UNIVERSIDAD DE ALCALÁ

ESCUELA POLITÉCNICA SUPERIOR

DEPARTAMENTO DE ELECTRÓNICA

TRABAJO DE INVESTIGACIÓN TUTELADO

LOCALIZACIÓN, SEGUIMIENTO Y POSE DE MÚLTIPLES INTERLOCUTORES UTILIZANDO FUSIÓN AUDIOVISUAL

Autor: Diego Alonso Jiménez

Directores: Dr. Manuel Mazo Quintas

Dr. Javier Macías Guarasa

Índice

Índice de figuras

Índice de tablas

1. Introducción

El desarrollo de este trabajo se enmarca dentro de la realización de una tesis doctoral en el campo de los "Espacios Inteligentes" y constituye por lo tanto un paso previo y fundamental para desarrollo de la misma.

Los "espacios inteligentes" es un término acuñado recientemente, y que se enmarca dentro de una tendencia de investigación más amplia que es la "computación ubicua", termino también de reciente creación. La "computación ubicua" plantea un espacio dotado de un conjunto de sistemas sensoriales, de comunicación, y de cómputo inteligente, que son transparentes e imperceptibles para el usuario, pero que están continuamente percibiendo el entorno y cooperando entre ellos para proporcionar la ayuda necesaria a cada persona. Este tipo de entornos se definen como "ambientes inteligentes" o "espacios inteligentes", y requieren ante todo de una apropiada red o conjunto de "redes de sensores" que sean capaces de obtener información relevante del entorno, y a partir de estos datos planificar el comportamiento que se debe desplegar en función de las demandas de los usuarios y de las capacidades de actuación con las que se cuente.

En una primera aproximación (en el apartado de objetivos se enumeran con mayor detalle los objetivos de la tesis), el objetivo que se persigue con la tesis que se pretende desarrollar es la integración de información procedente de sensores acústicos (micrófonos) y de visión (cámaras) ubicadas en el entorno de interés con el objetivo general de obtener la pose de los usuarios (y otro tipo de agentes) que se muevan en dicho entorno.

1.1 Antecedentes

En la literatura científica existen una gran cantidad de trabajos que emplean la información procedente de un único tipo de sensores con el objetivo de localizar

determinados targets[1] dentro de la escena bajo análisis, como por ejemplo videocámaras [Okuma 2004] [Lanz 2006], arrays de micrófonos [Lathoud 2005] [Castro 2007], infrarrojos…etc. Dado que cada tipo de sensores posee sus propias fortalezas y debilidades, la comunidad científica se ha planteado recientemente la utilización combinada de varios de ellos como una solución que aumente la robustez de los algoritmos de localización y tracking implementados.

Cuando se trata con aplicaciones del tipo videoconferencias "inteligentes", interfaz hombre máquina, análisis automático de escenas, vigilancia, apuntamiento automático de cámaras…etc, en las cuales se ven implicadas fuentes sonoras, la utilización de esta información puede ser muy interesante a la hora de realizar la detección, localización y tracking[2] de determinados targets dentro de la misma. Si además se tiene en cuenta que el habla es la forma natural del ser humano de comunicarse, es inmediata su utilización a la hora de facilitar la interacción hombre-máquina y por lo tanto a su inclusión dentro de los "espacios inteligentes".

Por otro lado, la localización empleando técnicas acústicas sólo puede utilizarse de una manera fiable cuando dichas señales son capturadas por arrays de micrófonos con unas características determinadas, siendo además su estimación típicamente poco precisa. Cuando nos enfrentamos con señales con una baja relación señal a ruido o generadas en entornos de alta reverberación, los sensores acústicos no son, por si mismos, capaces de generar una estimación fiable de la posición.

Por este motivo se considera básica introducir sensores adicionales, como por ejemplo las cámaras de visión, que proporcionen una estimación adicional de la pose (posición y orientación).

Los sensores de visión constituyen una alternativa fundamental en tareas como la estudiada en este trabajo, principalmente por la gran cantidad de información que

[1] El término inglés *"target"* significa, en general, objetivo, y en la literatura científica del área de interés se refiere específicamente al objeto u objetivo de seguimiento. Debido a su amplio uso en referencias bibliográficas del área, tanto en español como en inglés, este término como sinónimo del español.

[2] Dado que la noción de seguimiento corresponde con el término inglés *"track"* en la literatura científica de la rama, incluida la escrita en lengua española, se utilizarán los términos español e inglés por igual.

proporcionan, lo cual les otorga un potencial de precisión y robustez superior al de los sensores de audio, si bien la complejidad de los algoritmos también es elevada.

1.2 Objetivos

La finalidad general y última perseguida en la tesis que se pretende desarrollar a partir de este trabajo, es la de obtener una solución que permita detectar, localizar, seguir, determinar la pose y el estado (habla/silencio), de personas ubicadas dentro de un entorno complejo, dinámico y poblado por un número múltiple y variable de targets. A este tipo de aplicaciones se las conoce en la literatura científica como *Multiple Target Tracking (MTT).*

Este objetivo general se puede desglosar en los siguientes objetivos parciales:

1. *Estudio y justificación de las alternativas para la extracción de información del entorno,* que permita realizar de forma óptima las tareas de estimación perseguidas. Por los motivos señalados anteriormente se pretende emplear arrays de micrófonos y cámaras para este propósito.
2. *El número de targets* admitidos por el algoritmo podrá ser múltiple y variable en el tiempo.
3. *Ejecución en tiempo real*. La ejecución en tiempo real de los procesos y algoritmos implicados en la estimación de la posición y la pose, son fundamentales a la hora de garantizar la correcta funcionalidad de las aplicaciones que se integrarán en el entorno de los "espacios inteligentes".
4. *Verificación del aumento de la robustez, fiabilidad y tolerancia a fallos* de la técnica de estimación propuesta frente a otras que emplean un único tipo de sensores para la consecución de las mismas tareas.
5. *Validación de la técnica de fusión audiovisual propuesta* por comparación con otras técnicas audiovisuales basadas en diferentes algoritmos a los que se propondrán en la tesis, o de fusión sensorial que

empleen otro tipo de sensores como fuente de información, y ya hayan sido planteadas por la comunidad científica.

1.3 Aplicaciones

Las posibles aplicaciones que emplean la detección, localización, tracking y obtención de la pose de una persona u objeto en un entorno concreto son numerosas y a buen seguro su número se incrementará conforme los algoritmos proporcionados por la comunidad científica aumenten su precisión, robustez y fiabilidad.

Los ámbitos en los que en la actualidad se están empleando este tipo de técnicas son entre otros: vigilancia autónoma, interacción hombre máquina, apuntamiento automático de cámaras, guiado de robots móviles…etc y sobre todo el de los "espacios inteligentes" y los "meetings", entornos altamente dinámicos e interiores. En este último entorno, el conocer la ubicación de cada uno de los participantes así como si están hablando o no, es crucial para aumentar la interactividad del mismo tanto si estos se realizan de manera local o a través de videoconferencia.

Este aumento de la interactividad comentado anteriormente se consigue a través del procesamiento de la información capturada del medio y se traduce por ejemplo en la selección o apuntamiento de cámaras en función del sujeto que se encuentre hablando en cada momento [Wang y Chu 1997], en mejorar y realzar la señal acústica capturada por arrays de micrófonos, en identificar a cada uno de los participantes en el meeting empleando información procedente de distintos tipos de sensores, en poder identificar la tarea concreta que están realizando (e.g. una presentación),…etc.

Otro campo emergente para el que la localización y tracking de una individuo constituye un paso previo imprescindible es el de la comunicación no verbal de las personas (e.g. determinadas posiciones corporales, miradas o expresiones faciales), la cual permitiría por ejemplo analizar las interacciones entre los participantes en una reunión de manera automática.

1.4 Estructura del estado del arte

El presente trabajo se encuentra dividido en seis apartados claramente diferenciados.

En el *primer* apartado de introducción comienza con una justificación de la utilización de la fusión sensorial de sensores audiovisuales como un instrumento adecuado para robustecer los algoritmos de localización y tracking actuales, para a continuación fijar los objetivos que se perseguirán a lo largo de esta tesis y definir aplicaciones de interés para los sistemas propuestos, así como posibles problemas que pueden presentarse en este tipo de aplicaciones.

En el *segundo* apartado se aborda el estudio del estado del arte de algoritmos que emplean array de micrófonos para la localización y tracking de los target de interés.

En el *tercer* apartado se realiza un estudio análogo al desarrollado en el segundo, pero en este caso centrado en las cámaras de visión como sistema sensor.

En el *cuarto* apartado se revisa el estado del arte en cuanto a fusión audiovisual se refiere, para concluir con un *quinto* apartado en el que se recogen las conclusiones y futuras líneas de investigación.

En cada uno de los apartados reseñados anteriormente se tratará, en primer lugar, de explicar conceptos básicos para la comprensión de la temática concreta, para a continuación describir las principales líneas de investigación, así como los resultados y principales problemas presentados por cada una de ellas.

Adicionalmente, y debido a su importancia en los diferentes ámbitos (audio, visión y fusión) que componen el presente estado del arte, se incluye un *sexto* apartado en forma de *anexo* en el que se recoge la formalización matemática que permite la implementación de un PF *(Particle Filter)* básico, así como las distintas versiones y modificaciones presentadas por la comunidad científica para mejorar su fiabilidad, robustez y tiempos de cómputo.

2. Audio: Arrays de micrófonos

2.1 Introducción

La utilización conjunta de arrays de micrófonos y sofisticados mecanismos de procesamiento de señal, ha permitido a la comunidad científica la adquisición remota de señales de audio con una alta calidad. Estas aplicaciones explotan la capacidad de *filtrado espacial* del array para realzar la señal acústica producida por una persona y suprimir (al menos en parte) las señales sonoras producidas por otras personas o fuentes indeseadas. A este procedimiento se le conoce como *beamforming*[3] y permite dirigir el array hacia un punto o dirección fija del espacio o bien utilizar algoritmos que sigan la posición de uno o más hablantes, ajustando el enfoque del array en consecuencia. De esta manera se evita la utilización de micrófonos de mano, cabeza o solapa mucho más intrusivos en las aplicaciones finales.

Los arrays de micrófonos se utilizan en la actualidad principalmente en teleconferencias [Wang y Chu 1997], en reconocimiento de voz [Hughes y otros 1999], en identificación de personas [Sturim 2007], o para capturar voz en entornos altamente reverberantes [Varma 2002] y [Royo 2007].

Por otro lado, y como ya se apuntó en el apartado de introducción, la justificación de la utilización de arrays de micrófonos dentro de los "espacios inteligentes" es clara al ser el habla la forma habitual de comunicarse de los seres humanos y por lo tanto la forma natural de interactuar con diferentes dispositivos presentes en el "espacio inteligente". Sin embargo la utilización de cada tipo de sensor presenta una problemática inherente al mismo y al entorno que le rodea, que en determinados casos puede reducir la robustez y fiabilidad de los algoritmos empleados en una aplicación concreta. En el caso del audio y los arrays de micrófonos se tiene que:

1. En numerosas situaciones reales la fuente sonora se encuentra ubicada en un *entorno altamente reverberante*, donde la presunción de campo lejano

[3] El término inglés *"beamforming"* significa conformación del haz, y hace referencia a la capacidad de los micrófonos o arrays de micrófonos para dirigir su patrón de directividad a diferentes direcciones del espacio. Debido a su amplia utilización en la literatura científica se pretende utilizar ambos términos por igual.

puede no ser cierta y puede encontrarse en movimiento, lo que provoca que la consideración del emisor como una fuente puntual no sea correcta.

2. La señal recibida por los arrays de micrófonos presentan una *baja relación señal a ruido*, lo cual afecta negativamente a la estimación.
3. La voz humana es una *señal de banda ancha*.
4. La naturaleza de las conversaciones humanas y en particular de los meetings las transforma en *actividades muy dinámicas*, en las que 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.
5. Otro factor importante de la voz humana es que se trata de una *señal intermitente*, por lo que será necesario estimar de una manera fiable su presencia en cada instante de tiempo.
6. En aplicaciones de *tracking en tiempo real* que emplean beamforming, el movimiento de la persona debe ser insignificante respecto a la duración del segmento de datos utilizado para computar dicha estimación. De otra manera la *latencia acumulada debido a la utilización de de grandes segmentos* de datos podría producir retardos inaceptables entre la producción del sonido por la persona y la generación de la estimación del sistema.

2.2 Propagación de ondas acústicas

El estudio realizado a lo largo del presente estado del arte en referencia a ondas acústicas asume que estas se propagan según la ecuación lineal de onda [Ziomek 1995]:

$$\nabla^2 x(t,\vec{r}) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} x(t,\vec{r}) = 0 \qquad (2.1)$$

donde $x(t,\vec{r})$ representa la presión del sonido para cualquier tiempo y posición, ∇^2 el operador Laplaciana y c la velocidad de sonido en m/s.

Gracias a esto, el camino recorrido por la onda desde el emisor hasta los micrófonos puede ser modelado como un sistema lineal. Para que esta modelización sea totalmente válida, tanto el medio como la fuente emisora tienen que cumplir unas características que se conocen como *simple acoustic conditions* [Dibiase 2000] pag. 9:

- *La fuente emite ondas sonoras esféricas*. Se asume que el tamaño y la forma del emisor no afecta significativamente en la propagación del sonido. Para un análisis más realista del patrón de radiación de la cabeza humana se puede consultar [Meuse 1994].
- *El efecto Doppler es despreciable.* La fuente puede estar en movimiento, pero en ningún caso será comparable a la velocidad del sonido, por lo que la frecuencia permanece invariante.
- *El medio es homogéneo.* La velocidad de sonido se considera constante (sobre 342 m/s) independientemente de la temperatura, la humedad y los parámetros climatológicos. Para un estudio detallado de esta variación [Alonso 2004].
- *El medio es no dispersivo.* La dispersión produce cambios en la velocidad de propagación con la frecuencia, lo cual no encaja dentro de un sistema lineal, por lo que este efecto será despreciado.
- *El medio es sin pérdidas.* Un medio sin pérdidas no absorbe energía de las ondas sonoras, la atenuación de las mismas se debe exclusivamente a su propagación en forma de ondas esféricas y es independiente de la frecuencia.

Bajo los supuestos anteriores y teniendo en cuenta la propagación esférica de las ondas acústicas, la ecuación (2.1) se puede resolver por el método de separación de variables [McCowan 2001], obteniéndose:

$$x(t,\vec{r}) = -\frac{A}{4\pi r} e^{j(wt-kr)} \tag{2.2}$$

donde $r = |r|$ es la distancia radial desde la fuente y k es el número de onda dado por $2\pi / \lambda$. De la ecuación (2.2) se desprenden importantes implicaciones:

- En general, las ondas acústicas se propagan como ondas esféricas y con una amplitud que decae proporcionalmente a la distancia a la fuente.

- La señal acústica se puede reconstruir *temporalmente* muestreando la señal en cada posición del espacio, o *espacialmente* muestreando en cada instante de tiempo.
- El principio de superposición de ondas permite la concurrencia de múltiples señales sin interacción entre las mismas, por lo que utilizando en el conocimiento de sus características espacio-temporales se pueden diseñar algoritmos que distingan y separen las distintas señales.

2.2.1 Atenuación, ruido y reverberación

La propagación de las ondas como señales esféricas y la consiguiente *atenuación* de la amplitud con la distancia, producen una disminución de la relación señal a ruido (SNR), lo cual puede hacer que la potencia capturada por el micrófono sea insuficiente si éste se encuentra situado muy lejos de la fuente.

Otro factor que, al igual que la atenuación, afecta negativamente a la señal capturada por los micrófonos es el *ruido*[4]. En el ámbito específico de la localización empleando arrays de micrófonos, todo sonido excepto el emitido por el target a seguir es considerado ruido, el cual es puede clasificar en 2 grupos principales:

1. **Ruido no direccional**: Es debido principalmente al ruido de fondo y se puede modelar como ruido blanco. Según se recoge en [Svaizer 1997], este ruido reduce la SNR de la señal, pero no influye significativamente en la señal acústica dominante producida por el target objetivo.
2. **Ruido direccional**: Este tipo de ruido lo producen otros hablantes situados en el entorno del array de micrófonos o fuentes de ruido estacionarias procedentes de zonas concretas del espacio, como por ejemplo los aires acondicionados o los ventiladores de los ordenadores. Estas señales indeseadas se suman coherentemente con la señal deseada, pudiendo provocar errores en la estimación de la posición.

[4] En acústica el término ruido hace referencia al conjunto de señales sonoras indeseadas que sumadas a su llegada al micrófono desvirtúan la señal deseada.

Por último, la propagación de señales acústicas en espacios cerrados es generalmente multicamino, debido a lo cual la onda sonora alcanza los micrófonos siguiendo diferentes y múltiples caminos. Por ello, además de la contribución de la onda directa la señal captada contendrá múltiples copias de la señal retardadas, atenuadas y distorsionadas debido a las reflexiones y difracciones en los objetos y superficies de la habitación. A este fenómeno se le conoce como *reverberación* y está caracterizado por la respuesta al impulso de la habitación y depende principalmente del tamaño de esta (que define el tiempo durante el cual persiste la reverberación) y el tipo de superficies de la misma (que define la energía absorbida en cada reflexión). Por ello la respuesta al impulso de la habitación puede modelarse como una delta en el origen correspondiente a la onda directa y una serie de deltas retardadas, atenuadas y despolarizadas fruto de las ondas reflejadas [Alonso 2004], [Rodriguez 2006].

Por lo tanto, las reverberaciones provocan reflexiones de de la señal, la cual llega a los micrófonos por diferentes caminos y direcciones a la señal directa. Estas reflexiones sobre todo las conocidas como *"early reflection"* pueden producir errores en la estimación. La Figura 2.1 clarifica estos conceptos.

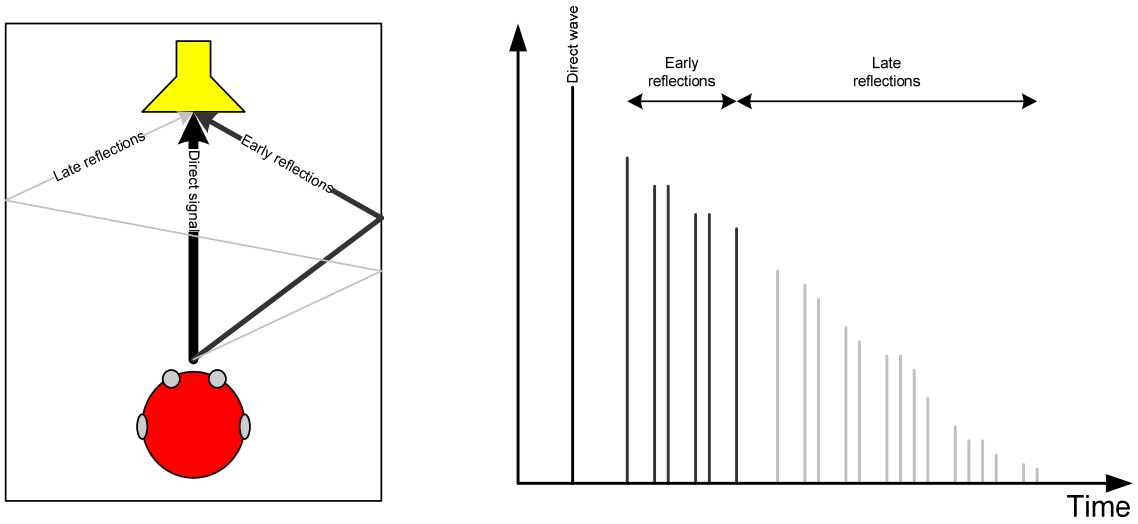


Figura 2.1 Ejemplo de multiphat y ecograma de una habitación.

2.2.2 Modelo de señal en los micrófonos

Tal y como se recoge en apartados anteriores, la señal recibida por los micrófonos está compuesta por la onda directa, por un conjunto de ondas reflejadas y ruido. En este apartado se trata de modelar matemáticamente dichos conceptos [Dibiase 2000] pp. 10-15 y [Castro 2007] pp. 9-11.

Bajo la presunción de "simple acoustic condition" se tiene que la *onda directa* recibida en el micrófono es una versión escalada y retardada de la señal $x(t)$ producida por la fuente según se recoge en (2.3), en la que se ha considerado el micrófono como el origen de coordenadas.

$$x_{direct}(r,t) = \frac{a}{r} \cdot s\left(t - \frac{r}{c}\right) = \frac{a}{r} \cdot s(t-\tau) \tag{2.3}$$

donde $x_{direct}(r,t)$ es la señal capturada en el micrófono teniendo en cuenta únicamente la onda directa, a es la amplitud de la onda, r es la distancia entre micrófono y la fuente, $s(t)$ es la señal original, c es la velocidad del sonido y τ es el tiempo de retardo entre emisor y receptor.

Aplicando la teoría de sistemas lineales se puede modelar la señal recibida en el micrófono $x(t)$, como la señal original $s(t)$ convolucionada con la respuesta al impulso del la habitación $h(t)$, obteniéndose:

$$x(\vec{r}_m, \vec{r}_s, t) = s(t) * h(\vec{r}_m, \vec{r}_s, t) \tag{2.4}$$

donde $\vec{r}_m$ y $\vec{r}_s$ son la posición de micrófono y fuente respectivamente.

La respuesta al impulso de la habitación caracteriza todos los caminos de señal entre fuente y receptor, pero es difícil de estimar en situaciones reales, por lo que la ecuación (2.4) se suele utilizar en una forma más útil en la que se recogen los caminos directo y reflejados:

$$x(\vec{r}_m, \vec{r}_s, t) \cong \frac{a}{r} \cdot s(t-\tau_m) + s(t) * u(\vec{r}_m, \vec{r}_s, t) \tag{2.5}$$

En (2.5) $u(t)$ representa la respuesta al impulso del sistema teniendo en cuenta únicamente las ondas reflejadas, las cuales se modelan como versiones filtradas de la

original. En esta expresión destaca que el modelo se define en base al parámetro τ_m que representa retardo de la onda entre fuente y destino, el cual es fácilmente medible bajo en condiciones de campo lejano.

En este punto se puede formular finalmente la señal recibida en el micrófono m añadiendo a (2.5) un término $v_m(t)$ correspondiente al ruido aditivo, con lo que se obtiene:

$$x_m(\vec{r}_m, \vec{r}_s, t) \cong \frac{a}{r_m} \cdot s(t - \tau_m) + s(t) * u(\vec{r}_m, \vec{r}_s, t) + v_m(t) \tag{2.6}$$

2.2.3 Campo cercano vs. campo lejano

Tal y como se comentó en el apartado 2.2 las ondas sonoras se propagan de forma esférica. Sin embargo cuando la distancia entre la fuente y el array de micrófonos r es mucho más grande que el tamaño físico del array R, las ondas que llegan al array pueden considerarse planas, ya que la curvatura de la onda es muy pequeña y puede ser despreciada, en comparación con el tamaño del array. A esta situación se le denomina de *campo lejano* y se cumple cuando se satisface la siguiente desigualdad [McCowan 2001]:

$$|r| \succ \frac{2R^2}{\lambda} \tag{2.7}$$

En estas condiciones y observando la Figura 2.2(a), se puede deducir que la distancia extra que tiene que viajar la onda para llegar al micrófono n respecto al micrófono 0 viene dada por:

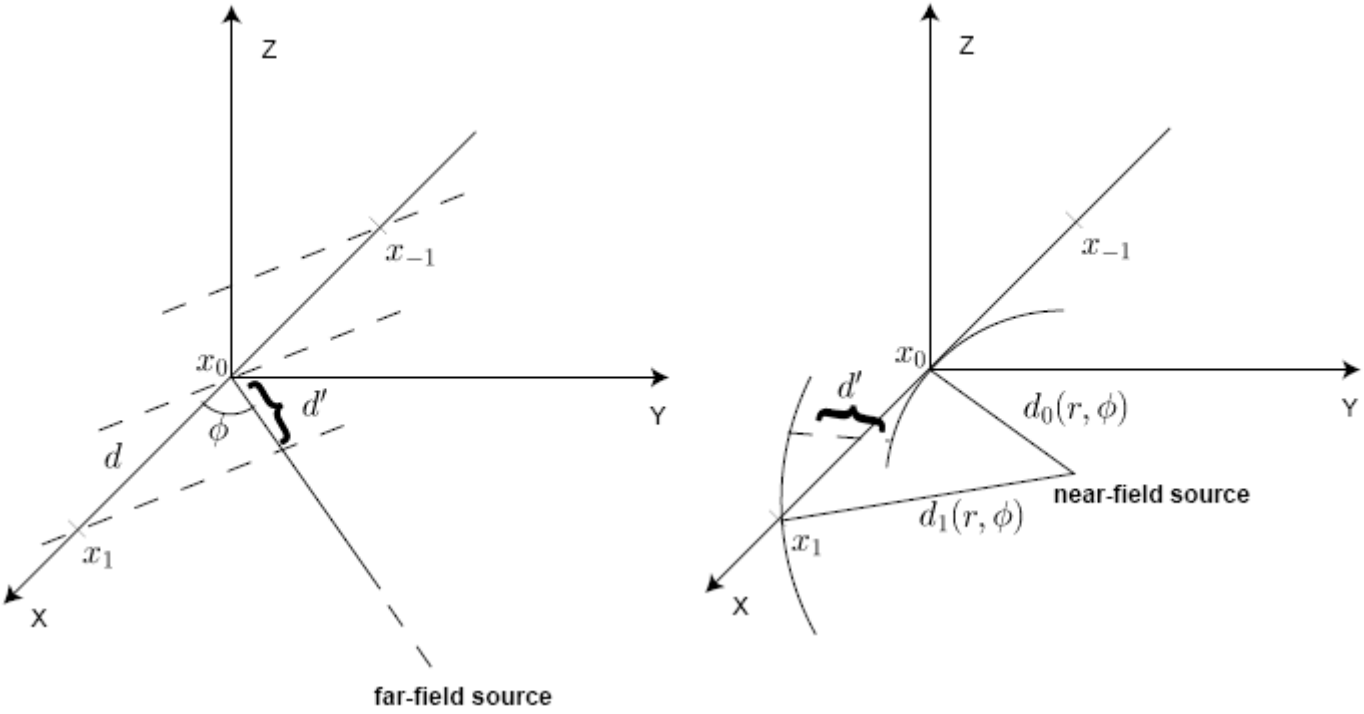
$$d' = nd\cos(\phi) \rightarrow \Delta\tau = \tau_0 - \tau_n = n \cdot \frac{d\cos(\phi)}{c} \tag{2.8}$$

La expresión anterior en la que $\Delta\tau$ representa la el retardo de la onda entre los micrófonos 0 y n, permite calcular de una manera sencilla la dirección de llegada (DOA) de la señal una vez estimado dicho retardo entre las señales.

En caso de que no se cumpla la condición de campo lejano nos encontramos en el caso representado en la Figura 2.2(b), denominada campo cercano, donde se puede

observar que no es posible obtener la DOA de una forma simple como sucedía en el caso (a), de hecho no existe dicho concepto en campo cercano, ya que la fuente emisora está tan cerca que la DOA es diferente para cada elemento del array.

Gran cantidad de algoritmos de localización basados en la estimación de los TDOA realizan una asunción de campo lejano. Sin embargo en entornos cerrados como los "espacios inteligentes", el área de campo lejano se encuentran en la mayoría de los casos fuera de los límites de la habitación, impidiendo así su aplicación. Para solucionar esta problemática aparecen los algoritmos SRP (2.4.2.3), basados en el apuntamiento del array a una zona concreta del entorno.



(a) (b)

Figura 2.2 Frente de ondas procedente de fuente de campo (a) lejano (b) cercano

2.3 Arrays de micrófonos

Un array de micrófonos está constituido por un conjunto de micrófonos situados en unas posiciones espaciales conocidas. Su utilización dentro de la ingeniería se asienta en su capacidad de filtrado espacio-temporal, gracias a la cual es posible la localización de personas dentro de un entorno y mejorar la señal recibida con respecto a la capturada con un sólo micrófono.

2.3.1 Conceptos básicos

En este apartado se presentan conceptos fundamentales para la comprensión del procesado de señal acústica presentado en posteriores apartados de este estado del arte. Para una revisión más profunda de estos y otros conceptos se puede consultar [McCowan 2001].

Posiblemente la característica más importante de los arrays de micrófonos es la de *filtrado espacial*, la cual nos permite capturar la señal procedente de la zona concreta del espacio donde se encuentra el target deseado, mientras que las señales procedentes de otros targets, fuentes indeseadas o ruido son filtradas. Esta característica constituye la base del Multiple Targets Tracking (MTT) realizado con arrays de micrófonos, por lo que en adelante nos centraremos en ver cómo estos son capaces de llevarla a cabo.

Un array de sensores puede considerarse como una versión muestreada de una apertura continua, donde dicha apertura sólo es excitada en un número finito de puntos. Además cada elemento puede ser considerado como una apertura continua, por lo que la respuesta global del array se obtiene por la superposición de todos y cada uno de los sensores. En base a esto se definen:

- La *longitud efectiva,* L, de un array uniforme, como la longitud de apertura continua que es muestreada, la cual viene dada por $L = N \cdot d$ siendo N el número de micrófonos y d la distancia entre micrófonos.
- La *longitud física*, como la distancia entre el primer y el último micrófono, la cual se expresa como $d \cdot (N-1)$.

2.3.2 Patrón de directividad

Tal y como se comento en el párrafo anterior, la respuesta global del array de micrófonos se obtiene por superposición de las respuestas individuales de cada uno. A esta respuesta se le denomina *patrón de directividad* el cual se puede expresar, para arrays lineales equiespaciados, como [McCowan 2001]:

$$D(f,\theta,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} \omega_n(f) e^{j\frac{2\pi}{\lambda}\sin(\theta)\cos(\phi)\cdot nd} \tag{2.9}$$

donde N es el número de elementos del array, $\omega_n(f)$ es el peso complejo conjugado de cada elemento n, d es la distancia entre micrófonos y el convenio utilizado para la representación en coordenadas esféricas se muestra en la Figura 2.3.

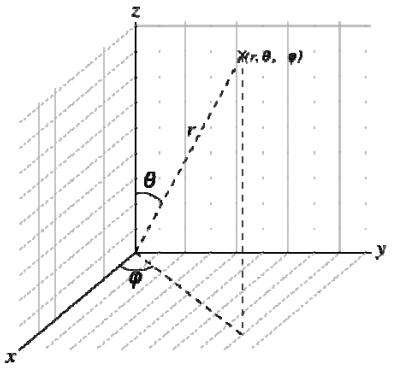


Figura 2.3 Convenio de representación en coordenadas esféricas

Considerando únicamente la dirección horizontal ($\theta = \pi/2$):

$$D(f,\phi) = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} \omega_n(f) e^{j\frac{2\pi}{\lambda}\cos(\phi)\cdot nd} = \sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} \omega_n(f) e^{j\frac{2\pi f}{c}\cos(\phi)\cdot nd} \qquad (2.10)$$

La fórmula anterior arroja interesantes conclusiones, ya que permite observar el comportamiento del patrón de directividad del array con la variación de diferentes parámetros:

- *Variación del número de elementos*, N, con L y f fijos: La Figura 2.4(a) muestra el descenso de los lóbulos laterales conforme aumenta N, con el consiguiente aumento de la directividad global.
- *Variación de la longitud efectiva*, $L = Nd$, con N y f fijos: La Figura 2.4(b) muestra la disminución de la anchura del lóbulo central conforme aumenta L, con el consiguiente aumento de la directividad global.
- *Variación de la frecuencia*, f, con N y L fijos: En la Figura 2.4(c) se observa la disminución de la anchura del lóbulo central y por tanto el aumento de la directividad global con el aumento de la frecuencia.

En base a lo anterior se puede concluir que la anchura del lóbulo principal es inversamente proporcional al producto $(N \cdot d \cdot f)$. Sin embargo en la mayoría de la

aplicaciones prácticas (especialmente en las relacionadas con localización como la que nos ocupa) es deseable tener una anchura constante y dado que N suele ser un parámetro fijo en la mayoría de ellas se debe asegurar que el producto $f \cdot d$ permanezca lo más constante posible.

Otra conclusión importante que se puede extraer, es la relación de compromiso que se debe alcanzar en sistemas que emplean arrays de micrófonos mediante una anchura de lóbulo principal lo suficiente estrecha para conseguir un filtrado espacial adecuado que elimine la mayor cantidad posible de señales indeseadas, sin que esta anchura sea excesivamente estrecha y provoque que el algoritmo de localización (el cual no conoce a priori la posición de los targets) filtre también la señal deseada y por lo tanto se produzcan errores en la estimación de la posición.

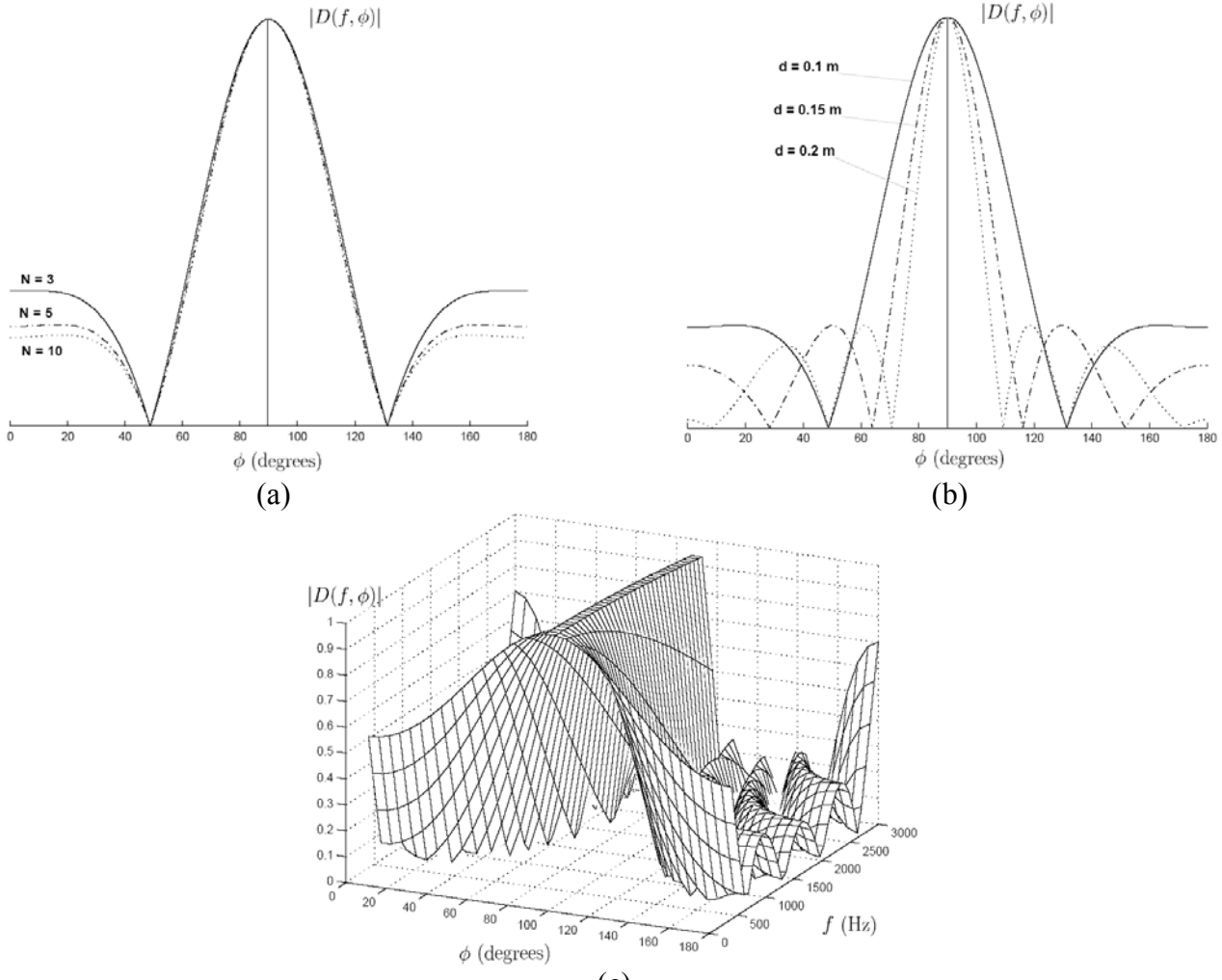


Figura 2.4 Patrón de directividad para variaciones de (a) N (f=1 kHz, L=0.5) (b) L (f=1kHz, N=5) (c) 400Hz≤f≤3000Hz (N=5, d=0.1m)

En el apartado 2.1 ya se señaló el gran ancho de banda de la señal de voz humana [20Hz – 20KHz] como uno de los principales problemas en las aplicaciones que utilizan la capacidad de filtrado espacial de los arrays de micrófonos con propósitos de localización. La Figura 2.4(c) describe a la perfección dicha problemática, por la cual las señales indeseadas procedentes de direcciones indeseadas son atenuadas en frecuencias altas en las que el lóbulo principal es más estrecho, mientras que dicha atenuación será menor conforme la frecuencia disminuye y la anchura del lóbulo principal aumenta, lo cual provoca grandes interferencias en frecuencias bajas.

2.3.3 Aliasing[5] espacial

De acuerdo con el principio de Nyquist, la frecuencia de muestreo debe ser mayor o igual que el doble de la máxima componente frecuencial de la señal para evitar aliasing en frecuencia. Análogamente podemos aplicar este principio a los arrays de micrófonos, que constituyen una versión muestreada de una apertura continua [McCowan 2001].

$$f_s = \frac{1}{T_s} \geq 2 \cdot f_{\max} \xrightarrow{arrays\_de\_mics} f_{x_s} = \frac{1}{d} \geq 2 \cdot f_{x_{\max}} \tag{2.11}$$

donde f_{x_s} es la frecuencia de muestreo espacial en muestras por metro, $f_{x_{\max}}$ es la mayor componente de frecuencia espacial y el array es lineal. La frecuencia a lo largo del eje x viene dada por:

$$f_{x_s} = \frac{\sin(\theta)\cos(\phi)}{\lambda} \xrightarrow{máximo\_para} f_{x_{\max}} = \frac{1}{\lambda_{\min}} \tag{2.12}$$

con lo que la condición final que se obtiene es:

$$d \leq \frac{\lambda_{\min}}{2} \tag{2.13}$$

donde $\lambda_{\min}$ representa la mínima longitud de onda de la señal de interés.

[5] El término inglés Aliasing significa solapamiento y está ampliamente extendido a lo largo de la literatura de señales y sistemas, por lo que en adelante se empleará el término inglés.

A la ecuación (2.13) se la conoce como *teorema de muestreo espacial* y se debe cumplir para evitar el aliasing espacial, el cual provoca la aparición de grandes lóbulos en direcciones indeseadas en el patrón de directividad, con la consiguiente integración de información procedente de dichas direcciones y los evidentes efectos negativos en cuanto a distorsión de la señal deseada y localización de los targets de la escena. La Figura 2.5 representa gráficamente dicha problemática (otras figuras ilustrativas a este respecto se pueden encontrar en [Seltzer 2003] pags 24-25):

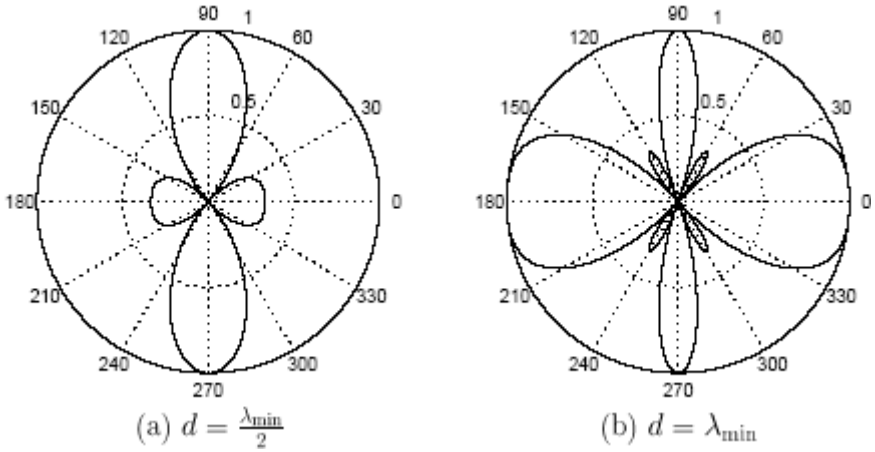


Figura 2.5 Patrón de directividad (a) sin (b) con aliasing espacial

Beamforming

Se conoce como *beamforming* a la técnica que permite dirigir el patrón de directividad del array hacia una zona concreta del espacio.

Con este objetivo, debemos fijarnos en el término $\omega_n(f)$ que representa el peso asignado a cada uno de los micrófonos del array dentro del patrón de directividad general (ecuación 2.9). Hasta este momento se han considerado pesos iguales para todos los micrófonos y por lo tanto $\omega_n(f) = 1/n$, pero este término es en general exponencial complejo y puede ser expresado como:

$$\omega_n(f) = a_n e^{j\varphi_n(f)} \tag{2.14}$$

donde $a_n(f)$ y $\varphi_n(f)$ representan la amplitud y la fase de la función de peso, por lo que modificando $a_n(f)$ se puede modificar la forma del patrón de directividad y modificando $\varphi_n(f)$ se puede controlar la localización angular del lóbulo principal.

Sustituyendo (2.14) en (2.9) se obtiene:

$$D(f,\theta,\phi)=\sum_{n=-\frac{N-1}{2}}^{\frac{N-1}{2}} a_n(f)e^{j\frac{2\pi}{\lambda}\sin(\theta)\cos(\phi)nd+\varphi_n(f)} \tag{2.15}$$

Usando como peso de fase

$$\varphi_n(f)=-2\pi\frac{\sin(\theta')\cos(\phi')}{\lambda}nd \tag{2.16}$$

el patrón de directividad se orienta en la dirección θ',ϕ', siendo reseñable que al no modificar los pesos de amplitud, la única diferencia entre el patrón de directividad original y el nuevo es la orientación del lóbulo principal, como muestra la Figura 2.6.

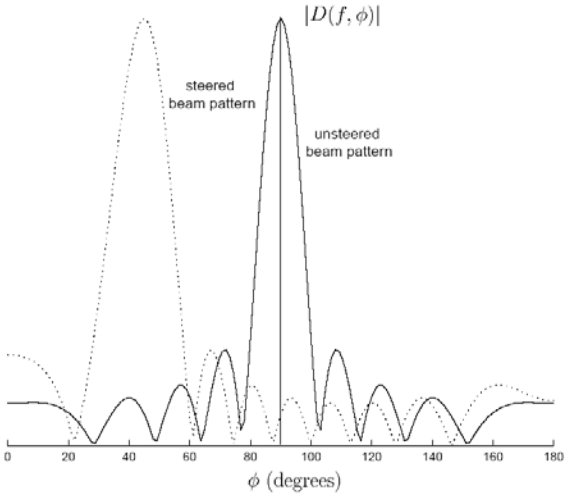


Figura 2.6 Patrón de directividad original y modificado (ϕ' = 45 deg , f=1KHz, N=10, d=0.15m)

A continuación se presentan las formas clásicas para la realización del *beamforming.*

2.3.3.1 Delay-and-sum beamformer

Es la técnica de beamforming más simple y puede ser interpretada desde el punto de vista de la frecuencia o del tiempo.

En el **dominio frecuencial** basta aplicar los conceptos presentados en 0, con lo que considerando únicamente el patrón de directividad horizontal quedaría el siguiente peso de fase para cada micrófono:

$$\varphi_n(f) = -2\pi \frac{(n-1)d\cos(\phi')f}{c} \tag{2.17}$$

Teniendo en cuenta (2.17) y una ponderación de pesos por igual para todos los micrófonos ($a_n(f) = \frac{1}{N}$), la función de peso de los micrófonos resultante es:

$$\omega_n(f) = \frac{1}{N} e^{j\frac{-2\pi f}{c}(n-1)d\cos(\phi')} \tag{2.18}$$

En base a esto se puede expresar la salida del array como la suma de la señal de salida de todos los micrófonos ponderados por la función de pesos anterior, obteniéndose:

$$y(f) = \frac{1}{N}\sum_{n=1}^{N} x_n(f) e^{j\frac{-2\pi f}{c}(n-1)d\cos(\phi')} \tag{2.19}$$

Nótese que por conveniencia se ha modificado el rango del sumatorio anterior con respecto a (2.15).

El desplazamiento en frecuencia presentado anteriormente puede ser implementado mediante el correspondiente retardo de tiempo en el **dominio temporal**, donde el retardo correspondiente al sensor n viene dado por:

$$\tau_n = \frac{(n-1)d\cos(\phi')}{c} \tag{2.20}$$

que representa el tiempo que el frente de ondas plano emplea para llegar desde el micrófono de referencia[6] al enésimo micrófono.

Por lo que la salida del array de micrófonos en el dominio temporal se puede expresar como:

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} x_n(t-\tau_n) \tag{2.21}$$

[6] Micrófono de referencia: El convenio más extendido consiste en considerar como micrófono de referencia al más cercano al hablante, de tal manera que los retardos temporales de la señal acústica en el resto de micrófonos se midan con respecto a este.

2.3.3.2 Filter-and-sum beamformer

El filter-and-sum beamformer constituye una generalización del delay-and-sum beamformer consistente en añadir prefiltrado a la entrada de cada uno de los micrófonos.

La ecuación (2.22) y la Figura 2.7 expresan matemática y gráficamente la salida de un filter-and-sum beamformer en el dominio de la frecuencia:

$$y(f) = \frac{1}{N}\sum_{n=1}^{N} W_n(f) x_n(f) e^{j\frac{-2\pi f}{c}(n-1)d\cos(\phi')} \tag{2.22}$$

En el dominio temporal se obtiene [Seltzer 2003]:

$$y[n] = \sum_{m=0}^{M-1}\sum_{p=0}^{P-1} h_m[p]\cdot x_m[n-p-\tau_m] \tag{2.23}$$

Tanto el delay-and-sum como el filter-and-sum son ejemplos de beamforming fijo, ya que los parámetros de procesamiento del algoritmo no cambian a lo largo del tiempo.

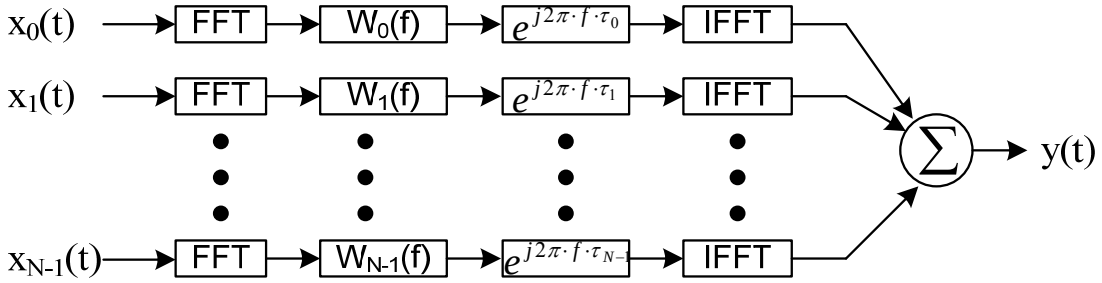


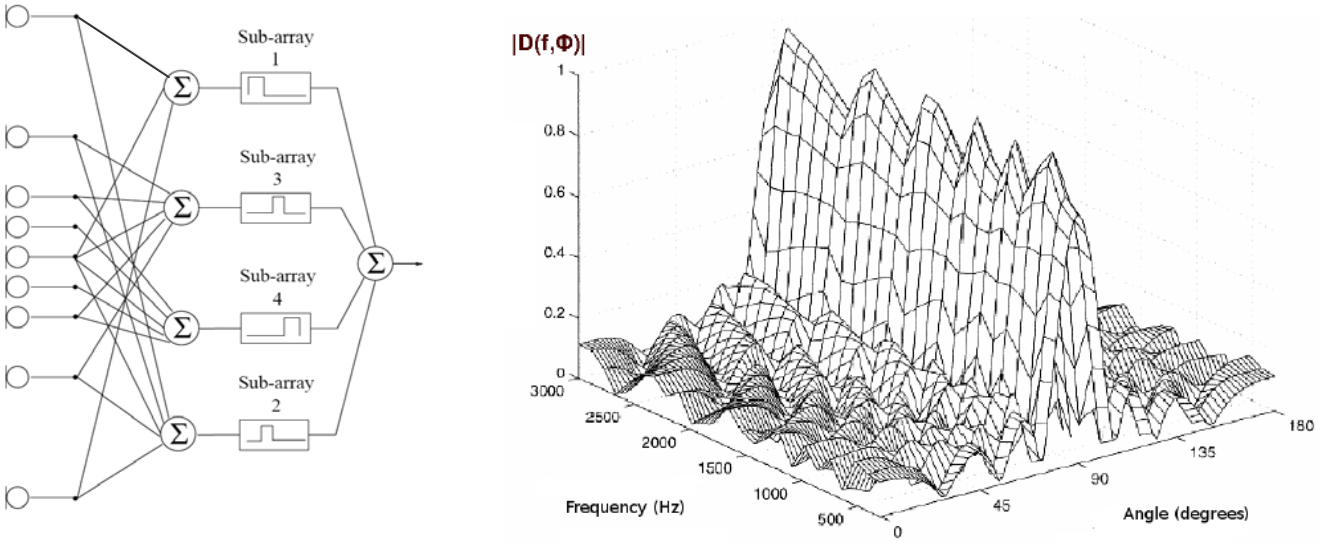
Figura 2.7 Filter-and-sum beamformer en el dominio de la frecuencia

2.3.3.3 Sub-array Beamforming

La ecuación del patrón de directividad (2.9) y su representación gráfica (Figura 2.4) muestran claramente la variación inversamente proporcional de la anchura del lóbulo principal del patrón de directividad con la frecuencia y la separación entre micrófonos (producto $f \cdot d$). Esta dependencia provoca que la anchura del lóbulo principal y los lóbulos secundarios sólo permanezcan constantes para anchos de banda

pequeños, en los cuales éste no es significativo con respecto a la frecuencia central. Sin embargo la voz humana es una señal de banda ancha (2Hz, 20KHz), por lo que la utilización de un array lineal equiespaciado no es adecuada si se desea obtener un patrón de directividad constante.

Una técnica que permite obtener un patrón de directividad constante en un amplio rango de frecuencias consiste en emplear arrays armónicos, cuyos elementos puedan formar parte de diferentes sub-arrays equiespaciados. En esta técnica, agrupada dentro de las denominadas *Constant Directivity Beamforming (CDB)*, *c*ada sub-array es restringido a una banda de frecuencia diferente mediante la aplicación de filtros paso banda y sumados a su salida para obtener la salida total. La Figura 2.8 muestra un ejemplo de este tipo de arrays.



(a) (b)

Figura 2.8 Ejemplo de Sub-array Beamforming y su patrón de directividad

Es importante recordar que según aumenta la frecuencia menor es la longitud del array necesario para mantener constante la anchura del lóbulo principal y que un número constante de elementos en el array proporciona lóbulos secundarios constantes.

2.3.3.4 Beamforming adaptativo

Los métodos presentados anteriormente constituyen los tipos más importantes de *beamforming fijo*, denominados así debido a que sus parámetros de procesamiento no cambian dinámicamente con el tiempo. En contrapunto a estos, en el *beamforming*

adaptativo dichos parámetros varían dinámicamente en base a ciertos criterios de optimización, bien muestra a muestra o frame a frame.

El primer ejemplo de beamforming adaptativo lo encontramos en [Frost 1972], en el cual los pesos aplicados a cada una de las señales del array son modificados dinámicamente con el objetivo de minimizar la potencia de ruido a la salida del mismo mientras se mantiene la respuesta en frecuencia deseada en la dirección del objetivo.

En [Griffiths y Jim 1982] se presenta el *Generalizad Sidelobe Canciller (GSC)* como una alternativa a la propuesta de Frost consistente en 2 estructuras. La primera implementa una estructura de beamforming clásico, mientras que la segunda, encargada de la parte adaptativa, trata de eliminar la señal deseada a su salida dejando únicamente señales indeseadas y ruido. Con ello teóricamente al restar a la señal de salida de la primera estructura la señal de salida de la segunda tendremos únicamente la señal deseadas sin ruido. La estrategia seguida para la asignación de pesos en la parte adaptativa esta orientada a eliminar la parte de la señal común a las 2 estructuras, es decir, la señal deseada. La estructura del GSC se presenta en la Figura 2.9.

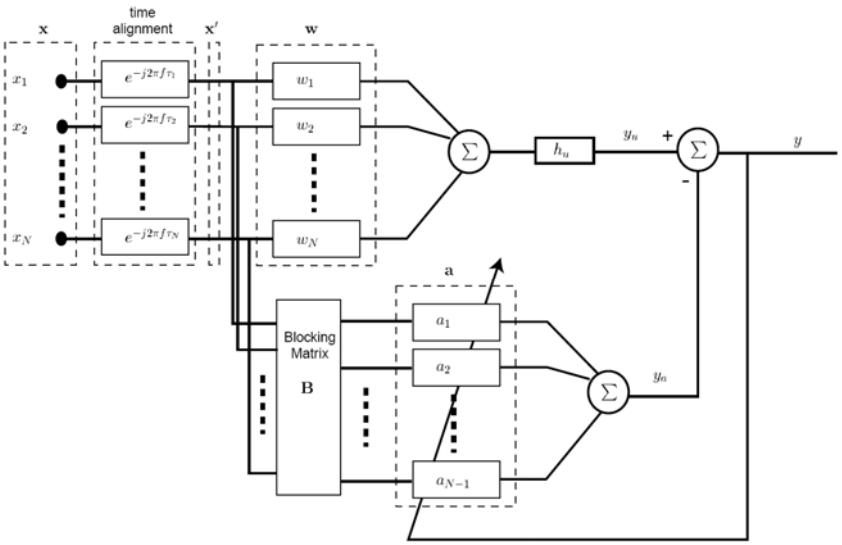


Figura 2.9 Estructura del Generalised Sidelobe Canceller

En la práctica esta estructura puede producir distorsiones de la señal deseada debido a que el bloque adaptativo no elimina totalmente la señal deseada, lo cual es particularmente problemático para señales de banda ancha como la voz humana.

En algunos casos los parámetros del filtro pueden ser calibrados para un entorno o usuario concreto como por ejemplo en [Nordholm 1999], donde se propone un esquema de calibración para un teléfono manos libres en interiores de coches. Para ello se capturan una serie de señales de habla dentro del coche y se utilizan para la calibración inicial del filter-and-sum beamforming. Estos parámetros son actualizados dinámicamente en base a las señales de calibración almacenadas y las señales de ruido actualizadas.

Uno de los problemas principales de los métodos adaptativos, es que en general asumen que la señal procedente del objetivo y la señal interferente son incorreladas. Cuando no se cumple esta condición (e.g. entornos altamente reverberantes), la cancelación de la señal deseada no se realiza correctamente, lo cual degrada seriamente la calidad de la señal de salida. Para solucionar esto [Van Compernolle 1990] demuestra que la cancelación de la señal realizada por el filtro adaptativo puede mejorar sensiblemente si la adaptación de los parámetros se realiza únicamente cuando no hay señal de voz en la captura de los micrófonos.

Adicionalmente, este tipo de técnicas implican una elevada carga computacional, lo cual dificulta su implementación en entornos reales.

De este apartado se puede concluir que son muchos los esfuerzos realizados enfocados a mejorar el comportamiento de los filtros adaptativos en cuanto a la eliminación de la señal deseada, no habiéndose encontrado hasta el momento un método óptimo en entornos reales y altamente reverberantes, por lo que estos métodos no han tenido una gran aceptación en la comunidad científica.

2.3.3.5 Post-filtering[7]

En la práctica, las técnicas de beamforming convencionales como el delay-and-sum o el filter-and-sum no proporcionan los resultados esperados teóricamente en cuanto a reducción de ruido.

[7] El término inglés Post-filteing se utiliza sin traducir debido a su amplia difusión en la literatura científica de la rama.

La etapa de post-filtering aquí tratada, tiene como principal objetivo la reducción de dicho ruido, para lo cual se introduce un post proceso adicional a la salida de los esquemas de beamforming clásicos, tal y como se recoge en [Brandstein 2001] pág. 39, [McCowan 2001] pág. 58 y [Abad 2007] pág 41 entre otros.

En [Zelinski 1988] se introduce por primera vez el denominado *Wiener* post-filter para un delay-and sum beamformer, el cual se representa gráficamente en la Figura 2.10, y en el que la señal recibida en el array de micrófonos es previamente alineada en módulo TDC (Time delay compensation).

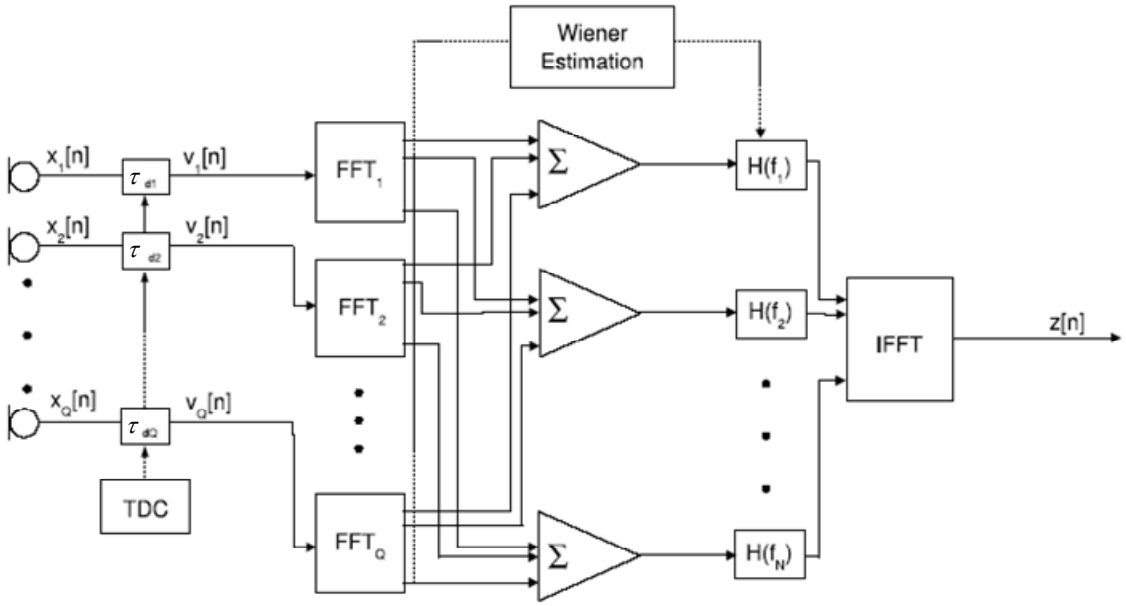


Figura 2.10 Esquema del Wiener post_filter para un delay-and-sum beamformer.

La incorporación de este postproceso al beamformer clásico nos permite utilizar la información obtenida a partir del filtrado espacial (Direction of arrival), para realizar al mismo tiempo un filtrado en frecuencia de la señal, lo cual se traduce en una mejora tanto a nivel especial como frecuencial.

La expresión general del filtro de Wiener para un array de micrófonos se presenta a continuación:

$$h_{opt}(f) = \frac{\phi_{ss}(f)}{\phi_{ss}(f) + \phi_{nn}(f)} \tag{2.24}$$

donde $\phi_{ss}(f)$ y $\phi_{nn}(f)$ representan las auto-spectral density de la señal deseada $s(f)$ y de ruido $n(f)$, a la salida del beamformer respectivamente.

Un problema típico de los filtros de Wiener es la estimación de $\phi_{ss}(f)$ y $\phi_{nn}(f)$. La disponibilidad de múltiples canales, típica en los arrays de micrófonos, posibilita una solución interesante bajo los siguientes supuestos:

1. La señal recibida en cada micrófono puede ser modelada como la suma de la señal deseada y ruido.
2. El ruido y la señal deseada son incorrelados.
3. La densidad espectral de potencia en cada micrófono es la misma $\left(\phi_{n_i n_i}(f) = \phi_{n\,n}(f), i = 1,\ldots, N\right)$.
4. El ruido entre los diferentes micrófonos es incorrelado $\left(\phi_{n_i n_j}(f) = 0, \forall i \neq j\right)$.
5. La señal de entrada se puede reconstruir perfectamente en fase.

   Bajo las condiciones anteriores se verifica que:

$$\phi_{v_i v_j}(f) = \phi_{ss}(f) \qquad\qquad \phi_{v_i v_j}(f) = \phi_{ss}(f) + \phi_{nn}(f) \tag{2.25}$$

   En cuyo caso $h(f)$ puede ser estimada mediante la siguiente expresión:

$$\hat{h}(f) = \frac{\dfrac{2}{Q(Q-1)}\Re\left\{\displaystyle\sum_{i=1}^{Q-1}\sum_{j=i+1}^{Q}\hat{\phi}_{v_i v_j}(f)\right\}}{\dfrac{1}{Q}\displaystyle\sum_{i=1}^{Q}\hat{\phi}_{v_i v_j}(f)} \tag{2.26}$$

Es evidente considerar que bajo las condiciones expuestas anteriormente la etapa de post-filter es particularmente interesante en presencia de ruido blanco, a pesar de lo cual es bastante útil en presencia de ruido difuso, el cual se aproxima razonablemente a las condiciones impuestas. En [Abad 2007] se demuestra que este tipo de filtro es efectivo en la eliminación de ruido incoherente, que el rechazo de ruido tanto correlado como incorrelado con la señal deseada es mayor si este no procede de la misma dirección que la señal deseada y que es robusto a pequeños errores en el apuntamiento del array.

La expresión general del filtro de Wiener se presenta a continuación:

$$\hat{h}(f) = \frac{\sum_{i=1}^{Q} |w_i(f)|^2}{\sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} w_i(f) w_j^*(f)} \cdot \frac{\Re\left\{ \sum_{i=1}^{Q-1} \sum_{j=i+1}^{Q} \hat{\phi}_{v_i v_j}(f) \right\}}{\frac{1}{Q} \sum_{i=1}^{Q} \hat{\phi}_{v_i v_j}(f)} \tag{2.27}$$

Este tipo de filtro propuesto por Zelinski ha sido ampliamente utilizado en el ámbito de los arrays de micrófonos, por ejemplo a partir de un GSC como beamformer ([Fischer 1995]) o en combinación con técnicas de eliminación de las reverberaciones en señales de voz basadas en el procesamiento por separado de las componentes de mínima fase y todas las fases de la señal de voz ([Gonzalez-Rodriguez 2000]).

2.4 Detección, localización, tracking y pose con sensores de audio

En general, el objetivo más importante de un sistema de localización es la precisión, que para el caso particular de arrays de micrófonos depende principalmente de cuatro factores.

1. Cantidad y calidad de los micrófonos empleados.
2. Ubicación relativa de los micrófonos entre sí y de estos con la fuente sonora.
3. Nivel de ruido y reverberación en el entorno.
4. El número de fuentes activas y el contenido espectral de sus emisiones.

El sistema deberá por tanto ser capaz de estimar el número de hablantes activos en cada instante de tiempo *(detección)*, su posición instantánea *(localización)*, su trayectoria espacio-temporal en el tiempo *(tracking)* y su orientación *(pose)*, para lo cual se distinguen **tres grupos de estrategias** claramente diferenciadas pero no necesariamente exclusivas [Amida 2007]:

1. *Time asynchrony*: Los tiempos de vuelo de las ondas acústicas desde los targets activos hasta los micrófonos son diferentes para cada uno de los micrófonos debido a las diferentes ubicaciones de los mismos en el espacio. Los métodos que se basan en este concepto, requieren normalmente un conocimiento muy

preciso de la geometría del array, pero sin embargo no requieren ningún conocimiento particular de la habitación, siendo los más utilizados en las aplicaciones prácticas. Algunas configuraciones típicas de los arrays de micrófonos en este tipo de técnica son *Uniform Linear Arrays (ULA), Uniform Circular Arrays (UCA) y Huge Microphone Array (HMA)* entre otras.

2. *Respuesta al impulso*: Esta técnica se basa en el concepto de que para una ubicación de target activo determinada, el camino recorrido por la onda sonora para llegar a cada uno de los micrófonos es diferente, por lo que la respuesta al impulso aplicada también lo será. Considerando que la respuesta al impulso del sistema es conocida de antemano, gracias a un proceso de calibración, es posible estimar la posición del target objetivo. Un inconveniente de este sistema es lo tedioso de dicha calibración, lo cual conduce a la aparición de diferentes técnicas para su realización automática, entre las que destaca la *bind Multiple Inputs Multiple Outputs (MIMO)* channel identification [Buchner 2004], realizada mediante un proceso online automático que no requiere un conocimiento de la geometría del array. El modelo de la habitación conseguido con esta técnica permite no sólo localización, sino también separar señales procedentes de múltiples targets simultáneamente. Esta línea de investigación está todavía en etapa de desarrollo, aunque algunos resultados preeliminares ya se pueden encontrar en [Buchner 2005]. La principal problemática de esta técnica es que la respuesta al impulso depende de la posición de los micrófonos y de las fuentes sonoras.

3. *Microphone channel*: Las técnicas más recientes propuestas por la comunidad científica van orientadas a la ubicación en una misma localización de varios micrófonos direccionales orientados hacia diferentes direcciones del espacio [Matsumoto 2006]. La localización del target se puede reconstruir en base a la función de transferencia de cada uno de los micrófonos, la cual se asume como conocida. Esta solución se puede combinar con las explicadas en el punto 1 [Rui 2005].

Tal y como se comentó anteriormente, el conjunto de soluciones que más aceptación tiene dentro de la comunidad científica son las pertenecientes al primer grupo, por lo que en adelante el presente estado del arte se centrará en las mismas.

Este grupo de soluciones se basan directa o indirectamente en la siguiente observación: Dadas 2 señales $x_1(t)$ y $x_2(t) = x_1(t - \tau_{12})$ recibidas por 2 micrófonos con un retardo relativo τ_{12}, la correlación de ambas señales $f(\tau) = \int_t x_1(t) x_2(t - \tau) dt$ presenta un máximo en $\tau = \tau_{12}$.

2.4.1 Detección

La señal de audio es por naturaleza una señal intermitente y por tanto no se deben realizar estimaciones durante los períodos de silencio.

Los detectores de voz tradicionales o Voice Activity Detectors (VAD) emplean características individuales del canal para calcular las métricas, como por ejemplo niveles de energía, cruces por cero,…etc. A partir de estas se aplican reglas de clasificación basadas en umbrales fijos o recalculados en los períodos de silencio. Esta forma de actuar hace que presenten problemas en entornos con baja SNR y especialmente con ruidos no estacionarios[8].

Otros métodos alternativos que consideran la correlación cruzada entre canales pueden encontrarse en [Lathoud 2005].

Otra técnica interesante para la implementación del VAD fue presentada en [Omologo 2003] basada en la Cross-Power Spectrum (CSP) de la señal capturada. Este factor resulta altamente interesante, ya que permite reutilizar la información de CSP calculada necesariamente durante el proceso de detección para la implementación del VAD, durante el proceso de localización, ya que los principales algoritmos de localización (GCC y SRP) se basan en el análisis de la CSP.

Una alternativa a los VAD en aplicaciones de seguimiento es la presentada por investigadores del IDIAP en [Lathoud y McCowan 2004] y denominada *short-term clustering*, la cual se estudia más adelante (2.4.3.3).

[8] Una señal se considera estacionaria cuando sus parámetros estadísticos no varían con el tiempo en un intervalo dado.

2.4.2 Localización

El termino localización dentro del contexto de arrays de micrófonos, se refiere a la obtención de la posición de varios targets activos en el espacio, durante el breve periodo de tiempo en el que la señal de voz puede ser considerada estacionaria (típicamente entre 20 y 30 ms).

A lo largo de este punto se presentaran las principales ventajas e inconvenientes de los algoritmos de localización que emplean arrays de micrófonos, los cuales pueden ser clasificados en tres grandes grupos [McCowan 2001]:

- Los que emplean *Time Difference of Arrival (TDOA).*
- Los basados en *High-resolution Spectral Estimation.*
- Los que utilizan *Steered Response Power (SRP).*

2.4.2.1 Basados en *Time Diference of Arrival (TDOA)*

Los algoritmos basados en este tipo de técnicas se dividen en dos fases claramente diferenciadas y secuenciales entre sí:

1. Se estiman los retardos en la señal acústica (*Time Delay Estimation* ≡ *TDE*) entre pares de micrófonos separados espacialmente. Según [Omologo 1994] hay tres formas principales para llevar a cabo esta primera etapa:
    a. *Correlación Cruzada Normalizada (CC).*
    b. *Correlación Cruzada Generalizada (GCC).*
    c. *Filtros adaptativos Least Mean Squared (LMS).*

    Comparativas realizadas entre los tres métodos anteriores en [Omologo 1994] demuestran que los métodos que emplean GCC (basados en CC) son los que presentan un mejor comportamiento, por lo que en apartados posteriores nos concentraremos únicamente en estas dos técnicas.
2. En base a los TDE obtenidos anteriormente y a la posición espacial de los micrófonos, se obtienen unas 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 y en base a determinados criterios de optimización, [Muñoz 2006] pags. 31-43, se obtiene la estimación de la posición.

A este respecto la comunidad científica ha desarrollado gran cantidad de variantes de este principio básico en base a la obtención de la estimación o la extensión en su aplicabilidad (2D vs. 3D, campo cercano vs. campo lejano…etc). Algunos ejemplos pueden encontrarse en [Varma 2002] pags. 24-27, [Brandstein 1997] o [Svaizer 1997].

Las principales ventajas de este tipo de técnicas son el bajo coste computacional y la simplicidad. Sin embargo su implementación práctica en sistemas reales se ve limitada debido al insuficiente comportamiento que presentan en entornos reverberantes o con alta SNR. Además de lo anterior, estos métodos suelen proporcionar una estimación de la *Direction Of Arrival (DOA)* en lugar de la localización espacial, por lo que en la práctica se prefieren los sistemas basados en beamforming que proporcionan una mayor robustez.

2.4.2.1.1 Correlación Cruzada Normalizada (CC)

Dado un array compuesto por N micrófonos, las señales acústicas recibidas en dos de ellos se pueden escribir como:

$$\begin{aligned} x_i(t) &= \alpha_i \cdot s(t) + n_i(t) \\ x_j(t) &= \alpha_j \cdot s(t + \tau_{ij}) + n_j(t) \end{aligned} \tag{2.28}$$

donde $s(t)$ representa la señal acústica recibida, $n_i(t)$ y $n_j(t)$ las componentes indeseadas capturadas por cada micrófono (ruido y reverberaciones), α_i y α_j las atenuaciones y τ_{ij} el retardo entre ambas señales debido a las diferentes ubicaciones espaciales de los micrófonos.

Con el objetivo de determinar τ_{ij} se recurre al cálculo de la correlación cruzada entre las 2 señales $C_{x_i x_j}(\tau)$, la cual analiza las similitudes entre las mismas y viene dada por:

$$C_{x_i x_j}(\tau) = E\left[x_i(t) \cdot x_j(t-\tau)\right] \tag{2.29}$$

En (2.25) E representa la esperanza, por lo que sustituyendo (2.24) en (2.25) se obtiene:

$$C_{x_i x_j}(\tau) = \alpha_i \alpha_j \cdot C_{s_i s_j}(\tau - \tau_{ij}) + C_{n_i n_j}(\tau) \tag{2.30}$$

Teóricamente τ_{ij} se puede obtener maximizando (2.26), pero en la práctica se trabaja con observaciones finitas en el tiempo, por lo que la función debe ser estimada dentro de una ventana temporal T. Esta estimación se denota como $\hat{C}_{x_i x_j}(\tau)$:

$$\hat{C}_{x_i x_j}(\tau) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} x_i(t) \cdot x_j(t-\tau)dt \tag{2.31}$$

Finalmente la estimación de TDOA, $\hat{\tau}_{ij}$, se puede obtener como:

$$\hat{\tau}_{ij} = \arg\max_{\tau} \hat{C}_{x_i x_j}(\tau) \tag{2.32}$$

La discusión anterior corresponde al dominio temporal. Aplicando el *teorema discreto de la correlación*, se obtiene el equivalente en el dominio de la frecuencia:

$$C_{x_i x_j}(\tau) = DFT(x_i(t)) \cdot DFT'(x_j(t)) \tag{2.33}$$

donde $DFT(x(t))$ representa la Transformada Discreta de Fourier de la señal $x(t)$, y "'" significa complejo conjugado.

La DFT puede ser calculada de manera eficiente mediante la *Fast Fourier Transform* (FFT), por lo que (2.29) puede ser rescrita como:

$$C_{x_i x_j}(\tau) = \mathrm{Re}\left[IFFT(FFT(x_i(t)) \cdot FFT'(x_j(t)))\right] \tag{2.34}$$

en la que Re[ ] representa la parte real de la función compleja.

La CC es una técnica adecuada para el cálculo de TDOA en entornos 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 autocorrelación, lo cual afecta negativamente a estimación de la posición.

2.4.2.1.2 Correlación Cruzada Generalizada (GCC)

Una formulación más general de (2.25) se obtiene mediante la Correlación Cruzada Generalizada (GCC), que consiste en aplicar un filtro previo a cada una de las señales de entrada antes de calcular la CC [Castro 2007]. A la GCC se la denota como $C_{x_i x_j}^{(g)}(\tau)$:

$$C_{x_i x_j}^{(g)}(\tau) = E\left[\left(h_i(t) * x_i(t)\right) \cdot \left(h_j(t-\tau) * x_j(t-\tau)\right)\right] \tag{2.35}$$

La transformada de Fourier de la GCC se denomina *Cross-Power Spectrum (CSP)* y se denota como $C_{x_i x_j}^{(g)}(\omega)$:

$$C_{x_i x_j}^{(g)}(\omega) = \int_{-\infty}^{\infty} C_{x_i x_j}^{(g)}(\tau) e^{-j\omega\tau} d\tau \tag{2.36}$$

Sustituyendo (2.31) en (2.32) y aplicando la propiedad de convolución de la transformada de Fourier, la Cross-Power Spectrum puede ser rescrita en función de la transformada de Fourier de las señales de entrada como:

$$C_{x_i x_j}^{(g)}(\omega) = \left(H_i(\omega) X_i(\omega)\right) * \left(H'_j(\omega) X'_j(\omega)\right) \tag{2.37}$$

Definiendo la función de pesos como $\Phi_{x_i x_j}(\omega) = H_i(\omega) H'_j(\omega)$ y aplicando la transformada inversa de Fourier a (2.33) se obtiene:

$$C_{x_i x_j}^{(g)}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{x_i x_j}(\omega) X_i(\omega) X'_j(\omega) e^{-j\omega\tau} d\omega \tag{2.38}$$

La aplicación de la función de pesos adecuada provoca la aparición de un máximo en la GCC correspondiente al TDOA, el cual se calcula como:

$$\hat{\tau}_{ij} = \arg\max_{\tau} \hat{C}_{x_i x_j}^{(g)}(\tau) \tag{2.39}$$

En [Knapp 1976] se proponen diferentes funciones de peso, siendo el principal objetivo de las mismas la de enfatizar en la GCC el valor real del TDOA sobre los máximos relativos indeseados fruto del ruido y las reverberaciones. Algunas de las más importantes son:

- *Roth Processor*: Suprime las frecuencias donde el espectro de potencia de ruido es mayor.

$$\Phi_{x_i x_j}^{Roth}(\omega) = \frac{1}{C_{x_i x_j}(\omega)} \tag{2.40}$$

- *Smoothed Coherente Transform (SCOT)*: Mejora los resultados obtenidos con Roth ponderando la señal con pesos dependientes de su SNR.

$$\Phi_{x_i x_j}^{SCOT}(\omega) = \frac{1}{\sqrt{C_{x_i x_j}(\omega) C_{x_j x_j}(\omega)}} \tag{2.41}$$

- *Maximun Likelihood (ML) Estimator*: En esta función de pesos la señal $S(\omega)$ y el ruido $N(\omega)$ deben ser conocidos, por lo que en la práctica deberán ser estimados.

$$\Phi_{x_i x_j}^{ML}(\omega) = \frac{|S(\omega)|}{|N_i(\omega)||N_j(\omega)|} \cdot \left(1 + \frac{|S(\omega)|}{|N_i(\omega)|} + \frac{|S(\omega)|}{|N_j(\omega)|} \right)^{-1} \tag{2.42}$$

- *Phase Transform (PHAT)*: Esta función de pesos "blanquea" 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 sub-optima en entornos reverberantes.

$$\Phi_{x_i x_j}^{PHAT}(\omega) = \frac{1}{|X_i(\omega) X_j'(\omega)|} \tag{2.43}$$

Un defecto aparente de *PHAT* consiste en que la ponderación de la señal de manera inversa a su módulo provoca que los errores se acentúen en situaciones en los que la potencia de señal es pequeña. Para tratar de solucionar este problema [Dibiase 2000] pag. 46, propone la utilización de un filtro paso banda (300Hz, 6KHz) en conjunción con *PHAT* de tal manera que se enfaticen únicamente las bandas de frecuencia en las que hay mayor potencia de señal acústica.

Diferentes estudios realizados por la comunidad científica como por ejemplo [Svaizer 1997], revelan que el método GCC-PHAT es el que presenta un mejor

comportamiento en situaciones reales, por lo que se dedicará el siguiente apartado a estudiar los detalles de su implementación ([Dibiase 2000] y [Castro 2007]).

2.4.2.1.3 Implementación de GCC-PHAT

A la hora de la implementación del algoritmo se deben tener en cuenta las siguientes consideraciones:

- No se pueden tener observaciones temporales infinitas, por lo que se toman ventanas temporales asumiendo que la localización de cada target se mantiene estacionaria durante la duración de la misma.
- Para el análisis se empleará una versión discretizada, fruto del muestreo de la señal analógica original.

Por lo tanto, cualquier técnica de localización comenzará segmentando la señal en bloques, a cada uno de los cuales se le aplica una ventana temporal (Rectangular, Bartlett, Hanning, Hamming o Blackman) con el objetivo de mejorar la representación espectral de las señales, de entre las que la Hamming demuestra ser la que presenta un mejor comportamiento a nivel práctico (2.4.2.4.4).

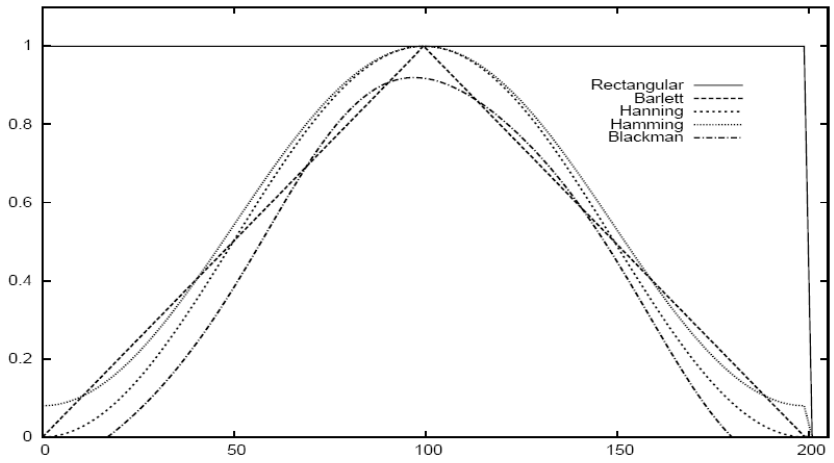


Figura 2.11 Ventanas temporales típicas

Es una práctica habitual el que los bloques de datos consecutivos estén solapados en el dominio del tiempo, de tal manera que los datos ubicados al final de cada bloque, y por tanto suprimidos por la ventana temporal, estén situados en el centro del siguiente bloque, sufriendo por tanto la menor atenuación. De esta manera se

garantiza que todos los que todos los datos tienen la misma importancia dentro del análisis. El uso de ventanas temporales es también necesario para reducir los efectos de distorsión en el espectro de la señal derivados de la longitud finita de los frames.

Otra consideración importante a la hora de la elección de la ventana temporal es el efecto sobre el patrón de directividad, por el cual para una ventana rectangular se obtiene una mayor anchura del lóbulo principal con unos lóbulos laterales no despreciables, mientras que para otros tipos de ventanas (e.g. Hamming) los efectos son los inversos, tal y como se puede observar en :

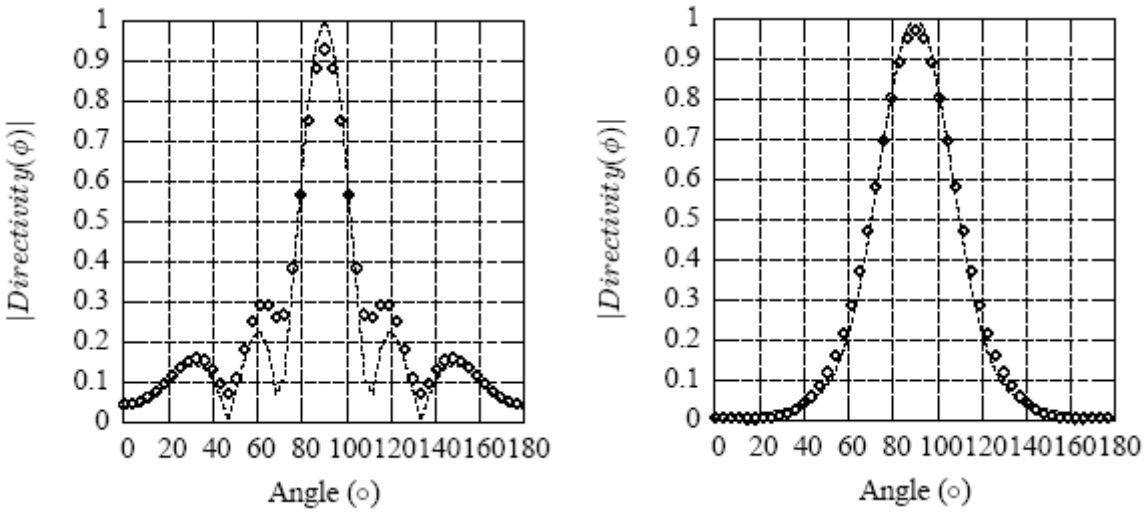


Figura 2.12 Patrón de directividad de un array de micrófonos con ventanas (a) rectangular (b) Hamming

Tras esto se calcula la DFT de cada uno de los bloques, de tal manera que el cómputo de cada uno de ellos nos permite trackear la posición del target objetivo. La frecuencia con la que el algoritmo produce estimaciones de la posición depende del avance de los bloques de datos, denominado *frame shift*, mientras que la latencia de cada una de las estimaciones depende del tamaño del bloque o *frame size.*

Por lo que si $x_1[n]..x_M[n]$ representan las señales discretas en cada uno de los micrófonos del array y dichas señales de entrada son segmentadas en bloques de longitud N, la señal del m-esimo micrófono para el bloque b, $x_{m,b}[n]$, se puede escribir como:

$$x_{m,b}[n] = \omega[n] \cdot x_m[bA + n] \quad para \quad n = 0...N-1 \tag{2.44}$$

donde $\omega[n]$ es la ventana temporal y A es el frame shift. Los bloques contiguos se solapan cuando $A \prec N$, escogiéndose típicamente $A = N/2$ por los motivos señalados anteriormente.

Por lo tanto, la DFT de (2.44) puede expresarse como:

$$X_{m,b}[k] = \sum_{n=0}^{N-1} x_{m,b}[n] \cdot e^{-jk\frac{2\pi}{K}n} \quad para \quad k = 0...K-1 \tag{2.45}$$

Obsérvese que la longitud de la DFT es K con $K \geq N$, por lo que cada bloque de datos se le debe aplicar un *zero-padded* para incrementar su longitud antes de realizar la DFT. Generalmente se suele elegir un K múltiplo de 2 que permita la obtención de la DFT a través de la Fast Fourier Transform (FFT). El zero-padding también es necesario a la hora de considerar la propiedad de desplazamiento, circular convolución circular y resolución espectral de la DFT.

Por lo tanto, sustituyendo (2.45) en (2.38) se obtiene $\hat{C}_{ij,b}(\hat{\tau})$, que representa la GCC-PHAT obtenida a partir de la DFT entre los micrófonos i y j en el bloque de datos b:

$$\hat{C}_{ij,b}(\hat{\tau}) = \frac{1}{K}\sum_{k=0}^{K-1} \Phi_{ij}[k] X_{i,b}[k] X'_{j,b}[k] e^{jk\frac{2\pi}{K}\hat{\tau}} = \frac{1}{K}\sum_{k=0}^{K-1} \Phi_{ij}[k] C_{ij,b}[k] e^{jk\frac{2\pi}{K}\hat{\tau}} \tag{2.46}$$

donde $\Phi_{ij}[k]$ representa la versión discreta de la función de pesos $\Phi_{ij}[\omega]$.

Teniendo en cuenta la relación (2.34) se obtiene una forma alternativa para calcular $\hat{C}_{ij,b}(\hat{\tau})$:

$$\hat{C}_{ij,b}(\hat{\tau}) = \mathrm{Re}\left[IFFT\left(\Phi_{ij}[k] \cdot X_{i,b}[k] X'_{j,b}[k]\right)\right](\hat{\tau}) = \mathrm{Re}\left[IFFT\left(\frac{X_{i,b}[k] X'_{j,b}[k]}{\left|X_{i,b}[k]\right| \left|X'_{j,b}[k]\right|} \right)\right](\hat{\tau}) \tag{2.47}$$

La expresión (2.47) muestra la forma de implementar GCC-PHAT a partir de las FFT de las señales capturadas por los micrófonos.

El tiempo en el que la expresión anterior se hace máxima constituirá la estimación de TDOA entre los micrófonos i y j para el bloque b.

$$\hat{\tau}_{ij,b} = \arg\max_{\tau} \hat{C}_{ij,b}(\hat{\tau}) \tag{2.48}$$

Por otro lado es importante reseñar que la distancia entre micrófonos, d, limita el rango de TDOA válidos de forma que $\tau \in [-d/c, d/c]$. A este respecto se debe hacer una observación importante relativa a la pérdida de precisión inherente al método GCC-PHAT basado en la DFT, debido al traslado de las unidades de tiempo τ, de la ecuación (2.34) a las muestras discretas $\hat{\tau}$ de (2.47) según se muestra a continuación:

$$\hat{\tau} = round\left(\tau[\sec] f_s[\sec^{-1}]\right) \tag{2.49}$$

Por lo que la ecuación (2.8), bajo la presunción de campo lejano, se puede rescribir para tiempo discreto como:

$$\hat{\tau} = round\left(\frac{d\cos(\phi)}{c} \cdot f_s\right) \tag{2.50}$$

A partir de 2.50 se deduce que la imprecisión cometida depende de diversos factores:

- *La distancia entre micrófonos*: A mayor distancia entre micrófonos menor error.
- *La frecuencia de muestreo:* En cuanto mayor sea, menor será el error. En base a esto aparecen distintas técnicas que tratan de aumentar artificialmente la frecuencia de muestreo, como por ejemplo como [Varma 2002] pag. 60, donde este aumento se realiza añadiendo ceros a la FFT (interpolación en frecuencia).
- *Dirección de llegada,* ϕ. A mayor ángulo, menor error. En el caso particular de $\phi = \pi/2 \rightarrow \tau = 0$ no se comete error de redondeo.
- *Función de redondeo*: Existen diferentes métodos como el redondeo al τ más cercano a cero, el redondeo al τ más cercano o la interpolación ([Castro 2007] Pág. 38).

2.4.2.2 Basados en High-resolution Spectral Estimation

El objetivo primero de este grupo de técnicas es la estimación de la distribución de potencia de la señal, para a partir de ella poder detectar los picos de potencia presentes. Estos 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.

La idea de dividir la señal en dos subespacios fue profundamente estudiada en [De Moor 1993] y ha dado lugar a una amplia gama de algoritmos de banda estrecha, siendo alguno de los más importantes:

- *MUSIC (MUltiple SIgnal Classification)*: [Schmidt 1986].
- *ESPRIT (Estimation of Signal Parameters via Rotacional Invariante Techniques):* [Roy 1989].
- *MIN-NORM (MINimun NORMs):* [Kumaresan 1983].
- *WSF (Weighted Subspace Fitting):* [Viberg 1991].

Estudios interesantes a este respecto fueron realizados por [Benesty 2000], donde se 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 2.13 muestra la configuración del sistema empleado en [Benesty 2000], el cual emplea altavoces ubicados en posiciones estáticas para generar las señales acústicas y un par de micrófonos separados 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.

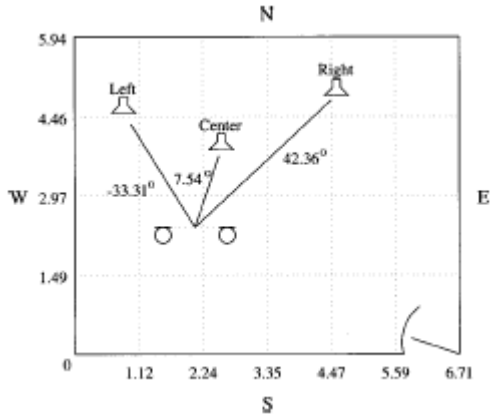


Figura 2.13 Configuración del sistema en los experimentos de [Benesty 2000]

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:

1. 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 por [Benesty 2000] 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.
2. Este conjunto de técnicas funciona 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, siendo uno de los más importantes el gran aumento de la carga computacional.

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 [Wang 1985] y desarrollado en [Hung 1988] y [Friedlander 1993]. 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 comparte (aproximadamente) 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.

2.4.2.3 Basados en Steered Response Power (SRP)

Muchas de las técnicas de procesamiento de señales digitales se basan en la capacidad de los arrays de micrófonos de apuntar a una dirección o posición determinadas de espacio.

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. Cuando la localización de la fuente no es conocida, 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* [Dibiase 2000].

La forma más simple de implementar una steered response consiste en la utilización de la salida de un delay-and-sum beamformer (2.3.3.1), en la que los diferentes retardos temporales (calculados en función de la posición del espacio a la que deseamos apuntar) son aplicados a la señal de entrada de cada uno de los micrófonos, de tal manera que se produce una alineación temporal de las mismas, tras lo cual se suman para obtener la salida del algoritmo. Un procedimiento más general, denominado filter-and-sum beamformer (2.3.3.2) aplica un filtro a cada una de las señales de entrada a los micrófonos como paso previo al alineamiento y posterior suma de las mismas.

En el caso de que el beamforming se aplique en contextos de captura de señales de voz (e.g. [Chou 1995]), los filtros no solamente deben suprimir el ruido y las señales indeseadas, sino que deben prestar especial atención en no distorsionar significativamente la señal deseada. Sin embargo, cuando el objetivo final es el de localización, los filtros únicamente deben tratar de trasladar la mayor cantidad de potencia posible de la señal deseada a la salida del beamformer, por lo que en este tipo de aplicaciones es muy interesante el uso de la *Phase Transform (PHAT) (2.4.2.1.2*) que

ha demostrado su buen comportamiento en cuanto a estimación de TDOA a pesar de que distorsiona las señales de entrada.

En apartados anteriores (2.4.2.1.2) se apuntó que las estimaciones de TDOA obtenidas a partir de GCC no proporcionan estimaciones suficientemente robustas bajo condiciones de baja SNR y entornos reverberantes. A este respecto en [Dibiase 2000] pag 83-84 se propone el aumento del número de micrófonos utilizados para la obtención de la estimación (empleo de más de 2) como un mecanismo que permita mejorar dichos resultados. Considerando los puntos reseñados anteriormente, en [Brandstein 2001] pag. 164-178 se propone una técnica robusta de localización basada en la equivalencia entre SRP y la suma de las GCC-PHAT de todas las posibles combinaciones de parejas de micrófonos. A esta técnica se la conoce 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 cubren dos de los objetivos más importantes presentados en 1.2 como de obligado cumplimiento:

- Posibilidad de aceptar un número variable y múltiple de targets simultáneos [Wax 1983].
- Alta robustez y fiabilidad, siendo en este caso superior a la proporcionada por los métodos basados en TDOA [Varma 2002] pag. 122-123.

La principal desventaja de este algoritmo reside en la alta carga computacional, la cual se ve incrementada con el aumento del número de micrófonos y con el número de localizaciones espaciales a las que se debe orientar el array (número de targets). Adicionalmente, el tamaño del grid (2.4.2.4.5) de búsqueda utilizado limita la precisión del algoritmo [Duraiswami 2001].

2.4.2.3.1 Algoritmo SRP-PHAT

La ecuación (2.22), representada gráficamente en la Figura 2.7, presenta la salida de un filter-and-sum beamformer, que expresado en función de ω queda como sigue:

$$Y(\omega,\vec{q}) = \sum_{n=1}^{M} W_n(\omega) X_n(\omega) e^{j\omega\Delta_n} \tag{2.51}$$

donde Δ_n es el retardo adecuado para enfocar el micrófono n a la posición del espacio $\vec{q}$ y $X_n(\omega)$ y $W_n(\omega)$ representan las transformadas de Fourier de la señal del n-esimo micrófono y su filtro asociado.

El equivalente de la expresión anterior en el dominio temporal puede ser utilizado para dirigir el array hacia puntos de interés específicos y analizar la potencia obtenida a la salida para cada uno de ellos. Cuando en el punto enfocado exista realmente un target activo, SRP debería proporcionar un máximo global. La salida del algoritmo SRP para una ubicación espacial específica $\vec{q}$ puede ser expresada como la potencia de salida del filter-and-sum beamformer [Dibiase 2000], obteniéndose:

$$P(\vec{q}) = \int_{-\infty}^{\infty} |Y(\omega)| |Y(\omega)|' \, d\omega = \int_{-\infty}^{\infty} |Y(\omega)|^2 \, d\omega \tag{2.52}$$

Por lo que la estimación de la localización $\hat{\vec{q}}_s$ se obtiene como:

$$\hat{\vec{q}}_s = \arg\max_{\mathrm{q}} P(\vec{q}) \tag{2.53}$$

Sin embargo, la función de potencia así definida puede presentar picos en localizaciones incorrectas debido a reverberaciones o ruido, lo cual conduciría a errores de estimación. La elección del filtro correcto puede ayudar a reducir este tipo de efectos, por lo que teniendo en cuenta 2.4.2.1.2 parece adecuado utilizar PHAT como filtro de entrada en cada uno de los micrófonos. De esta manera se unen las ventajas proporcionadas por SRP con la robustez proporcionada por PHAT, obteniéndose el algoritmo SRP-PHAT propuesto por primera vez en [Dibiase 2000] pag. 80-82 y que se formula tal y como se muestra a continuación:

$$P(\vec{q}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{-\infty}^{\infty} \Phi_{ij}(\omega) X_i(\omega) X_j'(\omega) e^{jw(\Delta_j - \Delta_i)} d\omega \quad \forall i \neq j \text{ y } \mathrm{j} \succ \mathrm{i} \tag{2.54}$$

donde $\Delta_j - \Delta_i = \tau_{ij}$ representa el TDOA entre los micrófonos i y j para la fuente sonora ubicada espacialmente en $\vec{q}$ y la función de pesos $\Phi_{ij}(\omega)$ que utiliza la transformada de fase (PHAT) como filtro asociado, obteniéndose:

$$\Phi_{ij}(\omega) = W_i(\omega)W_j'(\omega) = \frac{1}{\left|X_i(\omega)X_j'(\omega)\right|} \Leftrightarrow W_n(\omega) = \frac{1}{\left|X_n(\omega)\right|} \quad (2.55)$$

2.4.2.3.2 SRP como función de la GCC

Comparando las expresiones (2.34) y (2.50) para GCC y SRP respectivamente (presentada aquí nuevamente por claridad)

$$C_{x_i x_j}^{(g)}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{x_i x_j}(\omega)X_i(\omega)X_j'(\omega)e^{-j\omega\tau}d\omega \qquad \textbf{GCC}$$

$$P(\vec{q}) = \sum_{i=1}^{M}\sum_{j=1}^{M}\int_{-\infty}^{\infty} \Phi_{ij}(\omega)X_i(\omega)X_j'(\omega)e^{jw(\Delta_j-\Delta_i)}d\omega \quad \forall i \neq j \text{ y j} \succ \text{i} \quad \textbf{SRP}$$

se observa que SRP de un array de 2 elementos es equivalente a la GCC de esos 2 micrófonos. Por tanto, según se incrementa el número de micrófonos, SRP se extiende de manera natural a suma de la GCC de todas las posibles parejas de micrófonos.

Por lo tanto, si nombramos $C_{ij}(\tau_{ij})$ a la GCC-PHAT de las señales de los micrófonos i y j "apuntando" a una posición $\vec{q}$ en la que el retardo entre las señales capturadas por ambos es τ_{ij}, la expresión de SRP-PHAT en el dominio temporal puede ser rescrita como:

$$P(\vec{q}) = P(\Delta_1 ... \Delta_M) = 2\pi\sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}(\Delta_j - \Delta_i) = 2\pi\sum_{i=1}^{M}\sum_{j=1}^{M} C_{ij}(\tau_{ij}) \quad \forall i \neq j \text{ y j} \succ \text{i} \quad (2.56)$$

donde $\Delta_1 ... \Delta_M$ son los retardos apropiados para apuntar el array hacia la posición $\vec{q}$.

Se demuestra así que SRP es equivalente a la suma de la GCC-PHAT de todas las posibles combinaciones de parejas de micrófonos, siendo equivalentes y no iguales debido a que:

- El sumatorio presentado en (2.52) incluye la suma de las M autocorrelaciones, es decir la correlación de una señal consigo misma para un retardo cero. Esta componente añade únicamente un DC offset respecto a SRP, ya que dicho término es independiente del los retardos aplicados. Esta componente se elimina si se toma $i \neq j$.

- Dicha ecuación incluye las GCC del micrófono i con el j y viceversa, lo cual es equivalente a un escalado por 2 ($C_{ij}(\Delta_j - \Delta_i) = C_{ji}(\Delta_i - \Delta_j)$) con respecto a SRP, el cual sólo incluye una de las dos posibilidades. Estas componentes se eliminan si se toma $i \succ j$.

Por lo tanto se llega a la conclusión de que SRP es, con un factor de escalado y un offset constante, equivalente a la suma de la GCC-PHAT de todas las posibles combinaciones de parejas de micrófonos.

2.4.2.3.3 Implementación de SRP-PHAT

El procedimiento a seguir en la implementación de SRP-PHAT guarda gran cantidad de similitudes con el presentado para GCC-PHAT en 2.4.2.1.3 respecto a que debemos segmentar la señal en bloques, a cada uno de los cuales se le aplicará una ventana temporal (típicamente Hamming) y que los bloques contiguos se solapan en el dominio temporal para evitar los efectos de las discontinuidades.

La ecuación (2.52) define SRP en términos de GCC. Sustituyendo la expresión de GCC presentada en (2.42) en (2.52) obtenemos la estimación de SRP del bloque b, en un array de M micrófonos, enfocado hacia la posición definida por los retardos temporales $\Delta_1 ... \Delta_M$ en el dominio temporal $\hat{P}_b(\Delta_1 ... \Delta_M)$, y conocida como TSRP.

$$\hat{P}_b\left(\hat{\Delta}_1 ... \hat{\Delta}_M\right) = 2\pi \sum_{i=1}^{M}\sum_{j=1}^{M} \hat{C}_{ij,b}\left(\hat{\tau}_{ij}\right) = 2\pi \sum_{i=1}^{M}\sum_{j=1}^{M} \frac{1}{K}\sum_{k=0}^{K-1} \Phi_{ij}[k] X_{i,b}[k] X'_{j,b}[k] e^{jk\frac{2\pi}{K}\hat{\tau}_{ij}} \qquad (2.57)$$

Aplicando (2.43) podemos expresar TSRP en términos de la FFT como:

$$\hat{P}_b\left(\hat{\Delta}_1 ... \hat{\Delta}_M\right) = \sum_{i=1}^{M}\sum_{j=1}^{M} \mathrm{Re}\left[IFFT\left(\frac{X_{i,b}[k] X'_{j,b}[k]}{\left|X_{i,b}[k]\right| \left|X'_{j,b}[k]\right|} \right) \right] \left(\hat{\tau}_{ij}\right) \qquad (2.58)$$

Un efecto de imprecisión importante que afecta a TSRP es la *discretización*, por el que los retardos de tiempo continuos de (2.52) se transforman en retardos discretos en muestras en (2.53), los cuales están ligeramente desplazados respecto a los continuos. Para tratar de reducir la imprecisión producida por la discretización, en 2.4.2.1.3 pag. 35

se presentan diferentes técnicas de redondeo a partir de las cuales se puede predecir la información existente entre muestras consecutivas de la GCC.

La imprecisión de TSRP comentada en el párrafo anterior nos lleva a la búsqueda de una implementación alternativa que no presente dichos efectos, apareciendo la versión en el dominio de la frecuencia de SRP (FSRP), presentada en (2.50) y mostrada aquí nuevamente por claridad.

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

Aquí se puede apreciar como los retardos temporales en muestras aplicados por TSRP son sustituidos en el dominio de la frecuencia por multiplicaciones por exponenciales complejas evaluadas en los retardos apropiados en unidades reales y por lo tanto sin sufrir los efectos de la discretización.

A este respecto, los experimentos prácticos realizados en [Castro 2007] pag. 73 y 123, demuestran en la práctica la equivalencia teórica entre TSRP y FSRP presentada anteriormente. La diferencia principal estriba en que FSRP es más costoso a nivel de carga computacional que TSRP, lo cual justifica la utilización de esta última técnica a pesar de los errores de estimación debidos al redondeo que esta presenta.

2.4.2.4 Efecto de los distintos parámetros en SRP-PHAT

Como se ha podido comprobar en apartados anteriores del presente estado del arte, existen una gran cantidad de parámetros involucrados en la implementación práctica de un array de micrófonos y sus algoritmos de localización y tracking asociados. El principal objetivo de este apartado reside en presentar dicha dependencia, de manera que facilite la elección práctica de los mismos.

Para ello, se han seleccionado dos parámetros (MOTP[9] y A-MOTA[10]) de los propuestos en el proyecto CHIL dentro de su plan de evaluación [Mostefa y otros 2006], los cuales serán empleados para evaluar la dependencia del algoritmo para cada uno de los parámetros concretos. Para que los parámetros anteriores queden perfectamente definidos es necesario destacar que el proyecto CHIL clasifica los errores en dos grandes grupos: *Fine and Gross errors* considerándose que un error pertenece al primer grupo cuando la diferencia entre la estimación generada por el algoritmo y la posición real del target, o *ground truth*, es menor que 500 mm y al segundo en caso contrario.

Los datos numéricos presentados en los siguientes apartados han sido extraídos de [Castro 2007], a partir de los resultados extraídos de la experimentación con las bases de datos IDIAP AV16.3 [Gatica-Perez 2004] y EDECAN HIFI-MM1. El objetivo de su introducción dentro del presente estado del arte es el de permitir cuantificar el efecto que tiene la variación de cada uno de los parámetros, y así servir de ayuda a la hora de seleccionar los mismos en desarrollos que emplean arrays de micrófonos en tareas de seguimiento como la que se desea implementar en la tesis que aquí comienza.

2.4.2.4.1 Frecuencia de muestreo y técnicas de interpolación

La Tabla 2.1 muestra los resultados de la variación de la frecuencia de muestreo obtenidos bajo las siguientes condiciones:

- **BBDD:** HIFI-MM1
- **Array:** Lineal 4 elementos
- **Separación entre mics:** 200 mm
- **Frame size:** 320 ms
- **Tamaño grid de búsqueda:** 150 mm
- **Frecuencia de muestreo:** Variable

Efecto de la frecuencia de muestreo

[9] MOTP (*Multiple Object Tracking Precision)*: Representa la precisión del sistema en mm y se calcula como la distancia total euclidea entre la ground truth y la posición estimada, considerando únicamente las estimaciones con error fino. Muestra la habilidad del algoritmo a la hora de estimar la posición independientemente de los errores de tracking.

[10] A-MOTA (*Audiovisual Multiple Object Tracking Accuracy*): Representa la precision de tracking en % y se calcula como:

$$A-MOTA = 1 - \frac{gross\_errors + deletions}{\#ground\_truths}$$

Parámetro	Unidades	fs=16 KHz	fs=16x3 KHz	fs=48 KHz
MOTP	[mm]	230	209	206
	Red. error rel [%]		9.1	10.4
A-MOTA (%)	[%]	11 ± 0.6	51 ± 1	71 ± 0.9
	Red. error rel [%]		363.6	545.5

Tabla 2.1 Efectos de la frecuencia de muestreo

Como era de esperar, el aumento de la *frecuencia de muestreo* hace mejorar los resultados obtenidos por el algoritmo, lo cual se debe a que dicho aumento nos permite realizar una conversión de tiempo continuo a tiempo discreto más precisa.

Es también interesante observar, como la interpolación x3 realizada mejora las prestaciones del algoritmo, aunque sin alcanzar nunca los resultados obtenidos para los 48KHz originales.

Para reducir los efectos negativos de la conversión de tiempo continuo a tiempo discreto existen 3 técnicas principales. La Tabla 2.2 muestra los efectos de cada una, obtenidas bajo las mismas condiciones mostradas al inicio del apartado.

Efecto de las Técnicas de interpolación					
Parámetro	**Unidades**	**No rounding**	**Rounding**	**Interpolation**	**FSRP**
MOTP	[mm]	171	230	230	198
	Red. error rel [%]		-34.5	-34.5	-15.8
A-MOTA (%)	[%]	5 ± 0.4	11 ± 0.6	40 ± 1.0	58 ± 1.0
	Red. error rel [%]		6.3	36.8	55.8

Tabla 2.2 Efectos de las técnicas de interpolación

Como era de esperar FSRP proporciona los mejores resultados al operar con los retardos exactos para dirigir el array a un punto concreto del espacio 2.4.2.3.3. Sin embargo la principal desventaja de este método, y que en muchos casos impide su implementación práctica, es la elevada carga computacional (es necesario realizar tantas multiplicaciones complejas como puntos de la FFT por el número de puntos donde se desee apuntar el array).

De lo anterior se deduce que la interpolación y el redondeo son de alguna manera técnicas equivalentes, orientadas ambas al aumento de la precisión de los retardos aplicados, aunque se demuestra que sus efectos no son aditivos [Castro 2007] pág. 78. Sin embargo sus efectos en cuanto a carga computacional son opuestos, ya que

mientras que la interpolación hace aumentar el número de operaciones a computar en la FFT (una interpolación x2 prácticamente duplica el tiempo de proceso), el redondeo no requiere de aumento de operaciones en la FFT y por tanto mantiene constante el tiempo de ejecución.

2.4.2.4.2 Frame size

En este punto se trata de demostrar hasta que punto el aumento del frame size continua mejorando la precisión del algoritmo, así como evaluar el efecto de dicho aumento cuando se trata con targets móviles. Las condiciones bajo las que se desarrollan dichos estudios se muestran a continuación:

- **BBDD:** AV16.3
- **Frame size:** Variable
- **Array:** Circular 8 elementos
- **Tamaño grid de búsqueda:** 150 mm
- **Radio del array:** 100 mm
- **Frecuencia de muestreo:** 16 KHz

En un primer experimento se desea evaluar el efecto del frame size en el caso de tratar con *targets estáticos*. Los resultados se presentan en la Tabla 2.3.

Efecto de frame size con targets estáticos				
Parámetro	**Unidades**	**F. size=40 ms**	**F. size=320 ms**	**F. size=640 ms**
MOTP	[mm]	197	191	190
	Red. error rel [%]		3.0	3.6
A-MOTA (%)	[%]	61 ± 1.1	74 ± 1.0	87 ± 0.8
	Red. error rel [%]		21.3	42.6

Tabla 2.3 Efectos del frame size con targets estáticos

Los resultados presentados muestran la previsible mejora en los resultados del algoritmo con el aumento del frame size, si bien dicha mejora conlleva un aumento del tiempo de proceso debido al aumento del número de operaciones a realizar por la FFT.

A continuación se trata de realizar un estudio análogo al anterior, pero ahora con targets móviles (secuencia identificada como seq11 dentro de la BBDD AV16.3). Los resultados obtenidos se presentan en la Tabla 2.4.

Efecto de frame size con targets en movimiento (seq 11)				
Parámetro	**Unidades**	**F. size=40 ms**	**F. size=320 ms**	**F. size=640 ms**
MOTP	[mm]	201	214	258

	Red. error rel [%]		-6.5	-25.9
A-MOTA (%)	[%]	50 ± 4.5	74 ± 3.9	67 ± 4.2
	Red. error rel [%]		48.0	34.0

Tabla 2.4 Efectos del frame size con targets en movimiento

En este caso, a diferencia del caso estático, los resultados mejoran al pasar de frames de 40 ms a frames de 320, pero sin embargo empeoran al pasar de estas últimas a frames de 640. Esto se debe al movimiento de los targets en la grabación bajo estudio, ya que si tomamos tramas demasiado largas la distancia recorrida durante la misma no será despreciable produciendo el consiguiente error de estimación.

Unos cálculos básicos pueden ser realizados para ilustrar este efecto [Castro 2007]. Tal y como se comentó anteriormente, considerando que un error como *gross* cuando la distancia entre nuestra estimación y la ground truth es mayor que 500 mm. Considerando que la velocidad normal de una persona al caminar es de $5\,Km/h$ se obtiene:

$$500mm \cdot \frac{1m}{1000mm} \cdot \frac{1h}{5km} \cdot \frac{1km}{1000m} \cdot \frac{3600s}{1h} = 0.36s \qquad (2.59)$$

De (2.55) se deduce que para frame size de duración mayor que 360 ms los resultados obtenidos comienzan a empeorar.

2.4.2.4.3 Longitud de la FFT

En teoría, una mayor longitud de la FFT debería redundar en una representación más precisa del espectro de la señal. Sin embargo la Tabla 2.5, realizada bajo las premisas presentadas a continuación, concluye que no se produce ningún tipo de mejora en los resultados (siempre que la longitud de la FFT sea mayor que el Frame Size):

- **BBDD:** AV16.3
- **Frame size:** 512 ms
- **Array:** Circular 8 elementos
- **Tamaño grid de búsqueda:** 150 mm
- **Radio del array:** 100 mm
- **Frecuencia de muestreo:** 16 KHz

Efecto del frame size			
Parámetro	**Unidades**	**FFT size=8192**	**FFT size=16384**

MOTP	[mm]	223	224
	Red. error rel [%]		-0.4
A-MOTA (%)	[%]	76 ± 1.0	76 ± 1.0
	Red. error rel [%]		0.0

Tabla 2.5 Efectos del frame size con targets estáticos

Sin embargo una limitación importante impuesta por la longitud de la FFT es la de la distancia máxima a la que puede encontrarse el target objetivo para ser localizable por el algoritmo, ya que dicha longitud de la FFT define la longitud de la función de correlación en la cual se realiza la búsqueda de los valores máximos de potencia, en los cuales se localizan retardos temporales (en muestras) en los que se espera encontrar los targets de la escena.

Es por tanto obvio que para una longitud de la FFT N, es posible realizar la búsqueda anterior únicamente dentro del rango de retardos temporales comprendidos en $\left[-\frac{N-1}{2}, \frac{N}{2}\right]$, el cual puede ser trasladado a distancia máxima del target como:

$$\frac{N/2[muestras]\cdot c[m/s]}{f_s[1/s]} = dist[m] \tag{2.60}$$

Los casos más restrictivos a este respecto se producen para los menores frame size. Para un ejemplo concreto en el que se ha tomado un frame size de 40 ms y una $f_s = 16KHz$ se obtiene:

$$M_{\min} \cdot f_s = 40ms \cdot 16KHz = 640\ muestras$$

$$Condiciones: (\mathrm{N} \geq \mathrm{M}) \wedge N = 2^n \rightarrow N_{\min} = 1024 \text{ y retardo en muestras} \in [-511, 512]$$

$$dist = \frac{512[muestras]\cdot 345[m/s]}{16000[1/s]} = 11.04$$

2.4.2.4.4 Ventana temporal

La tarea de dividir la señal capturada en tramos para su posterior procesado puede realizarse de manera directa (equivalente a aplicar una ventana rectangular), o mediante otro tipo de ventanas temporales (Figura 2.11) que prevengan los posibles problemas que pueden aparecer en el dominio de la frecuencia debido a una terminación excesivamente abrupta de la de la señal en el dominio temporal.

- **BBDD:** AV16.3
- **Array:** Circular 8 elementos
- **Radio del array:** 100 mm
- **Frame size:** 500ms
- **Tamaño grid de búsqueda:** 150 mm
- **Frecuencia de muestreo:** 16 KHz

Efecto de las ventanas temporales						
Parámetro	**Unidades**	**Rectangular**	**Hamming**	**Hanning**	**Blackmann**	**Barlett**
MOTP	[mm]	224	224	223	224	224
	Red. error rel [%]		0.0	0.4	0.0	0.0
A-MOTA (%)	[%]	76 ± 1.0	76 ± 1.0	72 ± 1.0	71 ± 1.0	47 ± 1.1
	Red. error rel [%]		0.0	-5.3	-6.6	-38.2

Tabla 2.6 Efectos de las ventanas temporales

2.4.2.4.5 Grid de búsqueda

La forma habitual de funcionamiento de SRP comienza con una división del espacio mediante un grid, el cual define una serie de puntos en los que el algoritmo buscara targets activos, de lo que se deduce la importancia de la correcta elección de la separación entre dichos puntos para lograr un correcto funcionamiento. De antemano parece lógico pensar que a menor tamaño del grid menores errores se cometerán, lo cual queda refrendado en la Tabla 2.7.

- **BBDD:** AV16.3
- **Array:** Circular 8 elementos
- **Radio del array:** 100 mm
- **Frame size:** 640ms
- **Tamaño grid de búsqueda:** Variable
- **Frecuencia de muestreo:** 16 KHz

Efecto del tamaño del grid de búsqueda I						
Parámetro	**Unidades**	**50 mm**	**100 mm**	**150 mm**	**200 mm**	**250 mm**
MOTP	[mm]	99	156	192	228	255
	Red. error rel [%]		-57.6	-93.9	-130-3	-157.6
A-MOTA (%)	[%]	-93 ± 0.6	91 ± 0.7	90 ± 0.7	78 ± 1.0	75 ± 1.1
	Red. error rel [%]		-2.2	-3.2	-16.1	-19.4

Tabla 2.7 Efecto del tamaño del grid de búsqueda I

La Tabla 2.7 confirma las estimaciones iniciales respecto a los efectos de la variación del grid de búsqueda. Sin embargo resulta de sumo interés observar los

resultados mostrados en la Tabla 2.8, obtenidos con una BBDD y un entorno diferente a la anterior:

- **BBDD:** HIFI-MM1
- **Array:** Lineal 4 elementos
- **Separación entre mics:** 200 mm
- **Frame size:** 320 ms
- **Tamaño grid de búsqueda:** Variable
- **Frecuencia de muestreo:** 48 KHz

Efecto del tamaño del grid de búsqueda II						
Parámetro	**Unidades**	**250 mm**	**150 mm**	**100 mm**	**50 mm**	**25 mm**
MOTP	[mm]	216	196	193	203	228
	Red. error rel [%]		9.3	10.6	6.0	-5.6
A-MOTA (%)	[%]	57 ± 1.0	76 ± 0.9	41 ± 1.0	37 ± 1.0	37 ± 1.0
	Red. error rel [%]		33.3	-28.1	-35.1	-35.1

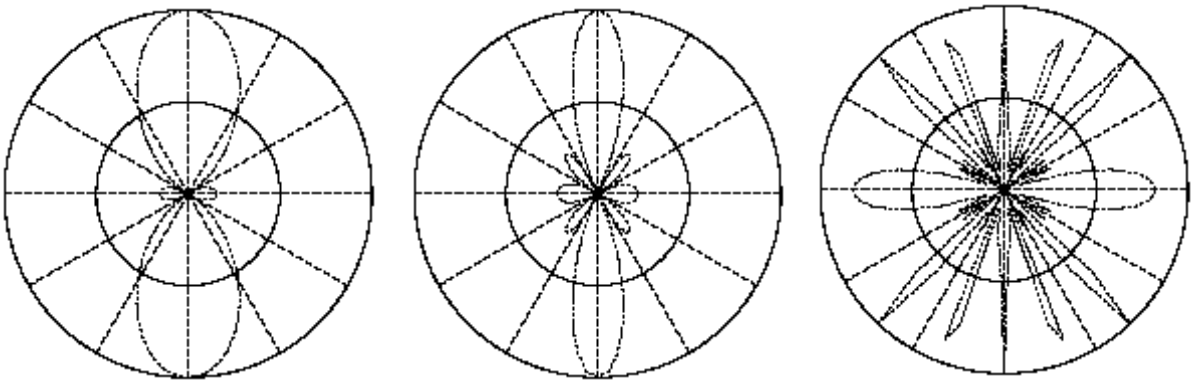
Tabla 2.8 Efecto del tamaño del grid de búsqueda II

Como se puede observar los resultados obtenidos presentan una mejora al pasar de 250 mm a 150 (como cabría esperar), pero sin embargo a partir de ese valor, la reducción del grid de búsqueda tiene un efecto contrario al esperado aumentando los errores.

El efecto presentado anteriormente se explica en base al pobre patrón de directividad del array de micrófonos utilizado para la captura de los datos, el cual se debe una gran separación entre micrófonos que provoca que según aumenta la frecuencia de la señal (con la consiguiente reducción de la anchura del lóbulo principal 2.3.2) se produzca aliasing espacial (2.3.3) y con el aparezcan grandes lóbulos en direcciones indeseadas del patrón de directividad.

A partir de (2.13) podemos obtener la frecuencia a partir de la cual se produce aliasing espacial, lo cual queda refrendado en la Figura 2.13, extraída de [Castro 2007] pags. 100 y 101, y que representan el patrón de directividad del array utilizado para diferentes frecuencias

$$d \le \frac{\lambda_{\min}}{2} \rightarrow f_{\max} \le \frac{c}{2d} = \frac{345[m/s]}{0.4[m]} = 862.5Hz$$



(a) (b) (c)

Figura 2.14 Patrón de directividad del array lineal de 4 elementos equiespaciados 200mm para frecuencias de (a) 500Hz (b) 1000Hz (c) 5000Hz

De lo presentado anteriormente se desprende que en el array aparecen dos efectos claramente diferenciados que dificultan las estimaciones de nuestro algoritmo:

1. Para las frecuencias bajas evitamos los efectos del aliasing espacial, pero la gran anchura del lóbulo principal impide un apuntamiento preciso del array. Debido a esta anchura excesiva el array integra información procedente de zonas del espacio demasiado grandes, no permitiendo al algoritmo discriminar la existencia de targets en zonas cercanas del espacio.

2. Para las frecuencias altas, y que en general proporcionan los mejores resultados de localización, la aparición de grandes lóbulos laterales en provoca que nuestro array integre información procedente de direcciones indeseadas, con la consiguiente corrupción de los resultados.

Algunas soluciones a estos problemas pueden ser la utilización de arrays de longitud efectiva mayor y menor distancia entre micrófonos (2.3.1) de manera que se evite el aliasing espacial o la utilización de CDB (2.3.3.3).

2.4.2.4.6 Geometría del array de micrófonos

Una decisión importante a la hora de definir el array de micrófonos es la elección de los 3 parámetros básicos que lo configuran (número de micrófonos, separación entre los mismos y longitud efectiva). En este apartado se trata de dar

respuesta a este aspecto evaluando los resultados obtenidos de manera práctica para ante la variación de dichos parámetros.

En primer lugar se estudian los efectos de la variación del *número de micrófonos*, consideraremos un array con una longitud física constante de 640 mm y para las configuraciones mostradas en la Tabla 2.9, en la cual se presentan también los resultados obtenidos.

- **BBDD:** HIFI-MM1 Simulada
- **Array:** Variable
- **Separación entre mics:** Variable
- **Frame size:** 640 ms
- **Tamaño grid de búsqueda:** 50 mm
- **Frecuencia de muestreo:** 48 KHz

Efecto del número de micrófonos (N)						
Parámetro	**Unidades**	**Lineal 33 d=33 mm**	**armónico 11**	**Lineal 5 d=160 mm**	**Lineal 3 d=320mm**	**Lineal 2 d=640mm**
MOTP	[mm]	103	182	173	164	113
	Red. error rel [%]		-76.7	-68	-59.2	-9.7
A-MOTA (%)	[%]	81 ± 0.8	72 ± 0.9	50 ± 1.0	46 ± 1.0	-38 ± 1.0
	Red. error rel [%]		-11.1	-38.3	-43.2	-46.9

Tabla 2.9 Efecto del número de micrófonos

La tabla anterior confirma los resultados esperados, ya que el aumento del número de micrófonos mejora los resultados obtenidos por el algoritmo, lo cual se debe principalmente a dos factores. En primer lugar la disminución de micrófonos del array para una longitud física constante se traduce en un aumento de la distancia entre micrófonos, lo cual provoca la aparición de aliasing espacial (2.3.3). El segundo en que se traduce esta disminución del número de micrófonos es un aumento de los lóbulos laterales del patrón de directividad (2.3.2).

El otro parámetro a considerar es la variación de la *longitud efectiva* para una distancia entre micrófonos fija de 20 mm. Los resultados se presentan en la Tabla 2.10.

Efecto de la longitud efectiva (L)							
Parámetro	**Unidades**	**Lineal 33 L=660mm**	**Lineal 25 L=500mm**	**Lineal 21 L=420mm**	**Lineal 17 L=340mm**	**Lineal 11 L=220mm**	**Lineal 5 L=100mm**
MOTP	[mm]	103	121	186	185	169	247
	Red. error rel [%]		-17.5	-80.6	-79.6	-64.1	-139.8

A-MOTA(%)	[%]	81 ± 0.8	71 ± 0.9	8 ± 0.5	6 ± 0.5	-43 ± 1.0	-48 ± 1.0
	Red. error rel [%]		-12.3	-90.1	-92.6	-153.1	-159.3

Tabla 2.10 Efecto de la longitud efectiva

El resultado más importante obtenido de la tabla anterior consiste en la variación de manera no lineal de los resultados obtenidos. Esta mejora escalonada da a entender que las mejoras en la localización se producen cuando la longitud efectiva proporciona patrones de directividad suficientemente estrechos como para distinguir entre puntos de búsqueda adyacentes, siendo este efecto más acusado cuando más pequeño hacemos el grid de búsqueda (en este caso 50 mm).

Por último se evalúan los efectos de la variación de la *distancia entre micrófonos*, obteniéndose los resultados mostrados en la Tabla 2.11.

- **BBDD:** HIFI-MM1 Simulada
- **Frame size:** 640 ms
- **Array:** Lineal
- **Tamaño grid de búsqueda:** 50 mm
- **Separación entre mics:** Variable
- **Frecuencia de muestreo:** 48 KHz

Efecto de la distancia entre micrófonos					
Parámetro	**Unidades**	**d=160 mm (L=800 mm)**	**d=80 mm (L=400 mm)**	**d=40 mm (L=200 mm)**	**d=20 mm (L=100 mm)**
MOTP	[mm]	173	102	174	247
	Red. error rel [%]		41.0	-0.6	-42.8
A-MOTA (%)	[%]	50 ± 1.0	-34 ± 1.0	-47 ± 1.0	-48 ± 1.0
	Red. error rel [%]		-168.0	-194.0	-196.0

Tabla 2.11 Efectos de la distancia entre micrófonos

Como se puede apreciar el aumento de la distancia entre micrófonos redunda en una mejora en la localización. Esto se debe a dos factores principales. En primer lugar este aumento se traduce en un aumento de la longitud efectiva, con la consiguiente reducción de la anchura del lóbulo principal del patrón de directividad, permitiendo un apuntamiento más preciso del array. En segundo lugar dicho aumento se traduce en el aumento de los retardos de las señales entre los micrófonos, lo cual reduce los efectos del redondeo al convertir las unidades de tiempo real a muestras.

2.4.2.4.7 Posición de los targets

Teóricamente cuanto mayor sea el ángulo de incidencia de la señal, mayor retardo habrá entre las señales recibidas por los distintos micrófonos, redunda en mejores estimaciones de la posición (2.4.2.1.3). Esto se debe a que pequeñas diferencias en las TDOA provocan mayores errores de redondeo al pasar de tiempo continuo a tiempo discreto, por lo que cuanto más asimétrica sea la ubicación del target respecto del array mejor debería ser la estimación de la localización. Un caso particular lo constituye la ubicación perpendicular al array, en la cual no se produce error de redondeo y por tanto debería ser la que mejor comportamiento presentara. Lo expuesto anteriormente se ve refrendado en la Tabla 2.12 bajo las condiciones presentadas.

- **BBDD:** HIFI-MM1
- **Array:** Lineal 4 elementos
- **Separación entre mics:** 200 mm
- **Frame size:** 320 ms
- **Tamaño grid de búsqueda:** 150 mm
- **Frecuencia de muestreo:** 48 KHz

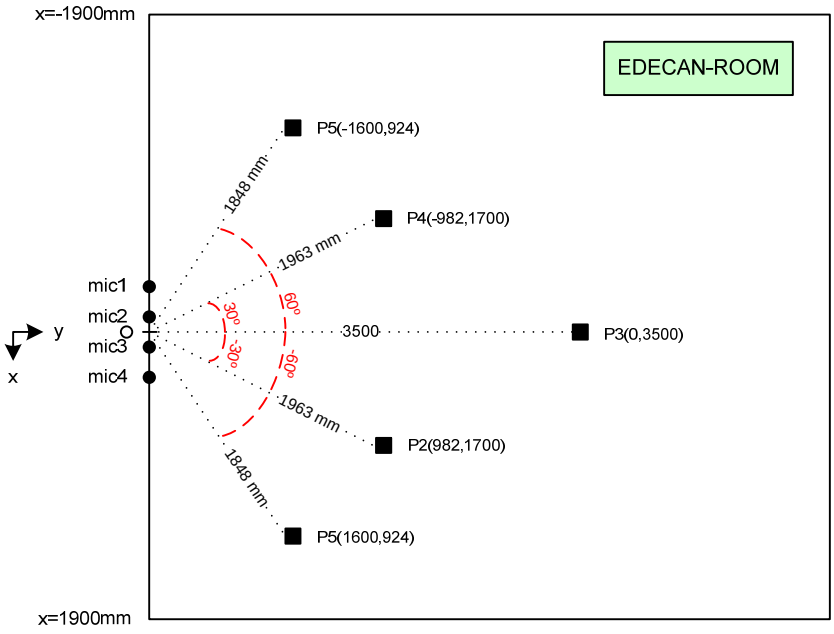


Figura 2.15 Posiciones de los targets en la EDECAN Room utilizadas en este ensayo

Efecto de la posición de los targets						
Parámetro	**Unidades**	**Posición 3**	**Posición 1**	**Posición 5**	**Posición 2**	**Posición 4**
MOTP	[mm]	187	217	197	168	207

	Red. error rel [%]		-16.0	-5.3	10.2	-10.7
A-MOTA (%)	[%]	83 ± 1.7	80 ± 1.8	75 ± 2.0	68 ± 2.2	71 ± 2.0
	Red. error rel [%]		-3.6	-9.6	-18.1	-14.5

Tabla 2.12 Efecto de la posición de los targets

2.4.2.4.8 Carga computacional

Uno de los objetivos inicialmente marcados (1.2) consiste en la ejecución del algoritmo en tiempo real, lo cual permitirá su implementación en sistemas reales, por lo que se hace necesario un estudio de los cuatro principales factores que influyen sobre el mismo:

1. *FSRP vs. TSRP*: A pesar de que ambos métodos son teóricamente equivalentes en diferentes apartados se ha aludido a la imprecisión inherente al método TSRP relativa a la conversión de tiempo continuo a tiempo discreto, la cual no se produce en FSRP. Como contrapartida FSRP incrementa el tiempo de cómputo varios órdenes de magnitud respecto a TSRP.

- **BBDD:** HIFI-MM1
- **Array:** Lineal 4 elementos
- **Separación entre mics:** 200 mm
- **Frame size:** 320 ms (FFT size:16348)
- **Grid de búsqueda:** 250 mm (429 localiz.)
- **Frecuencia de muestreo:** 48 KHz

Variación de la carga computacional FSRP vs. TSRP			
Parámetro	**Unidades**	**FSRP**	**TSRP**
MOTP	[mm]	231	217
	Red. error rel [%]		6.1
A-MOTA (%)	[%]	71 ± 0.9	56 ± 1.0
	Red. error rel [%]		-51.7
TIEMPO DE EJECUCIÓN	[U. de tiempo real]	62.00	0.25
	Reducción rel. [%]		-24700.0

Tabla 2.13 Variación de la carga computacional FSRP vs. TSRP

2. *Frame Size*: Como se indicó en 2.4.2.4.2, el aumento de este parámetro implica el aumento del número de puntos de la FFT, con el consiguiente aumento (lineal) de la carga computacional, el cual se cuantifica en la Tabla 2.14.

- **BBDD:** AV16.3
- **Array:** Circular 8 elementos
- **Radio del array:** 100 mm
- **Frame size:** Variable
- **Grid de búsqueda:** 150 mm (56867 localiz.)
- **Frecuencia de muestreo:** 16 KHz

Variación de la carga computacional con el Frame size						
Parámetro	**Unidades**	**640 ms**	**320 ms**	**160 ms**	**80 ms**	**40 ms**
MOTP	[mm]	190	191	192	194	197
	Red. error rel [%]		-0.5	-1.1	-2.1	-3.7
A-MOTA (%)	[%]	87 ± 0.8	74 ± 1.0	68 ± 1.1	65 ± 1.1	61 ± 1.1
	Red. error rel [%]		-14.95	-21.84	-25.3	-29.89
TIEMPO DE EJECUCIÓN	[U. de tiempo real]	3.38	1.80	0.88	0.48	0.30
	Reducción rel. [%]		-87.8	-284.1	-604.2	-1026.7

Tabla 2.14 Variación de la carga computacional con el Frame size

3. *Numero de micrófonos*: El aumento del número de micrófonos redunda en un aumento del número de parejas de las que se debe computar la GCC para en el algoritmo SRP. Los efectos se muestran en la siguiente tabla.

- **BBDD:** HIFI-MM1 Simulada
- **Array:** Variable
- **Separación entre mics:** Variable
- **Frame size:** 640 ms
- **Tamaño grid de búsqueda:** 50 mm
- **Frecuencia de muestreo:** 48 KHz

Variación de la carga computacional con el número de micrófonos (N)						
Parámetro	**Unidades**	**Lineal 33 L=660 mm**	**armónico 11**	**Lineal 5 L=800 mm**	**Lineal 3 L=960 mm**	**Lineal 2 L=1280**
MOTP	[mm]	103	182	173	164	113
	Red. error rel [%]		-76.7	-68	-59.2	-9.7
A-MOTA (%)	[%]	81 ± 0.8	72 ± 0.9	50 ± 1.0	46 ± 1.0	-38 ± 1.0
	Red. error rel [%]		-11.1	-38.3	-43.2	-46.9
TIEMPO DE EJECUCIÓN	[U. de tiempo real]	40.00	11.80	1.13	0.38	0.18
	Reducción rel. [%]		-239.0	-3439.8	-10426.3	22122.2

Tabla 2.15 Variación de la carga computacional con el número de micrófonos

4. *Tamaño del Grid*: De manera teórica su disminución debería aumentar la carga computacional. Los datos reales se presentan en la siguiente tabla.

- **BBDD:** AV16.3
- **Array:** Circular 8 elementos
- **Radio del array:** 100 mm
- **Frame size:** 640 ms
- **Grid de búsqueda:** Variable
- **Frecuencia de muestreo:** 16 KHz

Variación de la carga computacional con el tamaño del grid					
Parámetro	**Unidades**	**50 mm 56876 loc**	**100 mm 7770 loc**	**150 mm 2450 loc**	**250 mm 540 loc**
MOTP	[mm]	227	236	259	262
	Red. error rel [%]		-4.0	-14.1	-15.4
A-MOTA (%)	[%]	53 ± 3.2	51 ± 3.2	37 ± 3.1	39 ± 3.2
	Red. error rel [%]		-4.3	-34.0	-29.8
TIEMPO DE EJECUCIÓN	[U. de tiempo real]	6.85	3.55	3.38	3.15
	Reducción rel. [%]		-93.0	-102.7	-117.5

Tabla 2.16 Variación de la carga computacional con el tamaño del grid

La conclusión general de este apartado es que existe una clara relación de compromiso entre la eficiencia y robustez del algoritmo que presenta el algoritmo y el tiempo de CPU necesario para ejecutarlo.

2.4.2.5 Estrategias de mejora de SRP-PHAT

A lo largo de este apartado se presenta una serie de técnicas orientadas a incrementar la precisión, fiabilidad y robustez del algoritmo SRP básico.

2.4.2.5.1 Búsqueda Coarse to fine

Tal y como se demuestra en [Varma 2002] la principal desventaja de SRP-PHAT, con respecto a otros algoritmos como GCC-PHAT, es la alta carga computacional. Uno de los motivos de esta elevada carga es la gran cantidad de puntos del espacio en los que el algoritmo SRP realiza la búsqueda de targets activos, ya que típicamente se divide el espacio en un grid de puntos equiespaciados, para posteriormente aplicar SRP sobre cada uno de ellos.

Una técnica interesante desarrollada con el objetivo de reducir dichos tiempos de cómputo sin necesidad de perder la precisión asociada a los grid de búsqueda pequeños

(Tabla 2.16), fue introducida en [Duraiswami 2001] y desarrollada en [Zotkin 2004] donde se propone una búsqueda jerárquica en varios niveles, de los más grandes a los más pequeños. Esta técnica se basa en que la señal de voz es de banda ancha y en la relación espacio-frecuencial inherente a la misma, gracias a la cual:

- Frecuencias altas → Longitudes de onda pequeñas → permiten explorar el entorno de un modo fino y preciso.
- Frecuencias bajas → Longitudes de onda grandes → permiten integrar energía procedente de grandes zonas del espacio.

En base a esto, Zotkin y Duraiswami realizaron diferentes experimentos con su array con el objetivo de obtener una relación entre la anchura del lóbulo principal (b), el cual define la amplitud de la zona del espacio de la que se captura la energía, y la frecuencia de corte que se debe aplicar en el algoritmo a las señales capturadas para obtener dicha anchura, obteniéndose:

$$b[m] \cong \frac{2\lambda}{5} \tag{2.61}$$

A la hora de implementar el algoritmo, el primer paso es seleccionar una frecuencia de corte apropiada y relativamente baja, para dividir el especio en el número de grandes zonas que deseemos de acuerdo con la relación anterior. De estas zonas se seleccionan las que tienen una potencia acústica adecuada, y en las que por tanto puede haber un target activo, y se dividen en 8 cubos equiespaciados, cada uno de los cuales es explorado con una frecuencia de corte el doble de grande que en caso anterior (ya que el área a explorar es el doble de pequeña). Este proceso se itera el número de veces deseado en función de la precisión necesaria.

La técnica anterior promete ser mucho menos gravosa en términos de carga computacional que la búsqueda clásica, ya que permite desde un primer momento descartar grandes zonas espaciales de la búsqueda de targets.

Experimentos realizados posteriormente por [Castro 2007] conducen a que la relación (2.57) propuesta no es válida para la configuración de su array, de lo cual se puede inferir que distintos arrays presentan diferentes relaciones entre la frecuencia de

corte y la anchura del lóbulo principal. La relación obtenida por Castro para la configuración de micrófonos presentada en Figura 2.15 es:

$$b \cong 0.15 + 5\lambda \tag{2.62}$$

Bajo esta nueva relación se aprecian mejores resultados, pero todavía alejados de los obtenidos con el método de grid equiespaciado clásico.

- **BBDD:** HIFI-MM1
- **Frame size:** 500 ms
- **Array:** Lineal 4 elementos
- **Grid de búsqueda:** 100 mm (para clásico)
- **Separación entre mics:** 200 mm
- **Frecuencia de muestreo:** 48 KHz

Efecto de la búsqueda Coarse to fine			
Parámetro	**Unidades**	**Coarse to fine**	**Método clásico**
MOTP	[mm]	239	193
	Red. error rel [%]		19.2
A-MOTA (%)	[%]	16 ± 0.7	41 ± 1.0
	Red. error rel [%]		156.2

Tabla 2.17 Efecto de la búsqueda Coarse to fine

En la tabla anterior se aprecia que los resultados obtenidos son sensiblemente peores que con el método tradicional, lo cual se debe a que los resultados cometidos durante el primer paso del algoritmo en el que se seleccionan las grandes zonas de búsqueda son todavía grandes. Esto puede deberse a dos factores fundamentales:

1. El empleo de una versión paso bajo de la señal original con el objetivo de explorar en una única iteración grandes zonas del espacio, por lo que una parte importante de la información original no se usa en el proceso de localización, sobre todo en la primera iteración.
2. El empleo de un array lineal de únicamente 4 elementos equiespaciados 200 mm, lo cual implica un alto grado de aliasing desde frecuencias relativamente bajas (según 2.13 desde 862.5 Hz). Esto obliga a utilizar frecuencias menores que ésta en nuestra búsqueda corarse to fine para evitar el aliasing espacial, lo cual unido a (2.58) genera unos cubos de 2 metros de lado, comparables a la dimensión total de la estancia y por lo tanto inútiles.

A pesar de los resultados presentados por [Castro 2007] esta técnica no deja de ser igualmente prometedora bajo un array con una longitud efectiva superior que permita un apuntamiento más fino y con una menor separación entre micrófonos que evite el aliasing espacial.

Por otro lado, y con el objetivo de optimizar el algoritmo desde el punto de vista de carga computacional se pueden emplear estrategias adicionales. Por ejemplo y dado que la señal es filtrada con el objetivo de buscar targets en zonas amplias del espacio, no es necesario realizar la IFFT del total de puntos necesarios, sino que podemos utilizar la frecuencia de corte empleada para realizar una IFFT tan pequeña como sea posible en cada uno de los pasos del algoritmo.

2.4.2.5.2 Noise Masking

Dada la naturaleza intermitente de las señales de voz, puede suceder que en determinados bloques de procesamiento aparezcan pausas o potencias de señal acústica muy bajas, lo cuales generarán estimaciones de posición poco precisas. Este tipo de estimaciones erróneas pueden también ser producto de una potencia de ruido excesiva.

Para reducir los efectos anteriores es imprescindible disponer de un VAD que descarte los frames en los que no hay señal de voz. Sin embargo es posible que existiendo señal de voz, la potencia del ruido sea tan alta como para corromper las estimaciones. Por ello y tal y como se indica en [Dibiase 2000] pag. 28-34 es interesante introducir en el algoritmo consideraciones relacionadas con la relación SNR.

Estas técnicas se basan en una estimación de la potencia del ruido, realizada durante la primera frame capturada sin potencia acústica. En base a esta información se define un umbral, de tal manera que si alguna muestra de la FFT tiene una potencia inferior al umbral marcado no será considerada en el algoritmo de localización. Se definen 2 tipos de umbral:

1. *Umbral fijo:* Se calcula mediante un promediado de la potencia del ruido dentro del ancho de banda deseado.
2. *Umbral adaptativo:* En este caso se obtiene una potencia de ruido para cada frecuencia concreta, obteniéndose una mascara de potencia de ruido dependiente

de la frecuencia con la que comparar los espectros de potencia de las señales acústicas capturadas. En este caso sólo se descartará una muestra de la FFT si su potencia es inferior al valor de la máscara de ruido para dicha frecuencia.

La utilización de umbrales fijos demuestra ser de escasa utilidad en la práctica debido a que la potencia de la voz humana decrece considerablemente conforme aumenta la frecuencia, por lo que la utilización de este tipo de umbral como referencia universal provocará que se descarte gran parte de estas altas frecuencias, con el consiguiente descenso en la precisión de la estimación. Este efecto es aún más negativo de lo que podría parecer, ya que las frecuencias altas se asocian con anchuras del lóbulo principal pequeñas, siendo estas las que permiten un apuntamiento fino del array.

En contrapunto la utilización de umbrales adaptativos redunda en pequeñas mejoras bajo condiciones de ruido medio, siendo especialmente interesantes bajo entornos altamente ruidoso. Un importante parámetro a definir en este tipo de parámetros es el valor de umbral a partir del cual se empiezan a descartar muestras de la FFT (0 dB, 12 dB, 24 dB…etc).

2.4.2.5.3 Ponderación de las fortalezas del sistema

En apartados anteriores se han presentado circunstancias o condiciones en las que el comportamiento de nuestro sistema es más favorable, por lo que sería deseable el diseño de técnicas que pongan más énfasis en estas condiciones que proporcionan mejores resultados [Castro 2007].

Existen una gran cantidad de parámetros que afectan a la estimación final de la posición: Geometría del array, la señal en las distintas bandas de frecuencia, la SNR…etc. Sería deseable por tanto poder medir el efecto que cada una de ellas tiene en nuestra estimación y diseñar esquemas que ponderen positivamente las condiciones favorables y oculten los efectos negativos de cada una. Para realizar esta tarea existen dos esquemas básicos:

1. *Una función de pesos*, la cual realice funciones de filtro sobre un cierto parámetro en función de la fiabilidad que podamos otorgarle.

2. *Una red neuronal,* la cual puede ser entrenada para un conjunto de parámetros de entrada seleccionados por su influencia en la estimación de la posición. Una vez entrenada la NN (*Neuronal Network)* genera por sí misma la función de pesos más apropiada para cada uno de los parámetros, optimizando así la estimación generada.

Tomando como ejemplo concreto la distancia entre micrófonos, tal y como se presenta en 2.4.2.1.3 el aumento de la misma implica un aumento en la TDOA y el la consiguiente reducción de los errores de redondeo. Por otro lado, tal y como se indica en 2.4.2.3.2 SRP es equivalente a la suma de la GCC de todas las posibles combinaciones de parejas de micrófonos, por lo que podemos pensar en una función de pesos que pondere positivamente las GCC de los micrófonos más alejados y negativamente la de los más cercanos. Expresado matemáticamente obtenemos [Castro 2007].

- *Algoritmo SRP sin función de peso*

$$P = 2\pi \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}\left(\tau_{ij}\right) \tag{2.63}$$

- *Algoritmo SRP con función de peso*

$$P = 2\pi \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{ij} C_{ij}\left(\tau_{ij}\right) \quad \text{con} \quad \alpha_{ij} = \frac{d_{ij}}{d_{\max}} \tag{2.64}$$

donde P es la potencia estimada por SRP para una localización concreta, N es el número de micrófonos del array, $C_{ij}(\tau)$ es la GGC entre los micrófonos i y j y α_{ij} es un valor en el rango (0,1] que representa la función de pesos, la cual se calcula como la distancia entre los micrófonos i y j (d_{ij}) dividida entre la distancia máxima entre cualquier pareja de micrófonos $d_{\max}$.

- **BBDD:** HIFI-MM1 Simulada
- **Array:** (1) Lineal 33 elementos
  (2) Armónico 11 elementos
  (3) Lineal 5 elementos
- **Frame size:** 640 ms
- **Tamaño grid de búsqueda:** 50 mm

- **Separación entre mics:** (1) 20mm
  (2) Vble.
  (3) 160mm
- **Frecuencia de muestreo:** 48 KHz

Efecto de la aplicación o no de la función de pesos					
Parámetro	**Unidades**	**Lineal 33 Con F. pesos**	**Lineal 33 Sin F. pesos**	**Lineal 5 Con F. pesos**	**Lineal 5 Sin F. pesos**
MOTP	[mm]	103	111	173	164
	Red. error rel [%]		-7.8		5.2
A-MOTA(%)	[%]	81 ± 0.8	93 ± 0.5	50 ± 1.0	50 ± 1.0
	Red. error rel [%]		14.8		0.0

Tabla 2.18 Efecto de la aplicación o no de la función de pesos

Los resultados presentados en la Tabla 2.18 reflejan los buenos resultados derivados de la utilización de esta técnica, así como su dependencia del número de micrófonos, ya que mientras que la mejoría en el caso de 33 micrófonos es importante, en 5 es prácticamente nula.

Otra característica importante es que la aplicación de esta técnica no implica un aumento de la carga computacional. Todos estos factores unidos convierten a esta técnica en imprescindible cuando se trata con aplicaciones prácticas con requerimientos de tiempo real.

2.4.2.5.4 Técnicas de filtrado

Tal y como se ha indicado en apartados anteriores, la teoría indica que cada banda de frecuencia de la voz humana es útil según para que propósito. Las altas frecuencias son interesantes desde el punto de vista de localización, ya que redundan en un mejor patrón de directividad y en una menor anchura del lóbulo principal, lo cual redunda en un apuntamiento más fino. Sin embargo la potencia de la señal de voz decae fuertemente con el aumento de la frecuencia, de lo que se deduce una clara relación de compromiso entre altas frecuencias con un patrón de directividad adecuado pero pobre relación señal a ruido y bajas frecuencias con buena relación señal a ruido pero grandes lóbulos laterales del patrón de directividad.

Es tipo de técnicas fue introducido en [Dibiase 2000] pag. 46 y estudiado en profundidad en [Castro 2007] pag. 143-151. Algunas de las conclusiones más importantes de las extraídas en ambos trabajos se presentan a continuación:

- Es interesante eliminar las frecuencias por debajo de 300 Hz, ya que gran parte del ruido se localiza en esta banda de frecuencias.
- La eliminación de frecuencias por debajo de 2 KHz mejora los resultados obtenidos, ya que a pesar de que en estas frecuencias se encuentra la mayor cantidad de energía de la señal, implican una anchura excesiva del lóbulo principal del patrón de directividad, impidiendo un apuntamiento preciso. Es curioso que la eliminación de estas frecuencias hace prácticamente irreconocible la señal de voz original.
- La eliminación de frecuencias por debajo de 6 KHz empeora los resultados respecto al filtro paso alto anterior, lo cual indica que la potencia de la señal es tan baja que afecta a la precisión de la localización.
- Por otro lado las frecuencias en el rango [16-24 KHz] presentan una potencia acústica tan baja que las hacen de poca utilidad para la localización.

A la vista de lo anterior se puede concluir que un filtro paso banda [2-16 KHz] es el más adecuado en términos de localización, aunque siempre se debe tener en cuenta el gran efecto de la configuración del array que conlleva variaciones del patrón de directividad y aliasing espacial.

Esta información a priori puede ser muy útil a la hora de construir una función de pesos similar a la obtenida en el apartado anterior para la distancia entre micrófonos.

2.4.2.5.5 Uso de información del entorno

Otro tipo de información a priori de la que podemos disponer, bien gracias a una introducción manual o a partir de otro tipo de sensores como por ejemplo cámaras, es la del entorno en la que se encuentran los targets, como por ejemplo sus límites, su geometría, las zonas muertas…etc. La introducción de este tipo de información en nuestro algoritmo puede proporcionar mejoras como la reducción el tiempo de cómputo

al eliminar de terminadas zonas del grid de búsqueda o aumentar la fiabilidad definiendo zonas en las que el algoritmo no puede ubicar un target activo.

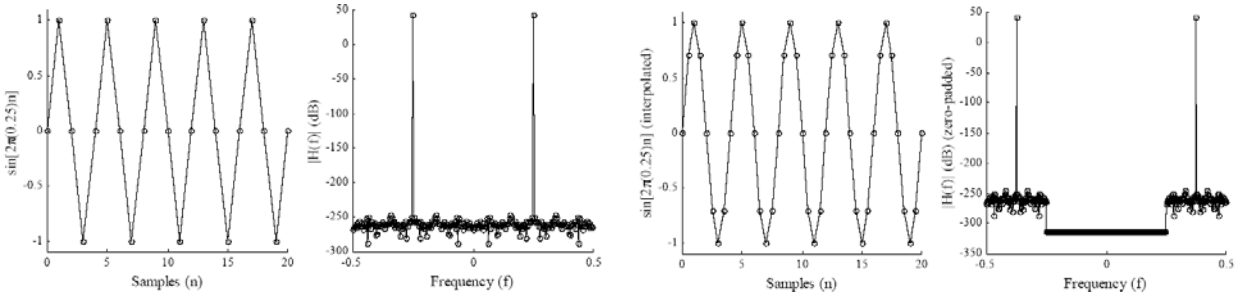
Este tipo de información se puede clasificar en dos grandes grupos:

- *Zonas muertas*: Identifica zonas en las que la probabilidad de encontrar un target activo es prácticamente nula, como por ejemplo encima de un armario o de una mesa. Cuando el algoritmo retorna las posiciones en las que ha encontrado un target activo y alguna de estas estimaciones cae en una zona muerta, se puede suponer que el algoritmo ha cometido un error y por lo tanto descartar dicha estimación, la cual puede ser sustituida por la anterior estimación válida.
- *Límites de la habitación y ubicaciones más probables:* Dado el conocimiento del entorno del que disponeos, debemos implementar el algoritmo de tal manera que realice la búsqueda únicamente dentro de los límites de la habitación. De una manera más elegante se pueden definir, dentro de estos límites de la habitación, zonas donde es más probable que encontremos targets activos (teniendo en cuenta la altura del ser humano, sillas alrededor de una mesa de reuniones,…etc) en las cuales se puede, por ejemplo definir un grid de búsqueda más fino que en el resto de la habitación.

2.4.2.5.6 Interpolación

Como se ya introdujo en 2.4.2.1.3, en casos en los que se trabaja con señales digitales, se debe asumir que el retardo entre dos señales es un número entero de muestras. Sin embargo en muchas ocasiones dicho retardos no son realmente un número entero de muestras, lo cual conlleva errores en la evaluación de las ecuaciones como por ejemplo en la GCC a partir de la cual obtenemos SRP.

Dado que la frecuencia de muestreo es fija en la mayoría de los sistemas, puede ser interesante la introducción de submuestreo ([Varma 2002] pag. 60-68). Para lo cual se realiza un zero-padding en el centro de la DFT de tal manera que la longitud de la misma sea el doble de la original, lo cual se traduce en el dominio temporal en una interpolación x2, tal y como se muestra en la Figura 2.16.



(a) (b)

Figura 2.16 Señal original (a) Vs. señal interpolada (b)

2.4.3 Tracking

Los métodos de tracking pueden ser vistos como una forma de filtrado de la posición instantánea estimada por el algoritmo de localización, de tal manera que generan una trayectoria espacial suavizada. De esta manera si el algoritmo de localización genera una estimación de posición que representa un cambio abrupto en la trayectoria del target, ésta será automáticamente corregida por comparación con la posición anterior y filtrada de manera que la nueva posición estimada este de acuerdo a comportamiento del target en el pasado.

En este apartado se presentan aspectos básicos y generales de algunos mecanismos de tracking basados en distintos trabajos realizados en el ámbito de los arrays de micrófonos. El desarrollo matemático de los mismos se presenta en 3.4.1.2 para el Filtro de Kalman (KF) y en el Anexo I para el Filtro de Partículas (PF).

2.4.3.1 Filtro de Kalman (KF)

El *Filtro de Kalman Lineal (KF)* consta de un conjunto de ecuaciones matemáticas que proporcionan un cálculo eficiente (recursivo) para la estimación del estado de un proceso, de tal manera que minimiza el error cuadrático medio. Para ello se basa en la comparación de las medidas de entrada con las estimaciones en curso para producir una estimación recursiva del estado del sistema. Las ganancias y pesos aplicados a los datos de entrada dependen de importantes factores como la precisión de las medidas y el movimiento del target.

El filtro es muy potente en varios aspectos, ya que proporciona estimaciones de los estados pasados, presentes y futuros, pudiendo realizar dichas estimaciones incluso cuando la naturaleza del sistema modelado es desconocida.

Este filtro fue presentado por primera vez en [Kalman 1960] y en [Welch 2006] se presenta de una manera clara y apoyada en ejemplos prácticos que facilitan su comprensión y utilización en entornos reales. En [Nadeu 2006] se puede encontrar una implementación de tracking simple basada en este filtro, donde la estimación de la localización se obtiene en base a SRP-PHAT que pondera la información de Cross-Power Spectrum (CSP) de los frames de habla anterior y actual, en base a un factor de suavizado adaptativo proporcionado por el KF.

Sin embargo el KF asume modelos con dinámicas lineales y gaussianas, por lo que no puede ser utilizado cuando se trata con comportamientos humanos típicamente no lineales como los cambios bruscos de turno de palabra, o otros efectos típicos de las conversaciones humanas son las expresiones cortas (típicamente de menos de un segundo), en los cuales el target activo cambia constantemente y los solapamientos entre targets representan una parte no despreciable de la conversación.

El *Filtro de Kalman Extendido (EKF)* se presenta como una solución que utiliza una linealización del modelo no lineal obtenida mediante una serie de Taylor [Brandstein 2001] pag. 210-212. Sin embargo la calibración de los parámetros del filtro constituye un problema en aplicaciones prácticas, tal y comos e demuestra en [Julier 1997].

El *Unscented Kalman Filter (UKF)* representa una solución que evita este proceso previo de linealización y se adapta a medidas de fuentes de ruido no gaussianas [Julier 1997]. En [Dvorkind 2005] se presenta una aplicación del UKF a localización de fuentes acústicas.

Otro trabajo interesante que emplea el KF para la estimación de un número múltiple de targets de la escena, en este caso procedente de múltiples arrays de micrófonos se presenta en [Potamitis 2004].

Aparte de integrar información procedente de distintos frames, el filtro de Kalman puede ser utilizado para integrar información procedente de sistemas de

estimación que emplean distintos tipos de sensores [Brandstein 2001] pag. 203-205, fusionando así las fortalezas de ambos, siendo este el objetivo último perseguido por la tesis que aquí comienza para el caso particular de un sistema audio-visual.

2.4.3.2 Filtro de Partículas (PF)

El *Filtro de partículas (PF),* también conocido como *Filtro de Monte-Carlo*, constituye una alternativa al filtro de Kalman expuesto anteriormente. El PF constituye una aproximación al filtro óptimo Bayesiano que representa la distribución de probabilidad a través de un conjunto finito de partículas ([Doucet 2001], [Arulampalam 2002] y [Van Der Merwe 2000]).

Para un modelo en el espacio de estados, el PF aproxima recursivamente, dadas unas observaciones, la distribución de estados del filtro empleando un modelo dinámico, un modelo de observación y técnicas de muestreo, para predecir la configuración de candidatos y medir su probabilidad [Lehmann 2004].

En [Vermaak 2001b] se presentan resultados a nivel de simulación empleando un array lineal de micrófonos separados 60 cm, empleando GCC y un PF con 50 partículas para el tracking. La inclusión del PF aumenta la robustez del algoritmo resultante. En [Ward y Lehmann 2003] encontramos una aplicación de localización y tracking para un único target basada en el PF, en este caso sobre datos reales.

Sin embargo los rápidos cambios en el turno de palabra hacen necesaria la utilización de modelos multifuente ([Larocque 2002]) o la adaptación del modelo simple para realizar una conmutación entre targets activos ([Lehmann 2004]). La estimación del número de fuentes activas constituye otro problema del PF, fuertemente relacionado con el problema de asociación. Por otro lado el número de targets activos varía rápidamente con el tiempo, lo cual implica la implementación de complicadas reglas de nacimiento/muerte.

En apartados sucesivos de este documento se realiza un estudio más detallado de este algoritmo debido a su amplia utilización en tareas de tracking que aquí nos ocupa, y a su amplia aceptación por parte de la comunidad científica.

2.4.3.3 Short-Term Clustering

Recientemente, el laboratorio de investigación suizo IDIAP ha propuesto en [Lathoud y McCowan 2004] una nueva e interesante técnica conocida como *Short-Term Clustering*, en la que no es necesario un conocimiento previo del número de targets.

Este sistema agrupa, de manera online y sin supervisión externa, estimaciones de localización cercanas entre si tanto en tiempo como en espacio y las separa de las que no lo están. En [Lathoud y Odobez 2004] se demuestra en base a datos reales que este sistema puede ser aplicado a tareas de segmentación de voz espontánea, sin necesidad de ningún tipo de restricción por parte de los participantes. Trabajos posteriores de los mismos autores relacionados con esta misma temática se presentan en [Lathoud y Odobez 2007].

La segmentación de voz resultante presenta una gran precisión, incluso cuando existen múltiples targets activos de manera concurrente. Adicionalmente esta técnica permite la detección y resolución del problema de cruces de trayectorias.

En [Gatica-Perez 2007] esta técnica se utiliza como alternativa a la implementación de los clásicos y normalmente poco robustos *Voice Activity Detectors* (VAD), para la detección de tramos de habla y silencio. Los buenos resultados obtenidos con esta técnica evitan que el mal funcionamiento del VAD de al traste con una buena implantación del algoritmo de estimación.

2.4.4 Pose

En el contexto de los arrays de micrófonos, la estimación de la pose se refiere a la obtención de la dirección hacia la que está enfocada la fuente sonora. Esta estimación de la dirección así definida puede ser aplicada a cualquier fuente sonora, a pesar de que en el ámbito en el que se realiza este trabajo hace que el interés se centre en personas como fuentes sonoras, o lo que es lo mismo, la estimación de la orientación de la cabeza.

El conocimiento de la orientación de los hablantes, unida a la ubicación de los mismos, puede conducir a mejoras importantes en aplicaciones que requieren de la

interacción hombre máquina, permitiendo un mejor entendimiento de lo que los usuarios quieren hacer o a lo que se refieren.

Hasta el momento, la mayoría de los trabajos relativos a estimación de pose se han desarrollado a partir de sensores de visión, siendo escasos los trabajos orientados a la resolución de este problema basados en arrays de micrófonos. Sin embargo, dado que los propios seres humanos emplean información tanto visual como acústica para inferir la orientación de las fuentes y el ámbito de "espacios inteligentes" en el que se plantea esta tesis, es inmediato concluir que la información de audio también debería ser utilizada para la obtención de dicha orientación.

La mayoría de los algoritmos de estimación de pose trabajan de manera cooperativa con robustos algoritmos de localización, lo cual se debe a que la estimación de la orientación de la cabeza es un factor que puede degradar considerablemente las características de dichos algoritmos de localización. Por ello estos algoritmos suelen trabajar en dos etapas: En la primera de ellas se trata de obtener la posición de la fuente y en la segunda la orientación de la misma.

Uno de los primeros trabajos que siguen la filosofía presentada anteriormente es [Mungamuru 2004], en el que tomando como base el algoritmo SRP-PHAT (2.4.2.3.3), se realiza una extensión de la búsqueda a la orientación de los targets por medio de una ponderación de la contribución de cada par de micrófonos para diferentes orientaciones posibles. Una aproximación similar basada en SRP-PHAT se presenta en [Brutti 2005], dando lugar al algoritmo *Oriented Global Coherente Field* (OGCF). Este tipo de algoritmos basados en SRP-PATH, presentan las ventajas de ser al mismo tiempo robustos al ruido y las reverberaciones y de ser casi independientes de la orientación del la cabeza del target (siempre que la red de micrófonos este correctamente distribuida por la habitación.

Un trabajo previo que no sigue la filosofía de extensión de un robusto sistema de localización como los tratados anteriormente se presenta en [Sachar 2004], el cual basa su estimación en medidas de energía acústica empleando un array con un gran número de micrófonos.

A continuación se presentan dos alternativas para la obtención de la pose: La primera de ellas basada en el algoritmo SRP-PHAT y la segunda en determinadas consideraciones de la directividad del hablante ([Abad 2007] pág. 112).

2.4.4.1 Estimación de la pose basada en SRP-PHAT

Esta técnica, introducida en [Mungamuru 2004] y [Brutti 2005] y desarrollada y evaluada en [Abad 2007] pág. 135-137, realiza la estimación de la pose a partir de la maximización conjunta de la función de verosimilitud espacial para las potenciales posiciones y orientaciones de la fuente. La función de verosimilitud para el algoritmo SRP-PHAT puede ser extendida para incorporar la pose de los targets de la siguiente manera ([Abad 2007]).

$$F(\vec{x};o) = F\left(\vec{T}(\vec{x}),\vec{O}(\vec{x};o)\right) = \sum_{p=1}^{P} \Omega_p \frac{X_{p1}(f)\cdot X_{p2}^{*}(f)}{\left|X_{p1}(f)\right|\cdot\left|X_{p2}^{*}(f)\right|}\cdot e^{-j2\pi f\tau_p} df \tag{2.65}$$

Para cada localización espacial $\vec{x}$ y orientación o, además del vector de retardos temporales $\vec{T}(\vec{x}) = \left[\tau_1(x),\tau_2(x),...,\tau_p(x)\right]$ constituido por los retardos teóricos entre cada par de micrófonos, se define un vector de pesos $\vec{O}(\vec{x},o) = \left[\Omega_1(x,o),\Omega_2(x,o),...,\Omega_p(x,o)\right]$ formado por los pesos adecuados para representar la influencia de cada CC en términos de orientación relativa.

Para el los valores de $\Omega_p(\vec{x},o)$, una opción razonable consiste en que el peso del p-esimo par de micrófonos dependa de la energía que debería recibir procedente de una fuente hipotética ubicada en una posición concreta. Una solución simple propuesta en [Abad 2007] consiste en calcular en primer lugar la diferencia de ángulos entre la orientación explorada y la línea que cruza el punto medio del par de micrófonos y la posición explorada ($\theta_p(x)$). Con ello, el peso $\Omega_p(x,o)$ se calcula a partir de la mencionada diferencia de ángulos, mediante la cual se aproxima el patrón de directividad del hablante. La expresión para $\Omega_p(x,o)$ y su representación gráfica se muestran a continuación:

$$\Delta\theta_p(\vec{x},o)=\theta_p(\vec{x})-o \rightarrow \Omega_p(\vec{x},o)=F\left(\Delta\theta_p(\vec{x},o)\right)=\frac{1}{1+0.6\cdot\sin^2\left(\Delta\theta_p(\vec{x},o)/2\right)} \quad (2.66)$$

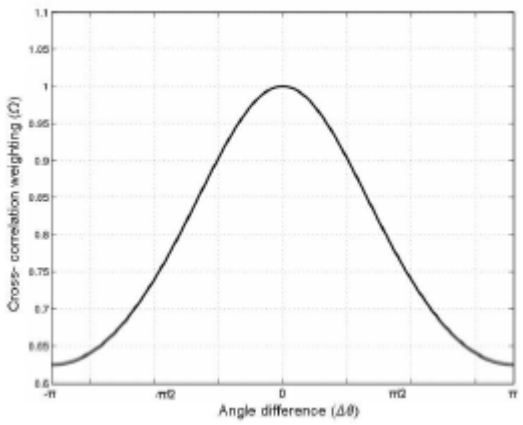


Figura 2.17 Función de directividad normalizada dependiente de la diferencia de ángulos para el pesado del par de micrófonos

La estimación conjunta de la posición y la orientación se obtiene a partir de la siguiente maximización:

$$\left\{\hat{\vec{x}},\hat{o}\right\}=\arg\max_{\vec{x},o} F\left(\vec{T}(\vec{x}),\vec{O}(\vec{x},o)\right) \quad (2.67)$$

La aplicación de este método proporciona una solución robusta y elegante a la estimación simultánea de la posición y la pose de un target a partir de información acústica. De hecho, la introducción del parámetro que nos permite estimar la pose debería robustecer la estimación de la posición generada, ya que la contribución de cada CC es ponderada de acuerdo con el grado de fiabilidad de la información que proporciona cada par de micrófonos en función de su posición relativa con respecto a la fuente.

En aplicaciones de ellos la implementación directa de éste algoritmos es inviable debido a la alta carga computacional que implica. Sin embargo [Abad 2007] se propone un método eficiente que si permite dicha implementación.

2.4.4.2 Estimación de la pose basada en HLBR *(High/Low Band Ratio)*

Este método, propuesto en [Segura 2007], es altamente eficiente en términos de carga computacional debido a su simplicidad.

Asumiendo que se dispone de una red de micrófonos correctamente distribuida, el conocimiento del patrón de radiación humano se puede utilizar para estimar la pose de los hablantes activos en la escena, mediante el cálculo de la energía recibida en cada micrófono y buscando el ángulo que mejor encaja el patrón de radiación con las medidas de energía.

Sin embargo, esta simple aproximación presenta diversos problemas en entornos reales, ya que la calibración de los arrays ha de ser perfecta y la atenuación de la señal en cada uno de los micrófonos es diferente debido al modelo de propagación. Para resolver los problemas anteriores manteniendo la simplicidad computacional, [Segura 2007] emplea una normalización de la energía acústica. Adicionalmente, este método no precisa arrays de micrófonos con un gran número de elementos.

Por otro lado, el patrón de radiación de la voz humana presenta una baja direccionalidad para bajas frecuencias y mientras que para altas frecuencias, dicha direccionalidad es alta. En base a este conocimiento se puede definir un HLBR (High/Low Band Ratio) del patrón de radiación, como una relación entre las bandas alta y baja del mismo. De acuerdo con observaciones experimentales, el rango considerado para la banda baja es de [200 Hz, 400 Hz], mientras que para la banda alta es de [3500 Hz, 4500 Hz].

En base a lo anterior, en lugar de calcular la energía absoluta recibida en cada micrófono, se propone calcular el HLBR de energía acústica para cada micrófono y así poder estimar la orientación que mejor concuerda con las medidas realizas.

Para que la implementación de este algoritmo sea posible, es preciso conocer la posición de la fuente $\vec{x}$, por lo que en implementaciones prácticas deberá ser estimada a través de algoritmos de localización. Una vez estimada dicha posición se puede calcula el vector $\vec{v}_q$ desde el hablante a cada micrófono m_q, cuyo módulo se representa como $\left|v_q\right|$ y su ángulo como θ_q. A partir de las variables anteriores es posible obtener la función $\vec{V}(\theta)$, que relaciona el HLBR de energía acústica de cada micrófono con cada ángulo a través de la siguiente expresión:

$$\vec{V}(\theta) = \sum_{q=1}^{Q} \left|\vec{v}_q\right| \delta\left(\theta - \theta_q\right) \tag{2.68}$$

La estimación de la pose del hablante se obtiene mediante la búsqueda del ángulo que maximiza la correlación entre el modelo matemático del HLBR del patrón de radiación $G\theta$ y el HLBR de energía acústica medida en cada micrófono. La expresión matemática que representa esta relación se presenta a continuación:

$$\hat{o} = \arg\max_{\theta} \vec{G}(\theta) * \vec{V}(\theta) = \sum_{q=1}^{Q} \left|\vec{v}_q\right| \vec{G}\left(\theta - \theta_q\right) \tag{2.69}$$

Para el modelado de $\vec{G}(\theta)$ se emplea la función representada en la Figura 2.17. A la técnica descrita por las ecuaciones (2.68) y (2.69) y el modelo anterior, se la conoce como HLBR básica (HLBR-B).

Otro método también basado en HLBR pero en este caso no dependiente del modelo del HLBR del patrón de radiación, consiste en calcular el vector suma de todos los vectores HLBR de cada micrófono. El ángulo del vector suma constituye la estimación de la pose obtenida con esta técnica, a la cual se la denomina vectorial HLBR (HLBR-V) y se expresa matemáticamente como sigue:

$$\vec{v}_{sum} = \sum_{q=1}^{Q} \vec{v}_q \qquad\qquad \hat{o} = \angle \vec{v}_{sum} \tag{2.70}$$

Además de su sencillez, una de las ventajas de las técnicas basadas en HLBR es que pueden proporcionar una alta resolución sin que esto implique un incremento significativo de la carga computacional. En el caso HLBR-B, la ecuación (2.69) puede ser evaluada para múltiples orientaciones diferentes debido al bajo coste computacional de cada evaluación, mientras que para HLBR-V la resolución se consigue de manera natural, ya que la estimación se obtiene a partir del ángulo del vector formado por la suma de los vectores HLBR de cada micrófono, y por consiguiente no está restringido a un set finito de posibles candidatos.

2.5 Líneas de investigación en el área de seguimiento acústico

A lo largo de este apartado dedicado a los arrays de micrófonos, se ha hecho especial hincapié en la identificación y descripción de las técnicas que han tenido una mayor aceptación por parte de la comunidad científica en el ámbito de la detección, localización, seguimiento y obtención de pose, así como en la introducción de determinados conceptos básicos para la comprensión de las mismas.

Durante de dichas exposiciones se han referenciado trabajos de diferentes grupos de investigación punteros en la materia. El objetivo de este apartado no es otro que el identificar, de manera clara y concisa, los grupos de investigación punteros en el trabajo con arrays de micrófonos, para así facilitar el seguimiento de los avances y trabajos que los mismos puedan aportar a la comunidad científica.

Dentro del **ámbito nacional**, los grupos de trabajo más importantes son:

1. *UPM (Universidad Politécnica de Madrid):* Trabajos referenciados a lo largo del presente estado del arte como [Muñoz 2006], [Rodriguez 2006] o [Castro 2007], realizados bajo la supervisión de Javier Macías Guarasa, son de gran interés especialmente en el ámbito de la localización
2. *UPC (Universidad Politécnica de Cataluña):* Los últimos trabajos realizados por este grupo como [Abad 2007] o [Segura 2007] son de especial interés en al ámbito de la obtención de la pose.

En el **ámbito internacional**, los grupos de trabajo punteros son:

1. ITC (Istituto Trentino di Cultura): Este grupo de investigación, liderado por Mauricio Omologo y en el que destacan otros investigadores como Alessio Brutti o Piergiorgio Svaizer, posee un reconocido prestigio internacional, avalado por trabajos como [Omologo 2003], [Brutti 2005] o [Brutti 2008], abarcando todos los ámbitos de trabajo con arrays de micrófonos (detección, localización, seguimiento y pose).

2. IDIAP (Institut Dalle Molle d'Intelligence Artificielle Perceptive) Research Institute: Este grupo constituye sin sombra de duda, el referente principal en lo que a avances en el campo de procesado de la señal acústica se refiere. Investigadores como Guillaume Lathoud, Jean-Marc Odobez o Iain McCowan poseen un reconocido prestigio internacional en esta temática, gracias a trabajos como [McCowan 2001], [Lathoud y McCowan 2004], [Lathoud y Odobez 2004] o [Lathoud y Odobez 2007] por citar algunos.

3. Visión

3.1 Introducción

La visión constituye uno de los mecanismos sensoriales de percepción más importantes (sino el más) en el ser humano, en cuanto a la cantidad de información que se puede obtener a partir del mismo. La utilización conjunta de cámaras de visión y computadoras de gran capacidad intentan emular este sentido humano y emplear la gran cantidad de información proporcionada dentro del ámbito concreto de aplicación, en nuestro caso los "espacios inteligentes".

La visión artificial se emplea en un amplio abanico de áreas, desde la industria [Hui-Fuang 2004], la medicina [Martins 2007] y [Duncan y Ayache 2000] o la robótica [Martins 2007], hasta la detección de elementos (e.g. caras) [Everingham 2005], la seguridad [Cucchiara 2005] o la navegación autónoma de vehículos [Akita 2007], pasando por supuesto por los "espacios inteligentes" [Pizarro 2007], [Marrón 2005], entre otros. Este tipo de sensores se han convertido en uno de los más empleados dentro de la comunidad científica a lo largo de los últimos años, debido principalmente a la gran cantidad de información que proporcionan.

Para el caso de las cámaras, al igual que para los arrays de micrófonos los problemas asociados a este tipo de sensores afectan en mayor o menor medida en función de la solución algorítmica concreta implementada. Algunos de los problemas más importantes asociados a este tipo de sensores son:

1. *Incertidumbres propias del sistema de adquisición* como cambios de iluminación (brillos, sombras, contraste), ruido (característicos del sensor y la óptica) o la aparición de outliers[11].

[11] El término inglés "*outlier"* se define como observaciones concretas del set de datos inconsistentes con el resto de datos del mismo, por lo que pueden ser considerados como un ruido esporádico que no aporta información relevante para el proceso de seguimiento. Este término se emplea sin traducir debido a su gran difusión.

2. *Oclusiones totales o parciales* que otros targets u objetos de la escena producen sobre el target de interés, las cuales dependen de la ubicación del sensor y de la vista generada por el mismo.
3. *Ambigüedad de las imágenes obtenidas* las cuales siempre proceden de una proyección en 2D de la escena 3D original.
4. *Variaciones naturales* en de los targets de la imagen: Color, forma, tamaño, textura, relaciones con el resto…etc.
5. *Asociación o identificación* de cada dato de posición con el target del que ha sido extraído, ya que el sensor no proporciona información sobre la identidad de la medida.
6. *La gran cantidad de información* proporcionada por este tipo de sensores, junto con la restricción de tiempo real impuesta como requerimiento de nuestro sistema, imponen la necesidad de una selección de la información a utilizar.

3.1.1 Obtención de la imagen

Los sensores empleados para la implementación de cámaras de visión pueden agruparse dentro de tres tipos principales [Bradner 2003]:

1. *CCD (charge coupled devides)*: Consiste en un circuito integrado que contiene un array con un número determinado de condensadores enlazados entre sí. La carga de cada condensador varía en función de la intensidad de luz recibida y proporciona una corriente eléctrica proporcional a la misma. Este tipo de sensores presentan problemas de *blooming*[12], el cual afecta al procesamiento de la imagen en situaciones de cambios de iluminación, incidiendo por tanto en la robustez del algoritmo de tracking implementado. Otro factor limitante puede ser su escaso rango dinámico (típicamente 8 bit → 48 dB).

[12] B*looming*: La recepción de una gran intensidad lumínica en un punto influye en los píxeles adyacentes, provocando la aparición de líneas blancas en la imagen. Se emplea sin traducir debido a su amplia difusión en la comunidad científica.

2. *CMOS*: Esta tecnología permite integrar la parte analógica y digital en un mismo chip, facilitando la implementación de cámaras de pequeño tamaño e interfaz simple. Existen dos ventajas principales de estos sobre los CCD:

    - Sensibilidad logarítmica: Su rango dinámico es de 20 bit (120 dB), evitando los problemas de blooming de los CMOS.
    - Acceso directo a los píxeles: Permite leer desde un sólo píxel a un grupo de ellos (área variable de imagen). Este aspecto es de especial interés en sistemas de alta velocidad, ya que si se selecciona un número reducido de píxeles su lectura requerirá menos tiempo.
    - Presentan un menor consumo y mayor velocidad que los CMOS.

    Las desventajas se pueden resumir en dos:

    - Presentan un mayor nivel de ruido que los CCD.
    - Presentan una menor sensibilidad a la luz, lo cual los hace inapropiados en condiciones de baja luminosidad. Su *fill factor* (porcentaje de píxel dedicado a captar luz) es del 70% frente al 100% de los CMOS.

3. *Linlog (Linear/Logarithmic)*: Presentan una respuesta lineal ante bajos niveles de iluminación y logarítmica ante valores altos de la misma, lo que permite un alto rango dinámico. Este ajuste puede ser realizado de manera online y la conversión es implementada a nivel de hardware, por lo que no implica un aumento de la carga computacional.

Por otro lado, en función de la disposición de los píxeles las cámaras pueden ser *lineales o matriciales* y en función de la información de salida pueden ser *en blanco y negro* (un elemento CCD o CMOS por cada píxel) *o en color* (tres elementos CCD o CMOS por cada píxel –componentes RGB-).

De lo anteriormente expuesto se deduce que la cámara a utilizar dependerá del tipo de aplicación a la que vaya destinada, no existiendo una cámara universal que cubra las necesidades de todas simultáneamente.

3.2 Extracción de información de la imagen

La entrada de un sistema de seguimiento visual, entendiendo seguimiento desde el punto de vista del presente estado del arte como la estimación de la posición y la velocidad, en tiempo real, de un número variable y desconocido de targets, consiste en una secuencia de imágenes en 2D extraídas de una escena 3D.

Tal y como se comentó en el apartado de introducción, la cantidad de información extraída de cada imagen es en la mayoría de los casos muy grande, y sin embargo, sólo determinadas subregiones de cada una son utilizadas en la localización y seguimiento de los targets objetivo. La extracción de estas subregiones se realiza en base al conocimiento de los targets *(foreground)* y del fondo en el cual se encuentran ubicados *(background)*. A este mecanismo se le conoce como seguimiento basado en características y es necesario para la implementación de tracking en tiempo real de cualquier tipo de targets, incluso los más simples y abarcan desde las más sencillas basadas en "blobs"[13] hasta las más complejas basadas en combinaciones de diferentes características.

Estas y otras técnicas serán utilizadas en los diferentes trabajos estudiados en 3.5 para la generación de sus respectivos modelos de observación.

Evidentemente, cada tipo de características conlleva asociadas unas ventajas y unos inconvenientes, haciéndolas adecuadas o no en función de la aplicación concreta que se desee abordar.

A continuación se presentan someramente algunos de los mecanismos más empleados para llevar a cabo esta tarea:

1. *Tracking basado en segmentación:* Uno de los elementos clave a la hora de realizar un sistema de tracking robusto en tiempo real empleando visión, es la extracción de características en un periodo de tiempo reducido. Uno de los mecanismos más sencillos y recurrentes de la literatura científica consultada es

[13] El término inglés "blob", cuya traducción directa podría ser mancha, se refiere a una zona de la imagen con unas determinadas características de color, intensidad luminosa o textura. Al ser un término utilizado recurrentemente en la literatura científica del área se utiliza sin traducir.

la segmentación. Por poner un ejemplo de la misma se propone el trabajo de Tomasi en [Tomasi 1994].

2. *Puntos característicos:* Las ubicaciones de la imagen en las que el nivel de gris cambia de manera significativa con respecto a los puntos que le rodean se denominan puntos de interés. Una amplia exposición de los mismos, así como de los operadores que nos permiten obtenerlos puede encontrarse en [Schmid 2000]. A diferencia de características de orden superior, estos mecanismos sólo proporcionan una escasa información geométrica, por lo que se requiere un alto grado de correspondencia para la implementación de algoritmos de seguimiento basados únicamente en este tipo de características.

3. *Bordes:* Un borde representa una región de la imagen en la que el gradiente de nivel de gris presenta un máximo local. Debido a la mayor extensión espacial de los mismos con respecto a los puntos de interés tienden a presentar una mayor estabilidad. Sin embargo, la variación de las condiciones de iluminación tiende a aumentar las probabilidades de falsa alarma. Ejemplos de este tipo de detectores se pueden encontrar en [Lu 2000].

4. *Tracking de Plantillas:* En entornos con gran cantidad de bordes (e.g. interiores) la utilización de puntos característicos presenta la ventaja de la reducción de la carga computacional. Sin embargo en entornos menos estructurados no es tan evidente el encontrar dichos puntos característicos, en cuyo caso se recurre a modelos predictivos como los presentados en [Isard 1996]. En casos como el anterior, el uso de plantillas o regiones de imágenes puede ser un mecanismo eficiente para el seguimiento de targets.

5. *Contornos activos:* Desde el punto de vista de tracking visual, los contornos activos permiten reducir los grados de libertad, al mismo tiempo que fuerzan al contorno resultante a cumplir una serie de restricciones impuestas por un determinado grupo de transformaciones (e.g. Transformación afín, proyectiva…etc).

El popular algoritmo *Condensation* (tratado en puntos posteriores) emplea contornos activos para el tracking de una cabeza basado en la utilización de curvas spline[14].

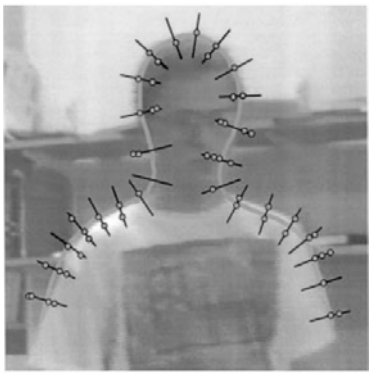


Figura 3.1 Modelo de contorno activo empleado en el tracking de cabezas

6. *Tracking de características híbridas:* Como su propio nombre indica, este mecanismo propone la fusión del resultado de varios de los métodos expuestos con anterioridad, con el objetivo de robustecer el algoritmo de seguimiento.

3.3 Modelado de Targets

El modelado de targets requiere, de manera general, de la definición de un *vector de estados*, y de unos *modelos de actuación* y *obsevación*.

En el *vector de estados* se identifican las variables estimadas a la salida del algoritmo de seguimiento, implementado a partir de las características de observación definidas en el *modelo de observación* y con una evolución temporal caracterizada por el *modelo de actuación.*

De una apropiada elección de estos modelos dependerán en gran medida los resultados obtenidos por los algoritmos de seguimiento implementados. A este respecto la comunidad científica se encuentra dividida entre dos tendencias:

[14]Un *"spline"* es una curva que se define a trozos a través de polinomios y se utiliza recurrentemente para la aproximación de formas complejas debido a la simplicidad de la representación y a la baja carga computacional que conllevan.

- *Modelos complejos y específicos*: La mayor parte de aplicaciones de MTT de la literatura científica consultada, se centran en el aprovechamiento de las modernas técnicas de procesado de imagen para obtener un modelo específico y robusto de los targets a seguir, basados en el color, la forma, el contorno,…etc. Como contrapartida, en muchos casos se simplifica el algoritmo de estimación de la posición. Un ejemplo de este tipo de modelos se utiliza en [Isard 1998a] y [Zhao 2004].

- *Modelos sencillos y generales*: Este tipo de modelos focalizan su esfuerzo en el algoritmo de estimación de la posición y seguimiento empleando modelos sencillos, lo cual redunda a la par en una generalidad de los targets a trackear y en la reducción de los tiempos de cómputo, ayudando a conseguir las especificaciones de tiempo real, la cual constituye una de las especificaciones más importantes del presente trabajo. Esta especificación, junto con la propuesta de utilización de fusión audiovisual para realizar el tracking de los targets, aconseja en principio la utilización de este tipo de modelos sencillos y generales. Este tipo de modelos se proponen por ejemplo en [Ng 2006] y [Marrón 2008].

La diferencia fundamental entre el tracking y la localización de targets, consiste en que el primero de ellos explota la dinámica de los targets en busca de una mayor eficiencia, eficacia y robustez.

Un modelo dinámico puede tener un mayor o menor grado de complejidad, de acuerdo a la naturaleza del movimiento que se desea modelar, o dicho de otra manera a lo predecible que sea su dinámica. Este modelado suele realizarse a través de un proceso Autorregresivo (ARP), pudiendo ser diferente en función del número de observaciones del vector de estados que se utilicen para estimar el vector de estados actual.

En el caso que aquí nos ocupa, parece evidente la utilización de un *espacio de observación* definido en $[x, y, z]$ (al que podremos añadir más parámetros en caso de necesidad), ya que esta es la información proporcionada por las cámaras dentro de un espacio inteligente. Por lo tanto, y de una manera general, la extracción de características de los targets de entorno se puede plantear como la obtención del vector de medidas de posición $[x, y, z]$, respecto a un sistema de coordenadas local, a partir de

la información sensorial proporcionada por las cámaras ubicadas en el entorno. Al vector de medidas así definido, se le denomina vector de estados ($\vec{x}_t$) [Ogata 1998] pag. 70, al cual se podrán añadir posteriormente los parámetros adicionales (velocidad, color…etc) necesarios hasta definir completamente el estado de cada uno de los targets y obtener la pose de cada uno, lo cual es otro de los objetivos de la tesis que aquí comienza.

Además del modelo *Constant Velocity (CV),* presentado en el siguiente apartado, existen diferentes modelos de actuación lineales que pueden ser utilizados para modelar la evolución del vector de estados (*Tethered, Brownian, Constrained Brownian, Damped oscillation, Constant Acceleration*…etc) [Springer 2005] pag. 302.

3.3.1 Modelo Constant Velocity (CV)

En este apartado se aborda con mayor grado de detalle el modelo de CV. Este modelo discreto de primer orden, permite obtener la evolución de la posición y velocidad de un target cualquiera, a partir de la información de posición cartesiana relativa al sistema de adquisición de medidas en un espacio tridimensional $[x, y, z]$. Las ecuaciones que definen el comportamiento de este modelo se presentan a continuación, en las cuales se supone que estamos trabajando en un plano con $y = cte$:

$$\begin{aligned} x_{t+1} &= x_t + vx_t \cdot T_s \\ y_{t+1} &= y_t \\ z_{t+1} &= z_t + vz_t \cdot T_s \end{aligned} \tag{3.1}$$

Las expresiones presentadas en (3.1) definen el comportamiento discreto del modelo planteado, cuya representación matemática en variables de estado permite describir cualquier sistema discreto, lineal e invariante, a través de las ecuaciones genéricas de estado y salida presentadas a continuación y gráficamente en la Figura 3.2:

$$\begin{aligned} \vec{x}_t &= A \cdot \vec{x}_{t-1} + B \cdot \vec{u}_t \\ \vec{y}_t &= C \cdot \vec{x}_t + D \cdot \vec{u}_t \end{aligned} \tag{3.2}$$

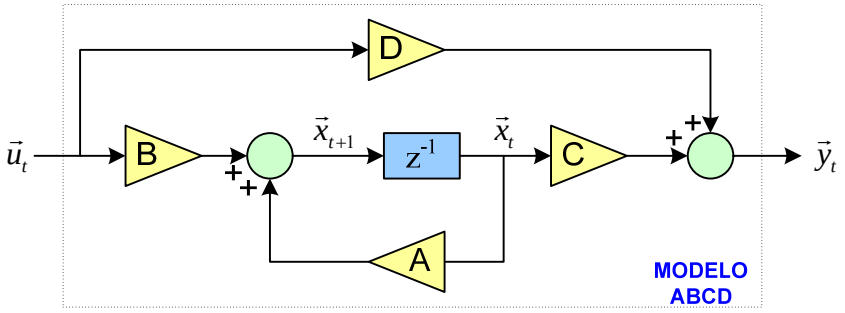


Figura 3.2 Diagrama de bloques ABCD de un sistema expresado en variables de estado

donde:

- $t \in N$ es el índice de tiempo.
- $\vec{x}_t \in R^n$ es el vector de estados.
- $\vec{u}_t \in R^r$ es el vector de control.
- $\vec{y}_t \in R^r$ es el vector de medida.
- $A \in R^{nxn}$ matriz de transición de estados.
- $B \in R^{nxr}$ matriz de aplicación de control
- $C \in R^{mxn}$ matriz de observación
- $D \in R^{mxr}$ matriz de transición directa

Haciendo $\vec{x}_t = [x_t \quad y_t \quad z_t \quad vx_t \quad vz_t]^T$, $\vec{y}_t = [x_t \quad y_t \quad z_t]^T$ y $\vec{u} = [0 \quad 0]^T$ se particulariza la formulación genérica (3.2) según las ecuaciones (3.1), obteniéndose:

$$\vec{x}_{t+1} = A \cdot \vec{x}_t \rightarrow \begin{bmatrix} x_{t+1} \\ y_{t+1} \\ z_{t+1} \\ vx_{t+1} \\ vy_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & T_s & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & T_s \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ z_t \\ vx_t \\ vz_t \end{bmatrix} \tag{3.3}$$

$$\vec{y}_t = C \cdot \vec{x}_t \rightarrow \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ z_t \\ vx_t \\ vz_t \end{bmatrix} \tag{3.4}$$

El modelo descrito por las ecuaciones (3.3) y (3.4) se denomina *Constant Velocity (CV)*.

Para evitar que el modelo sea válido únicamente para targets con movimiento constante se introduce un vector de ruido $\vec{v}_t = \begin{bmatrix} v_{x,t} & v_{y,t} & v_{z,t} & v_{vx,t} & v_{vz,t} \end{bmatrix}$ a la ecuación de estados (3.3), cuyas componentes se modelan como ruido blanco Gaussiano, de tal manera que las mismas permiten caracterizar las variaciones de velocidad del target así como el resto de incertidumbres propias del sistema de estimación.

Finalmente se completa la definición del sistema añadiendo otro vector de ruido blanco Gaussiano $\vec{o}_t = \begin{bmatrix} o_{x,t} & o_{y,t} & o_{z,t} \end{bmatrix}^T$ a las medidas de posición $\vec{y}_t$, obteniéndose las ecuaciones presentadas a continuación y gráficamente en la Figura 3.3:

$$\vec{x}_{t+1} = A \cdot \vec{x}_t + \vec{v}_t \rightarrow \begin{bmatrix} x_{t+1} \\ y_{t+1} \\ z_{t+1} \\ vx_{t+1} \\ vy_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & T_s & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & T_s \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ z_t \\ vx_t \\ vz_t \end{bmatrix} + \begin{bmatrix} v_{x,t} \\ v_{y,t} \\ v_{z,t} \\ v_{vx,t} \\ v_{vy,t} \end{bmatrix} \tag{3.5}$$

$$\vec{y}_t = C \cdot \vec{x}_t + \vec{o}_t \rightarrow \begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ z_t \\ vx_t \\ vz_t \end{bmatrix} + \begin{bmatrix} o_{x,t} \\ o_{y,t} \\ o_{z,t} \end{bmatrix} \tag{3.6}$$

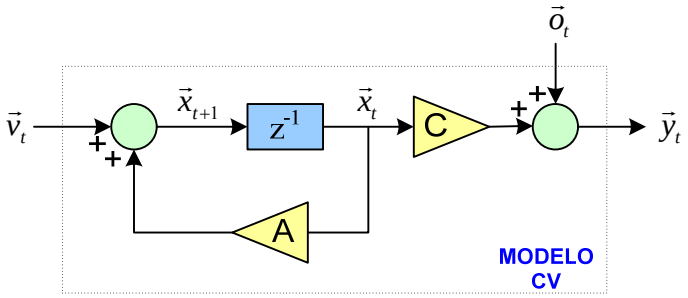


Figura 3.3 Diagrama de bloques del modelo CV

Es importante hacer notar en este punto que el movimiento de los targets puede ser en ciertos casos no lineal, como por ejemplo para personas, las cuales no siempre presentan un comportamiento lineal, con lo cual podría parecer incongruente utilizar un modelado lineal como el aquí presentado. La explicación de esto se encuentra en la alta frecuencia con la que se adquieren las imágenes, la cual permite modelar ciertos

comportamientos no lineales mediante sistemas lineales sin una pérdida apreciable de precisión.

Por otro lado el vector de estados empleado por el modelo CV anterior utiliza únicamente posiciones y velocidades. Dado que en nuestro caso los targets a seguir serán siempre personas, que la estimación de posición procedente del sistema de audio a partir de la cual se realizará la fusión sensorial hará siempre referencia a la posición de la boca de la persona, y que la cabeza humana tiene una forma de elipse "fácilmente" reconocible en sistemas de visión, podría ser interesante introducir algún parámetro relativo a dicha elipse dentro del vector de estados (e.g. centroide de la elipse, eje mayor, eje menor…etc). Otros parámetros interesantes podrían hacer referencia a blobs de color de piel o identificación de oclusiones entre targets, ambas utilizadas en trabajos estudiados posteriormente como [Khan 2003] o [Gatica-Perez 2007].

Es importante resaltar que la elección de este modelo probabilístico de posición hace necesario que el algoritmo de estimación empleado aproveche la incertidumbre incluida en el modelo para aumentar la robustez del seguimiento.

3.4 Algoritmos de estimación y asociación

Del estudio de la literatura científica relacionada, se desprende que el problema de seguimiento de múltiples targets abordado en el presente trabajo es un problema complejo no resuelto aún de forma general por la comunidad científica. La principal razón de esta complejidad radica en las difíciles características del entorno, lo cual implica la división del problema en dos procesos interrelacionados [Hue 2001]: La *estimación del vector de estados de cada target* y *la asociación de los mismos con alguno de los targets presentes en el entorno*.

1. *Proceso de estimación*: En este proceso, el modelo elegido para caracterizar el movimiento de los targets a seguir, las medidas de posición extraídas del entorno o el propio algoritmo de estimación constituyen los factores fundamentales de los que dependerá el éxito de esta tarea. La elección de modelos (desde el punto de vista de la teoría de control) y algoritmos que contemplen la incertidumbre

implícita en el entorno es de gran importancia para el correcto funcionamiento del mismo.

2. *Proceso de asociación:* Este proceso consiste en identificar cada dato de posición con el target del que ha sido extraído, siendo en general un problema pobremente resuelto por la comunidad científica debido principalmente a la alta carga computacional que conlleva. Otro de los problemas asociados es el de la creación y eliminación de hipótesis (procesos de *birth* y *death),* necesario en la aplicación de MTT aquí abordada tal y como se recoge en los objetivos generales presentados en 1.2.

Estos procesos se pueden llevar a cabo mediante *algoritmos determinísticos, probabilísticos* o una *combinación de ambos* [Marrón 2008]. A este respecto parece necesaria la utilización de algoritmos probabilísticos para el algoritmo de estimación, que aproveche la incertidumbre introducida en el modelado del movimiento para aumentar la robustez del algoritmo, mientras que para la parte de asociación esta elección no es tan inmediata y depende de las especificaciones y objetivos de la aplicación concreta.

A continuación se presenta un estudio teórico de las distintas soluciones aportadas por la comunidad científica para cada uno de los procesos anteriores, prestando especial atención al cumplimiento o no de las especificaciones indicadas en el apartado de introducción (e.g. multimodalidad, tiempo real, robustez…etc).

3.4.1 Algoritmos de estimación

El estudio aquí realizado se centra en los algoritmos de estimación *probabilísticos* por los motivos indicados en párrafos anteriores, los cuales están basados en la teoría de la probabilidad y más concretamente en el *"Filtro de Bayes"* [Thrun 2000] .

Las distintas particularizaciones de esta teoría dan lugar a distintos algoritmos probabilísticos [Arulampalam 2002], cuyo ámbito de aplicación depende de la aplicación concreta y que serán estudiados a continuación.

Por otro lado, e independientemente de las diferencias existentes entre las distintas particularizaciones, todas se caracterizan por una ejecución en dos etapas (*predicción* y *corrección*), y por implicar una representación de cualquier variable involucrada en el proceso a través de *funciones densidad de probabilidad (PDF)*. La diferencia entre unas particularizaciones y otras será el uso de funciones continuas o discretas para dichas PDF tal y como se muestra en la Figura 3.4.

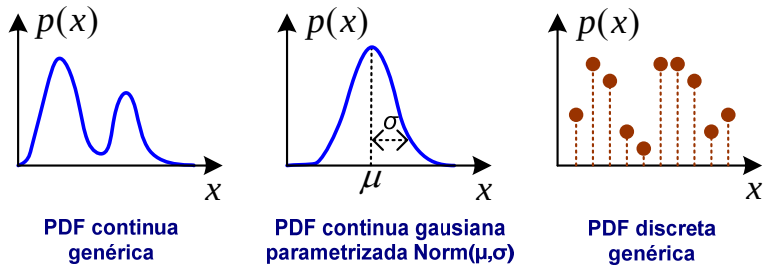


Figura 3.4 Ejemplos de las PDF más habituales

En la figura anterior, la función de la izquierda representa una PFD en su versión más general. Sin embargo el tratamiento de este tipo de variables es matemáticamente imposible en un contexto práctico, por lo que se recurre a otros formatos dentro de las funciones continuas, como el de la función central típica del KF (Kalman Filter) o recurriendo a funciones discretas como la de la derecha, típica de los PFs (Particle Filter), POMDPs (Partially Observable Markov Decision Process) y DBNs (Dynamic Bayesian Network).

3.4.1.1 Estimadores clásicos

Basados en el modelo en variables de estados del sistema a estimar (e.g el modelo CV mostrado en la Figura 3.2) constituyen el estimador probabilístico más simple.

El objetivo de este estimador recursivo, desde el punto de vista de la teoría de control aplicada al seguimiento de targets, consiste en estimar el vector de estados de dicho target. La formulación del mismo se presenta en la Figura 3.5, en la que el estimador se identifica con un recuadro en línea discontinua:

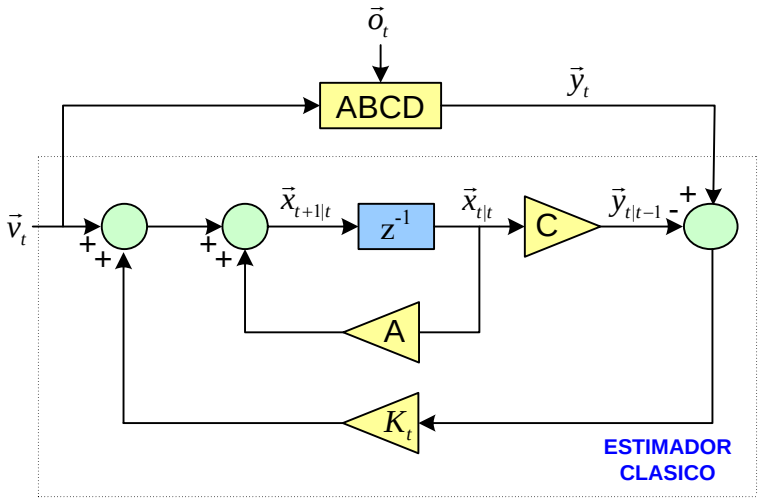


Figura 3.5 Diagrama de bloques del estimador clásico

En la figura anterior se pueden apreciar las dos etapas típicas de los estimadores probabilísticos:

- *Predicción* (desarrollada en el instante $t-1$): En esta primera etapa se utiliza el modelo del sistema para obtener una predicción del estado $\vec{x}_{t|t-1}$.
- *Corrección* (desarrollada en el instante t): En la segunda etapa se compara salida predicha $\vec{y}_{t|t-1}$ con la real $\vec{y}_t$, y a partir de este error y de la matriz de estimación K_t se corrige la evolución del estimador y se obtiene la estimación corregida del vector de estados $\vec{x}_{t|t}$.

La elección de K_t depende del algoritmo elegido, pero en cualquier caso siempre trata de minimizar el error, expresado matemáticamente como:

$$error_t = \left| \vec{x}_t - \vec{x}_{t|t} \right| \tag{3.7}$$

El estimador clásico presenta una serie de problemáticas que impiden su utilización en tareas de MTT:

- Es un estimador *unimodal*, por lo que se requeriría un estimador por cada target.
- *No es robusto ni fiable* ante incertidumbres del modelo, ya que el ruido que afecta a las medidas $\vec{o}_t$ y al estado $\vec{v}_t$ no se utiliza para estimar el vector de

estados. En el modelo CV presentado en 3.3.1 el ruido $\vec{v}_t$ sí se tiene en cuenta para la estimación del vector de estados, ya que en este caso se introduce de forma controlada.

3.4.1.2 El filtro de Kalman

El filtro de Kalman (KF) ya introducido en 2.4.3.1 permite obtener la solución que minimiza $error_t$ mediante la formalización matemática del estimador clásico recursivo presentado en el apartado anterior.

La formulación del filtro de Kalman impone las siguientes restricciones al modelo:

- El modelo debe ser lineal: El no cumplimiento de esta restricción por gran cantidad de modelos reales provoca la aparición del EKF y el UKF ya tratados con anterioridad (3.4.1.2).
- Se debe incorporar incertidumbre al modelo mediante sendos vectores de ruido, uno asociado al modelo ABCD ($\vec{v}_t$) y otro a las medidas ($\vec{o}_t$), los cuales deben estar caracterizados por Funciones Densidad de Probabilidad (PDF) normales de media nula, independientes y con matrices de covarianza V_t y O_t (calculadas mediante métodos de identificación de ruido), expresadas matemáticamente conforme a (3.8):

$$\begin{aligned} \vec{v}_t &= norm(0, V_t) \\ \vec{o}_t &= norm(0, O_t) \end{aligned} \tag{3.8}$$

Las etapas de predicción y corrección vienen definidas en este caso por [Kalman 1960]:

- *Predicción* (desarrollada en el instante $t-1$): Esta etapa utiliza la ecuación de estado del modelo para realizar una estimación a priori del vector de estados $\vec{x}_{t|t-1}$ y de la matriz de covarianza del error de estimación $P_{t|t-1}$.

$$\begin{aligned} \vec{x}_{t|t-1} &= A \cdot \vec{x}_{t-1|t-1} + B \cdot \vec{u}_{t-1} \\ P_{t|t-1} &= A \cdot P_{t-1} \cdot A^T + V_{t-1} \end{aligned} \tag{3.9}$$

- *Corrección* (desarrollada en el instante t): En esta etapa se emplean el nuevo conjunto de medidas $\vec{y}_t$, la ecuación de salida del modelo, y la matriz de covarianza de error de estimación $P_{t|t-1}$, para calcular la matriz de estimación K_t (en este caso denominada matriz de Kalman), a partir de la cual se actualiza la estimación del vector de estado $\vec{x}_{t|t}$ y la covarianza del error de estimación para ese instante P_t.

$$\begin{aligned} K_t &= P_{t|t-1}\left(C \cdot P_{t|t-1} \cdot C^T + O_t\right)^{-1} \\ \vec{x}_{t|t} &= \vec{x}_{t|t-1} + K_t \cdot \left(\vec{y}_t - C \cdot \vec{x}_{t|t-1}\right) \\ P_t &= \left(I - K_t C\right) \cdot P_{t|t-1} \end{aligned} \tag{3.10}$$

La formulación anteriormente presentada hace que el error de estimación también siga una distribución normal parametrizada según:

$$error_t = \left|\vec{x}_t - \vec{x}_{t|t}\right| = norm\left(0, P_t\right) \tag{3.11}$$

A la vista de lo presentado anteriormente el KF no es el algoritmo más adecuado para utilizarlo en aplicaciones de seguimiento como la propuesta debido a que:

- Al tratarse de un estimador unimodal impide su utilización para aplicaciones de MTT (se requeriría un estimador por cada target de la escena).
- La no linealidad de los sistemas implicados hace necesaria una linealización de los mismos, lo que conlleva una perdida de precisión y un proceso de calibración de parámetros no siempre trivial.

De cualquier manera la comunidad científica ha utilizado este tipo de filtros en tareas de seguimiento empleando para ello un estimador por cada target de la escena como por ejemplo en [Rosales 1999] o [Zhao 2004].

3.4.1.3 Estimadores Bayesianos

La formalización de este filtro pasa por establecer un nuevo modelo para el sistema, que en lugar de definirse por una ecuación de estado y otra de salida se define en base a varias PDF:

- *Probabilidad a posteriori o creencia,* $p(\vec{x}_t \mid \vec{y}_{1:t})$: Representa la salida del estimados bayesiano y representa la probabilidad de que el vector de estados del sistema sea $\vec{x}_t$, teniendo en cuenta toda la historia del vector de medidas $\vec{y}_t$.
- *Modelo de actuación o de estado,* $p(\vec{x}_t \mid \vec{x}_{t-1})$: Caracteriza la evolución del vector de estados.
- *Verosimilitud, Modelo de percepción u observación,* $p(\vec{y}_t \mid \vec{x}_t)$: Caracteriza el proceso de medida y viene definida por el modelo de observación del sistema. Cuantifica la probabilidad de una medida $\vec{y}_t$ teniendo en cuenta que el mundo observado se encuentra en el estado $\vec{x}_t$.
- *Probabilidad a priori,* $p(\vec{x}_t \mid \vec{y}_{1:t-1})$: Puede considerarse como una primera aproximación a la creencia.
- *Probabilidad total del vector de medidas,* $p(\vec{y}_t \mid \vec{y}_{1:t-1})$.

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

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \frac{p(\vec{y}_t \mid \vec{x}_t, \vec{y}_{1:t-1}) \cdot p(\vec{x}_t \mid \vec{y}_{1:t-1})}{p(\vec{y}_t \mid \vec{y}_{1:t-1})} \tag{3.12}$$

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

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \frac{p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{y}_{1:t-1})}{p(\vec{y}_t \mid \vec{y}_{1:t-1})} \tag{3.13}$$

Si se tiene en cuenta que el conjunto completo de creencias ha de sumar la unidad, el valor del denominador se suele utilizar como factor de normalización η, obteniéndose:

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \eta \cdot p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{y}_{1:t-1}) \tag{3.14}$$

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

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \eta \cdot p(\vec{y}_t \mid \vec{x}_t) \cdot \int p(\vec{x}_t \mid \vec{x}_{t-1}, \vec{y}_{1:t-1}) \cdot p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1}) \partial \vec{x} \quad (3.15)$$

Aplicando de nuevo la *condición de Markov* se obtiene la *formulación compacta del filtro de Bayes*:

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \eta \cdot p(\vec{y}_t \mid \vec{x}_t) \cdot \int p(\vec{x}_t \mid \vec{x}_{t-1}) \cdot p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1}) \partial \vec{x} \quad (3.16)$$

En la formulación anterior se identifican de nuevo las dos etapas típicas de los filtros probabilísticos:

- *Predicción:* En la que la estimación inicial del vector de estados (probabilidad a priori $p(\vec{x}_t \mid \vec{y}_{1:t-1})$) en el instante t se calcula a partir de la creencia obtenida en el instante $t-1$ ($p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1})$) y del modelo de actuación del sistema $p(\vec{x}_t \mid \vec{x}_{t-1})$.

$$p(\vec{x}_t \mid \vec{y}_{1:t-1}) = \int p(\vec{x}_t \mid \vec{x}_{t-1}) \cdot p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1}) \partial \vec{x} \quad (3.17)$$

- *Corrección:* Realizada en función de la función de verosimilitud $p(\vec{y}_t \mid \vec{x}_t)$ y que permite obtener la creencia en el instante $t-1$.

Tal y como se ha comentado en anteriores ocasiones, el KF es una particularización del Filtro de Bayes en el que:

- Las PDF asociadas al modelo de estados $p(\vec{x}_t \mid \vec{x}_{t-1})$ y a la verosimilitud $p(\vec{y}_t \mid \vec{x}_t)$ son continuas y gaussianas.
- La PDF asociada a la creencia es continua y normal: $p(\vec{x}_t \mid \vec{y}_{1:t}) = norm(\vec{x}_{t|t}, P_t)$ y su media determina la salida determinística del KF.

La aplicación directa de la fórmula compacta del filtro de Bayes en el espacio continuo es complicada debido a que introduce en la etapa de predicción una integral que debe contemplar todos los posibles valores del vector de estados. Por ello la única versión en el espacio continuo del filtro de Bayes es el KF, el cual introduce una simplificación que restringe las PDF utilizables a gaussianas parametrizadas.

El resto de versiones del filtro de Bayes se obtienen mediante la discretización de alguna de las PDF implicadas en el mismo, por lo que se pueden clasificar en base a ello:

- *Localización de Markov (Markov Localization, ML)*: Propone una discretización del espacio de estados (las distintas posiciones dentro del espacio de interés) mediante un enfoque métrico y de la creencia $p(\vec{x}_t \mid \vec{y}_{1:t})$, obteniéndose en ambos sendos enrejillados. Esta solución se emplea sobre todo en mapeado, en el que la rejilla representa el valor estocástico de ocupación de todos los posibles localizaciones del entorno (estados), y en menor medida en aplicaciones de SLAM (Simultaneous Location and Maping) e incluso localización y tracking.

- *Procesos de Decisión de Markov (Markov Decisión Processes, MDP) y Procesos de Markov Parcialmente Observables (Partially Observable Markov Decisión Processes, POMDP):* En este caso la discretización del espacio de estados se realiza desde un punto de vista topológico, siendo necesaria también la discretización del modelo de actuación $p(\vec{x}_t \mid \vec{x}_{t-1})$, de la función de verosimilitud $p(\vec{y}_t \mid \vec{x}_t)$ y de la creencia $p(\vec{x}_t \mid \vec{y}_{1:t})$. Este tipo de solución se utiliza habitualmente tareas de planificación.

- *Filtro de Partículas (Particle Filter, PF) o Localizador de Monte-Carlo (Monte-Carlo Localization, MCL):* Esta solución requiere únicamente la discretización de la creencia $p(\vec{x}_t \mid \vec{y}_{1:t})$ mediante un conjunto de muestras n denominadas partículas ($S_t = \{\vec{s}_{i,t} / i = 1:n\}$), que son copias del vector de estado $\vec{x}_t$ modificadas por la incertidumbre propia de la creencia y que tienen un peso asociado a cada una de ellas proporcional a su probabilidad individual ($\vec{w}_t = \{w_t^{(i)} / i = 1:n\}$). La formalización matemática de la creencia queda definida como:

$$\begin{gathered} p(\vec{x}_t \mid \vec{y}_{1:t}) \cong \{S_t, \vec{w}_t\} = \{\vec{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{n} \\ S_t = \{\vec{s}_{i,t} = \vec{x}_t^{(i)} / i = 1:n\} \\ \vec{w}_t = \{w_t^{(i)} / i = 1:n\} \end{gathered} \tag{3.18}$$

La discretización de las PDF propuestas por cada uno de los métodos conlleva unas ventajas e inconvenientes que se recogen a continuación:

- Ventajas: Facilita el proceso de cálculo de la etapa de predicción del filtro, representado por la parte integral de la formula compacta del Filtro de Bayes, el cual pasa a definirse en todos los casos como sigue:

$$p(\vec{x}_t \mid \vec{y}_{1:t-1}) = \sum_{\forall \vec{x}} p(\vec{x}_t^{(i)} \mid \vec{x}_{t-1}^{(i)}) \cdot p(\vec{x}_t^{(i)} \mid \vec{y}_{1:t-1}) \tag{3.19}$$

- Inconvenientes: Ineficiencia en el coste computacional e inexactitud del algoritmo final.

Los modelos de rejilla propios de la discretización impuesta por el ML, implican un compromiso entre precisión y coste computacional. Además, en aplicaciones de localización como la que nos ocupa, la rejilla se ve infrautilizada debido al espacio métrico de representación que utiliza.

El enfoque topológico propuesto por MDPs y los POMDPs sólo permite que se incluyan en el espacio de estados discretizado los valores de interés para la aplicación. Sin embargo, la necesidad de almacenar en tablas los modelos de actuación y observación a través de un proceso off-line, para luego ser utilizados en el proceso on-line del filtro implica problemas de coste computacional. Adicionalmente, la utilización de un enfoque topológico en lugar de métrico impide definir su exactitud.

La principal ventaja del PF sobre el resto reside en la no discretización del espacio de representación del vector de estados, requiriendo únicamente la discretización de la creencia $p(\vec{x}_t \mid \vec{y}_{1:t})$, en cuya representación sólo aparecen muestras de los valores más probables de la misma, mejorando así la exactitud en la estimación con respecto al resto de soluciones, conservando la robustez frente a incertidumbres del modelo que caracteriza a los estimadores bayesianos y con una carga computacional que le permite ejecutarse en tiempo real.

Las características anteriores, unidas a la multimodalidad del filtro que permite plantear múltiples hipótesis de estimación con un único algoritmo, hacen que el PF sea el algoritmo utilizado de manera más recurrente por la comunidad científica para tareas de seguimiento de múltiples targets como la aquí abordada.

Debido a las ventajas del PF enumeradas en párrafos anteriores y a su amplia utilización en aplicaciones de MTT como la aquí presentada, en el Anexo I se presenta un estudio detallado del mismo.

3.4.2 Algoritmos de asociación

Una vez expuestos los principales algoritmos de estimación, se aborda el análisis de las técnicas de asociación, que tal y como se introdujo anteriormente son las encargadas de identificar cada una de las m medidas incluidas en el set de datos $Y = \{\vec{y}_i / i = 1:m\}$, con los targets k presentes en la escena $X = \{\vec{x}_j / j = 1:k\}$ de manera que puedan ser utilizadas por el algoritmo de estimación. A cada una de las posibles asociaciones de medidas y targets se la conoce como hipótesis y se representa matemáticamente como $\theta_{i,j} / i = 1:m, j = 1:k$.

Tal y como se puede intuir, el algoritmo de asociación tiene una importancia vital sobre el resultado final del algoritmo completo, ya que la asociación de una medida con el target incorrecto puede hacer divergir la posterior estimación de éste y el resto de los targets.

Por otro lado, tal y como se presentó en 3.1, la incertidumbre propia del sistema de medida puede provocar la aparición de outliers dentro de nuestro set de datos, los cuales no deben ser asignados a ningún target. Una solución habitualmente implementada por la comunidad científica consiste en definir un target adicional $k+1$, relacionado con todos estos outliers y que sólo se tendrá en cuenta en el proceso de asociación y no en el de estimación, con lo que las hipótesis de asociación pasan a definirse como $\theta_{i,j} / i = 1:m, j = 0:k$.

Desde el punto de vista de carga computacional nos encontramos ante un problema complejo, ya que el algoritmo debe seleccionar la hipótesis de asociación óptima siendo el número de estas exponencial y función del número de targets $k+1$ y de medidas m (generalmente alto): $(k+1)!^m$.

Adicionalmente, existen determinados factores que complican el algoritmo de asociación en nuestra aplicación concreta:

- *El número de muestras* a asociar a cada target es variable, desconocido y generalmente alto.
- *El número de targets* es variable en las aplicaciones de MTT como la de interés, con lo que no se puede aplicar la simplificación de considerar su número constante, siendo necesario o bien generalizar su ejecución para adaptarla a un número variable o diseñar procesos creación y eliminación de targets, añadiéndolos al algoritmo de asociación básico. Otra posible técnica para resolver esta problemática, aunque menos eficiente que la anterior, consiste en suponer un número constante y suficientemente alto de targets que en principio se encuentran ocultos e irlos haciendo visibles conforme aparezcan nuevos targets en la escena.

Una vez presentadas las consideraciones anteriores se procede al estudio de los diferentes algoritmos de asociación [Bar-Shalom 1995], que al igual que en el caso de los de estimación se pueden dividir en *determinísticos u orientados a medidas* y *probabilísticos u orientados a targets*.

3.4.2.1 Determinísticos – Orientados a medidas

La denominación como orientados a medidas de los algoritmos determinísticos, se debe a que responden a la pregunta de qué medida(s) $\vec{y}_i / i = 1:m$ asociar a cada target $\vec{x}_j / j = 1:k$ y permiten obtener la solución óptima de asociación para número de targets k desconocido y variable.

Al ser determinísticos, este tipo de algoritmos asignan a cada hipótesis de asociación una certeza absoluta, por lo que para cada $\theta_{i,j}$ se obtiene una $p_{i,j} = [0,1]$, la cual se mantiene a lo largo del tiempo, $1:t$. Como contrapartida, su implementación es más costosa en términos de carga computacional que en el caso de los probabilísticos.

3.4.2.1.1 Multiple Hypotheses Tracker (MHT)

Propuesto por primera vez en [Reid 1979], este algoritmo determinístico que calcula en cada instante de tiempo t las $(k+1)!^m$ hipótesis de asociación, está

especialmente indicado para tareas de MTT pues supone que cada medida puede proceder de alguno de los targets $\vec{x}_{1:k}$ supuestos en el instante $t-1$, del clutter o de un nuevo target, incluyendo así de forma implícita el proceso de creación y eliminación de targets.

Debido a la cantidad de hipótesis manejadas, el algoritmo presenta una carga computacional excesiva y lo hace impracticable en pocas iteraciones (sobre todo si el número de targets y medidas es elevado). Para solucionar este problema se proponen en la literatura científica diferentes simplificaciones, que hacen que el algoritmo deje de ser global, pero lo hace implementable en tiempo real. Las más importantes se presentan a continuación:

a. *Procesos de pruning y merging:* Los primeros permiten eliminar las hipótesis menos probables, mientras que los segundos aúnan las semejantes en una sola.

b. *Ventanas temporales:* Esta técnica se basa en el establecimiento de una ventana temporal del conjunto de hipótesis $\theta_{1:m,0:k}$, de manera que no se tenga en cuenta toda la historia temporal de medidas $Y_{1:t}$ y de targets $X_{1:t}$ en el proceso de asociación, sino sólo las que siguen manteniéndose en las últimas iteraciones del algoritmo ($t-a:t$), siendo el resto eliminadas.

c. *Procesos de gating:* Consiste en establecer un perímetro de semejanza alrededor de cada target $\vec{x}_{1:k}$ en el espacio de definición del proceso de asociación, de manera que sólo las medidas que se encuentran dentro de dicho perímetro son asociadas a él.

   Este perímetro de gating se puede utilizar también para limitar la creación de nuevos targets a situaciones en las que las medidas caen fuera del perímetro de todos los targets.

   La reducción de la carga computacional proporcionada por este proceso es mayor que la de los dos anteriores, ya que este proceso se realiza previamente al cálculo de las hipótesis $\theta_{1:m,0:k}$, con lo que se consigue una reducción de las

mismas a $\theta_{1:m_j,0:k}$, donde m_j representa el número de medidas que se encuentran dentro del perímetro de cada target.

Sin embargo, esta reducción de la carga computacional se traduce en una disminución de la robustez del algoritmo (posible pérdida de hipótesis probables), el cual se hace muy sensible al valor de gating seleccionado. Dicho perímetro se suele elegir en base a la incertidumbre asociada a los modelos de actuación y observación.

d. Procesos de clustering: Se basan en una reducción del número de hipótesis, conseguida mediante la asociación a cada hipótesis con un conjunto de medidas en lugar de una sola. La complejidad reside en este caso en definir el número de clases en las que se organiza el set de medidas, de lo que dependerá en gran medida robustez del algoritmo.

La utilización de este tipo de algoritmo de asociación en combinación alguno de los algoritmos de estimación propuestos anteriormente es muy común en la literatura científica, como por ejemplo en [Reid 1979] donde se combina con un KF, en [Coraluppi 2004] donde se usa un EKF, o en [Hue 2001] en combinación con un PF.

3.4.2.1.2 Nearest Neighbour (NN)

El NN puede considerarse como una simplificación del MHT en la que sólo se mantiene una hipótesis de asociación por cada target $\vec{x}_{1:k}, \theta_{0:k}$, siendo la hipótesis seleccionada la que asocia cada target con la medida o medidas que se encuentran más próximas dentro del espacio de asociación en cada instante t de ejecución del algoritmo.

Este algoritmo genera por tanto una solución subóptima, al proporcionar únicamente las k teóricas mejores hipótesis de asociación por cada target, con lo que elimina totalmente la incertidumbre en el proceso de asociación y consiguientemente la robustez del método antes situaciones complejas como oclusiones o movimientos inciertos.

El principal punto fuerte de este algoritmo reside en su simplicidad, lo cual implica bajos tiempos de cómputo y lo hace útil en determinadas situaciones como problemas de asociación sencillos (pocos targets con modelos de verosimilitud muy específicos) o un número de medidas lo suficientemente alto como para impedir la implementación de cualquier otro algoritmo en tiempo real.

Además el NN es un método válido para un número k de targets conocido, por lo que su utilización dentro del entorno de MTT implica la implementación de un proceso de gating para la eliminación y creación de targets y la asociación de las medidas que se encuentren fuera del gate de todos los targets al clutter.

Otra situación problemática se produce cuando una misma medida $\vec{y}_i$ es la más cercana a dos o más target $\vec{x}_{1:k}$ en el espacio de clasificación. Una solución posible consiste en incorporar un proceso de selección, basado en búsqueda de la organización que minimiza la suma de las distancias entre cada medida $\vec{y}_i$ y el objeto $\vec{x}_j$ al que esta se asigna.

3.4.2.2 Probabilísticos – Orientados a targets

En el caso de los algoritmos orientados a target, la pregunta a la que responden es que target $\vec{x}_j / j = 1:k$ asociar a cada medida $\vec{y}_i / i = 1:m$, generando siempre soluciones subóptimas en las que es necesario añadir procesos específicos para la creación y eliminación de targets.

Su carácter probabilístico conlleva que a su salida se obtenga una certidumbre $p_{i,j}$ para cada hipótesis $\theta_{i,j}$. Este tipo de algoritmos están más extendidos dentro de la literatura científica que los determinísticos e implican una carga computacional menor a pesar de tener que calcular los valores de las probabilidades $p_{1:m,0:k}$.

3.4.2.2.1 Probabilistic Data Association (PDA)

El objetivo de este algoritmo consiste en obtener la probabilidad de que cada medida i esté asociada a un único target k_1 o bien al clutter ($p_{1:m,j} / j = 0,1$), de lo que se deduce que este algoritmo es válido para un único target y que genera únicamente

dos hipótesis de asociación $\theta_{0:1}$, caracterizadas cada una de ellas por un conjunto m de valores de probabilidad.

Con el objetivo de reducir los tiempos de cómputo es habitual el aplicar al PDA procesos de gating y pruning.

En tareas de seguimiento de un único target, la combinación del PDA con el KF constituye la solución más recurrentemente utilizada por la comunidad científica [Rasmussen 2001]. La combinación de ambos recibe el nombre de PDAF (Probabilistic Data Association Filter). Los casos de aplicaciones con modelos de movimiento no lineales se solucionan sustituyendo el KF por un EKF, el UKF o un PF.

3.4.2.2.2 Joint Probabilistic Data Association (JPDA)

El JPDA representa la extensión del PDA para aplicaciones de MTT $(K+1>2)$, y se basa en la implementación de tantos PDAs como targets existentes en la escena, incorporando además la probabilidad conjunta de existencia de las distintas hipótesis (lo cual permite resolver las interacciones entre targets como cruces u oclusiones).

En base a esto, cada valor de probabilidad se calcula condicionado al resto, obteniéndose $(k+1)!^m$ valores de probabilidad conjunta $p_{1:m,0:k}$ que caracterizan a las $k+1$ hipótesis de asociación $\theta_{0:k}$, en lugar de m valores de probabilidad disjunta $p_{1:m,j}$ para cada una de ellas $\theta_j \,/\, j=0:k$, lo que daría lugar a k PDAs independientes.

Al igual que para los algoritmos anteriores, es habitual la implementación de procesos de gating y pruning para reducir la elevada carga computacional del algoritmo, la cual constituye la mayor problemática del mismo.

El JPDA representa uno de los procesos de asociación más utilizados en tareas de MTT debido a su gran fiabilidad, equiparable a la solución global aportada por el MHT pero con una carga computacional sensiblemente inferior.

La combinación del JPDA con el KF da lugar al JPDAF (Joint Probabilistic Data Association Filter). Esta combinación consiste en la aplicación de las hipótesis de asociación $\theta_{0:k}$ generadas por el JPDA a k procesos de estimación de la posición

individuales (uno para cada target) basados en el KF. El JPDAF se utiliza en diferentes trabajos de la literatura científica, como por ejemplo [Rasmussen 2001].

La principal problemática del JPDAF aparece cuando se trabaja en el seguimiento de targets con un modelo de movimiento no gaussiano (no lineal) y que por tanto no se ajusta a las especificaciones del filtro de Kalman utilizado en la etapa de estimación. En estos casos el EKF y el UKF no aseguran la convergencia del estimador, lo que da lugar a la aparición del S-JPDAF (Sampled JPDAF) [Schulz 2001] que emplea un PF como algoritmo de estimación.

Los métodos de estimación que emplean el S-JPDAF suelen proponer la utilización de un vector de estados aumentado que contenga el de cada uno de los targets implicados $\chi_t = \{\vec{x}_{1,t}, \vec{x}_{2,t}, ..., \vec{x}_{k,t}\} = \bigcup_{j=1}^{k} \vec{x}_{j,t}$, de modo que cada hipótesis θ_j queda representada por una o más partículas del set S_t, quedando definida la probabilidad de cada hipótesis por el peso $w_t^{(i)}$ asociado a cada partícula $\vec{s}_{i,t}$.

Las expresiones matemáticas de los procesos de predicción y corrección del S-JPDAF se presentan a continuación:

$$\begin{aligned} &predicción \rightarrow p(\vec{x}_{j,t} \mid \vec{y}_{1:t-1}) = \int p(\vec{x}_{j,t} \mid \vec{x}_{j,t-1}) \cdot p(\vec{x}_{j,t-1} \mid \vec{y}_{1:t}) \cdot \partial \vec{x}_j \\ &correccion \rightarrow p(\vec{x}_{j,t} \mid \vec{y}_{1:t}) = \eta \cdot p(\vec{x}_{j,t} \mid \vec{y}_{1:t-1}) \cdot \sum_{i=1}^{m_j} p_{i,j} \cdot p(\vec{y}_{i,t} \mid \vec{x}_{j,t}) \end{aligned} \tag{3.20}$$

La etapa de predicción es la típica de los estimadores bayesianos presentados en 3.4.1.3, mientras que en la etapa de corrección se utilizan los valores de probabilidad $p_{i,j}$ obtenidos por el algoritmo de asociación para ponderar el valor de verosimilitud $p(\vec{y}_{i,t} \mid \vec{x}_{j,t})$ asociado a cada medida $\vec{y}_i$ en el entorno de gating del target j.

Al igual que sucedía con el JPDAF, el principal inconveniente de los estimadores basados en S-JPDAF y la expansión del vector de estados es el crecimiento exponencial de la carga computacional con el número de targets k de la escena, siendo incluso más acusado en el caso del S-JPDAF ya que el tiempo de ejecución el PF supera con facilidad al del KF del JPDAF con número de partículas moderado.

Como alternativa al S-JPDAF con vector de estados extendido χ_t, se presenta en [Tweed 2002] o [Khan 2003] entre otros, donde se propone la utilización de PFs "casi" independientes para el proceso de estimación, donde la dependencia entre los mismos la aporta el JPDA empleado en el proceso de evolución y validación de las hipótesis de asociación. De esta forma se consigue reducir la carga computacional y conseguir que su crecimiento sea lineal en lugar de exponencial con el número de targets k.

Una solución elegante a la problemática presentada con anterioridad se presenta en [Marrón 2008], donde la autora propone utilizar la multimodalidad de la creencia $p(\vec{x}_{j,t} \mid \vec{y}_{1:t})$ del PF para caracterizar con una única PDF la posición de un número múltiple de targets, en combinación con un método de asociación determinístico (k-medias secuencial), de modo que cada modo de la creencia represente la posición de un target, manteniendo la flexibilidad, fiabilidad y robustez de otras alternativas pero cumpliendo con la restricción de tiempo real, también objetivo de la tesis que aquí comienza y que no cumplen ni los algoritmos basados en MHT y en JPDA.

3.5 Líneas de investigación en el área de seguimiento visual

Los trabajos presentados en este apartado se prestan a una primera división en *función del número de targets* que son capaces de seguir, centrándose en primer lugar en los seguidores de un único target y posteriormente en los de múltiples targets.

A su vez, los trabajos en los que se aborda la tarea de MTT como la de interés, se dividen *en función del algoritmo implementado* para realizar dicho seguimiento tal y como se detalla en apartados sucesivos.

3.5.1 Seguimiento de un único target

Este tipo de aplicaciones y sus algoritmos asociados representan el punto de partida de las tareas de MTT, en las cuales se simplifica el proceso de asociación (cada medida puede proceder únicamente del target de interés o del clutter).

Uno de los primeros trabajos en el área de seguimiento de targets empleando visión como sensor para la extracción de información de la escena y PFs como algoritmos de estimación, fue el desarrollado por investigadores de la Universidad de Oxford en **[Isard 1998a]** en el que propone por primera vez el popular el algoritmo *Condensation* (Condicional Density Propagation).

En este trabajo se plantea la implementación de un algoritmo de seguimiento probabilístico de objetos de la escena, los cuales están modelados mediante una curva (en este caso mediante una spline). Este tipo de modelos no lineales impiden la implementación de un KF como algoritmo de estimación, lo cual conduce a la implementación de un PF.

La validez del algoritmo se evalúa bajo diferentes secuencias de imágenes complejas, una de las cuales se muestra en la Figura 3.6:

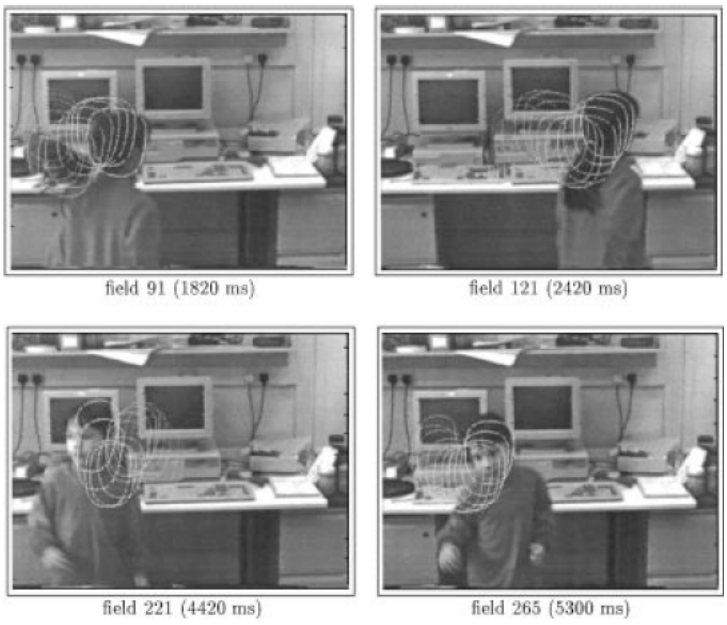


Figura 3.6 Resultados del algoritmo Condensation [Isard 1998a]

En los resultados presentados en la figura anterior, se realiza el tracking de la cabeza de un niño mientras baila. La forma de la curva que modela la forma de la cabeza (modelo de observación) se establece bien de manera manual al inicio del proceso de seguimiento, o mediante un proceso de aprendizaje off-line basado en PCA (Principal Components Analysis). Las elipses blancas de la figura representan la

evolución de la posición de la cabeza extraída en tiempo real (50fps), en un procesador a 200MHz y 100 partículas.

Este tipo de algoritmo de observación entraría dentro de los denominados en 3.3 como complejos y específicos, e implica necesariamente un aumento del número de partículas para obtener una aproximación fiable de cada uno de los parámetros que lo componen, lo cual los hace impracticables en aplicaciones de tiempo real ([Hue 2001]).

Posteriores trabajos de este grupo de la universidad de Oxford se centran en el área del reconocimiento visual de gestos, haciendo necesaria una mayor complicación de los modelos de observación utilizados y por consiguiente el aumento del número de partículas. Por ello se hace necesario modificar el algoritmo Condensation y así permitir la ejecución de los algoritmos en tiempo real, para lo cual este grupo de investigación propone dos soluciones que serán posteriormente utilizadas de manera recurrente por la comunidad científica.

La primera de ellas se conoce como *ICondensation* (Importante Sampling Condensation) **[Isard 1998b]** (Figura 3.7), el cual incorpora al Condensation una función de verosimilitud mejorada basada en distintas fuentes de medida (blobs de color, forma, contorno…etc de los gestos de la mano), mejorando así la etapa de corrección del PF y permitiéndole adaptarse a movimientos bruscos de las manos. En la Figura 3.7 se emplean del orden de 400 partículas.

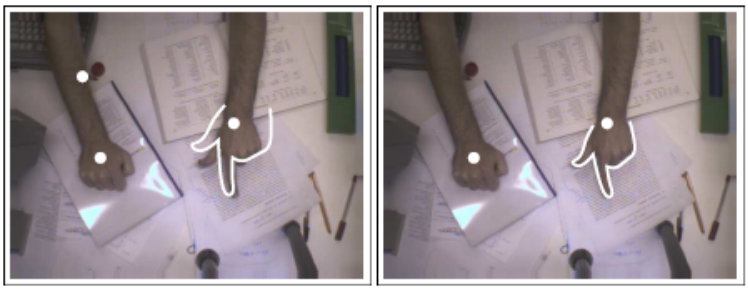


Figura 3.7 Resultados del algoritmo ICondensation [Isard 1998b]

La segunda se conoce como *Partitioned Sampling* (PF con Muestreo Segmentado) **[MacCormick 2000]** (Figura 3.8) y consiste en una descomposición de la función de verosimilitud $p(\vec{y}_t \mid \vec{x}_t)$ en varias funciones con el objetivo de hacer más eficiente el paso de corrección del PF y por tanto reducir el número de partículas

necesarias. Según el autor, el PF con muestreo segmentado proporciona una fiabilidad semejante a la del algoritmo básico con la cuarta parte de partículas, consiguiendo así el objetivo de reducción de la carga computacional.

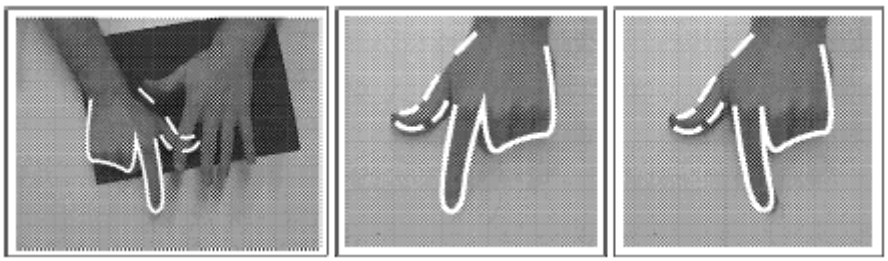


Figura 3.8 Resultados del algoritmo Partitioned Sampling [MacCormick 2000]

En **[Sullivan 2001]** investigadores de la universidad KTH de Estocolmo, en colaboración con los creadores del algoritmo Condensation de la universidad de Oxford, proponen la combinación de métodos determinísticos con los probabilísticos del PF para realizar el seguimiento de los targets de la escena. Un ejemplo de estos trabajos se presenta en la Figura 3.9.

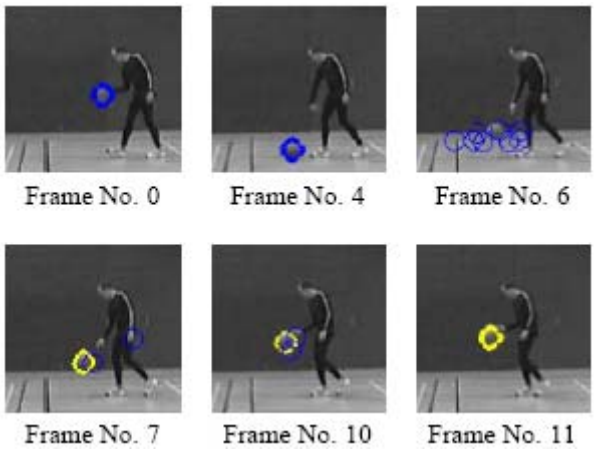


Figura 3.9 Resultado en la combinación de métodos probabilísticos y determinísticos en [Sullivan 2001]

La combinación de métodos probabilísticos y determinísticos, permite la utilización de un modelo de actuación o movimiento simple como el CV (Constant Velocity) presentado en 3.3.1. Sin embargo, el modelo de observación implementado es complejo y basado en la forma y el color del target a seguir, lo cual lo hace específico para la aplicación concreta.

Una desventaja de este trabajo es que la inicialización del proceso de seguimiento se realiza de forma manual, no proporcionando tampoco resultados de tiempo de ejecución ni del número de partículas usadas, por lo que no se puede establecer si el mismo es ejecutable en tiempo real.

Trabajos posteriores basadas en el ICondensation tienden a implementar modelos más complejos y específicos si cabe, como en **[Chen 2002]** Figura 3.10, donde el modelo de observación consiste en una elipse de 30 parámetros, por lo que el estimador trabajando con 20 partículas se ejecuta a una velocidad de 10fps en un procesador de 933 MHz.

Figura 3.10 Ejemplo de resultado en seguimiento de caras basado en ICondensation [Chen 2002]

El éxito del ICondensation en tareas de seguimiento de caras se debe a que permite la combinación de modelos de observación sencillos (e.g. blobs de color de piel) con otros más complejos (e.g. contorno de la cabeza). Este es el caso de **[Hu 2004]** Figura 3.11 donde se implementa simultáneamente un seguimiento de caras y el reconocimiento de gestos en las mismas, empleando para ello un modelo de formas y de deformación de caras.

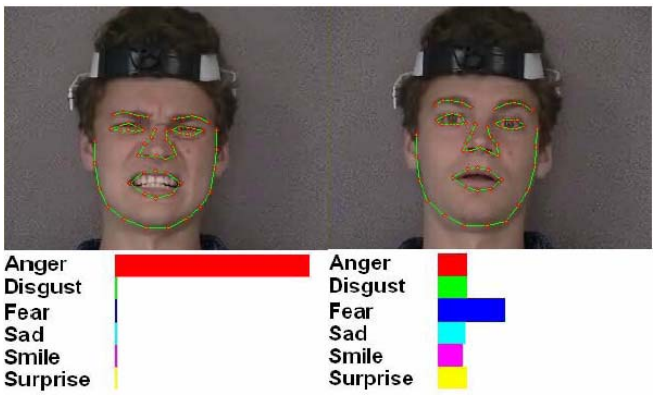


Figura 3.11 Resultados en seguimiento de caras e identificación de gestos basados en el algoritmo ICondensation [Hu 2004]

Otra línea de investigación con gran aceptación en la comunidad científica, es la de simplificar el PF de manera que éste sólo estima parte del vector de estados y que se conoce como *Rao-Blackwellized PF* (RBPF). Valga como ejemplo el trabajo de **[Torma 2003]** Figura 3.12.

El modelo de observación se divide en una parte determinística y otra estocástica realizada por un RBPF, las cuales se combinan mediante el ICondensation. La caracterización del target se realiza a través de una curva spline.

Figura 3.12 Resultados en seguimiento de objetos mediante RBPF en [Torma 2003]

El experimento representado en la figura anterior se ejecuta en un procesador de 1.4 GHz a una frecuencia de 30fps y con 400 partículas (por las 2000 Condensation propuesto en [Isard 1998a] según los autores), con lo que es más fácil alcanzar la especificación de tiempo real.

En referencia a los algoritmos de asociación utilizados, su implementación se simplifica considerablemente al encontrar únicamente un target en la escena y al uso de modelos de observación complejos y específicos, por lo que la mayoría de los trabajos se decantan por un NN de manera que se elige la hipótesis de asociación más semejante a la predicha según el modelo de seguimiento.

Otro grupo de investigación puntero en lo que a localización y tracking de personas se refiere es el laboratorio IDIAP (Suiza), los cuales no sólo utilizan visión como elemento sensor, sino también arrays de micrófonos (algunos de sus trabajos ya han sido referenciados en la parte de audio (e.g. [Lathoud y Odobez 2007]) y fusión de información procedente ambos tipos de sensores (estudiado en el siguiente apartado), por lo cual sus trabajos son de gran interés para la tesis que aquí comienza.

En **[Odobez 2004]** (Figura 3.13) este grupo de investigadores plantean incluir en el modelo de actuación del target a seguir componentes de velocidad con el objetivo de

robustecer el modelo de movimiento, las cuales son obtenidas gracias a un proceso de segmentación del conjunto de partículas.

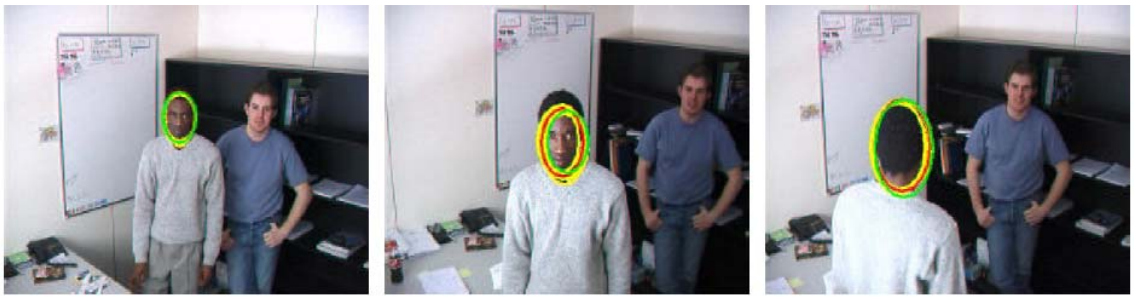


Figura 3.13 Resultados en seguimiento obtenidos por el IDIAP en [Odobez 2004]

Los experimentos presentados en la figura anterior han sido realizados con un procesador de 2.5GHz, a una frecuencia de 4fps y empleando 200 partículas. En ellos se presenta con una elipse amarilla la propuesta arrojada por las partículas más probables, en rojo la media y en verde la moda de la misma distribución. Según los autores del artículo, la comparativa de su propuesta frente algoritmo Condensation arroja importantes ventajas en su favor en cuanto a precisión, fiabilidad y robustez.

3.5.2 Seguimiento de múltiples targets

En el presente estado del arte, la tarea de seguimiento de múltiples targets se clasifica en función del número de estimadores necesarios y de la configuración del vector de estados elegida, dando lugar a la siguiente división:

- *Un estimador por cada target.*
- *Un estimador basado en la extensión del vector de estados.*
- *Un único estimador multimodal.*

3.5.2.1 Empleando un estimador por cada target

Dado el éxito obtenido por el algoritmo Condensation en sus diferentes versiones en el tracking de una única persona, parece evidente que una primera aproximación orientada al seguimiento de múltiples personas pasaría por su extensión a un número múltiple de targets.

En base a esto, investigadores de la universidad de Rennes (IRISA), de nuevo en colaboración con los creadores del Condensation, proponen en [Perez 2002] (Figura 3.14) la utilización de estimadores independientes, basados en el algoritmo Condensation, para cada uno de los targets de la escena.

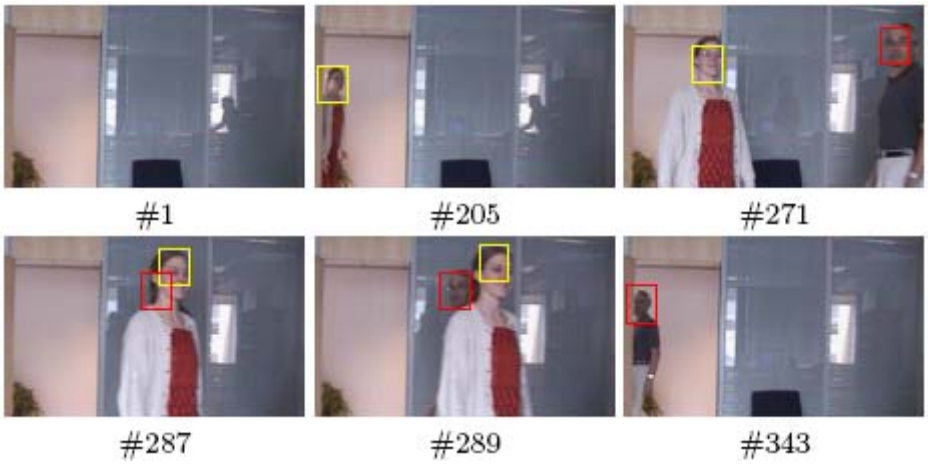


Figura 3.14 Resultados obtenidos en seguimiento de un número variable de targets empleando el algoritmo Condensation en [Perez 2002]

El modelo de observación utilizado en este trabajo consiste en añadir al modelo de contorno de la cabeza propuesto en [Isard 1998b], información procedente del histograma de color de la piel.

La información del histograma de color no sólo se utiliza en el comentado modelo de observación, sino que también se emplea para la detección de nuevos targets en la escena. Para ello se obtiene el histograma de color del fondo y se buscan cambios significativos en el mismo que puedan representar la aparición de un nuevo target.

En cuanto al algoritmo de asociación, se podría definir como un NN en un espacio de representación de histograma de color.

De lo presentado anteriormente se desprende la gran dependencia de los distintos bloques que componen el algoritmo de seguimiento con el histograma de color, lo cual lo convierte a priori en un algoritmo poco robusto ante cambios en las condiciones del mismo (e.g. iluminación) y poco flexible debido a la especificidad del modelo.

Por último destacar que no se presenta en el artículo información referente al tiempo de ejecución del algoritmo ni del número de partículas utilizadas. Sin embargo

es probable que la utilización del histograma de color dificulte la implementación del mismo en tiempo real.

El autor del trabajo anterior en colaboración con investigadores de la universidad de Cambridge propone en [Vermaak 2003] una modificación interesante del PF para abordar la tarea de MTT, consistente en un proceso de segmentación de partículas, a cada una de las cuales se las asigna un identificador que las clasifica dentro de lo que los autores denominan *"mixture components"*.

En la Figura 3.15 se presenta el funcionamiento del dicho algoritmo aplicado al seguimiento de jugadores en un partido de fútbol. En la parte superior de la figura se recoge el resultado del tracking de un PF estándar de 200 partículas y en la inferior el PF propuesto en el artículo, con un máximo de 10 "*mixture components*" y 20 partículas en cada una. En la parte del PF estándar se observa como las partículas se desplazan hasta combinarse en uno de los modos, mientras que la propuesta de los autores permite mantener la multimodalidad temporal del algoritmo.

La diferencia entre el algoritmo propuesto y otro que implementara un PF por cada target de la escena, la encontramos en el paso de corrección, en el que las partículas son ponderadas de manera conjunta en función de un algoritmo de asociación probabilístico similar al JPDA presentado en 3.4.2.2.2.

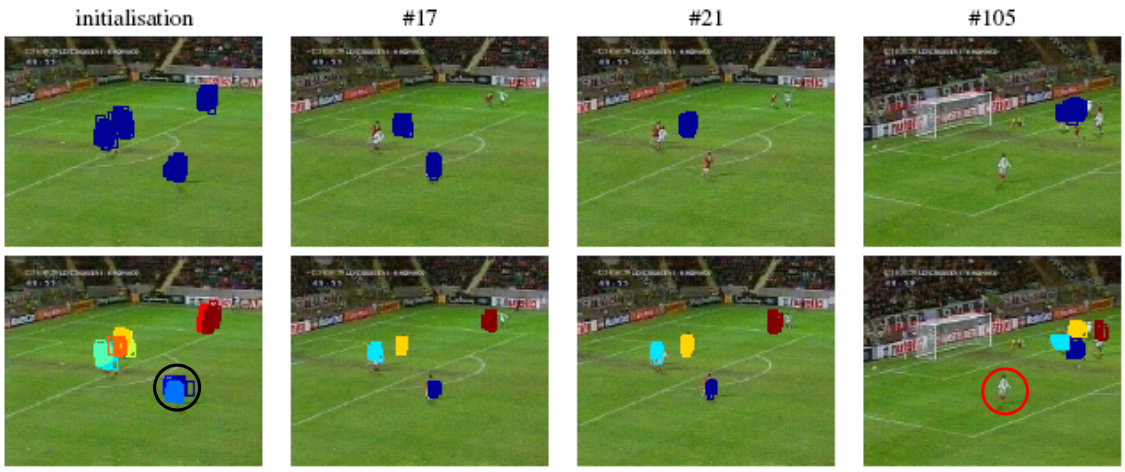


Figura 3.15 Comparativa de los resultados obtenidos en el seguimiento con un PF estándar (fila superior) y el propuesto en [Vermaak 2003] (fila inferior)

Se puede apreciar que en general cada target de la escena puede estar representado por más de un "*mixture component*" (marcado con circulo negro) y que el

número de targets para los que se define el algoritmo es constante, no permitiendo por tanto el seguimiento de nuevos jugadores (marcado con circulo rojo).

Otra desventaja del algoritmo es que su inicialización debe realizarse de manera manual, la cual se utiliza además para obtener el modelo de observación de los targets basado en histograma de color.

Un par de años más tarde este grupo de investigación propone en **[Ng 2005]**, un proceso de segmentación del set de medidas de entrada, aunque en un entorno de simulación y con datos no asociados a sensores de visión.

A la clasificación del set de medidas de entrada a los PF usados para estimar la posición de cada uno de los targets se le aplican procesos de creación y eliminación de filtros, permitiendo así existencia de un número variable de objetos.

Otro concepto interesante implementado en este trabajo es la utilización de un método determinístico basado en la persistencia en el tiempo de las clases de medida para controlar los procesos de creación y eliminación de targets anteriormente comentado, permitiendo la eliminación de los outliers.

Respecto al algoritmo de asociación de cada una de las clases con los hipotéticos targets se propone la utilización de un NN.

Estos conceptos serán posteriormente trasladados en [Marrón 2008] (analizado posteriormente) al seguimiento de objetos empleando sensores de visión, con la diferencia de que en el mismo se propone utilizar la multimodalidad inherente al filtro Bayesiano para estimar la posición de todos los targets de la escena, en lugar de utilizar estimadores independientes para cada target.

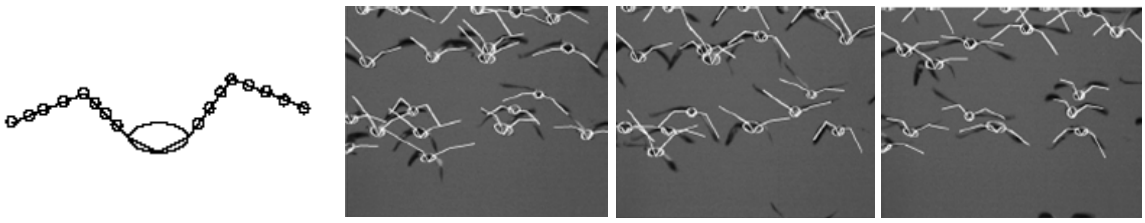
El mismo autor del trabajo presentado anteriormente propone en [Ng 2006] una idea similar a la presentada en su trabajo anterior utilizando un único estimador, pero basado en la técnica de vector de estados aumentados, con lo cual persiste el problema de carga computacional.

Para terminar este apartado se presentan dos soluciones, citadas anteriormente en el presente estado del arte, que proponen la utilización de PFs “casi” independientes para el proceso de estimación.

En la primera de ellas grupo de investigadores de la Universidad de Bristol propone en [Tweed 2002] una modificación del algoritmo Condensation propuesto en [Isard 1998a] a la que denominan *Subordinated Condensation.*

La idea más innovadora presentada en este artículo consiste en la utilización de un modelo de relación entre las partículas, el cual añade restricciones al proceso de asociación de las medidas con los targets de la imagen. El principal objetivo de esta idea es la resolución de eventuales oclusiones que puedan aparecer en la secuencia de imágenes.

El modelo de observación empleado es complejo y específico para la aplicación concreta de seguimiento de pájaros en vuelo (Figura 3.16(b)), el cual utiliza la estructura corporal de los pájaros para definirlos como una elipse central (para el cuerpo) y una serie de puntos espaciados (para las alas), tal y como se representa en la Figura 3.16(a).



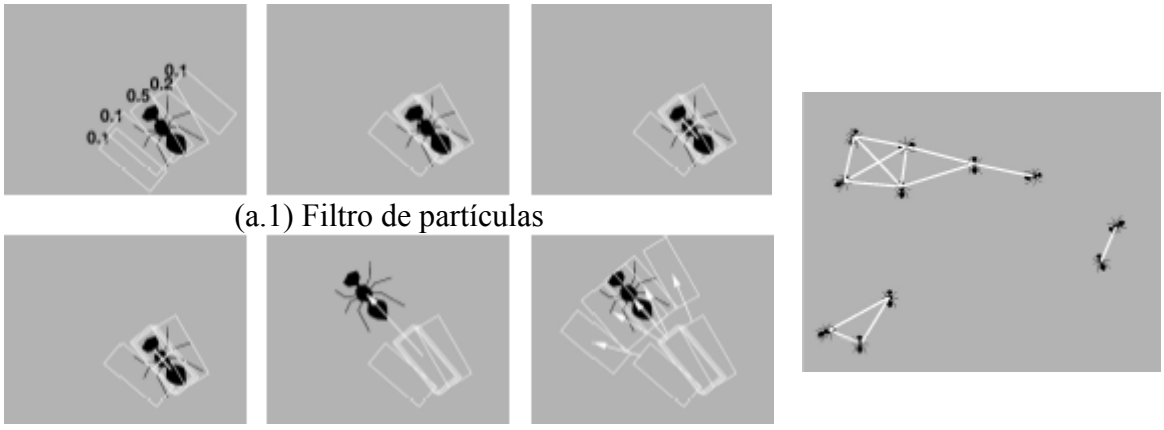
(a) (b)

Figura 3.16 Modelo de seguimiento (a) y resultado del proceso de seguimiento (b) del algoritmo Subordinated Condensation propuesto en [Tweed 2002]

El PF utilizado para la estimación se organiza en tres niveles jerárquicos con un número constante de partículas cada uno. En el caso de la Figura 3.16 se emplean 1200, 6000 y 6000 partículas respectivamente para en cada uno de los niveles del PF (unas 400 por cada pájaro), lo cual nos hace intuir (aunque el autor no lo menciona en el artículo) que la ejecución del algoritmo no se realiza en tiempo real.

La segunda es desarrollada por investigadores de GATECH (USA) en **[Khan 2003]**, donde se trata de encontrar soluciones de asociación que conlleven una menor carga computacional que el JPDA. Para ello se utilizan procesos diferentes para modelar, por un lado el movimiento de cada target, y por otro la relación de posición entre ellos.

La estimación de la posición de cada uno de los targets se realiza mediante un PF, donde se incluye un modelo de asociación de la posición de los distintos targets de la escena basado en un Markov Random Field (MRF) como parte del modelo de observación. Esta información de relación se incluye dentro del vector de estados de cada target.



(a.1) Filtro de partículas

(a.2) Modelo de movimiento

(a) (b)

Figura 3.17 Proceso de evolución de partículas (a) y MRF (b) en los experimentos de [Khan 2003]

La Figura 3.17(a.1) muestra el conjunto de partículas (identificadas con rectángulos blancos) cada una con un peso normalizado asociado función del modelo de observación. En el siguiente paso las partículas sufren un proceso de resampling de acuerdo con los pesos anteriormente referidos, para finalmente obtener la posición estimada como la media de las partículas obtenidas fruto del resampling.

La Figura 3.17(a.2) presenta en primer lugar el frame anterior de la ejecución (izq) y la evolución del mismo (central), tras lo cual se produce el movimiento de las partículas de acuerdo con el modelo estocástico de movimiento (dcha) y a continuación se les asignan los pesos normalizados, tras lo cual se vuelve a repetir el proceso.

En la Figura 3.17(b) se presenta un ejemplo del MRF antes referido, en el que se refleja la relación entre los targets. Sólo se representan las uniones entre targets con una probabilidad alta de interacción en su movimiento. De este modo se generan restricciones de asociación para incrementar o decrementar la probabilidad de las hipótesis de posición entregadas por el PF de cada target, en lugar de estimar todas las posibles hipótesis de asociación mediante un único PF. De ahí el calificativo de "casi"

independientes, ya que en este caso la posición de cada objeto se usa en la ponderación de las partículas asociadas a los targets que se relacionan con el a través del MRF.

La Figura 3.18 muestra un ejemplo de ejecución del algoritmo comentado, en el que se realiza un seguimiento de dos hormigas utilizando 400 partículas (su número depende del número de targets), con un tiempo de ejecución de 8fps en un procesador de 2.5GHz. El frame central de la figura presenta un alto grado de oclusión entre los dos targets, solucionado positivamente gracias al MRF.

Figura 3.18 Resultados del algoritmo propuesto por [Khan 2003]

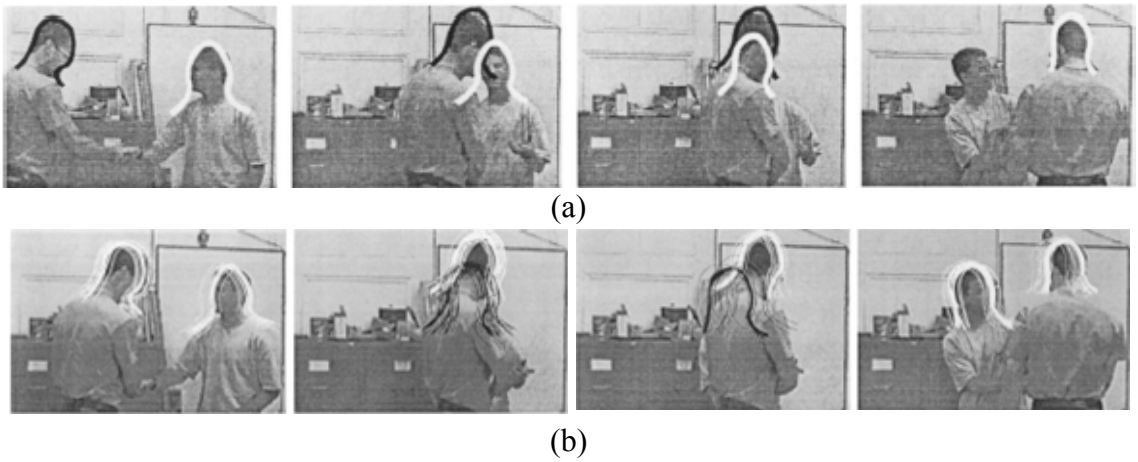
El artículo demuestra que la solución planteada permite reducir el tiempo de ejecución 150 veces respecto al JPDAF, permitiendo así un seguimiento de hasta 20 targets simultáneos. Sin embargo, que el número de partículas sea función del número de targets se convierte en un factor negativo, ya que también su tiempo de ejecución será variable y no constante (como es deseable).

Trabajos posteriores de este mismo grupo se presentan en [Khan 2004] en la misma línea de investigación de seguimiento de hormigas, en el cual se trabaja en el refinamiento de las técnicas expuestas anteriormente.

3.5.2.2 Empleando un único estimador basado en la extensión del vector de estados

El grupo de investigación de [Isard 1998a] introducen en **[MacCormick 1999]** el concepto de *principio de exclusión*, el cual modifica la PDF de verosimilitud del algoritmo Condensation con el objetivo de evitar los problemas derivados de las oclusiones totales o parciales de unos targets sobre otros, debido a las cuales las medidas procedentes de un objeto total y parcialmente incrementan la probabilidad a posteriori de otro diferente.

El autor propone una expansión del vector de estados χ_t (cuyo principal inconveniente en general es la difícil gestión computacional de un vector de estados de dimensiones dinámicas) de manera que se incluyan en el las componentes de todos los targets de la escena así como un identificador que indica si cada objeto oculta a otro. Este modelo de relación entre targets se usa, además de en la modificación de la PDF de verosimilitud del PF, en el proceso de asociación en caso de oclusión.



(a)

(b)

Figura 3.19 Resultados del algoritmo Condensation (a) sin aplicar el principio de exclusión y (b) aplicando dicho principio

En la figura anterior se aprecia como la aplicación del principio de exclusión permite al estimador conservar los dos targets (imagen de la derecha) a pesar de la oclusión total mostrada en el frame anterior.

El algoritmo así planteado utiliza una probabilidad a posteriori multimodal, cuya evolución se simplifica gracias al principio de exclusión. A pesar de ello, continúa sin ser implementable en tiempo real a pesar de la disminución del número de partículas necesarias, debido principalmente a la complejidad del modelo de observación (B-spline) y a la utilización de un vector de estados aumentado.

El autor resuelve este problema de carga computacional, al menos en parte, mediante el partitioned sampling ([MacCormick 2000]) abordado en el apartado anterior, a pesar de lo cual no se resuelve el problema de seguimiento de forma global al no permitirse un número variable de targets.

Otros trabajos como **[Tao 1999]**, proponen modificaciones al algoritmo básico de tal manera que se incorpore a la creencia información relativa a la relación entre los

targets de la escena (de manera similar a los trabajos ya tratados de [Tweed 2002] y [Khan 2003]), con el objetivo de simplificar el número de hipótesis de estimación válidas. Esta información también se utiliza en el proceso de asociación.

Este trabajo permite el seguimiento de un número variable de targets, para lo cual introduce dos factores que modelan la probabilidad de incorporación y eliminación de los mismos.

Se propone de nuevo un modelo de observación complejo y específico a la tarea de seguimiento concreta, basado en la forma del objeto, e implementado a través de un proceso jerárquico.

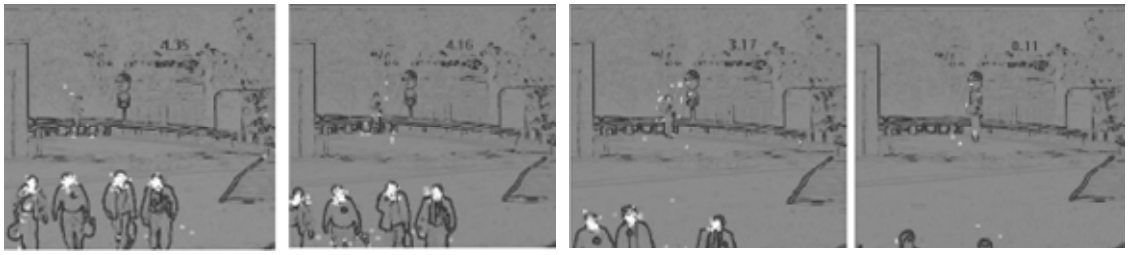


Figura 3.20 Resultados obtenidos en [Tao 1999]

El ejemplo presentado en la figura anterior utiliza 300 partículas, lo cual le permite ejecutarse en un procesador de 400 MHz a una velocidad de 1fps (no en tiempo real).

Otra de las ideas más interesantes presentadas en el área de MTT es de nuevo proporcionada por los creadores del Condensation en **[Isard 2001].** En el se propone como idea más interesante el algoritmo BraMBLe, en el que destaca un novedoso modelo de observación basado en blobs en el espacio 3D, los cuales se obtienen mediante un entrenamiento off-line del sistema de detección.

La PDF obtenida mediante el proceso de entrenamiento off-line antes referido se utiliza en el proceso online para la detección de nuevos targets en la escena, los cuales son introducidos dentro del vector de estados aumentado propuesto en este trabajo.

El algoritmo de asociación propuesto es un JPDA, el cual se ve favorecido en cuanto a reducción de carga computacional, del trabajo con blobs aquí propuesto.

Al igual que en otros trabajos anteriores, y con el objetivo de aumentar la robustez del algoritmo ante oclusiones totales y parciales, se introduce en el vector de estados información relativa a la interacción entre targets.

En la Figura 3.21 se presentan los resultados obtenidos en este trabajo, en el que la escena contiene un número variable de targets (2 o 3). El algoritmo se ejecuta en un procesador de 477MHz a una velocidad de 15fps, empleando para ello 1000 partículas.

Figura 3.21 Resultados del algoritmo BraMBle en [Isard 2001]

La posible problemática asociada a este algoritmo reside en que no se le pone a prueba con un número mayor de targets y en una escena más complicada, lo cual podría aumentar considerablemente el número de partículas necesarias y no permitir su ejecución en tiempo real.

El algoritmo BraMBle aquí presentado es utilizado en posteriores trabajos de la comunidad científica, como por ejemplo en **[Branson 2005]**, empleado en este caso en el seguimiento de ratones.

Por otro lado, los mismos autores de [Perez 2002] proponen en **[Hue 2001]** la introducción en el vector de estados de una variable que caracteriza la probabilidad que tiene cada nueva medida de ser asociada a un target concreto de la escena. El valor de esta variable se obtiene a partir de un PMHT (versión probabilística del MHT tratado en 3.4.2.1.1). Los resultados obtenidos se muestran en la Figura 3.22.

Figura 3.22 Ejemplo de resultados obtenidos en [Hue 2001]

El modelo de observación consiste en una representación de la silueta de los objetos basado en descriptores de Fourier, el cual se aprende en un proceso off-line.

El algoritmo aquí presentado es válido únicamente para un número constante de targets en la escena, lo cual trata de ser resuelto por la autora en [Hue 2002], no mostrándose en este caso resultados prácticos de la propuesta.

Respecto a su tiempo de ejecución, la autora no presenta datos concretos, pero sí especifica que para la Figura 3.22 se utilizan 100 partículas y 9 coeficientes de Fourier.

Para concluir este apartado, el grupo de investigación del IDIAP autores de [Odobez 2004] plantean en **[Smith 2005]** una nueva modificación del PF básico, basado en un modelo de observación organizado en dos niveles jerárquicos (uno superior a nivel de blob de color y otro inferior a nivel de píxel).

Los autores también plantean la utilización de una PDF para modelar la relación de posición de los distintos targets de la escena, idea ya propuesta en [Khan 2003] y mejorada aquí.

El algoritmo permite trabajar con un número variable de targets, si bien su numero no es elevado (un máximo de 4 en los ensayos presentados) y la escena no es excesivamente compleja. El artículo no informa sobre la velocidad de ejecución del algoritmo, si bien se supone una elevada carga computacional debido al modelado de los targets.

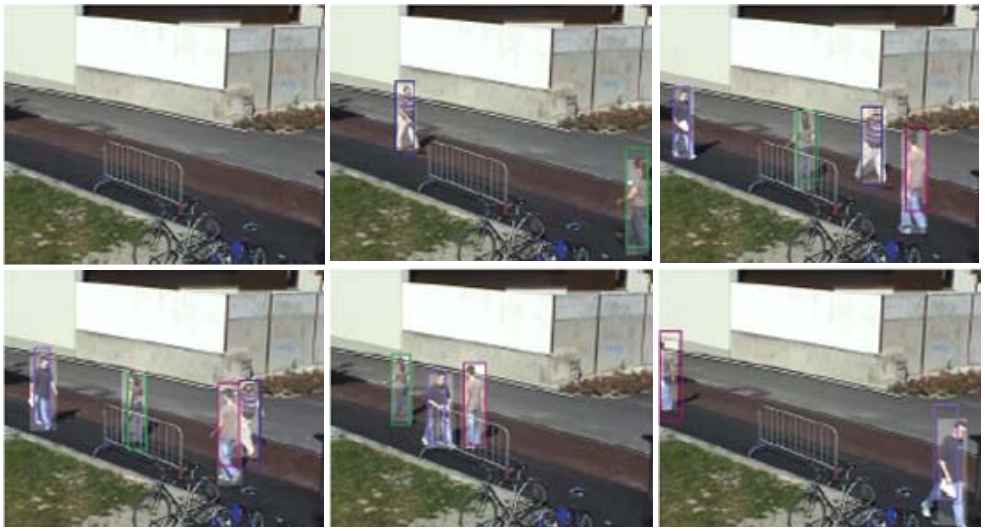


Figura 3.23 Resultados obtenidos en los experimentos de [Smith 2005]

La Figura 3.23 presenta los resultados obtenidos con este algoritmo, en el que la estimación de los diferentes targets se marca con rectángulos de diferentes colores.

Por otro lado, un grupo de investigadores del Istituto Trentino di Cultura (ITC) italiana, en colaboración con investigadores de la Universidad de California, proponen en [Lanz 2005] el algoritmo *Hybrid Joint-Separable (HJS)*, que será posteriormente mejorado en **[Lanz 2006]** para aplicaciones de MTT de personas. Los resultados obtenidos en este trabajo son muy interesantes, por lo que algunos detalles de la implementación se presentan a continuación.

La primera contribución importante realizada en este trabajo se basa en la representación eficiente de las PDF a priori $p(\vec{x}_t \mid \vec{z}_{1:t-1})$ y a posteriori $p(\vec{x}_t \mid \vec{z}_{1:t})$ a través del producto de sus *marginales* mediante la siguiente igualdad:

$$p(\vec{x}_t \mid \vec{z}_{1:\tau}) \cong \prod_k p(x_t^k \mid z_{1:\tau}) \tag{3.21}$$

donde $p(x_t^k \mid z_{1:t})$ se define como:

$$p(x_t^k \mid z_{1:\tau}) = \int p(\vec{x}_t \mid z_{1:\tau}) d\vec{x}_t^{\neg k} \tag{3.22}$$

con $\vec{x}_t^{\neg k}$ representando el vector $\vec{x}_t$ con la $k-esima$ componente eliminada y τ tomando valores $t-1$ y t para las distribuciones a priori y a posteriori respectivamente.

En base a las representaciones anteriormente mostradas se propone la utilización de una representación eficiente de las distribuciones a priori y a posteriori, mientras que las etapas de predicción y corrección se realizan en base a modelos de actuación y observación conjuntos, tal y como se presenta en la siguiente figura.

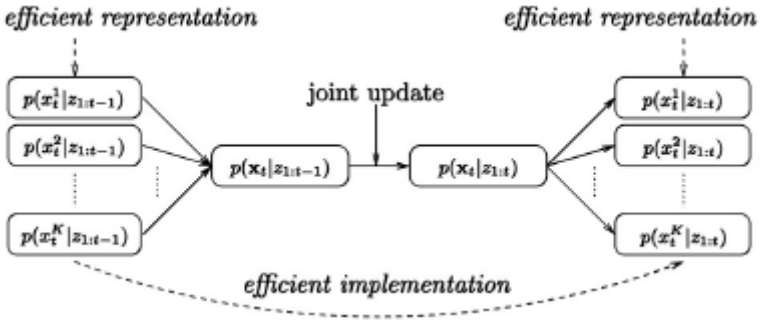


Figura 3.24 Diagrama de bloques del algoritmo recursivo implementado en [Lanz 2006]

La segunda contribución consiste en la presentación de un robusto sistema de oclusiones, basado en un modelo de apariencia obtenido de acuerdo a los principios de formación de la imagen e implementado a nivel de píxel.

El modelo de observación empleado por este sistema se basa en la utilización de tres imágenes del target tomadas desde diferentes ángulos (frontal, lateral y trasero), sobre cada una de las cuales se realiza una extracción de fondo a objeto de obtener un modelo 3D de apariencia de los targets en el que se identifican los histogramas de color de cabeza, torso y piernas, empleándose además un modelo cilíndrico de forma para cada target. Este proceso se realiza previamente al comienzo del proceso de tracking y obliga a que los targets entren a la escena siempre por una misma zona y se queden parados en la misma durante un tiempo hasta que el algoritmo extrae estas características (comportamiento que se puede observar en diferentes videos extraídos de [FBK Lanz]), lo cual constituye una limitación del sistema.

Una vez obtenidos estos modelos de observación, se comparan diferentes coeficientes (en base a la distancia de Bhattacharyya) de los mismos y los histogramas candidatos procedentes de la escena obtenidos gracias a un proceso de renderización.

La Figura 3.25 representa los procesos de renderización (a) y de obtención de modelos de targets (b) anteriormente comentados.

(a) (b)

Figura 3.25 Procesos de renderización (a) y de obtención de modelos (b) en [Lanz 2006]

Adicionalmente, la dinámica conjunta del sistema se describe por medio de un *forward Markov Random Field (MRF)*

El algoritmo propuesto permite obtener un espacio de representación que crece linealmente con el número de targets, y una complejidad computacional que crece de manera cuadrática ($O(K^2 N)$).

En la Figura 3.26 presenta resultados reales obtenidos con este algoritmo, obtenidos con una cámara calibrada con una velocidad de captura de 15 fps y una

resolución de 200x150 tras un proceso de submuestreo. Se emplea un vector de estados de 4 dimensiones formado por la posición y la velocidad en el plano definido por el suelo de la habitación. En las tres esquinas restantes de la habitación se ubican sendas cámaras para la obtención de diferentes vistas, las cuales no son utilizadas a efectos de tracking según el autor.

El algoritmo se ejecuta en un procesador de 3GHz utilizando unas 100 partículas por cada target, demuestra un seguimiento fiable, preciso y robusto a las oclusiones, con velocidad de ejecución dependiente del número de targets que varía entre 15 fps para 3 targets y entre 3 y 10 fps (dependiendo de las oclusiones) para 5 targets. Esta dependencia de la velocidad de ejecución del algoritmo en función del número de targets constituye el otro inconveniente de este algoritmo.

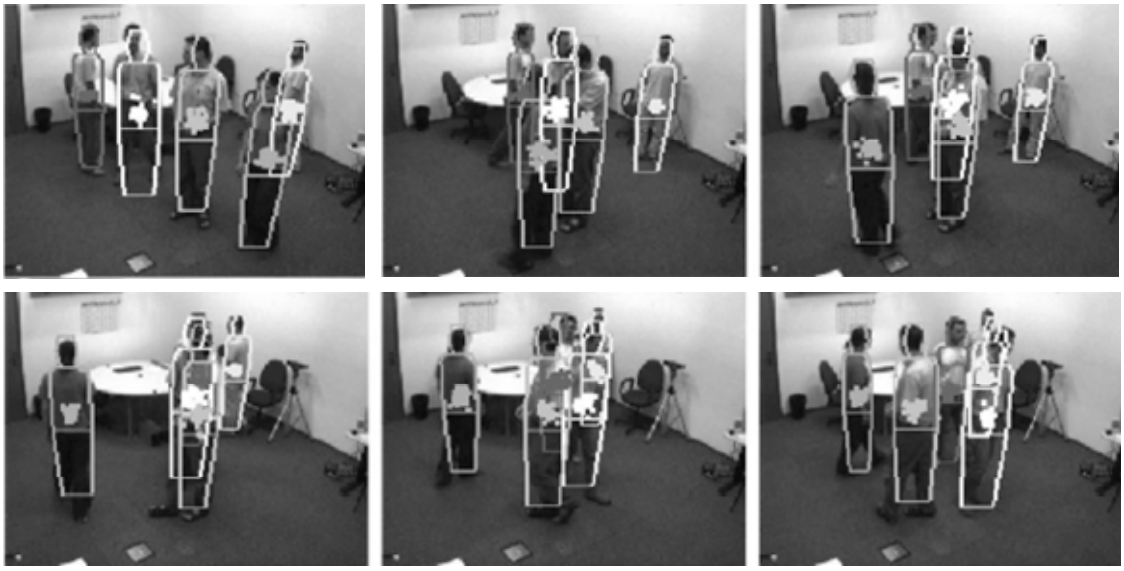


Figura 3.26 Resultados obtenidos en los experimentos de [Lanz 2006]

3.5.2.3 Empleando un único estimador multimodal

La primera referencia de la literatura científica en la que se presenta la idea de emplear la multimodalidad de la creencia para modelar la estimación de posición de todos los targets de la escena se encuentra en **[Koller-Meier 2001]**, realizado por investigadores del ETH de Zurich en colaboración con el GATECH. El PF así modificado recibe el nombre de *Filtro de Partículas Extendido* (XPF).

En este trabajo sólo se presentan resultados de simulación debido principalmente a los problemas de robustez del estimador, que se traduce en una pérdida habitual de hipótesis de estimación. Para resolver esto, la comunidad científica propone en [Okuma 2004] y [Marrón 2008] la introducción en el algoritmo de un clasificador de medidas, tal y como podremos comprobar posteriormente en este mismo apartado.

En el primero de estos trabajos, **[Okuma 2004]**, llevado a cabo por investigadores de la Universidad British Columbia, se propone un clasificador de medidas al que denominan *Adaboost* basado en el aprendizaje off-line de modelos de color (análogo a alguno de los propuestos en el apartado anterior). Este detector se utiliza además como base del algoritmo de asociación y en la detección de nuevos objetos en la escena, por lo que el algoritmo resultante es válido para un número variable de targets.

El algoritmo resultante es denominado por autores como *Boosted Particle Filter* (BPF). En la Figura 3.27 se presentan los resultados obtenidos por dicho algoritmo aplicado al seguimiento de jugadores de jockey sobre hielo en una secuencia de imágenes. Los rectángulos blancos representan la estimación obtenida como salida del algoritmo *Adaboost*, mientras que los rectángulos negros representan las partículas utilizadas por el BPF.

La secuencia de imágenes permite ver como el algoritmo es capaz de adaptarse a la aparición de nuevos jugadores (dos primeras imágenes para el jugador de arriba a la izquierda) y a la desaparición de los mismos (tercera y cuarta imagen en la parte izquierda de la escena).

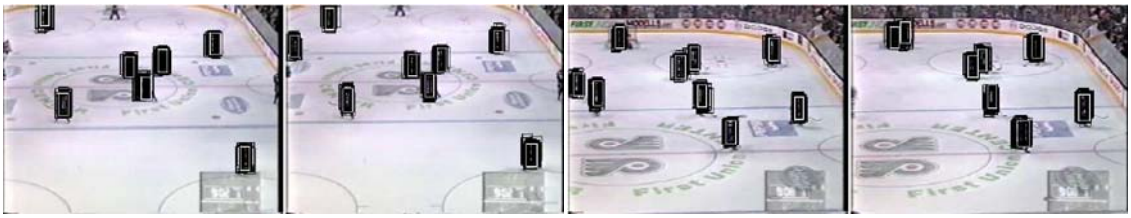


Figura 3.27 Resultado en el seguimiento de jugadores de jockey empleando el BPF propuesto en [Okuma 2004]

Dentro del grupo de algoritmos que emplean un único estimador multimodal para modelar la posición de todos los targets de la escena, destaca la propuesta realizada en [Marrón 2008], donde se propone un algoritmo que la autora denomina XPFCP (*Filtro de Partículas Extendido con Proceso de Clasificación).*

Este algoritmo propone la combinación de algoritmos probabilísticos de estimación, que emplea como base un XPF (*eXtended Particle Filter*), y determinísticos de asociación, implementado a partir de un clasificador de medidas, con el objetivo de robustecer el comportamiento del algoritmo global.

La salida determinística final del algoritmo se obtiene a partir de un nuevo proceso de clasificación, en este caso de las partículas procedentes del PF, el cual segmenta la PDF de salida en función de sus picos de máxima probabilidad, de manera que se obtiene un vector de estado $\vec{x}_t$ de un número variable de targets.

Otro factor a destacar de este trabajo es la utilización de modelos sencillos y generales (en este caso un modelo CV -3.3.1-), lo cual facilita su implementación para el seguimiento de diferentes targets a la par que reduce la carga computacional del algoritmo, facilitando su implementación en tiempo real.

En la Figura 3.28 se presentan resultados reales obtenidos con el XPFCP, en los cuales se ha empleado un set de 600 partículas con una veloidad de ejecución de 15fps, pudiendo observarse la correcta adaptación a un número variable de targets, mientras que la Figura 3.29 presenta un diagrama de bloques general del algoritmo.

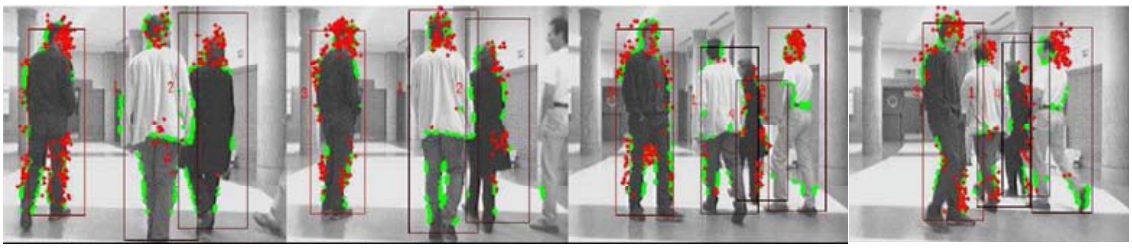


Figura 3.28 Resultados en el seguimiento de personas empleando el algoritmo XPFCP propuesto en [Marrón 2008]

En la Figura 3.29 se identifican en color verde los pasos básicos del PF, mientras que en amarillo se representan los pasos adicionales propuestos por la autora.

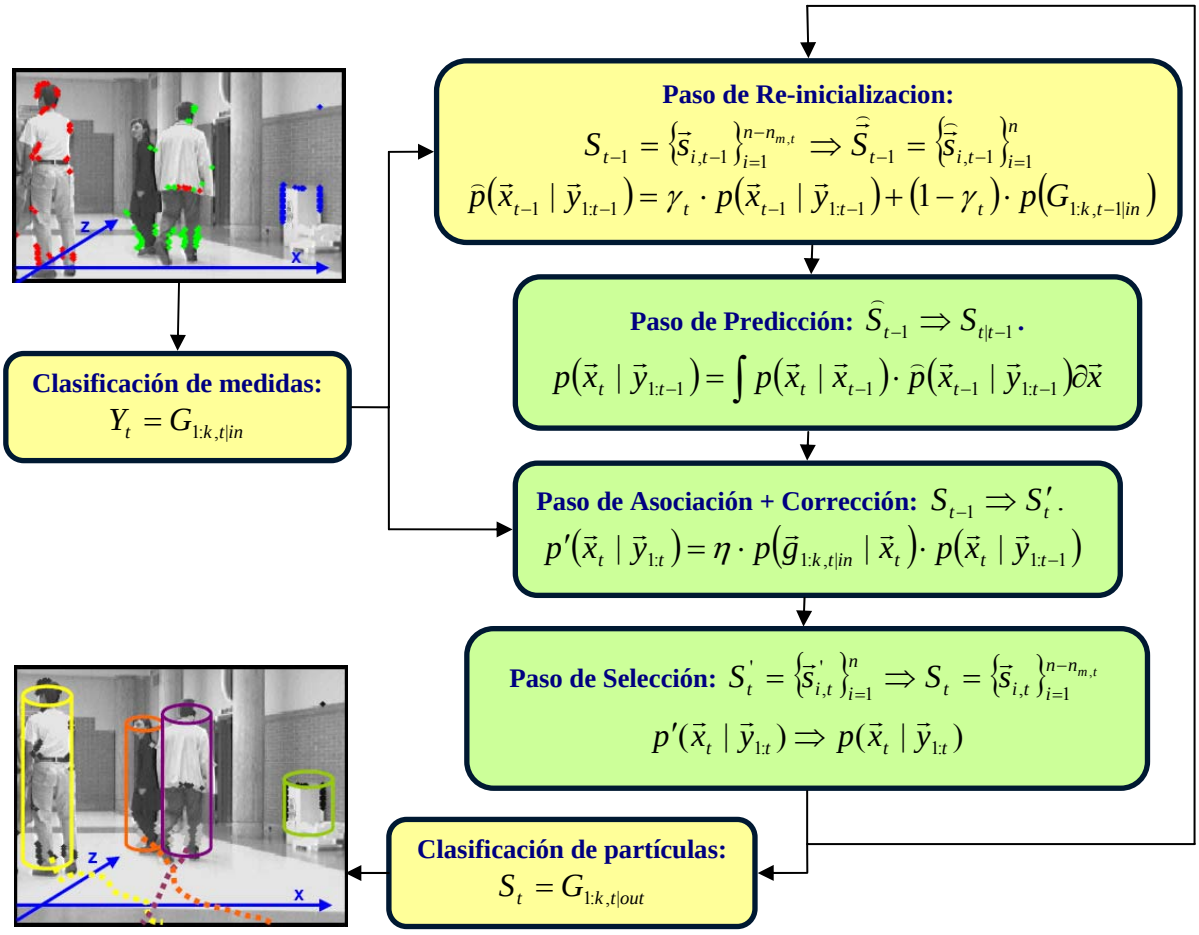


Figura 3.29 Diagrama general de bloques del XPFCP

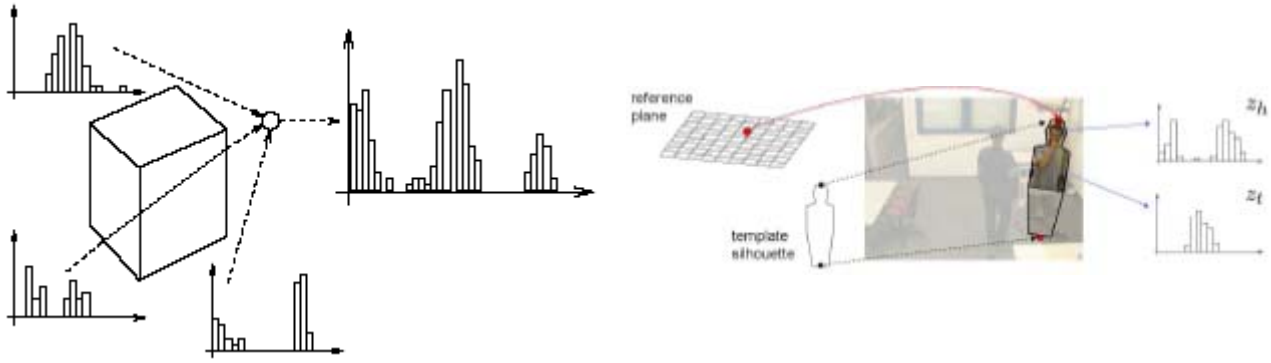
3.6 Pose

La cantidad de trabajos orientados a la obtención de la pose de una persona a partir de información visual son sensiblemente superiores que para el caso se audio (2.4.4). En este punto, destacan los trabajos realizados por dos grupos de investigación, los cuales se describen brevemente a continuación:

El primero de ellos es el desarrollado por el ITC (Istituto Trentino di Cultura) en **[Lanz 2006b]**. Este trabajo, aplica técnicas análogas a las presentadas en [Lanz 2006] como la utilización del PF con un vector de estados de pequeñas dimensiones (en este caso 5 incluyendo posiciones y velocidades en los ejes x,y y la orientación de la cabeza), un modelo de observación que emplea el color y forma del target y que a través de un proceso de renderización previa, permite obtener la verosimilitud a través de la

comparación de la imagen real con el modelo renderizado o la utilización de modelos de representación eficiente basadas en la densidades de probabilidad marginales introducidas en el algoritmo HJS (Hybrid Joint Separable) en [Lanz 2005].

El algoritmo consta de dos etapas bien diferenciadas. En la primera (Figura 3.30 (a)) de ellas tiene como objetivo la obtención del modelo de renderización, para lo cual se asume que se disponen de múltiples vistas del target, adquiridas previamente, y de cada una de las cuales se obtiene el histograma de color. A cada vista se le asigna un peso proporcional a la cantidad de área visible (la cual es proporcional al coseno de desplazamiento angular de las nuevas vistas). Los modelos de apariencia de las nuevas vistas se calculan mediante una interpolación ponderada de los histogramas de los que se dispone. En la segunda etapa (Figura 3.30 (b)) se utiliza un modelo de forma para identificar los targets de la imagen, de los cuales se extrae el torso y la cabeza para posteriormente calcular su histograma de color. La comparación de estos histogramas con el modelo renderizado obtenido en la primera etapa proporciona la estimación de la pose buscada.



(a) (b)

Figura 3.30 Proceso previo de renderización a partir de las diferentes vistas del target basada en histograma de color (a) y proceso de extracción de histogramas candidatos (b) en [Lanz 2006b]

La Figura 3.31 muestra resultados reales obtenidos en [Lanz 2006b] en el que el algoritmo se ejecuta a una velocidad de 10 fps con una resolución de 266x200 píxeles, el filtro de partículas utiliza 500 partículas para su funcionamiento y el algoritmo se ejecuta en un PC Dual-Core de 3GHz con unas cargas computacionales de 40%, 60% y 90% para 1,2 y 4 cámaras respectivamente. En el caso de 4 cámaras, el error obtenido nunca sobrepasa los 10º.

Figura 3.31 Resultados en la estimación de pose obtenidos en [Lanz 2006b]

En segundo lugar es imprescindible mencionar los trabajos realizados en el instituto suizo IDIAP relativos a la estimación de pose, de entre los cuales se puede destacar: [Ba 2004], [Ba 2005], [Odobez 2007], [Ba 2008] y [Smith 2008].

En **[Ba 2005]**, se propone la utilización de un Rao-Blackwellized PF para la obtención de la pose de los targets de la escena, a partir de las cuales se puede obtener los VFOA (Visual Focus of Attention). Para ello se define un espacio de estados de 3 dimensiones, con un modelo de observación basado en textura y color y un modelo de actuación autorregresivo de primer orden. El PF así definido y configurado para la utilización de 100 partículas, genera los resultados presentados en la siguiente figura:

Figura 3.32 Resultados obtenidos en la estimación de la pose en [Ba 2005]

En el artículo se especifica que los resultados presentados han sido obtenidos a partir de las implementaciones en Matlab de los algoritmos definidos y que por lo tanto no se puede hablar sobre ejecución en tiempo real. Sin embargo el autor argumenta que el bajo número de partículas y la existencia de implementaciones eficientes del Rao_Blackwellized PF, les hacen ser optimistas a este respecto. Los errores en la estimación de la pose son de un 7% en *pan* y *roll* y un 14% en *tilt*.

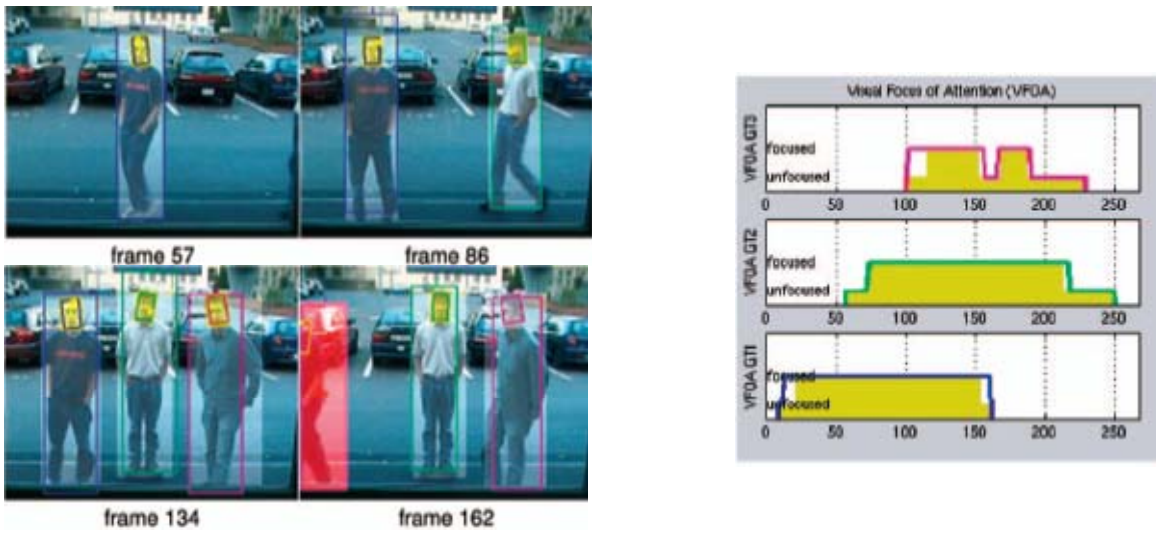
En **[Ba 2008]**, el autor parte de los resultados presentados anteriormente y redirige sus esfuerzos hacia la obtención del VFOA del conjunto de target de la escena, dentro del contexto de los "meetings". Para ello se propone el empleo de un

Input_Output Hidden Markov Model (IOHMM), en el que los estados ocultos son los VFOA del conjunto de participantes en el meeting y las observaciones son las poses individuales de cada uno, extraídas a partir del algoritmo de su trabajo anterior.

El desarrollo del algoritmo se asienta sobre dos propiedades fundamentales: La primera consiste en que un grupo de personas tiende de manera natural a compartir un VFOA común, lo cual proporciona pistas a un observador externo sobre cual será la pose de un participante en un meeting si el resto están mirando en la misma dirección. La segunda se basa en la información contextual (e.g. proyectores, participante que está hablando), los cuales suelen atraer la atención de los participantes en un meeting.

Por último, en el trabajo presentado en **[Smith 2008]** se trata de obtener el VFOA de un número variable de personas, en este caso en un entorno exterior y en el que el objetivo consiste en determinar si los usuarios que cruzan la escena fijan o no su atención en determinados objetos presentes en la misma.

El algoritmo emplea una Dynamic Bayesian Network que obtiene de manera simultánea el número de personas, su posición y pose. Para inferir de manera eficiente cada uno de los parámetros del vector de espacios extendido se propone la utilización de un Reversible-Jump Markov Chain Monte Carlo (RJMCMC) junto con un novedoso modelo de observación que proporciona el número de personas en la escena y su localización. Los resultados obtenidos se presentan en la siguiente figura.



(a) (b)

Figura 3.33 Resultados obtenidos en [Smith 2008]: (a) Imágenes de la escena y (b) Representación gráfica de los resultados

4. Fusión audiovisual

4.1 Introducción

La fusión multisensorial se ha revelado en los últimos años como un campo de investigación en auge, debido principalmente a que cada tipo de sensor empleado en la misma es capaz de compensar la debilidad del resto y viceversa, en un proceso que tiene como objetivo el aumento de la robustez y fiabilidad del algoritmo final.

La comunidad científica ha destacado en los últimos años la importancia de la fusión multisensorial con los objetivos anteriormente comentados, a pesar de lo cual, las soluciones de tracking para múltiples hablantes empleando sensores de audio abordadas desde el procesamiento de señal (e.g. [Potamitis 2004], [Vo 2004]), y las presentadas para múltiples targets del entorno basadas en visión computacional (e.g. [Perez 2002], [Yu 2004], [Zhao 2004]) de manera independiente, superan ampliamente en número a las propuestas de fusión audiovisual. Debido a ello, las aportaciones que se pueden realizar en este ámbito pueden considerarse más importantes si cabe que las propuestas para sensores independientes.

En el caso concreto de la tesis que aquí comienza se propone la utilización conjunta de sensores de audio y visión. En apartados anteriores ya se estudiaron en detalle las fortalezas y debilidades de cada uno de ellos, destacando las oclusiones para el caso de las cámaras y el ruido y las reverberaciones para el caso de los arrays de micrófonos, las cuales pueden ser minimizadas o en el mejor de los casos solucionadas gracias a la fusión multisensorial.

Las principales aplicaciones de tracking de targets del entorno empleando técnicas de fusión audiovisual se desarrollan principalmente en los campos de las teleconferencias, la vigilancia, los “espacios inteligentes” y los interfaces hombre máquina, de forma similar a lo presentado en los apartados dedicados a audio y visión de manera independiente.

4.2 Clasificación de las técnicas de fusión audiovisual

De manera general, una primera clasificación de los trabajos presentados por la comunidad científica en el ámbito de la fusión audiovisual surge debido a las diferentes metas perseguidas por cada uno (tracking de un único o múltiples targets), la configuración de sensores audiovisuales utilizada y el entorno específico en el que se desarrolla.

Sin embargo, desde el punto de vista del presente trabajo, la clasificación más importante de algoritmos de fusión es la que se presenta a continuación, representada de forma gráfica en la Figura 4.1:

- Orientados a Sistema *(System Oriented)*: Este tipo de algoritmos se centran en la integración de módulos, o lo que es lo mismo, tratan de aprovechar el gran desarrollo de algoritmos de tracking que emplean sensores de audio o de visión independientes entre si, para a partir de la estimación arrojada por cada uno de ellos encontrar la manera óptima de combinar dicha información en forma de estimación conjunta.
- Orientados a Modelo *(Model Oriented):* Por el contrario, este tipo de algoritmos tratan de obtener una formulación matemática conjunta que aproveche al máximo las fortalezas de cada tipo de sensor en las diferentes etapas del algoritmo, de tal manera que se genere directamente una estimación óptima conjunta.

En la Figura 4.1 se puede observar la captación de información de cada tipo de sensor procedente de la escena y la entrega de la misma ($y_a(t)$ en el caso de audio e $y_v(t)$ en el caso de la visión) bien a algoritmos independientes de tracking de audio y visión en el caso orientado a sistema, o bien a algoritmos de tracking que emplean la información conjunta de ambos tipos de sensores para el orientado a modelo.

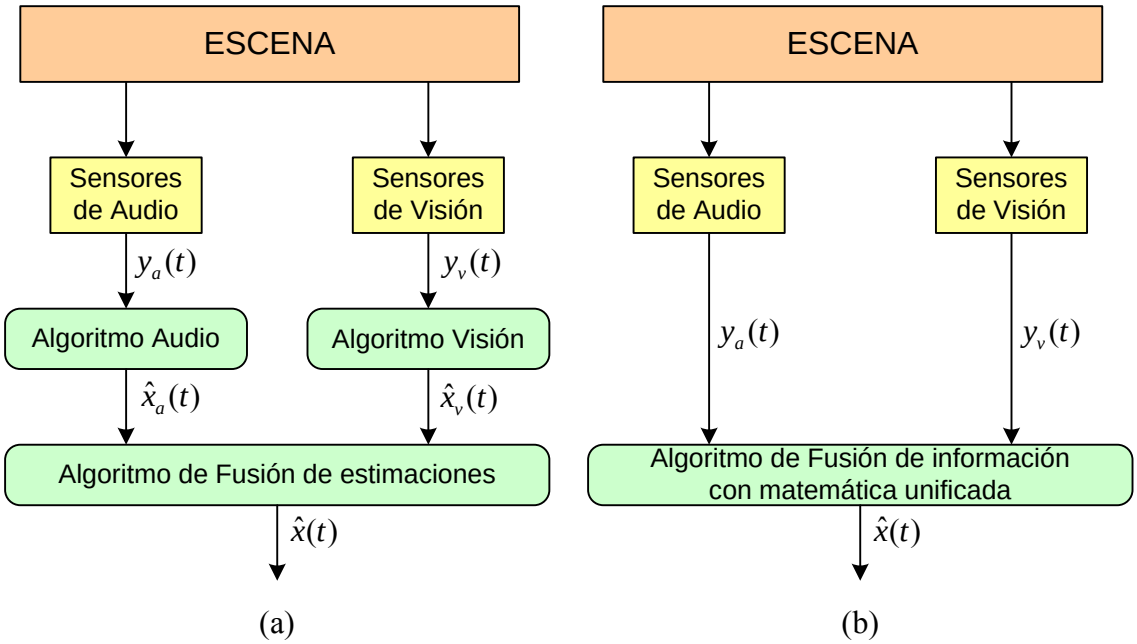


Figura 4.1 Clasificación de algoritmos de fusión audiovisual: (a) Orientados a sistema, (b) Orientados a Modelo

En el caso del system oriented el algoritmo de audio generará una estimación $\hat{x}_a(t)$ y el de visión otra $\hat{x}_v(t)$, la cual deberá ser fusionada por un algoritmo de más alto nivel en una salida de estimación única $\hat{x}(t)$, la cual es generada directamente a partir de la información de los sistemas sensores en el caso model oriented.

4.3 Líneas de investigación en el área de fusión audiovisual

En este apartado se presentan las líneas de investigación más importantes relacionadas con la fusión de información de sensores de audio y visión con propósitos de tracking de targets en una escena.

Una primera división de las mismas se realiza en función del número de targets de la escena a seguir (un único targets o múltiples targets), mientras que éstas últimas se dividen a su vez en orientadas a sistema u orientadas a modelo conforme a las definiciones presentadas en el apartado anterior.

4.3.1 Seguimiento de un único target

4.3.1.1 Seguimiento de un único target en escenas con un sólo target

Los primeros trabajos de tracking empleando fusión audiovisual partían de la configuración de escena más sencilla posible, compuesta únicamente por un sólo target.

En esta línea, investigadores de la Universidad de Stanford, en colaboración con la universidad de Toronto, proponen en **[Aarabi 2001]** un sistema de tracking basado en 2 cámaras y un array lineal de micrófonos de 3 elementos.

El sistema de visión emplea una segmentación de fondo a partir de una imagen de fondo conocida, mientras que el sistema de audio se basa en la GCC para obtener una estimación de los TDOA, ambos con el objetivo de obtener una SLF *"Spatial Likelihood Function" (SLF)*, la cual representa la probabilidad de encontrar un target en las distintas posiciones de la habitación.

Por lo tanto, este trabajo sigue una filosofía orientada a sistema, en el que los sistemas de audio y visión trabajan de manera independiente con el objetivo de generar dos SLF, las cuales serán fusionadas en una para obtener la salida global del algoritmo de estimación.

Los resultados obtenidos muestran una importante mejora de precisión y robustez del sistema global respecto a los sistemas de audio y visión independientes, los cuales son obtenidos en entornos de baja relación señal a ruido (0.5 dB).

Otro trabajo muy citado por la comunidad científica es **[Cutler 2000]**. Desarrollado por investigadores de la universidad de Maryland, este trabajo emplea únicamente una cámara y un sensor de audio (ambos de bajo coste), de manera que cuando el target emite sonidos el sistema es capaz de identificar el momento y la posición desde la que lo hizo.

El algoritmo implementado se basa en la alta correlación entre el movimiento de la boca y el correspondiente sonido generado cuando una persona habla, el cual se utiliza para el entrenamiento de una red neuronal.

La Figura 4.2 ejemplifica los resultados obtenidos en este trabajo, en el cual se emplean una cámara de 30fps de resolución y un micrófono de sobremesa cuya señal es muestreada a 20KHz. Según el artículo, el algoritmo es capaz de ejecutarse en tiempo real en un PC estándar.

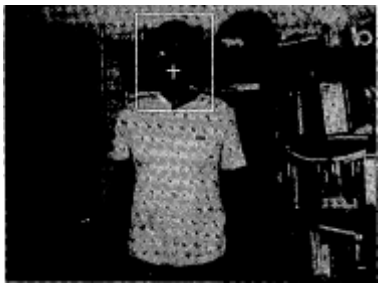


Figura 4.2 Resultados obtenidos en los experimentos de [Cutler 2000]

4.3.1.2 Seguimiento de un único target en escenas con múltiples targets

Los trabajos estudiados hasta el momento consideraban una escena compuesta únicamente por un target. Una variante de la misma consiste en una escena compuesta por múltiples targets, de los cuales nada más se desea seguir al que está hablando en cada momento. Un ejemplo de este tipo de aplicaciones es la desarrollada por investigadores de la universidad de Cambridge (colaboradores en el desarrollo el algoritmo Condensation tratado en la parte de visión -3.5.1-) en **[Vermaak 2001]**.

En este trabajo se propone la utilización de una configuración *STAC*[15] como sistema sensor (Figura 4.3(a)), siendo consideradas ambas fuentes de información como complementarias desde el punto de vista de que el audio es interesante para la inicialización (donde la visión es computacionalmente costosa) mientras que la visión es útil desde el punto de vista de localización (donde el audio es menos preciso).

En base a esto se implementan dos modelos de observación. El de audio, compuesto por todos los TDOA obtenidos a través de GCC, y una función de verosimilitud que será capaz de rechazar los tiempos erróneos procedentes de las reverberaciones en el PF que se implementará en la parte de fusión. Y el de visión,

[15] STAC (Stereo Audio and Cycloptic vision sensor): Sistema de sensors compuesto por un par de micrófonos ubicados a ambos lados y de manera simétrica a una cámara.

consistente en una curva que define el contorno de la cabeza obtenida a partir de puntos característicos de la imagen (Figura 4.3(b)), a la cual se la asocia también una función de verosimilitud.

(a) (b)

Figura 4.3 Sistema sensor (a) y modelo de observación para visión (b) en [Vermaak 2001]

El espacio de estados aquí utilizado consta de las coordenadas (x, y) del centroide de la curva que define el contorno de la cabeza y la propia curva.

Una vez obtenidos los datos procedentes de los sensores de audio y visión de la forma comentada anteriormente, la fusión de ambas se realiza mediante un PF empleando muestreo segmentado ([MacCormick 2000]) que emplea las PDF de verosimilitud de los sensores de audio y video en un proceso con dos pasos, en los que la PDF de verosimilitud de audio se emplea para la estimación de la coordenada x y la de visión para la coordenada y.

La Figura 4.4 muestra los resultados obtenidos, en el que el target que esta hablando en cada momento se representa con la curva que marca su contorno. En los experimentos se utiliza un set de 20 partículas.

Figura 4.4 Resultados obtenidos en los experimentos de [Vermaak 2001]

En **[Zotkin 2001]** se vuelve a plantear la utilización de métodos probabilísticos basado en el Teorema de Bayes y más concretamente el PF, para realizar la fusión de los datos de audio y visión, los cuales en este caso proceden de múltiples sensores de audio y múltiples cámaras calibradas.

En **[Beal 2002]** se recurre a la misma configuración de sensores STAC ya utilizada en [Vermaak 2001]con la diferencia de que aquí el método probabilístico de fusión propuesto es el EM (Expectation-Maximization).

Por otro lado el laboratorio IDIAP suizo (del que ya se referenciaron trabajos de tracking tanto en la parte de audio como en la de visión) presenta en **[Gatica-Perez 2003a]** una solución de tracking empleando en este caso fusión audiovisual. Este grupo constituye uno de los más punteros en cuanto a tracking audiovisual se refiere, por lo además del presente, otros trabajos suyos serán revisados más adelante.

El sistema de sensores utilizado en este trabajo consta de un array circular de 8 micrófonos ($f_s = 16KHz$) y una cámara de visión no calibrada (25fps), situados en el entorno de la manera presentada en la Figura 4.5. En este entorno se plantea un proceso de calibración AV basado en diversas secuencias de entrenamiento.

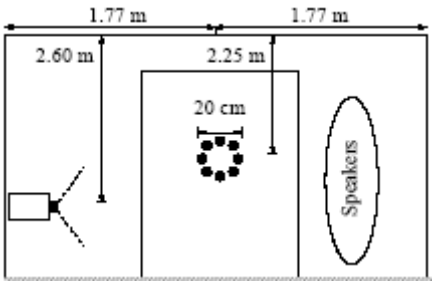


Figura 4.5 Test setup utilizado en [Gatica-Perez 2003a]

La forma básica a seguir en la escena se modela como una elipse parametrizada, adecuada para el tracking de cabezas, en base a la cual se define un subespacio de transformación cuyos parámetros constituyen el *espacio de estados* $x_t = \left(T_t^x, T_t^y, s_t\right)$, donde T_t^x y T_t^y representan las traslaciones y s el escalado.

El modelo de actuación se define como $\vec{x}_t = A\vec{x}_{t-1} + B\vec{w}_t$, donde A y B son parámetros del modelo y $\vec{w}$ es ruido blanco.

El *modelo de observación visual* ($p(\vec{y}_t^{vid} \mid \vec{x}_t)$) se basa en una extracción de bordes a partir de los cuales se identifican contornos candidatos a elipses.

Respecto al *modelo de observación de audio* ($p(\vec{y}_t^{aud} \mid \vec{x}_t)$) se propone la utilización de GCC para la estimación de los TDOA, a la cual se le aplica una transformación de fase PHAT para reducir el efecto de las reverberaciones, de cuya combinación surge el algoritmo GCC-PHAT (2.4.2.1.3).

La combinación de ambos modelos de observación se define como:

$$p(\vec{y}_t \mid \vec{x}_t) = p(\vec{y}_t^{vid} \mid \vec{x}_t) \cdot p(\vec{y}_t^{aud} \mid \vec{x}_t) \tag{4.1}$$

La fusión de la información se realiza mediante un Importance PF (I-Condensation), de manera que la información de audio se emplea en los pasos de corrección y reinicialización, de tal manera que se resuelven los procesos de inicialización y recuperación ante fallos (más problemáticos en visión), a la par que se obtiene una localización precisa gracias a la información de visión (principal problema de los arrays de micrófonos).

En la Figura 4.6 se presentan resultados reales obtenidos en este trabajo, en los cuales se emplea un set de 500 partículas y el tracker es iniciado automáticamente cuando el target entra en la escena. El algoritmo resultante permite una ejecución en tiempo real.

(a) (b)

Figura 4.6 Experimentos extraídos de [Gatica-Perez 2003a]. (a) Clutter Visual y oclusiones AV (b) Cambios de turno de palabra en una conversación

Siguiendo esta línea de investigación en **[Gatica-Perez 2003b]** se presenta un trabajo similar al anterior pero en este caso empleando 3 cámaras no calibradas, ubicadas de la forma representada en la Figura 4.7. El objetivo de esta nueva topología es implementar en el algoritmo un procedimiento que permita el switcheo automático entre cámaras en función de la que presente una vista más adecuada.

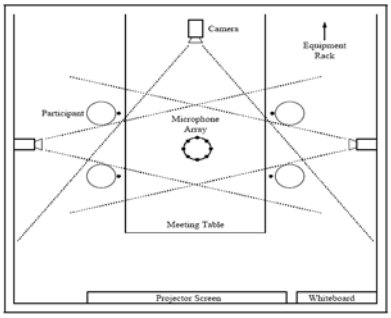


Figura 4.7 Test setup utilizado en [Gatica-Perez 2003b]

Una de las conclusiones más importantes que se pueden obtener de este apartado, es la gran aceptación por parte de la comunidad científica del PF como algoritmo de fusión de información audiovisual y los contrastados resultados proporcionados por el mismo.

4.3.2 Seguimiento de múltiples targets (MTT)

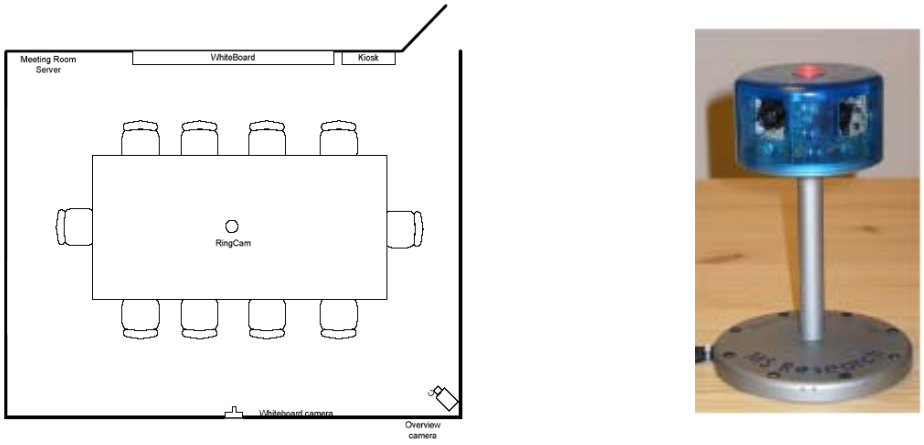
En este punto cabe destacar, que ninguno de los trabajos presentados hasta el momento en este apartado permite realizar el seguimiento, a partir de datos de audio y visión, de la posición y estado (hablando o en silencio) de un número múltiple y variable de targets ubicados en un entorno real.

A continuación se realiza un estudio de los trabajos más destacables de la comunidad científica que tratan de abordar este problema, los cuales se dividen en orientados a sistema y orientados a modelo tal y como se especifica en 4.2.

4.3.2.1 Orientados a sistema (System Oriented)

Uno de los primeros trabajos presentados en este ámbito es el desarrollado por investigadores de Microsoft en **[Cutler 2002]**, en el ámbito de aplicación de las teleconferencias y la grabación de reuniones.

El sistema sensor propuesto está compuesto por un array circular de 8 micrófonos omnidireccionales montados sobre la base de la estructura que contiene 5 cámaras calibradas ubicadas sobre un patrón pentagonal, las cuales proporcionan un field of view de 360º (Figura 4.8(b)). Adicionalmente se utilizan una "overview camera" y una "whiteboard camera" con el objetivo de captar una visión global de la escena y de la pizarra respectivamente. La configuración de sensores propuesta se representa en la Figura 4.8(a).



(a) (b)

Figura 4.8 Test setup (a) y Ring camera (b) utilizadas en [Cutler 2002]

Para la localización basada en señales acústicas se implementa un algoritmo GCC-PHAT (2.4.2.1.3). A dicha señal acústica se le aplica previamente un filtro paso banda [200-4000]Hz (2.4.2.5.4) con el objetivo de enfatizar las frecuencias en las que se encuentra mayor cantidad de potencia acústica y la reducción del ruido estacionario (e.g. aire acondicionado o proyectores).

El tracking visual se implementa a partir de un Hidden Markov Models (HMM), que mediante la expansión del vector de observación utilizado permite incorporar de manera probabilística diferentes características (e.g. contorno, histograma de color de fondo y target) al mencionado tracker visual. La necesidad de diferentes características para el tracking visual se justifica en el artículo en base al gran clutter presente en el entorno.

El sistema se inicializa de manera automática en base a la información procedente del sistema de audio o de visión, por lo que es válido para un número variable de targets.

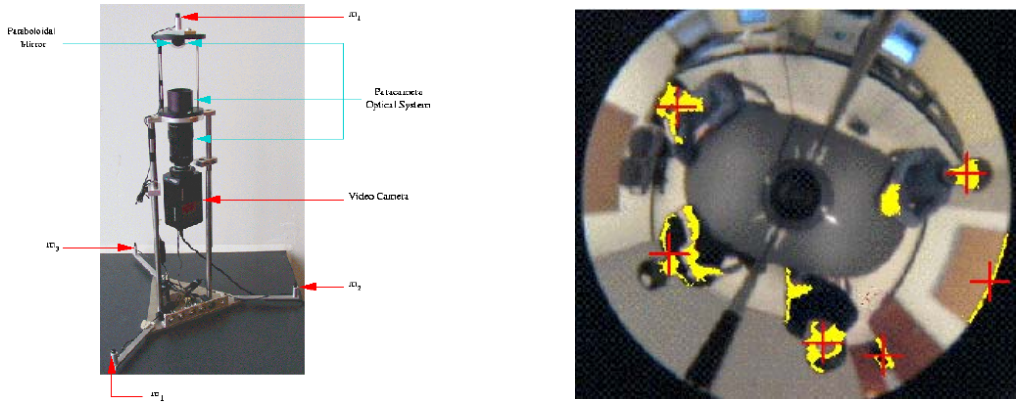
Por otro lado, con el objetivo de robustecer el algoritmo se implementa una verificación jerárquica del sistema de tracking estructurado en dos niveles. El nivel más bajo esta basado en el histograma de color y presenta un bajo tiempo de ejecución, adecuado para aplicaciones de tiempo real. Si el target no pasa este primer nivel, se procede a una verificación más segura aunque lenta, que en caso de arrojar un resultado negativo descartará el objeto.

La fusión se realiza de nuevo mediante un PF (en configuración bootstrap) no informando el artículo sobre el número de partículas utilizadas ni la velocidad de ejecución del algoritmo, el cual es muy probable que no se ejecute en tiempo real debido a la finalidad última de la aplicación es la grabación y posterior revisión de reuniones.

El artículo es también interesante desde el punto de vista de HW, ya que proporciona referencias y modelos de los sensores y equipos utilizados para la implementación del sistema.

En **[Kapralos 2003]**, investigadores canadienses de la universidad de York proponen un sistema diferente al resto debido al su carácter determinístico, en lugar de probabilístico de los trabajos presentado hasta el momento.

El sistema sensor propuesto se presenta en la Figura 4.9(a), y está compuesto por una cámara omnidireccional (de resolución limitada) y un array de 4 micrófonos ubicados de forma piramidal y calibrados entre si.



(a) (b)

Figura 4.9 Sistema sensor (a) y resultado del algoritmo (b) obtenido en [Kapralos 2003]

Para cada frame obtenido con la cámara omnidirecional, el algoritmo de visión realiza una extracción de blobs de color de piel, la cual se realiza en el espacio transformado HSV[16] con el objetivo de minimizar la sensibilidad del algoritmo ante cambios de iluminación. Para la extracción de estos blobs de color se aplica un proceso de eliminación de fondo, gracias al cual las partes de la imagen que no sufren un cambio significativo no son procesadas por el algoritmo. Posteriormente, y una vez identificados los píxeles de la imagen como candidatos a piel, se realiza una agrupación en regiones que elimina los píxeles que no se encuentran dentro de un umbral determinado.

Una vez obtenidas las regiones anteriormente comentadas, se realiza un apuntamiento del array de micrófonos hacia esas zonas concretas empleando un *delay and sum beamforming* con objeto de identificar las personas que están hablando en cada momento. Los resultados obtenidos se muestran en la Figura 4.9(b).

En **[Siracusa 2003],** investigadores del instituto MIT de Massachussets proponen un sistema audiovisual basado en una cámara estereo y un array lineal de micrófonos de 2 elementos, con el objetivo de obtener, a partir de métodos probabilísticos, no sólo la posición de los targets del entorno e identificar el que está hablando en cada momento, sino también la pose del hablante a partir de la cual se puede inferir a quien está hablando.

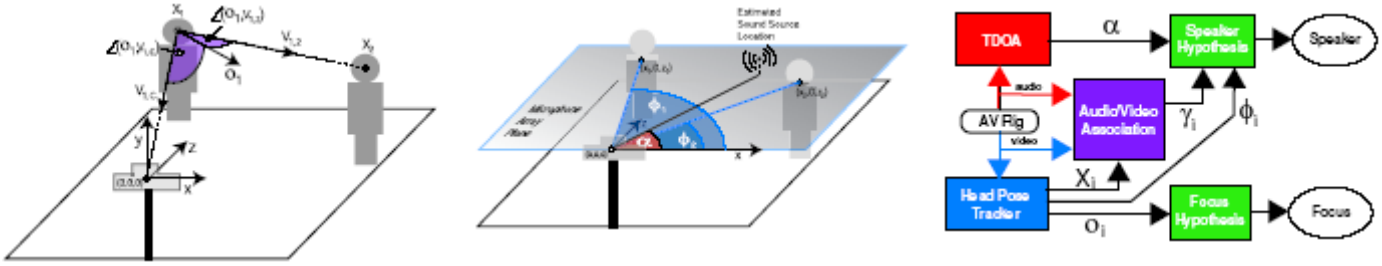
El sistema consta de 3 módulos claramente diferenciados (Figura 4.10):

1. *El módulo de visión* se encarga de la estimación de la posición $(x, y, z,)$, la pose O_i y las orientaciones hacia la cámara $V_{i,c}$ y hacia la otra persona que compone la escena $V_{i,j}$, para un total de seis grados de libertad $(x, y, z, O_i, V_{i,j}, V_{i,c})$, de cada una de las personas del entorno de manera independiente (las escenas siempre se componen de 2 o 3 individuos).

[16] El HSV (Hue, Saturation, Value) o tonalidad, saturación, valor, define un de espacio de color transformado representado en coordenadas cilíndricas en función de sus componentes constitutivas.

2. *El módulo de audio* se encarga de la estimación de la DOA (Direction Of Arrival) de la señal de audio, la cual se obtiene a partir de los TDOA obtenidos mediante la GCC implementada en este caso en el dominio del tiempo.

3. *El tercer módulo* se encarga de combinar la información de los dos anteriores con el propósito de sincronizarlos.



(a) (b) (c)

Figura 4.10 Modelo de visión (a) de audio (b) y sistema completo (c) para [Siracusa 2003]

Las hipótesis se dividen en dos sets con el objetivo de tomar decisiones independientes respecto a si cada uno de los targets está hablando o no y hacia donde está mirando, para lo cual se emplean modelos estadísticos sencillos y las observaciones obtenidas de cada uno de los módulos anteriores.

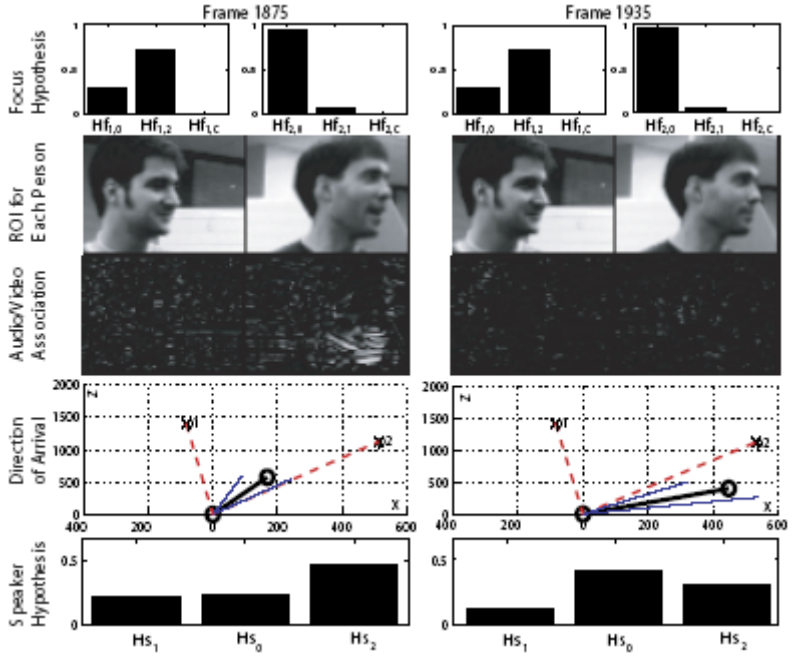
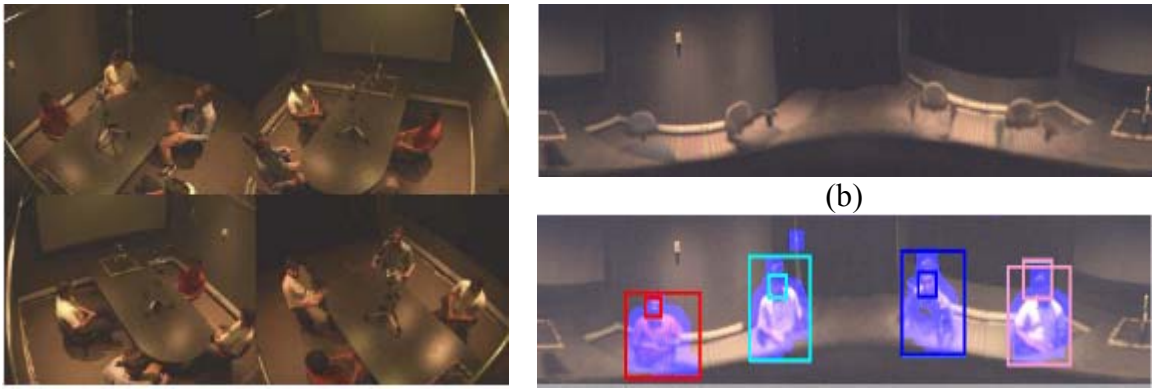


Figura 4.11 Resultados extraídos de los experimentos de [Siracusa 2003]

La Figura 4.11 presenta los resultados obtenidos con el algoritmo anterior, en el que las cámaras estereo tienen una resolución de 320x240 píxel y una frecuencia de adquisición de 25fps, y el array de micrófonos empela una $f_s = 44.1KHz$.

Las escenas utilizadas para los test son simples, desde el punto de vista de que sólo se utilizan un número fijo de 2 o 3 individuos, los cuales no hablan de manera simultánea y el sistema de visión no se ve sometido a oclusiones totales o parciales.

En **[Busso 2005]**, investigadores de la University of Southern California (USC) proponen un nuevo sistema que combina sensores de audio y visión. El sistema de visión consta de 4 cámaras calibradas y situadas en las esquinas de la habitación, capaces de adquirir a una velocidad de 15fps y una cámara omnidireccional situada en el centro de la mesa de reuniones, mientras que el sistema acústico implementa un array de 16 micrófonos omnidirecionales ($f_s = 48KHz$), de los cuales 14 están distribuidos a lo largo de un cuadrado de 50x50cm y dos de ellos se encuentran en el centro del mismo en una posición más elevada que los anteriores. En la Figura 4.12 (a) y (b) se observan las imágenes obtenidas por las 4 cámaras de las esquinas y la panorámica central, lo cual permite hacerse una idea de la información de visión disponible y del entorno en que se realizan los ensayos.



(a) (b) (c)

Figura 4.12 Imágenes de las 4 cámaras de las esquinas (a) de la omnidireccional (b) y resultado del algoritmo (c) en los experimentos extraídos de [Busso 2005]

El sistema de audio emplea el algoritmo *Fraccional Lower Order Statistics-Phase Transform Method* (FLOS-PHAT) para estimar los TDOA entre cada pareja de micrófonos, a partir de los cuales obtiene la posición de los diferentes targets de la escena mediante el algoritmo *One step Least Squares* (OSLS).

El sistema de visión se basa en una extracción de fondo para identificar las zonas de la imagen que cambian, las cuales son combinadas en regiones. De las regiones así obtenidas se eliminan las sombras y se realiza una representación poligonal aproximada de las siluetas resultantes, la cual facilita la implementación del algoritmo en tiempo real. Los máximos locales de cada uno de dichos polígonos son identificados como las cabezas de los individuos.

La información de ambos sistemas es fusionada para obtener la ubicación de cada uno de los participantes en el meeting además de su identidad, obteniéndose los resultados presentados en la Figura 4.12(c). El algoritmo resultante permite su implementación en tiempo real a una velocidad de 13fps en un PC de 2.8 GHz.

4.3.2.2 Orientados a modelo (Model Oriented)

En este apartado se presenta una segunda generación de algoritmos con una filosofía radicalmente opuesta a la anterior, en la cual se trata de encontrar una formulación matemática conjunta que permita obtener la estimación de la posición de los targets en el entorno de manera directa a partir de las medidas de los sistemas sensores (en nuestro caso audio y visión).

Todas estas técnicas basan su formulación matemática en el filtro de partículas (PF) pero difieren en la elección del espacio de estados, los modelos de actuación y observación y la técnica de muestreo.

Tal y como se muestra a continuación, los resultados obtenidos con estas técnicas superan ampliamente a los presentados en el apartado anterior.

El primero de los trabajos aquí presentados, desarrollado por investigadores del Instituto MIT en **[Checka 2004]**, permite el seguimiento y obtención del estado (silencio o hablando) de manera simultanea de un número variable de individuos (2 o 3 en los ensayos presentados).

Para ello utilizan un sistema sensor basado en 2 cámaras calibradas y 4 sub-arrays lineales de 4 micrófonos cada uno, en un entorno de 6x6 metros como el presentado en la Figura 4.13.

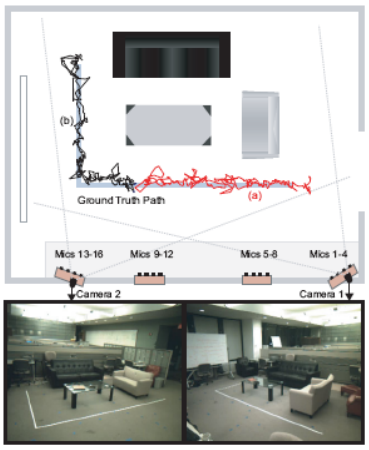


Figura 4.13 Test setup utilizadas utilizado en [Checka 2004]

Este trabajo propone la utilización del siguiente *vector de estados* aumentado:

$$\vec{X}_t = \left(\vec{n}_t, \vec{x}_t^1, ..., \vec{x}_t^n\right) \xrightarrow{con} \vec{x}_t^i = [x, y, h, s] \tag{4.2}$$

donde n_t representa el número de targets de la escena, mientras que el par $[x, y]$ indica la posición respecto al plano formado por el suelo, h la altura y s estado activo o inactivo (en cuanto a habla) para cada uno de los targets.

Para el *modelo de estados* se elige un *"zeroth order"* que aplica fuerzas gausianas aleatorias a cada una de las variables x, y, h, mientras que la s se actualiza de manera conjunta para todos los targets, lo que permite modelar la dependencia entre los individuos. La adicción y eliminación de targets se realiza mediante un proceso de predicción que actualiza las probabilidades v_{remain} y v_{add} para cada instante de tiempo de manera similar a lo propuesto en [Isard 2001].

El *modelo de observación* completo se define a partir de dos términos: El primero para visión, basado en una extracción de fondo realizada a partir del conocimiento del mismo en ausencia de individuos, gracias a un proceso de entrenamiento previo y que proporciona un modelo cilíndrico de los targets caracterizados por un radio r y una altura h. El segundo para audio, obtenido a partir de la Short-Time Fourier Transform (STFT) de las señales recibidas en cada uno de los micrófonos.

La fusión se realiza mediante un proceso de Importante Sampling básico, por lo que es probable que el algoritmo se vuelva rápidamente ineficiente con el aumento del

número de targets, lo cual explicaría que en los ensayos presentados en el trabajo sólo se utilicen escenas con 2 o 3 targets.

La Figura 4.14 muestra los resultados obtenidos por el algoritmo, en el que la velocidad de adquisición de las cámaras es de 20 fps, $f_s = 16KHz$ para los arrays de micrófonos y se utilizan 200 partículas inicializadas de forma manual. Los individuos que están hablando se identifican con un rectángulo de color azul claro y los que no con un rectángulo de color azul oscuro.

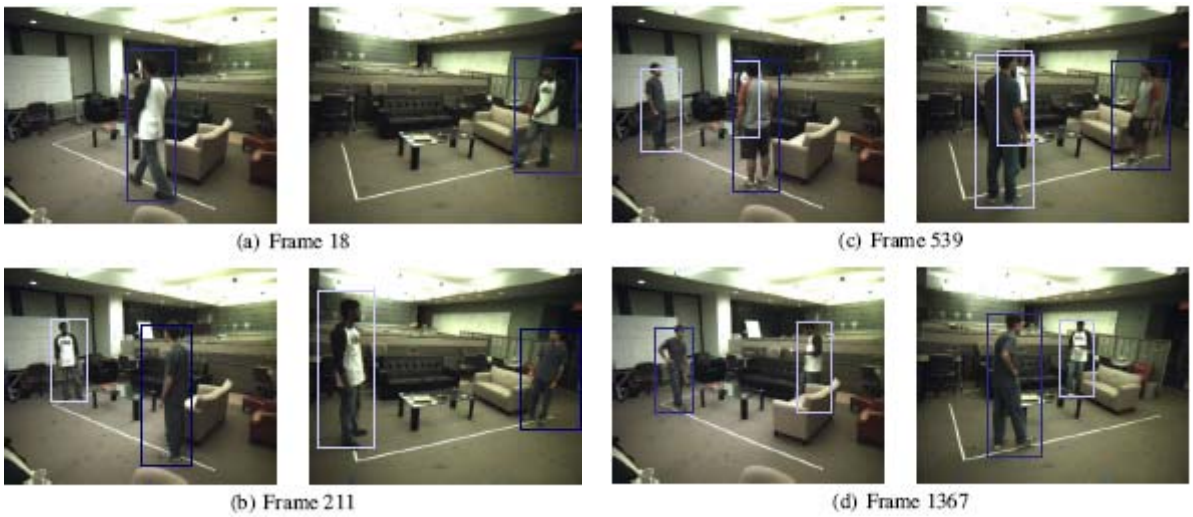


Figura 4.14 Resultados de los experimentos realizados en [Checka 2004]

El trabajo presentado en **[Chen 2004]** utiliza la misma configuración de sensores que [Cutler 2002] (Figura 4.8). El diagrama de bloques implementado por el algoritmo se presenta en la Figura 4.15.

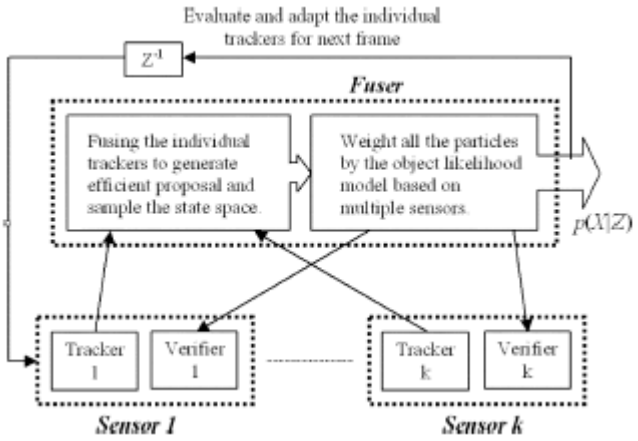


Figura 4.15 Diagrama de bloques del algoritmo implementado en [Chen 2004]

En la figura anterior se aprecia claramente que Chen propone la utilización de un PF independiente para el seguimiento de cada target del entorno, cada uno de los cuales se define en base a un vector de estados para la representación de la cabeza humana, $\vec{X}_t = \left(x_t^c, y_t^c, \alpha_t\right)$, compuesto por la posición espacial del centroide de la elipse y su eje mayor respectivamente. Para la obtención de la estimación del vector de estados, cada PF utiliza una mezcla de componentes, obtenida a partir de 3 seguidores de características individuales basados en: contorno, color y audio.

Las estimaciones de cada PF individual constituyen la entrada de otro PF global, encargado de realizar la fusión de la información y que será el responsable de realizar el muestreo del espacio de estados y de calcular los pesos de las partículas en función de los modelos de verosimilitud. Este proceso permite aumentar la robustez en el tracking de los PFs individuales y proporciona información a los verificadores de los PF individuales, permitiendo así cerrar el lazo.

Los resultados obtenidos con el algoritmo se presentan en la Figura 4.16 y según el autor permite realizar un tracking en tiempo real de 5 o 6 personas en un PC de 2.2 GHz. Los cuadrados representan las salidas de los trackers de color y contorno, la línea negra la del tracker de audio y la gris la fusión de los 3.

Figura 4.16 Resultados de los experimentos realizados en [Chen 2004]

El autor no proporciona datos del número de partículas utilizada por cada uno de los trackers individuales, ni por el PF de fusión del nivel superior. Tampoco se dice específicamente que el número de targets del entorno pueda ser variable, de lo cual se deduce que no debe serlo. Adicionalmente existen individuos del entorno (los más alejados del sistema sensor) que no son detectados.

Tal y como se puso de manifiesto en los apartados dedicados a tracking basado sensores de visión y audio de manera independiente, el laboratorio suizo IDIAP constituye en ambos casos un grupo puntero en nuevas aportaciones a la comunidad científica, tal y como se demuestra en trabajos como [Lathoud y Odobez 2007] (en la parte de audio) o [Odobez 2004] (en la parte de visión), estudiados con anterioridad en el presente estado del arte. Esto se pone aún más de manifiesto para el caso de la fusión audiovisual, en la trabajos como **[Gatica-Perez 2007]** (desarrollado a partir de [Gatica-Perez 2005]) pueden ser considerados como el estado del arte actual en materia de localización audiovisual, lo cual implica la necesidad de un estudio en profundidad de dicho trabajo.

Este trabajo, llevado a cabo en el contexto del proyecto AMI (Augmented Multi-party Interaction), se desarrolla en el mismo entorno ya presentado en la Figura 4.7 y que utiliza como sistemas sensores tres cámaras no calibradas con un field-of-view prácticamente no solapado y un array circular de 8 micrófonos situados encima de la mesa de reuniones.

En el mismo se propone la utilización de un *vector de estados aumentado,* en el que la posición de la persona se representa en el plano imagen a partir de una transformación afín (traslación + escalado) a través de $\left(u_{i,t}, v_{i,t}, s_{i,t}\right)$ y se añade una variable que indica si cada uno de los targets esta hablando $\left(k_{i,t}\right)$ en el instante de tiempo t (0: silencio, 1: hablando).

$$\vec{X}_t = \left(\vec{X}_{i,t}\right) \xrightarrow{con} \vec{X}_{i,t} = \left(u_{i,t}, v_{i,t}, s, k_{i,t}\right) \tag{4.3}$$

El vector de estados anterior implica la elección de modelos de observación acordes con el mismo y la utilización de métodos que eviten que el vector de estados aumentado (con el aumento exponencial de carga computacional que implica) impida una ejecución en tiempo real del algoritmo final.

El *modelo de actuación o modelo dinámico* consta de dos factores:

1. El primero de ellos describe el modelo de actuación de cada uno de los targets de manera independiente, para lo cual el modelo de cada uno se factoriza como:

$$p\left(\vec{X}_{i,t} \mid \vec{X}_{i,t-1}\right) = p\left(\vec{x}_{i,t} \mid \vec{x}_{i,t-1}\right) \cdot p\left(k_{i,t} \mid k_{i,t-1}\right) \tag{4.4}$$

Donde la distribución continua $p\left(k_{i,t} \mid k_{i,t-1}\right)$ es modelada como un modelo autorregresivo de segundo orden ([Isard 1998a]) con:

$$p\left(k_{i,t} \mid k_{i,t-1}\right) = \begin{bmatrix} \beta_{00} & \beta_{01} \\ \beta_{10} & \beta_{11} \end{bmatrix} \tag{4.5}$$

donde a $p\left(k_{i,t} \mid k_{i,t-1}\right)$ se la denomina matriz de transición probabilística (TPM) con parámetros $\beta_{01} = 1 - \beta_{00}$ y $\beta_{10} = 1 - \beta_{11}$.

2. El segundo describe las interacciones (e.g. oclusiones) entre los distintos individuos de la escena, las cuales se modelan a partir de un MRF ([Khan 2004]) basado en la información de visión y en el que los vértices representan los targets y los enlaces se definen entre cada pareja de objetos en cada instante de tiempo.

   La medida de oclusión espacial utilizada se conoce como "*precision*" (ν) "*recall*" (ρ), ligeramente modificada respecto a su formulación original, en la que un valor 0 para ambas variables identifica un solapamiento total y un 1 un solapamiento nulo.

El *modelo de observación* se define como la combinación de los modelos audio, forma y estructura espacial como $\vec{Y}_t = \left(\vec{Y}_t^{\,a}, \vec{Y}_t^{\,sh}, \vec{Y}_t^{\,st}\right)$. Cada uno de los términos se define a continuación:

$$p\left(\vec{Y}_t \mid \vec{X}_t\right) = \prod_{i \in \mathrm{I}_t} p\left(\vec{Y}_{i,t}^{\,a} \mid \vec{X}_{i,t}\right) \cdot p\left(\vec{Y}_{i,t}^{\,sh} \mid \vec{X}_{i,t}\right) \cdot p\left(\vec{Y}_{i,t}^{\,st} \mid \vec{X}_{i,t}\right) \tag{4.6}$$

1. M*odelo de observación de audio:* se define en tres pasos:

   a. *Localización de targets*: Se encarga de la generación de las posiciones candidatas a ubicar personas hablando en el entorno. Esta tarea se lleva a cabo mediante el algoritmo SRP-PHAT (2.4.2.3) implementado en el dominio de la frecuencia, a partir de la suma de las GCC-PHAT de cada

pareja de micrófonos (2.4.2.3.2). El algoritmo realiza una búsqueda en un grid de puntos fijos de dicho entorno (2.4.2.4.5).

b. *Clasificación de las localizaciones anteriores como habla/ruido:* Aplicando el algoritmo *short-term clustering* (2.4.3.3) sobre las localizaciones anteriores se identifica cuales proceden de un target y cuales de ruido (las cuales son eliminadas). Según el artículo, esta implementación mejora la eficiencia y robustez del algoritmo respecto a la más típica basada en un VAD.

c. *Calibración AV*: Se requiere para el mapeo de las localizaciones del sistema de audio en el plano imagen de las cámaras y se realiza en base a una serie de secuencias de entrenamiento en un proceso offline.

2. *Modelo de Observación de forma:* Se basa en una detección de bordes sobre la cual se realiza una búsqueda de contorno elípticos en base al modelo definido en [Isard 1998a]. El resultado es un vector de candidatos, del cual se extraerán los targets reales en base a un modelo de verosimilitud clásico similar al implementado en [Isard 1998a].

3. *Modelo de Observación de estructura:* Se basa en una representación paramétrica del solapamiento entre blobs de piel y la forma de la cabeza, basada en que la presencia de píxeles de piel en la cabeza se limita a unas zonas específicas dentro de la configuración elíptica de la misma, típicamente las zonas central e inferior, aunque también fuera de la misma (e.g. cuello, manos). La Figura 4.17 representa este concepto.

(a) (b) (c)

Figura 4.17 Estructura espacial en [Gatica-Perez 2007]. Estructura espacial dada (a), blobs de piel (b) partes de división de la elipse (c).

Los blobs de piel se obtienen a partir de procedimientos estándar, para lo cual se obtiene una *Gaussian mixture model* (GMM) de 20 componentes de color de

piel en el espacio RGB, a partir de secuencias de entrenamiento con personas de diferentes razas. Los blobs de piel se clasifican en función de un umbral de verosimilitud, al que le sigue un proceso de extracción de blobs basado en consideraciones morfológicas.

Por otro lado, tal y como se ha indicado en apartados anteriores, la alta dimensión del espacio de estados obtenida fruto de la expansión del mismo provoca un aumento exponencial de la carga computacional, lo cual lo hace no implementable en tiempo real. Para solucionar este problema en el artículo se propone la utilización de un *Markov Chan Monte Carlo Particle Filter* (MCMC-PF), el cual utiliza un *Metropolis-Hastings* (MH) en cada instante de tiempo ,con objetivo la ubicación de las partículas de forma eficiente en zonas de alta verosimilitud,

Para los procesos de nacimiento y muerte *(birth/death)* se decide implementar un mecanismo sencillo que en cada instante de tiempo establece los targets de la escena y a los que posteriormente se aplica el MCMC-PF para su detección. Una formulación alternativa podría consistir en aplicar el MCMC-PF para la obtención de nacimiento y muerte (lo cual es realizable teóricamente), a través de un *reversible-jump MCMC*. Sin embargo en el artículo se basan en la ineficiencia de las partículas que referencian a un número incorrecto de objetos como base para desestimar su utilización.

La identificación de nuevos targets (*birth)* se realiza en base a la obtención de blobs piel, buscados en zonas de alta probabilidad de aparición de individuos. De este proceso se obtienen un conjunto de candidatos, que en base a su verosimilitud respecto del modelo se observación visual (color y estructura) identificará los nuevos targets reales y filtrará el resto.

La eliminación de targets *(death)* se produce cuando alguno de ellos abandona la imagen o cuando la verosimilitud visual que presentan es muy baja.

Los resultados presentados en el trabajo muestran que el algoritmo es ejecutable en tiempo real utilizando un número fijo de 500 partículas y 20 iteraciones del MRF en todos los ensayos. La Figura 4.18 muestra un ejemplo de uno de ellos, en el que se realiza un tracking de 4 personas de manera simultanea, además de indicar en cada momento si los diferentes targets está hablando o no.

(a) (b)

Figura 4.18 Resultados de los experimentos de [Gatica-Perez 2007] sin la aplicación del MRF (a) y con la aplicación del MRF (b)

En el artículo se presentan resultados adicionales bajo diferentes configuraciones del entorno y orientados a evaluar las mejoras obtenidas gracias a las diferentes aportaciones realizadas en el artículo. Adicionalmente en la página [IDIAP] se pueden encontrar los videos de los ensayos realizados en referencia a este trabajo.

Otros trabajos interesantes se presentan en [Maganti 2006], [Maganti 2006a] y [Wolfel 2005], los dos primeros llevados a cabo por el Instituto IDIAP.

4.4 Pose

Tal y como se comentó en el apartado de audio, los trabajos centrados en la obtención de la pose a partir de información acústica son muy escasos, de lo cual se deduce que esta misma situación es trasladable al entorno de la fusión audiovisual.

El trabajo más interesante de los encontrados que trata de resolver esta problemática es el presentado en **[Segura 2007]**, el cual sigue un esquema orientado a sistema como el presentado en la siguiente figura:

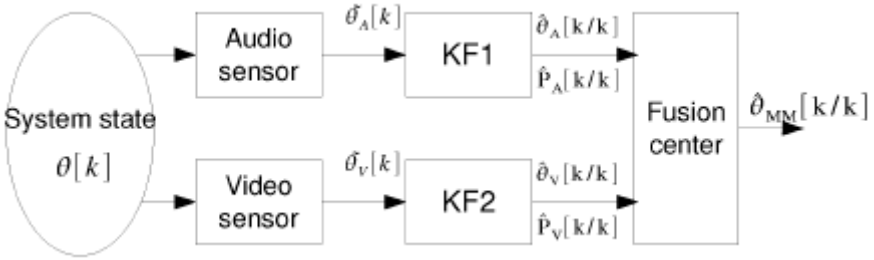
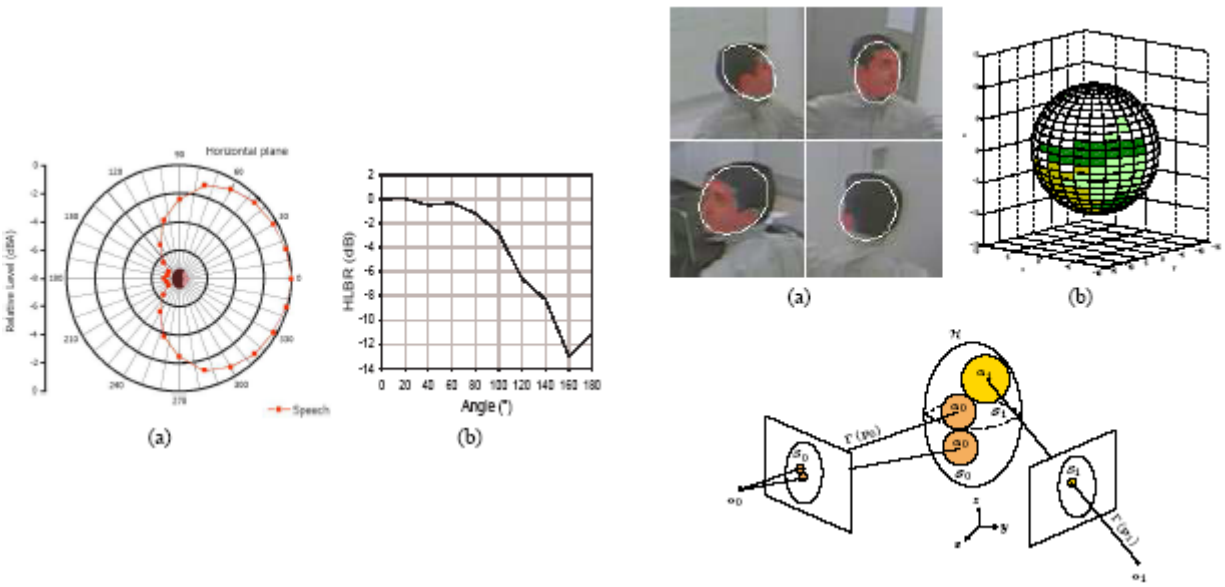


Figura 4.19 Esquema orientado a sistema empleado en [Segura 2007]

La estimación de pose a partir de la información acústica se realiza mediante el algoritmo HLVR ya tratado en la parte de audio (2.4.4.2), mientras que la de visión se obtiene mediante la extracción de blobs de piel de cada una de las vistas del target, a partir de las cuales se realiza una reconstrucción que permite obtener la pose deseada. La figura siguiente aclara los conceptos aplicados para cada uno de los algoritmos:



(a) (b)

Figura 4.20 Aclaración de los algoritmos implementados para Audio (a) y visión (b) en [Segura 2007]

La Figura 4.21 presenta los resultados reales obtenidos por el algoritmo, que en palabras del autor presenta un 41.45 % de reducción del error entre el mejor de los estimadores monomodales de pose (visión) y la fusión visual propuesta. En referencia a las métricas utilizadas para evaluar el algoritmo, cabe destacar que se ajustan a las definidas en el CLEAR Evaluation Plan dentro del proyecto CHIL ([CLEAR 2007b]) y presentadas más adelante en este mismo trabajo.

Figura 4.21 Resultados obtenidos en [Segura 2007]. En verde resultado de la estimación acústica y en rojo la visual.

4.5 BBDD y métricas de evaluación

4.5.1 BBDD

Para concluir el apartado dedicado a fusión audiovisual, resaltar la importancia de contar con secuencias de audio y video (sincronizados entre sí) que permitan llevar a cabo el testeo, maduración e implementación de los algoritmos propuestos.

A este respecto, la situación ideal consistiría en poder contar con un sistema que permitiera generar nuestras propias secuencias de test, pero hasta ese momento una alternativa interesante puede consistir en la utilización de BBDD ya creadas.

En este apartado se hace especial hincapié en las BBDD que proporcionan información audiovisual sincronizada entre sí, al encontrarse más en línea con el objetivo perseguido en la tesis que aquí comienza, a pesar de lo cual otras BBDD que proporcionan información de un único tipo de sensor (e.g. HIFI-MM1 para audio utilizada en el presente estado del arte) pueden ser interesantes para testear determinados aspectos del algoritmo. Las BBDD audiovisuales más importantes son:

- *AV 16.3:* Esta base de datos, implementada dentro del proyecto AMI, esta disponible gratuitamente y proporciona secuencias de audio y video generados en la sala de ensayos del proyecto IDIAP, presentada en la Figura 4.7, a partir de 3 cámaras de visión y un array de micrófonos circular de 8 elementos. Esta base de datos ha sido utilizada para la generación de parte de las tablas presentadas en el apartado de audio. Para una explicación más detallada de la misma ver [Gatica-Perez 2004].

- *AV 16.7:* Esta base de datos ha sido generada específicamente para el testeo de algoritmos de MTT y contiene secuencias de hasta 4 personas hablando y moviéndose de manera simultanea, por lo que su utilización en el entorno de esta tesis debería ser valorada. Para una descripción más amplia de la misma ver [Smith 2006].
- *Multi-Channel Wall Street Journal Audio-Visual (MC-WSJ-AV)*: Incluye secuencias únicamente de grabaciones de audio, con diferente número de hablantes y bajo situaciones variadas (e.g. moviéndose, parados…etc).

Otra fuente posible de secuencias de análisis es coordinada por el proyecto europeo *CHIL* (Computers in the Human Interaction Loop) dentro del proyecto CLEAR.

4.5.2 Métricas de evaluación

A la hora de desarrollar un algoritmo, es de vital importancia el disponer de un conjunto de métricas estándar que nos permitan evaluar, su precisión, fiabilidad y el grado de mejora obtenido mediante la introducción de determinadas modificaciones, así como poder comparar dichos resultados con los arrojados por otros algoritmos y verificar cual cubre de manera más óptima las especificaciones de cada sistema concreto.

El proyecto CHIL, ya nombrado anteriormente en el presente estudio, lleva varios años trabajando con éxito en esta línea y ha definido una serie de parámetros de evaluación para tracking acústico, visual y multimodal, el último de los cuales es el más interesante desde la perspectiva de este estado del arte, así como para la evaluación de las estimaciones de la pose de personas.

Los detalles de las métricas estandarizadas para tracking acústico, visual y multimodal se detallan en [Bernardin 2006] y [CLEAR 2007a], mientras que los definidos para pose de personas se describen en [CLEAR 2007b].

Es importante destacar que en el ámbito de este proyecto se organizan competiciones orientadas a la mejora continua de los algoritmos y técnicas existentes, en las cuales se ponen a disposición de los participantes secuencias de prueba y test, con

lo cual todos los algoritmos son testeados no solo bajo las mismas métricas, sino también respecto a unos datos comunes. Dichos datos pueden constituir otra posible fuente de información a utilizar a modo de BBDD durante el desarrollo de la tesis que aquí comienza.

5. Conclusiones y futuras líneas de investigación

En referencia a la parte de *audio*, la principal conclusión extraída a partir del estudio realizado es la mayor fiabilidad, precisión y robustez del algoritmo *Steered Response Power* (SRP) respecto al resto de algoritmos presentados por la comunidad científica, lo cual lo convierte en este momento en el principal candidato para ser utilizado en la tesis que aquí comienza.

Una de los principales problemas que se encuentran en aplicaciones de detección y localización con arrays de micrófonos es el de las reverberaciones, los cuales reducen sobremanera la fiabilidad de los algoritmos en entornos reales. Para mitigar este problema la comunidad científica propone la utilización de la transformada de fase (PHAT) en combinación con SRP, de cuya unión se obtiene el algoritmo conocido como SRP-PHAT (2.4.2.3.1).

Una implementación eficiente del mismo se obtiene a partir de la suma de las GCC-PHAT de todas las posibles parejas de micrófonos, implementada con éxito en diversos trabajos aquí estudiados (e.g. [Castro 2007] y [Gatica-Perez 2007]).

Otra conclusión importante consiste en la gran importancia de una adecuada elección de los parámetros, tanto de configuración del array (e.g. número de micrófonos, distancia entre micrófonos, configuración del array…etc), como del propio algoritmo (e.g. frecuencia de muestreo, frame size, ventana temporal…etc) dado su elevado impacto en la estimación final generada. En el apartado dedicado a audio (2.4.2.4) del presente estado del arte se muestran diversas tablas que permiten la cuantificación de dichos efectos (obtenidas a partir de [Castro 2007]).

Dado que nos encontramos en un espacio inteligente la ubicación de arrays de micrófonos en cada una de las paredes de la habitación (e.g. [Lehmann 2004] pag 32) podría ser interesante a la hora de diversificar la captura de datos y de reducir los efectos de las reverberaciones. La filosofía y topología de esta solución, que podría obtendría la estimación de la posición a partir de la fusión de las estimaciones generadas por cada uno de los arrays [Potamitis 2004], se enmarca sin ningún problema dentro de

los sistemas inteligentes, por lo que una implementación de este estilo debería ser considerada.

Adicionalmente se presentan estrategias de mejora del algoritmo SRP-PHAT básico (e.g. búsqueda "Coarse to fine", Noise Masking, ponderación…etc), de cuya implementación depende en gran medida el éxito del algoritmo final.

La revisión en profundidad de los trabajos relacionados con el algoritmo *Shor-Term Clustering* debería ser otra de las prioridades de los siguientes trabajos a desarrollar, debido a los excelentes resultados arrojados en trabajos como [Lathoud y Odobez 2007] y [Gatica-Perez 2007]. De esta forma evitaríamos en gran medida los problemas derivados de la implementación de un *Voice Activity Detector* (VAD) para la detección de los tramos de habla y silencio, así como que un mal funcionamiento del mismo de al traste con una buena implementación del algoritmo de estimación.

De la parte de *visión* se desprende que el filtro de partículas (PF), es el algoritmo de estimación con mayor aceptación por parte de la comunidad científica en tareas de MTT como las aquí perseguidas, debido principalmente a que permite realizar un modelado de targets con un comportamiento distinto del gaussiano (cubierto por el KF) de una manera sencilla y eficiente. Debido a su gran importancia y a su más que probable utilización en la tesis que aquí comienza, en el Anexo I se presenta la teoría básica que define su implementación.

La elección del PF como algoritmo de estimación implica otra decisión que condicionará en gran medida el resto del algoritmo y es la elección del tipo de representación del vector de estados. Descartados los que utilizan un vector de estados independiente (y por lo tanto un traker) para cada target, debido principalmente a su nula robustez ante oclusiones, la decisión esta entre implementar un vector de estados aumentado (e.g. [Lanz 2006] o [Gatica-Perez 2007]) o utilizar la multimodalidad inherente del PF (e.g. [Okuma 2004] o [Marrón 2008]) para modelar la creencia de todos los targets de la imagen. En este momento esta decisión no está clara, siendo necesario un estudio más en profundidad para tomarla.

De igual manera, la utilización de un PF requiere de la definición de un vector de estados y unos modelos de actuación (e.g. *Constante Velocity (CV)*, Autorregresivo

de orden N…etc) y observación (elipse de cabezas, blob de color de piel, splines que definen un contorno, modelos específicos de detección de caras…etc) que lo caractericen. En este punto la comunidad científica se encuentra dividida entre la utilización de modelos generales y sencillos ([Marrón 2008]) o específicos y complejos ([Tao 1999]). De una correcta decisión en este punto y en el anterior, relativas al PF, dependerá en un porcentaje muy alto el correcto funcionamiento de nuestro algoritmo.

Dentro del apartado dedicado a visión se describen también diferentes soluciones al problema de asociación, tan importante como el de estimación y si cabe peor resuelto que este debido a la menor atención prestada al mismo por parte de la comunidad científica. En este punto la elección entre algoritmos determinísticos y probabilísticos no está tan clara como en la parte de estimación, por lo que se presentan diferentes posibilidades de implementación de ambos tipos de algoritmos.

Otra decisión de diseño del algoritmo, en este caso específica de las aplicaciones de MTT, es la definición de un modelo de oclusiones o de interacción entre targets de la escena que permita la estimación de los objetos cuando la visión de los mismos es incompleta o nula. A este respecto el *Markov Random File* (MRF) se ha utilizado con buenos resultados en trabajos tan importantes como [Khan 2004], [Lanz 2006] o [Gatica-Perez 2007], lo cual le sitúa como un importante candidato a ser implementado a lo largo de la tesis.

En este punto es importante hacer notar que los MRF propuestos en los tres trabajos anteriores hacen uso exclusivamente de información procedente del sistema de visión. Una aportación interesante podría consistir en introducir información procedente del sistema de audio dentro de dicho algoritmo.

La elección del tipo de la técnica de creación y eliminación de targets (*birth / death)* depende del tipo de vector de estados, ya que mientras que un *Extended Particle Filter* (XPF) que utiliza la multimodalidad inherente del PF no requiere de un proceso de estas características, el basado en vector de estados aumentados necesita de un mecanismo específico que inserte dentro del vector de estados los nuevos targets. En este último caso se debería utilizar la información tanto de audio como de visión para una detección óptima.

Respecto al apartado de *fusión* parece clara la elección de un algoritmo *model oriented* (en prejuicio de los system oriented) debido en primer lugar a la tendencia de la comunidad científica hacia algoritmos de este tipo (e.g. [Gatica-Perez 2007]) y en segundo a que los resultados obtenidos con este tipo de solución son considerablemente mejores en cuanto a robustez fiabilidad e incluso simplicidad.

Una puntualización importante referente al vector de estados se deriva de la revisión de los objetivos perseguidos en la tarea de MTT que nos ocupa y consiste en que este deberá contener al menos las posiciones (x, y, z) y velocidades (v_x, v_z) (para la estimación de pose) de los targets, por lo que en principio parece conveniente definir un espacio de observación cartesiano (x, y, z) (en contra del espacio imagen (u, v) definido en [Gatica-Perez 2007]). Esto se debe a que la estimación proporcionada tanto por los arrays de sensores del sistema de audio, como por las cámaras de visión (suponemos aquí que al trabajar en un espacio inteligente se cuenta con visión estereoscópica y por tanto información de profundidad) se puede obtener de manera directa en este espacio de coordenadas. Otra variable introducida con éxito en otros trabajos es la que define si cada objeto concreto esta sometido a oclusiones o no (sólo tendría sentido en para el caso de vector de estados aumentado).

En el punto en el que nos encontramos parece interesante decantarnos en principio por la utilización de modelos sencillos que no sobrecarguen la carga computacional del algoritmo y así cumplir la especificación de tiempo real, pero sin embargo no estaríamos utilizando de manera óptima la información procedente del sistema de visión si no introdujéramos en el mismo ciertos parámetros derivados del conocimiento a priori del tipo de target a seguir, como por ejemplo podrían ser el centroide de la elipse definida por la cabeza del individuo o información de blobs de color de piel, por nombrar alguna. Esto nos lleva a la implementación de una solución de compromiso entre ambos extremos.

En caso de utilizar un XPFCP del estilo del propuesto en [Marrón 2008] una propuesta de fusión, conceptualmente directa, entre la información de audio y de visión consistiría en introducir en los pasos de reinicialización y asociación+corrección de la Figura 3.29 información procedente tanto de la de la parte de visión, obtenida a la salida

del algoritmo determinístico de clasificación de partículas, como de la parte de audio, procedente de la salida de un Short Term clustering aplicado a la salida del SRP-PHAT a modo de clasificador de partículas (habla/no habla). La Figura 5.1 muestra gráficamente este posible esquema de fusión.

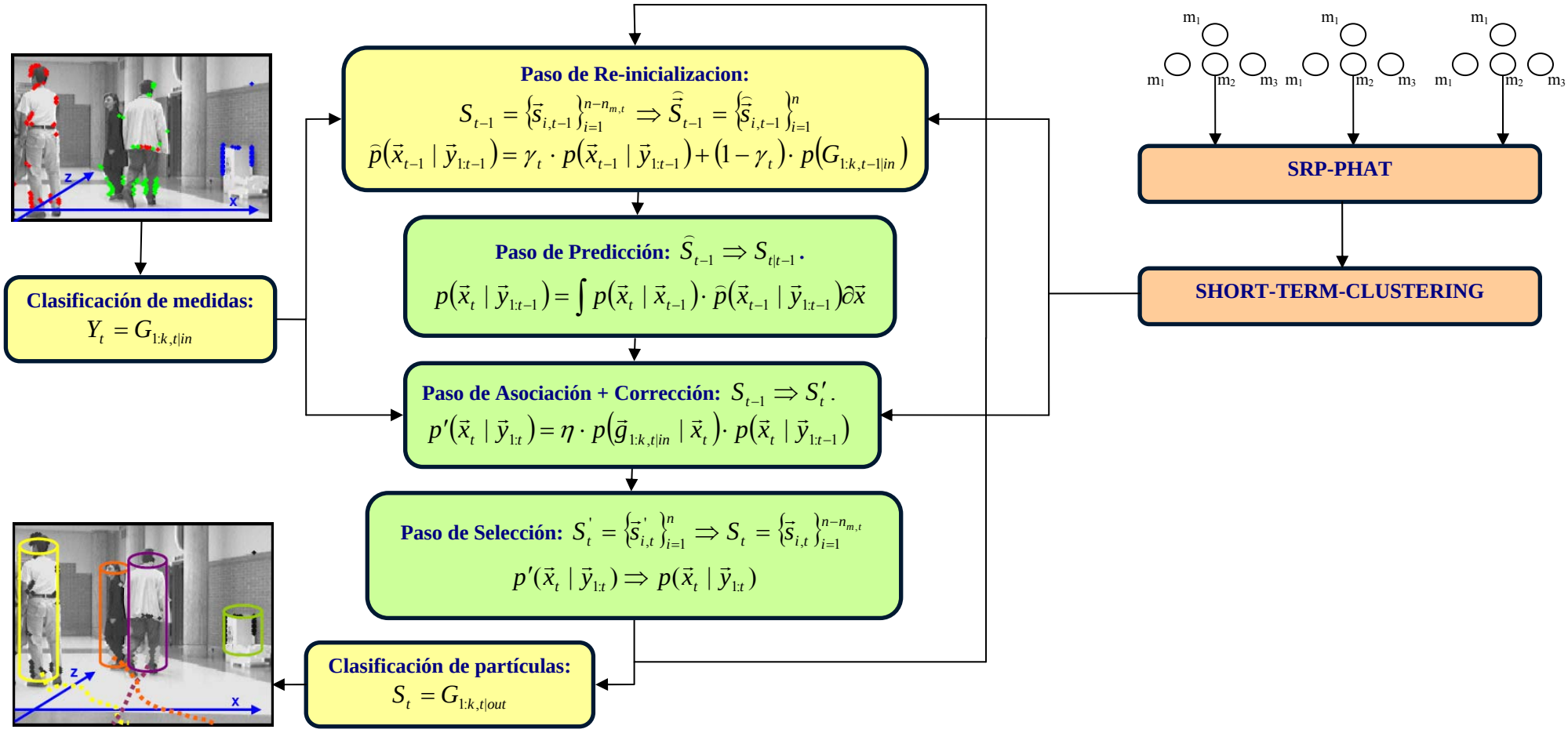


Figura 5.1 Posible implementación del algoritmo de fusión a partir de un XPF

ANEXO I: El Filtro de Partículas (PF)

A1.1. Introducción

La gran aceptación por parte de la comunidad científica de esta versión del filtro óptimo bayesiano tanto en la parte de audio, como en las de visión y fusión, hacen necesario dedicarle especial atención dentro del presente estado del arte debido a la más que probable utilización del mismo durante el desarrollo de la tesis que aquí comienza.

Por ello a lo largo de este anexo se expondrán la formalización matemática que permite la implementación de un PF básico, así como las distintas versiones y modificaciones presentadas por la comunidad científica para mejorar su fiabilidad, robustez y tiempos de cómputo.

Información adicional a la aquí presentada se puede encontrar en, [Van Der Merwe 2000], [Doucet 2001], [Arulampalam 2002] y [Marrón 2008] entre otros.

A1.2. Descripción del PF básico

El PF se define como un estimador probabilístico recursivo, fruto de una particularización del filtro de Bayes (3.4.1.3) en el que se discretiza la función de creencia a estimar $p(\vec{x}_t, \vec{y}_{1:t})$ gracias a un conjunto, que denominaremos set, de n partículas definidas como $S_t = \left\{ \vec{s}_{i,t} = \vec{x}_t^{(i)} / i = 1:n \right\}$. A cada una de estas partículas se le asocia un peso normalizado $\vec{w}_t = \left\{ \tilde{w}_t^{(i)} / i = 1:n \right\}$ dentro del set que caracteriza su probabilidad dentro de la PDF global. Con lo cual la creencia discreta, salida final del algoritmo de estimación, queda definida como:

$$p(\vec{x}_t^{(1:n)}, \vec{y}_{1:t}) \cong S_t = \left\{ \vec{x}_t^{(i)}, \tilde{w}_t^{(i)} \right\}_{i=1}^{n} \tag{A1.1}$$

Partiendo de la creencia discreta presentada anteriormente, la estimación final del algoritmo se puede obtener bien a partir de la envolvente de dicha PDF discreta, o a partir de algún estadístico de la misma (e.g. mediana, media), junto con la varianza, la cual informa sobre la fiabilidad del valor estimado.

La discretización de la función de creencia a través de un conjunto de muestras ponderadas $p(\vec{x}_t^{(1:n)}, \vec{y}_{1:t}) \cong S_t = \left\{\vec{x}_t^{(i)}, \tilde{w}_t^{(i)}\right\}_{i=1}^{n}$, se realiza a través del muestreo de Monte-Carlo estudiado a continuación.

A1.2.1. Muestreo de Monte-Carlo

El muestreo de Monte-Carlo tiene como objetivo aproximar la función de creencia continua mediante un conjunto de n muestras, a partir de la siguiente expresión:

$$p(\vec{x}_{0:t}, \vec{y}_{1:t}) \equiv \frac{1}{n}\sum_{i=1}^{n}\delta(\vec{x}_{0:t}^{(i)}) \tag{A1.2}$$

donde $S_t = \left\{\vec{s}_{i,t} = \vec{x}_t^{(i)} / i = 1:n\right\}$ representa un conjunto de muestras independientes e idénticamente distribuidas (i.i.d.) de la densidad de creencia continua, y δ es la función Delta de Dirac.

Desarrollando matemáticamente la idea anteriormente expuesta, se llega a la conclusión de que el muestreo de Monte-Carlo permite obtener una formulación discreta de la creencia $p(\vec{x}_t^{(1:n)}, \vec{y}_{1:t})$ a partir del set de partículas ponderadas $S_t = \left\{\vec{x}_t^{(i)}, \tilde{w}_t^{(i)}\right\}_{i=1}^{n}$ a través de la siguiente expresión:

$$p(\vec{x}_t^{(1:n)}, \vec{y}_{1:t}) = \sum_{i=1}^{n}\tilde{w}_t^{(i)} \cdot \delta(\vec{x}_t^{(i)}) \tag{A1.3}$$

donde el valor de los pesos $\vec{w}_t = \left[\tilde{w}_t^{(1:n)}\right]$ se obtiene a partir de la función de ponderación de las muestras $w(\vec{x}_{0:t})$, la cual se calcula de forma recursiva a partir de los modelos de observación $p(\vec{y}_t \mid \vec{x}_t)$ y estado $p(\vec{x}_t \mid \vec{x}_{t-1})$ y de la función de aproximación a la creencia $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ a partir de la siguiente expresión:

$$w(\vec{x}_{0:t}) = w(\vec{x}_{0:t-1}) \cdot \frac{p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{x}_{t-1})}{q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})} \tag{A1.4}$$

los cuales han sido previamente normalizados a través de:

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^{n} w_t^{(i)}} \tag{A1.5}$$

A1.2.2. El algoritmo SIS: versión básica del PF

Las ecuaciones presentadas en el apartado anterior permiten definir las etapas de *predicción* y *corrección*, típicas de cualquier estimador recursivo, para dar lugar al PF básico tal y como se muestra a continuación:

- *Predicción*: En esta etapa cada una de las partículas se propaga al siguiente instante temporal a través del modelo de estados $p(\vec{x}_t \mid \vec{x}_{t-1})$ para obtener la probabilidad a priori $p(\vec{x}_t \mid \vec{y}_{1:t-1})$, la cual se puede representar, según la teoría expuesta en apartados anteriores, como un conjunto de muestras i.i.d $S_{t|t-1} = \left\{\vec{s}_{i,t|t-1} = \vec{x}_{t|t-1}^{(i)} / i = 1:n\right\}$ manteniendo su peso asociado $S_{t|t-1} = \left\{\vec{x}_{t|t-1}^{(i)}, \tilde{w}_{t-1}\right\}_{i=1}^{n}$ a través de la expresión:

$$\vec{x}_{t|t-1}^{(i)} = p(\vec{x}_t, \vec{x}_{t-1}) \cdot \vec{x}_{t-1}^{(i)} = f(\vec{x}_{t-1}^{(i)}, \vec{u}_{t-1}, \vec{v}_{t-1}) \tag{A1.6}$$

La cual sustituye a las más genéricas (3.17) y (3.19), presentadas de nuevo aquí por claridad.

$$p(\vec{x}_t \mid \vec{y}_{1:t-1}) = \int p(\vec{x}_t \mid \vec{x}_{t-1}) \cdot p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1}) \partial \vec{x}$$

$$p(\vec{x}_t \mid \vec{y}_{1:t-1}) = \sum_{\forall \vec{x}} p(\vec{x}_t^{(i)} \mid \vec{x}_{t-1}^{(i)}) \cdot p(\vec{x}_t^{(i)} \mid \vec{y}_{1:t-1})$$

- *Corrección*: También denominada *ponderación de muestras* o *Importance Sampling*, parte de la versión muestreada de la PDF a priori $p(\vec{x}_t^{(1:n)} \mid \vec{y}_{1:t-1})$ obtenida en la etapa de predicción, sobre la cual se aplica el muestreo de Monte-Carlo (A1.3) para obtener la creencia o PDF a posteriori $p(\vec{x}_t^{(1:n)} \mid \vec{y}_{1:t})$. La expresión así obtenida sustituye a la más general (3.14) del filtro de Bayes, presentada de nuevo por claridad.

$$p(\vec{x}_t \mid \vec{y}_{1:t}) = \eta \cdot p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{y}_{1:t-1})$$

  El objetivo único de este paso de corrección es obtener un peso $\tilde{w}_t^{(i)}$ para cada partícula $\vec{x}_t^{(i)}$ que valore la probabilidad de acierto de cada una como representante del proceso de estimación y así poder completar la definición del

set de partículas como $S_t = \{\vec{x}_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^{n}$, lo cual se realiza a partir de la expresiones (A1.4) y (A1.5), no modificándose en este paso el valor de las partículas. El set $S_t = \{\vec{x}_t^{(i)}, \tilde{w}_t^{(i)}\}_{i=1}^{n}$ así definido se utiliza para describir a la creencia continua $p(\vec{x}_t \mid \vec{y}_{1:t})$.

El principal problema del PF así definido reside en la denominada *degeneración del set*, consistente en que con la evolución temporal de la ejecución recursiva del filtro, gran parte de las partículas acaban por tener un peso nulo $\tilde{w}_t^{(i)} = 0$, concentrándose el peso total sólo en unas pocas. Esta degeneración del set se traduce en una ineficiencia computacional del algoritmo, ya que las partículas de peso nulo no aportan información a la estimación.

Esta problemática hace que la comunidad científica se vuelque en la búsqueda de soluciones que lo resuelvan, por lo que se requiere la definición de un concepto que evalúe los resultados de cada propuesta. A este efecto se define el *número de partículas eficientes* (n_{eff}) como:

$$n_{eff} = \frac{n}{1 + \Omega^2\left(w_{opt}(\vec{x}_{0:t})\right)} \tag{A1.7}$$

donde $\Omega^2\left(w_{opt}(\vec{x}_{0:t})\right)$ representa la varianza de la función de ponderación óptima, obtenida mediante la densidad de aproximación de la creencia óptima definida como sigue:

$$w_{opt}(\vec{x}_{0:t}) = w_{opt}(\vec{x}_{0:t-1}) \cdot \frac{p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{x}_{t-1})}{q_{opt}(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})} \tag{A1.8}$$

Teóricamente se demuestra que el valor óptimo de la función de aproximación a la creencia $q_{opt}(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ es aquel que minimiza la varianza de los pesos $\vec{w}_t$, demostrándose que la función que cumple este requisito de manera óptima es $q_{opt}(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t}) = p(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$, la cual es justamente la incógnita en un PF. Para evitar esta problemática se define el *valor aproximado del número de partículas eficaces* como:

$$\hat{n}_{eff} = \frac{1}{\sum_{i=1}^{n} \left(\tilde{w}_t^{(i)}\right)^2} \tag{A1.9}$$

De lo expuesto anteriormente se deduce que la eficiencia del PF depende fuertemente de la función de aproximación a la creencia $q_{opt}(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ utilizada. Sin embargo una de las soluciones más recurrentes en la comunidad científica para evitar el problema de empobrecimiento del set consiste en añadir un tercer paso al PF conocido como *resampling,* el cual se trata en el siguiente apartado.

A1.2.3. El algoritmo SIR: PF básico + selección

El algoritmo SIR (*Squential Importance Resampling*), se obtiene como resultado de añadir al PF básico presentado en el apartado anterior un proceso de *selección* o *resampling*, en el cual las partículas con bajo peso asociado son eliminadas, mientras que las que tienen un peso alto, y por tanto representan las hipótesis más probables, son reproducidas para sustituir a las anteriores.

La formalización matemática de la explicación anterior parte del set $S_t' = \left\{\vec{x}_t^{(i)}, \tilde{w}_t^{(i)}\right\}_{i=1}^{n}$ obtenido a la salida del paso de corrección, del que se repiten n_i veces cada partícula $\vec{x}_t^{(i)}$, de modo que el número de partículas de salida que corresponden a cada partícula original este relacionado con su peso asociado $\tilde{w}_t^{(i)}$, a través de la siguiente expresión:

$$\mathrm{E}\{n_i\} = n \cdot \tilde{w}_t^{(i)} \quad \Rightarrow \quad \sum_{i=1}^{n} n_i = n \tag{A1.10}$$

A la salida de este del proceso así definido se obtiene un set de partículas $S_t = \left\{\vec{s}_{i,t} = \vec{x}_t^{(i)} \,/\, i = 1:n\right\}$, cada una de ellas con un peso asociado idéntico e igual a $\tilde{w}_t^{(1:n)} = 1/n$.

Por otro lado, la expresión (A1.10) en la que se especifica el número reproducciones n_i para cada partícula, no es única, por lo que aparecen en la comunidad científica distintas propuestas (basadas por ejemplo en el $\hat{n}_{eff}$ presentado en el apartado anterior) orientadas a diferentes objetivos como por ejemplo reducir la varianza o

disminuir la carga computacional. Sin embargo se demuestra ([Van Der Merwe 2000]), que el tipo de implementación del proceso de selección tiene poca incidencia en la eficiencia del estimador final.

Sin embargo, la solución del problema de *degeneración del set* a partir del mencionado paso de selección, produce la aparición de una nueva problemática conocida como *empobrecimiento del set* que consiste en que todas las partículas de set S_t terminan concentrándose en una misma hipótesis, es decir $n_{i1} = n$ para una única partícula $\vec{x}_t^{(i1)}$ mientras que $n_{i2} = 0$ para el resto de ellas $\left\{\vec{x}_t^{(i2)} / i2 = 1 : n - i1\right\}$.

Esta problemática hace que el PF pierda la diversidad que le hace versátil, flexible, robusto y por supuesto multimodal siendo especialmente relevante en caso de que el ruido asociado al modelo de estados $\vec{v}_t$ sea pequeño.

La solución pasa en la mayoría de los casos por modificar la función de aproximación a la creencia $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ o por introducir nuevos pasos en el PF. Una solución simple consiste en dispersar las partículas del set S_t por el espacio de estado a través de un modelo de ruido de estado, que realice la función de diversificación que desarrollaría, en caso de existir, el ruido real $\vec{v}_t$ asociado al modelo de estado del sistema bajo estudio.

Es interesante reseñar que muchas soluciones aportadas por la comunidad científica al problema de empobrecimiento reseñado, implican una alta carga computacional que los hacen incompatibles con la especificación de tiempo real aquí perseguida.

Debido a la importancia del algoritmo SIR expuesto y a objeto de facilitar la comprensión del mismo, en la Figura A1.5.2 presenta un diagrama funcional de los distintos pasos del mismo y la Figura A1.5.3 la representación unidimensional de la evolución de un set de partículas.

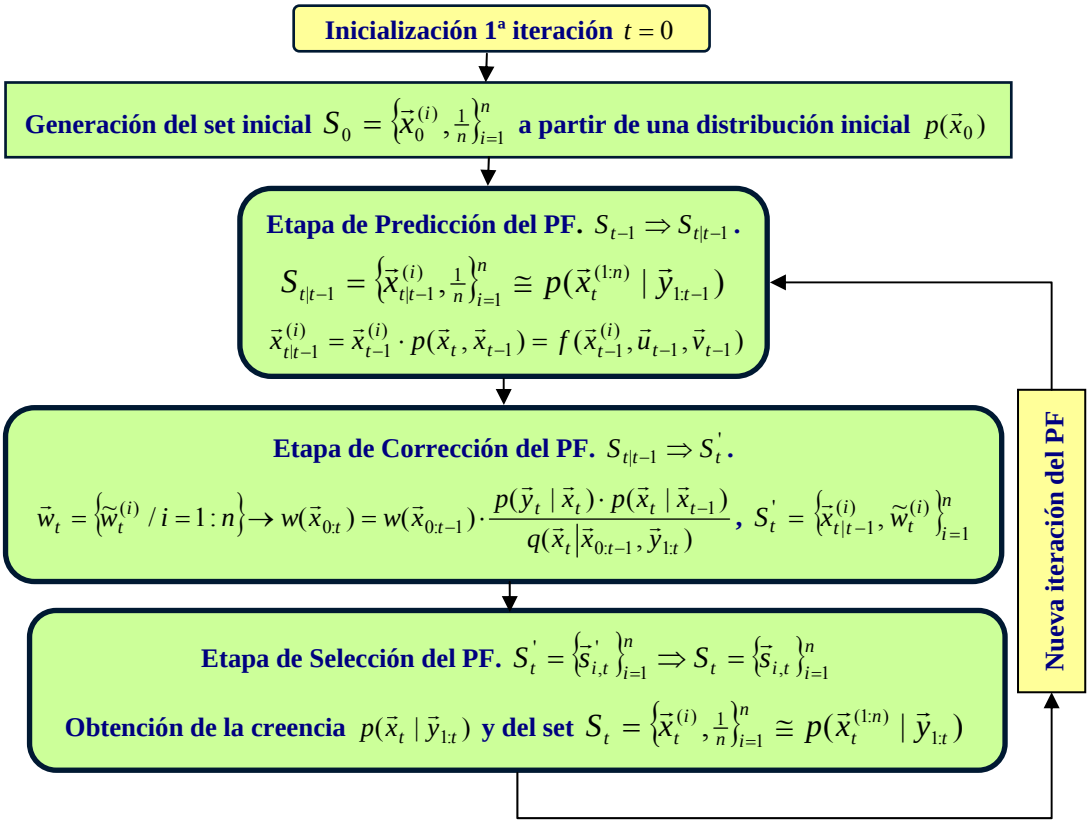


Figura A1.5.2 Diagrama funcional del PF básico (algoritmo SIR)

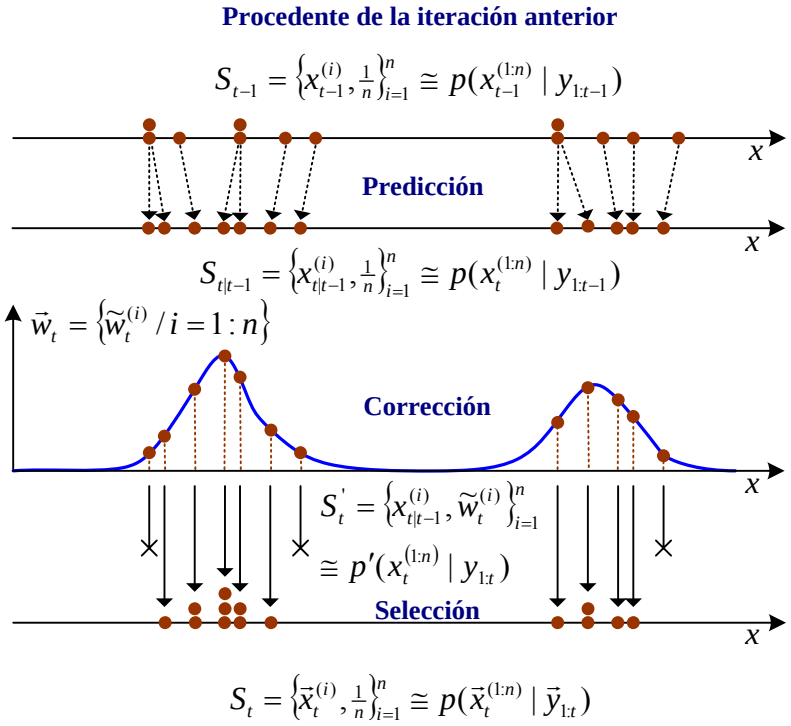


Figura A1.5.3 Representación unidimensional de la evolución de un set de partículas en las diferentes etapas del PF básico

El funcionamiento presentado de cada una de las etapas del PF básico presentado en las figuras anteriores se describe a continuación:

0. *Inicialización:* Ejecutado únicamente en la primera iteración del filtro, su objetivo es obtener una versión muestreada de la probabilidad a priori del vector de estados a estimar $p(\vec{x}_0^{(1:n)})$. Las opciones más típicas teniendo en cuenta el desconocimiento del valor inicial son:

   - Utilizar una distribución aleatoria y uniforme a lo largo de todo el espacio de estados (opción elegida en la Figura A1.5.2), según la siguiente expresión:

$$S_0 = \left\{\vec{x}_0^{(i)}, \tfrac{1}{n}\right\}_{i=1}^{n} \tag{A1.11}$$

   - Elegir como probabilidad a priori una distribución proporcional a la función de verosimilitud del sistema bajo estudio, con el objetivo de que el PF converja más rápidamente hacia el valor real de la creencia, y que por lo tanto cumple:

$$p(\vec{x}_0) \propto p(\vec{y}_0 \mid \vec{x}_0) \tag{A1.12}$$

1. *Predicción:* A la entrada de este proceso se tienen un conjunto de muestras ponderadas por igual e i.i.d. $S_{t-1} = \left\{\vec{x}_{t-1}^{(i)}, \tilde{w}_{t-1}^{(i)}\right\}_{i=1}^{n} = \left\{\vec{x}_{t-1}^{(i)}, \tfrac{1}{n}\right\}_{i=1}^{n}$, (bien procedentes del paso de inicialización, o de la ejecución recursiva del PF), que representan el valor de la creencia estimada por el PF en el instante $t-1$: $p(\vec{x}_{t-1} \mid \vec{y}_{1:t-1})$.

   En el proceso de predicción, estas partículas son propagadas mediante el modelo de actuación o de estados del sistema $p(\vec{x}_t \mid \vec{x}_{t-1})$, obteniendo a partir de (A1.6) un nuevo set de partículas $S_{t|t-1} = \left\{\vec{s}_{i,t|t-1} = \vec{x}_{t|t-1}^{(i)} / i = 1:n\right\}$ con pesos idénticos $\vec{w}_{t-1} = \left\{w_{t-1}^{(i)} = \tfrac{1}{n} \middle/ i = 1:n\right\}$ que caracteriza la estimación a priori de la creencia $p(\vec{x}_t \mid \vec{y}_{1:t-1})$. El peso de las partículas no se modifica en esta etapa, quedando por tanto el set representado como $S_{t|t-1} = \left\{\vec{x}_{t|t-1}^{(i)}, \tfrac{1}{n}\right\}_{i=1}^{n}$.

2. *Corrección (Importante Sampling):* En este paso se recalcula el peso de las partículas $\vec{w}_t = \{w_t^{(i)} / i = 1:n\}$ mediante la función de aproximación a la creencia $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ y las PDFs características de los modelos de actuación $p(\vec{x}_t \mid \vec{x}_{t-1})$ y de observación $p(\vec{y}_t \mid \vec{x}_t)$, gracias a la expresión (A1.8), que permite calcular la función de ponderación de muestras $w(\vec{x}_{0:t})$. El set de partículas a la salida de esta etapa se caracteriza como $S_t^{'} = \{\vec{x}_{t|t-1}^{(i)}, \widetilde{w}_t^{(i)}\}_{i=1}^{n}$, la cual representa a la función de creencia $p'(\vec{x}_t \mid \vec{y}_{1:t})$ (donde se utiliza "'" por tratarse de la salida de la etapa de corrección y no la global del PF).
3. *Selección:* Empleando el esquema de resampling elegido y el vector de pesos calculado en la etapa anterior $\vec{w}_t = \{w_t^{(i)} / i = 1:n\}$, se obtiene el nuevo conjunto de partículas $S_t = \{\vec{x}_t^{(i)}, \frac{1}{n}\}_{i=1}^{n}$ y que constituye la salida discreta proporcionada por el PF como estimación final de la PDF a posteriori $p(\vec{x}_t \mid \vec{y}_{1:t})$ para el instante t, tal y como expresa la ecuación (A1.3).

A1.2.4. El algoritmo Bootstrap: Versión SIR más extendida

Este algoritmo, presentado por primera vez en [Gordon 1993] (como parte del algoritmo SIR presentado en el apartado anterior) y utilizado en trabajos tan importantes y representativos como el algoritmo *Condensation* de [Isard 1998a], constituye uno de los más utilizados dentro de la comunidad científica. Este hecho se debe principalmente a dos factores: Su sencillez y su bajo tiempo de cómputo (lo cual le hace especialmente apto en aplicaciones de tiempo real).

Su principal aportación consiste en usar como función de aproximación a la creencia $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$ el modelo de actuación del sistema $p(\vec{x}_t \mid \vec{x}_{t-1})$, con lo que el proceso para calcular el peso asociado a las partículas en la etapa de corrección se simplifica, obteniéndose:

$$w(\vec{x}_{0:t}) = w(\vec{x}_{0:t-1}) \cdot \frac{p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{x}_{t-1})}{q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})} = w(\vec{x}_{0:t-1}) \cdot \frac{p(\vec{y}_t \mid \vec{x}_t) \cdot p(\vec{x}_t \mid \vec{x}_{t-1})}{p(\vec{x}_t \mid \vec{x}_{t-1})} \qquad \text{(A1.13)}$$
$$= w(\vec{x}_{0:t-1}) \cdot p(\vec{y}_t \mid \vec{x}_t)$$

La principal problemática asociada a la elección anterior reside en que se elimina de la función de pesos cualquier información relativa al modelo de actuación implicado en el proceso de estimación. En casos en los que el modelo de actuación $p(\vec{x}_t \mid \vec{x}_{t-1})$ no es bueno, o la función de verosimilitud $p(\vec{y}_t|\vec{x}_t)$ es estrecha, puede suceder que la mayoría de las partículas obtenidas a la salida del proceso de predicción (caracterizadas por la PFD a priori $p(\vec{x}_t|\vec{y}_{1:t-1})$ - marcada en marrón sobre el paso de corrección -) se encuentren en zonas en las que la PDF de verosimilitud $p(\vec{y}_t|\vec{x}_t)$ que las pondera tiene un valor bajo, con lo que el peso asociado a dichas partículas será también bajo. Este concepto se representa en la siguiente figura extraída de [Van Der Merwe 2000].

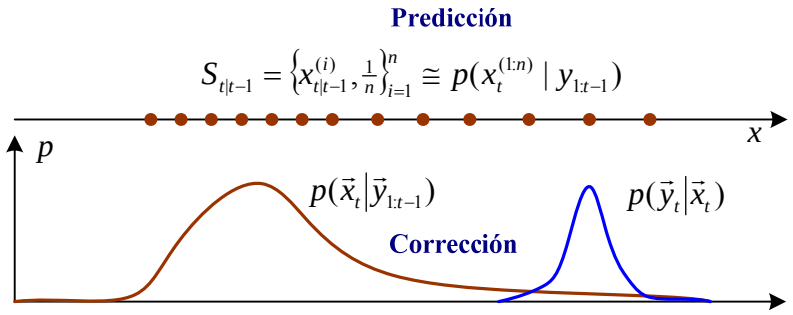


Figura A1.5.4 Representación unidimensional y unimodal del problema de elegir $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t}) = p(\vec{x}_t \mid \vec{x}_{t-1})$ característica del algoritmo *Bootstrap*

El efecto presentado en la Figura A1.5.4 conduce a tres problemas principales:

- Rápida divergencia del PF debido al empobrecimiento del set de partículas S_t' a la salida del paso de corrección.
- La estimación se hace muy sensible a ruidos en las medidas $\vec{o}_t$, lo cual lo vuelve fácilmente inestable.
- Dificultad para adaptarse a movimientos rápidos del target a seguir (especialmente importante en tareas de seguimiento como la aquí abordada).

La principal conclusión que se debe extraer de este apartado es la gran importancia de la función de aproximación a la creencia $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$, en la estimación final proporcionada por el PF.

Con el objetivo de mejorar la eficiencia del estimador y/o eliminar el problema de empobrecimiento del set del algoritmo *Bootstrap* anteriormente presentado, la comunidad científica propone diversas soluciones, muchas de ellas aplicadas en trabajos presentados en apartados anteriores. Estas modificaciones se pueden agrupar en dos ramas principales, tal y como se recoge en la Tabla A1.1. La explicación detallada de cada uno de estos métodos se considera en principio fuera del ámbito del presente estado del arte.

Mejoras al algoritmo Bootstrap	
Basadas en El proceso de selección	**Basadas en La función de aproximación a la creencia** $q(\vec{x}_t \mid \vec{x}_{0:t-1}, \vec{y}_{1:t})$
Re-muestreo Multinomial	Edición previa y rechazo selectivo
Re-muestreo Residual	Markov Chain Monte-Carlo Step (MCMC)
Tree Based Branching	Muestreo segmentado
Re-muestreo Estratificado	Auxiliary Particle Filter (APF)
Método de mínima varianza	Extended Kalman Particle Filter (EKPF)
	Importante Sampling Condensation (ICondensation)

Tabla A1.1 Mejoras al algoritmo *Bootstrap*

Referencias

AUDIO

[Abad 2007]

Abad, F.J. ***A Multi-Microphone approach to speech processing in a smart-room environment***. Phd thesis, Universidad Politécnica de Cataluña. Febrero 2007.

[Alonso 2004]

Alonso, D. ***Sistema de posicionamiento local (LPS): Estudio de un sistema de balizamiento por ultrasonidos aplicando técnicas multimodo***. Master thesis, Escuela Politécnica Superior, Universidad de Alcalá. Año 2004.

[Amida 2007]

Augmented Multi-party Interaction with Distance Access. ***Localization and Tracking of Multiple Interlocutors with Multiple Sensors***. Enero 2007.

[Benesty 2000]

J. Benesty. ***Adaptative eigenvalue decomposition algorithm for passive acoustic source localization.*** Journal of the Acoustic Society of America. Enero 2000, Volumen 107, Page(s):384-391

[Brandstein 1997]

Brandstein, M.S.; Adcock, J.E.; Silverman, H.F. ***A closed-form location estimator for use with room environment microphone arrays***. IEEE Transactions on Speech and Audio Processing. Enero 1997, Volumen 5, Page(s):45-50.

[Brandstein 2001]

M. Brandstein and D. Ward. ***Microphone Arrays.***
Digital Signal Processing. Springer, 2001.

[Brutti 2005]

Brutti, A. ; Omologo, M.; Svaizer, P. ***Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays***. INTERSPEECH 2005. Septiembre 2005, Page(s):2337-2340.

[Brutti 2008]

Brutti, A. ; Omologo, M.; Svaizer, P. ***Localization of multiple speakers based on a two step acoustic map analysis***. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008). Abril 2008, Page(s):4349-4352.

[Buchner 2004]

H. Buchner, R. Aichner y W. Kellerman. ***Trinicon: A versatile framework for multichannel blind signal processing***. IEEE International Conference on Acoustics, Speech and Signal Processing. Mayo 2004, Volumen 3, Page(s):889-892.

[Buchner 2005]

H. Buchner, R. Aichner y W. Kellerman. ***Simultaneous localization of multiple sound sources using blind adaptive mimo filtering***. IEEE International Conference on Acoustics, Speech and Signal Processing. Marzo 2005, Volumen 3, Page(s):18-23.

[Castro 2007]

Carlos Castro García. ***Speaker Localization Techniques in Reverberant Acoustic Environments.*** Master thesis, Royal Institute of Technology (KTH), Stockholm 2007.

[Chou 1995]

Chou, T. ***Frequency-independent beamformer with low response error.*** IEEE International Conference on Acoustics, Speech, and Signal Processing. Mayo 1995, Volumen 5, Page(s):2295-2298.

[De Moor 1993]

De Moor, B. ***The singular value decomposition and long and short spaces of noisy matrices***. IEEE Transactions on Signals Processing. Septiembre 1993, Volumen 41, Page(s):2826-2838.

[Dibiase 2000]

Joseph Hector Dibiase. ***A High-Accuracy, Low-Latency Technique for Talker Localization in reverberant Environments Using Microphone Arrays***. Phd thesis Brown University, 2000.

[Duraiswami 2001]

Duraiswami, N; Zotkin, D. ***Active speech source localization by a dual coarse-to-fine search.*** IEEE Transaction on Acoustics, Speech, and Signal Processing. Mayo 2001, Volumen 5, Page(s):3309-3312.

[Dvorkind 2005]

Dvorkind, T.V.; Gannot, S. **Speaker localization using the unscented Kalman filter**. Proceedings on HSCMA Workshop, Marzo 2005.

[Fischer 1995]

Fischer, S.; Simmer, K.U. ***An adaptive microphone array for hands-free communication.*** An International Workshop on Acoustic Echo and Noise Control. Año 1993, Page(s):44-47.

[Friedlander 1993]

Friedlander, B.; Weiss, A.J. ***Direction finding for wide-band signals using an interpolated array.*** IEEE Transactions on Signal Processing. Abril 1993, Volumen 41, Page(s):1618-1634.

[Frost 1972]

Frost, O.L. ***An algorithm for linearly constrained adaptive array processing.***
IEEE Issue 8. Agosto 1972, Volumen 60, Page(s):926-935.

[Gatica-Perez 2004]

Lathoud, G.; Odobez J.M.; Gatica-Perez, D. ***AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking***. IDIAP Research Institute, Switzerland, Junio 2004.

[Gonzalez-Rodriguez 2000]

Gonzalez-Rogriguez, J.;Sanchez-Bote, J.L., Ortega-García, J. Speech dereverberation and noise reduction with a combined microphone array approach. IEEE International Conference on Acoustics, Speech, and Signal Processing. Junio 2000, Volumen 3, Page(s):1037-1040.

[Griffiths y Jim 1982]

L. Griffiths, C. Jim. ***An alternative approach to linearly constrained adaptative beamforming***.
IEEE Trans. On Antennas and Propagation. Junio 1982, Volumen 30(1), Page(s):27-32.

[Hughes y otros 1999]

Hughes, T.B.; Hong-Seok Kim; DiBiase, J.H. Silverman, H.F. ***Performance of an HMM speech recognizer using a real-time tracking microphone array as input***. IEEE Transactions on Speech and Audio Processing. Mayo 1999, Volumen 7, Page(s):246-349.

[Hung 1988]

Hung, H.; Kaveh, M. ***Focussing matrices for coherent signal-subspace processing.*** IEEE Trans. on Acoustics, Speech, and Signal Processing. Agosto 1988, Volumen 36, Page(s):1272-1281.

[Julier 1997]

Julier, S.J.; Uhlmann, J.K. ***A new extension of the Kalman filter to nonlinear systems***. Proceeding on Aerospace/Defence Sensing, Simulation and Controls (AeroSense), 1997.

[Kalman 1960]

Kalman, R.E. ***A new approach to linear filtering and prediction problems.*** Transactions of the ASME-Journal of Basic Engineering. Marzo 1960, Volumen 82, Page(s):35-45.

[Knapp 1976]

Knapp, C.; Carter, G. ***The generalized correlation method for estimation of time delay***.
IEEE Acoustics, Speech, and Signal Processing. Agosto 1976, Volumen 24, Page(s):320-327.

[Kumaresan 1983]

Kumaresan, R; Tufts, D.W. ***Estimating the Angles of Arrival of Multiple Plane Waves***.
IEEE Transactions on Aerospace and Electronics Systems. Junio 1983, Volumen AES-19,
Page(s):134-139.

[Larocque 2002]

Larocque, R.J.; Reilly, J.P.; Ng, W. ***Particle filters for tracking an unknown number of sources.***
IEEE Trans. On Signal Processing. Diciembre 2002, Volumen 50, Page(s):2926-2937.

[Lathoud 2005]

Lathoud, G.; Magimai-Doss, M. ***A sector-based, frequency-domain approach to detection and localization of multiple speakers***. IEEE International Conference on Acoustics, Speech and Signal Processing. Marzo 2005, Volumen 3, Page(s):265-268.

[Lathoud y McCowan 2004]

Lathoud, G.; McCowan, I.A.; Odobez, J.M. ***Unsupervised location-based segmentation of multi-party speech***. Proceedings of NIST-ICASSP Meeting Recognition Workshop. Abril 2004.

[Lathoud y Odobez 2004]

Lathoud, G.; Odobez, J.M.; McCowan, I.A. ***Short-term spatio-temporal clustering of sporadic and concurrent events***. Proceedings of NIST-ICASSP Meeting Recognition Workshop. Abril 2004.

[Lathoud y Odobez 2007]

Lathoud, G.; Odobez, J.M. ***Short-term spatio-temporal clustering applied to Multiple Moving Speakers***. IEEE Transaction on Audio, Speech and Language processing. Julio 2007, Volumen 15, Page(s):1696-1710.

[Lehmann 2004]

Lehmann, E. ***Particle Filtering Methods for Acoustic Source Localization and Tracking***. Phd thesis, Australian National University. Julio 2004.

[Matsumoto 2006]

Mitsuharu Matsumoto. ***A Study on Miniaturizing Microphone Array Utilizing Aggregated Microphones***. Major in Pure and Applied Physics. School of Science and Engineering Waseda University. Marzo 2006.

[McCowan 2001]

Iain A. McCowan. ***Robust Speech Recognition Using Microphone Arrays.***
Phd thesis, Queensland University of Technology, Australia, 2001.

[Meuse 1994]

P.C. Meuse and H.F. Silverman. ***Characterization of Talker Radiation Pattern Using a Microphone Array***. IEEE International Conference on Acoustics, Speech and Signal Processing. Abril 1994, Volumen 2, Page(s):257-260.

[Mostefa y otros 2006]

Mostefa, D.; Garcia M.J.; y otros. ***Clear Evaluation Plan***. Projecto CHIL (Computers in the Human Interaction Loop). Febrero 2006.

[Mungamuru 2004]

Mungamuru, B.; Aarabi, P.; ***Enhanced sound localization.*** IEEE Transactions Systems, Man, and Cybernetics. Issue 3 Junio 2004, Volumen 34, Page(s):1526-1540.

[Muñoz 2006]

Eva 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 thesis, Technical University of Madrid, 2006.

[Nadeu 2006]

Nadeu, C.; Segura, C.; Abad, D.; Hernando, J. ***Multimodal person tracking in a smart-room environment.*** IV Conference on Speech Technology. Zaragoza Noviembre2006, Page(s):271-276.

[Nordholm 1999]

Nordholm, S.; Claesson, I. Dahl, M.; ***Adaptive microphone array employing calibration signals: an analytical evaluation.*** IEEE Transactions on Speech and Audio Processing. Issue 3 Mayo 1999, Volumen 7, Page(s):241-252.

[Omologo 1994]

Omologo, M.; Svaizer, P. ***Acoustic event localization using a crosspower-spectrum phase based technique***. IEEE International Conference on Acoustics, Speech and Signal Processing. Abril 1994, Volumen 2, Page(s):273-276.

[Omologo 2003]

Omologo, M.; Arman, L.; Matassoni, M.; Sveizer, Piergiorgio. ***Use of a csp-based voice activity detector for distant-talking asr.*** Proceedings of the European Conference on Speech Communication and Technology (Eurospeech). Año 2003, Page(s):501-504.

[Rodriguez 2006]

Rodriguez Lorenzo, Vanessa. ***Diseño, implementación y evaluación de herramientas de simulación de entornos acústicos reverberates: aplicación a sistemas de reconocimiento automático de habla***. Proyecto fin de carrera, Escuela Técnica Superior de Ingenieros de Telecomunicación, 2006.

[Roy 1989]

Roy, R.; Dailath, T. ***ESPRIT- Estimation of signal parameters via rotational invariance techniques***. IEEE Transactions on Acoustics, Speech and Signal Processing. Julio 1989, Volumen 37, Page(s): 984-995.

[Royo 2007]

Royo Fernández, Francisco José. ***Design, Implementation and Evaluation of Techniques for Speech Signal Improvement in Reverberant Environments: Application in Automatic Speech Recognition System***. Master thesis, Technical University of Madrid, 2007.

[Rui 2005]

Rui, Y.; Florencio,D.; Lam, W. Jinyan Su. ***Sound source localization for circular arrays of directional microphones***. IEEE International Conference on Acoustics, Speech and Signal Processing. Marzo 2005, Volumen 3, Page(s):93-96.

[Sachar 2004]

Sachar, J.M.; Silverman, H.F. ***A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array***. IEEE International Conference on Acoustics, Speech and Signal Processing. Mayo 2004, Volumen 4, Page(s):65-68.

[Schmidt 1986]

Schmidt, R. ***Multiple emitter location and signal parameter estimation***.
IEEE Transactions on Antennas and Propagation. Marzo 1986, Volumen 34, Page(s):276-280.

[Seltzer 2003]

Michael L. Seltzer. ***Microphone Array Processing for Robust Speech Recognition***. Phd thesis Carnegie Mellon University, Pittsburgh, Julio 2003.

[Segura 2007]

Segura, C.; Canton-Ferrer ,C.; Abad, A.; Casas, J.R.; Hernando, J. ***Multimodal Head Orientation Towards Attention Tracking in Smartrooms***. IEEE International Conference on Acoustics, Speech and Signal Processing. Abril 2007, Volumen 2, Page(s):681-684.

[Sturim 2007]

Christian Müller (Ed.) ***Speaker Classification I:*** Fundamentals, Features & Methods. Springer, 2007. D.E. Sturim, W.M.Campbell, and D.A. Reynolds ***Classification Methods for Speaker Recognition*** Page(s):278,297.

[Svaizer 1997]

Svaizer, P.; Matassoni, M.; Omologo, M. ***Acoustic source localization in a three-dimensional space using crosspower spectrum phase.*** IEEE International Conference on Acoustics, Speech and Signal Processing. Abril 1997, Volumen 1, Page(s):231-234.

[Van Compernolle 1990]

Van Compernolle, D. ***Switching adaptativa filters for enhancing noisy and reverberant speech from microphone array recordings.*** IEEE International Conference on Acoustics, Speech and Signal Processing. Abril 1990, Volumen 2, Page(s):833-836.

[Varma 2002]

Krishnaraj Varma. ***Time-Delay-Estimate Based Direction-of-Arrival Estimation for Speech in Reverberant Environments***. Phd thesis, Virginia Polytechnic Institute, 2002.

[Vermaak 2001b]

Vermaak, J.; Blake, A.; ***Nonlinear filtering for speaker tracking in noisy and reverberant environments.*** IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP01). Mayo 2001, Volumen 5, Page(s):3021-3024.

[Viberg 1991]

Viberg, M.; Ottersten, B.; Kailath, T. ***Detection and estimation in sensor arrays using weighted subspace fitting***. IEEE Transactions on Signal Processing. Noviembre 1991, Volumen 39, Page(s):2436-2449.

[Wang 1985]

Wang, H.; Kaveh, M. ***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. Agosto 1985, Volumen 33, Page(s): 823-831.

[Wang y Chu 1997]

H. Wang y P. Chu. ***Voice source localization for automatic camera pointing system in videoconferencing.*** IEEE International Conference on Acoustics, Speech, and Signal Processing. Abril 1997, Volumen 1, Page(s): 187-190.

[Ward y Lehmann 2003]

Ward, D.; Lehmann, E.; Williamson, R. ***Particle filtering algorithms for tracking an acoustic source in a reverberant environment***. IEEE Transaction on Speech and Audio Processing. Noviembre 2003, Volumen 12, Page(s):826-836.

[Wax 1983]

Wax, M.; Dailath, T. ***Optimum localization of multiple sources by passive arrays.*** IEEE Transaction on Acoustics, Speech, and Signal Processing. Octubre 1983, Volumen 31, Page(s):1210-1217.

[Welch 2006]

Welch, G.; Bishop, G. ***An introduction to the Kalman Filter***. TR 95-041. Julio 2006. Department of Computer Science University of North Carolina at Chapel Hill.

[Zelinski 1988]

Zelinski, R.; ***A microphone array with adaptive post-filtering for noise reduction in reverberant rooms.*** IEEE International Conference on Acoustics, Speech, and Signal Processing. Abril 1988, Volumen 5, Page(s):2578-2581.

[Ziomek 1995]

Ziomek L.J. ***Fundamentals of Acoustic Field Theory and Space-Time Signal Processing.*** Boca Raton, CRC Press 1995. ISBN: 0849394554.

[Zotkin 2001]

Zotkin, D.; Duraiswami, R.; Davis, L. ***Multimodal 3-D tracking and event detection via the particle filter***. IEEE International Conference on Computer Vision, Workshop on Detection and Recognition of Events in Video (ICCV-EVENT), Julio 2001, Page(s):20-27.

[Zotkin 2004]

Zotkin, D.; Duraiswami, R. ***Accelerated speech source localization via a hierarchical search of steered response power.*** IEEE Transaction on Acoustics, Speech, and Signal Processing. Septiembre 2004, Volumen 12, Page(s):499-508.

VISIÓN

[Akita 2007]

Akita, T.; Inagaki, S.; Suzuki, T.; Hyakawa, S.; Tsuchida, N. ***Analysis of vehicle following behaviour of human driver based on hybrid dynamical system model***. IEEE International Conference on Control Applications. Octubre 2007, Page(s):1233-1238.

[Arulampalam 2002]

Arulampalam, M.S.; Maskell, S.; Gordon, N.; Clapp, T. ***A tutorial on particle filters for online nonlinear non-gaussian Bayesian tracking***. IEEE Transactions on Signal Processing. Febrero 2002, Volumen 50, Page(s):174-188.

[Ba 2004]

Ba, S.O.; Odobez, J.M. ***A probabilistic framework for joint head tracking and pose estimation***. Proceedings of the 17th Int.Conf. on Pattern Recognition. Agosto 2004, Volumen 4, Page(s):264-267.

[Ba 2005]

Ba, S.O.; Odobez, J.M. ***A Rao-Blackwellized Mixed State Particle Fitler for Head Pose Tracking in Meetings***. ICMI Workshop on Multimodal Multiparty Meetings. Año 2005, Page(s):9-16.

[Ba 2008]

Ba, S.O.; Odobez, J.M. ***Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues***. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008). Marzo 2008, Volumen 4, Page(s):2221-2224.

[Bar-Shalom 1995]

Bar-Shalom, Y.; Li, X. ***Multitarget-Multisensor Tracking: Principles and techniques.***
YBS Publishing, Storrs, CT (1995).

[Bradner 2003]

Bradner, M.; Ribo, M.; Pinz, A. ***State of the art of vision-based self-localisation***. IEEE 1st International Workshop on Robotic Sensing. Junio 2003, Page(s):6

[Branson 2005]

Branson, K; Belongie, S. ***Tracking multiple mouse contours (without too many samples)***. Proceedings of the Fourth IEEE Conference on Computer Vision and Pattern Recognition (CVPR05), ISBN: 0-7695-2372-2, San Diego, Junio 2005, Volumen 1, Page(s):1039-1046.

[Chen 2002]

Chen, Y.; Huang, T.S.; Rui, Y. ***Mode-based multi-hypothesis head tracking using parametric contours***. Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR02). ISBN: 0-7695-1602-5, Washington, Mayo 2002.

[Coraluppi 2004]

Coraluppi, S.; Carthel, C.; ***Multi-hypothesis sonar tracking***. Proc. of the seventh Int. Conference on Information Fusion (FUSION04). ISBN:91-7056-114-X, Junio 2004, Volumen 1, Page(s): 33-40.

[Cucchiara 2005]

Cucchiara, R.; Grana, C.; Prati, A.; Vezzani, R. ***Computer vision system for in-house video surveillance***. IEEE Proceedings on Vision, Image and Signal Processing. Abril 2005, Volumen 152, Page(s):242-249.

[Doucet 2001]

Doucet, A.; De Freitas, N.; Gordon, N. ***Sequential MonteCarlo methods in practice***. Springer-Verlag, ISBN: 0-389-95146-6, New York, Año 2001

[Duncan y Ayache 2000]

Duncan, J.; Ayache, N. ***Medical Image Analysis: Progress over Two Decades and Challenges Ahead***. IEEE Transactions on Pattern Analysis and Machine Intelligence. Enero 2000, Volumen 22, Page(s):85-106.

[Everingham 2005]

Everingham, M.; Zisserman, A. ***Automated detection and identification of persons in video using a coarse 3D head model and multiple texture maps***. IEEE Proceedings on Vision, Image and Signal Processing. Diciembre 2005, Volumen 152, Page(s):902-910.

[FBK Lanz]

Fundación Bruno Kessler (FBK): http://tev.fbk.eu/smartrack/

[Ferrari 2001]

Ferrari, V; Tuytellars, T.; Van Gool, L. ***Markerless augmented reality with a real-time affine region tracker***. IEEE International Symposium on Augmented Reality. Octubre 2001, Page(s):87-96.

[Fortmann 1983]

Fortmann, T.; Bar-Shalom, Y.; Scheffe, M. ***Sonar tracking of multiple targets using joint probabilistic data association***. IEEE Journal of Oceanic Eng. Julio 1983, Vol 8, Page(s):173-184.

[Gordon 1993]

Gordon, N.J.; Salmond, D.J.; Smith, A.F.M. ***Novel approach to nonlinear/nongaussian bayesian state estimation.*** IEE Proceedings-F in Radar and Signal Processing, Abril 1993, Volumen 140, Page(s):107-113.

[Hu 2004]

Chen, Y; Huang, T.S.; Feris, R.; Turk, M. ***Manifold Based Analysis of Facial Expression***. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04). ISBN:1063-6919/04, Año 2004.

[Hue 2001]

Hue, C.; Le Cadre, P.; Pérez, P. ***A particle filter to track multiple objects***. IEEE Workshop on Multi-Object Tracking. Julio 2001, Page(s):61-68.

[Hue 2002]

Hue, C; le Cadre, J.P.; Pérez, P. ***Tracking multiple objects with particle filtering***. IEEE Transactions on Aerospace and Electronics Systems, Julio 2002, Volumen 38, Page(s):791-812.

[Hui-Fuang 2004]

Hui-Fuang Ng. ***Automatic thresholding for defect detection***. Proceedings of Third International Conference on Image and Graphics. Diciembre 2004, Page(s):532-535.

[Isard 1996]

Isard, M; Blake, A. ***Contour tracking by stochastic propagation of conditional density***. Proceeding of European Conference on Computer Vision. Año 1996, Page(s):343-356.

[Isard 1998a]

Isard, M; Blake, A. ***Condensation – Conditional Density Propagation for Visual Tracking***. International Journal of Computer Vision. Junio 1998, Page(s):5-28.

[Isard 1998b]

Isard, M.; Blake, A. ***ICondensation: Unifying low-level and high-level tracking in a stochastic framework***. Proceedings of the Fifth European Conference on Computer Vision (ECCV98). Año 1998, Volumen 1, Page(s):893-908.

[Isard 2001]

Isard, M; MacCormick, J. ***BraMBLe: A Bayesian multiple-blob tracker***. Proceedings of the Eighth IEEE Int. Conf. on Computer Vision (ICCV01), Vancouver, Julio 2001, Volumen 2, Page(s):34-41.

[Khan 2003]

Khan, Z.; Balch, T.; Delaert, F. ***Efficient particle filter-based tracking of multiple interacting***. Proceedings of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'03), Las Vegas, Octubre 2003, Page(s):980-986.

[Khan 2004]

Khan, Z.; Balch, T.; Delaert, F. ***An MCMC-based particle filter for tracking multiple interacting targets***. Proc. of the Eur. Conf. Comput. Vision (ECCV), Praga, Mayo 2004, Page(s):279-290.

[Koller-Meier 2001]

Koller-Meier, E.B.; Ade, F. ***Tracking multiple objects using a condensation algorithm.*** Journal of Robotics and Autonomous Systems, Febrero 2001, Volumen 34, Page(s):93-105.

[Lanz 2005]

Lanz, O.; Manduchi, R. ***Hybrid joint-separable multibody tracking***. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Junio 2005, Volumen 1, Page(s):413-420.

[Lanz 2006]

Lanz, O. ***Aproximate Bayesian Multibody Tracking***. IEEE Transactions on Pattern Analysis and Machine Intelligence, Septiembre 2006, Volumen 28, Parte I, Page(s):1436-1449.

[Lanz 2006b]

Lanz, O; Brunelli, R. ***Dynamic Head Location and Pose from Video***. IEEE International Conference on Multisensor Fusion and Integration from Intelligent Systems, Septiembre 2006, Page(s):47-52.

[Lu 2000]

Lu, C.P; Hager, G.D.; Mjolsness, E. ***Fast and globally convergent pose estimation from video images***. IEEE Transactions on Pattern Analysis and Machine Intelligence. Junio 2000, Volumen 22, Page(s):610-622.

[MacCormick 1999]

MacCormick, J.; Isard, M.; ***Partitioned sampling, articulated objects, and interface-quality hand tracking.*** Proceedings of the Sixth European Conference on Computer Vision (eccv00), Junio 2000, Volumen 2, Page(s):3-19.

[MacCormick 2000]

Hu, C.; Chang, Y. ***Partitioned Sampling, Articulated Objects, and Interface-Quality Hand Tracking***. Computer Vision – ECCV 2000, Ed. Springer, ISBN: 978-3-540-67686-7 Año 2000, Page(s):3-19.

[Marrón 2005]

Marrón, M.; García, J.C.; Sotelo, M.A.; Fernandez, D.; Pizarro, D. ***"XPFCP": An extended particle filter for tracking multiple and dynamic objects in complex environments***. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005). Agosto 2005, Page(s):2474-2479.

[Marrón 2008]

Marrón, M. ***Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos***. Phd thesis, Departamento de electrónica, Universidad de Alcalá, 2008.

[Martins 2007]

Martins, A.; Almeida, J.M.; Ferreira, H.; y otros. ***Autonomous Surfrace Vehicle Docking Manoeuvre with Visual Information***. IEEE International Conference on Robotics and Automation. Abril 2007, Page(s):4994-4999.

[Ng 2005]

Ng, W.; Li, J.; Godsill, S.; Vermaak, J.; ***A review of recents results in multiple target tracking.*** Proceedings of the Fourth International Symposium on Image and Signal Processing and Analysis (ISPA05), ISBN: 953-184-089-X, Zagreb, Septiembre 2005, Page(s):40-45.

[Ng 2006]

Ng, W.; Li, J.; Godsill, J. ***Online multisensor-multitarget detection and tracking.*** Proceedings of the 2006 IEEE Aerospace Conference, ISBN: 0-7803-9545-X, Montana, Marzo 2006.

[Odobez 2004]

Odobez, J.M.; Gatica-Perez, D. ***Embedding motion model-based stochastic tracking***. Proceedings of the Seventeenth International Conference on Pattern Recognition (ICPR04). Agosto 2004, Volumen 2, Page(s):815-818.

[Odobez 2007]

Odobez, J.M.; Ba, S. ***A Cognitive and Unsupervised Map Adaptation Approach to the Recognition of the Focus of Attention from Head Pose***. IEEE International Conference on Multimedia and Expo 2007. Julio 2007, Page(s):1379-1382.

[Ogata 1998]

Ogata, K. ***Ingeniería de Control Moderna*** – Tercera Edición.
Pearson Education - Prentice Hall, 1998.

[Okuma 2004]

Okuma, K.; Taleghani, A.; De Freitas, N.; Little, JJ.; Lowe, D.G. ***A boosted particle filter: multi-target detection and tracking***. Proceedings of the Eight European Conference on Computer Vision (ECCV04), Lecture Notes in Computer Science, ISBN: 3-540-21984-6, Praga, Mayo 2004, Volumen 3021, Parte I, Page(s):28-39.

[Perez 2002]

Pérez, P.; Hue, C.; Vermaak, J.; Gangnet, M. ***Color-based probabilistic tracking.*** Proceedings of the Seventh European Conference on Computer Vision (ECCV02), Lecture Notes in Computer Science. ISBN:3-540-43745-2, Copenhage, Mayo 2002, Volumen 2350, Parte I, Page(s):661-675

[Pizarro 2007]

Pizarro, D.; Santiso, E.; Mazo, M.; Marrón, M. ***Pose and Sparse Structure of a Mobile Robot using an External Camera***. IEEE International Symposium on Intelligent Signal Processing. Octubre 2007, Page(s):1-6.

[Rasmussen 2001]

Rasmussen, C.; Hager, G.D. ***Probabilistic data association methods for tracking complex visual objetcs***. IEEE Transactions on Pattern Analysis and Machine Intelligence. Junio 2001, Volumen 23, Page(s):560-576.

[Reid 1979]

Reid, D. ***An algorithm for tracking multiple targets***. IEEE Transactions on Automatic Control. Diciembre 1979, Volumen 24, Page(s):843-854.

[Rosales 1999]

Rosales, R.; Sclaroff, S. ***3D trayectory recovery for tracking multiple objects and trajectory guided recognition of actions***. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR99). Junio 1999, Volumen 2, Page(s):117-123.

[Schmid 2000]

Schmid, C.; Mohr, R.; Bauckhage, Ch. ***Evaluation of Interest Point Detectors***. International Journal of Computer Vision. Año 2000, Volumen 37, Page(s):46.

[Schulz 2001]

Schulz, D.; Burgard, W.; Fox, D.; Cremers, A.B. ***Tracking multiple moving objects with a mobile robot***. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2001), ISBN: 0-7695-1272-0, Diciembre 2001, Volumen 1, Page(s):371-377.

[Smith 2005]

Smith, K; Gatica-Perez, D.; Odobez, J.M. ***Using particles to track varying numbers of interacting people***. Proceedings of the Fourth IEEE Conference on Computer Vision and Pattern Recognition (CVPR05), San Diego, Junio 2005, Page(s):962-969.

[Smith 2008]

Smith, K.; Ba, S.; Odobez, J.M.; Gatica-Perez, D. ***Tracking the Visual Focus of Attention for a Varing Number of Wandering People***. IEEE Transactions on Pattern Analysis and Machine Intelligence, Julio 2008, Page(s):1212-1229.

[Springer 2005]

Blake, A. ***Visual tracking: A short research roadmap.*** Handbook of Mathematical Models of Computer Vision: The Handbook. Faugeras, O.; Chen, Y.; Paragios, N.; Ed Springer

[Sullivan 2001]

Sullivan, J.; Rittscher, J. ***Guiding Random Samples by Deterministic Search.*** Proceedings of International Conference on Computer Vision (ICCV 2001), Vancouver, Canada, Año 2001.

[Tao 1999]

Tao, H; Sawhney, H.S.; Kumar, R. ***A sampling algorithm for tracking multiple objects***. Proceedings of the International Workshop on Vision Algorithms at (ICCV99), Lecture Notes in Computer Science, ISBN: 3-540-67973-1, Corfu, Septiembre 1999, Volumen 1883, Page(s):53-68.

[Thrun 2000]

Thrun, S. ***Probabilistic algorithms in robotics***. Artificial Intelligent Magazine. Abril 2000, Volumen 21, Page(s):93-109.

[Tomasi 1994]

Jianbo, Shi.; Tomasi, C. ***Good features to track***. IEEE Proceedings of Computer Vision and Pattern Recognition. Junio 1994, Page(s):593-600.

[Torma 2003]

Torma, P; Szepesvári, C. ***Sequential importance sampling for visual tracking reconsidered***. Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics. ISBN: 0-9727358-0.1, Key West, Junio 2003.

[Tweed 2002]

Tweed, D.; Calway, A. ***Tracking Many Objects Using Subordinated Condensation***. Proceedings of the British Machine Vision Conference (BMVC2002), ISBN: 1901725197, Octubre 2002, Page(s):283-292.

[Van Der Merwe 2000]

Van der Merwe, R.; De Freitas, N.; Doucet, A.; Wan, E.; ***The unscented particle filter***. Technical Report CUED/F-INFEG/TR 380. Cambridge University Engineering Department. Agosto 2000, Page(s):46.

[Vermaak 2001]

Vermaak, J.; Gagnet, M.; Blake, A.; Perez, P. ***Sequential Monte Carlo fusion of sound and vision for speaker tracking***. Proceeding of the IEEE International Conference on Computer Vision (ICCV), Julio 2001, Page(s):6.

[Vermaak 2003]

Vermaak, J.; Doucet, A.; Pérez, P. ***Maintaining multi-modality through mixture tracking.*** Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV03), Junio 2003, Volumen 2, Page(s):1110-1116.

[Zhao 2004]

Zhao, T.; Nevatia, R. ***Tracking multiple humans in crowded environment***. Proceedings of the Third IEEE Conference on Computer Vision and Pattern Recognition (CVPR04). Junio 2004, Volumen 2, Page(s):406-413.

FUSIÓN

[Aarabi 2001]

Aarabi, P.; Zaky, S. ***Robust sound localization using multi-source audiovisual information fusion***. Information Fusion, Septiembre 2001, Volumen 3, Page(s):209-223.

[Beal 2002]

Beal, M.; Attias, H.; Jojic, N. ***Audio-video sensor fusion with probabilistic graphical models***. Proceeding of European Conference on Computer Vision, Mayo 2002, Page(s):736-750.

[Bernardin 2006]

Bernardin, K. ***Multiple Object Tracking Performance Metrics and Evaluation in Smart Room Environment***. The Sixth IEEE International Workshop on Visual Surveillance (VS 2006), Graz, Austria, Mayo 2006.

[Busso 2005]

Busso, C.; Hernanz, S.; Chu, C.-W.; Kwon, S.-I.; Lee, S.; Georgiou, P.; Cohen, I.; Narayanan, S. ***Smart room: Participant and speaker localization and identification***. Proceeding IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP05), Marzo 2005, Volumen 2, Page(s):1117-1120.

[Canton-Ferrer 2008]

Canton-Ferrer, C.; Segura, C.; Casas, J.R.; Pardàs, M.; Hernando, J. ***Audiovisual Head Orientation Estimation with Particle Fitlering in Multisensor Scenarios***. EURASIP Journal on Advances in Signal Processing, Enero 2008, Page(s):12.

[Checka 2004]

Checka, N.; Wilson, K.; Siracusa, M.; Darrell, T. ***Multiple person and speaker activity tracking with a particle filter***. Proceeding IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP04), Mayo 2004, Volumen 5, Page(s):881-884.

[Chen 2004]

Chen, Y.; Rui, Y. ***Real-time speaker tracking using paticle filter sensor fusion***. Proceeding of the IEEE, Marzo 2004, Volumen 92, Page(s):485-494.

[CLEAR 2007a]

Bernardin, K. ***CLEAR 2007 Evaluation Plan. 3D Person Tracking Task***. Proyecto CHIL (Computers in the Human Interaction Loop). Marzo 2007.

[CLEAR 2007b]

Voit, M. ***CLEAR 2007 Evaluation Plan. Head Pose Estimation***. Proyecto CHIL (Computers in the Human Interaction Loop). Febrero 2007.

[Cutler 2000]

Cutler, R.; Davis, L. ***Look who's talking: Speaker detection using video and audio correlation***. Proceeding of the IEEE International Conference on Multimedia (ICME), Julio 2000, Volumen 3, Page(s):1589-1592.

[Cutler 2002]

Cutler, R.; Rui, Y. Gupta, A.; Cadiz, J.; Tashev, L.; He, L.; Colburn, A.; Zhang, Z.; Liu,Z.; Silverberg, S. ***Distributed meetings: a meeting capture and broadcasting system***. Proceeding of the ACM International Conference on Multimedia (MM), Diciembre 2002, Page(s):10.

[Gatica-Perez 2003a]

Gatica-Perez, D.; Lathoud, G.; McCowan, I.; Odobez, J-M.; Moore D.; ***Audio-visual speaker tracking with importance particle filters***. Proceeding of the International Conference on Image Processing (ICIP), Octubre 2005, Page(s):8.

[Gatica-Perez 2003b]

Gatica-Perez, D.; Lathoud, G.; McCowan, I.; Odobez, J-M.; ***A mixed-state i-particle filter for multi-camera speaker tracking***. Proc. of the IEEE Int. Conf. on Computer Vision, Workshop on Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC), Octubre 2005, Page(s):4.

[Gatica-Perez 2005]

Gatica-Perez, D.; Lathoud, G.; Odobez, J-M.; McCowan, I. ***Multimodal multispeaker probabilistic tracking in meetings***. Proceeding of the International Conference on Multimodal Interfaces (ICMI), Octubre 2005, Page(s):8.

[Gatica-Perez 2007]

Gatica-Perez, D.; Lathoud, G.; Odobez, J-M.; McCowan, I. ***Audio-Visual Probabilistic Tracking of Multiple Speakers in Meetings***. IEEE Transactions on Audio, Speech, and Language Processing, Febrero 2007, Volumen 15, Page(s):601,616.

[IDIAP]

IDIAP research institute: http://www.idiap.ch/~gatica/av-tracking-multiperson.html

[Kapralos 2003]

Kapralos, B.; Jenkin, M.; Milios, E. ***Audio-visual localization of multiple speakers in a video teleconferencing setting***. Int. J. Imaging Syst. And Tech, Año 2003, Volumen 13, Page(s):95-105.

[Lincoln 2005]

Lincoln, M.; McCowan, I.; Vepa, J.; Maganti, H.-K. ***The Multi-Channel Wall Street Journal Audio-Visual Corpus (MC-WSJ-AV): Specifications and Initial Experiments***. Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU), Diciembre 2005, Page(s):6.

[Maganti 2006a]

Maganti, H.-K.; Gatica-Perez, D. ***Speaker Localization for Microphone-Array-Based ASR: the Effect of Accuracy on Overlapping Speech***. Proceeding of the International Conference on Multimodal Interfaces (ICMI), Noviembre 2005, Page(s):4.

[Maganti 2006]

Maganti, H.-K.; Gatica-Perez, D.; McCowan, I. **Speech Enhancement and Recognition in Meetings with an Audio-Visual Sensor Array**. IDIAP Research Report IDIAP-RR-06-24, submitted to IEEE, Abril 2006, Page(s):13.

[Potamitis 2004]

Potamitis, I.; Chan, H.; Tremoulis, G. ***Tracking of multiple moving speakers with multiple microphone arrays***. IEEE Transactions on Speech and Audio Processing, Septiembre 2004, Volumen 12, Page(s):520,529.

[Siracusa 2003]

Siracusa, M.; Morency, L.-P.; Wilson, J.; Fisher, J.; Darrell, T. ***A multi-modal approach for determining speaker location and focus***. Proceeding of the International Conference on Multimodal Interfaces (ICMI), Noviembre 2003, Page(s):4.

[Smith 2006]

Smith, K.; Schreiber, S.; Potucek, I.; Beran, V.; Rigoll, G.; Gatica-Perez, D. ***2D Multi-Person Tracking: A Comparative Study in AMI Meetings***. Proceeding of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Mayo 2006, Page(s):331,344.

[Vo 2004]

Vo, B.; Singh, S.; Ma, W.K. ***Tracking multiple speakers using random sets***. Proceeding of the IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), Mayo 2004, Page(s):357-360.

[Wolfel 2005]

Wolfel, M.; Nickel, K.; McDonough, J. ***Microphone Array Driven Speech Recognition: Influence of Localization in the Word Error Rate***. Proceeding of the Workshop on Machine Learning for Multimodal Interaction (MLMI), Julio 2005, Page(s):320-331.

[Yu 2004]

Yu, T.; Wu, Y. ***Collaborative tracking of multiple targets***. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Junio 2004, Volumen 1, Page(s):834-841.

[Zotkin 2001]

Zotkin, D.; Duraiswami, R.; Davis, L. ***Multimodal 3-D tracking and event detection via the particle filter***. IEEE International Conference on Computer Vision, Workshop on Detection and Recognition of Events in Video (ICCV-EVENT), Julio 2001, Page(s):20-27.