ANTEPROYECTO DE TESIS DE MÁSTER

Alejandro Legrá Ríos

22 de enero de 2010

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

DEPARTAMENTO: Electrónica

AUTOR: Alejandro Legrá Ríos

DIRECTORES: Javier Macías Guarasa y Daniel Pizarro Pérez

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 captada por arrays de micrófonos, y la procedente de capturas de vídeo del entorno, para mejorar la interacción de estos en espacios inteligentes.

El trabajo que aquí se propone está orientado a analizar la viabilidad de una nueva estrategia de uso de la información acústica, a través del modelado de los arrays de micrófonos como cámaras de perspectiva, y la explotación posterior de dicha información visual con técnicas de procesamiento de imágenes. Como aplicaciones objetivo iniciales pueden citarse las referidas a la autocalibración de los arrays [7] y las de localización y seguimiento de hablantes en un espacio inteligente [8].

En esta tesis de máster pretendemos partir de trabajos iniciados por los Proyectos Fin de Carrera de Eva Muñoz Herraiz [9] ("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 [10] ("Speaker Localization Techniques in Reverberant Acoustic Environments"), María Cabello Aguilar [11] ("Comparativa teórica y empírica de métodos de estimación de la posición de múltiples objetos") [12] ("Diseño, implementación y evaluación de un sistema de localización de locutores basado en fusión audiovisual"), y la Tesis Doctoral de Marta Marrón Romera [8] ("Seguimiento de múltiples objetos en entornos interiores muy poblados basado en la combinación de métodos probabilísticos y determinísticos").

2. Objetivos

Los objetivos de la tesis de máster son:

- Diseñar e implementar un sistema de generación de imágenes a partir de la información acústica procedente de múltiples agrupaciones de micrófonos siguiendo el esquema de bloques mostrado en la figura 1.
- Desarrollo de algoritmos de tratamiento de las imágenes acústicas para la localización de fuentes acústicas y reducción de ruido.
- Aplicar técnicas de reconstrucción tridimensional mediante múltiples cámaras:
  - Autocalibración de los arrays de micrófonos a partir de patrones de geometría conocida y su detección en las imágenes generadas por los arrays de micrófonos.
  - Localización y seguimiento de fuentes acústicas mediantes múltiples arrays de micrófonos utilizando mapas de ocupación tridimensional basados en el concepto de Visual Hull [13].
- Evaluar los algoritmos implementados, realizando experimentos utilizando el software desarrollado y las bases de datos multimodales disponibles en el Grupo. La evaluación cumplirá los siguientes requisitos:
  - Medir las prestaciones de la algorítmica desarrollada en las aplicaciones que se generen, en diferentes condiciones acústicas 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).

[1] En principio se plantea el uso de los datos de la evaluación CLEAR 2006 y 2007, pero se analizarán alternativas.

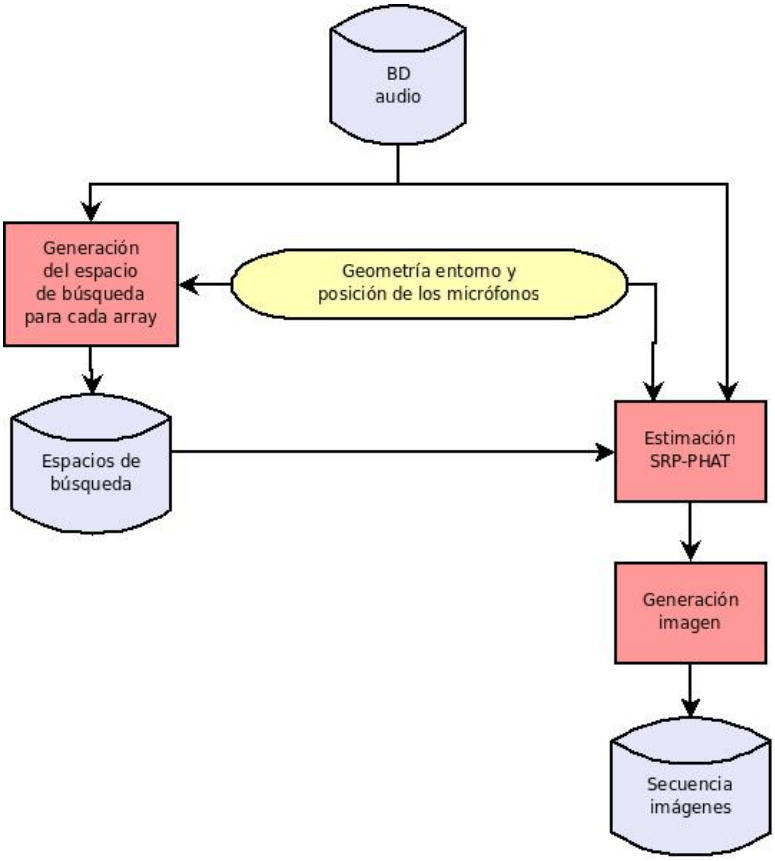

Figura 1: Arquitectura del sistema de generación de imágenes a partir de información acústica.

Los requisitos que debe cumplir el trabajo propuesto son los siguientes:

- Incorporar los procesos que procedan en los sistemas de procesamiento de audio y vídeo 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.

3. Fases del desarrollo

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

- Formación inicial (1 mes)
  - Formación en técnicas de programación y en el entorno operativo en el que se desarrollará la tesis de máster.
  - Consulta bibliográfica de los distintos métodos de procesamiento de información acústica (en tareas de localización y seguimiento de personas)
  - Formación en la algorítmica de soporte ya desarrollada y disponible en el Grupo para las tareas de procesamiento de información acústica o visual.
- Diseño, implementación y adaptación de los módulos software necesarios (6 meses):
  - Herramientas y algorítmica para la generación de información visual a partir de la acústica.
  - Herramientas y algorítmica de modelado de la información visual generada en las aplicaciones correspondientes.
- Pruebas y evaluación (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 la algorítmica generada, sobre las bases de datos y tareas definidas.
- Documentación.

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

4. Herramientas y recursos

Las herramientas necesarias para la elaboración de la tesis de máster son:

- PC compatible
- Cluster de proceso de alto rendimiento
- Sistema operativo GNU/Linux [14]
- Entorno de desarrollo Emacs [15]
- Entorno de desarrollo KDevelop [16]
- Procesador de textos LaTeX [17]
- Lenguaje de procesamiento matemático Octave [18]
- Control de versiones CVS [19]
- Compilador C/C++ gcc [20]
- Gestor de compilaciones make [21]

Otros recursos necesarios para la elaboración de la tesis de máster 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 [22][23] y 2007 [24][25]
  - Base de datos pública "AV 16.3" de IDIAP [26]
  - 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)
- Algorítmica de procesamiento de imágenes 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] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, Cambridge, UK, second edition edition, 2003.

[8] 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.

[9] 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.

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

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

[12] María Cabello 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.

[13] Kong Man Cheung, Simon Baker, and Takeo Kanade. Shape-from-silhouette across time part i: Theory and algorithms. *IJCV*, 62(3):221 – 247, May 2005.

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

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

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

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

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

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

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

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

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

[23] 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.

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

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

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