Tesis de Máster

Estudio, implementación y evaluación de un sistema de localización de locutores basado en el modelado de arrays de micrófonos como cámaras de perspectiva

Alejandro Legrá Rios

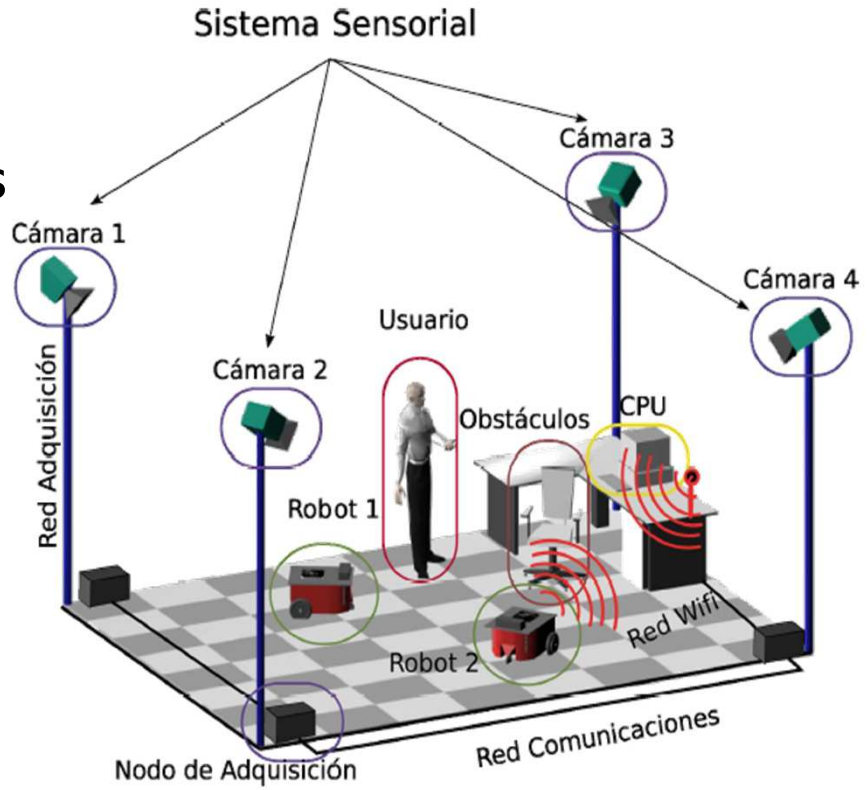
Directores: Javier Macías Guarasa & Daniel Pizarro Pérez

Universidad de Alcalá. Departamento de Electrónica. GEINTRA
Máster Oficial en Sistemas Electrónicos Avanzados. Sistemas Inteligentes

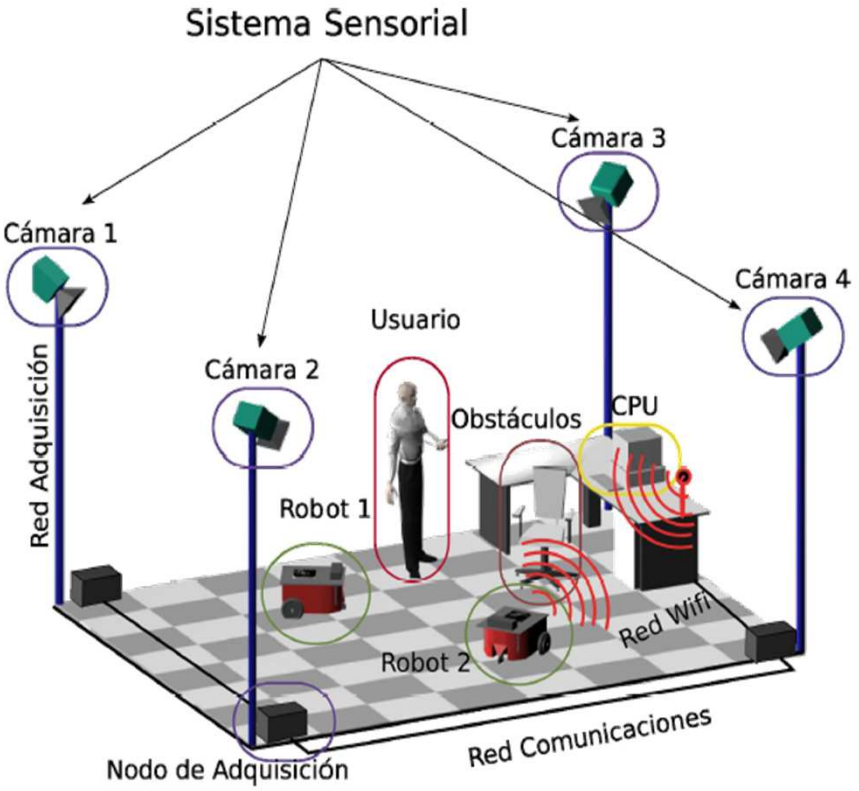
Sumario

- Presentación
- Estudio teórico
- Desarrollo
- Resultados
- Conclusiones
- Líneas futuras

- Análisis automático de los espacios inteligentes a partir del procesamiento de múltiples sensores
- Importancia de la detección, localización y seguimiento de personas en espacios inteligentes
- Trabajos de Fusión de señales de audio y de video para mejorar la interacción en el entorno



- Varios *arrays* de micrófonos
  - Posibilidad de usar SRP
  - Cada *array* de micrófonos modelado como una cámara
  - Una imagen por *array* Información relacionada con la potencia acústica "vista" por el *array*
  - Aplicar técnicas de visión para la localización



- Diseñar e implementar un sistema de generación de imágenes, a partir de información acústica
- Desarrollar algoritmos de tratamiento de imágenes, para la localización de hablantes
- Evaluar los algoritmos implementados, sobre las bases de datos multimodales disponibles en GEINTRA
- Estudiar los efectos de distintas resoluciones del espacio de búsqueda en los resultados obtenidos
- Estudiar los efectos de la eliminación de los errores en los máximos encontrados

- Modelado de *arrays* de micrófonos como cámara

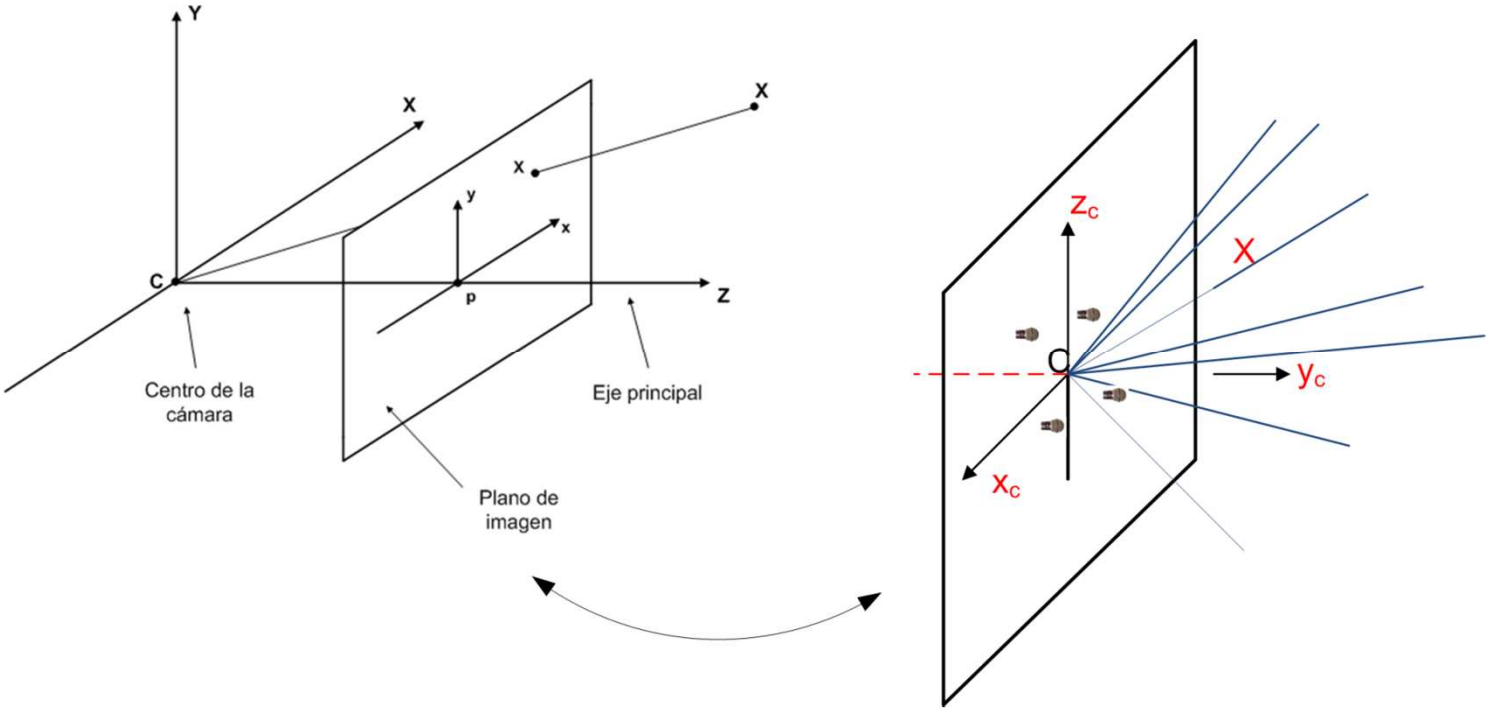
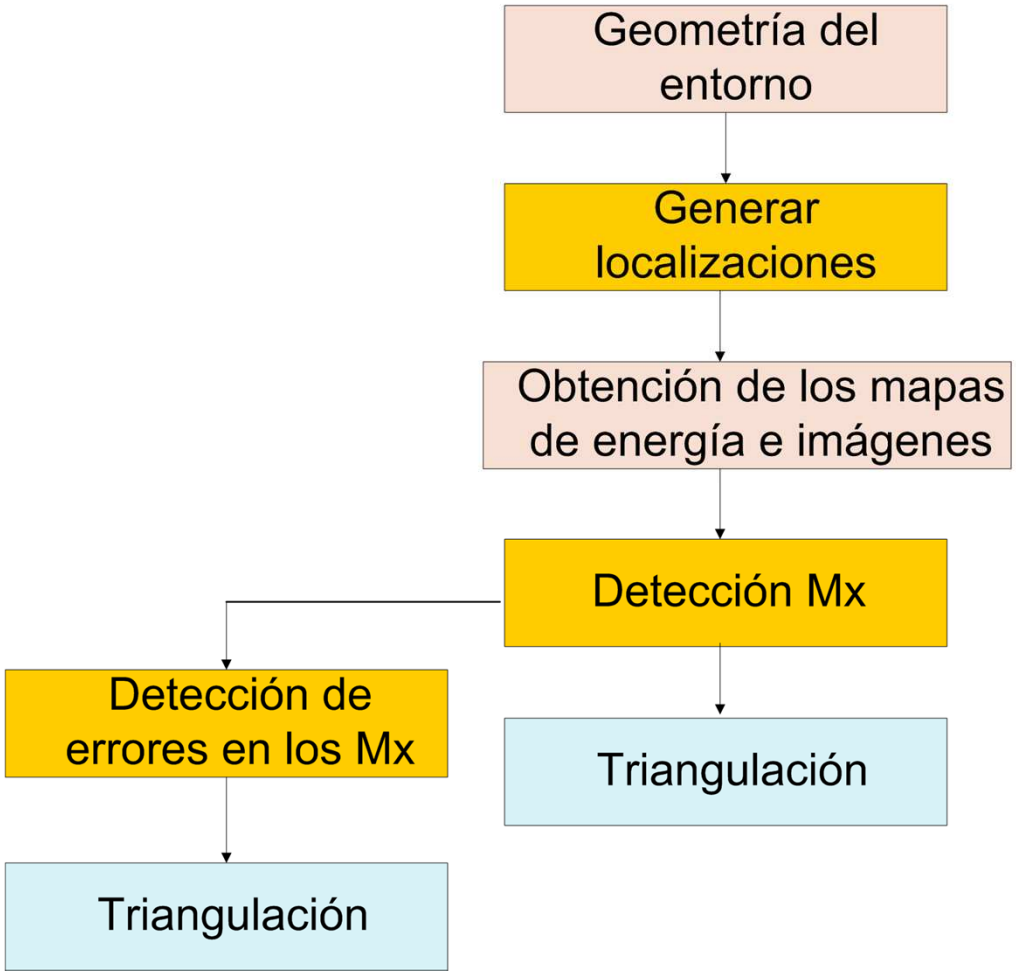
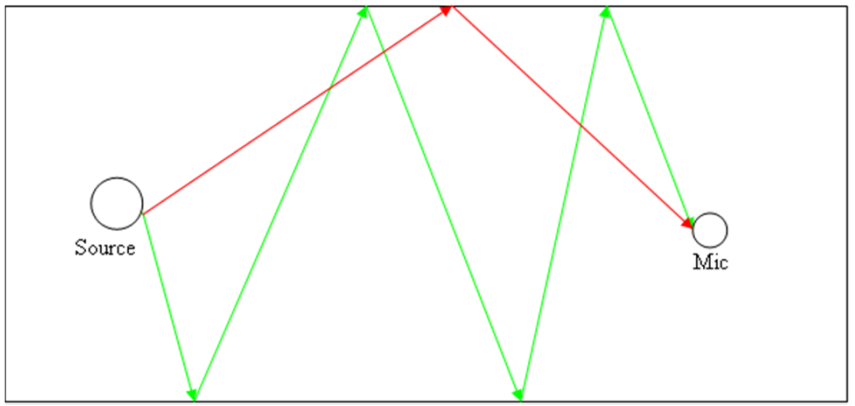


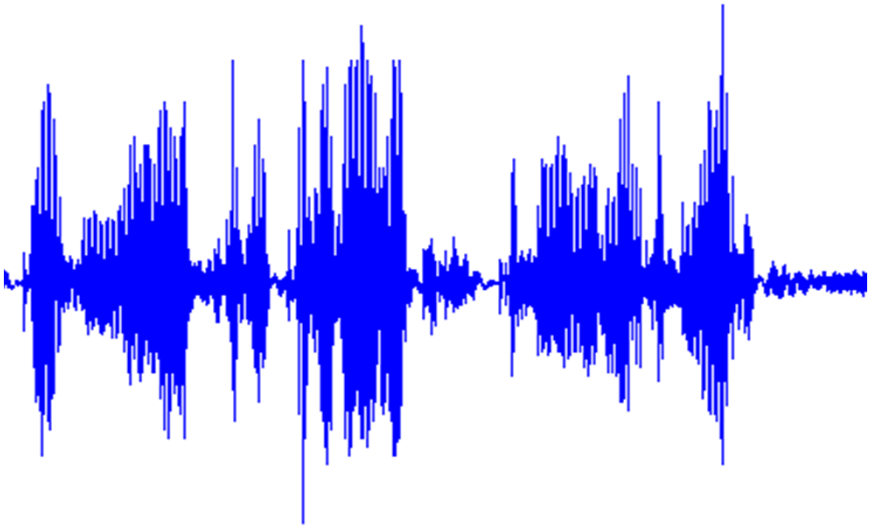
Diagrama General



❑ Problemáticas

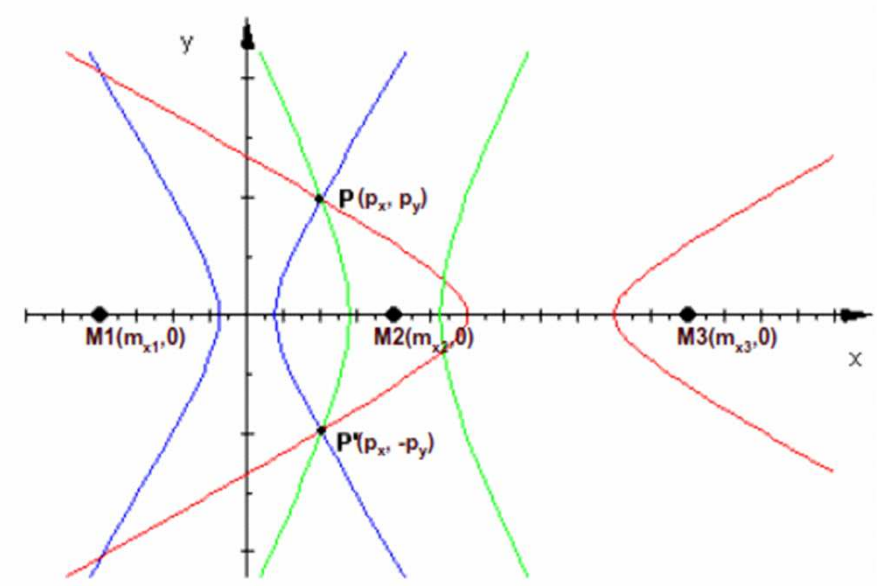
- Entornos reverberantes
- Baja SNR, debido a la distancia y ruido de fondo
- Señal de voz de banda ancha e intermitente
- Conversaciones humanas muy dinámicas

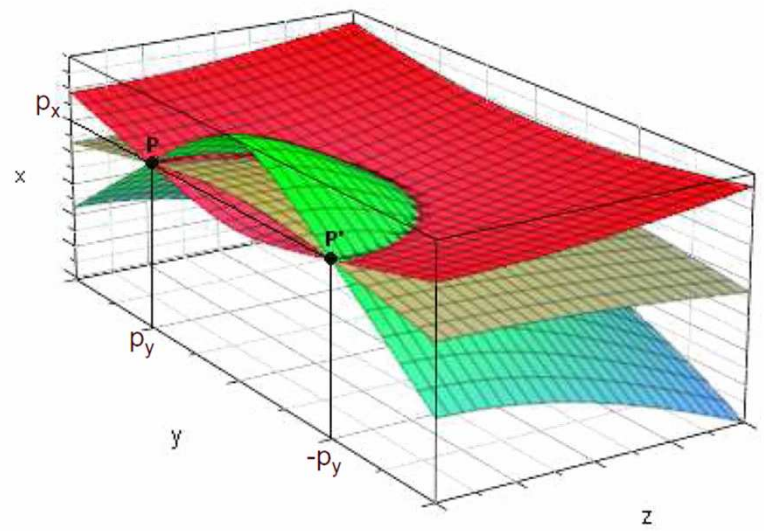




- Variantes
  - TDOA (Time Different of array). Problemas en entornos reverberantes
    - CC
    - GCC
    - Mejora con filtro PHAT: $\Phi_{x_i x_j}^{PHAT}(\omega) = \frac{1}{|C_{x_i x_j}(\omega)|} = \frac{1}{|X_i(\omega)X_j'(\omega)|}$

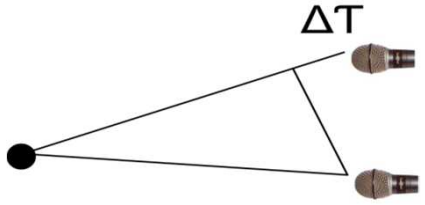
$$c_{x_i x_j}^{(g)}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_{x_i x_j}(\omega) X_i(\omega) X_j'(\omega) e^{-j\omega\tau} d\omega$$





- Basados en Steered Response Power (SRP)

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



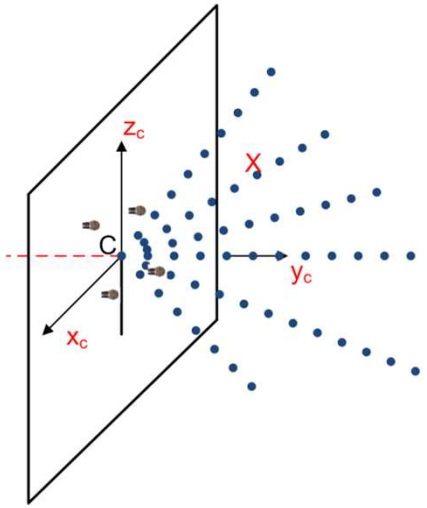
$$P(\mathbf{q}) = \int_{-\infty}^{\infty} |Y(\omega)|^2 d\omega = \int_{-\infty}^{\infty} Y(\omega) Y'(\omega) d\omega$$

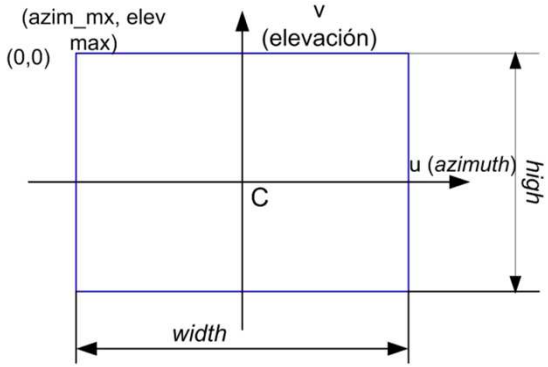
- SRP en función de GCC

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

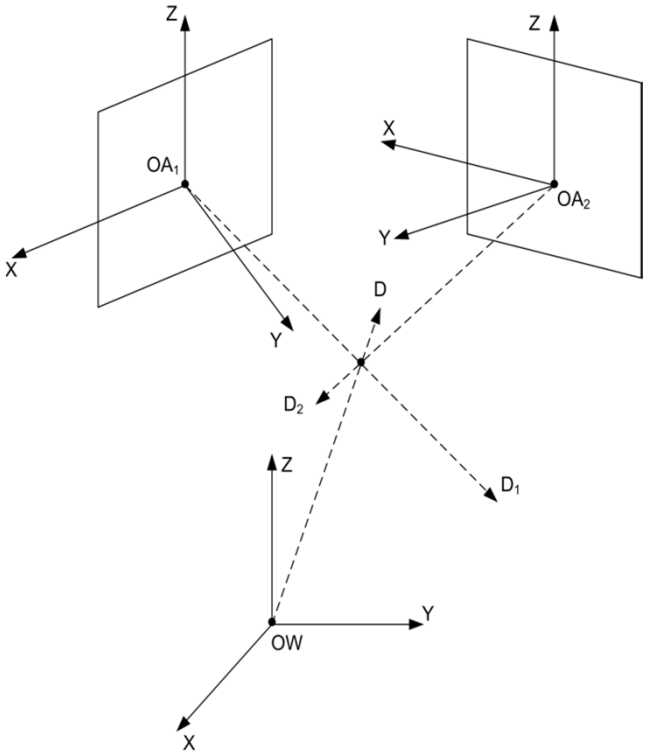
$$\hat{\mathbf{q}}_\mathbf{s} = arg\, max_q\, P(\mathbf{q})$$

Estudio Teórico
Formación de la imagen y triangulación





❑Triangulación



$$R_1 = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad T = \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix}$$

$$D^{OA_1} \times X^{OA_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad X^{OW} = \left(A^T A\right)^{-1} A^T B$$

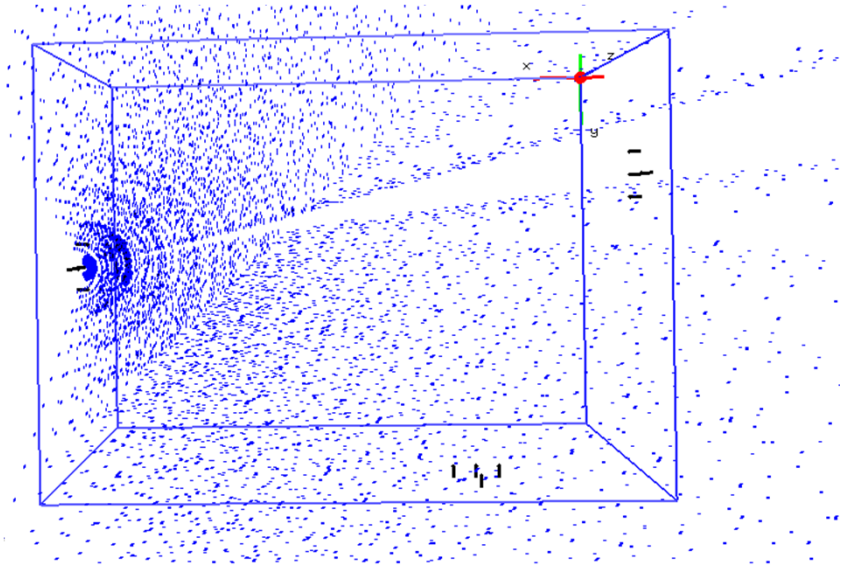
$$D^{OA_1} \times \left(R_1 \cdot [X^{OW} - T^{OA_1}]\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad A_1 X^{OW} = B_1$$

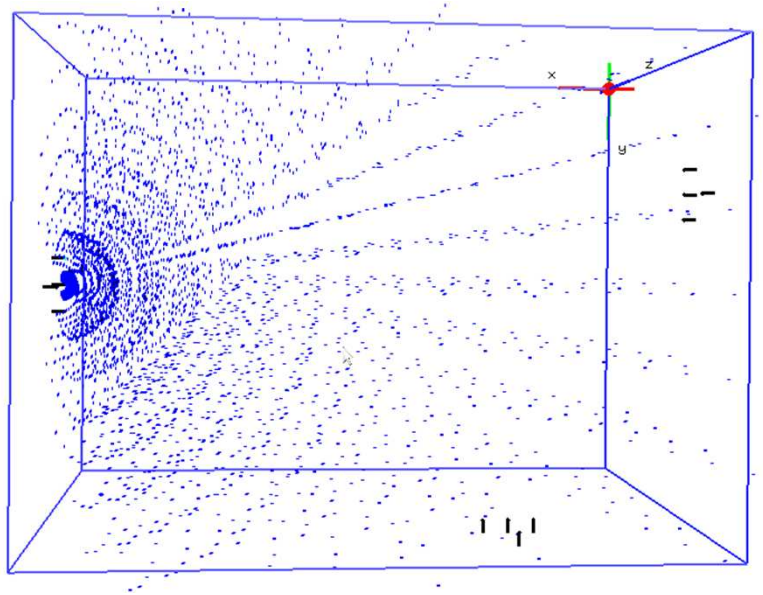
- Generación de localizaciones

$$d_{max} = \sqrt{x_{max}^2 + y_{max}^2 + z_{max}^2}$$

  - Generación en esférica

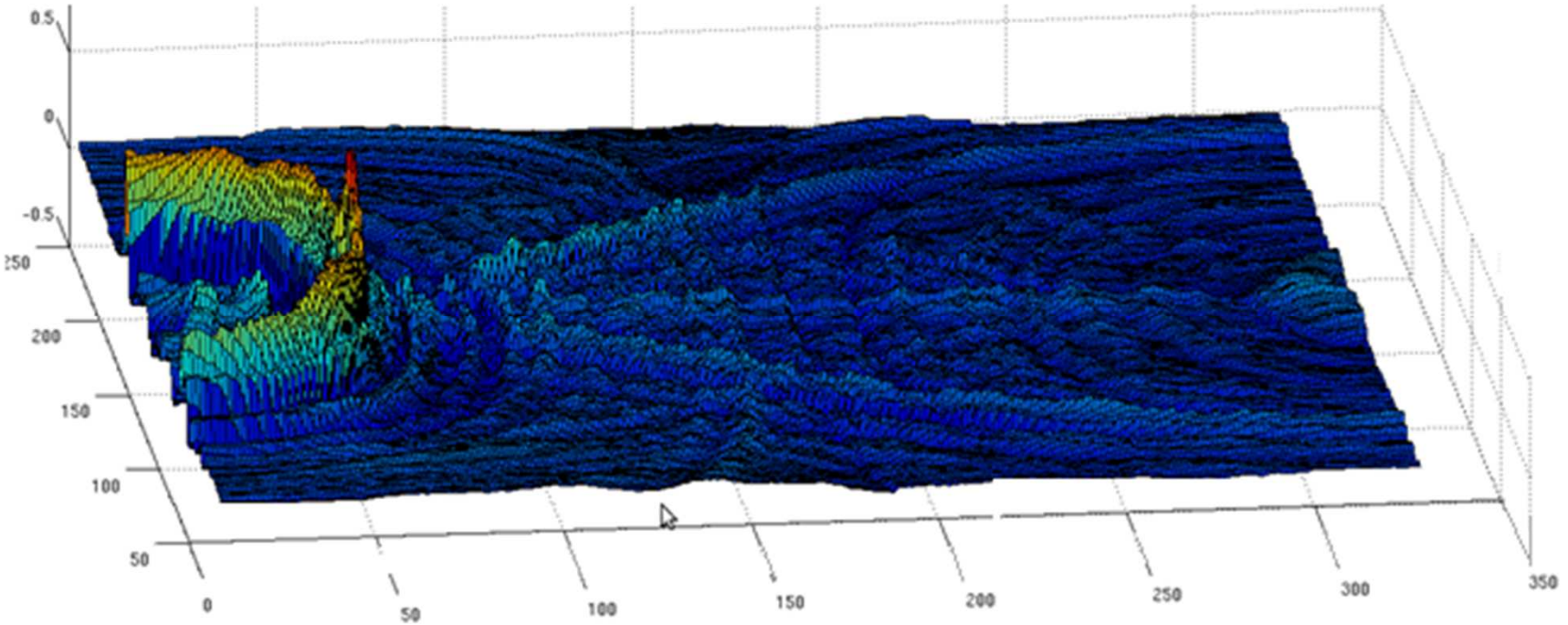
  - Generación no esférica. Límites del entornos



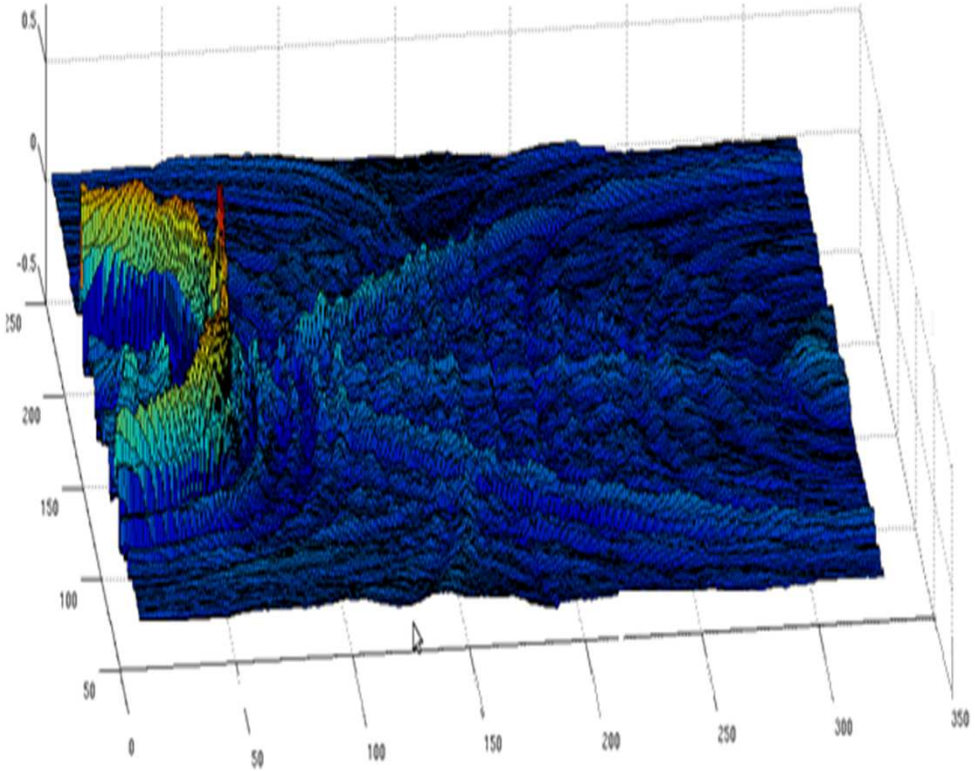


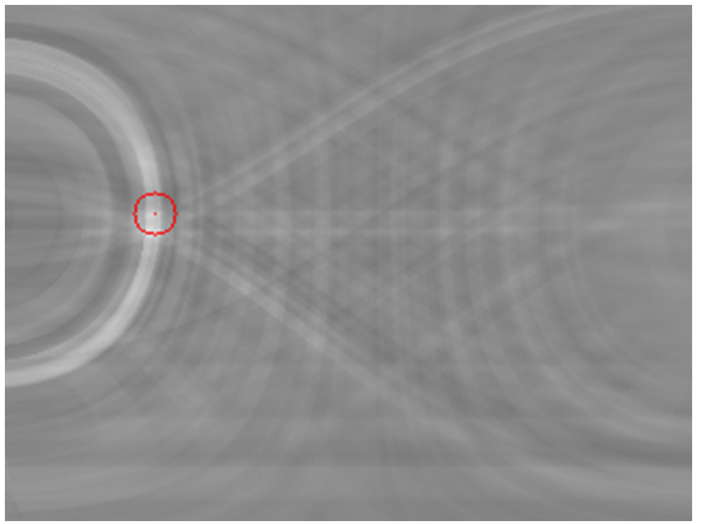
$$P_l = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \widehat{c}_{ij}(\tau)$$

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



- Problemas de escalado
- Uso del mapa de potencias

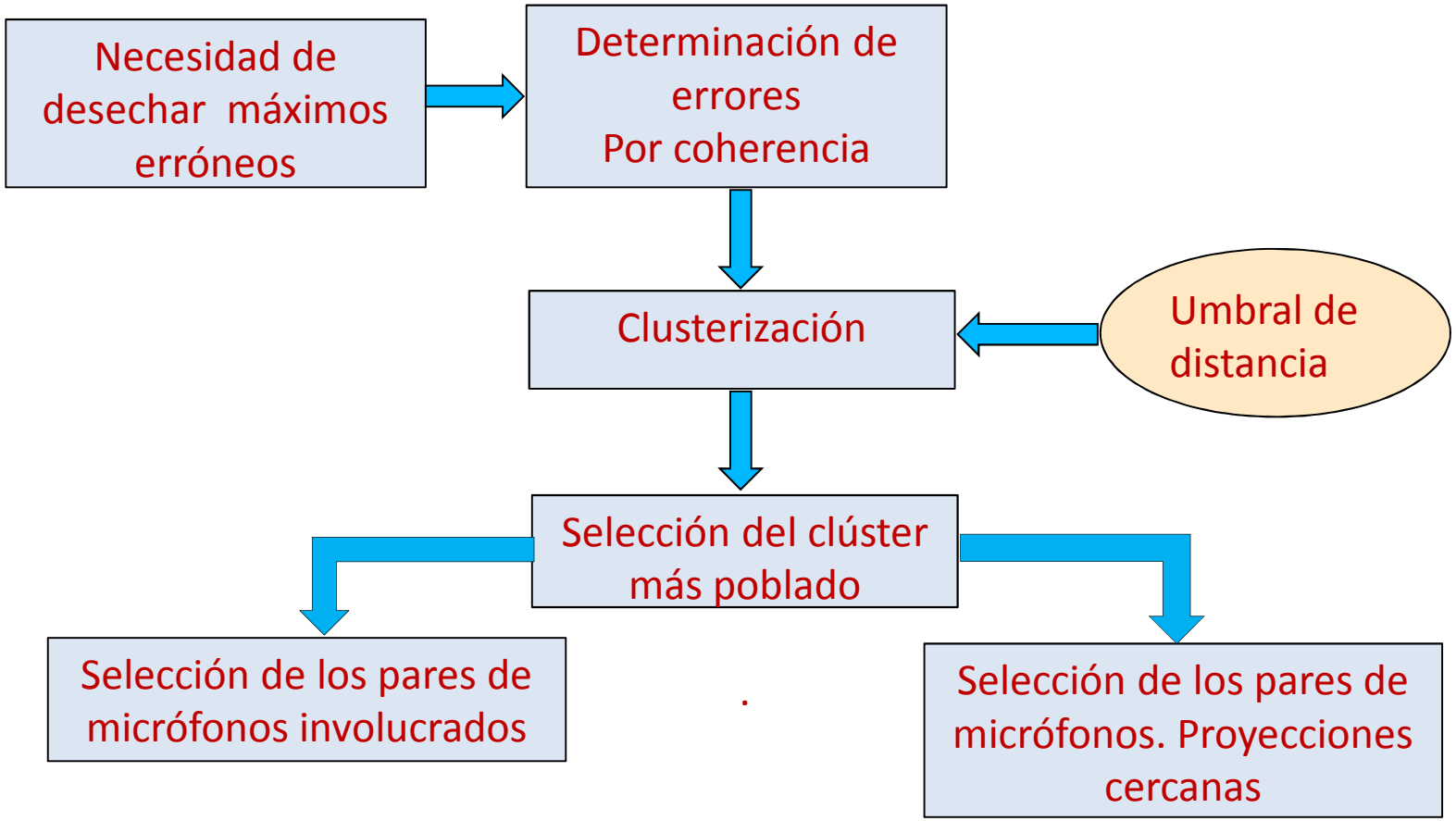




- Nom Maximun Supression
  - Umbralización
  - Radio de Vecindad
  - Posiciones enteras. Dentro de las imágenes

- Nom Maximun Supression con aproximación subpixélica
  - Forma de los máximos
  - Funciones cuadráticas, gaussianas ect..
  - Aproximación supixélica

Técnicas de estimación de coherencia



- Bases de datos del Proyecto CHIL (2007) en conjunto con la campaña CLEAR(2007)
  - Campaña de evaluación internacional común
- Campaña oficial CLEAR 2007
  - Evalúa varias tecnologías divididas en distintas áreas:
    - Visión: Detección y seguimiento 2D de rostro, seguimiento 2D de personas entre otos
    - Audio: Seguimiento de la persona que esta hablando.
    - Fusión de audio y video
- Está compuesto por 7 bases de datos de ellas se utilizan:
  - AIT e ITC

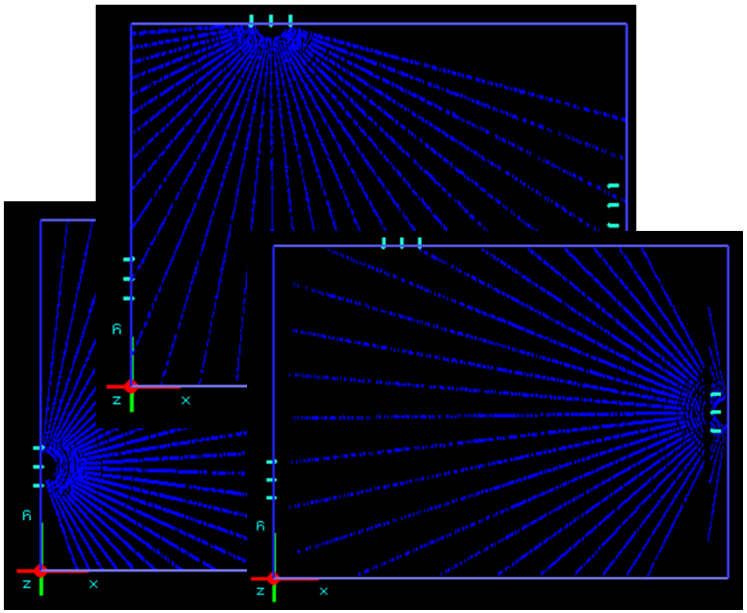
- Estudio del efecto de las localizaciones siguientes:
  - Efectos del barrido en ángulos
    - Efectos del proceso de búsqueda de máximos
    - Efectos del proceso de triangulación
    - Métricas globales
  - Efectos del barrido en profundidad
    - Efectos del proceso de búsqueda de máximos
    - Efectos del proceso de triangulación
    - Métricas globales
  - Efectos de la estrategia de generación de rayos
    - Efectos del proceso de búsqueda de máximos
    - Efectos del proceso de triangulación
    - Métricas globales
- Estudio del efecto de las variantes de localizaciones

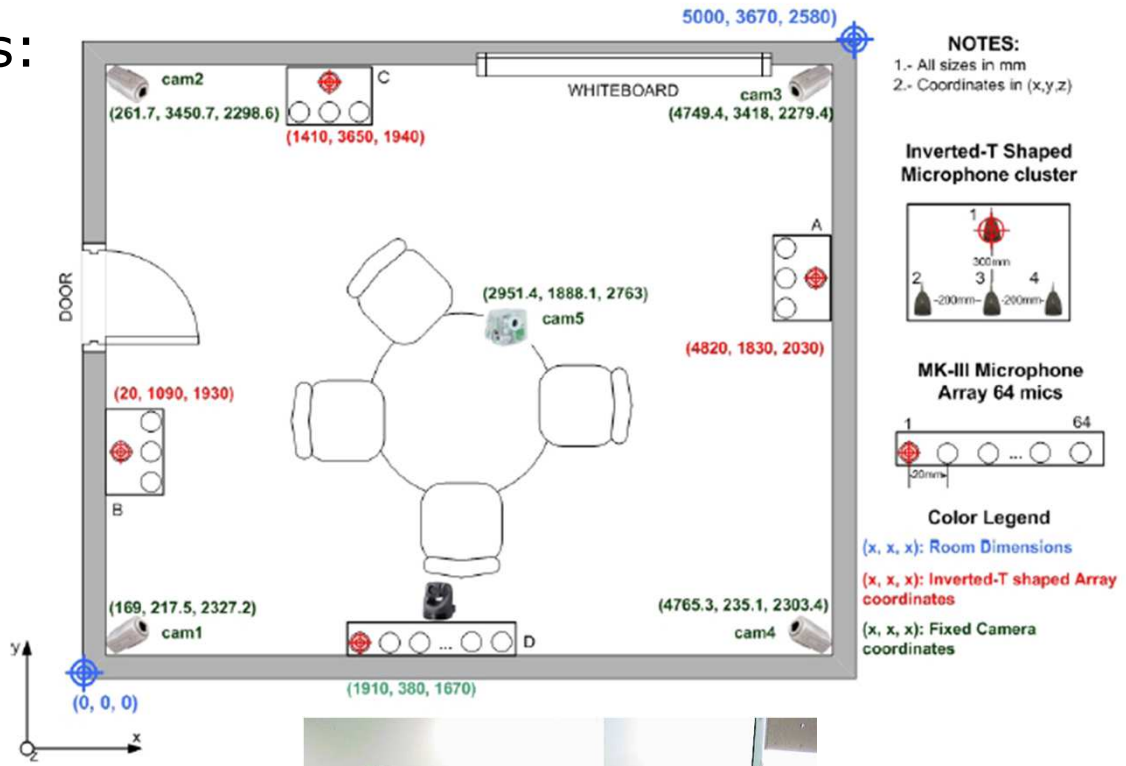
Resultados Experimentales
Métricas

- Pcor: Por ciento de *frames* del total de aciertos en los que el posicionamiento es menor de 50 cm
- Bias fine: Promedio de los errores cometidos que son menores de 50 cm
- Bias fine+gross: Promedio de todos los errores cometidos
- Deletion: Por ciento de *frames* en los que no se da estimación, encontrándose un locutor activo
- Media del error en grados. Promedio de los errores absolutos en la estimación de los máximos de potencia
- Error >20º. Por ciento de los errores mayores de 20º cometidos en la estimación de los máximos

Resultados Experimentales
Experimento base AIT

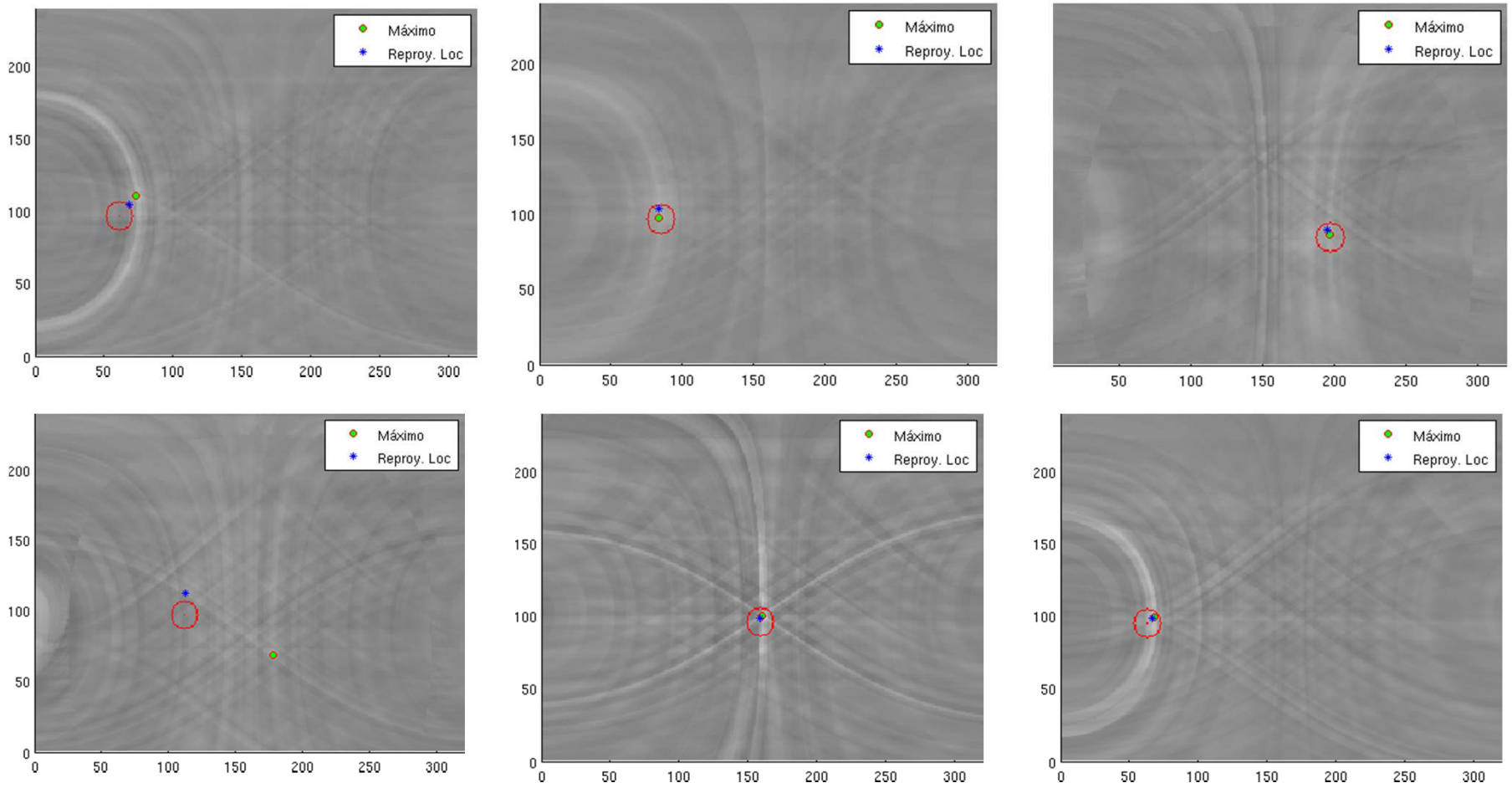
- Resolución de los Barridos:
  - Azimuth: 320 ptos
  - Elevación: 240 ptos
  - Δr = 100 mm
- Generación no esférica





Resultados Experimentales
Experimento base AIT

❑Imágenes Resultantes



Resultados Experimentales
Experimento base AIT

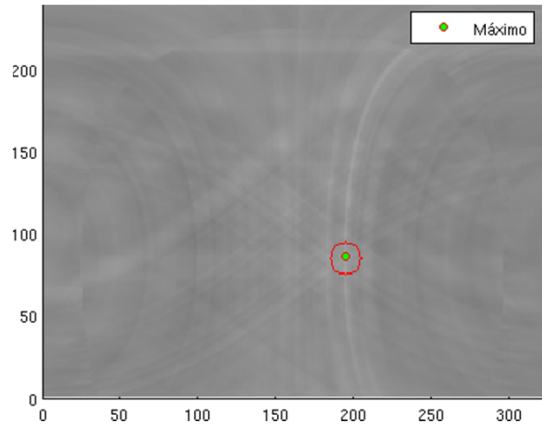
	Azimuth			Elevación		
Métricas	Array_0	Array_1	Array_2	Array_0	Array_1	Array_2
Media del error	11,70º	10,33º	11,82º	5,84º	4,63º	7,56º
error >20º	20,61 %	18,80 %	19,94 %	7,44 %	4,96 %	8,73 %

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

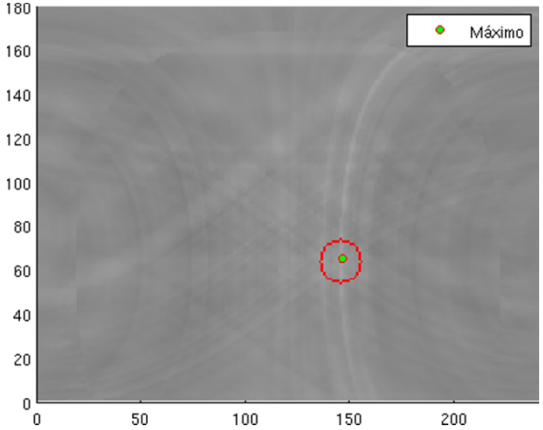
Resultados Experimentales
Estudio del efecto del barrido en ángulos

- Imágenes Resultantes

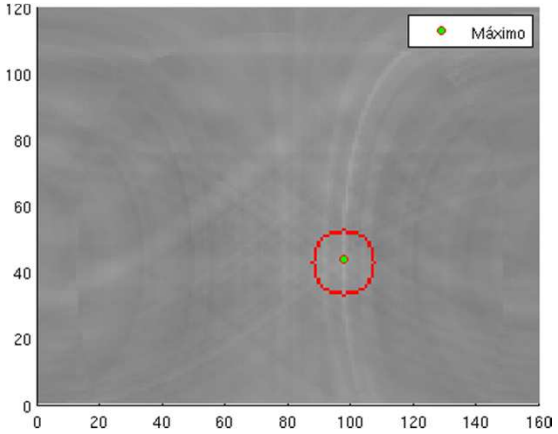
320 x 240



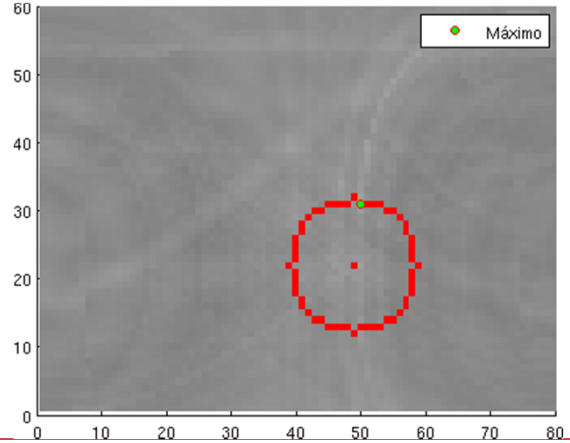
240 x 180



160 x 120



80x 60



Resultados Experimentales
Estudio del efecto del barrido en ángulos

Resumen de los errores cometidos en los máximos

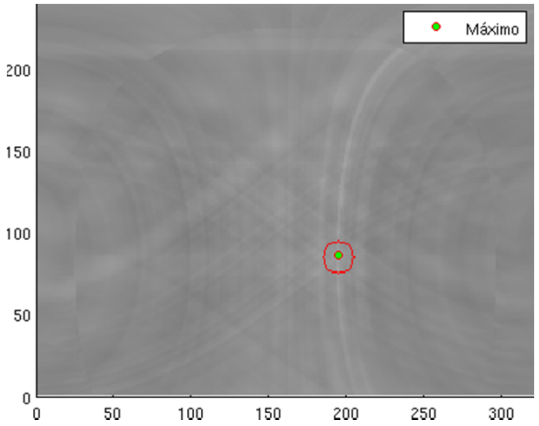
Métricas	320 × 240		240 × 180		160 × 120		80 × 60	
	azimuth	elevación	*azimuth*	elevación	*azimuth*	elevación	*azimuth*	elevación
Media del error	11,28º	6,01º	12,69º	5,87º	11,65º	6,21º	13,13º	7,00º
error >20º	19,78 %	7,04 %	22,41 %	6,47 %	20,90 %	7,75 %	24,17 %	10,66 %

Resultados globales para distintas resoluciones

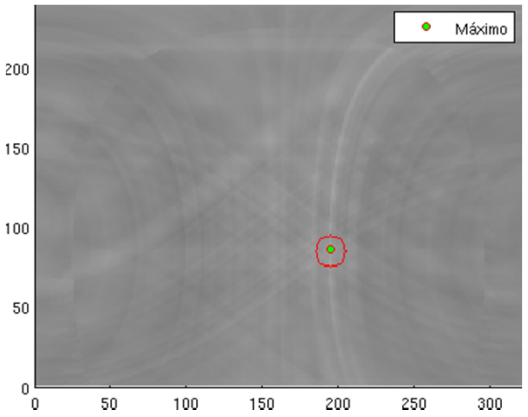
	320 × 240	240 × 180	160 × 120	80 × 60
Pcor	55,0 ± 2,7 %	56,0 ± 2,7 %	53,0 ± 2,7 %	50,0 ± 2,7 %
Rel. error reduction		1,8 %	−3,6 %	−9,1 %
Bias fine (x:y:z) [mm]	−1 : −29 : −118	−1 : −38 : −119	−4 : −37 : −118	−16 : −34 : −111
Bias fine+gross (x,y,z) [mm]	−23 : −59 : −171	−37 : −58 : −175	−83 : −92 : −173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

❑Imágenes resultantes

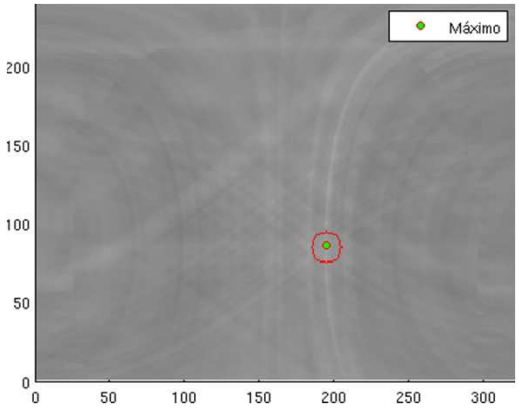
320 x 240 x 50



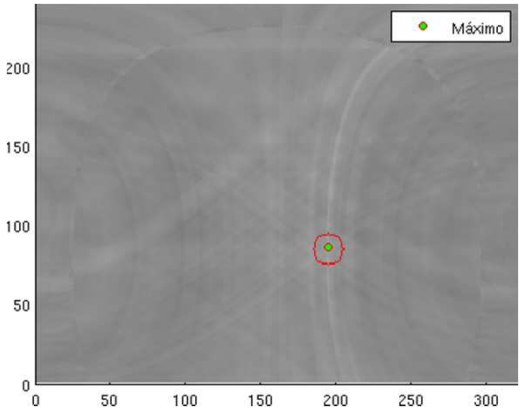
320 x 240 x 100



320 x 240 x 200



320 x 240 x 250



Resultados Experimentales
Estudio del efecto del barrido en profundidad

Resumen de los errores cometidos en los máximos

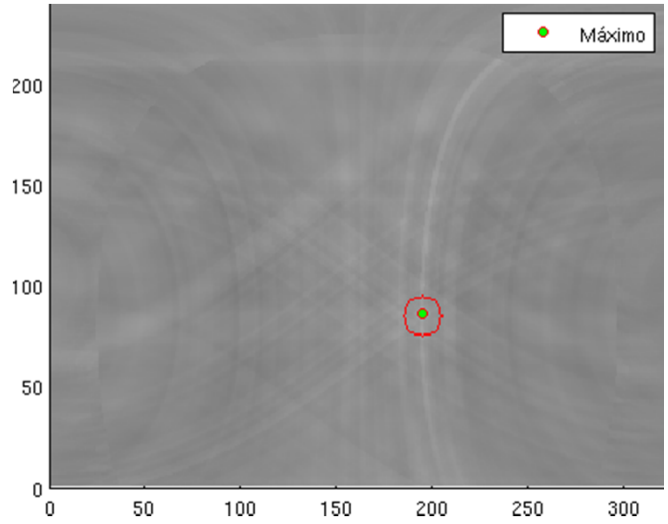
Métricas	320 × 240 Δr 50 mm		320 × 240 Δr 100 mm		320 × 240 Δr 200 mm		320 × 240 Δr 250 mm	
	azimuth	elevación	*azimuth*	elevación	*azimuth*	elevación	*azimuth*	elevación
Media del error	11,18°	5,98°	11,28°	6,01°	11,22°	6,06°	11,47°	6,15°
error >20°	19,60 %	6,99 %	19,78 %	7,04 %	19,51 %	7,17 %	19,85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

Resultados globales para distintas resoluciones

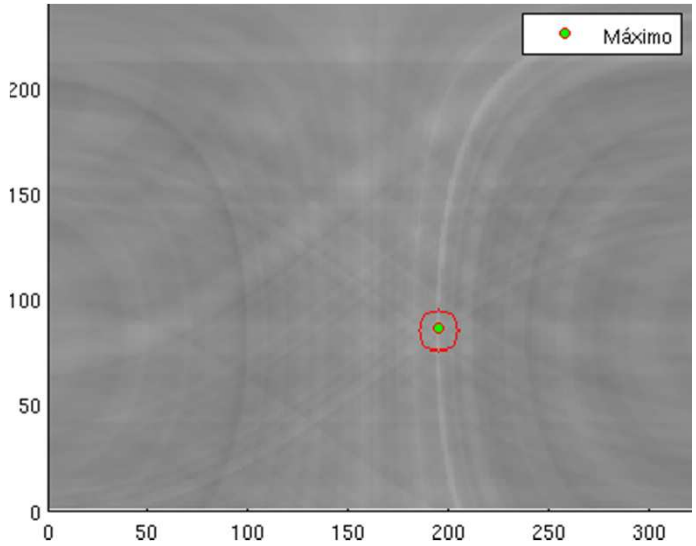
Métricas de CHIL	320 × 240 Δr 50 mm	320 × 240 Δr 100 mm	320 × 240 Δr 200 mm	320 × 240 Δr 250 mm
Pcor	55,0 ± 2,7 %	55,0 ± 2,7 %	56,0 ± 2,7 %	55,0 ± 2,7 %
Rel. error reduction		0,0 %	1,8 %	0,0 %
Bias fine (x:y:z) [mm]	−4 : −29 : −118	−1 : −29 : −118	4 : −38 : −113	−2 : −18 : −110
Bias fine+gross (x,y,z)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

- Resultados de las imágenes obtenidas

No esférica 320 x 240



Esférica 320 x 240



❑Métricas

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

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

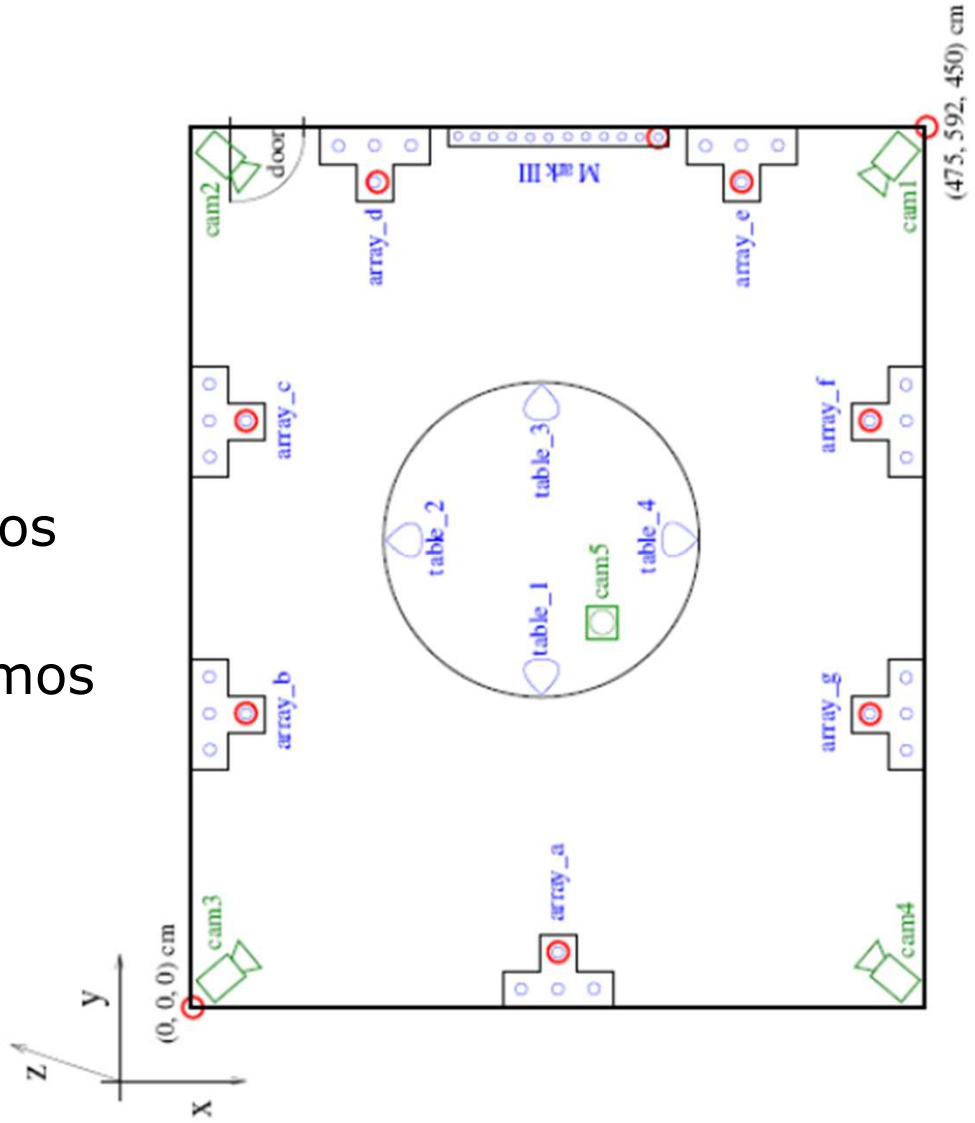
Resultados Experimentales
Experimento base ITC

- Resolución de los Barridos
  - Azimuth: 320 ptos
  - Elevación: 240 ptos
  - Δr = 100 mm

- Generación no esférica

- Posee 7 arrays de micrófonos

- Posibilidad de estimar máximos erróneos por coherencia



❑ Resultados obtenidos con las métricas generales utilizando y sin utilizar coherencia

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

Conclusiones

- Se ha diseñado implementado un sistema novedoso de localización basado en señales acústicas, a través del modelado de *arrays* de micrófonos como cámaras de perspectivas

- Se utilizó la técnica SRP-PHAT para calcular la potencia acústica en posiciones generadas en forma de rayos que parten desde el centroide

- Se obtuvieron imágenes con información de potencia acústica

- Aplicación del algoritmo Nom Maximun Supression con aproximación subpixélica con el objetivo de obtener los máximos locales

Conclusiones

- El sistema diseñado fue evaluado con las bases de datos de AIT e ITC pertenecientes al proyecto de CHIL bajo la campaña

- Se evaluaron los efectos de distintas resoluciones en las localizaciones, sobre las prestaciones del sistema implementado, en el cual se observó una degradación paulatina de las métricas de evaluación a medida que estas disminuían

- Evaluación del sistema de estimación de coherencia sobre la base de datos de ITC, consiguiéndose mejoras importantes en las prestaciones del sistema

Líneas Futuras

- Analizar la utilización de filtros a la señal de audio recibida por los micrófonos, con el objetivo de mejorar la respuesta del algoritmo SRP-PHAT

- Analizar la respuesta de SRP-PHAT, utilizando distintos valores de tamaño de ventana de análisis, tiempo en el cual se considera que el locutor no cambia su posición

- Trabajar en nuevas estrategias de definición del espacio de búsqueda, donde las pérdidas de resolución a distancias mayores del *array* de micrófonos sean menores

- Analizar la distribución de los niveles en los mapas de energía, con el fin de obtener imágenes, que permitan aplicar otras técnicas mas sofisticadas de tratamientos de imágenes, para la búsqueda de los máximos de energía de manera mas robusta

Líneas Futuras

- Experimentar con otras técnicas mas elaboradas de algoritmos de detección de coherencia a los máximos encontrados, con el objetivo de eliminar los máximos erróneos del sistema de triangulación

- Estudiar la influencia que tiene sobre los errores, la posición del locutor en el espacio

- Utilizar técnicas de seguimiento (tracking), con el objetivo de hacer un filtrado espacio temporal de los resultados

- Realizar la evaluación del sistema propuesto en nuevas bases