Resumen

Esta tesis se enmarca en el campo del procesado de señales acústicas y sus aplicaciones para entornos de comunicación emergentes. El procesado de señales acústicas es un área de investigación muy amplia que abarca el diseño de algoritmos para el tratamiento de una o varias señales acústicas con el fin de realizar una tarea determinada, como puede ser: la localización de la fuente de sonido que originó las señales acústicas adquiridas, la mejora de la relación señal a ruido de las mismas, la separación de señales de interés a partir de un conjunto de fuentes interferentes o el reconocimiento del tipo de fuente y/o el contenido del mensaje. Entre las tareas anteriores, la localización de fuente de sonidos (SSL, *Sound Source Localization*) y el reconocimiento automático de voz (ASR, *Automatic Speech Recognition*) han sido especialmente tratados en esta tesis. De hecho, la localización de fuentes de sonido en una habitación ha recibido mucha atención por parte de la comunidad científica en las últimas décadas. La mayoría de las aplicaciones reales de arrays de micrófonos necesitan localizar una o más fuentes de sonido activas en condiciones adversas (baja relación señal-ruido y una alta reverberación). Algunas de estas aplicaciones son los sistemas de teleconferencia, videojuegos, robots autónomos, sistemas remotos de vigilancia, la adquisición de señal en modo manos libres, etc. De hecho, la localización robusta de fuentes de sonido bajo condiciones de alto nivel de ruido y reverberación sigue siendo un reto. Uno de los algoritmos más conocidos para la localización de fuentes en entornos ruidosos y reverberantes es el *Steered Response Power - Phase Transform* (SRP-PHAT), que constituye el marco de referencia para las contribuciones que se proponen en esta tesis. Otro desafío en el diseño de algoritmos de SSL es lograr su funcionamiento en tiempo real con una alta precisión en la localización y con un número razonable de micrófonos a un coste computacional reducido. Aunque el algoritmo SRP-PHAT ha demostrado ser un algoritmo de localización efectivo en entornos reales, su aplicación práctica se basa por lo general en un procedimiento costoso de búsqueda por mallado, por lo que el coste computacional de este método supone un problema a considerar. Es por ello que diversas modificaciones y optimizaciones se han propuesto en la literatura para mejorar su rendimiento y aplicabilidad. En esta tesis se propone una nueva estrategia que extiende eficazmente el comportamiento del algoritmo SRP-PHAT convencional. Este nuevo método realiza una exploración completa del espacio muestreado en lugar de calcular el SRP en posiciones espaciales discretas, aumentando así su robustez y permitiendo un mallado espacial más ancho que reduce el coste computacional requerido en una aplicación práctica, reduciendo también el coste en hardware (menor número de micrófonos). Esta estrategia permite implementar aplicaciones en tiempo real basándose en la información de las posiciones estimadas, como por ejemplo redirigir de forma automática la posición de una cámara o la detección de fragmentos de habla / no habla en sistemas avanzados de videoconferencia.

Como se ha comentado anteriormente, además de las contribuciones relacionadas con SSL, esta tesis está también relacionada con el campo del reconocimiento automático de voz (ASR). Esta tecnología permite a un ordenador o dispositivo electrónico identificar las palabras pronunciadas por una persona para que el mensaje se pueda almacenar y procesar de una forma útil. ASR es utilizado en el día a día en una serie de aplicaciones y servicios, como interfaces hombre-máquina naturales, sistemas de dictado, traductores electrónicos y mostradores de información automática. Sin embargo, aún existen algunos desafíos que hay que resolver. Un problema importante en ASR es reconocer a las personas que están hablando en una habitación mediante el uso de micrófonos a distancia. En el reconocimiento de voz distante, los micrófonos no sólo reciben la señal vía directa de las fuentes de sonido, sino que también reciben réplicas retardadas como resultado de la propagación multitrayecto. Por otra parte, también existen

múltiples situaciones en las teleconferencias en las que varios oradores hablan simultáneamente. En este contexto, cuando múltiples señales de voz están presentes simultáneamente, los métodos de separación de fuentes de sonido (SSS, *Sound Source Separation*) pueden emplearse con éxito para mejorar el rendimiento del reconocimiento automático de voz en escenarios con múltiples fuentes. Con el objetivo de mejorar este tipo de situaciones, en esta tesis se ha propuesto un método de entrenamiento diferente. Este entrenamiento, el cual se basa en un modelo robusto construido a partir de voces previamente separadas en diversos entornos acústicos, utiliza las técnicas de separación como una etapa de mejora del habla que suprime las interferencias no deseadas. Se ha estudiado la combinación de la separación de fuentes y el uso de este entrenamiento específico para la mejora del reconocimiento de voz en diferentes condiciones acústicas, dando lugar a mejoras de hasta un 35% en la tasa final de reconocimiento.

Palabras Clave: Localización de fuentes de sonido, separación de fuentes de sonido, SRP-PHAT, array de micrófonos, detección de habla, reconocimiento automático de voz.