

Trabajo Fin de Carrera

“Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual”

Autora: María Cabello Aguilar
Tutores: Javier Macías Guarasa
Marta Marrón Romera



Septiembre 2010



- 1. Introducción**
2. Estudio teórico
3. Desarrollo
4. Resultados obtenidos
5. Conclusiones
6. Líneas futuras



Introducción: Presentación

3/35

- Análisis automático de **Espacios Inteligentes** a partir de múltiples sensores.



- Tareas de detección, localización y seguimiento de personas.
- Explotación de información acústica y procedente de vídeo.



Introducción: Objetivos

4/35

- **Mejorar las prestaciones** de las herramientas y algoritmos en sistemas de procesamiento de audio (localización y seguimiento de locutores).
- Desarrollar un nuevo sistema de experimentación para la localización basada en información acústica.
- Mejorar el sistema de experimentación basado en visión que se va a emplear.
- Diseñar e implementar un sistema de localización de hablantes basado en **fusión audiovisual**.
- **Evaluar** los algoritmos de localización y seguimiento implementados.



1. Introducción
2. **Estudio teórico**
3. Desarrollo
4. Resultados obtenidos
5. Conclusiones
6. Líneas futuras



Estudio teórico

Localización basada en información acústica

6/35

- Propagación de las ondas → ondas esféricas

- **3 factores influyentes:**

- Atenuación

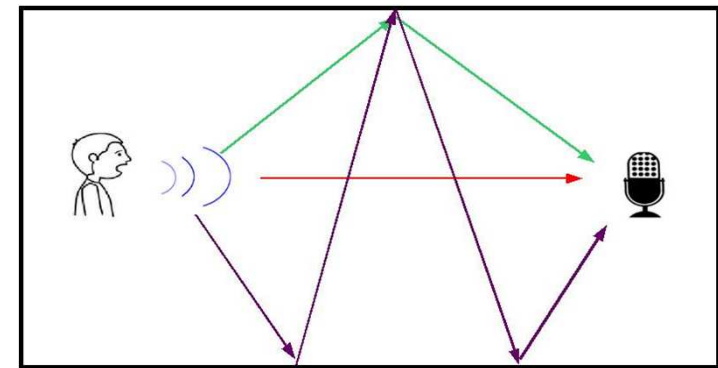
- Señal decrece con la distancia

- Ruido

- No direccional: “de fondo”
 - Direccional: ambigüedad en la estimación

- Reverberación

- Señal recibida contiene aportación de la señal original y numerosas copias (reflexión y difracción de las ondas)





Localización basada en información acústica

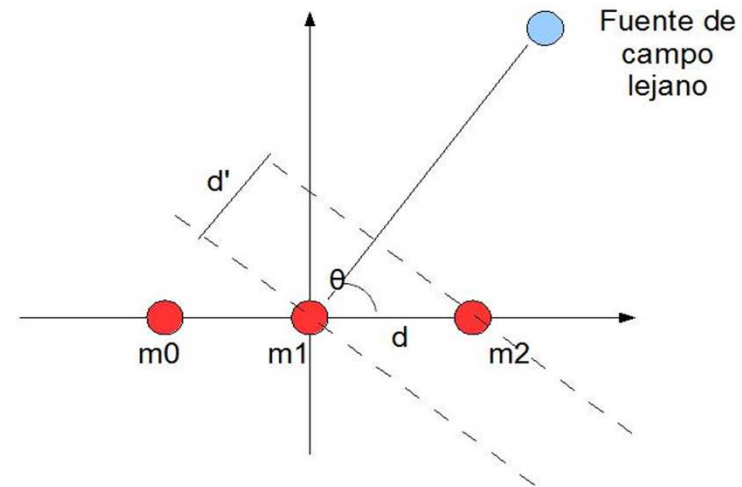
- Modelo de señal de un **micrófono** (entorno reverberante)

$$x_m(t) \cong \frac{a}{r_m} \cdot s(t - \tau_m) + \tilde{v}_m(t)$$

- Condición de **campo lejano**:

$$|r| > \frac{2R^2}{\lambda}$$

$$\Delta t = \tau_0 - \tau_1 = \frac{d'}{c} = \frac{d \cdot \cos \theta}{c}$$

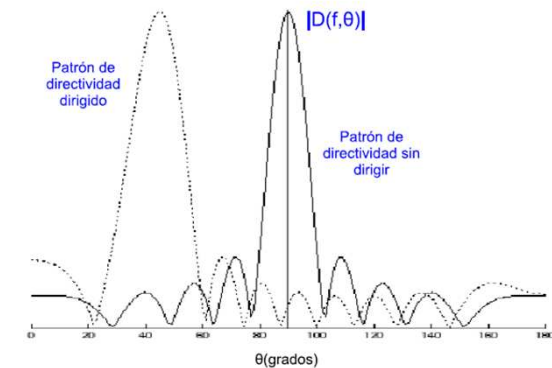


- Arrays de micrófonos**: mejoran la calidad de la señal

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



- Algoritmos de localización
 - Solución basada en la estimación de la diferencia del tiempo de llegada TDOA
 - Solución basada en **SRP**
 - Beamformer → dirigir el array
 - Maximum Likelihood → máximo pico de potencia
 - SRP-PHAT → elimina aportación amplitud



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

- SRP en términos de GCC (correlación cruzada generalizada)

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

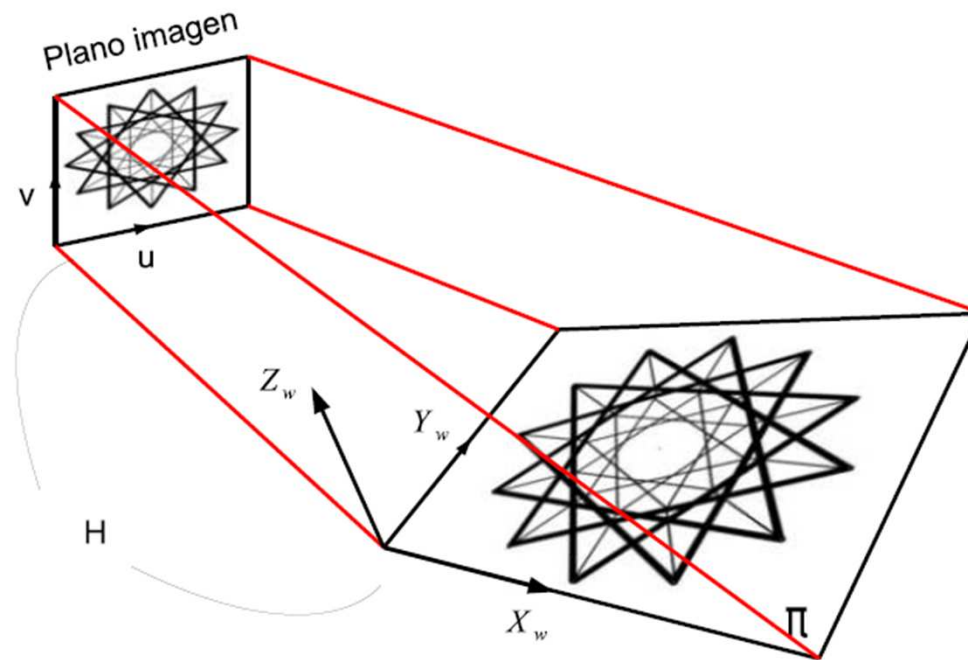


Estudio teórico

9/35

Localización basada en información procedente de visión

- Imágenes de entrada $\rightarrow$ **espacio de ocupación**
- Proyección de imágenes mediante matriz de homografía



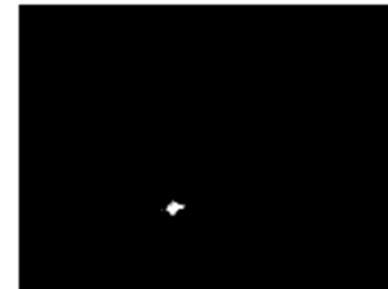


Estudio teórico

10/35

Localización basada en información procedente de visión

- **Generación de un grid** partiendo de múltiples cámaras





- **Filtro de Partículas:** modela el comportamiento mediante funciones de probabilidad multimodales.

$$\vec{x}_{t+1|t} = A \cdot \vec{x}_t + \vec{q}_t = \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ vx_t \\ vy_t \end{bmatrix} + \vec{q}_t$$

$$\vec{y}_t = H \cdot \vec{x}_t + \vec{r}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_t \\ y_t \\ vx_t \\ vy_t \end{bmatrix}$$

- Etapas:

- Predicción: set de partículas $\rightarrow$ distribución *a priori*
- Corrección: cálculo del peso de cada partícula
- Selección: set de partículas $\rightarrow$ distribución *a posteriori*





Estudio teórico

12/35

Localización basada en fusión audiovisual

- Sensores con debilidades y fortalezas → **fusión**
- Información visual no afectada por entorno acústico.



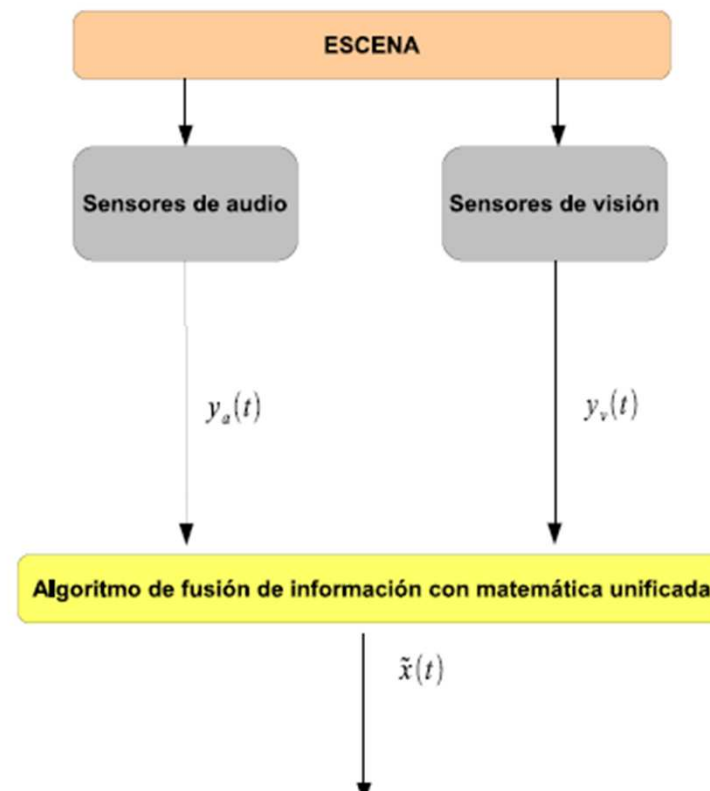
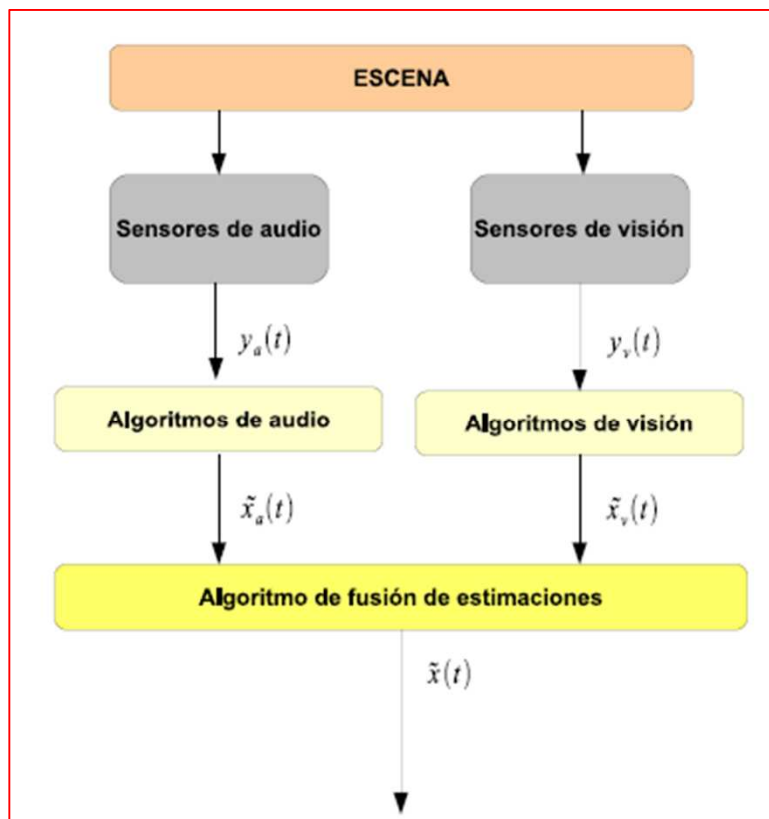
eleva prestaciones en sistemas de procesamiento
del habla

- 3 razones para añadir información visual a sistemas de localización basados en audio:
 - Ayuda a localizar al locutor
 - Naturaleza intermitente del habla
 - Resuelve ambigüedades



- Clasificación

- Orientados a sistema: combina estimaciones
- Orientados a modelo: estimación conjunta





1. Introducción
2. Estudio teórico
3. **Desarrollo**
4. Resultados obtenidos
5. Conclusiones
6. Líneas futuras



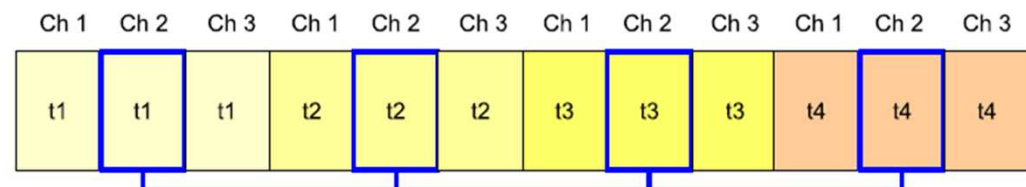
- Sistema SRP de partida con **limitaciones**
- Mejoras:
 - Ficheros de audio de entrada monocanal o multicanal (1 canal $\longrightarrow$ 1 micrófono)

- 1 canal $\longrightarrow$ 1 fichero

MONOCANAL



MULTICANAL



- Varios canales $\longrightarrow$ 1 fichero

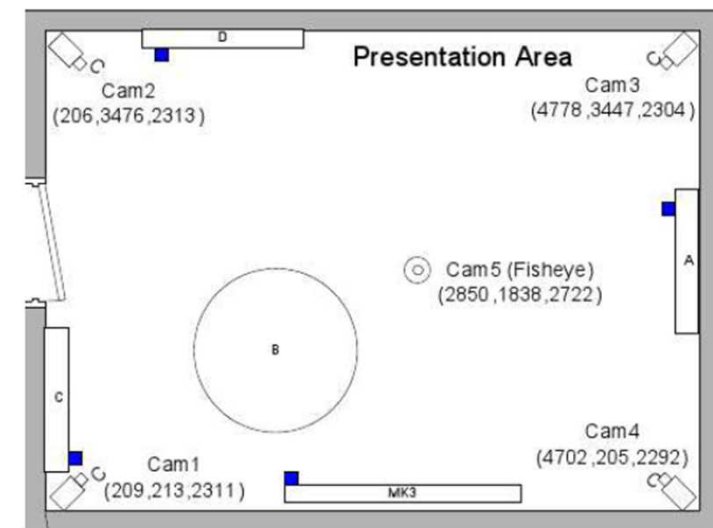


- Sistema SRP de partida con **limitaciones**
- Mejoras:
 - Agrupaciones de micrófonos para el cálculo de la potencia

$$P[q] = \sum_{i=0}^{i=n} P[i][q]$$

$$P[i][q] = \sum_{j=0}^{j=m} P[j][q]$$

	Sin subarrays	Con 3 subarrays
Pcor	30,0 ± 7,7 %	59,0 ± 8,3 %
Rel. error reduction		96,7 %
Bias fine (x:y:z) [mm]	-73 : 92 : -91	-13 : -27 : -102
Bias fine+gross (x,y,z) [mm]	-268 : -78 : -184	-47 : -142 : -146
AEE fine [mm] = MOTP	293	233
Rel. AEE reduction		20,5 %
Fine+gross [mm]	1017	585
Rel. BIAS f+g reduction		42,5 %
Loc. frames	135	135
Ref. duration (s)	309,0	309,0





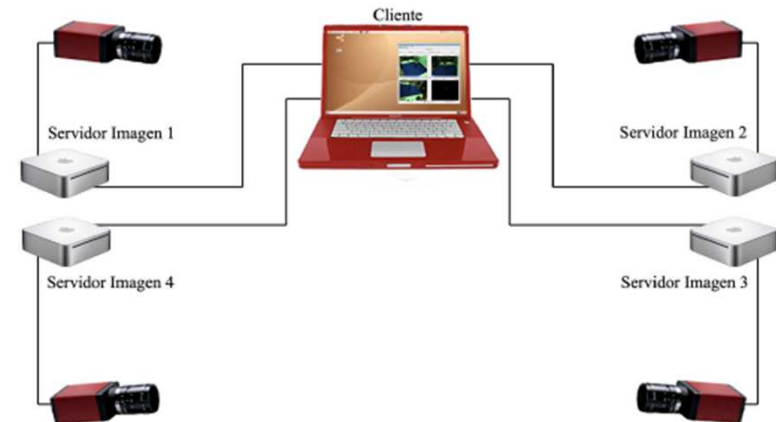
- Lanzar experimento → **tarea tediosa y manual**
- Necesidad de mejorar automatizando el sistema
- Objetivos logrados:
 - Sistema de experimentación sencillo
 - Sistema automático
 - Scripts de configuración
 - Ejecución sencilla de SRP y la evaluación
 - Generación de logs
 - Facilitación de adición de nuevas bases de datos



tarea sencilla y automática



- Arquitectura **cliente-servidor**
- Existencia de limitaciones
- Adición de mejoras y nuevas funcionalidades:
 - Parámetros de entrada
 - Cálculo automático de la matriz de homografía
 - Funcionamiento en tiempo real y modo simulación
 - Ejecución de varios servidores desde un mismo host
 - Inclusión de *groundtruth*, etiquetado de posiciones.
 - Envío de información cliente-servidor
 - Almacenamiento de salida con formato CLEAR



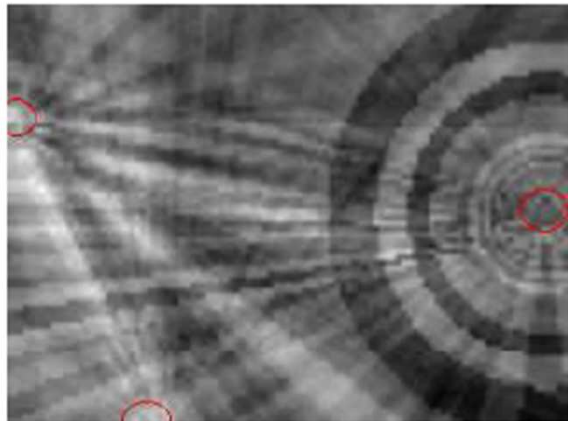
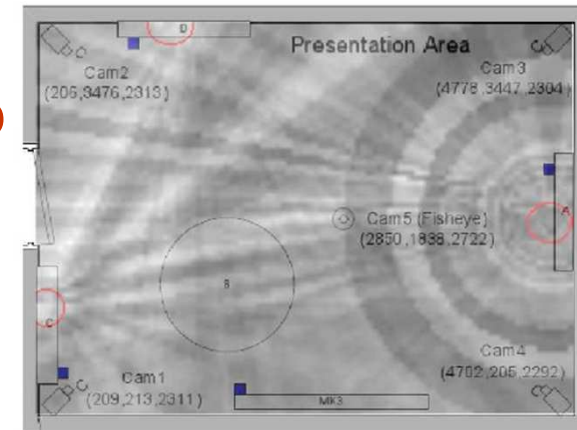


Desarrollo

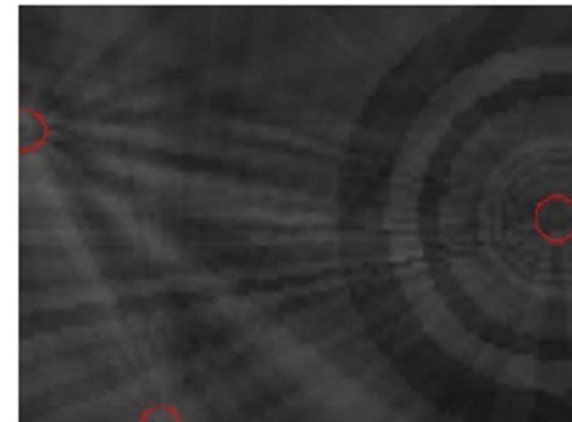
19/35

Sistema de localización basado en fusión audiovisual

- Arquitectura cliente-servidor: imagen de audio “procede de otra cámara”
- Formación de la **imagen de audio**
 - Fichero de potencias para cada localización



normalización





- Formación de la **imagen de audio**

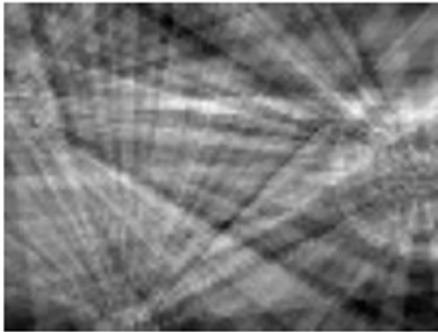


Imagen original



Cálculo de los máximos
Nivel de gris 250



GRID



Puntos máximos dilatados



Filtrado nivel de gris 200

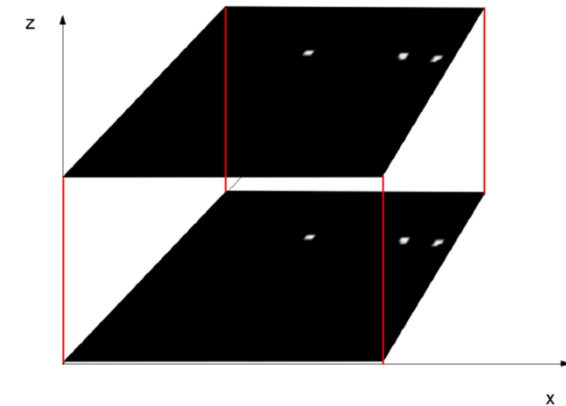


Desarrollo

21/35

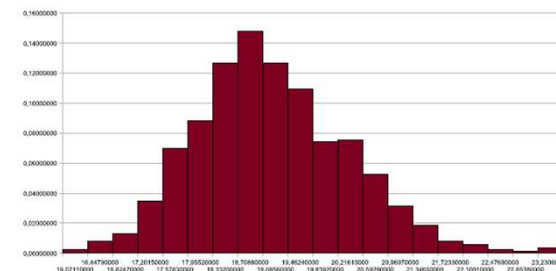
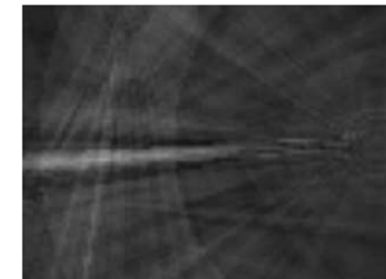
Sistema de localización basado en fusión audiovisual

- Aplicación de la matriz de homografía



- **Validación del grid de audio**

- Variabilidad del nivel de gris
- Cálculo de la desviación típica del nivel de gris en frames de background
- Validación en función de la desviación típica



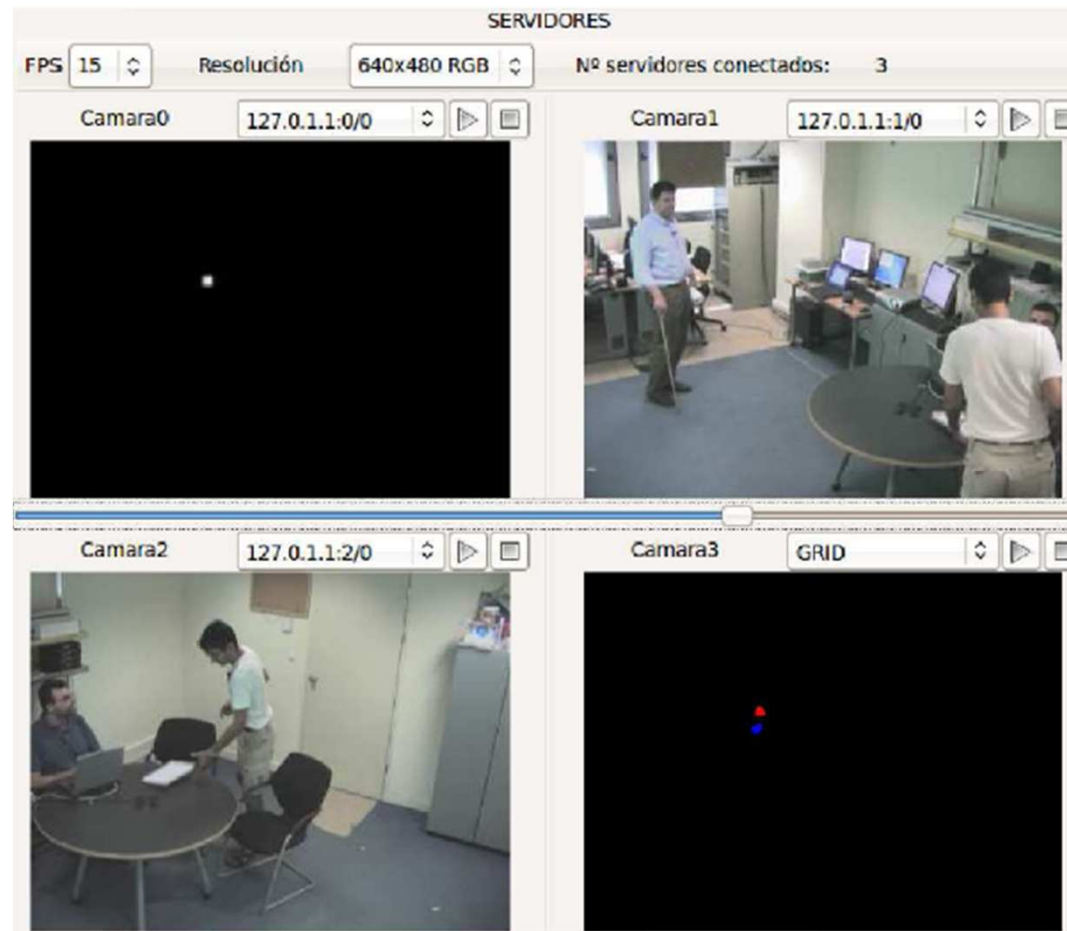


Desarrollo

22/35

Sistema de localización basado en fusión audiovisual

- Obtención del **grid audiovisual**
 - Unión de grid de audio + grid de vídeo
 - OR lógica





1. Introducción
2. Estudio teórico
3. Desarrollo
- 4. Resultados obtenidos**
5. Conclusiones
6. Líneas futuras



Resultados

24/35

Evaluación con diferentes bases de datos

- Evaluación de los algoritmos de localización y seguimiento implementados.
- Diferentes bases de datos audiovisuales:
CLEAR 2006/2007 (AIT, IBM , UPC, UKA...), Av-16.3, HIFI.
- **Métricas CLEAR** de evaluación
 - Pcor
 - Bias fine
 - Bias fine+gross
 - AEE Fine = MOTP
- Evaluación de resultados de sistemas de localización basados en **audio** y visión de forma **manual** o a través de la evaluación CLEAR.

$$P_{cor} = \frac{n^{\circ} \text{ error fine}}{n^{\circ} \text{ frames con estimación}}$$

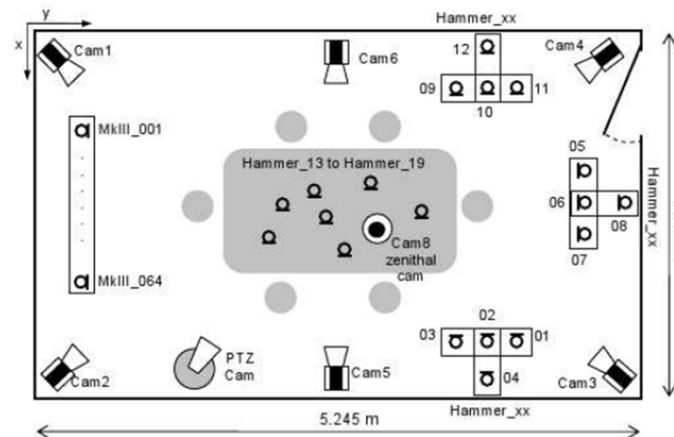


Resultados

Sistema de localización acústica

25/35

- Comparación con resultados obtenidos por institutos



	UKA	ITC	AIT	UPC	IBM
Pcor	67,0 ± 4,2 %	55,0 ± 2,3 %	65,0 ± 2,6 %	68,0 ± 2,1 %	63,0 ± 2,2 %
Bias fine (x:y:z) [mm]	-13 : -31 : -46	-4 : -45 : -57	-5 : -30 : -107	36 : 50 : -34	96 : -46 : -62
Bias fine+gross (x,y,z) [mm]	57 : -92 : -82	128 : -161 : -138	-9 : -81 : -163	75 : -80 : -139	470 : -200 : -100
AEE fine [mm] = MOTP	172	151	226	159	236
Fine+gross [mm]	613	643	620	628	904
Loc. frames	476	1822	1273	1819	1891
Ref. duration (s)	2398,0	2470,0	2364,0	2481,0	2427,0



Resultados

26/35

Sistema de experimentación basado en fusión audiovisual

- Evaluación de los resultados:
 - Manera visual
 - Evaluación CLEAR
- Base de datos AIT → 5 cámaras
(AIT_20060728) 3 arrays de 4 micrófonos
1 array Mark III 64 canales
- Empleo de etiquetados de posiciones reales



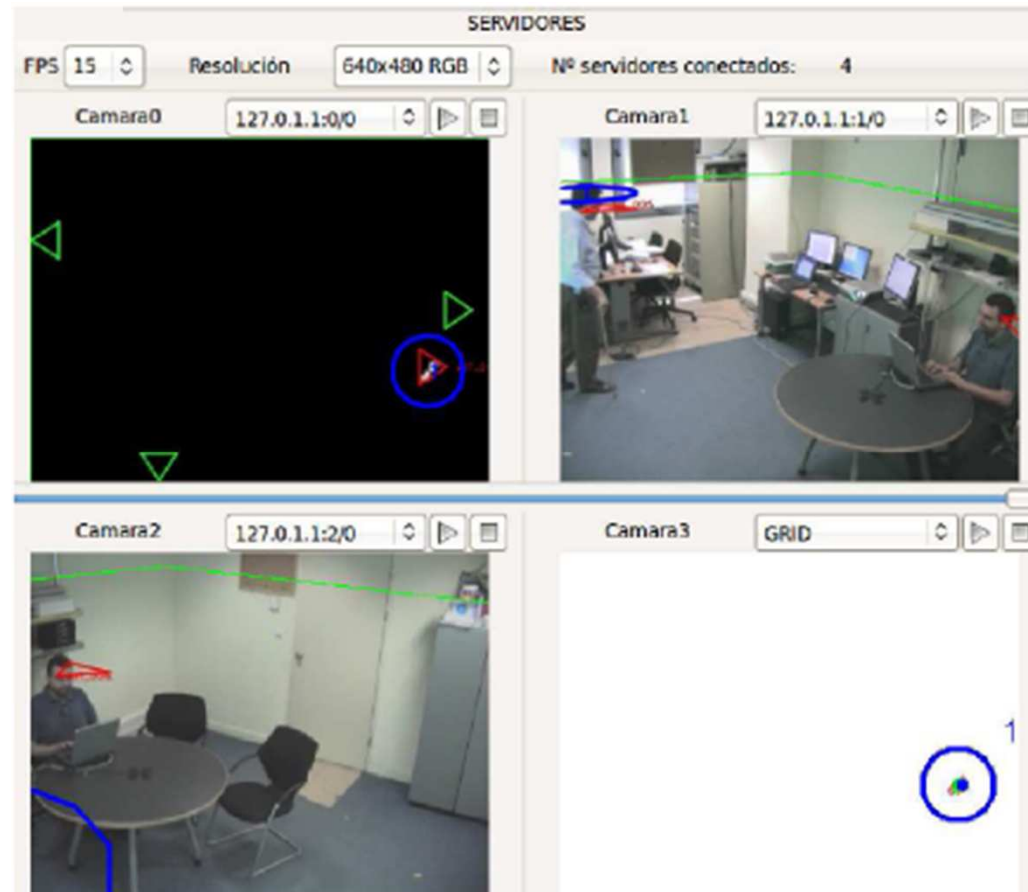


Resultados

27/35

Sistema de experimentación basado en fusión audiovisual

- Persona localizada gracias al grid de audio



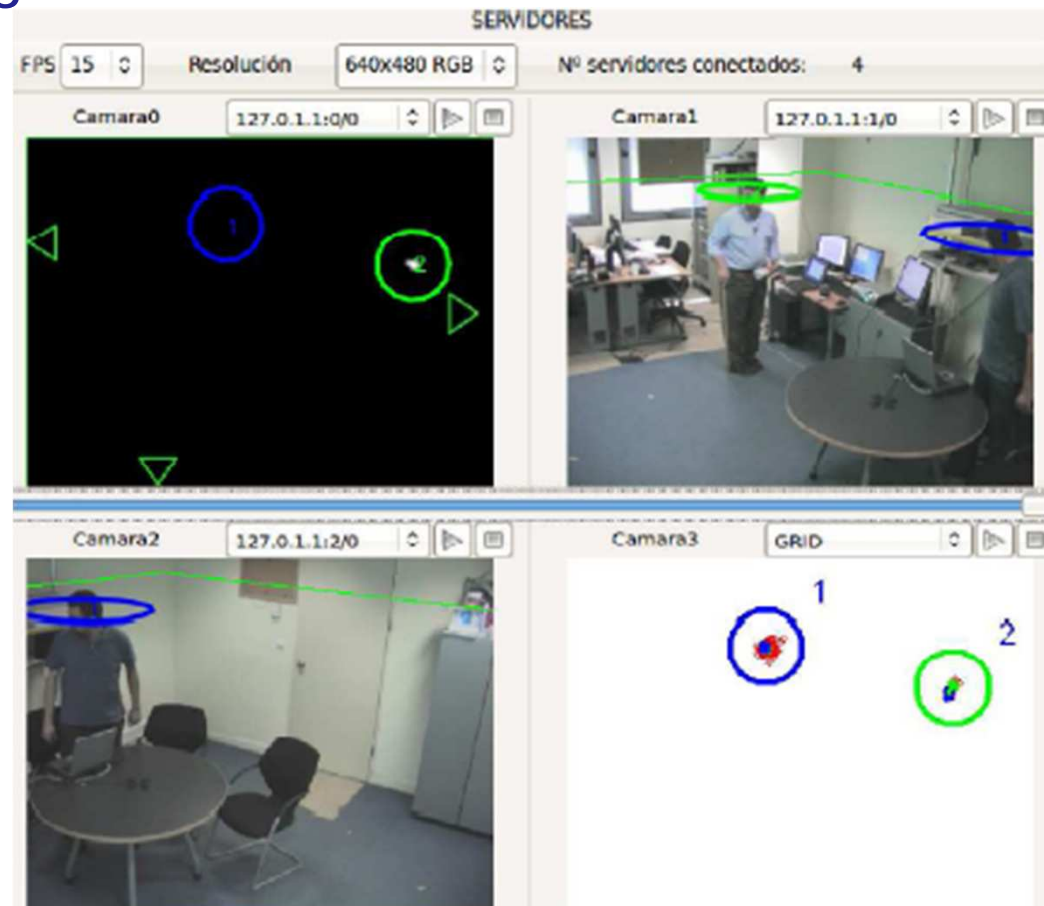


Resultados

28/35

Sistema de experimentación basado en fusión audiovisual

- Una persona localizada gracias al grid de audio y otra gracias al grid de vídeo



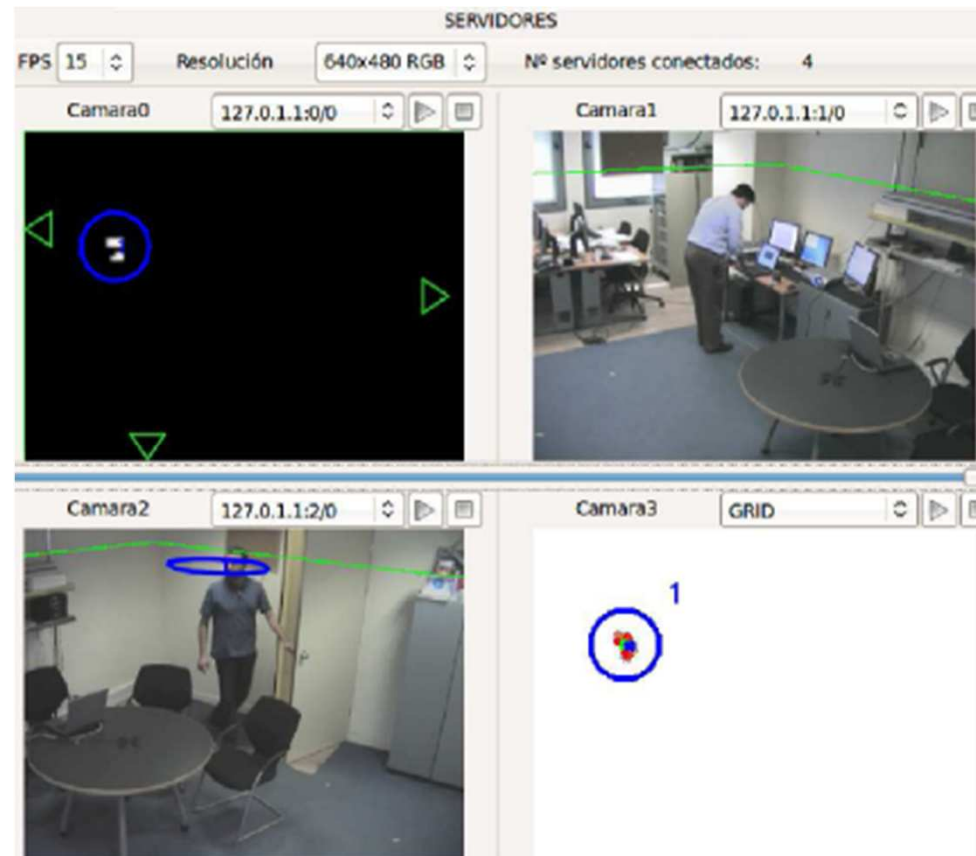


Resultados

29/35

Sistema de experimentación basado en fusión audiovisual

- Ruido validado que provoca la validación errónea de una persona



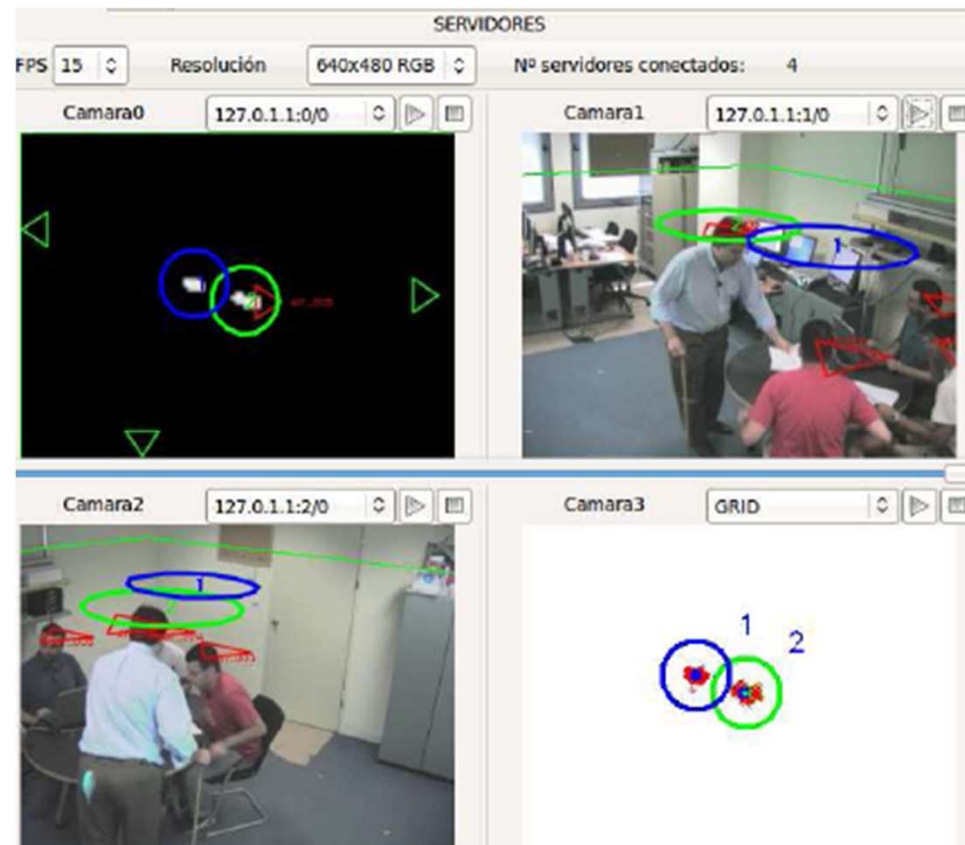


Resultados

30/35

Sistema de experimentación basado en fusión audiovisual

- Error en la localización de locutores y personas sin localizar





1. Introducción
2. Estudio teórico
3. Desarrollo
4. Resultados obtenidos
- 5. Conclusiones**
6. Líneas futuras



- Sistema de localización acústica mejorado, obteniendo mejores resultados en evaluación CLEAR.
- Sistema de experimentación de procesamiento de audio versátil
- Sistema de experimentación basado en visión mejorado
- Nuevo sistema de estimación de la posición de objetos basado en **fusión audiovisual**
- Difícil generación de grid a partir de archivos de potencia acústica
- Necesidad de discriminar ruido de voz humana



1. Introducción
2. Estudio teórico
3. Desarrollo
4. Resultados obtenidos
5. Conclusiones
- 6. Líneas futuras**



- Ajuste de parámetros de SRP para cada base de datos.
- Validación de la estrategia de obtención de grid de audio.
- Validación de la estrategia que permite validar el grid de audio.
- Fichero de configuración de parámetros.
- Discriminación de potencia de ruidos.
- Prueba del algoritmo de fusión con más bases de datos.
- **Evaluación** de los resultados de la fusión mediante técnicas CHIL.
- Asignación de pesos a partículas de audio o vídeo.
- Implementación en **tiempo real**.



Fin

35/35

Trabajo Fin de Carrera

“Diseño, implementación y evaluación de un sistema de locutores basado en fusión audiovisual”

¡¡GRACIAS POR SU ATENCIÓN!!