionamiento en los tres mapas de energía generados por los *arrays*



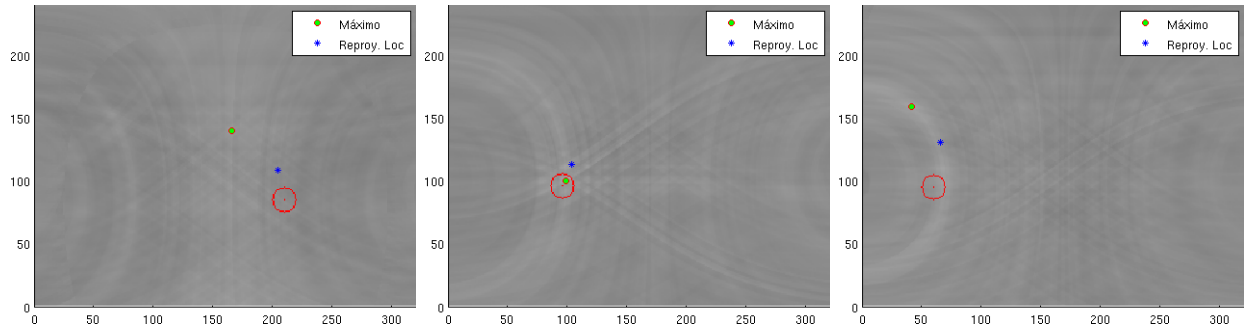
(a) Array 1, Frame 153 (b) Array 2, Frame 153 (c) Array 3, Frame 153

Figura 4.13: Representación de los resultados del proceso de triangulación en un *frame* donde el error en la estimación es pequeño, a pesar que en el *array 3* la estimación del máximo de energía no fue bueno

4.5.4. Resultados globales (con las métricas finales)

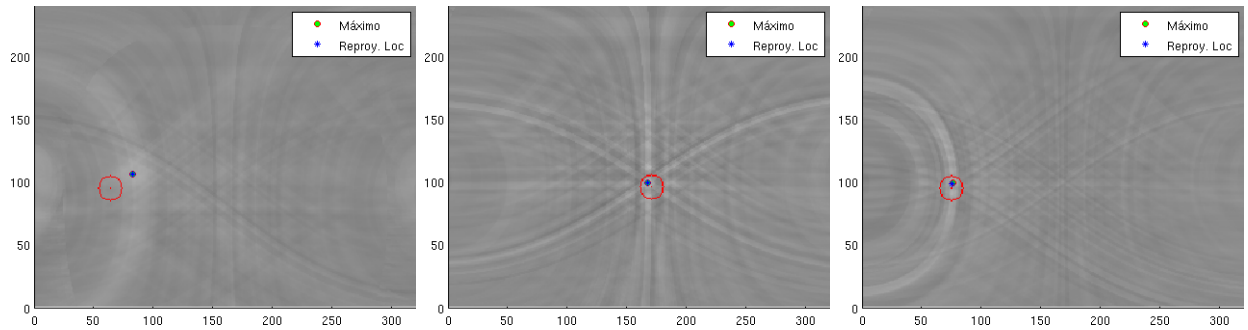
En esta sección se muestran los resultados globales del sistema de localización obtenidos con las métricas de CLEAR de los experimentos de AIT e ITC tomados como *baseline*. Las métricas de CHIL muestran que los errores cometidos en la base de datos de AIT son menores que en ITC, (ver tabla **??**). En AIT los errores en la elevación (eje z) son considerablemente mayores que los detectados en los ejes x e y, lo cual es una contradicción pues las imprecisiones de elevación que se tuvieron en la estimación de los máximos de potencia, (ver tabla 4.1 de la página 68) fueron menores que los de *azimuth*.

Las grabaciones en la base de datos de ITC se obtuvieron a partir de 7 *arrays* de micrófonos, debido a esto los errores cometidos pueden ser inherentes a la forma en que se realiza el posicionamiento. Osea si ocurren errores muy grandes en la búsqueda de los máximos de energía en los



(a) Array 1, Frame 102 (b) Array 2, Frame 102 (c) Array 3, Frame 102

Figura 4.14: Representación de los resultados del proceso de triangulación en un *frame* en el que ocurre un error grande en la estimación de la posición debido a que la estimación del máximo de energía acústica en los 3 *arrays* es grande



(a) Array 1, Frame 140 (b) Array 2, Frame 140 (c) Array 3, Frame 140

Figura 4.15: Representación de los resultados del proceso de triangulación donde la proyección de la posición es casi perfecta, a pesar de cometerse un error apreciable en la estimación del máximo en el *array 1*

mapas de potencia la triangulación se cometen errores muy grandes en la triangulación. Debido a esto, se plantea el uso de una técnica de estimación de coherencia, que permita separar del proceso de triangulación los máximos erróneos *(outliers)*.

4.6. Estudio del efecto del barrido en ángulos

En esta se realiza un estudio detallado de los efectos que tiene la resolución en el barrido de ángulos, en los pasos en los que está dividido el sistema de localización implementado, y en los resultados globales medidos por las métricas de CHIL. Para ello se parte del experimento *baseline* cuya resolución es 320 en *azimuth* y 240 en elevación, se repiten los experimentos con cada una de las siguientes resoluciones: (240×180) (160×120) (80×60) y se van mostrando en cada apartado los resultados obtenidos.

4.6.1. Resultados del proceso de cálculo de máximos

En este apartado se realizará un estudio del efecto que tiene el uso de distintas resoluciones en el proceso de búsqueda del máximo global en las imágenes que se obtienen de cada *array* de micrófonos. En la figura 4.16 se puede observar como a medida que disminuye la resolución de los barridos en *azimuth* y elevación las imágenes se van granulando. En las tablas 4.2, 4.3, 4.4, 4.5 se muestran el promedio del error absoluto y el por ciento de *frames* en los que dicho error es menor de 20º. Además se puede observar que los errores cometidos en elevación son menores que los de *azimuth* y van aumentando gradualmente a medida que disminuye la resolución.

En la tabla 4.6 se representa un resumen de los errores cometidos en *azimuth* y elevación para las distintas resoluciones, donde se promedian los errores cometidos por los tres *arrays*.

	80 × 60			160 × 120		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	12,29º	9,12º	17,97º	11,70º	10,02º	13,24º
error >20º	22,12 %	17,27 %	33,12 %	20,54 %	18,29 %	23,88 %

Tabla 4.2: Errores de *Azimuth* cometidos en la estimación de los máximos en la base de datos de AIT utilizando distintas resoluciones

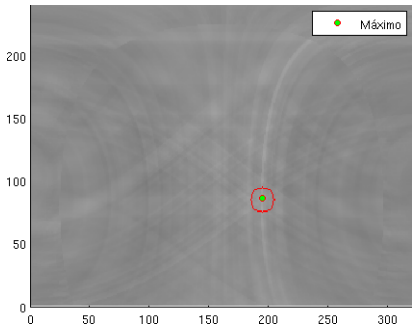
	240 × 180			320 × 240		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	13,51º	11,08º	13,47º	11,70º	10,33º	11,82º
error >20º	24,48 %	19,60 %	23,15 %	20,61 %	18,80 %	19,94 %

Tabla 4.3: Errores de *Azimuth* cometidos en la estimación de los máximos en la base de datos de AIT utilizando distintas resoluciones

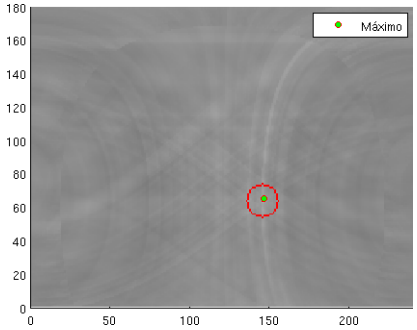
4.6.2. Resultados globales (con las métricas finales)

En este apartado se realiza un estudio del efecto que tiene el uso de distintas resoluciones en el proceso de búsqueda del máximo global mediante las métricas globales de CLEAR.

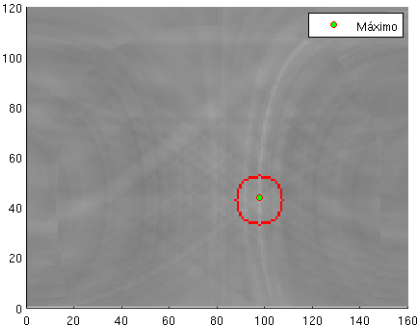
Las métricas de CHIL muestran un aumento en los errores de manera gradual, a medida que disminuye la resolución en los barridos de *azimuth* y elevación, (ver tabla 4.7). Los errores



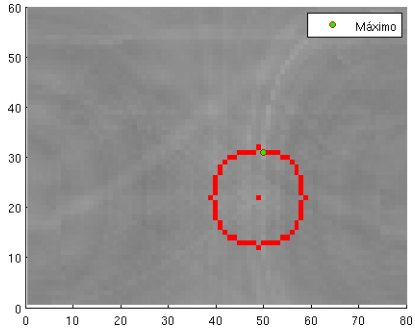
(a) SubArray 0, Frame 65, Resolución 320×240



(b) Array 1, Frame 65, Resolución 240×180



(c) Array 2, Frame 65, Resolución 160×120



(d) Array 3, Frame 65, Resolución 80×60

Figura 4.16: Representación de los resultados del proceso de búsqueda de máximos para distintas resoluciones de los mapas de energía acústica partiendo de un *frame* donde el error en la estimación de este es pequeño

	80×60			160×120		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	6,86º	4,77º	9,34º	5,83º	4,75º	8,06º
error >20º	10,13 %	6,07 %	15,77 %	7,51 %	5,19 %	10,54 %

Tabla 4.4: Errores de elevación cometidos en la estimación de los máximos en la base de datos de AIT utilizando distintas resoluciones

	240×180			320×240		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	5,85º	4,72º	7,06º	5,84º	4,63º	7,56º
error >20º	6,85 %	4,49 %	8,09 %	7,44 %	4,96 %	8,73 %

Tabla 4.5: Errores de elevación cometidos en la estimación de los máximos en la base de datos de AIT utilizando distintas resoluciones

Métricas	320×240		240×180		160×120		80×60	
	azimuth	elevación	*azimuth*	elevación	*azimuth*	elevación	*azimuth*	elevación
Media del error	11,28º	6,01º	12,69º	5,87º	11,65º	6,21º	13,13º	7,00º
error >20º	19,78 %	7,04 %	22,41 %	6,47 %	20,90 %	7,75 %	24,17 %	10,66 %

Tabla 4.6: Resumen de los errores cometidos para distintas resoluciones en el barrido horizontal

cometidos en elevación son considerablemente mayores que en los otros ejes. También se puede observar una disminución apreciable del tiempo de procesamiento (Tiempo real).

	320 × 240	240 × 180	160 × 120	80 × 60
Pcor	55,0 ± 2,7 %	56,0 ± 2,7 %	53,0 ± 2,7 %	50,0 ± 2,7 %
Rel. error reduction		1,8 %	−3,6 %	−9,1 %
Bias fine (x:y:z) [mm]	−1 : −29 : −118	−1 : −38 : −119	−4 : −37 : −118	−16 : −34 : −111
Bias fine+gross (x,y,z) [mm]	−23 : −59 : −171	−37 : −58 : −175	−83 : −92 : −173	−190 : −113 : −184
Bias AEE fine [mm] = MOTP	246	253	253	249
Rel. AEE reduction		−2,8 %	−2,8 %	−1,2 %
Bias fine+gross [mm]	634	637	654	738
Rel. BIAS f+g reduction		−0,5 %	−3,2 %	−16,4 %
Deletion rate	0 %	0 %	0 %	0 %
Rel. Del. rate reduction		*nan* %	*nan* %	*nan* %
Loc. frames	1271	1271	1272	1272
Tiempo Real	3,17	1,75	0,84	0,27

Tabla 4.7: Resultados globales obtenidos con las base de datos AIT para distintas resoluciones

4.7. Estudio del efecto del barrido en profundidad

En esta sección se realiza un estudio de los efectos que tiene el uso de distintas resoluciones en el barrido de profundidad, en los pasos que está dividido el sistema de localización implementado y en las estadísticas obtenidas a partir de las métricas CHIL. Para ello se parte del experimento *baseline* en el que cada rayo está formado por una localización cada 100 mm realizando experimentos para los siguientes paso en profundidad de la generación de los puntos: 50 mm, 200 mm, 250 mm.

4.7.1. Resultados del proceso de cálculo de máximos

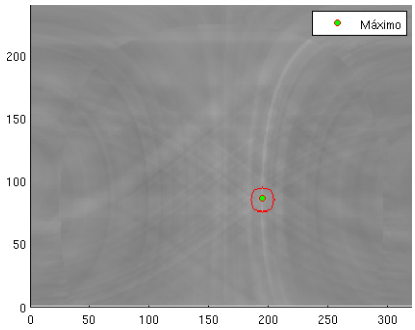
En este apartado se realizará un estudio del efecto que tiene el uso de distintas resoluciones de profundidad en el proceso de búsqueda del máximo global en las imágenes que se obtienen de cada *array* de micrófonos. En la Figura 4.17 se representa una de las imágenes resultantes para cada una de las resoluciones en el barrido de profundidad, donde no se observan diferencias significativas entre estas.

En las tablas 4.8, 4.9, 4.10, 4.11 se puede observar que los errores cometidos en elevación son menores que los de *azimuth*. Además no se observan diferencias considerables en dichos errores para las distintas resoluciones de profundidad.

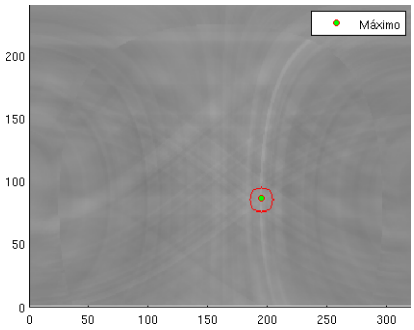
En la tabla 4.12 se representa un resumen de los errores cometidos en *azimuth* y elevación para las distintas resoluciones, donde que se promedian los errores cometidos por los tres *arrays*.

	320 × 240 Δr 50 mm			320 × 240 Δr 100 mm		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	11,44º	10,31º	11,80º	11,70º	10,33º	11,82º
error >20º	20,13 %	18,80 %	19,86 %	20,61 %	18,80 %	19,94 %

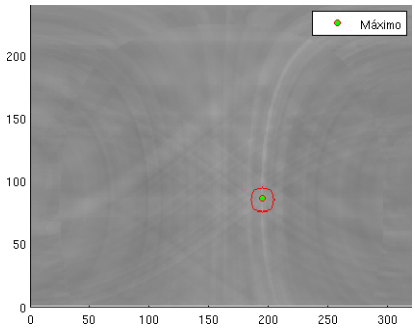
Tabla 4.8: Errores de *azimuth* cometidos en la estimación de los máximos en la base de datos de AIT para distintos barridos en profundidad.



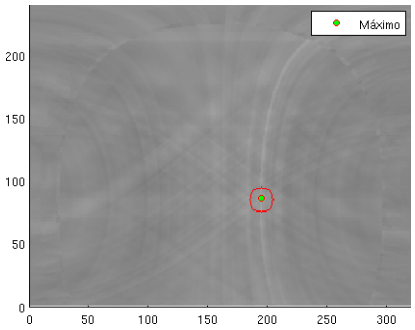
(a) Array 1, Frame 65, Resolución $320 \times 240 \times 100$



(b) Array 1, Frame 65, Resolución $320 \times 240 \times 50$



(c) Array 1, Frame 65, Resolución $320 \times 240 \times 200$



(d) Array 1, Frame 65, Resolución $320 \times 240 \times 250$

Figura 4.17: Representación de los resultados del proceso de búsqueda de máximos para distintas resoluciones en profundidad de los mapas de energía acústica partiendo de un *frame* donde el error en la estimación de este es pequeño

	320×240 Δr 200 mm			320×240 Δr 250 mm		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	11,63º	10,28º	11,75º	12,22º	10,44º	11,74º
error >20º	20,17 %	18,56 %	19,82 %	21,00 %	18,72 %	19,82 %

Tabla 4.9: Errores en *azimuth* cometidos en la estimación de los máximos en la base de datos de AIT para distintos barridos en profundidad.

	320×240 Δr 50 mm			320×240 Δr 100 mm		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	5,80º	4,61º	7,52º	5,84º	4,63º	7,56º
error >20º	7,47 %	4,84 %	8,65 %	7,44 %	4,96 %	8,73 %

Tabla 4.10: Errores de elevación cometidos en la estimación de los máximos en la base de datos de AIT para distintos barridos en profundidad.

	320 × 240 Δr 200 mm			320 × 240 Δr 250 mm		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	5,87º	4,64º	7,67º	6,20º	4,70º	7,54º
error >20º	7,43 %	4,92 %	9,16 %	7,63 %	4,92 %	8,57 %

Tabla 4.11: Errores de elevación cometidos en la estimación de los máximos en la base de datos de AIT para distintos barridos en profundidad.

Métricas	320 × 240 Δr 50 mm		320 × 240 Δr 100 mm		320 × 240 Δr 200 mm		320 × 240 Δr 250 mm	
	azimuth	elevación	*azimuth*	elevación	*azimuth*	elevación	*azimuth*	elevación
Media del error	11,18º	5,98º	11,28º	6,01º	11,22º	6,06º	11,47º	6,15º
error >20º	19,60 %	6,99 %	19,78 %	7,04 %	19,51 %	7,17 %	19,85 %	7,04 %

Tabla 4.12: Resumen de los errores cometidos para distintas resoluciones en el barrido de profundidad

4.7.2. Resultados globales

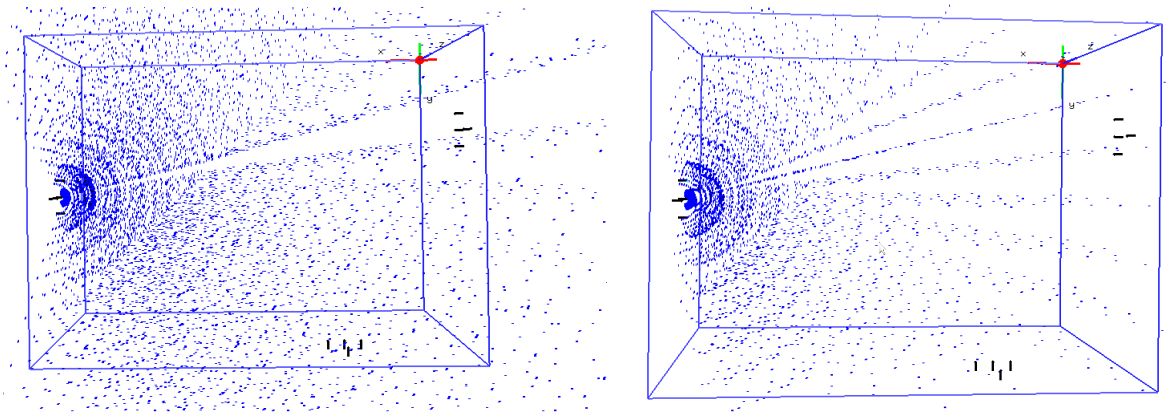
En la tabla 4.13 se pueden ver que las métricas de CLEAR reflejan resultados muy similares para cada una de las resoluciones en el barrido de profundidad.

Métricas de CHIL	320 × 240 Δr 50 mm	320 × 240 Δr 100 mm	320 × 240 Δr 200 mm	320 × 240 Δr 250 mm
Pcor	55,0 ± 2,7 %	55,0 ± 2,7 %	56,0 ± 2,7 %	55,0 ± 2,7 %
Rel. error reduction		0,0 %	1,8 %	0,0 %
Bias fine (x:y:z) [mm]	−4 : −29 : −118	−1 : −29 : −118	4 : −38 : −113	−2 : −18 : −110
Bias fine+gross (x,y,z) [mm]	−27 : −57 : −170	−23 : −59 : −171	−17 : −60 : −167	−5 : −61 : −172
Bias AEE fine [mm] = MOTP	247	246	249	251
Rel. AEE reduction		0,4 %	−0,8 %	−1,6 %
Bias fine+gross [mm]	631	634	625	639
Rel. BIAS f+g reduction		−0,5 %	1,0 %	−1,3 %
Deletion rate	0 %	0 %	0 %	0 %
Rel. Del. rate reduction		*nan* %	*nan* %	*nan* %
Loc. frames	1271	1271	1270	1271
Tiempo real	5,23	3,17	1,93	1,63

Tabla 4.13: Resultados globales obtenidos con la base de datos de AIT para distintas variaciones de los puntos en profundidad

4.8. Estudio del efecto de la estrategia de generación de rayos

Durante el proceso de desarrollo se generaron dos variantes de generación de los rayos: la primera fue la esférica (ver Figura 4.18(a))en el que se generan la misma cantidad de puntos en cada una de las direcciones y la segunda, el *baseline*, en la que los rayos terminan en la intersección de las superficies que limitan la sala donde se desarrolla la grabación, (ver Figura 4.18(b)). En este apartado se realiza un estudio de los efectos de estas dos variantes en el sistema de localización.



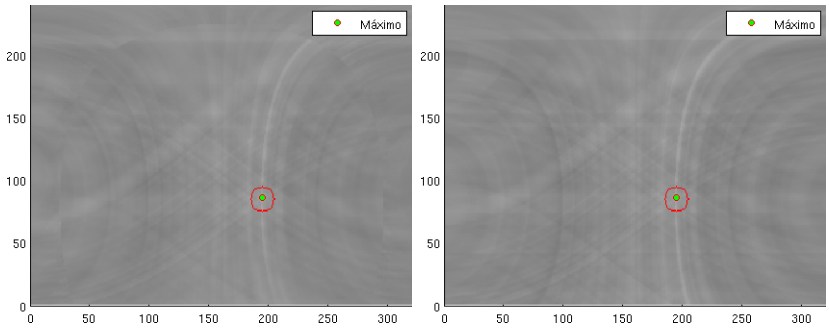
(a) Variante 1. Localizaciones esféricas (b) Variante 2. Localizaciones no esféricas

Figura 4.18: Representación gráfica de las variantes de generación de puntos.

4.8.1. Resultados del proceso de cálculo de máximos

En esta sección se muestran el efecto de las dos estrategias de generación de los rayos en el cálculo de las proyecciones de los máximos en las imágenes.

En la Figura 4.19 se muestra como la apariencia de las imágenes y la reproyección del máximo no sufre cambios significativos salvo en el tiempo de ejecución, donde la variante esférica se demora 7,20 veces el Tiempo real, frente a un 3,17en la variante no esférica.



(a) Array 1, Frame 65, Generación esférica (b) Array 1, Frame 65, Generación no esférica

Figura 4.19: Representación de los resultados del proceso de búsqueda de máximos en las dos estrategias de generación de rayos

En las tablas 4.14, 4.15 y 4.16 se puede observar que tanto la media del error, como el por ciento de *frames* donde el error es mayor de 20º son ligeramente superiores en la generación esférica.

	Generación esférica			Generación no esférica		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	12,54º	15,29º	13,12º	11,70º	10,33º	11,82º
error >20º	21,31 %	23,44 %	22,57 %	20,61 %	18,80 %	19,94 %

Tabla 4.14: Errores de *azimuth* cometidos en la estimación de los máximos en la base de datos de AIT para las dos variantes de generación de rayos implementas

	Generación esférica			Generación no esférica		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	7,18º	5,84º	9,14º	5,84º	4,63º	7,56º
error >20º	10,81 %	8,69 %	12,62 %	7,43 %	4,95 %	8,73 %

Tabla 4.15: Errores de elevación cometidos en la estimación de los máximos en la base de datos de AIT para las dos variantes de generación de rayos implementas

Métricas	Generación esférica		Generación no esférica	
	azimuth	elevación	*azimuth*	elevación
Media del error	13,65	7,39	11,28	6,01
error >20º	22,44 %	10,71 %	19,78 %	7,04 %

Tabla 4.16: Resumen de los errores cometidos utilizando generación de rayos esférica y no esférica

4.8.2. Resultados globales (con las métricas finales)

Las métricas de CLEAR reflejan resultados ligeramente mejores en el experimento *baseline* donde se usa la estrategia de generación de rayos de forma no esférica, ver tabla 4.17:

	Generación no esférica	Generación esférica
Pcor	$55{,}0 \pm 2{,}7\,\%$	$52{,}0 \pm 2{,}8\,\%$
Rel. error reduction		$-5{,}5\,\%$
Bias fine (x:y:z) [mm]	$-1 : -29 : -118$	$-16 : -47 : -142$
Bias fine+gross (x,y,z) [mm]	$-23 : -59 : -171$	$-103 : -47 : -197$
Bias AEE fine [mm] = MOTP	246	258
Rel. AEE reduction		$-4{,}9\,\%$
Bias fine+gross [mm]	634	687
Rel. BIAS f+g reduction		$-8{,}4\,\%$
Deletion rate	0 %	1 %
Rel. Del. rate reduction		$-inf\,\%$
Loc. frames	1271	1253
Ref. duration (s)	2364,0	2364,0
Tiempo real	3,17	7,20

Tabla 4.17: Resultados globales obtenidos con la base de datos de AIT con la generación de puntos de forma esférica y el *baseline*

4.9. Resultados del proceso de estimación de coherencia

La posición final se obtiene a partir de un proceso de triangulación donde un error en la proyección del máximo de energía en algunos de los *arrays* puede provocar grandes errores en el posicionamiento. Por esta razón se decidió utilizar los seminarios correspondientes a ITC, cuyo entorno está compuesto por 7 *arrays* lo que permite a través de un algoritmo de estimación de coherencia eliminar del proceso de triangulación los *array* de micrófonos que proporcionen proyecciones del máximo de potencia acústicas erróneos.

En la tabla 4.18 se muestran los resultados del proceso de localización usando y sin usar el

	Coherencia	Sin Coherencia
Pcor Rel. error reduction	$60{,}0 \pm 3{,}5\,\%$	$12{,}0 \pm 2{,}0\,\%$ $-80{,}0\,\%$
Bias fine (x:y:z) [mm]	$-151 : 5 : 24$	$-139 : 210 : 82$
Bias fine+gross (x,y,z) [mm]	$-351 : 306 : 14$	$-352 : 787 : 126$
Bias AEE fine [mm] = MOTP Rel. AEE reduction	198	348 $-75{,}8\,\%$
Bias fine+gross [mm] Rel. BIAS f+g reduction	759	1020 $-34{,}4\,\%$
Deletion rate Rel. Del. rate reduction	$25\,\%$	$0\,\%$ $100{,}0\,\%$
Loc. frames	747	993
Ref. duration (s)	1208,0	1208,0

Tabla 4.18: Resultados obtenidos eliminando los arrays que no realizan una proyección del máximo de potencia de forma coherente

proceso de estimación de coherencia. En ésta se observan las mejoras obtenidas en el proceso de estimación de coherencia en todos las métricas de CHIL excepto en que se obtiene un $25\,\%$ de *Deletion*. Esto ocurre, debido a que en dicha variante para dar una estimación, se necesita que al menos hayan 4 *arrays* de micrófonos cuyas proyecciones de máximos de potencia acústica sean coherentes.

4.10. Propuesta y resultados del sistema final seleccionado

Después de analizar los resultados obtenidos durante todo el desarrollo experimental, apoyándose en las métricas de CLEAR y en el tiempo de duración de los experimentos se decidió que el sistema final tuviera las siguientes características:

- Resolución en los barridos de *azimuth* y elevación de: 160×120. En el estudio del efecto del barrido de los ángulos respecto al experimento *baseline* los resultados son muy similares sin embargo el tiempo de duración de experimento es considerablemente menor.
- Paso en el barrido de profundidad de 250 mm. Los cambios de este parámetro afectaba muy poco los resultados.
- Generación de rayos no esférica. Los resultados obtenidos en este experimento resultaron mejores y con un tiempo de duración menor.

A continuación se muestra la tabla 4.19 con las métricas globales de la propuesta final del sistema.

4.11. Conclusiones

En este capítulo se hizo una introducción al proyecto CHIL, que en conjunto con la campaña de evaluación CLEAR, proporcionó las bases de datos con las grabaciones de audio, toda la información necesaria para ejecutar el sistema de localización y el conjunto de métricas comunes para la comunidad científica internacional que permiten evaluar los resultados del sistema de

	AIT propuesta final resolución: (160×120), Δr 250 mm
Pcor Rel. error reduction	$54{,}0 \pm 2{,}7\,\%$
Bias fine (x:y:z) [mm]	$-19 : -31 : -115$
Bias fine+gross (x,y,z) [mm]	$-74 : -93 : -179$
Bias AEE fine [mm] = MOTP Rel. AEE reduction	252
Bias fine+gross [mm] Rel. BIAS f+g reduction	646
Deletion rate Rel. Del. rate reduction	$0\,\%$
Loc. frames	1273
Ref. duration (s)	2364,0
Tiempo real	0,48

Tabla 4.19: Resultados obtenidos de la propuesta final para la base de datos de AIT

localización a partir de señales acústicas. Además se define la estrategia de evaluación llevada a cabo para evaluar los resultados obtenidos. Se describen las características de las bases de datos AIT e ITC de CHIL utilizadas en todo el proceso de experimentación y los parámetros de configuración que definen un experimento tomado como base *baselline*, a partir del cual, se hace un estudio del efecto en el cálculo de la potencia acústica por la técnica SRP-PHAT, que tiene la utilización de distintas resoluciones en amplitud y profundidad, sobre los resultados obtenidos en las etapas de: generación de imágenes, búsqueda de los máximos de energía acústica, triangulación y en las métricas de evaluación global definidas en CHIL. Se hizo un estudio sobre las mejoras obtenidas en los resultados, la eliminación de *outliers* (máximos de potencia acústica erróneos) en el proceso de estimación de coherencia. Acorde con los distintos resultados obtenidos en cada una de las etapas, se propone y evalúa un sistema final, con los parámetros de configuración que mejores resultados brindaron.

Capítulo 5

Conclusiones

5.1. Conclusiones

En esta Tesis de Máster se ha diseñado, implementado y evaluado un nuevo sistema de localización basado en información de audio a través del modelado de *arrays* de micrófonos como cámaras de perspectiva, utilizando la técnica SRP-PHAT para calcular la potencia acústica en posiciones generadas en forma de rayos que parten del centroide de los *arrays*. Esto permite obtener la potencia acústica en todas las direcciones que conforman el espacio de búsqueda y formar imágenes que contienen información de la potencia acústica y poder aplicar el algoritmo de *No Maximun Supression* para encontrar de manera robusta las direcciones de máxima energía donde se debe encontrar el locutor. Ello da paso a la aplicación del algoritmo de triangulación DLT muy usado en aplicaciones de localización de objetos a partir de imágenes obtenidas por varias cámaras. El sistema implementado fue evaluado a partir de un experimento base con las bases de datos de AIT de CHIL con las métricas de CLEAR. Se evaluó el efecto de la resolución con la que se hace el escaneo en ángulos del espacio de localización, comprobando la poca influencia que tuvo en los resultados la resolución en distancia con las que se generaron los rayos. A partir de los resultados obtenidos se demuestra la influencia que tiene sobre la precisión del sistema los errores cometidos en la búsqueda de las direcciones de máxima potencia acústica, para lo cual se implementó un sistema de estimación de coherencia que permite eliminar del proceso de triangulación los arrays de micrófonos que brindan medidas erróneas. Para evaluar este proceso se utilizó la base de datos ITC de CHIL de la cual se tienen grabaciones procedentes de 7 *arrays* de micrófonos lográndose mejoras considerables en la precisión final del sistema. Como resultados del proceso de experimentación se propuso un sistema final para la base de datos de AIT.

5.2. Líneas futuras

Se plantean las siguientes líneas de trabajo futuro:

- Analizar la utilización de filtros a la señal de audio recibida por los micrófonos, con el objetivo de mejorar la la respuesta del algoritmo SRP-PHAT.
- Analizar la respuesta de SRP-PHAT, utilizando distintos valores de tamaño de ventana de análisis, tiempo en el cual se considera que el locutor no cambia su posición.
- Trabajar en nuevas estrategias de definición del espacio de búsqueda, donde las pérdidas de resolución a distancias mayores del *array* de micrófonos sean menores.
- Analizar la distribución de los niveles en los mapas de energía, con el fin de obtener imágenes en las que se represente mejor la información, que permitan aplicar otras técnicas más sofisticadas de tratamientos de imágenes, para la búsqueda de los máximos de energía de manera más robusta.
- Experimentar con otras técnicas más elaboradas de algoritmos de detección de coherencia a los máximos encontrados, con el objetivo de eliminar los máximos erróneos del sistema de triangulación.
- Estudiar la influencia que tiene sobre los errores, la posición del locutor en el espacio.
- Utilizar técnicas de seguimiento *(tracking)*, con el objetivo de hacer un filtrado espacio temporal de los resultados.
- Realizar la evaluación del sistema propuesto en nuevas bases.

Parte IV

Apéndices

Como hacer un Experimento

En este apartado se explica cómo se generan las imágenes basadas en potencia acústica para una determinada base de datos. Todo el proceso está basado en la definición de un conjunto de parámetros que controlan las variaciones soportadas por el sistema de localización y del uso de una serie de *scripts* de propósito general que facilitan la generación y ejecución de los experimentos para una determinada base de datos.

Generación y configuración del experimento

Para la generación del experimento primero se realizan una serie de pasos para configurar su comportamiento general los cuales se enumerarán a continuación:

1. Dentro del directorio de trabajo "`far-field`" ir al directorio autoGenExp y en el fichero de configuración genExp.cfg y chequear que:
   - El directorio de trabajo (HOME_DIR) coincida con el de su sistema.
   - El directorio donde se encuentra la librería que genera las imágenes
     SRP_BIN=$Home_DIR/generateVisualAudio/genVA
2. Crear un fichero de configuración el directorio ..autoGenExp/exp_cfg_files en el que se definan las siguientes variables:
   - EXP_ID la cual posee el nombre del experimento, se sugiere que el formato sea <DATABASE>-<EXPERIMENT>.cfg
   - OBJETIVE a la cual se le asigna un comentario que brinde información del objetivo del experimeto a realizar.
   - DB_ID se le asigna el nombre de la base de datos con la que se va a trabajar. ver el fichero supported_db_ids.list para ver los posibles nombres correctos de base de datos disponibles.
3. Se sugiere comprobar que existen los siguientes ficheros:
   - ..autonGenExp/lists/DB_ID.list que posee el listado de los ficheros de audio de la base de datos.
   - ..autoGenExp/simFiles/<DATABASE>.sim que posee los parámetros de configuración del experimento.
4. Ejecute genExp.sh seguido del fichero .cfg creado en el punto anterior.
5. Si todo fue correcto debe de haberse creado un directorio con el mismo nombre del identificador EXP_ID proporcionado en el fichero de configuración. El directorio debe de contener los siguientes ficheros:
   - db-id.cfg. El cual define la variable DB_ID con el nombre de la base de datos.
   - genExp.cfg
   - go-EVAL.cfg
   - go-EVAL.sh
   - go-GENDBLIS.cfg

- go-GENDBLIS.sh
- go-SRP.cfg
- go-SRP.sh
- info.cfg
- DB_ID.list
- DB_ID.sim
- DB_ID-subarrays.list
- README
- README.autoGenExp
- .cvsignore

Tareas y configuraciones que dependen de la base de datos

1. En el directorio del experimento abrir el fichero go-GENDBLIS.cfg y verificar el valor de las variables que dependen de la base de datos:
   - DB_HOME_DIR. Directorio de la base de datos a procesar.
   - SIZECOMMONFILENAME. Debe tener el número de caracteres en común que poseen los ficheros de audio.
   - NUMFILESPERUTT. El numero de ficheros de audio por uterancia.
   - NUM_MICS. Proporciona el número de micrófonos por cada subarray de micros envueltos en la simulación.
   - CHANNELS. Si se utilizan los ficheros de audio multicanales se debe de especificar que canales serán utilizados, por ejemplo CHANNELS = (1 2 3 4). Si solamente se utiliza un fichero de un solo canal: CHANNELS =().
2. Ejecutar go-GENDBLIS.sh y verificar que se crea el fichero audiosource con el siguiente formato <audioSource>-<EXP_ID>. Este fichero contiene el path de cada uno de los ficheros de audio a analizar.

Generación de las localizaciones

Para la generación de las localizaciones es necesario mirar algunas variables que se encuentran el fichero de configuración del experimento, <DB_ID.sim>. Cada base de datos tiene un fichero de simulación por defecto el cual debe ser retocado para adaptarlo al experimento en particular. En esta sección se comentará el significado de cada uno de estos parámetros y se explicará como generar las localizaciones.

1. Abrir el fichero de simulación <DB_ID.sim>e ir a la sección [SimulationConfigInfo] y verificar que la variable fileEnvironment tiene el nombre del fichero de configuración del entorno donde se han hecho las grabaciones de la base de datos.
2. En la sección [Directories] se configurarán los directorios donde se encuentran todos los elementos utilizados en el experimento el cual está compuesto por las siguientes variables:
   - dirEnvironments. Define el directorio donde se encuentra el fichero de configuración del entorno donde se hicieron las grabaciones en la base de datos.

- dirMicArrays. Define el directorio donde se encuentran los ficheros que determinan la distribución de cada uno de los micrófonos en el entorno.
- dirSrcLocationResults. Define el directorio donde se almacenarán los resultados del experimento así como donde se ubicarán los ficheros que contendrán las localizaciones.
- dirSrcLocationTruth. Define el directorio donde se encuentra almacenado el fichero de etiquetado del grountruth.

3. En la sección [MicArrays] se definen las variables:

   - numMicArrays. Define la cantidad de arrays de micrófonos del entorno.
   - Habrá la misma cantidad de micarray[i] como numMicArrays y este define el número de subarrays, el número de micrófonos por subarray y el índice de micrófonos envueltos en cada subarray. Este índice está relacionado con la posición de cada micrófono en el fichero .arr.

4. En la sección [SrcLocationTruth]. Se definen las variables que tienen que ver con las características y localización del fichero de *groundtruth* estas variables son:

   - numSrcLocationTruth. Define el número de fuentes de *groundtruth.*
   - frameShift. Define la variación del tiempo del etiquetado de *groundtruth.*
   - srcLocationTruth0. Define el nombre del fichero fuente del etiquetado de **groundtruth.**
   - resetStartTime. Si se define en 1. El etiquetado de tiempo en el fichero de *groundtruth* es inicializado en 0.

5. En la sección [VisualAudioSearchSpace]. Se definen los variables utilizados para la generación de las localizaciones las cuales son:

   - numSearchSpaceFiles. Define la cantidad de ficheros de localizaciones el cual es igual a la cantidad de subArray.
   - searchSpaceFile(0 hasta numSearchSpaceFiles - 1). Define el nombre de cada uno de los ficheros de localizaciones.

6. Ejecutar el programa que genera las localizaciones donde va a ser evaluado el algoritmo de srp para generar las imágenes. Para ello es necesario ir al directorio de trabajo y dentro de este a ..generateSubArraySearchSpace/ y dentro de este ejecutar ./generateSubArraySearchSpace -v <pathname completo del fichero de simulación>

7. Si todo fue bien, verificar que se generaron las localizaciones en el directorio dirSrcLocationResults definido anteriormente. También se puede utilizar la librería visualSimulation que permite observar la ubicación de las localizaciones generadas en el entorno.

8. Se pueden utilizar dos variantes en las localizaciones para más detalle ir a...

Tareas que dependen de la configuración del experimento

1. En el fichero de configuración go-SRP.cfg es necesario confirmar una serie de parámetros que necesita el algoritmo genVA.c. A continuación se caracterizan los parámetros más importantes:

- Fs. Es la frecuencia de muestreo que será aplicada en caso de que el fichero de audio sea un .raw. En caso contrario en programa leerá el fichero de audio y considerará la frecuencia de muestreo con la que fue grabada.
- FFT_SIZE: Es el tamaño de la FFT y siempre debe de ser mayor o igual que el tamaño de la trama. Además para lograr la mayor eficiencia esta debe de ser potencia de 2.
- DIR_INPUT_FILES: Especifica el directorio donde el programa puede encontrar los ficheros de entrada requeridos (audiosource y ficheros .sim).
- INI_AUDIO_FILE, END_AUDIO_FILE :determina el intervalo de tiempo en segundos de la ventana del fichero de audio que se analizará. Estos valores deben de ser positivos y END_AUDIO_FILE debe de ser mayor que INI_AUDIO_FILE.
- ROUND_FLAG: Puede tomar los siguientes valores: 0(no se realiza redondeo), 1(se redondea al entero más cercano), o mayor que 1 (el redondeo se hace acorde a la interpolación lineal).
- LOW_FREQ: Este valor no es usado en esta variante.
- FILTER_FLAG: Debe de estar puesto en 0, para que no se realice el filtrado.
- CHANNELS: Especifica los canales que se usarán en caso de que se esté en un fichero de audio multicanal. El formato es CHANNELS="6,7,8,....º CHANNELS=.ªll"si los ficheros de audio tienen un solo canal o se analizará cada canal por separado.

2. Ejecutar bash go-SRP.sh. Se debe de crear un fichero .log con la información relacionada a la ejecución del programa y los posibles errores que pudieron ocurrir su formato es: <date&time-EXP_ID>.log. También se debió de crear y/o actualizar un fichero con el formato <EXP_ID>.run con la información de la línea de comando utilizada (útil para ver los parámetros de ejecución del algoritmo) para ejecutar el algoritmo.

3. Si todo fue bien se deben de crear para cada *frame* en todos los subarray dos ficheros con el formato Image_<EXP_ID>_subarray[subarray]_frame[frame].dat (que contiene el promedio de potencia acústica calculado para cada dirección) y Image_<EXP_ID>_subarray[subarray]_frame[frame] (la imagen de potencia acústica).

.

Ejemplo

En esta sección se ejecutará un experimento real donde se crean las imágenes basadas en potencia acústica. Para el experimento se utilizó la base de datos ITC de Chil.

Generación y configuración del experimento

Generar el directorio del experimento con los correspondientes ficheros

1. Dentro del directorio de trabajo far-field ir al directorio autoGenExp y en el fichero de configuración genExp.cfg chequear los valores de las siguientes variables:

   - HOME_DIR="/home/alegra/reposito/proyecto/far.field/"
   - SRP_BIN="$HOME_DIR/generateVisualAudio/genVA"

2. Crear un fichero de configuración sampleExp.cfg y guardarlo dento del directorio $HOME_DIR/autoGenExp/exp_cfg_files/sampleExp.cfg con las siguientes variables:
   - EXP_ID="ITC-Chil07-Prueba"
   - OBJETIVE="Test generate experiment with ITC-DEV-2007"
   - DB_ID="ITC-DEV-2007"
3. Chequear que el fichero $HOME_DIR/autoGenExp/supported_db_ids.list incluye:ITC-DEV-2007.list
4. Chequear que existe el fichero $HOME_DIR/autoGenExp/simFiles/ITC-DEV-2007.sim
5. Ejecutar en una terminal de comando genExp.sh seguido del fichero de configuración:
   - bash genExp.sh sampleExp.cfg
6. Ahora en el direcrorio experiments se debe de haber creado el directorio del experimento con el mismo nombre de EXP_ID ($HOME_DIR/experiments/ITC-Chil07-Prueba/) con los siguientes ficheros:
   - db-id.cfg. Cuyo contenido es: DB_ID="ITC-DEV-2007"
   - genExp.cfg
   - go-EVAL.cfg
   - go-EVAL.sh
   - go-GENDBLIS.cfg
   - go-GENDBLIS.sh
   - go-SRP.cfg
   - go-SRP.sh
   - info.cfg
   - ITC-DEV-2007.list
   - ITC-DEV-2007.sim
   - ITC-DEV-2007-subarrays.list
   - README
   - README.autoGenExp
   - .cvsignore

Tareas que dependen de la configuración de la base de datos

1. Abrir el fichero go-GENDBLIS.cfg y chequear las siguientes variables:
   - El directorio donde se encuentra la base de datos en este caso: DB_HOME_DIR="/usr/share/geintra/databases/mmodal/CHIL_D7_14/CHIL_D7_14/DATA/DEV/SE
   - La cantidad de caracteres que tienen en común los ficheros de audio de la base de datos en cuestión: SIZECOMMONFILENAME=15
   - La cantidad de ficheros a procesar por uterancia: NUMFILEPERUTT=28
   - La cantidad de micrófonos por cada subarray de micrófonos: NUM_MICS=4
   - CHANNELS=(1 2 3 4)
2. Ejecutar en la línea de comando el scripts: $bash go-GENDBLIS.sh si todo fue bien se debió de haber creado el fichero audioSource-ITC-Chil07-Prueba.

Generación de las localizaciones

1. Abrir el fichero de simulación: ITC-DEV-2007.sim y verificar que el contenido es:

[**SimulationConfigInfo**]
content = Visual Simulation Definition file
version = 1.1
comment = ISL 2004 room (described in CHIL document D7_6.pdf (CHIL_D7_6/DOCUMENTATION/D7_6
fileEnvironment = environment.env

[**Directories**]
dirEnvironments = environments/ITCChilRoom/
dirMicArrays = environments/ITCChilRoom/
dirSrcArrays = environments/ITCChilRoom/
dirGeometricFiles =
dirSrcLocationResults = experiments/ITC-Chil07-Prueba/
dirSrcLocationTruth = /usr/share/geintra/databases/mmodal/CHIL_D7_14/CHIL_D7_14/DEV/SEMINA
dirPowerPlots = /media/H/CHIL_D7_14/CHIL_D7_14/DATA/TEST/SEMINARS/video_labels/

[**MicArrays**]
numMicArrays = 1
micArray0 = allMics.subarr

[**SrcArrays**]
numSrcArrays = 0
[**GeometricFiles**]
numGeometricFiles = 0
[**SrcLocationTruth**]
numSrcLocationTruth = 0
frameShift = 1.00
fileFormat = CHIL_PERSON_TRACKING_REFERENCE2007
srcLocationTruth0 = ITC_20060714_Acoustic3d_label.txt
resetStartTime = 1

[**SrcLocationResults**]
numSrcLocationResults = 0
fileFormat = CHIL_PERSON_TRACKING_REFERENCE2007
histog_min = -0.8
histog_max = 0.8
resetStartTime = 1
[**SearchSpace**]

fileSearchSpace = searchSpace50_trialed.txt

[**VisualAudioSearchSpace**]
stepsAzimElevRange = 0.1 0.1 100
numSearchSpaceFiles = 7
searchSpaceFile0 = locations_subArray0.sim
searchSpaceFile1 = locations_subArray1.sim
searchSpaceFile2 = locations_subArray2.sim
searchSpaceFile3 = locations_subArray3.sim
searchSpaceFile4 = locations_subArray4.sim
searchSpaceFile5 = locations_subArray5.sim
searchSpaceFile6 = locations_subArray6.sim

3. Generar las localizaciones para ello ir al directorio $HOME_DIR/generateSubarraySearcSpace y ejecutar el scrpits:
   - ./generateSearcSpace -v ../experiments/ITC-Chil07-Prueba/ITC-DEV-2007.sim
4. Verificar que se generaron los ficheros de localizaciones:
   - locations_subarray0.sim
   - locations_subarray1.sim
   - locations_subarray2.sim
   - locations_subarray3.sim
   - locations_subarray4.sim
   - locations_subarray5.sim
   - locations_subarray6.sim
5. Utilizar la librería de visualización del entorno para comprobar si la generación de las localizaciones se generaron correctamente. Para ello ir al directorio ../VisualSimulation y ejecutar el scripts:./visualSimulation -v ../experiments/ITC-Chil07-Prueba/ITC-DEV-2007.sim
6. Una vez que salga la pantalla con el entorno apretar la tecla “q” y se deben de dibujar las localizaciones generadas correspondientes al fichero locations_subarray5.sim (subArray5) como se muestra en la Figura 5.1. Para ir mirando las demás apretar la tecla “1”.

Tareas que dependen de la configuración del experimento

- Abrir el fichero go-SRP.cfg para configurar los parámetros necesarios para ejecutar el algoritmo genVA. El fichero go-SRP.cfg debe de quedar como:
- Correr el experimento para ello ejecutar el scripts: bash go-SRP.sh. Si todo fue bien se deben de crear para cada *frame* en todos los subarray dos ficheros con el formato Image_ITC_Chil07_Prueba_subarray[0-6subarray]_frame[0-...].dat (que contiene el promedio de potencia acústica calculado para cada dirección) y Image_ITC_Chil07_Prueba_subarray[0-6subarray]_frame[0-...].jpg (la imagen de potencia acústica).

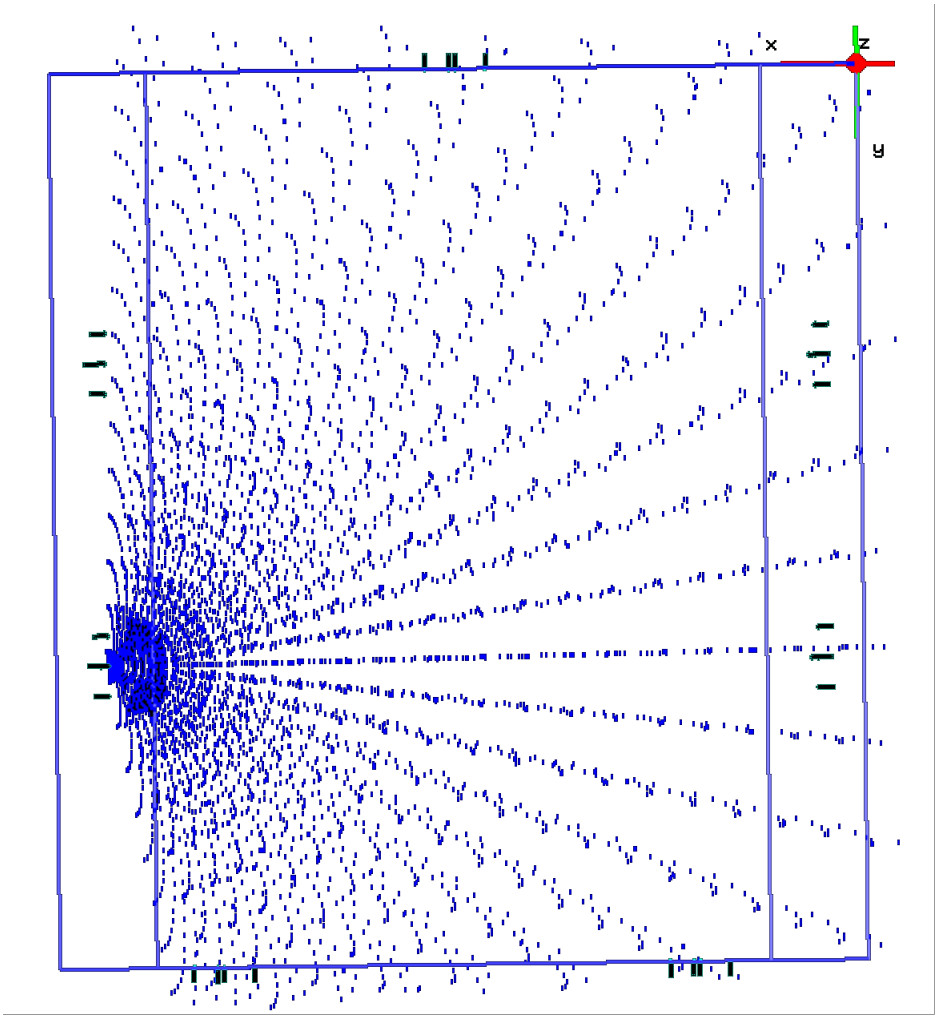


Figura 5.1: Localizaciones relativas al subarray5 del entorno de ITC-DEV-2007

Parte V

Bibliografía

Bibliografía

[1] "Augmented multi-party interacion (ami) project. state of the art overview: Localization and tracking of multiple interlocutors with multiple sensors," Technical report, Tech. Rep., 2007.

[2] M. L. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Carnegie Mellon University, 2003.

[3] W. Herbordt, *Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing.* Springer, Heidelberg, Germany, March, 2005.

[4] D. Gelbart and N. Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," *In International Conference on Spoken Language Processing (ICSLP)*, 2002.

[5] S. Kochkin and T. Wickstrom, "Headsets, far field and handheld microphones: Their impact on continuous speech recognition," Technical report, EMKAY, a division of Knowles Electronics, Tech. Rep., 2002.

[6] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," *In proceedings of the ACM International Conference on Multimedia*, pp. 1061–1070, 2008.

[7] E. Munnoz Herraiz, "Disenno, implementacion y evaluacion de tecnicas de localizacion de fuente y de mejora de la sennal de habla en entornos acusticos reverberantes: aplicacion a sistemas de reconocimiento automatico de habla," Master's thesis, Universidad Politecnica de Madrid, Spain, 2005.

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

[9] M. C. Aguilar, "Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos," Tech. Rep., 2007.

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

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision. Second Edition.* Cambridge University Press 2000, 2003, 2003.

[12] M. Weiser, "The computer for the 21st century," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 3, pp. 3–11, 1999.

[13] ——, "Parc builds a world saturated with computation," *Science (AAAS)*, 1993.

[14] M. Coen, "Design principles for intelligent environments," *In Proceedings of the National Conference on Artificial Intelligence*, pp. 547–554, 1998.

[15] G.Look, B.Kottahachchi, R. Laddaga, and H. Shrobe, "A location representation for generating descriptive walking directions," *International Conference on Intelligent User Interfaces*, pp. 122–129, 2005.

[16] G.Look and H. Shrobe, "Towards intelligent mapping applications: a study of elements found in cognitive maps," *International Conference on Intelligent User Interfaces*, pp. 309–312, 2007.

[17] A. Pentland, "Smart rooms," *Scientific American*, vol. 274, pp. 54–62, 2007.

[18] K. Ara, N. Kanehira, D. Olguín, B. N. Waber, T. Kim, A. Mohan, P. Gloor, R. Laubacher, D. Oster, A. Pentland, and K. Yano, "Sensible organizations: Changing our businesses and work styles through sensor data," *Information and Media Technologies*, vol. 3, pp. 604–615, 2008.

[19] S. Shafer, J. Krumm, B. Brumitt, B. Meyers, M. Czerwinski, and D. Robbins, "The new easyliving project at microsoft research," *Proceedings of the 1998 DARPA / NIST Smart Spaces Workshop*, pp. 127–130, July 1998.

[20] T. Sogo, H. Ishiguro, and T. Ishida, "Acquisition of qualitative spatial representation by visual observation," *Proceedings of the 16th international joint conference on Artificial intelligence*, vol. 16, pp. 1054–1060, 1999.

[21] H. Hashimoto, J. Lee, and N. Ando, "Self-identification of distributed intelligent networked device in intelligent space," *Proceedings in IEEE International Conference on Robotics and Automation, 2003*, vol. 3, pp. 4172–4177, Nov. 2003.

[22] P. Steinhaus, M. Ehrenmann, and R. Dillmann, "Mephisto: A modular and extensible path planning system using observation," *Lecture notes in computer science*, pp. 361–375, 1999.

[23] P. Steinhaus, M. Strand, and R. Dillmann, "Autonomous robot navigation in human-centered environments based on 3d data fusion," *EURASIP Journal on Applied Signal Processing*, vol. 2007, pp. 224–224, Jan. 2007.

[24] J. Villadangos, J. Urena, M. Mazo, A. Hernandeza, F. Alvarez, J. García, C. Marziani, and D. Alonso, "Improvement of ultrasonic beacon based local position system using multi-access techniques," *Proceedings of IEEE International Symposium on Intelligent Signal Processing*, 2005.

[25] D. Pizarro, E. Santiso, and M. Mazo, "Simultaneous localization and structure reconstruction of mobile robots with external cameras," *International Symposium on Industrial Electronics ISIE05*, June 2005.

[26] I. Fernandez, M. Mazo, J. Lazaro, D. Pizarro, E. Santiso, P. Martin, and C. Losada, "Guidance of a mobile robot using an array of static cameras located in the environment," *Autonomous Robots*, vol. 23, pp. 305–324, Nov. 2007.

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

[28] A. Hoover and B. D. Olsen., "A real-time occupancy map from multiple video streams," *Proceedings in IEE International Conference on Robotics and Automation, 1999*, vol. 3, pp. 2261–2266, May 1999.

[29] E. Kruse and F. Wahl, "Camera-based observation of obstacle motions to derive statistical data for mobile robot motion planning," *Proceedings in IEE International Conference on Robotics and Automation, 1998*, vol. 1, pp. 662–667, May 1998.

[30] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, pp. 1497–1498, Oct. 1973.

[31] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.

[32] G. Bienvenu, "Eigensystem properties of the sampled space correlation matrix," *Proceedings of the IEEE International, Conference on Acoustics, Speech and Signal Processing*, vol. 8, pp. 332–335, Apr. 1983.

[33] M. Wax, T.-J. Shan, and T. Kailath, "Spatio-temporal spectral analysis by eigenstructure methods," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 817–827, Aug. 1984.

[34] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 823–831, Aug. 1985.

[35] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 1661–1669, Aug. 1987.

[36] J.-S. Hu, C.-C. Cheng, W.-H. Liu, and T. M. Su, "A speaker tracking system with distance estimation using microphone array," *Proceedings of the IEEE/ASME International Conference on Advanced Manufacturing Technologies and Education*, vol. 35, pp. 485–494, Aug. 2002.

[37] J.-S. Hu, T. Su, C.-C. Cheng, W.-H. Liu, and T. I.Wu, "Aself-calibrated speaker tracking system using both audio and video data," *Proceedings of the IEEE Conference on Control Applications*, vol. 2, pp. 731–735, Sept. 2002.

[38] I. A. McCowan, "Robust speech recognition using microphone arrays," Ph.D. dissertation, Queensland University of Technology, Australia, 2001.

[39] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 273–276, Aug. 1994.

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

[41] E. Muñoz Herraiz, "Design, implementation and evaluation of source localization and speech signal improvement techniques in acoustic reverberant environments: Application to automatic speech recognition system," Master's thesis, Technical University of Madrid, Madrid, 2006.

[42] B. D. Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Transactions on Signals Processing*, vol. 41, pp. 2826–2838, Sept. 1993.

[43] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Mar. 1986.

[44] R. Roy and T. Dailath, "Esprit- estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 984–995, July 1989.

[45] R. Kumaresan and D. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Transactions on Aerospace and Electronics Systems*, vol. AES-19, pp. 134–139, July 1983.

[46] M. Viberg, B. Ottersten, and T. Kailath, "Detection and estimation in sensor arrays using weighted subspace fitting," *IEEE Transactions on Signal Processing*, vol. 39, pp. 2436–2449, Nov. 1991.

[47] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Journal of the Acoustic Society of America*, vol. 107, pp. 384–391, Jan. 2000.

[48] H. Hung and M. Kaveh, "Focussing matrices for coherent signal-subspace processing," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1272–1281, 1988.

[49] B. Friedlander and A. Weiss, "Focussing matrices for coherent signal-subspace processing," *IEEE Transactions on Signal Processing*, vol. 41, pp. 1618–1634, 1993.

[50] M. Brandstein and S. Griebel, "Explicit speech modeling for microphone array applications," *Microphone Arrays: Signal Processing Techniques and Applications, Springer Verlag*, 2001.

[51] M. W. T. Dailath, "Optimum localization of multiple sources by passive arrays," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 31, pp. 1210–1217, Oct. 1983.

[52] K. Varma, "Time-delay-estimate based direction-of-arrival estimation for speech in reverberant environments," Ph.D. dissertation, Virginia Polytechnic Institute, 2002.

[53] R. Duraiswami and D. Zotkin, "Active speech source localization by a dual coarse-to-fine search," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3309–3312, May 2001.

[54] 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, pp. 520–529, 2004.

[55] P.-J. Chung, J. F. Bsohme, and A. O. Hero, "Tracking of multiple moving sources using recursive em algorithm," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 50–60, 2005.

[56] R. Duraiswami and J. Neumann, "Microphone arrays as generalized cameras for integrated audio visual processing," *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, July 2007.

[57] A. Redondi, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Geometric calibration of distributed microphone arrays," *Proceedings IEEE International Workshop on Multimedia Signal Processing 2009*, pp. 1–5, Oct. 2009.

[58] L. Kitchen and A. Rosenfeld, "Gray-level corner detection," *Pattern Recognition Letters*, vol. 1, no. 2, pp. 95–102, 1982.

[59] J. Knight, A. Davison, and I. Reid., "Towards constant time slam using postponement," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2001, pp. 406–412.

[60] P. A. Beardsley, A. Zisserman, and D. W. Murray, "Navigation using affine structure from motion," *Proceedings of the third European conference on Computer Vision.*

[61] O. D. Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig?" *Proceedings of the third European conference on Computer Vision.*

[62] D. Hawkins, *Identification of outliers.* Chapman and Hall, 1980.

[63] V. Barner and T. Lewis, *Identification of outliers.* Wiley and Sons, New York, 1994.

[64] P. Rousseeuw and A. Leroy, *Robust Regression and Outliers Detection.* Wiley and Sons, New York, 1987.

[65] K. Yamanishi, J. Takeuchi, G. Williams, and P. Milne, "On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms," *Data Mining and Knowledge Discovery*, vol. 8, pp. 275–300, May 2004.

[66] K. Yamanishi and J. Takeuchi, "Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner," *International Conference on Knowledge Discovery and Data Mining*, pp. 389–394, 2001.

[67] I. Ruts and P. J. Rousseeuw, "Computing depth contours of bivariate point clouds," *Computational Statistics and Data Analysis*, vol. 23, pp. 275–300, Nov. 1996.

[68] T. Johnson, I. Kwok, and R. Ng, "Fast computation of 2-dimensional depth contours," 1998.

[69] E. M. Knorr and R. T. Ng, "Finding intensional knowledge of distance-based outliers," *Proceedings of the 25th International Conference on Very Large Data Bases*, pp. 211–222, 1999.

[70] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: Algorithms and applications," 2000.

[71] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," vol. 29, pp. 427–438, 2000.

[72] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 15–26, 2002.

[73] S. D. Bay and M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule," *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 29–38, 2003.

[74] Z. He, X. Xu, and S. Deng, "A fast grredy algorithm for outlier mining," 2007.

[75] M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," 2000.

[76] J. Tang, Z. Chen, A. Fu, and D. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," *Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 535–548, 2002.

[77] M. F. Jaing, S. S. Tseng, and C. M. Su, "Two-phase clustering process for outliers detection," *Pattern Recognition Letters*, vol. 22, pp. 691–700, May 2001.

[78] D. Yu, G. Sheikholeslami, and A. Zhang, "Finding outliers in very large datasets," *Computer Science*, vol. 4, pp. 387–412, 2002.

[79] C. Aggarwal and P. Yu, "Outlier detection for high dimensional data," *ACM SIGMOD Record*, vol. 30, pp. 37–46, June 2001.

[80] D. M. J. Tax, A. Ypma, and R. P. W. Duin, "Support vector data description applied to machine vibration analysis," 1999.

[81] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, pp. 1443–1471, 2001.

[82] R. Jalvo, "Reconstrucción volumétrica exacta a partir de múltiples cámaras y su aplicación a los espacios inteligentes," Master's thesis, Universidad de Alcalá, Alcalá de Henares, 2007.

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

[84] R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007.* Springer, 2008.

[85] M. M. Romera, "Seguimiento de múlltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos," Ph.D. dissertation, Universidad de Alcalá, 2009.

[86] R. Hartley, R. Gupta, and T. Chang, "Stereo from uncalibrated cameras," *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 761–764, 1992.

[87] K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time part i: Theory and algorithms," *IJCV*, vol. 62, pp. 221–247, May 2005.

[88] E. M. Knorr and R. T. Ng, "A unified notion of outliers: Properties and computation," pp. 219–222, 1997.

[89] Amida, *Augmented Multi-party Interaction with Distance Access. Localization and Tracking of Multiple Interlocutors with Multiple Sensors.* January, 2007.

[90] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 273–276, Apr. 1994.

[91] S. of the art overview: Localization and tracking of multiple interlocutors with multiple sensors, "Augmented multi-party interacion (ami) project." Technical report, Tech. Rep., 2007.