Modelado de *arrays* de micrófonos como cámaras de perspectiva en aplicaciones de localización de locutores

Alejandro Legrá-Ríos, Javier Macías-Guarasa, Daniel Pizarro y Marta Marrón-Romera
Departamento de Electrónica — Universidad de Alcalá
Alcalá de Henares-España
email: {alejandro.legra,macias,pizarro,marta}@depeca.uah.es

***Resumen*—Las tareas de localización y seguimiento de personas en espacios inteligentes son esenciales para mejorar la interacción entre el sistema y en entorno. En este trabajo, se presenta un sistema de localización de personas usando la información proporcionada por varias agrupaciones (*arrays*) de micrófonos situados en el espacio inteligente. Se ha implementado una estrategia basada en el modelado del *array* de micrófonos como cámaras de perspectiva, que generan imágenes relacionadas con la potencia acústica recibida por cada *array*. El sistema usa técnicas de visión computacional para estimar finalmente las posiciones de los usuarios. La evaluación preliminar se ha realizado con datos procedentes de la campaña de evaluación CLEAR 2007, generada en el proyecto CHIL, financiado por la Unión Europea.**

I. INTRODUCCIÓN

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

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

En la literatura científica existe una gran cantidad de trabajos que emplean la información procedente de un único tipo de sensores con el objetivo de localizar determinados objetivos dentro de la escena bajo análisis, como por ejemplo cámaras de vídeo [7] [8], agrupaciones (*arrays*) de micrófonos [9] [10], balizas infrarrojo, etc. Entre todos ellos, los dos primeros han sido los más empleados.

Los sistemas basados en sensores de visión presentan importante dificultades, debidas principalmente a las transformaciones que sufren las imágenes ante variaciones naturales de la escena, como pueden ser los cambios de iluminación, sombras, contraste, o los cambios naturales sufridos por los objetos (color, forma, tamaño, textura,..). Aparte de dichas variaciones, la pérdida de dimensionalidad que implica la proyección del mundo tridimensional en una cámara, da lugar a oclusiones y ambigüedades difíciles de tratar.

Por el contrario, las señales de audio, son poco direccionales, especialmente a baja frecuencia, por lo que puede servir de complemento cuando el sistema de visión no puede seguir al objetivo.

Este artículo describe la propuesta y evaluación inicial de un módulo de localización basado en audio, mediante el modelado de *arrays* de micrófonos como cámaras de perspectiva. La información visual obtenida, se utiliza para realizar la estimación de la posición de la persona que está hablando.

En la literatura hay muy pocos trabajos en esta línea, entre los que destacan los descritos en [11], restringido al caso de un *array* de micrófonos esférico modelado como una cámara de proyección central y en el que no hay una evaluación objetiva ni se aborda el tratamiento de múltiples arrays de micrófonos. Otros autores plantean propuestas de generación de imágenes a partir de la información acústica que proporcionan arrays de micrófonos, pero restringidos a tareas de calibración, ofreciendo resultados únicamente en simulación [12] o para entornos acústicos no reverberantes [13].

En este trabajo se realiza una propuesta de integración de la información visual generada por múltiples arrays de micrófonos, aplicando técnicas de visión computacional, y se evalúa sobre bases de datos de referencia en la comunidad científica.

II. DESCRIPCIÓN DEL SISTEMA

II-A. Esquema general propuesto

Una cámara de vídeo proporciona información del mundo 3D sobre una imagen en 2 dimensiones, siendo necesario realizar una transformación de un espacio a otro [14]. La forma en la que se realiza dicha transformación, se explica mediante el modelo de una cámara de perspectiva, en el que cada punto de la imagen obtenida es el resultado de la proyección del rayo que pasa por el centro de la lente e intersecta el plano imagen.

En nuestra propuesta, y de manera análoga, se generan una serie de localizaciones que parten del centro del *array* de micrófonos formando rayos "acústicos" que barren todo el espacio de búsqueda, como se muestra en la Figura 1. En cada rayo acústico, se calculará la potencia acústica que recibe el *array*, utilizando la información obtenida a partir de la aplicación del algoritmo de estimación de la respuesta en potencia dirigida SRP-PHAT (*Steered Response Power with Phase Transform*) [15], de modo que aquellas direcciones del espacio (con respecto al *array* de micrófonos) en las que haya una fuente acústica activa aparecerán con un mayor valor de intensidad en la imagen generada.

El sistema propuesto se divide en las siguientes etapas:

- Construcción de mapas tridimensionales con información de la energía acústica de las distintas direcciones que conforman todo el espacio donde se espera encontrar el locutor activo.
- Generación de imágenes bidimensionales con información de la potencia acústica de las distintas direcciones que conforman todo el espacio donde se espera encontrar el locutor activo.
- Búsqueda de la dirección del máximo de potencia acústica, utilizando el algoritmo de *Non Maximun Supression* con aproximación subpixélica, a partir de una función cuadrática.
- Cálculo de la posición 3D a partir de la técnica de triangulación lineal DLT (*Direct Lineal Transform*), combinada con una variante de la técnica RANSAC (*Random Sample Consensus*) para eliminar las estimaciones de máximos de potencia que no sean coherentes con el resto (en las salas donde haya más de 4 *arrays* de micrófonos).

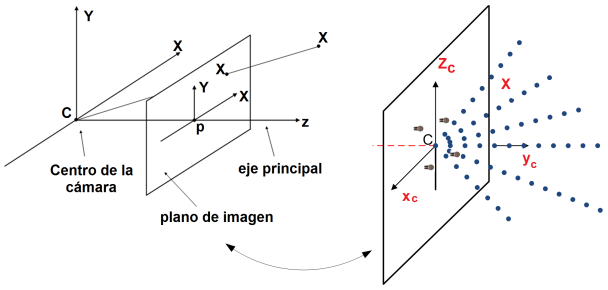

Figura 1. Modelado del *array* de micrófonos como cámara de perspectiva

II-B. Cálculo de la potencia acústica en un punto en el espacio y construcción de los mapas de potencia

En el cálculo de la potencia acústica, se utiliza la técnica SRP con filtrado PHAT (SRP-PHAT) [15], que permite redirigir el patrón de recepción de un *array* de micrófonos hacia una posición determinada y es la que mejores prestaciones muestra, en entornos reverberantes [10].

Para implementar SRP-PHAT se utiliza la salida de un *filter-delay-and-sum beamformer*, cuya expresión en el dominio de la frecuencia viene dada por la ecuación (1).

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

donde Δ_n es el *retardo* adecuado aplicado al micrófono n para dirigir el patrón de recepción del *array* a la localización espacial q, y $X_n(\omega)$ y $W_n(\omega)$ son las Transformadas de Fourier de la señal recibida en el micrófono n y su filtro asociado, respectivamente.

La expresión de la potencia de salida recibida al apuntar a una posición $\mathbf{q}$ viene dada por la ecuación (2)

$$P(\mathbf{q}) = \int_{-\infty}^{\infty} |Y(\omega)|^2 d\omega = \int_{-\infty}^{\infty} Y(\omega) Y'(\omega) d\omega \quad (2)$$

Para mejorar los efectos perjudiciales provocados por el ruido y la reverberación, el filtro W_n consiste en una función de pesado PHAT (*Phase Transform*), propuesta en [15], donde también se demuestra que, finalmente, la expresión de la potencia recibida es la mostrada en la ecuación (3):

$$P(\mathbf{q}) = \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{-\infty}^{\infty} \Phi_{ij}(\omega) X_i(\omega) X_j'(\omega) e^{j\omega(\Delta_j - \Delta_i)} d\omega \quad (3)$$

donde:

$$\Phi_{ij}(\omega) = W_i(\omega) W_j'(\omega) = \frac{1}{|X_i(\omega) X_j'(\omega)|} \quad (4)$$

son los filtros PHAT

Esta estrategia, a pesar de ser la que mejores resultados presenta en los entornos reales, tiene como principal desventaja que la ponderación se realiza con el inverso de su módulo, por lo que los errores se acentúan, cuando la potencia de la señal es baja [10].

Para la implementación sobre señales digitales, se aplica el tradicional enventanado temporal, utilizando la ventana de Hamming, dividiéndolas en bloques de tamaño fijo antes de aplicarles la Transformada Rápida de Fourier (FFT). El sistema de cálculo de potencia realiza una estimación en cada bloque de datos, asumiendo que el locutor no se moverá durante ese periodo de tiempo.

El algoritmo SRP-PHAT es evaluado en cada una de las localizaciones que conforman el espacio de búsqueda. Dichas localizaciones están alineadas según una serie de rayos "visuales" que parten desde el centro del *array* de micrófonos y realizan un barrido en azimut y elevación, tal y como se muestra en la Figura 2. Estimando el valor de intensidad recibida según cada uno de los rayos, se construye el mapa visual de la potencia acústica "vista" por el *array* de micrófonos considerado.

Nuestra primera aproximación al cálculo del valor de intensidad de cada uno de los rayos definidos (que corresponden a un valor del par (azimut, elevación) dentro de ese mapa de potencia), se basa en el cálculo del promedio de las potencias estimadas en todas las localizaciones que pertenecen a cada rayo, a través de la ecuación (5):

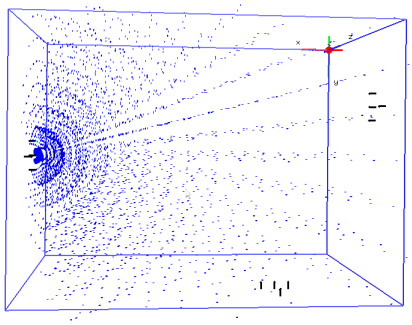

Figura 2. Localizaciones generadas en una habitación para calcular el mapa de potencia acústica en uno de los *arrays* de micrófonos

$$P_{(\text{azimut,elevación})} = \frac{\sum_{k=1}^{k=n_{(\text{azimut,elevación})}} P(\mathbf{q}_k)}{n_{(\text{azimut,elevación})}} \quad (5)$$

donde $n_{(azimut,elevación)}$ representa el número de localizaciones que conforman el rayo considerado y $P(\mathbf{q}_k)$ la potencia obtenida en la localización $\mathbf{q}_k$ a la que pertenece el rayo.

En la Figura 3 se muestra un ejemplo de mapa de potencia obtenido a partir de un *array* de micrófonos en una ventana acústica en la que hay un locutor activo.

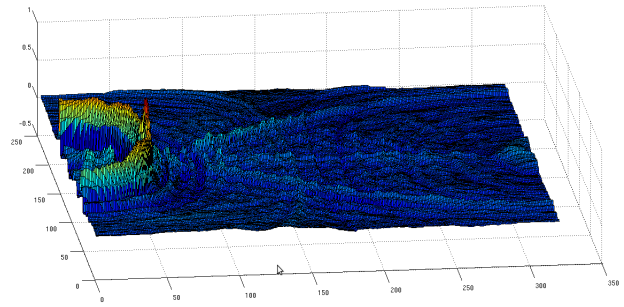

Figura 3. Mapa de potencia obtenido a partir de un *array* de micrófonos, con un locutor activo (ejes x e y medidos en píxeles, y z proporcional a potencia acústica)

II-C. Generación de imágenes a partir de los mapas de potencia acústica

Las imágenes de potencia acústica fueron generadas a partir de los mapas de potencias descritos anteriormente, en las que cada píxel equivale a un punto en el mapa de potencia, de coordenadas $n_{(\text{azimut,elevación})}$ y cuya intensidad es proporcional a la potencia acústica calculada en esa dirección. La altura de la imagen corresponde al barrido en elevación, y la anchura al barrido en azimut.

Las imágenes se generaron en escala de grises con una profundidad de 8 bits por píxel. Debido a que los niveles de potencia son valores reales, es necesario obtener un rango de variación entre 0 y 255, usando un escalado que tiene en cuenta los niveles de potencia obtenidos en todos los mapas de la base de datos disponible. En la Figura 4 se muestra una imagen obtenida en una ventana acústica con presencia de un locutor activo.

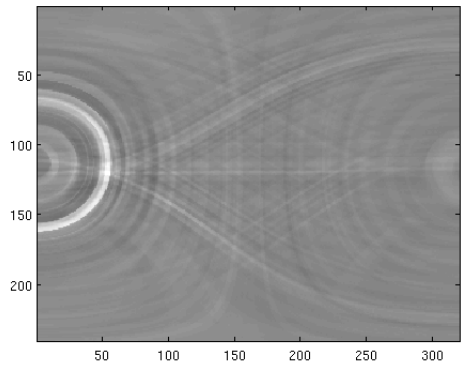

Figura 4. Imagen de potencia acústica obtenida a partir de un *array* de micrófonos, con un locutor activo (ejes medidos en píxeles)

II-D. Búsqueda de la dirección de máxima potencia acústica

A partir de las imágenes o de los mapas de potencias obtenidos para cada ventana de análisis, se ha aplicado técnica *Non Maximum Supression* con aproximación subpixélica (descrito en [16]) para buscar las direcciones de máxima potencia acústica. La suposición del método propuesto es que en dichas direcciones de máximo de potencia se deben encontrar las fuentes activas de sonido, vistas por cada uno de los *arrays* de micrófonos.

En este estudio preliminar, la búsqueda se restringió a un único máximo, lo que permitiría detectar a un único locutor activo.

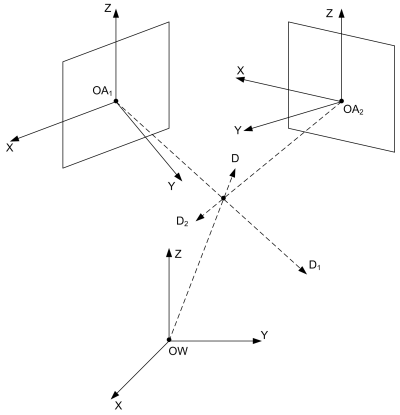

Figura 5. Triangulación lineal

II-E. Cálculo de la posición 3D

Una vez modelado cada array de micrófonos como una cámara de perspectiva, se propone obtener la posición 3D del locutor mediante un enfoque de triangulación a partir de múltiples imágenes. Esta técnica ha sido ampliamente utilizada en el ámbito de la visión artificial [14].

Se supone un conjunto compuesto por N arrays de micrófonos, dispuestos en diferentes posiciones y orientaciones del espacio. Fijando un origen de coordenadas global OW, cada array i se caracteriza por la posición $T^{OA_i} \in \mathbb{R}^3$ y orientación $R^{OA_i} \in SO(3)$ de su centro de proyección OA_i con respecto OW. De este modo, una posición tridimensional

cualquiera X^{OW}, tomada con respecto al origen OW, se puede expresar en el origen de coordenadas de cada array:

$$X^{OA_i} = R^{OA_i}[X^{OW} - T^{OA_i}] \tag{6}$$

Una vez obtenidas las direcciones de los máximos de potencia para cada array, descritos por los vectores de 3 dimensiones D_i con respecto a los orígenes de coordenadas de cada array (ver Figura 5 para el caso de $N = 2$), se trata de estimar el vector de posición X^{OW} de la fuente activa de sonido en las coordenadas absolutas el mundo (con origen OW). Para ello, nuestra propuesta utiliza el algoritmo de triangulación lineal DLT, descrito en [14] y que se resume a continuación:

Para un array i un punto X^{OA_i}, en la dirección D_i debe ser colineal a dicha dirección:

$$D_i \times X^{OA_i} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{7}$$

Sustituyendo (6) en (7):

$$D_i \times \left(R^{OA_i}[X^{OW} - T^{OA_i}]\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{8}$$

A partir de (8) se obtiene un sistema lineal de tres ecuaciones en función de X^{OW}:

$$A_i X^{OW} = B_i \qquad A_i \in \mathbb{R}^{3\times 3} \qquad B_i \in \mathbb{R}^3 \tag{9}$$

Se puede demostrar que el sistema (9) presenta sólo 2 ecuaciones linealmente independientes para 3 incógnitas. Combinándose los sistemas de ecuaciones de los N arrays y eliminando una ecuación de cada uno de ellos, se obtiene un sistema con $2N$ ecuaciones y 3 incógnitas.

$$AX^{OW} = B \qquad A \in \mathbb{R}^{2N\times 3} \qquad B_i \in \mathbb{R}^{2N} \tag{10}$$

Si $N > 2$ el sistema tiene suficiente rango y las coordenadas estimadas del locutor (X^{OW}) se pueden obtener como la solución de mínimo error cuadrático, representada en la ecuación (11).

$$X^{OW} = \left(A^T A\right)^{-1} A^T B \tag{11}$$

La estimación de la posición del locutor mediante el algoritmo de triangulación, se ve muy afectado por los errores en la estimación de los máximos de potencia acústica, obtenidos en los *arrays* de micrófonos. Por esta razón se utiliza una variante del algoritmo RANSAC para eliminar los resultados incoherentes del proceso de triangulación [17]. En esta variante, se estima la localización por cada par de *arrays* de micrófono posible, luego se agrupan utilizando un umbral de distancia y se selecciona el grupo más poblado como aquel que con mayor probabilidad contiene medidas coherentes. Para obtener la posición final, se seleccionan únicamente los arrays de micrófonos incluidos en ese grupo más poblado.

III. DESARROLLO EXPERIMENTAL

III-A. Bases de datos

Los resultados fueron evaluados con parte de las bases de datos generadas en el proyecto *Computer in the Human Interaction Loop* (CHIL), conjuntamente con la Campaña de Evaluación *Classification of Events, Activities and Relationships* (CLEAR 2007) [18]. Dicha campaña constituye un esfuerzo internacional para evaluar sistemas que son diseñados para analizar a las personas, sus identidades, actividades, interacciones y en escenarios con interacción hombre-hombre. En dicha evaluación se analizaron varias tecnologías agrupadas fundamentalmente en el área de audio y visión. Dentro del área de audio se incluyen tecnologías que permiten la identificación, localización y seguimiento del locutor activo.

Los datos evaluados son:

- De los seminarios de AIT en la base de datos CLEAR 2007 se usó el conjunto de test (de unos 40 minutos, con unas 1586 tramas etiquetadas de locutores activos).
  En estos seminarios se usaron tres *arrays* de micrófonos, cada uno de los cuales está compuesto por 4 micrófonos, 3 de ellos ubicados en configuración lineal y el otro en la parte superior, formando el conjunto una T invertida.
- De los seminarios de ITC en la base de datos CLEAR 2007 se usó el conjunto de desarrollo, de aproximadamente una hora de duración. Las grabaciones se realizan en este caso con 28 micrófonos, agrupados en 7 *arrays*, con la misma configuración que los de AIT. El hecho de que existan 7 *arrays* permite utilizar la técnica de estimación de coherencia descrita anteriormente.

III-B. Métricas de evaluación

Para la evaluación se utilizaron las métricas definidas en el proyecto CHIL (y seleccionadas también en la evaluación CLEAR 2007) [19]. En dicho esquema, los errores cometidos por el sistema de localización implementado se dividen en dos grupos (se mantiene la nomenclatura inglesa por ser de difícil traducción y comúnmente utilizada en este tipo de evaluaciones):

- *fine error*: Si la distancia que existe entre la posición real y la estimada es menor o igual de 500 mm.
- *gross error*: Si la distancia entre la posición real y la estimada es mayor de 500 mm.

Las métricas fundamentales son las siguientes:

- Pcor: Es la relación entre la cantidad de *frames* en que el error del posicionamiento es considerado como *fine error* y la cantidad de *frames* en los que hay algún locutor activo.
- *Bias/AEE fine+gross*: El error medio (por coordenada o conjunto) cometido en todas las estimaciones realizadas.
- *Bias/AEE fine*: El error medio (por coordenada o conjunto) en las estimaciones definidas como *fine errors*.
- Deletions: se define como el porcentaje (%) de la cantidad de *frames* en los que el sistema de localización no devuelve resultado respecto a la cantidad total de tramas etiquetadas.

III-C. Experimento Base

Se define un experimento base *(baseline)*, que se utiliza como patrón de comparación con otros experimentos. Este está definido por un conjunto de parámetros divididos en 2 grupos: el primero de ellos permanecerá invariante durante todo el proceso de experimentación y está relacionado con el proceso del cálculo de SRP, el segundo está relacionado con el proceso de generación de localizaciones en el que se calculará la potencia acústica definiendo la resolución con la que se construyen las imágenes y los mapas de potencia.

En el *baseline* se usan ventanas de anchura $0,5$ segundos y el mismo desplazamiento, y los siguientes parámetros de control:

- Número de puntos utilizados en el barrido horizontal (azimut): 320 puntos.
- Cantidad de puntos utilizados en el barrido vertical (elevación): 240 puntos.
- Paso en profundidad: 100 mm (separación entre las localizaciones dentro de un mismo rayo "visual".
- Alturas máxima y mínima a la que se puede encontrar una persona hablando: $2000\ mm$ *y* $750\ mm$.

A partir del experimento base, se analizó el efecto que tiene sobre los resultados la resolución con la que se generan las imágenes, utilizando los datos de AIT.

Además se realizaron pruebas adicionales con la base de datos ITC, para comprobar las posibles mejoras que provocaría el uso de un algoritmo de estimación de coherencia que elimine del proceso de localización los *arrays* que aportan máximos de potencias incoherentes con el resto.

III-D. Estudio del efecto del barrido en ángulos

Se realizó un estudio detallado de los efectos que tiene la resolución en el barrido de ángulos. Para ello se parte del experimento *baseline* cuya resolución es 320 en azimut y 240 en elevación, y se repiten los experimentos con cada una de las siguientes resoluciones: (240×180) (160×120). Las métricas de CHIL muestran un aumento en los errores de manera gradual, a medida que disminuye la resolución en los barridos de azimut y elevación, (ver tabla I), como era previsible, aunque las diferencias no son estadísticamente significativas, lo que apunta a trabajos futuros orientados a la optimización de la exploración del espacio de búsqueda. Los errores cometidos en el eje z son considerablemente mayores que en los otros ejes, lo que se justifica por la menor resolución vertical de la estructura de los *arrays* usados. También se puede observar una disminución apreciable del tiempo de procesamiento (Tiempo real).

Para establecer una comparación con sistemas tradicionales de localización basada en SRP-PHAT evaluando el espacio de búsqueda completo con un grid de estimaciones de 5 cm de separación entre puntos y las mismas condiciones en el resto de los parámetros de control, se obtiene un $Pcor = 0.63$, $AEE\ fine\ [mm] = 233$ y $AEEfine + gross\ [mm] = 572$. Los resultados del trabajo presentado en este artículo son del mismo orden que estos, pero siendo un trabajo preliminar sin haber profundizado en un estudio de otras alternativas ni ajuste de elementos de control, resultan muy prometedores.

Métricas de CHIL	320×240	240×180	160×120
Pcor	$55.0 \pm 2.7\,\%$	$56.0 \pm 2.7\,\%$	$53.0 \pm 2.7\,\%$
Bias fine (x:y:z)	$1:29:118$	$1:38:119$	$4:37:118$
Bias fine+gross (x:y:z)	$23:59:171$	$37:58:175$	$83:92:173$
AEE fine	246 mm	253 mm	253 mm
AEE fine+gross	634 mm	637 mm	654 mm
Deletion rate	$0\,\%$	$0\,\%$	$0\,\%$
Veces tiempo real	3.17	1.75	0.84

Tabla I
RESULTADOS GLOBALES OBTENIDOS CON LAS BASE DE DATOS AIT PARA DISTINTAS RESOLUCIONES

III-E. Estudio del efecto del barrido en profundidad

En esta sección se realiza un estudio de los efectos que tiene el uso de distintas resoluciones en el barrido de profundidad. Para ello se parte del experimento *baseline* en el que cada rayo está formado por una localización cada $100\ mm$ realizando experimentos para los siguientes pasos en profundidad de la generación de los puntos: $50\ mm$, $200\ mm$. En la tabla II se pueden ver que las métricas de CLEAR reflejan resultados muy similares para cada una de las resoluciones en el barrido de profundidad, lo que muestra que no es un parámetro crítico del sistema y de nuevo sugiere la necesidad de trabajar más la selección del espacio de búsqueda general y dentro de la exploración para cada rayo.

Métricas de CHIL	320×240 $\Delta r\ 50\ mm$	320×240 $\Delta r\ 100\ mm$	320×240 $\Delta r\ 200\ mm$
Pcor	$55.0 \pm 2.7\,\%$	$55.0 \pm 2.7\,\%$	$56.0 \pm 2.7\,\%$
Bias fine (x:y:z)	$4:29:118$	$1:29:118$	$4:38:113$
Bias fine+gross (x:y:z)	$27:57:170$	$23:59:171$	$17:60:167$
AEE fine	247 mm	246 mm	249 mm
AEE fine+gross	631 mm	634 mm	625 mm
Deletion rate	$0\,\%$	$0\,\%$	$0\,\%$
Veces tiempo real	5.23	3.17	1.93

Tabla II
RESULTADOS GLOBALES OBTENIDOS CON LA BASE DE DATOS DE AIT PARA DISTINTAS VARIACIONES DE LOS PUNTOS EN PROFUNDIDAD

III-F. Resultados del proceso de estimación de coherencia

En la tabla III se muestran los resultados del proceso de localización usando y sin usar el proceso de estimación de coherencia sobre la base de datos de ITC.

En ésta se observan las mejoras obtenidas por la aportación del proceso de análisis de coherencia en todas las métricas de CHIL excepto en la que se refiere al número de borrados, que sube al 25 %. Esto ocurre, debido a que en el proceso de análisis de coherencia no se generan estimaciones si no hay como mínimo 4 *arrays* de micrófonos cuyas proyecciones de máximos de potencia acústica son coherentes.

El resultado apunta a la necesidad de dichas técnicas en sistemas con un número elevado de *arrays*, si bien es necesario solucionar las estimaciones incoherentes que eviten el elevado número de borrados.

Para establecer una comparación con sistemas tradicionales de localización basada en SRP-PHAT evaluando el espacio de

Métricas de CHIL	Coherencia	Sin Coherencia
Pcor	$60.0 \pm 3.5\,\%$	$12.0 \pm 2.0\,\%$
Bias fine (x:y:z)	$151 : 5 : 24$	$139 : 210 : 82$
Bias fine+gross (x:y:z)	$351 : 306 : 14$	$352 : 787 : 126$
Bias AEE fine	198 mm	348 mm
Bias fine+gross	759 mm	1020 mm
Deletion rate	$25\,\%$	$0\,\%$

Tabla III
RESULTADOS OBTENIDOS ELIMINANDO LOS ARRAYS QUE NO REALIZAN UNA PROYECCIÓN DEL MÁXIMO DE POTENCIA DE FORMA COHERENTE

búsqueda completo con un grid de estimaciones de 5 cm de separación entre puntos y las mismas condiciones en el resto de los parámetros de control, se obtiene un $Pcor = 0.62$, $AEE\ fine\ [mm] = 175$ y $AEE\ fine + gross\ [mm] = 726$, sobre los que cabe hacer el mismo comentario que en el apartado III-D.

IV. CONCLUSIONES Y LÍNEAS FUTURAS

En este trabajo se ha propuesto, implementado y evaluado un sistema de localización basado en información de audio a través del modelado de *arrays* de micrófonos como cámaras de perspectiva. Se construyeron imágenes y mapas con la información de potencia acústica en las diferentes direcciones en que se esperaba encontrar el locutor activo, y a partir de los máximos de potencia encontrados por el algoritmo *Non Maximun Supression* se realizó una estimación de la posición para cada *frame*, utilizando el algoritmo de triangulación lineal DLT, muy usado en aplicaciones de localización de objetos a partir de imágenes obtenidas por varias cámaras.

El sistema implementado fue evaluado a partir de un experimento base con parte de las bases de datos de la evaluación CLEAR 2007. Se evaluó el efecto de distintos elementos de control de la propuesta, incluyendo el efecto de la aplicación de un sistema de estimación de coherencia que permite eliminar del proceso de triangulación los *arrays* de micrófonos que proporcionan medidas erróneas.

Los resultados obtenidos, si bien no mejoran los que presentan algoritmos clásicos de localización basados en SRP-PHAT exhaustivo, son prometedores al abrir un buen número de estudios futuros.

Como trabajo futuro se plantea una evaluación exhaustiva de distintas mejoras sobre el algoritmo SRP, la definición de nuevas estrategias de generación del espacio de búsqueda (para fuentes en campo lejano y cercano), la utilización de técnicas más elaboradas de algoritmos de detección de coherencia en las medidas de los máximos encontrados, estudios detallados de la influencia que tiene la posición y orientación del locutor en el espacio sobre el error generado, así como el uso de técnicas de seguimiento para realizar un filtrado espacio temporal de los resultados y mejorar las prestaciones, incluyendo el tratamiento de escenas con múltiples locutores activos.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación con los proyectos VISNÚ (ref. TIN2009-08984) y SDTEAM-UAH (ref. TIN2008-06856-C05-05), y por la Comunidad de Madrid y la Universidad de Alcalá bajo el proyecto FUVA (CCG10-UAH/TIC-5988).

REFERENCIAS

[1] A. P. Consortium, "Augmented multi-party interacion (ami) project. state of the art overview: Localization and tracking of multiple interlocutors with multiple sensors," Technical report, Tech. Rep., 2007.
[2] M. L. Seltzer, "Microphone array processing for robust speech recognition," Ph.D. dissertation, Carnegie Mellon University, 2003.
[3] W. Herbordt, *Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing*. Springer, Heidelberg, Germany, March, 2005.
[4] D. Gelbart and N. Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," *In International Conference on Spoken Language Processing (ICSLP)*, 2002.
[5] S. Kochkin and T. Wickstrom, "Headsets, far field and handheld microphones: Their impact on continuous speech recognition," Technical report, EMKAY, a division of Knowles Electronics, Tech. Rep., 2002.
[6] A. Vinciarelli, H. Salamin, and M. Pantic, "Social signal processing: Understanding social actions through nonverbal behaviour analysis," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, vol. 3. Los Alamitos: IEEE Computer Society Press, 2009, pp. 42–49.
[7] D. Pizarro, M. Mazo, E. Santiso, M. Marron, and I. Fernandez, "Localization and geometric reconstruction of mobile robots using a camera ring," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, pp. 2396–2409, 2009.
[8] O. Lanz, "Approximate bayesian multibody tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1436–1449, 2006.
[9] G. Lathoud and M. Magimai-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 265–268, mar 2005.
[10] C. Castro García, "Speaker localization techniques in reverberant acoustic environments," Master's thesis, Royal Institute of Technology (KTH), Stockholm, 2007.
[11] A. O'Donovan and R. Duraiswami, "Microphone arrays as generalized cameras for integrated audio visual processing," *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jul. 2007.
[12] A. Redondi, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Geometric calibration of distributed microphone arrays," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing 2009*, Rio De Janeiro, Oct. 2009, pp. 1–5.
[13] S. Valente, F. Antonacci, M. Tagliasacchi, A. Sarti, and S. Tubaro, "Self-calibration of two microphone arrays from volumetric acoustic maps in non-reverberant rooms," in *Proc. of the International Symposium on Communications, Control and Signal Processing*, Limassol, Cyprus, March 2010.
[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision. Second Edition*. Cambridge University Press, 2003.
[15] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown University, 2000.
[16] J. Knight, A. Davison, and I. Reid., "Towards constant time slam using postponement," in *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2001, pp. 406–412.
[17] A. Legrá Rios, "Estudio, implementación y evaluación de un sistema de localización de locutores basado en el modelado de arrays de micrófonos como cámaras de perspectiva," Master's thesis, Universidad de Alcalá, Alcalá de Henares. España, 2010.
[18] R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds., *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*. Springer, 2008.
[19] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment,," in *Proc. of the Sixth IEEE International Workshop on Visual Surveillance (in conjunction with ECCV)*, Graz, Austria, May 2006.