Seguimiento audiovisual de locutor usando un filtro de partículas extendido con proceso de clasificación

Frank Sanabria-Macías*, Javier Macías-Guarasa†, Marta Marrón-Romera†, Daniel Pizarro† y Enrique Marañón-Reyes*
*Grupo de Procesamiento de Voz, CENPIS — Universidad de Oriente, Santiago de Cuba — Cuba
email: fsanm77@fie.uo.edu.cu
†Departamento de Electrónica — Universidad de Alcalá — Alcalá de Henares — España
email: {macias,marta,pizarro}@depeca.uah.es

***Resumen*—En este trabajo se describe el diseño, implementación y evaluación de un sistema de seguimiento de locutores usando fusión audiovisual. La información de audio y vídeo es obtenida a partir de agrupaciones (*arrays*) de micrófonos y cámaras de vídeo situados en el entorno. El sistema propuesto está compuesto por dos bloques de extracción de la información de audio y vídeo respectivamente, y otro que fusiona esta información en un plano de ocupación paralelo y a una altura determinada del suelo. Un filtro de partículas que opera sobre la información fusionada permite obtener finalmente la localización estimada del locutor en cada instante de tiempo. Como bloque de extracción de la información de audio se usa un algoritmo de detección de actividad acústica por sectores (volúmenes cónicos alrededor de cada *array*) sobre el plano de ocupación definido, y posteriormente busca en el interior de las intersecciones de los sectores activos la región de máxima potencia acústica, usando el algoritmo *Steered Response Power (SRP)*. El bloque de extracción de la información de visión detecta rostros humanos en las imágenes obtenidas con las cámaras de vídeo, con una versión multi-pose del algoritmo *Viola and Jones*, y proyecta estas detecciones sobre el plano de ocupación generado. El sistema ha sido evaluado usando la base de datos AV16.3 con resultados prometedores.**

I. Introducción

Los "Espacios Inteligentes" (*Intelligent Spaces*, IS) son entornos dotados de un conjunto de sistemas sensoriales, de comunicación, y de cómputo, transparentes e imperceptibles para el usuario, que están continuamente percibiendo el entorno y cooperando entre ellos para proporcionar la ayuda necesaria a cada persona en su interacción con este espacio. En este contexto en el que se persigue una interacción natural entre el entorno y los usuarios, es imprescindible disponer de información precisa acerca de la existencia y posición de dichos usuarios, así como de si están hablando o no. Para ello es necesario detectar, localizar y seguir a cada usuario en la escena en cuestión [1]. Para resolver dichas tareas se usa la información procedente del conjunto de sensores situados en estos entornos. Debido a que las informaciones de audio y vídeo presentan características complementarias, la comunidad científica se ha planteado la utilización combinada de éstas [2], para aumentar la robustez de los algoritmos de detección, localización y seguimiento incluidos en el IS, denominando al sistema global "proceso de seguimiento multimodal".

Así, dentro del esquema multimodal planteado es, por tanto, necesario resolver tres tareas específicas: la tarea de detección se encarga de discriminar la información de interés dentro de todo el volumen de información recibida por los sensores; el objetivo de la de localización es ubicar en el espacio de coordenadas del IS las fuentes de información, mientras que el seguimiento se encarga de estimar y filtrar la salida de localización obtenida a lo largo del tiempo de acuerdo a un conjunto de modelos determinados.

I-A. Sistemas de seguimiento basados en audio

Los métodos de detección tradicionales basados en audio (*Voice Activity Detectors (VAD)*), emplean características individuales del canal de audio para calcular distintas métricas (por ejemplo niveles de energía, cruces por cero, etc.) sobre las que se aplican reglas de clasificación basadas en umbrales fijos, precalculados o adaptativos (estimados en los períodos de silencio). Estos algoritmos presentan, por tanto, problemas en entornos con baja relación señal a ruido.

Por otro lado, muchos de los algoritmos de localización con audio aprovechan las diferencias de tiempos de llegada de la voz a distintos micrófonos de un *array* para completar el cálculo de la ubicación de la fuente. Una clase de estos algoritmos, conocidos como *Steered Response Power* (*SRP*), hace uso de técnicas de análisis de la orientación del patrón de radiación del *array* (*beamforming*), para evaluar una serie de localizaciones espaciales determinadas como posibles fuentes activas de sonido [3].

La mayor desventaja de los métodos descritos, es que su precisión depende de la densidad de localizaciones evaluadas. Esto implica un alto coste computacional en la obtención de resultados precisos, limitando la posibilidad de su ejecución en tiempo real. Una de las alternativas propuestas por la comunidad científica para solucionar esta dificultad consiste en dividir el espacio de búsqueda en sectores [4], determinar en cuáles de éstos hay actividad acústica y restringir la búsqueda del algoritmo *SRP* a los sectores activos.

Finalmente, los métodos de seguimiento, en general, pueden considerarse una forma de filtrado de la posición instantánea obtenida por el algoritmo de localización, además de resolver ocluciones y resolver oclusiones y problemas de asociación. Para poder resolver todas las tareas comentadas, el proceso de seguimiento implica la definición de un espacio de estados y de observación [5], y de modelos de actuación y observación.

En el vector de estados se incluyen las variables a obtener a la salida del algoritmo de seguimiento (por ejemplo la posición espacial y el estado de habla o silencio del locutor). El proceso de estimación incluido en el seguidor se realiza a partir del vector de observación (información de localización obtenida de la fuente), del modelo de observación y evaluando su evolución temporal caracterizada por el modelo de actuación.

I-B. Sistemas de seguimiento basados en vídeo

Para detectar la presencia de locutores en cada imagen de la secuencia de vídeo, se pueden utilizar algoritmos de detección de rostros, muy eficaces dada la riqueza de características del rostro humano. Estos detectores se basan fundamentalmente en el reconocimiento de patrones de iluminación [6] o en el uso de histogramas de iluminación, color y textura de la piel (*blobs*). Los primeros se caracterizan por ser más robustos frente a cambios de iluminación pero son más complejos computacionalmente, mientras que los segundos son menos sensibles a cambios de pose.

Una de las alternativas para desarrollar el proceso de localización a partir de imágenes es utilizar técnicas de *visual hull* [7], basadas en proyectar las zonas de interés detectadas en la imagen obtenida con cada una de las cámaras, como un volumen cónico en el espacio común tridimensional que observan. La intersección de los volúmenes generados contiene el objeto de interés.

La obtención del volumen completo y exacto del objetivo es computacionalmente costoso y no es estrictamente necesario en tareas de localización y seguimiento como la de interés. Por esta razón son comunes alternativas que sólo obtienen la intersección de dicho volumen con un plano horizontal ubicado a determinada altura del suelo [8], reduciendo de esta forma el tiempo de cómputo del algoritmo.

En el proceso de seguimiento a partir de información de vídeo se utilizan las mismas técnicas que para el caso de audio. En [5] se propone una alternativa al algoritmo clásico de seguimiento basado en filtro de partículas denominada "Filtro de Partículas Extendido con Proceso de Clasificación" *eXtended Particle Filter with Clasification Process (XPFCP)*, que permite manejar un número variable de objetivos en un contexto de seguimiento a partir de información visual.

I-C. Esquemas de fusión audiovisual

La principal clasificación de los algoritmos de fusión audiovisual desarrollados por la comunidad científica se refiere a cómo integran la información obtenida de las dos fuentes sensoriales, distinguiéndose fundamentalmente dos clases de métodos: los "orientados a sistema" y los "orientados a modelo" Los métodos "orientados a sistema" realizan el énfasis de la fusión en la integración de los resultados obtenidos al hacer seguimiento sólo con datos de visión y de audio de forma independiente, aprovechando el desarrollo del estado del arte en algoritmos de seguimiento unimodal, para luego combinar la salida de los seguidores para obtener la solución definitiva. Los métodos "orientados a modelo" por el contrario, se basan en obtener una formulación matemática conjunta de seguimiento que aproveche al máximo las fortalezas de cada tipo de sensor en las diferentes etapas del algoritmo de estimación usado como seguidor, de tal manera que se genere directamente un valor óptimo de estimación conjunta.

En este trabajo se propone un sistema seguimiento de locutores con fusión audiovisual "orientados a sistema". La propuesta aquí planteada está basada en combinar la información de *visual hull* y de actividad acústica generados por los algoritmos de localización en vídeo y audio respectivamente, para luego realizar el seguimiento de la combinación. Además del método de combinación de la información audiovisual, una de las novedades planteadas en el trabajo aquí descrito es el uso del XPFCP mencionado anteriormente como algoritmo de seguimiento en el contexto de fusión audiovisual.

Para el bloque de audio, en este trabajo se propone la combinación del método de detección y localización conjunta basada en sectores (*Sector Based Detection (SBD)*) propuesto en Lathoud, 2006 [4] y el de localización puntual basada en *SRP*. Igualmente, el trabajo aquí expuesto extiende explícitamente el método de detección basado en sectores al uso de varios *arrays* de micrófonos. También se aborda la evaluación de la eficiencia del localizador basado en la combinación *SBD+SRP* frente a la del basado únicamente en *SRP* o el basado en el método propuesto en Lathoud, 2006 [4], *SBD+SCG*, que combina SBD con un algoritmo de optimización basado en el uso del gradiente conjugado (*Scaled Conjugated Gradient (SCG)*).

II. Propuesta desarrollada

II-A. Arquitectura general

En este trabajo se propone combinar en un *mapa de ocupación*, la información de actividad visual generada analizando las imágenes procedentes de múltiples cámaras, con la de actividad acústica generado a partir de la información de audio procedente de múltiples *arrays* de micrófonos, similar a la propuesta en [9].

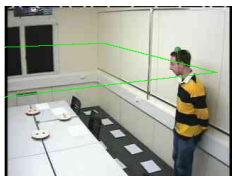


Figura 1: Plano de localización a 1.7m de altura

Este mapa se construye en un plano paralelo al suelo (x, y) a una altura constante $z = h$, que se extiende sobre todo el espacio de seguimiento de los locutores. La altura h del plano se ha de seleccionar de modo que coincida aproximadamente con la de la fuente de actividad, en este caso la boca de los locutores (figura 1). Por ello, el uso del sistema propuesto se orienta a aplicaciones como reuniones o conferencias en las que se espera que la altura h de la fuente sonora permanezca dentro de un rango de variación pequeño.

En la figura 2 se muestra el esquema general del sistema de fusión audiovisual. En los siguientes apartados se describe

cada uno de los bloques funcionales que componen el sistema propuesto.

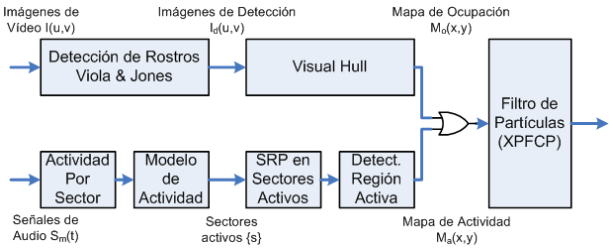
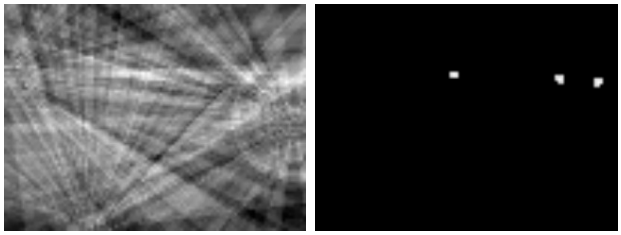


Figura 2: Esquema de fusión audiovisual propuesto



(a) (b)

Figura 3: (a) Imagen original *SRP*, (b) mapa acústico final

II-B. Mapa de actividad acústica

Para la obtención del mapa de actividad acústica se aplica en primer lugar el algoritmo de detección basado en sectores para cada *array*. Los sectores fueron definidos con forma de sector esférico, para intervalos dados de azimut y elevación. En éstos, el índice de actividad se calculó a partir de la evaluación de la métrica *SAM SPARSE MEAN* (SSM) [4], que evalúa el número de frecuencias en las que un sector ha sido *dominante* frente a los demás (donde la dominancia tiene que ver con un mayor valor de la potencia acústica detectada (aunque en [4] se usa realmente una métrica en el dominio de la fase que se demuestra que es funcionalmente equivalente a la potencia calculada por *SRP*)).

Como primera aproximación en este estudio preliminar, el detector de actividad utilizado finalmente aplica un umbral constante, seleccionado a partir de datos de entrenamiento, a la métrica SSM para decidir si esa ventana de habla se caracteriza como *activo (presencia de voz)* o *no activo (ausencia de voz)*.

En las regiones del plano de búsqueda donde se intersectan los sectores detectados como *activos* por todos los *arrays* de micrófonos, se aplica un algoritmo de búsqueda puntual que devuelve una localización $l_i \in \Re^3$ por cada intersección activa. Como algoritmos de búsqueda puntual se evaluaron:

- *SBD+SRP*: La búsqueda del máximo de la salida de potencia de *SRP* (usando una malla de puntos rectangular, como se muestra en la figura 4), evaluada sólo en las intersecciones activas
- *SBD+SCG*: La búsqueda del mínimo de la métrica *Phase Domain Metric (PDM)* [4] mencionada anteriormente, con un algoritmos (*Scaled Conjugated Gradient (SCG)*), también dentro de las intersecciones activas.

En ambos casos, las localizaciones encontradas se hacen *crecer* en el mapa de actividad acústica para que tengan un tamaño adecuado para la fusión con la información visual, usando un método expuesto en [10], con resultados como los mostrados en la figura 3.

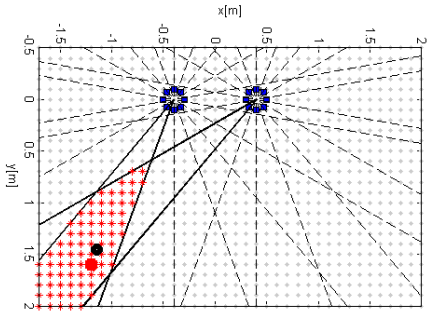


Figura 4: *grid* (gris), intersección de sectores activos (* rojo), máximo *SRP* (círculo rojo) y posición etiquetada manualmente (círculo negro), micrófonos (azul)

II-C. Mapa de actividad visual

Para la obtención del mapa de actividad visual es preciso detectar la presencia del locutor en el plano imagen de cada una de las cámaras. Para ello se utiliza un sistema de búsqueda de rostros basado en una variante multi-pose del algoritmo *Viola and Jones* [6]. Esta variante utiliza varios detectores en paralelo, entrenados para rostros frontales y de perfil, derecho e izquierdo respectivamente. Posteriormente se construye una imagen donde a las regiones activas detectadas en cada pose se les asignan la etiqueta de activas visualmente (ver en la figura 5b, con color blanco), mientras que el resto se le asigna la de visualmente inactivas (en negro en la misma imagen).

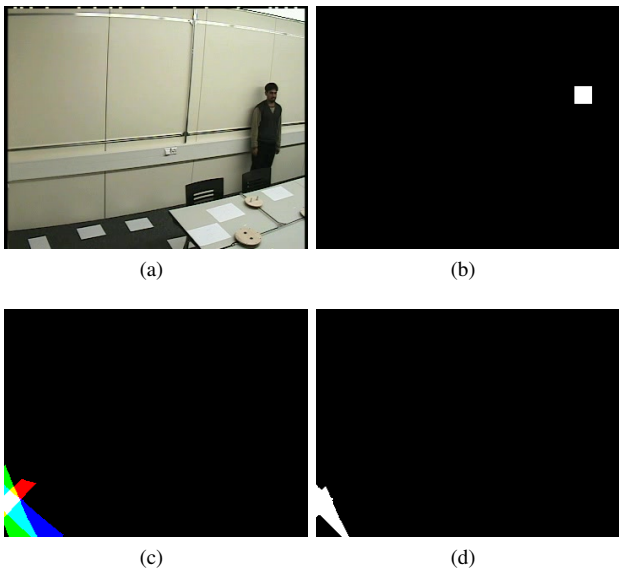
Las imágenes generadas en cada cámara a partir de la original (figura 5a) por el algoritmo de detección visual (ejemplo de una de ellas en la figura 5b) se proyectan al plano de ocupación a partir de la relación de homografía entre esta imagen y la del mapa de actividad visual (figura 5c).

Para combinar las proyecciones de las imágenes de detección de cada cámara, se propone la unión (OR lógica) de las intersecciones (AND lógica) de cada par de proyecciones (ver la figura 5d). Esta propuesta fue seleccionada para evitar que rostros no detectados en una vista, generaran un falso negativo, como ocurre si se usa simplemente la intersección de todas las proyecciones.

II-D. Fusión audiovisual y XPFCP

En este trabajo, y como primera aproximación a la estrategia de fusión, hemos planteado la combinación del mapa de actividad visual y el mapa de actividad acústica, realizando una unión de las regiones activas detectadas en ambos (operación lógica OR de las regiones detectadas).

El resultado de la fusión de los dos mapas de ocupación se usa como entrada al algoritmo XPFCP [5] que realiza la tarea



(a) (b) (c) (d)

Figura 5: Resultados del algoritmo de detección de rostros (a) Imagen Original, (b) Resultado de la detección (c) Proyección de las detecciones sobre el mapa de actividad, (d) Unión de intersecciones dos a dos

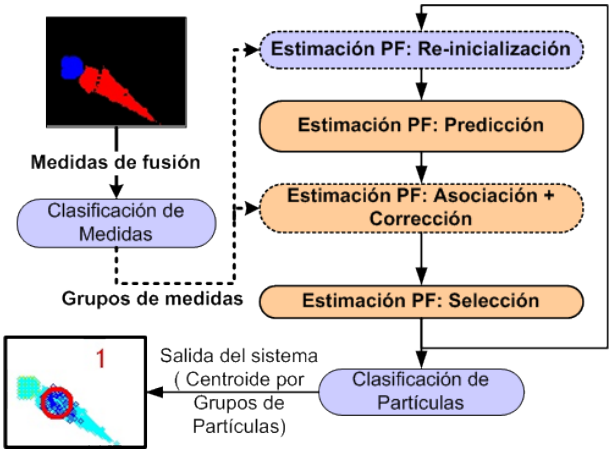


Figura 6: Flujograma del Filtro de Partículas Extendido con Proceso de Clasificación (XPFCP)

de seguimiento del locutor en el IS a lo largo del tiempo, entregando en cada instante la localización más probable (estimación de posición) del locutor. El XPFCP incluye varias modificaciones al filtro de partículas clásico para poder seguir de forma fiable varios objetivos usando un único filtro.

Entre las mejoras significativas que introduce el XPFCP, merece la pena citar la inclusión de dos bloques de clasificación, uno para insertar de forma organizada el conjunto de observaciones (procedentes en este caso del mapa de ocupación resultado de la fusión) en el PF, y otro para generar una salida de estimación determinista, ordenando en conjuntos las partículas del PF y asignándolos a los distintos objetivos de seguimiento.

En la figura 6 se muestra un flujograma general del XPFCP

Nombre	duración (s)	Comportamiento
$seq01-1p-0000$	217	ST
$seq02-1p-0000$	189	ST
$seq03-1p-0000$	242	ST
$seq11-1p-0100$	30	MV
$seq15-1p-0100$	35	MV

Cuadro I: Secuencias utilizadas. ST significa locutor estático, y MV significa locutor en movimiento

aplicado al problema de interés. Como se observa en la figura la entrada del sistema de seguimiento la compone el mapa de ocupación resultado de la fusión audiovisual (en rojo las zonas con actividad sonora positiva y en azul las de actividad visual positiva) y la salida el conjunto de partículas del PF (azul oscuro) y la organización de éstas en clases (círculo rojo), cuyo centro constituye la estimación final de la localización del locutor obtenida por el seguidor.

III. Configuración Experimental

III-A. Base de datos

Para evaluar los resultados generados por cada uno de los bloques desarrollados, así como los generados por el sistema de seguimiento completo propuesto se ha utilizado la base de datos AV16.3 [11].

AV16.3 utiliza una configuración de 3 cámaras de vídeo y dos *arrays* circulares de ocho micrófonos cada uno. Las secuencias de vídeo fueron grabadas con una frecuencia de 25 imágenes por segundo, y las de audio a $16kHz$.

En este trabajo se seleccionaron cinco secuencias de la base de datos, con un único locutor activo y con las características descritas en la tabla I.

El procesamiento de las señales de audio se realizó usando ventanas de audio (tramas, *frames*) de $40ms$ de longitud (no solapadas), de manera que a cada trama acústica se le asocie una única trama visual (fotograma) por cada cámara de vídeo.

III-B. Métricas de evaluación

A los resultados parciales obtenidos, se le han aplicado un conjunto de métricas estándar, definidas en el proyecto *Computers in the Human Interaction Loop* (CHIL) [12], siendo fundamentalmente las siguientes.

- ***Pcor***: tasa de localización medida como el porcentaje de tramas activas con un error inferior a $50cm$..
- **Error promedio de localización**: Promedio de los errores de localización con respecto a la posición etiquetada manualmente [mm].
- **Tasa de borrados (*Deletion Rate*)**: Falsos negativos, ventanas acústicamente activas no detectadas como tales).
- **TPR**: Tasa de verdaderos positivos, calculada como el porcentaje de tramas con actividad de voz detectados como activos.
- **FPR**: Tasa de falsos positivos, calculada como el porcentaje de tramas sin actividad de voz detectados como activos.

IV. Resultados y discusión

IV-A. Evaluación del sistema basado en audio

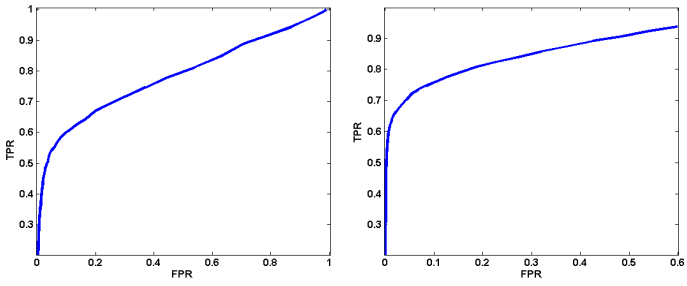
Para analizar el comportamiento el sistema de localización basado en audio debemos evaluar sus dos bloques funcionales, el detector de sectores activos, y los algoritmos de localización de fuentes puntuales.

IV-A1. Evaluación del bloque de detección por sectores: Este bloque fue evaluado de acuerdo a dos perspectivas:

Pensando en el sistema como un tradicional detector de actividad de voz (VAD), se consideró cada trama como elemento de evaluación. Una trama t se considera "activa" si algún volumen de intersección de sectores I, esta activo en t y "no activo" en caso contrario. En la evaluación, se considera que para una trama t se da un acierto (verdadero positivo *true positive (TP)*, si en las anotaciones (posiciones etiquetadas manualmente (*groundtruth*)) de dicha trama t hay una posición etiquetada. De esta forma se pretende analizar las prestaciones de la detección de actividad del algoritmo sin tener en cuenta la precisión en la localización.

Pensando en el sistema como detector-localizador se consideraron todos los pares intersección-trama (I, t) evaluados. Los pares (I, t) detectados por el algoritmo con actividad, son considerando "activos" y el resto "no activos". En la evaluación, se considera que un par (I, t) es un acierto (verdadero positivo *true positive (TP)*, si en las anotaciones (posiciones etiquetadas manualmente (*groundtruth*)) de dicha trama t existe una localización que pertenece a la intersección I. En este caso atendemos tanto a la precisión del proceso de detección, como al de localización de la fuente acústica en la zona espacial correcta.

En la figura 7 se muestran las curvas *Receiver Operating Characteristic* (ROC) del algoritmo (*TPR* vs. *FPR*), como detector y como detector-localizador.



(a) ROC por trama (b) ROC por trama-intersección

Figura 7: *ROC* (a) como detector de voz y (b) como detector-localizador.

Como se aprecia en las curvas ROC, el algoritmo de detección basado en sectores no presenta buenas características ni como detector, ni como detección-localización conjunta para valores de TPR mayores del 60 y 70 % respectivamente. Esto implica que para lograr tasas altas de TPR, es necesario permitir un elevado número de falsos positivos.

Las curvas ROC de cada *array* considerado de forma independiente presentaron un comportamiento similar, lo cual sugiere que la perdida de la efectividad no se debe a la inclusión del concepto de intersección de sectores. Analizando cualitativamente la ubicación de las tramas detectadas en un sector, se observa que muchos de los errores de detección se presentan en el inicio y final de los intervalos de actividad, donde la señal tiene menor energía. Esto sugiere que un umbral fijo no es suficiente para modelar este comportamiento y que es necesario incluir un mecanismo de umbral adaptativo, como el usado en [4], o bien combinar el valor de la métrica SSM con algún otro indicador característico de la señal de voz, lo que queda planteado para trabajos futuros.

IV-A2. Evaluación del bloque de localización puntual: En esta sección se presenta la evaluación de los dos algoritmos de localización puntual desarrollados, *SBD+SRP* y *SBD+SCG*. Estos además son comparados con *SRP* explorando la totalidad del espacio de búsqueda (sin restringirlo a la intersección activa). En la tabla II se muestran los resultados promedios de las cinco secuencias para estos algoritmos.

	SBD+SCG	SBD+SRP	SRP
Pcor[%]	76,0	96,0	79,0
Error promedio [mm]	524	161	478
Tasa de borrados [%]	33	33	0

Cuadro II: Resultados de los sistemas basados en audio

Los resultados indican que el método combinado de *SBD-SRP* presenta mejores índices de localización, tanto en *Pcor* como de *error promedio de localización* que los otros dos.

En la comparación de *SBD+SRP* con *SBD+SCG*, el incremento del error del segundo puede ser debido a mínimos locales y a soluciones encontradas fuera de las regiones de intersección de partida (que no se restringían a priori en el algoritmo).

En lo que respecta al *Pcor*, *SBD+SRP* ha mostrado ser significativamente mejor que *SRP*, a costa de un incremento en la tasa de borrados hasta un 33 % lo que puede interpretarse atendiendo a dos efectos posibles: El primero sería que *SBD+SRP* elimina regiones del espacio de búsqueda que implican posibles errores para *SRP*, con lo que el uso de *SBD+SRP* presentaría la ventaja adicional de "filtrar" zonas problemáticas para *SRP*, mejorando los resultados. El segundo sería que los tramas que *SBD+SRP* ha eliminado, coincidan con aquellos en los que *SRP* comete los mayores errores, lo que supone un problema importante para la comparación objetiva de ambas aproximaciones.

IV-B. Evaluación del sistema de seguimiento basado en audio, vídeo y fusión audiovisual

Con el objetivo de evaluar el sistema de seguimiento en su totalidad, y más concretamente el efecto en éste de la fusión de información audiovisual, se analizan los resultados generados por el XPFCP en tres situaciones: usando información sólo de audio (A), sólo de vídeo (V) y combinada de audio y vídeo (AV), con los resultados mostrados en la tabla III.

	A	V	AV
Pcor [%]	91	100	99
Error promedio [mm]	264	171	170
Tasa de borrados [%]	80	33	31

Cuadro III: Resultados del XPFCP usando información de audio (A), vídeo (V) y audio+vídeo (AV)

Como se observa en la tabla III, el método de seguimiento basado en información audiovisual mejora significativamente al que se basa únicamente en información de audio, y se comporta de manera similar al sistema de basado sólo en vídeo, tanto en detección (*tasa de borrados*) como en localización (*Pcor* y en *error promedio de localización*).

En cuanto a los resultados de detección, la fusión audio-vídeo fue capaz de mejorar ligeramente el resultado de la métrica de borrados (31 %) superando los resultados obtenidos por cada fuente por separado. Esto resulta lógico dado que la combinación a través de la unión permite que las situaciones no detectadas por un bloque, puedan compensarse con las detecciones del otro.

El hecho que la fusión audiovisual no supere al método usando sólo vídeo se debe fundamentalmente a dos factores: los problemas del bloque de detección de audio (en muchas de las tramas la información de audio no llega al bloque de fusión) y a la combinación "plana" de estimaciones de audio y vídeo, sin asignar ningún peso relativo a cada una de ellas.

Adicionalmente, es importante recordar que la especificación de una altura constante para la definición del mapa de ocupación añade un error adicional al sistema de seguimiento cuando la boca del locutor no se encuentra ubicada a esta altura seleccionada.

V. Conclusiones y líneas futuras

En este trabajo se ha presentado un sistema de fusión audiovisual para la localización y seguimiento de locutores. Esta propuesta aporta dos novedades en el contexto de interés: Primero, la extensión del concepto de detección de "sectores activos", dado un *array* de micrófonos, al de detección de "intersección de sectores activos" de múltiples *arrays* de micrófonos, con lo que se logra una reducción mayor del espacio de búsqueda. En segundo lugar, el uso por primera vez del XPFCP en un contexto de fusión audiovisual.

Los resultados obtenidos con esta propuesta de estimación se comparan con los obtenidos mediante sistemas de seguimiento unimodal (audio o vídeo), resultando que la propuesta de fusión multimodal mejora la utilización exclusiva de información acústica, pero no los resultados de la versión basada únicamente en vídeo.

Si embargo, el hecho de que la inclusión del proceso probabilístico de estimación basado en el XPFCP con un modelo de fusión tan simple arroje resultados prometedores, abre una interesante línea futura de trabajo: la incorporación de un modelo de fusión más depurado que pondere mejor la complementariedad de las observaciones de audio y vídeo, y aproveche a través de éste el carácter probabilístico del XPFCP. En cualquier caso, una primera tarea de trabajo futuro se centrará en el análisis exhaustivo de los resultados obtenidos, y proponer estrategias de mejora.

Con respecto a la evaluación del bloque de detección de actividad acústica y el de localización basado en sectores, en su función conjunta como detector y localizador, se abordarán estrategias complementarias que incluyan el modelado explícito de características de la señal de habla, en la línea de lo propuesto en [4]. Igualmente queda pendiente para trabajos futuros el análisis exhaustivo de las diferencias de rendimiento entre *SBD+SRP* y *SRP*, y la propuesta y evaluación de estrategias que mejoren el elevado número de falsos negativos que presentan algunos métodos.

Finalmente, como propuesta de solución al problema de los errores de localización en altura (fijada manualmente), se plantea la extensión del sistema al espacio 3D o bien el uso de un valor adaptativo para este parámetro de altura, basado en una estimación inicial de la altura de las fuentes de voz, o de la cara del locutor.

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. M. party Interaction (AMI) project, "State of the art overview: Localization and tracking of multiple intelocutors with multiple sensors," Augmented Multi-party Interaction (AMI) project, Tech. Rep., 2007.

[2] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.(ICASSP)*, pp. 881—-884, may 2004.

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

[4] G. Lathoud, "Spatio-Temporal Analysis of Spontaneous Speech with Microphone Arrays," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2006, phD Thesis #3689 at the École Polytechnique Fédérale de Lausanne (IDIAP-RR 06-77).

[5] M. Marrón-Romera, "Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos," Ph.D. dissertation, Escuela Politécnica Superior. Universidad de Alcalá. Spain, 2009.

[6] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE CVPR 2001*. Citeseer, 2001.

[7] R. J. Peñín, "Reconstrucción volumétrica exacta a partir de múltiples cámaras y su aplicación a los espacios inteligentes," Tesis de Master, Universidad de Alcalá, 2010.

[8] V. E. Gómez, "Sistema de detección de obstáculos y robots mediante múltiples cámaras en espacios inteligentes," Master's thesis, Escuela Politécnica de Alcalá de Henares, dec 2008.

[9] N. Checka, K. Wilson, V. Rangarajan, and T. Darrell, "A probabilistic framework for multi-modal multi-person tracking," *Computer Vision and Pattern Recognition Workshop*, vol. 9, p. 100, 2003.

[10] M. C. Aguilar, "Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual," Master's thesis, Escuela Politécnica Superior. Universidad de Alcalá. Spain, 2010.

[11] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16.3: An audio-visual corpus for speaker localization and tracking," in *MLMI*, 2004, pp. 182–195.

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