ANTEPROYECTO FIN DE CARRERA

María Cabello Aguilar

8 de mayo de 2009

TÍTULO: Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual

DEPARTAMENTO: Electrónica

AUTOR: María Cabello Aguilar

DIRECTORES: Marta Marrón Romera y Javier Macías Guarasa

1. Introducción

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 (por ejemplo robots móviles [6]).

El Grupo de Ingeniería Electrónica aplicada a Espacios Inteligentes y Transporte del Departamento de Electrónica de la Universidad de Alcalá ha arrancado una línea de actividad en la que se plantean trabajos orientados a la explotación conjunta (fusión) de la información acústica generada por hablantes y la procedente de capturas de vídeo del entorno, para mejorar la interacción de estos en espacios inteligentes, una de cuyas primeras aplicaciones será la localización robusta de locutores.

El trabajo que aquí se propone representa el primer paso en esta línea, orientado fundamentalmente a diseñar, implementar y evaluar un sistema de fusión de información acústica y visual para tareas de localización y seguimiento de hablantes en un espacio inteligente.

En este Proyecto Fin de Carrera pretendemos partir de trabajos iniciados por los Proyectos Fin de Carrera de Eva Muñoz Herraiz [7] ("Diseño, implementación y evaluación

de técnicas de localización de fuente y de mejora de la señal de habla en entornos acústicos reverberantes: aplicación a sistemas de reconocimiento automático de habla"), Carlos Castro González [8] ("Speaker Localization Techniques in Reverberant Acoustic Environments") y María Cabello Aguilar [9] ("Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos"), y especialmente de la Tesis Doctoral de Marta Marrón Romera [10] ("Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos").

2. Objetivos

Los objetivos del proyecto son:

- Mejorar las prestaciones de las herramientas y algorítmica disponibles en el Grupo en sistemas de procesamiento de audio para tareas de localización y seguimiento de locutores (sistemas de ayuda a la experimentación, sistemas de detección de actividad (voz/no voz), etc.). En concreto se abordarán trabajos orientados a:
  - Disponer de un sistema de detección de actividad (voz/no voz) para su uso en entornos acústicos reverberantes.
  - Mejorar las prestaciones del sistema de localización acústica de partida para incorporarle facilidades de procesamiento de ficheros de audio multicanal y de procesamiento de múltiples agrupaciones de micrófonos de forma integrada o independiente.
  - Disponer de un sistema de experimentación versátil que facilite la realización sistemática de experimentos sobre múltiples bases de datos (incluyendo tareas de configuración, ejecución y evaluación)
- Diseñar e implementar una sistema de localización de hablantes combinando la información acústica procedente de múltiples agrupaciones de micrófonos con la información visual capturada con múltiples cámaras en un espacio inteligente, siguiendo el esquema de bloques mostrado en la figura.

  Los requisitos que debe cumplir el trabajo propuesto son los siguientes:
  - Incorporar los procesos que procedan en los sistemas de localización y seguimiento de múltiples locutores, existentes o en desarrollo dentro del Grupo, con vistas a la mejora de las tasas de fiabilidad obtenidas.
  - Ser flexible en el sentido de permitir modificar con facilidad los parámetros de control disponibles en los algoritmos de estimación utilizados.
  - Ser flexible en el sentido de permitir la cómoda incorporación y control de nuevos algoritmos de estimación de fiabilidad en localización y seguimiento.
  - Estar bien documentado para facilitar su utilización en futuros proyectos.
  - Disponer de un software eficiente y robusto.
- Evaluar los algoritmos de localización y seguimiento implementados, realizando experimentos utilizando el software desarrollado y las bases de datos multimodales disponibles en el Grupo. La evaluación cumplirá los siguientes requisitos:

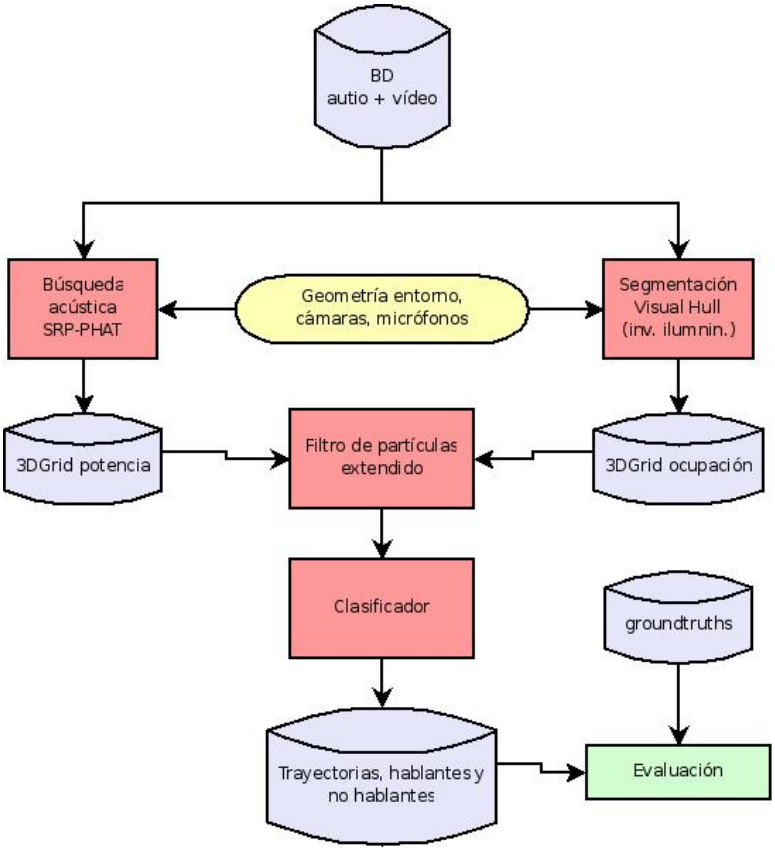

Figura 1: Arquitectura del sistema de fusión propuesto.

- Medir las prestaciones de los sistemas de localización y seguimiento utilizando únicamente información acústica, visual y la combinación de ambas. Se utilizarán las estrategias de evaluación y métricas de calidad propuestas dentro del proyecto CHIL [11] (en las evaluaciones CLEAR).
- Medir las prestaciones de las técnicas de localizacion y seguimiento implementadas en diferentes condiciones acústicas y visuales reales (en función de las bases de datos disponibles[1]).
- Buscar conclusiones razonadas sobre la validez de los resultados obtenidos con las técnicas implementadas. Además, se hará un estudio detallado que ofrezca información sobre la relevancia de los parámetros de control de la experimentación desde un punto de vista práctico.
- Interpretar los resultados obtenidos a la vista de su fiabilidad estadística, considerando en su justa medida las mejoras o degradaciones observadas (respecto a los sistemas de partida).

3. Fases del desarrollo

Las fases que se van a seguir para el diseño, desarrollo y evaluación del sistema son:

- Formación inicial (0,5 meses)
  - Formación en técnicas de programación y en el entorno operativo en el que se desarrollará el Proyecto.
  - Consulta bibliográfica de los distintos métodos de localización y seguimiento de personas basada en información acústica, visual y la combinación de ambas.
  - Formación en la algorítmica de soporte ya desarrollada y disponible en el Grupo para las tareas de localización y seguimiento de locutores usando información acústica o visual.
- Diseño, implementación y adaptación de los módulos software necesarios (3,5 meses):
  - Mejoras en las herramientas y algorítmica disponibles en el Grupo (1 mes).
  - Sistemas de localización y seguimiento de locutores basado en audio, vídeo y fusión audiovisual (2,5 meses).
- Pruebas y evaluación del sistema (1 mes):
  - Análisis de las bases de datos disponibles en el Grupo y selección de las más relevantes para las tareas de evaluación propuestas.
  - Diseño del enfoque experimental: bases de datos, tareas y métricas.
  - Estudio del rendimiento de las técnicas de localización y seguimiento, sobre las bases de datos y tareas definidas.

[1] En principio se plantea el uso de los datos de la evaluación CLEAR 2007, pero se analizarán alternativas.

- Documentación.

Por supuesto, las fases de diseño, desarrollo, pruebas y documentación son cíclicas y abarcan todo el periodo del vida del proyecto.

4. Herramientas y recursos

Las herramientas necesarias para la elaboración del proyecto son:

- PC compatible
- Sistema operativo GNU/Linux [12]
- Entorno de desarrollo Emacs [13]
- Entorno de desarrollo KDevelop [14]
- Procesador de textos LaTeX [15]
- Lenguaje de procesamiento matemático Octave [16]
- Control de versiones CVS [17]
- Compilador C/C++ gcc [18]
- Gestor de compilaciones make [19]

Otros recursos necesarios para la elaboración del proyecto son:

- Bases de datos de habla disponibles en el Grupo
  - Bases de datos generadas en el proyecto CHIL, para las evaluaciones CLEAR 2004, 2005, 2006 [20][21] y 2007 [22][23]
  - Base de datos pública "AV 16.3" de IDIAP [24]
  - Base de datos "HIFI-MM1" del GTH
  - Base de datos "HIFI-AV1" del GTH
  - Base de datos "HIFI-AV2" del GTH
- Algorítmica de procesamiento de habla disponible en el Grupo (incluyendo desarrollos propios y herramientas externas)

Referencias

[1] Augmented Multi party Interaction (AMI) project. State of the art overview: Localization and tracking of multiple intelocutors with multiple sensors. Technical report, 2007.

[2] Michael L. Seltzer. *Microphone Array Processing for Robust Speech Recognition.* PhD thesis, Carnegie Mellon University, 2003.

[3] Wolfgang Herbordt. *Sound capture for human/machine interfaces - Practical aspects of microphone array signal processing.* Springer, Heidelberg, Germany, March 2005.

[4] David Gelbart and Nelson Morgan. Double the trouble: Handling noise and reverberation in far-field automatic speech recognition. In *International Conference on Spoken Language Processing (ICSLP)*, 2002.

[5] Sergei Kochkin and Tim Wickstrom. Headsets, far field and handheld microphones: Their impact on continuous speech recognition. Technical report, EMKAY, a division of Knowles Electronics, 2002.

[6] Alessandro Vinciarelli, Maja Pantic, Hervé Bourlard, and Alex Pentland. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceeding of the 16th ACM international conference on Multimedia*, pages 1061–1070, 2008.

[7] Eva Muñoz Herraiz. Diseño, implementación y evaluación de técnicas de localización de fuente y de mejora de la señal de habla en entornos acústicos reverberantes: aplicación a sistemas de reconocimiento automático de habla. Master's thesis, ETSI Telecomunicación. Universidad Politécnica de Madrid. Spain, 2005.

[8] Carlos Castro. Speaker localization techniques in reverberant acoustic environments. Master's thesis, School of Electrical Engineering. Royal Institute of Technology (KTH). Sweden, 2007.

[9] María Cabello Aguilar. Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos, 2007.

[10] Marta Marrón-Romera. *Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos.* PhD thesis, Escuela Politécnica Superior. Universidad de Alcalá. Spain, 2009.

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

[12] Información sobre gnu/linux en wikipedia. http://es.wikipedia.org/wiki/GNU/Linux [último acceso mayo 2009].

[13] Página de la aplicación emacs. http://savannah.gnu.org/projects/emacs/ [último acceso mayo 2009].

[14] Página de la aplicación kdevelop. http://www.kdevelop.org [último acceso mayo 2009].

[15] Leslie Lamport. *LaTeX: A Document Preparation System, 2nd edition.* Addison Wesley Professional, 1994.

[16] Página de la aplicación octave. http://www.octave.org [último acceso mayor 2009].

[17] Página de la aplicación cvs. http://savannah.nongnu.org/projects/cvs/ [último acceso mayo 2009].

[18] Página de la aplicación gcc. http://savannah.gnu.org/projects/gcc/ [último acceso mayo 2009].

[19] Página de la aplicación make. http://savannah.gnu.org/projects/make/ [último acceso mayo 2009].

[20] Clear 2006 evaluation. http://isl.ira.uka.de/clear06/ [último acceso mayo 2009].

[21] Rainer Stiefelhagen and John Garofolo, editors. *Multimodal Technologies for Perception of Humans.Multimodal Technologies for Perception of Humans First International Evaluation Workshop on Classification of Events, Activities and Relationships, CLEAR 2006*. Springer, 2006.

[22] Clear 2007 evaluation. http://www.clear-evaluation.org [último acceso mayo 2009].

[23] Rainer Stiefelhagen, Rachel Bowers, and Jonathan Fiscus, editors. *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*. Springer, 2008.

[24] Av16.3: an audio-visual corpus for speaker localization and tracking. http://mmm.idiap.ch/Lathoud/av16.3_v6/ [último acceso mayo 2009].