Universidad Nacional Autónoma de México

Facultad de Ingeniería

Localización y rastreo de múltiples hablantes para robots de servicio usando un arreglo triangular de micrófonos

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Maestro en Ingeniería

PRESENTA:

Luis Miguel Gato Díaz

DIRECTOR DE TESIS:

Dr. Caleb Antonio Rascón Estebané

Ciudad de México, 2020

Dedico esta modesta obra a mi amada esposa Yazmín,
y a nuestro primer hijo: Luis Alejandro.

Reconocimientos

Quisiera agradecer a mi tutor, el Dr. Caleb Rascón, cuya guía certera, experiencia y valores humanos fueron decisivos para llevar a buen término la presente investigación. También, mi agradecimiento al Maestro Larry Escobar, con cuyo asesoramiento pude contar en todo momento.

Un agradecimiento especial al gobierno de México, al Consejo Nacional de Ciencia y Tecnología, y a la Universidad Nacional Autónoma de México, por permitirme estudiar en este hermoso país que me ha acogido como otro más de sus hijos.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Luis Miguel Gato Díaz. Ciudad de México, 2020

Resumen

En este trabajo de tesis se presenta el diseño e implementación de un sistema digital capaz de localizar, en términos de acimut, entre uno y cuatro hablantes simultáneamente. Pensado para ser usado por un robot de servicio, se persigue que sea una alternativa más ligera que las soluciones disponibles actualmente. Haciendo uso de un arreglo de tres micrófonos en los vértices de un triángulo, se estiman las direcciones de arribo de las voces midiendo la correlación que existe entre cada par de canales de audio. De acuerdo a las investigaciones realizadas, este método resulta más ligero que otras alternativas basadas en análisis espectral o basadas en formación de haz. Para llevar un seguimiento de las estimaciones, y mejorar la estabilidad de los resultados, se emplea el algoritmo de aprendizaje no supervisado k-*means*, y se realiza un rastreo de cada hablante con sendos filtros de Kalman. Se encontró que las alternativas más robustas a este método de rastreo implicaban un costo adicional superior que no justificaba una ganancia modesta en el desempeño. Los resultados de este trabajo permiten concluir que el diseño propuesto es mucho más eficiente que competidores del Estado del Arte. En efecto, usando varias métricas de desempeño empleadas también por otros autores para evaluar este tipo de algoritmos, y reproduciendo los resultados en exactamente las mismas condiciones gracias a contar con un *corpus* de audio especialmente diseñado para este fin, fue posible verificar que nuestra solución proporciona un desempeño en escenarios prácticos bastante similar a aquellos. Por otro lado, si comparamos el costo computacional que representan, resulta que al usar la implementación propuesta en el presente trabajo se ahorra aproximadamente la mitad en cuanto a ocupación del microprocesador. Si a esto le sumamos que estamos además reduciendo el número de micrófonos que necesita el robot para resolver la tarea de localizar hasta cuatro interlocutores, entonces se puede decir que estamos ante una propuesta en general más conveniente para aplicarse a robots de servicio.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La atención auditiva es la capacidad de la percepción humana de centrarse en un determinado sonido en presencia de otros sonidos de distracción. Un fenómeno interesante donde se manifiesta esta capacidad es el efecto de la reunión de cóctel, o *cocktail party* en inglés [3], que evidencia cómo nuestro cerebro es capaz de concentrar su atención auditiva en un estímulo sonoro en particular, mientras se filtran el resto de los sonidos que le rodean. Así, por ejemplo, en una actividad social en donde varias personas establecen diferentes diálogos entre ellos, cada cual puede concentrarse en la conversación que le interesa, ignorando el resto, a pesar de que coincidan en espacio y tiempo. Esto explica además porqué en ciertas situaciones somos capaces de detectar palabras aisladas, como nuestro nombre, o palabras tabú, en conversaciones ajenas a pesar de que no le estemos prestando toda la atención [4].

Este mecanismo de atención auditiva nos resulta sumamente natural, sin embargo, cuando se intenta replicar en un robot nos percatamos de que en realidad es un proceso muy complejo. De hecho, aún no se ha logrado que las máquinas alcancen un desempeño similar al de una persona promedio, en cuanto a imitar la atención auditiva que realiza nuestro cerebro. Este tipo de tareas, sin embargo, es de una importancia crucial para el futuro de distintas aplicaciones de la inteligencia artificial, como por ejemplo, para la robótica de servicio.

Incluso los seres humanos no somos capaces de percibir, entender, o recordar un sonido, si no le estamos prestando atención. Así que, para que un robot realice correctamente las tareas de reconocimiento del habla y posteriormente de interpretar y ejecutar instrucciones, es de suma importancia mejorar su capacidad para concentrar su atención auditiva en su interlocutor. La localización de las fuentes sonoras de interés constituye pues solamente una etapa previa a otros sistemas de audición robótica, y se busca que mantenga un reducido costo computacional para garantizar un buen desempeño global en un robot de servicio.

1.1. Presentación del problema

Muchos de los métodos más eficientes y de menor costo computacional de separación de fuentes de voz requieren conocer previamente la dirección de arribo de las ondas sonoras de interés para realizar un filtrado espacial [5]. Son métodos que realizan procesamiento simultáneo de múltiples señales provenientes de un arreglo de micrófonos. En la naturaleza, la selección natural ha puesto en evidencia la utilidad de tener más de un sensor de ondas acústicas. Los humanos, y los vertebrados en general, contamos con sistemas auditivos compuestos por dos oídos. Esta referencia binaural resulta determinante para la localización de fuentes sonoras y

para concentrar nuestra atención hacia uno entre varios estímulos sonoros [6]. Gracias a que contamos con dos oídos, percibimos dos versiones distintas de los estímulos sonoros que nos rodean, por un lado debido a la diferencia de amplitud, y por otro lado debido a la diferencia de fase entre ambas percepciones sonoras. Varios estudios médicos señalan que las personas que tienen dificultades para escuchar de uno de los oídos encuentran más difícil mantener la atención auditiva ante la presencia de fuentes sonoras interferentes [3]. Partiendo de esta premisa, hemos identificado la utilización de un arreglo de sensores (micrófonos) como un factor decisivo para que un robot pueda imitar la atención auditiva de los seres humanos.

El procesamiento digital de señales usando arreglos de sensores ha abierto un amplio espectro de aplicaciones ingenieriles en las últimas décadas. Así como los radares usan ondas electromagnéticas en el aire, arreglos de sonares son usados bajo el agua para localizar objetos distantes [7]. Diversos tipos de sensores, incluidos los sensores de ultrasonido y de luz infrarroja, forman parte de tecnologías emergentes en redes de sensores, con el propósito de analizar un entorno y localizar el origen de una determinada señal.

En la Figura 1.1 se ilustra un escenario en donde hay dos fuentes sonoras, y mediante el procesamiento de las observaciones obtenidas por un arreglo de tres sensores (por ejemplo, unos micrófonos), se realiza la localización de las fuentes como requisito para su posterior separación. No se han representado algunas etapas secundarias, como las de pre-procesamiento, para simplificar la figura. En este ejemplo, existen dos fuentes sonoras principales, s_1 y s_2, que son registradas por un arreglo de tres micrófonos, y que provienen de las direcciones θ_1 y θ_2, respectivamente. En un caso más general, puede tratarse de M sensores y N estímulos sonoros.

El modelo de mezclas convolutivas es el más aceptado en la comunidad científica para abordar este tipo de problemas, por ser uno de los más generales [5]. En el instante de tiempo discreto t, una mezcla de N fuentes sonoras $\mathbf{s}(t) = \{s_1(t), s_2(t), ..., s_N(t)\}$ es grabada por un arreglo de M micrófonos, obteniéndose las grabaciones $\mathbf{x}(t) = \{x_1(t), x_2(t), ..., x_M(t)\}$. Las mediciones de los sensores son alimentadas a un sistema de localización de fuentes sonoras. Este sistema debe presentar cierta robustez ante la presencia de reverberación en el entorno, que ha sido represento en la Figura 1.1 por las señales s_1', s_2' y s_2''. La salida de esta etapa es un vector con los ángulos estimados $\hat{\theta}_1$ y $\hat{\theta}_2$ de procedencia de los estímulos sonoros. La segunda etapa recibe como entrada el vector de ángulos estimados e implementa algún método de separación de fuentes sonoras para entregar por separado los estimados de los estímulos sonoros de interés $\hat{s}_1$ y $\hat{s}_2$. Una tercera etapa se encargaría de las tareas de reconocimiento de voz y la interpretación de las instrucciones contenidas en esta.

Este es solamente un ejemplo para ilustrar cómo la localización de fuentes sonoras muchas veces no es el fin, sino que es simplemente un paso intermedio y muchas veces necesario para que el robot ejecute otras tareas más complejas, como la separación de las voces de los distintos interlocutores, identificar a cada uno por su voz, y posteriormente realizar un reconocimiento de las instrucciones o comandos de voz que han sido emitidos. La localización de los interlocutores puede entonces resultar determinante en escenarios de ese tipo, pues de su desempeño dependen otros algoritmos de audición robótica.

1.1.1. Motivación

En un trabajo reciente [8], el Dr. Caleb Rascón *et al.* vinculados al IIMAS[1] describen un método de localización de múltiples fuentes de voz con un reducido costo computacional. El método allí propuesto, que alcanza una precisión de hasta el 100 % en la localización de una

[1] IIMAS es el acrónimo de Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (https://www.iimas.unam.mx).

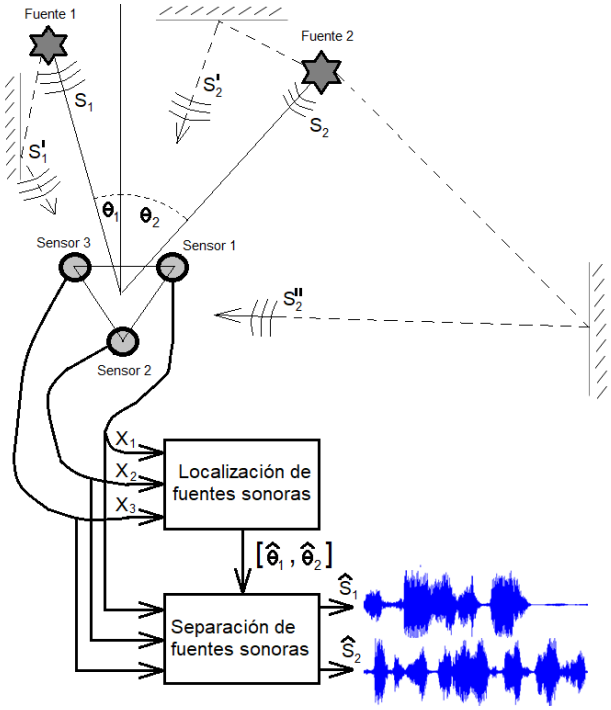


Figura 1.1: Caso de estudio: localización de dos fuentes sonoras con un arreglo de tres micrófonos, seguido de la separación de las voces para posteriormente realizar reconocimiento de voz. Las saetas en líneas discontinuas representan la reverberación.

fuente móvil, y superior al 90 % en la localización simultánea de dos fuentes móviles, ha obtenido varios reconocimientos internacionales.

Específicamente, se trata de la aplicación robótica que ganó el premio *Innovation Award* en la competencia *RoboCup@Home*[1] en el año 2013, como parte del Grupo Golem[2]. Por otro lado, la capacidad de ese sistema para localizar un número de fuentes mayor que el número de micrófonos disponibles, y su potencial impacto en áreas de bioacústica, dispositivos de ayuda auditiva, robótica, etc., le mereció a su autor principal la distinción *Innovator under 35* de la prestigiosa revista de tecnología *MIT Technology Review*, Edición México 2014.

Recientemente han aparecido trabajos que se presentan como una alternativa de bajo costo computacional al trabajo de [8], como por ejemplo [9], donde se presenta un novedoso sistema de localización y rastreo de múltiples fuentes sonoras basado en un paradigma jerárquico, que va aumentando gradualmente la resolución en cada nueva iteración. Sin embargo, estos competidores sugieren emplear sus sistemas con 8 y hasta 16 micrófonos para obtener resultados

[1] El premio "*Innovation Award*" honra los logros científicos y técnicos sobresalientes, prestando especial atención a la disponibilidad de componentes y tecnología que puede ser utilizables por la comunidad de robótica de servicio (https://athome.robocup.org/awards).

[2] Golem-II es un robot de servicio para asistir a las personas en su quehacer diario, con una orientación fuerte a la Inteligencia Artificial y a la interacción Humano-Robot, desarrollado por investigadores del Departamento de Ciencias de la Computación del IIMAS (http://golem.iimas.unam.mx).

satisfactorios. No queda claro cuál será el desempeño de sus propuestas al operar en igualdad de condiciones que un sistema como el de [8], que emplea únicamente tres micrófonos. Tampoco queda claro bajo qué condiciones prácticas de nivel de ruido y reverberación su propuesta es superior y siguen siendo válidos los resultados que ellos presentan. De igual forma, queda la interrogante de si la ganancia en precisión que se obtendría de usar su propuesta es tan grande como para justificar el significativo aumento tanto en consumo de recursos como en costo computacional.

1.2. Hipótesis

Es posible obtener un sistema de localización y rastreo de múltiples hablantes para robots de servicio usando un arreglo de tres micrófonos, cuya precisión sea similar, y aún así con un costo computacional más bajo, que variantes del Estado del Arte que requieren de un número mucho mayor de micrófonos.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar e implementar, para un robot de servicio, un sistema de localización y rastreo de múltiples hablantes usando un arreglo triangular de micrófonos.

1.3.2. Objetivos específicos

- Realizar un estudio del Estado del Arte de la audición robótica, y en particular de las técnicas de localización y rastreo de fuentes sonoras.
- Realizar un análisis comparativo entre los métodos identificados para determinar cuáles se ajustan mejor al problema de investigación.
- Implementar los métodos seleccionados y evaluar su desempeño ante la presencia de múltiples estímulos sonoros.
- Seleccionar el método con mejor desempeño y ajustarlo a los requerimientos de un robot de servicio.

Capítulo 2

Localización y rastreo de múltiples fuentes de voz

En el presente Capítulo, abordamos los principales fundamentos, métodos y métricas de desempeño asociados a las localización y rastreo de fuentes sonoras, y de voces humanas en particular. Específicamente, en la sección 2.1 realizamos una breve descripción del efecto de precedencia, para reconocer cómo de forma natural muchos animales, y los seres humanos, localizamos a nuestro objetivo incluso en escenarios reverberantes. En la sección 2.2 describimos el modelo geométrico de propagación de señales sobre el cual se basan la mayoría de los métodos de localización en robots descritos en el Estado del Arte. En la sección 2.3 realizamos un análisis descriptivo y comparativo de los métodos más importantes de localización en robots usando arreglos de sensores, tomando especial interés en las variantes de optimización que permiten reducir el costo computacional. En la sección 2.4 presentamos las principales métricas de desempeño usadas para evaluar los métodos de localización de fuentes sonoras, y en especial de fuentes de voz. En la sección 2.5 realizamos un análisis comparativo entre los métodos que reportan un mejor desempeño en la literatura científica para el rastreo de hablantes por su voz, identificando sus potencialidades y su limitaciones para aplicarse en robótica de servicio en escenarios reales, y esbozamos el soporte matemático y estadístico sobre el cual operan. Finalmente, en la sección 2.6, finalizamos con un resumen del Capítulo.

2.1. Localización de fuentes sonoras en animales y seres humanos: Efecto de precedencia

Tanto las propiedades espectrales como temporales de una señal acústica puede distorsionarse o degradarse por la acción del entorno entre el emisor y el receptor. El medio de propagación absorbe la energía sonora, en una proporción dependiente de la frecuencia. Los objetos con una impedancia acústica distinta a la del medio provocan que el sonido se disperse, también dependiendo de la frecuencia. La reflexión de las ondas sonoras suele producir reverberación, de tal forma que la percepción sonora cambia [10].

Precisamente, el sistema auditivo humano muestra un interesante efecto ante la presencia de reverberación que varios autores han denominado efecto de precedencia en la localización

sonora. Se le atribuye su primera descripción formal al médico alemán Helmut Haas en su tesis de doctorado en 1949, y en trabajos posteriores como [11]. Antes de Haas, varios descubrimientos independientes de este efecto datan de mediados del siglo XIX. Como consecuencia, ha recibido varias denominaciones, por ejemplo: ley del primer frente de onda, efecto del primer arribo, efecto de supresión auditiva, y efecto Haas. Este consiste en que cuando dos sonidos binaurales se presentan con un retardo muy breve entre sí, y son percibidos como un único evento sonoro, la localización de dicho evento queda determinada en gran medida por el sonido que arriba de primero [6].

Oyentes sanos tienen poca dificultad para localizar fuentes de sonido en ambientes reverberantes. En una habitación con paredes y techo lisos, la percepción sonora queda determinada en gran medida por la referencia binaural que provee el sonido que llega primero. Las reflexiones del sonido en las paredes y el techo llegan con un breve retardo, y el sistema nervioso las filtra para que no interfieran con la localización de la principal fuente sonora. La efectividad de este mecanismo es mayor para los retardos más breves [12]. Retardos muy prolongados son percibidos como eco, y se asocia a cada estímulo sonoro una localización diferente [13].

Sistemáticamente, se ha empleado como método de evaluación del efecto de precedencia un sistema compuesto por dos altavoces alimentados por la misma fuente, excepto porque una de ellas puede recibir un retardo ajustable con relación a la otra, mientras que la otra presenta una atenuación ajustable. El estudio se realiza, para un retardo determinado, ajustando la atenuación de la segunda fuente hasta que el oyente localiza la fuente sonora en el punto medio entre ambos altavoces.

En la configuración descrita anteriormente, investigaciones han mostrado que con un retardo de entre 0.5 ms y 10 ms, se precisa atenuar la segunda fuente entre 5 dB y 8 dB, para que el oyente localice la fuente en el punto medio entre los altavoces. Más específicamente, se ha observado que la atenuación requerida aumenta de 5 dB hasta 12 dB para retardos que aumentan desde 1.0 ms hasta 20 ms. Para retardos mayores a 20 ms se escuchan por separado las fuentes como si fuera una un eco de la otra. De igual forma, a medida que el retardo se reduce, la atenuación va hacia cero. Sin embargo, para retardos inferiores a 1 ms no se obtienen resultados coherentes entre varios oyentes [6].

Se debe tener en cuenta que el efecto de precedencia depende de la naturaleza del estímulo sonoro, dígase voz, música, ruido, etc., y del contenido espectral de este. Se ha observado un límite perceptual en el retardo entre las fuentes. Es decir, que algunos estímulos, como los impulsivos, por ejemplo, requieren más de 1 ms (típicamente más de 5 ms) para distinguir el retardo. Por otro lado, la presencia de interferencia de ruido blanco o de banda ancha en general puede cancelar el efecto de precedencia [6].

En personas con lesiones en lóbulos temporales del cerebro, con esclerosis múltiple, o con dislexia, así como en animales con ablaciones de cortex auditivo, se ha observado un deterioro de la capacidad de localizar una fuente sonora, mostrando una pérdida de la percepción del efecto de precedencia. Además, se ha observado una dependencia con la edad del oyente. En general la capacidad de localizar la fuente ante reverberaciones se ve deteriorada con el aumento de la edad [12].

El efecto de precedencia resulta de interés por varias razones. Por ejemplo, aparece cuando una persona debe localizar a una fuente sonora en un ambiente reverberante. Si el individuo tiene visibilidad directa con la fuente sonora, el frente de onda directo llegará primero que las reflexiones provenientes de otras direcciones. Debido a que se le da un mayor peso en la percepción de la onda directa con relación a las reflexiones, la localización de la fuente sonora suele ser bastante precisa [6]. Este efecto ha recibido una atención considerable en los campos de la acústica arquitectónica y en la estereofonía, donde tiene gran importancia la localización de un evento sonoro inducido por la reproducción de varias fuentes sonoras coherentes.

Además de su importancia práctica, el efecto de precedencia ha sido un tema de interés teórico para la psicoacústica, debido a lo que revela sobre el proceso de localización sonora [6]. A pesar del gran número de estudios del efecto de precedencia, no se ha logrado una comprensión cabal de este fenómeno aunado a las teorías existentes de audición binaural.

En cualquier caso, se ha comprobado que no se trata de un fenómeno particular de los seres humanos. Los estudios comparativos de este tipo con animales son mucho más difíciles de realizar que con humanos, dado que a los animales no se les puede preguntar directamente de hacia donde localizan la fuente sonora. A pesar de esto, varios estudios de comportamiento y estudios fisiológicos han obtenido resultados exitosos en la caracterización del efecto de precedencia varias especies de animales vertebrados, como gatos, ratones, búhos y pericos. Por ejemplo, se han medido los umbrales de retardo del sonido percibido por ambos oídos de un gato, para los cuales se distingue la reverberación como un eco, tanto en ejemplares adultos como en jóvenes [13]. En el caso particular de las aves pequeñas, como los pericos, debido al pequeño tamaño de sus cabezas, y por tanto de la separación de sus oídos, se esperaría una limitada capacidad para localizar fuentes sonoras. Solamente pueden percibir pequeños retardos entre ambos oídos. Sin embargo, se ha observado una capacidad de localización muy similar a la de los humanos, que tenemos una separación mucho mayor entre nuestros oídos. No se tiene claro qué mecanismo evolutivo han desarrollado las aves para lograr esto, pero se supone que esté relacionado con la existencia de un canal interaural [13].

Han habido varios trabajos que intentan imitar en robots de servicio el efecto de precedencia para localizar a fuentes sonoras en entornos reverberantes. En [14], por ejemplo, se logran localizar hasta dos fuentes usando un modelo de efecto de reverberación con respuesta al impulso en forma exponencial. El error de localización es notablemente inferior al obtenido sin usar dicho modelo.

En una investigación más reciente [15], se realiza un análisis comparativo entre distintos modelos de precedencia que combinan el modelo de caída exponencial en la respuesta al impulso del recinto, con un modelo de diferencia de nivel sonoro interaural. Insertando estos en varios algoritmos de separación de fuentes, evalúan su desempeño en recintos con tiempos de reverberación de hasta 890 ms. Todos los modelos descritos muestran mejoras significativas en el desempeño de la separación. A pesar de no conocer cabalmente el fenómeno del efecto de precedencia en la naturaleza, los intentos por imitarlo en robots de servicio han mostrado grandes potencialidades para mejorar los métodos de localización reportados en el Estado del Arte.

2.2. Modelo geométrico de propagación de las señales

En esta sección presentamos el modelo geométrico sobre el cual parten los principales algoritmos de localización por voz, y que también hemos de adoptar en aras de simplificar el planteamiento del problema y sus posibles soluciones. Como mismo en la naturaleza se ha verificado la utilidad de tener más de un sensor para percibir las ondas acústicas en la localización de la fuente sonora, su extensión hacia la audición robótica ha mostrado ser igual de conveniente, con el desarrollo del procesamiento digital de señales usando arreglos de sensores.

El problema genérico del procesamiento digital de señales usando arreglos de sensores consiste en estimar los parámetros de ondas incidentes. Típicamente se trata de ondas provenientes de fuentes distantes, de forma tal que se considera el modelo de campo lejano. Tanto en el caso de ondas electro-magnéticas transmitidas y capturadas por antenas como en las ondas sonoras generadas por altavoces en el aire u otros transductores en el medio acuático, se supone que

el medio de propagación es homogéneo y no dispersivo. Bajo esta premisa, la señal $E(t, \mathbf{r})$, dígase magnitud del campo electro-magnético o presión sonora, en el instante t y en el punto $\mathbf{r} = [x, y, z]^T$ está regida por la ecuación de onda:

$$\frac{\partial^2 E(t,\mathbf{r})}{\partial x^2} + \frac{\partial^2 E(t,\mathbf{r})}{\partial y^2} + \frac{\partial^2 E(t,\mathbf{r})}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E(t,\mathbf{r})}{\partial t^2} \tag{2.1}$$

donde c representa la velocidad de propagación. Esta ecuación se simplifica a la ecuación de onda esférica suponiendo que la fuente es un emisor puntual, quedando como [16]:

$$\frac{\partial^2 \{rE(t,\mathbf{r})\}}{\partial r^2} = \frac{1}{c^2}\frac{\partial^2 \{rE(t,\mathbf{r})\}}{\partial t^2} \tag{2.2}$$

siendo r la distancia que separa al punto de observación de la fuente. La solución de la Ecuación 2.2 es de la forma:

$$E(t,\mathbf{r}) = \frac{1}{r}s_+(t - r/c) + \frac{1}{r}s_-(t + r/c) \tag{2.3}$$

donde $s_+(t - r/c)$ y $s_-(t + r/c)$ son dos funciones arbitrarias, cuya forma depende de las condiciones iniciales. Teniendo en cuenta solamente la parte de la solución asociada a una onda que sale de la fuente, y suponiendo que la separación entre los sensores es mucho menor que la distancia que los separa de la fuente, entonces podemos simplificar la Ecuación 2.3 a:

$$E(t,\mathbf{r}) = s(t - r/c) \tag{2.4}$$

donde s incluye el factor de atenuación $\frac{1}{r}$ actuando sobre s_+ en la Ecuación 2.3.

Si generalizamos al caso de señales complejas, aunque restringiendo a señales de banda estrecha, la señal de interés puede ser de la forma: $s(t) = Ae^{j\omega t}$, de forma tal que [16]:

$$E(t,\mathbf{r}) = Ae^{j(\omega t - \mathbf{k}\cdot\mathbf{r})} \tag{2.5}$$

siendo ω la frecuencia angular de la señal, y $\mathbf{k}$ el vector de onda, que en coordenadas cartesianas queda dado por:

$$\mathbf{k} = \begin{bmatrix} k_x \\ k_y \\ k_z \end{bmatrix} = -\frac{\omega}{c}\begin{bmatrix} \cos(\phi)\cos(\theta) \\ \cos(\phi)\text{sen}(\theta) \\ \text{sen}(\phi) \end{bmatrix} \tag{2.6}$$

donde θ y ϕ representan, respectivamente, el ángulo de acimut y el ángulo de elevación de la fuente con relación al sistema de coordenadas fijado al arreglo de sensores, como se indica en la Figura 2.1.

Por tanto, el problema de estimar la dirección de arribo de una señal se puede traducir a un problema de estimación de dos parámetros: θ y ϕ. Si esta estimación se realiza a partir de las observaciones provenientes de M sensores que conforman un arreglo, entonces la salida del arreglo en el instante t, de acuerdo al presente modelo geométrico, puede expresarse como:

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_M(t) \end{bmatrix} = \mathbf{a}(\theta,\phi)s(t) \tag{2.7}$$

donde el vector de dirección queda dado por [16]:

$$\mathbf{a}(\theta,\phi) = \begin{bmatrix} e^{-j\mathbf{k}(\theta,\phi)\cdot\mathbf{r}_1} \\ e^{-j\mathbf{k}(\theta,\phi)\cdot\mathbf{r}_2} \\ \vdots \\ e^{-j\mathbf{k}(\theta,\phi)\cdot\mathbf{r}_M} \end{bmatrix} \tag{2.8}$$

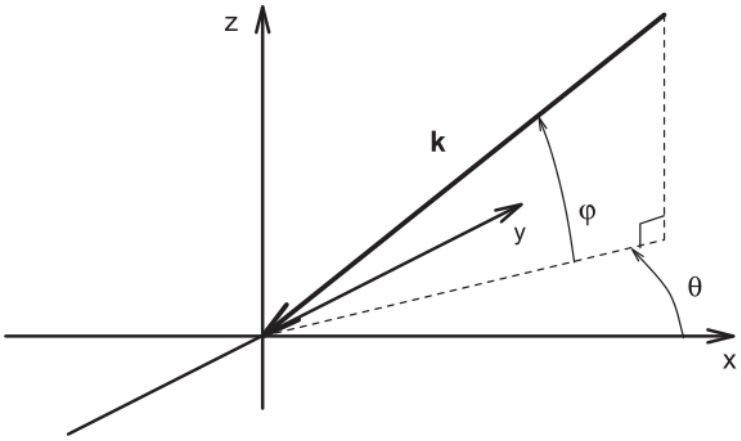


Figura 2.1: Sistema de coordenadas usado en la definición del vector de onda (Ecuación 2.6). El acimut es medido en sentido opuesto a las agujas del reloj y relativo al eje x, mientras que la elevación se mide relativa al plano xy.

2.2.1. Modelo geométrico bidimensional

En múltiples aplicaciones el escenario es bidimensional, de forma tal que la dirección de arribo queda determinada únicamente por un parámetro, ya sea el acimut o la elevación. Para simplificar la notación, en las siguientes secciones se supondrá que solamente se busca estimar el acimut θ en la determinación de la dirección de arribo de la señal de interés, ya sea porque la elevación de la fuente sea conocida de antemano, o porque la fuente y el eje del arreglo de sensores se ubican en un mismo plano:

$$\mathbf{a}(\theta, \phi) = \mathbf{a}(\theta) \tag{2.9}$$

2.2.2. Detección de varias fuentes contaminadas con ruido

La Ecuación 2.7 puede ser generalizada al caso de N fuentes ubicadas en distintas direcciones, más ruido. Las señales correspondientes se superponen al arribar al arreglo de sensores y, bajo el caso bidimensional descrito anteriormente, el vector de observaciones será de la forma:

$$\mathbf{x}(t) = \mathbf{A}(\boldsymbol{\theta})\mathbf{s}(t) + \mathbf{w}(t) \tag{2.10}$$

donde $\mathbf{A}(\boldsymbol{\theta})$ es una matriz de $M \times N$ elementos, compuesta por los vectores de dirección linealmente independientes $\mathbf{a}(\theta_i)$, para $i = 1, 2, \ldots, D$. Las N amplitudes complejas de las señales incidiendo en el instante t sobre el arreglo de sensores están contenidas en el vector $\mathbf{s}$. El vector $\mathbf{w}$ contiene muestras complejas de un proceso aleatorio Gaussiano de media cero y varianza σ^2.

2.2.3. Detección de varias fuentes contaminadas con ruido en un recinto reverberante

Si incluimos la presencia de reverberación, obtenemos el siguiente modelo de mezclas convolutivas [5]:

$$\mathbf{x}(t) = \sum_{k} \mathbf{A}_k(\boldsymbol{\theta})\mathbf{s}(t-k) + \mathbf{w}(t) \tag{2.11}$$

siendo $\mathbf{A}_k(\boldsymbol{\theta})$ la matriz de $M \times N$ que contiene el k-ésimo conjunto de vectores de dirección asociado a la k-ésima señal reflejada dentro del recinto.

2.3. Estimación de la dirección de arribo de múltiples señales usando arreglos de sensores

De acuerdo a la variada gama de literatura científica consultada, las técnicas existentes de localización pueden clasificarse en al menos tres categorías: (1) las basadas en maximizar la potencia a la salida de un formador de haz, denominadas genéricamente *beamforming*; (2) las basadas en la estimación espectral de alta resolución; y (3) las basadas en la estimación de las diferencias de tiempos de arribo de las distintas señales.

2.3.1. Métodos basados en formación de haz

Dentro de esta primera categoría se encuentran los métodos que "apuntan" el patrón del arreglo de sensores hacia varias direcciones y detecta dónde hay una mayor potencia. La variante más simple se obtiene realizando retardos y sumas de las distintas observaciones alineadas, como se observa en la Figura 2.2, denominado también formación de haz convencional. Los retardos se aplican a las señales provenientes de los distintos micrófonos para compensar la diferencia de tiempo de arribo de la señal de interés hacia cada micrófono. De esta forma, luego de procesar N muestras de cada micrófono, se obtiene una señal a la salida con potencia:

$$P(\theta) = \frac{1}{N}\sum_{n=0}^{N-1} |\mathbf{w}^H(\theta)\mathbf{x}[n]|^2 \tag{2.12}$$

donde $\mathbf{x}$ representa el vector de observaciones para determinada componente de frecuencia, y $\mathbf{w}(\theta)$ es el vector de pesos que se aplican a las observaciones, para introducir los retardos correspondientes a la dirección de arribo θ. La direcciones que proporcionen la mayor potencia corresponden a las direcciones de arribo de las señales de interés.

Se ha mostrado que este método proporciona el estimador óptimo de máxima verosimilitud para señales de banda estrecha [18]. Sin embargo, requiere un número elevado de sensores para obtener un bajo error de estimación. Adicionalmente, extender la solución óptima hacia el dominio de señales con un mayor contenido espectral como la voz, implicaría resolver simultáneamente varios problemas de optimización no lineales para las distintas frecuencias, incrementando notablemente el costo computacional. Como consecuencia, los métodos de localización basados en formación de haz se restringen casi exclusivamente a señales de una sola frecuencia o un contenido espectral bastante estrecho, y a sistemas con un elevado número de micrófonos.

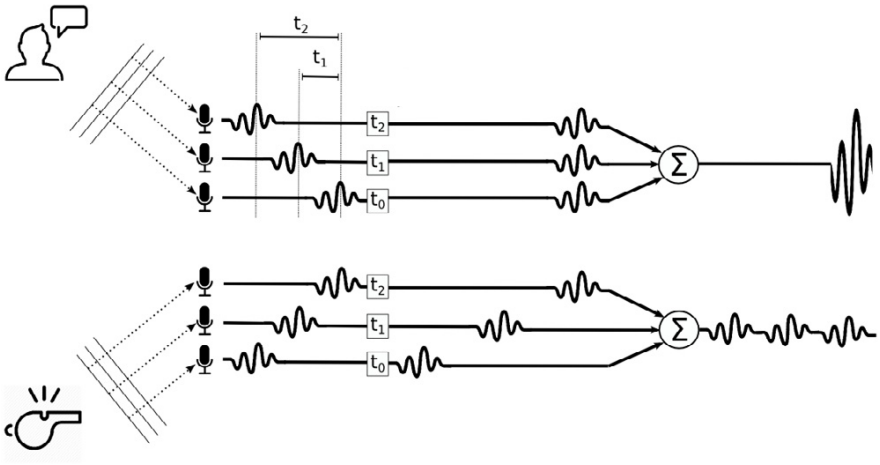


Figura 2.2: Diagrama del formador de haz convencional. Tomado de [17].

En [19] se analiza el desempeño del método de máxima verosimilitud para la localización de múltiples fuentes de banda estrecha. Además, se propone un método Bayesiano de localización de múltiples fuentes, que supera al de máxima verosimilitud, estimando el acimut de cada una a partir de un arreglo de sensores. Constituye uno de los pocos métodos en la literatura estudiada que permite estimar simultáneamente la dirección de arribo de las señales y el número de fuentes que las emiten usando arreglos de sensores. Además, dicho trabajo ha tenido un notable impacto en la comunidad científica, siendo una referencia importante para muchas investigaciones posteriores dentro de ese perfil. Una de las cualidades del método Bayesiano propuesto en [19] es que requiere un reducido número de observaciones para converger hacia una solución, y su estimador presenta un notable desempeño en términos de error cuadrático medio.

Varias investigaciones recientes citan al formador de haz Bayesiano de [19] como un importante pionero en la solución de este tipo de problemas, por ejemplo [20] y [21]. También en la tesis de doctorado [22] se cita a [19] y distintas variaciones de este método para la estimación simultánea de la dirección de arribo, y del número de fuentes. Incluso en una colección reciente [23] de avances en la estimación de la dirección de arribo y en la localización de fuentes, aparece [19] como una de las primeras referencias, lo cual es testimonio una vez más del impacto del trabajo en cuestión.

2.3.1.1. Optimización y robustez del algoritmo en escenarios prácticos

Otro tipo de formador de haz, conocido por la siglas en inglés como GSC (*generalized sidelobe canceller*), proporciona una mejor adaptación a cambios en los niveles de ruido e interferencia, que el formador de haz convencional. Este método, descrito inicialmente en [24], ha sido recientemente aplicado a localización en un robot de servicio [25]. Como se aprecia en la Figura 2.3, se emplea un formador de haz convencional, cuya salida es compensada con una estimación de ruido. Dicha estimación se realiza descartando las señales provenientes de otras direcciones aparte de la dirección que se está evaluando. Con una matriz de bloqueo se rechazan esas señales, y con un banco de filtros adaptativos es posible mejorar la respuesta ante cambios en el entorno.

En la literatura consultada, se han encontrado diversas variantes de formadores de haz para

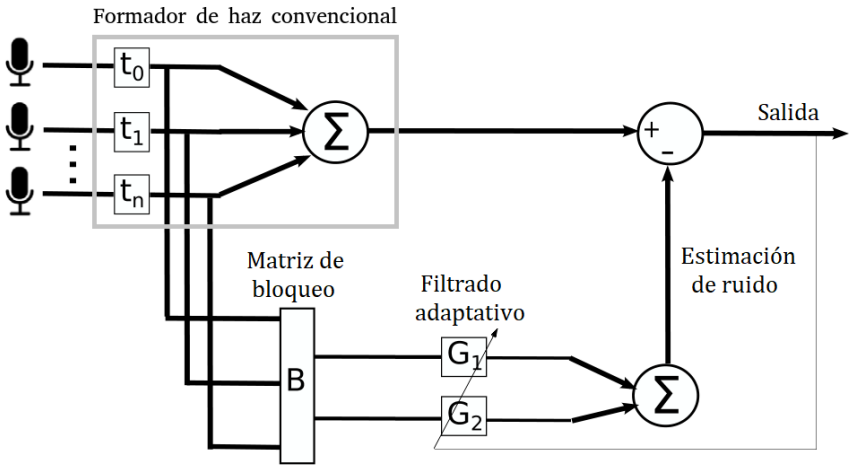


Figura 2.3: Diagrama del método GSC. Tomado de [17].

la localización de fuentes de voz en robots de servicio. En [26] se utiliza para estimar el acimut con un arreglo circular de 8 micrófonos omnidireccionales. En [27] se utiliza en robots móviles con un arreglo cúbico de 8 micrófonos omnidireccionales para localizar simultáneamente hasta 4 fuentes sonoras móviles y 7 fuentes sonoras estáticas. Para implementar un sonar pasivo, en [28] hacen uso de un formador de haz en un robot móvil, que es capaz de localizar una fuente sonora con un arreglo de 8 micrófonos omnidireccionales. En [29] proponen el diseño de un sistema de localización para audición robótica compuesto por un arreglo de micrófonos formado a su vez por tres sub-arreglos circulares. El método de formación de haz les permite localizar y separar dos fuentes sonoras. La integración sobre la superficie de un robot de servicio de un arreglo de 8 micrófonos, y el uso de un formador de haz como método de localización, permite en [30] localizar hasta 3 fuentes móviles. En [31] se implementa un formador de haz con 64 micrófonos en un arreglo esférico, que permite localizar fuentes con elevada precisión a distancias de decenas de metros. Un robot móvil en [32] y en [33] implementa un formador de haz convencional de 32 micrófonos combinado con una técnica de selección de bandas de frecuencias de acuerdo a su intensidad, para localizar y separar múltiples fuentes en movimiento.

Una de las principales desventajas de las distintas implementaciones de formadores de haz, es su desempeño pobre en bajas frecuencias. Por ejemplo, para un arreglo lineal y uniforme de micrófonos, el ancho del lóbulo principal del patrón del arreglo es [16]:

$$\Delta\theta \approx \frac{c}{M\ f\ d\ cos(\theta_0)} \tag{2.13}$$

siendo c la velocidad de propagación del sonido, M la cantidad de micrófonos del arreglo, f la frecuencia de la onda sonora, d la separación entre micrófonos, y θ_0 la dirección hacia donde apunta el arreglo. Queda claro que el lóbulo principal se ensancha a medida que la frecuencia de la onda sonora es menor. Esto implica una reducción en la capacidad de discriminar entre múltiples fuentes en distintas direcciones, es decir, una reducción en la resolución. También se pierde resolución en la medida que θ_0 se aproxima a $\pm\pi/2$, ya que el patrón del arreglo se ensancha en esas direcciones.

En la Figura 2.4 se ilustra también la pérdida de resolución para un arreglo de tres micrófonos en los vértices de un triángulo equilátero con 18 centímetros de lado. El lóbulo principal que proporciona el formador de haz convencional apuntando hacia $\theta = 20^o$, como se puede apreciar,

se ensancha a medida que la frecuencia disminuye, con lo cual la resolución del localizador se verá directamente afectada.

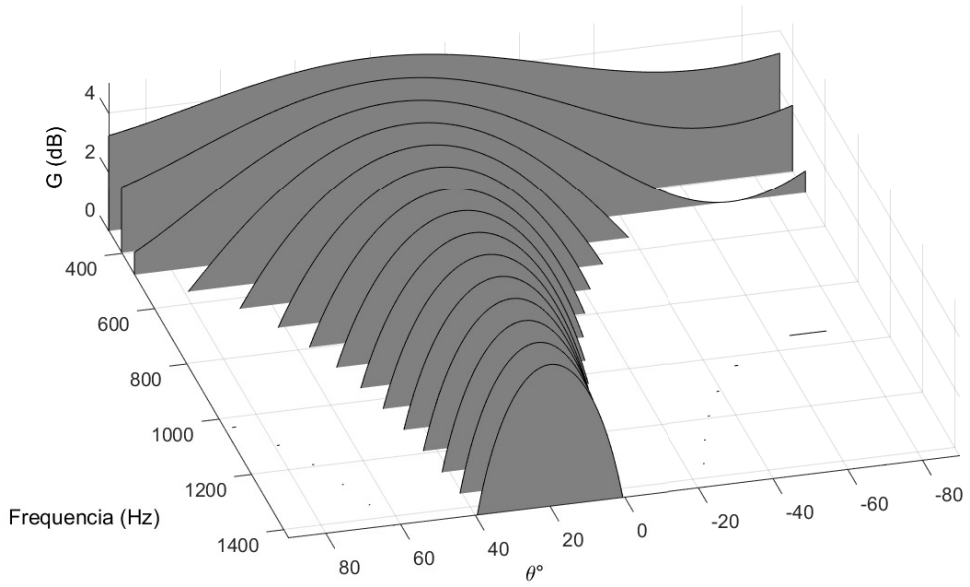


Figura 2.4: Ganancia contra ángulo de arribo y contra frecuencia, para un formador de haz convencional en arreglo de tres micrófonos en los vértices de un triángulo equilátero.

En este sentido, es de crítica importancia diseñar patrones estrechos en bajas frecuencias, al menos dentro del rango de frecuencias de la voz, y con alto rechazo a las señales en las direcciones no deseadas. En [34] sintetizan patrones de formación de haces para aplicaciones robóticas aprovechando las similitudes con el diseño de filtros lineales e invariantes en el tiempo. También en [35] y en [36] proponen un método, tanto para campo lejano como para campo cercano, de diseño de un formador de haz con pesos, que permite hacer más estrecho el patrón de un arreglo lineal, a la vez que minimiza los lóbulos laterales.

2.3.2. Métodos basados en análisis espectral de alta resolución

En contraste con la localización por formación de haz, las técnicas dentro de esta categoría realizan la estimación de la dirección de arribo a partir de la detección de picos en un dominio pseudo-espectral, que requiere de la matriz de correlación espectral entre los distintos sensores. Por ejemplo, el método de respuesta de mínima varianza sin distorsión (*minimum variance distortionless response*, MVDR) realiza la estimación de las direcciones de arribo detectando los picos en la función [16]:

$$P_{\text{MVDR}}(\theta) = \frac{1}{\mathbf{a}^H(\theta)\mathbf{R}_x^{-1}\mathbf{a}(\theta)} \tag{2.14}$$

siendo $\mathbf{a}(\theta)$ el vector de dirección de la Ecuación 2.8, $\{*\}^H$ representa tomar la transpuesta conjugada, y $\mathbf{R}_x$ es la matriz de correlación espectral de las observaciones, para una determinada

frecuencia de interés. Si no se conoce *a priori* dicha matriz, como es el caso más común, se debe realizar una estimación a partir de las observaciones. Para un estimado confiable, se requiere que durante el tiempo de observación las señales de interés y el ruido se comporten como procesos estacionarios, y se supone que la ubicación de las fuentes no cambia en dicho intervalo. Ambas restricciones son difíciles de garantizar en escenarios prácticos de localización de fuentes de voz, y limitan su aplicación casi exclusivamente a escenarios en mayor o menor medida controlados [18].

De forma similar, para la localización basada en MUSIC[1], se buscan picos en el pseudo-espectro siguiente [16]:

$$P_{\text{MUSIC}}(\theta) = \frac{1}{\mathbf{a}^H(\theta)\mathbf{E}_n\mathbf{E}_n^H\mathbf{a}(\theta)} \tag{2.15}$$

donde $\mathbf{E}_n = [\mathbf{e}_{N+1}, \mathbf{e}_{N+2}, ..., \mathbf{e}_M]$ contiene los vectores propios de la matriz de correlación espectral del ruido. Aquí se supone que existen N fuentes de voz, y M micrófonos. De ahí la restricción de que este tipo de métodos requiere que el número de fuentes a localizar sea menor que el número de micrófonos disponibles, tal que sea posible la descomposición en distintos sub-espacios vectoriales asociados a las distintas fuentes de voz, y al ruido. Además, el desempeño de estos algoritmos se deteriora rápidamente ante reverberación, es decir, ante propagación multi-trayecto, pues se requiere que la matriz de correlación espectral tenga rango completo [16].

Con respecto al costo computacional, al igual que los formadores de haz convencionales, la extensión hacia señales de banda ancha, como es el caso de la voz, implica un aumento significativo del costo computacional, pues la localización se debe realizar para el conjunto de frecuencias asociadas a la voz. Teniendo en cuenta que requieren abundante cálculo matricial, incluyendo el cálculo de los valores propios para cada frecuencia, la ejecución de estos métodos suele ser muy demandante para su uso en tiempo real.

A pesar de estos inconvenientes, MUSIC ha sido aplicado con éxito a la localización de fuentes sonoras en robots de servicio. En [38], por ejemplo, un robot móvil está equipado de 8 micrófonos sobre una circunferencia de 30 cm de diámetro, para localizar y dar seguimiento a dos hablantes en movimiento y a una fuente musical fija. Usan un modelo Bayesiano de mezclas Gaussianas para identificar la presencia de una fuente de voz en el pseudo-espectro que se obtiene de MUSIC.

También en una tesis de Maestría reciente [39], realizada en el mismo programa de posgrado en que se inscribe nuestro trabajo, se aplicó una variante de MUSIC para la localización de hasta 3 fuentes de voz hablando simultáneamente, usando dos configuraciones de arreglos de 6 de micrófonos: uno lineal y otro circular. Aunque las pruebas prácticas con que validaron sus resultados fueron limitadas en cantidad y en variedad, la precisión que se logró es bastante buena al menos para ese escenario en donde se hicieron las pruebas.

2.3.2.1. Optimización y robustez del algoritmo en escenarios prácticos

Una variante de MUSIC más robusta en escenarios prácticos, donde el ruido está correlacionado espacialmente, ha sido propuesta en [40]. La denominada *generalized eigenvalue decomposition* (GEVD), aplica un blanqueo del ruido antes de realizar el cálculo de los valores propios de la matriz de correlación espectral. Esto les permite identificar la dirección de la señal de interés suponiendo que el valor propio asociado a esta es el único mucho mayor que la unidad.

[1] MUSIC es un acrónimo para el algoritmo *Multiple Signal Classification*, que proporciona estimadores no sesgados para el número de señales presentes, sus direcciones de arribo, las proporciones en que dichas señales se hallan en una mezcla, y la relación señal a ruido. [37].

Además de que aumenta el costo computacional que MUSIC representa de por sí, esta variante requiere estimar previamente la matriz de covarianza del ruido. Dicha estimación se realizaría en intervalos de tiempo en los cuales las fuentes de voz están en silencio. Una variante de menor costo computacional, basada en *generalized singular value decomposition* (GSVD), ha permitido la ejecución de MUSIC en tiempo real. Para ilustrar el costo computacional referido, se debe tener en cuenta que dentro del sistema de audición robótica *HARK*[1], la etapa de localización con GSVD-MUSIC de hasta tres fuentes sonoras ocupa en promedio cerca del 50 % de una CPU a 2.4 GHz durante todo el tiempo que está corriendo en tiempo real el módulo de localización [41]. Esto sin contar el costo de posteriores etapas de audición, como son la separación y la extracción de rasgos de fuentes sonoras. Ejecutar estos algoritmos en un robot de servicio resultaría un importante reto, si tenemos en cuenta que debe ejecutar simultáneamente otras tareas de audición robótica, e incluso tareas de otra índole, como la visión computacional.

En un trabajo más reciente [42], se propone GSVD para la localización usando un arreglo lineal de 4 micrófonos, en una implementación similar a la descrita en [41], donde un *Kinect*[2] de *Xbox* es integrado a un robot móvil tipo *Turtlebot*[3]. Se localiza una fuente de voz fija, y esto permite al robot orientarse y dirigirse hacia el hablante. También en [43] un robot de servicio realiza localización de un hablante usando como arreglo de micrófonos un *Kinect*, e implementa la misma variante GSVD de MUSIC descrita en [41]. De esta forma, el robot logra exitosamente ir hacia el encuentro de su interlocutor.

2.3.3. Métodos basados en diferencias de tiempo de arribo

Probablemente la categoría de métodos de localización más referida en la literatura científica, es la que incluye distintas variaciones de localización a partir de la estimación de las diferencias de tiempo de arribo de las señales de interés, al arribar a los distintos elementos del arreglo de micrófonos. En el caso más general, donde las señales de interés no son deterministas, lo más habitual es medir la correlación de las observaciones por pares de micrófonos, y determinar a cual retardo corresponde el mayor grado de correlación. Aunque en arreglos unidimensionales esta información proporciona en general cierta ambigüedad en la estimación de las direcciones de arribo, en arreglos bidimensionales de micrófonos, esta información puede proporcionar soluciones unívocas.

En principio, el coeficiente de correlación de Pearson es el método habitual para calcular el vector de correlación normalizado entre dos conjuntos de N observaciones reales $(\mathbf{x}_i, \mathbf{x}_j)$ [17]:

$$\mathbf{R}_{i,j}[m] = \begin{cases} \dfrac{\sum_{n=0}^{N-m-1}(\mathbf{x}_i[n]-\overline{\mathbf{x}_i})(\mathbf{x}_j[n+m]-\overline{\mathbf{x}_j})}{\sqrt{\sum_{n=0}^{N-m-1}(\mathbf{x}_i[n]-\overline{\mathbf{x}_i})^2}\sqrt{\sum_{n=0}^{N-m-1}(\mathbf{x}_j[n+m]-\overline{\mathbf{x}_j})^2}} & \text{para } m \geq 0, \\ \mathbf{R}_{j,i}[-m] & \text{para } m < 0. \end{cases} \tag{2.16}$$

siendo m el retardo que se desea evaluar, y $(\overline{\mathbf{x}_i}, \overline{\mathbf{x}_j})$ las medias aritméticas de las observaciones. El valor de m que maximiza el coeficiente de correlación corresponde al intervalo de tiempo que está adelantada una señal de interés en la observación $\mathbf{x}_i$ respecto a la observación $\mathbf{x}_j$. Posterior a la estimación de retardos $m_{i,j}$ entre pares de micrófonos, se traduce esa información

[1] *HARK* es un software de audición robótica de código abierto. Incluye algoritmos de localización, de separación y de reconocimiento del lenguaje hablado. Es fácilmente integrable a robots que estén equipados con prácticamente cualquier tipo de arreglo de micrófonos [41].

[2] https://en.wikipedia.org/wiki/Kinect

[3] https://www.turtlebot.com

al dominio de las direcciones de arribo. En un escenario donde es válido la aproximación de campo lejano y frente de ondas plano, el acimut en donde se localiza la señal de interés, medido con respecto al eje de simetría de cada par de micrófonos (i, j), es de la forma [17]:

$$\theta_{i,j} = \text{sen}^{-1}\left(\frac{c}{d}\,\frac{m_{i,j}}{f_s}\right) \tag{2.17}$$

donde nuevamente c y d son la velocidad de propagación del sonido en metros por segundo, y la separación entre los micrófonos en metros, respectivamente. La fracción $m_{i,j}/f_s$ representa el tiempo de retardo entre observaciones i y j medido en segundos. Para pasar del tiempo discreto $m_{i,j}$ hacia su equivalente en tiempo continuo se realiza la división por la tasa de captura de muestras de audio (f_s) medida en Hertz.

2.3.3.1. Optimización y robustez del algoritmo en escenarios prácticos

Aprovechando que las señales de audio obtenidas por los micrófonos tendrán media cero, e ignorando por ahora la normalización que se hace en la Ecuación 2.16, obtenemos la siguiente forma equivalente:

$$\mathbf{R}_{i,j}[m] = \begin{cases} \sum_{n=0}^{N-m-1} \mathbf{x}_i[n]\mathbf{x}_j[n+m] & \text{para } m \geq 0, \\ \mathbf{R}_{j,i}[-m] & \text{para } m < 0. \end{cases} \tag{2.18}$$

No solamente el cálculo del vector de correlación usando la Ecuación 2.18 de por sí es menos costoso computacionalmente, sino que además admite una implementación inmediata basada en el algoritmo eficiente de la transformada rápida de Fourier (*fast Fourier transform*, FFT) [44] para operar en el dominio de la frecuencia. A pesar de que la eficiencia de una determinada implementación varía de una arquitectura de computadora a otra, la biblioteca de software libre FFTW (*the "fastest Fourier transform in the West"*)[1] es capaz de adaptar los distintos algoritmos de FFT a una arquitectura de computadora específica, para maximizar así su eficiencia.

En escenarios prácticos de audición robótica la estimación de retardos temporales se deteriora significativamente en presencia de reverberación. El solapamiento de los multi-trayectos luego de la correlación dificulta la estimación precisa de las diferencias de tiempo de arribo. Este problema no es nuevo y se han propuesto diversas variantes que proporcionan una mayor resolución temporal, y así distinguen mucho mejor la onda directa de las ondas reflejadas. En [45], por ejemplo, se describen variantes para enfrentar no solamente la reverberación, sino también el efecto de ruido en general. Se analiza el uso del filtro óptimo de Wiener-Hopf, filtros de blanqueo, transformación de fase, filtro de Eckart, y el estimador de retardo de máxima verosimilitud. En definitiva, lo que todos tienen en común es que aplican una función de peso $\psi(f)$ antes de regresar del dominio de Fourier al dominio del retardo temporal, durante el cálculo de la correlación.

La correlación mejorada se obtiene entonces como:

$$\mathbf{R}_{i,j}[m] = \mathcal{F}^{-1}\left\{\psi(f)\ \mathcal{X}_i(f)\mathcal{X}_j^*(f)\right\} \tag{2.19}$$

siendo $\mathcal{X}(f)$ la transformada de Fourier del vector de observaciones, $(^*)$ es la operación de tomar el complejo conjugado, y $\mathcal{F}^{-1}\{\cdot\}$ representa tomar la transformada inversa de Fourier. En la Tabla 2.1 se presentan las principales funciones de peso $\psi(f)$ que se aplican en las variantes más robustas para el cálculo de la correlación entre dos observaciones. Las expresiones $|\mathcal{S}(f)|^2$

[1] http://www.fftw.org

y $|\mathcal{N}(f)|^2$ representan las densidades espectrales de potencia de la señal de interés y del ruido, respectivamente, mientras que $\gamma_{i,j}(f)$ representa la función de coherencia espectral:

$$\gamma_{i,j}(f) \triangleq \frac{\mathcal{X}_i(f)\mathcal{X}_j^*(f)}{|\mathcal{X}_i(f)||\mathcal{X}_j(f)|} \tag{2.20}$$

Tabla 2.1: Funciones de peso aplicadas en el dominio de la frecuencia para una estimación más robusta de retardos mediante correlación.

Variante de correlación	**Función de peso:** $\psi(f)$
Correlación de Pearson	1
Filtro de Wiener-Hopf	$\frac{1}{\lvert\mathcal{X}_i(f)\rvert^2}$
Filtros de blanqueo	$\frac{1}{\lvert\mathcal{X}_i(f)\rvert\lvert\mathcal{X}_j(f)\rvert}$
Transformación de fase	$\frac{1}{\lvert\mathcal{X}_i(f)\mathcal{X}_j^*(f)\rvert}$
Filtro de Eckart	$\frac{\lvert\mathcal{S}_i(f)\rvert^2}{\lvert\mathcal{N}_i(f)\rvert^2\lvert\mathcal{N}_j(f)\rvert^2}$
Máxima verosimilitud	$\frac{\lvert\gamma_{i,j}(f)\rvert^2}{\lvert\mathcal{X}_i(f)\mathcal{X}_j^*(f)\rvert(1-\lvert\gamma_{i,j}(f)\rvert^2)}$

En [45] se realiza un análisis comparativo de los seis métodos de estimación usando las funciones de peso de la Tabla 2.1. El estimador con filtro de Wiener-Hopf tiene la propiedad favorable de atenuar aquellas componentes de frecuencia en donde el ruido es más intenso, que es donde la incertidumbre del estimador es mayor. Sin embargo, no mejora la resolución temporal del estimador, de forma tal que en general no mejora el desempeño del correlador ante reverberación. Algo similar ocurre con el estimador que emplea sendos filtros de blanqueo. Por otro lado, el estimador con transformación de fase proporciona en el dominio de la frecuencia una respuesta de magnitud unitaria, y que en el dominio del tiempo idealmente se compone de una suma de impulsos asociados a las distintas direcciones de arribo de la señal de interés, con lo cual proporciona la mayor resolución temporal. Como en la mayoría de los casos prácticos, existe cierto grado de correlación entre las observaciones de ruido, la salida se contamina con otros impulsos de falsas direcciones de arribo. El desempeño del estimador se deteriora rápidamente al reducir la relación señal a ruido, llegando al punto en que la estimación de las diferencias de tiempos de arribo se vuelve inestable. Esta respuesta pudiera ser mejorada combinando la transformación de fase con alguna de las funciones de peso que atenúa las componentes de ruido, por ejemplo, el filtro de Wiener-Hopf o el filtro de Eckart [45]. El filtro de Eckart maximiza la relación señal a ruido a la salida del correlador, suponiendo observaciones de ruido estacionario incorrelacionadas entre sí, pero su dificultad de uso práctico radica en que requiere conocer o estimar la densidad espectral de potencia del ruido y de la señal de interés. En [46] se describe una forma de relajar esta restricción, a cambio de cierta degradación en su desempeño. En [45] se demuestra que ante una relación señal a ruido baja, el desempeño del estimador de máxima

verosimilitud, que está basado en la función de coherencia espectral, es equivalente al estimador con filtro de Eckart.

Múltiples aplicaciones de estos métodos de localización se pueden encontrar en la literatura científica. Por ejemplo, en [47] se aplica a un robot usando dos micrófonos para la localización de un hablante ante la presencia de ruido coherente y reverberación. Para discriminar entre la dirección de arribo de señal útil y del ruido, se aplican además máscaras de frecuencia. Este método ha sido propuesto también para su inserción en el módulo de localización de *ManyEars*[1]. En [8] se estiman las direcciones de arribo de las señales de voz de interés estimando los retardos entre pares de micrófonos a partir de la correlación entre las observaciones, aplicando transformación de fase como función de peso. Con un arreglo de tres micrófonos formando un triángulo equilátero, se logran localizar hasta cuatro hablantes en reposo, y dos en movimiento. Los mismos autores describen en [48] cómo la integración de dicho método en un robot de servicio ha permitido facilitar la interacción de este con varias personas dirigiéndole la palabra.

Teniendo en cuenta el aumento de resolución y la robustez en escenarios reverberantes que proporciona la variante de transformación de fase a los métodos basados en diferencias de tiempos de arribo, varios autores han extendido su aplicación a métodos de formación de haz. La potencia a la salida del formador de haz se calcula en el dominio de la frecuencia, aplicando la función de peso de transformación de fase. Esta variante, denominada en ocasiones SRP-PHAT (*steered response power phase transform*) [18] conlleva un mayor costo computacional, pero proporciona una precisión significativamente superior al formador de haz convencional en escenarios prácticos. Si se dispone de un estimado de la relación señal a ruido, en [49] y en [50] proponen una variante mejorada de SRP-PHAT que denominan RW-PHAT (*reliability weighted phase transform*) aplicada a robots de servicio usando la plataforma *ManyEars*.

2.4. Evaluación de desempeño del sistema de localización

Para evaluar de forma objetiva el desempeño de un sistema de localización de fuentes de voz, en la literatura científica se han definido tanto criterios de evaluación directa, como indirecta. Los primeros miden la precisión del estimador, y los segundos miden cómo impacta esta precisión a posteriores etapas de audición robótica. De esta forma, es posible comparar entre sí distintos algoritmos de localización para tomar una decisión sobre cuál es el más apropiado para una aplicación específica.

2.4.1. Evaluación directa

Si se conoce de antemano el vector de direcciones en donde se ubican las distintas fuentes de voz que desean ser localizadas, es posible evaluar directamente el desempeño del sistema de localización comparando el vector de acimut estimado y con el vector de acimut actual. Así, la métrica de desempeño más empleada para sistemas de localización de múltiples fuentes de voz en escenarios reverberantes es el error cuadrático medio del estimador. Por ejemplo, en [19] se realiza una comparación entre el desempeño de dos estimadores, el de máxima verosimilitud y el Bayesiano, respecto al error cuadrático medio. Es también la métrica empleada en [51], en [52], y en [53]. En [54] evalúan un sistema de localización tanto con el error cuadrático medio

[1] *ManyEars* ha puesto a disposición de la comunidad científica su biblioteca en C de software libre para el procesamiento de arreglos de micrófonos que incluye localización y separación de fuentes sonoras. (https://github.com/introlab/manyears).

del estimador, como con el promedio de su error absoluto. En [18] sugieren el uso del error cuadrático medio de las direcciones de arribo para evaluar métodos de localización en escenarios reverberantes. Adicionalmente, comparan las funciones de distribución de probabilidad acumuladas para los errores de estimación. Estas curvas se obtienen mediante el análisis de sus histogramas.

En [55] se propone un método de localización de múltiples fuentes sonoras basado en MUSIC, y evalúan su desempeño a partir del error medio del estimador. Definen el error como la distancia angular entre las verdaderas direcciones de las fuentes y las direcciones estimadas. Esa misma métrica de desempeño se emplea en [56]. La distancia Euclidiana normalizada entre las coordenadas correspondientes a los valores de acimut verdaderos y los estimados, es la métrica de desempeño que se utiliza en [9]. En el caso de múltiples fuentes, se realiza un promedio de dicha métrica.

El promedio de los errores absolutos es la métrica de desempeño que utilizan en [57] y en [8]. En este último también se emplean otras medidas típicas de problemas de clasificación: la precisión, la exhaustividad y el F_1-*score*. En [58] evalúan un método de localización binaural basado en aprendizaje profundo y miden, bajo una resolución de 5°, la precisión del algoritmo como el porciento de veces que se estimó correctamente el acimut de hasta 3 fuentes de voz.

Finalmente, en [59], al igual que en [60] se evalúa el desempeño de localización de fuentes sonoras siguiendo los siguientes criterios:

- `E5`: tasa de detecciones con error absoluto inferior a 5°,
- `E10`: tasa de detecciones con error absoluto entre 5° y 10°,
- `E15`: tasa de detecciones con error absoluto entre 10° y 15°,
- `E30`: tasa de detecciones con error absoluto entre 15° y 30°,
- `I`: tasa de inserción de una fuente de voz, se considera que hubo una inserción cuando el error absoluto es mayor a 30°.

2.4.2. Evaluación indirecta

En el marco de audición robótica como un todo, es de interés determinar cómo afecta el error de estimación del método de localización a posteriores tareas a ejecutar por un robot de servicio. Sin medir directamente el desempeño del método de localización, se pueden evaluar los índices de desempeño de las etapas de separación de fuentes sonoras y de reconocimiento del habla, y así estudiar cómo afecta el desempeño de uno sobre los otros. Consideramos que esta evaluación indirecta del método de localización podría resultar un complemento valioso a los índices de desempeño directos considerados anteriormente, una vez que haya sido integrado el sistema de localización al módulo de audición robótica.

En la etapa de separación de fuentes sonoras, debemos reconocer la presencia de al menos tres causas de deterioro de la separación de cierta fuente s_{target}: la interferencia de otras fuentes (e_{interf}), el ruido natural (e_{noise}), y el ruido artificial (e_{artif}) introducido, por ejemplo, por distorsiones y no linealidades del método empleado en la separación. De esta forma, podemos formular la señal separada como:

$$\hat{s} = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \tag{2.21}$$

En métodos de separación que requieren una localización previa de las fuentes, la componente que depende directamente del desempeño del método de localización es e_{interf}. Por ejemplo, la

separación de fuentes sonoras por métodos formadores de haz se deteriora drásticamente ante errores de estimación de la dirección de arribo de las señales de interés. Precisamente, el patrón del arreglo de micrófonos debe "apuntar" con la mayor precisión posible hacia el objetivo, para obtener un buen rechazo a fuentes interferentes.

Bajo el modelo descrito por la Ecuación 2.21, podemos referirnos a las siguientes métricas de calidad de la separación [61]:

- Relación señal a distorsión (*Signal to Distorsion Ratio*):

$$\text{SDR} = 10\ log_{10}\left(\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2}\right) \tag{2.22}$$

- Relación señal a interferencia (*Signal to Interference Ratio*):

$$\text{SIR} = 10\ log_{10}\left(\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}\right) \tag{2.23}$$

- Relación señal a ruido (*Signal to Noise Ratio*):

$$\text{SNR} = 10\ log_{10}\left(\frac{||s_{\text{target}} + e_{\text{interf}}||^2}{||e_{\text{noise}}||^2}\right) \tag{2.24}$$

- Relación señal a artefactos (*Signal to Artifacts Ratio*):

$$\text{SAR} = 10\ log_{10}\left(\frac{||s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}}||^2}{||e_{\text{artif}}||^2}\right) \tag{2.25}$$

Existen un conjunto de herramientas distribuidas bajo la licencia de software libre GNU-GPL, y agrupadas bajo el nombre BSS_EVAL [62], que son muy eficaces para evaluar el desempeño de la separación de fuentes sonoras bajo los criterios descritos anteriormente.

En la literatura consultada se abordan otros criterios de evaluación, como el denominado criterio D, que toma valores entre 0 y 2 para representar la calidad de la separación. Sin embargo, dichos criterios sufren de algunas limitaciones abordadas en [61], que no afectan directamente los índices 2.22-2.25.

Con respecto a la etapa de reconocimiento de voz, luego de la localización y la separación, es de interés medir la tasa de palabras erróneas (WER, *word error rate*). Aunque no necesariamente una alta WER en robots de servicio implica una mejor comprensión del lenguaje hablado [63], este criterio sirve como indicador de las mejoras que introduce la separación de fuentes sonoras, y por tanto la localización de las fuentes, sobre la etapa de reconocimiento del habla. Entre las publicaciones que utilizan esta métrica de desempeño para evaluar sistemas de localización específicamente diseñado para robots de servicio están [26], [64], [65] y [41]. En algunos casos no emplean directamente la WER, sino su complemento, la tasa de palabras correctas, denominada a veces tasa de reconocimiento de voz.

2.5. Rastreo de múltiples fuentes de voz

En la presente sección, analizaremos los principales algoritmos de rastreo y métricas de evaluación reportadas en el Estado del Arte.

2.5.1. Algoritmos de rastreo de múltiples fuentes de voz

Los métodos de localización descritos en la sección 2.3 del presente trabajo proporcionan estimadores instantáneos de la dirección de arribo de la fuente de voz, de forma independiente a la información de las observaciones pasadas. Las estimaciones no quedan asociadas a su respectivo hablante, y no permiten directamente identificar ni rastrear las fuentes de voz presentes. Para determinar las trayectorias realizadas por los distintos hablantes a partir de las estimaciones individuales de direcciones de arribo, los algoritmos de rastreo más eficaces dividen el problema en dos etapas:

1. predecir, a partir de las estimaciones pasadas, la localización de las fuentes de voz en el futuro inmediato; y

2. corregir la localización predicha haciendo uso de la estimación actual.

Los principales métodos para tratar este tipo de problemas constituyen alguna variante del filtro Bayesiano, ya que este representa la solución óptima en el sentido de Bayes para la predicción ante procesos aleatorios. En la literatura se encuentran disímiles implementaciones con mayor o menor grado de simplificación para obtener soluciones eficaces en escenarios reales. Si se puede modelar la movilidad de las fuentes, así como el ruido, como procesos Gaussianos, y si la medición implica únicamente operaciones lineales, entonces se conduce al filtro de Kalman [66], introducido en 1960. Tantos años de trabajo de la comunidad científica en este sentido lo convierten en la solución más inmediata por su robustez y reducido costo computacional. Sin embargo, con la disponibilidad de computadoras modernas con cada vez mayores recursos y potencia de cómputo, otras soluciones más pesadas se han puesto en práctica y han proporcionado resultados también relevantes. En particular, estas soluciones cumplen su papel más importante en aplicaciones donde la linealidad y la Gaussianidad no pueden ser garantizadas. En este sentido, el filtrado secuencial de Monte Carlo, popularizado como filtro de partículas, ha permitido el rastreo con baja latencia de una o varias fuentes de voz, incluso en movimiento.

Recientemente, el Comité Técnico de Audio y Procesamiento de Señales Acústicas de IEEE[1] lanzó a la comunidad científica un reto[2] para competir en las áreas de localización y rastreo de fuentes de voz. Presentaron un *corpus* de grabaciones realizadas con varios arreglos de micrófonos, y abarcando múltiples situaciones para un entorno realista de audición robótica [67]. El *corpus* contiene las grabaciones multicanal para seis tareas distintas, desde la localización de una fuente sonora estática con un arreglo de micrófonos estático, hasta el rastreo de múltiples hablantes en movimiento con un arreglo micrófonos en movimiento. Los arreglos comprendidos son: un arreglo armónico de 15 micrófonos, un arreglo esférico de 32 micrófonos, un arreglo pseudo-esférico de 12 micrófonos sobre la superficie de un prototipo de cabeza humanoide, y un par de audífonos asistentes para débiles auditivos, con 2 micrófonos cada uno. De esta forma, y dada la variedad de circunstancias disponibles, fue posible medir y comparar objetivamente el desempeño de distintos algoritmos de localización y rastreo disponibles en el Estado del Arte. De entre los métodos en competencia, los que empleaban como algoritmo de rastreo filtro de Kalman o filtro de partículas fueron los que presentaron mejor desempeño en cuanto a error de acimut promedio en los distintos escenarios de prueba evaluados.

Fuera de estos dos métodos principales se pueden encontrar diversas ideas con cierto éxito en algunas aplicaciones prácticas, aunque no con tanto éxito para audición robótica en robots de servicio. Por ejemplo, en [68] se propone un método de rastreo basado en decodificación de Viterbi para seleccionar las direcciones de arribo que más se corresponden con una trayectoria

[1] *Audio and Acoustic Signal Processing Technical Committee* (AASP-IEEE).

[2] *Location and tracking* (LOCATA) *challenge*: https://locata.lms.tf.fau.de

detectada. Dicho algoritmo implica un elevado costo computacional y una latencia que no son compatibles con aplicaciones de tiempo real para robots de servicio. También se ha comenzado a aplicar más recientemente aprendizaje automático para el rastreo de fuentes de voz en escenarios complejos, por ejemplo, afectados con una marcada presencia de reverberación. En principio, la inteligencia artificial es capaz de caracterizar un escenario acústico para perfeccionar el modelo de predicción usado en el rastreo [69]. Sin embargo, los resultados se limitan a escenarios controlados y requiere de una etapa de entrenamiento para cada nuevo entorno [70] [71].

2.5.1.1. Filtro de Kalman

Dentro de esta categoría encontramos distintas variaciones de la teoría de estimadores lineales óptimos descrita en el trabajo pionero de R. E. Kalman [66]. Por ejemplo, en [8] se emplea un estimador invariante en el tiempo para el rastreo de hasta cuatro hablantes en reposo y dos en movimiento, usando únicamente tres micrófonos. Los autores indican que ha permitido reducir la influencia sobre la etapa de rastreo que tienen las estimaciones erróneas de la etapa de localización. Para modelar el movimiento de la i-ésima fuente de voz, consideran el siguiente modelo de predicción lineal no sesgado de mínima varianza:

$$\boldsymbol{\xi}_i(t) = \mathbf{F}\boldsymbol{\xi}_i(t-\tau) + \mathbf{w} \tag{2.26}$$

$$\mathbf{z}_i(t) = \mathbf{H}\boldsymbol{\xi}_i(t) + \mathbf{v} \tag{2.27}$$

donde $\boldsymbol{\xi}_i(t) = [\xi_i^x(t)\ \xi_i^y(t)\ \dot{\xi}_i^x(t)\ \dot{\xi}_i^y(t)]^T$ y $\mathbf{z}_i(t) = [z_i^x(t)\ z_i^y(t)]^T$ son los vectores de estado y de medición en coordenadas Cartesianas en el instante t, respectivamente. El intervalo de observación, τ, es la duración de una ventana de audio procesado. Implícitamente, se supone que tanto el ruido del proceso ($\mathbf{w}$) como el ruido de medición ($\mathbf{v}$) pueden ser considerados procesos aleatorios blancos Gaussianos de media cero y matrices de covarianza conocidas. Bajo estas condiciones, se tienen las siguientes matrices de covarianza para los ruidos:

$$Cov\{\mathbf{w}\} = E\{\mathbf{w}\mathbf{w}^T\} = \begin{bmatrix} \sigma_{w_x}^2 & 0 & 0 & 0 \\ 0 & \sigma_{w_y}^2 & 0 & 0 \\ 0 & 0 & \sigma_{w'_x}^2 & 0 \\ 0 & 0 & 0 & \sigma_{w'_y}^2 \end{bmatrix} \tag{2.28}$$

$$Cov\{\mathbf{v}\} = E\{\mathbf{v}\mathbf{v}^T\} = \begin{bmatrix} \sigma_{v_x}^2 & 0 \\ 0 & \sigma_{v_y}^2 \end{bmatrix} \tag{2.29}$$

siendo $(\sigma_{w_x}^2, \sigma_{w_y}^2)$ las varianzas del ruido en las componentes de posición, y $(\sigma_{w'_x}^2, \sigma_{w'_y}^2)$ las varianzas del ruido en las componentes de velocidad del vector de estado. De forma similar, $(\sigma_{v_x}^2, \sigma_{v_y}^2)$ son las varianzas del ruido en las dos componentes del ruido de medición. En general, es admisible la consideración de que las dos componentes Cartesianas aportan ruido con la misma varianza en el ruido de medición, aunque para el ruido del proceso no es tan acertado igualar el ruido en las componentes de posición con las de velocidad, puesto que son unidades de medida distintas. En la práctica, las matrices de covarianza deben estimarse o, como en muchos casos, se ajustan de forma empírica los elementos de la diagonal para obtener al menos un estimador sub-óptimo.

Para visualizar la relación que guardan los estados ($\boldsymbol{\xi}_i$) en el modelo de Kalman con las direcciones de arribo ($\boldsymbol{\theta}_i$) en sí, en la Figura 2.5 se muestra el diagrama polar de localización para dos fuentes de voz. Para las Ecuaciones 2.26 y 2.27, las matrices de transición ($\mathbf{F}$) y

de medición ($\mathbf{H}$) quedan definidas en [8] como invariantes en el tiempo, de forma tal que corresponden a un movimiento circular uniforme sobre el diagrama polar de la Figura 2.5, y son de la forma:

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & \tau & 0 \\ 0 & 1 & 0 & \tau \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.30}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{2.31}$$

Por tanto, tenemos las siguientes ecuaciones para los estados del sistema, que corresponden al rastreo de las fuentes de voz:

$$\xi_i^x(t) = \sin(\theta_i(t)) \tag{2.32}$$

$$\xi_i^y(t) = \cos(\theta_i(t)) \tag{2.33}$$

$$\dot{\xi}_i^x(t) = \frac{\xi_i^x(t) - \xi_i^x(t-\tau)}{\tau} \tag{2.34}$$

$$\dot{\xi}_i^y(t) = \frac{\xi_i^y(t) - \xi_i^y(t-\tau)}{\tau} \tag{2.35}$$

En realidad, el rastreo se va a realizar con las observaciones $\mathbf{z}_i(t)$ y no directamente con los estados, ya que que estamos tolerando la presencia de determinado ruido de medición. En cualquier caso, se debe inicializar el algoritmo para $t = 0$, y se hace generalmente colocando las componentes de velocidad en cero y las de posición tal que $\theta = 0$, si no se dispone de información adicional que determine otro estado inicial.

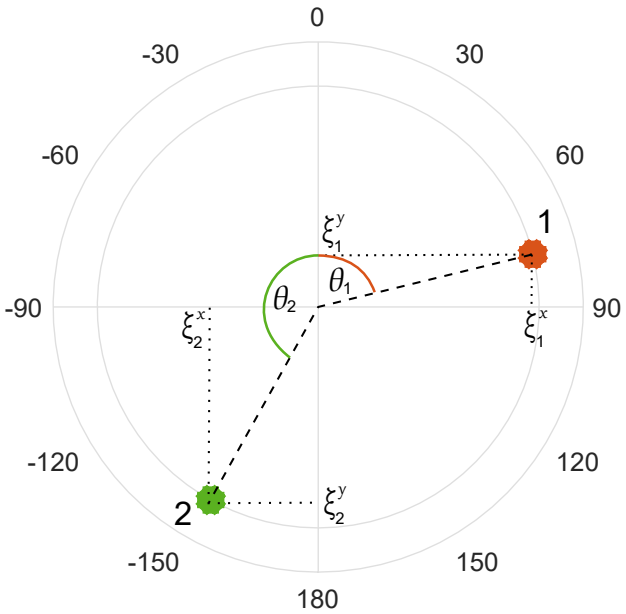


Figura 2.5: Diagrama polar de localización para el rastreo de dos fuentes de voz con filtro de Kalman.

Más recientemente, en [9], se ha usado un filtro de Kalman para el rastreo de múltiples hablantes tanto en el dominio del acimut, como en el dominio de la elevación. Esta variante viene a reemplazar un filtro de partículas que los autores habían usado previamente en [50]. El costo computacional del filtro de Kalman muestra ser notablemente inferior, y aún así les permite rastrear con una mayor precisión cuatro hablantes en movimiento, y nueve en reposo. El hecho de realizar la localización en un espacio tridimensional implica que el vector de estados del modelo de Kalman comprende seis componentes, tres coordenadas cartesianas para la posición, y tres para la velocidad. Esto también tiene como consecuencia que se requiere aumentar el número de micrófonos para mantener una precisión equivalente al modelo bidimensional. De hecho, en [9] se emplean arreglos de 8 y 16 micrófonos. Por otro lado, tiene la ventaja de que el desempeño del estimador se deteriora menos con una variación de la elevación. El rastreo por acimut tiende a perder resolución en la medida en que la elevación de las fuentes de voz se aleja de la elevación del arreglo de micrófonos. En el modelo usado en [9], el ruido del proceso solamente afecta a las componentes de velocidad del vector de estado, y el ruido de medición afecta por igual las tres componentes de las observaciones, de forma tal que las matrices de covarianza corresponden a:

$$Cov\{\mathbf{w}\} = E\{\mathbf{w}\mathbf{w}^T\} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_w^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_w^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_w^2 \end{bmatrix} \tag{2.36}$$

$$Cov\{\mathbf{v}\} = E\{\mathbf{v}\mathbf{v}^T\} = \begin{bmatrix} \sigma_v^2 & 0 & 0 \\ 0 & \sigma_v^2 & 0 \\ 0 & 0 & \sigma_v^2 \end{bmatrix} \tag{2.37}$$

Este modelo simplificado les permite ajustar empíricamente el sistema de rastreo modificando únicamente dos parámetros: σ_w^2 y σ_v^2.

Otros trabajos recientes emplean también alguna variación del filtro de Kalman para el rastreo de fuentes sonoras. En [43], por ejemplo, un robot de servicio realiza el rastreo de un hablante que le dirige la palabra desde una posición de reposo. El rastreo de la fuente localizada se realiza con un banco de filtros de Kalman, con un modelo de medición obtenido de forma empírica. Entre los concursantes del reto de localización y rastreo de [67], los métodos descritos en [72] y en [73] son capaces de rastrear a una fuente de voz mediante sendas implementaciones del filtro de Kalman que modela el movimiento del objetivo.

Tanto en aplicaciones de rastreo de fuentes de voz, como en rastreo de personas en un video, varios autores han empleado una implementación alternativa del filtro de Kalman, que resulta en un algoritmo de estimación lineal de máxima esperanza aplicando cálculo de variaciones. La principal diferencia de esta aproximación consiste en estimar las matrices de covarianza de los ruidos del proceso y de la medición a partir de las propias observaciones. Por ejemplo, dentro del reto de IEEE, uno de los métodos competidores [74] aplica dicho algoritmo de maximización. Hay que resaltar que esta última solución fue la única dentro del reto de IEEE capaz de rastrear con un arreglo móvil de micrófonos a múltiples hablantes en movimiento en forma simultánea. Los autores se inspiran en un trabajo previo [75] que aplica esta técnica al seguimiento visual automatizado de personas.

2.5.1.2. Filtro de partículas

Los filtros de partículas son filtros sub-óptimos que realizan de forma secuencial una estimación Monte Carlo basada en representaciones puntuales de la densidad de probabilidad. Aunque su introducción en la teoría de estimación tiene varias décadas de ser estudiado, su uso en aplicaciones prácticas es más reciente, debido al importante costo computacional que implica. La principal motivación para su uso está en que, en contraste con el filtro de Kalman que supone la presencia de una distribución normal y linealidad, permite tratar problemas cuyo estado sigue una distribución estadística arbitraria, y no se limita al estimador lineal. La función de densidad de probabilidad se evalúa en puntos (partículas) escogidos de forma aleatoria, y a los cuales se les asigna un peso proporcional a la densidad probabilidad correspondiente. De esta forma, se puede interpretar como un método que obtiene aproximaciones a la densidad de probabilidad presente.

Para modelar el movimiento de la i-ésima fuente de voz, consideremos el siguiente modelo de predicción:

$$\boldsymbol{\xi}_i(t) = \mathbf{F}\boldsymbol{\xi}_i(t-\tau) + \mathbf{G}\mathbf{w} \tag{2.38}$$

$$z_i(t) = \arctan\left(\frac{\xi_i^x}{\xi_i^y}\right) + v \tag{2.39}$$

donde la matriz $\mathbf{F}$ y el vector de estado $\boldsymbol{\xi}_i(t)$ coinciden con el modelo lineal de Kalman, y $\mathbf{G}$ permite extender el modelo a movimientos con aceleración constante dentro de cada intervalo de observación [76]:

$$\mathbf{G} = \begin{bmatrix} \frac{\tau^2}{2} & 0 \\ 0 & \frac{\tau^2}{2} \\ \tau & 0 \\ 0 & \tau \end{bmatrix} \tag{2.40}$$

Hasta aquí no hay diferencias con el modelo visto en la sección anterior, porque la matriz $\mathbf{G}$ puede ser introducida también en la Ecuación 2.26 sin que el modelo lineal pierda validez, sino que generaliza el modelo de movimiento a uno uniformemente acelerado. La no linealidad de la estimación se introduce en el modelo de medición de la Ecuación 2.39, que representa ahora directamente el acimut del objetivo ($z_i(t) = \theta_i(t)$). Aunque el estado del sistema siga una distribución normal, la no linealidad de la medición modifica la distribución estadística de las observaciones hacia una desconocida, de forma tal que la predicción óptima usando un filtro de Kalman convencional no puede ser aplicada en este caso.

Para los robots humanoides de [77] y [78] se empleó un sistema de localización basado en formación de haces, que usa filtrado de partículas como algoritmo de rastreo. La arquitectura propuesta, que forma parte del proyecto *ManyEars*, proporciona un sistema de localización y rastreo de fuentes de voz bastante robusto usando un arreglo de 8 micrófonos [79]. Así, han logrado que el robot interactúe de una forma más natural con personas en escenarios de la vida cotidiana. También filtros de partículas es la solución que se da en [38], localizando con el método MUSIC hasta dos hablantes y una fuente musical de forma simultánea. En [49] y en [50], luego de la localización basada en SRP-PHAT, emplean filtros de partículas como algoritmo de rastreo en lugar de filtro de Kalman, argumentando que la distribución estadística observada no es Gaussiana.

Entre los concursantes del reto de localización y rastreo de [67], el método descrito en [80] emplea un filtro de partículas para el rastreo de una fuente de voz en movimiento, siendo las estimaciones de las direcciones de arribo instantáneas las entradas del filtro, y asociando la salida del filtro como la trayectoria del objetivo. Otro competidor [81] emplea un algoritmo que combina un modelo de flujo de partículas con el algoritmo de agrupamiento *k-means*, siendo

capaz de rastrear a múltiples hablantes de forma simultánea, aunque únicamente si estos se encuentran en reposo.

Hay que resaltar que los algoritmos de filtrado de partículas se caracterizan por comprender un elevado costo computacional. Una forma de atacar esta limitación ha sido usar computadoras con arquitectura de *pipeline* y computación paralela para algunas de las etapas del algoritmo que lo admiten [76]. Sin embargo, su aplicación en robótica de servicio para el rastreo de hablantes en tiempo real continúa siendo limitada. Esto teniendo en cuenta que el robot debe realizar simultáneamente otras tareas ajenas a la localización y rastreo de hablantes por su voz, de igual o mayor complejidad. Adicionalmente, se le atribuye a este algoritmo un desempeño variable y hasta cierta medida impredecible, debido a su naturaleza estocástica de optimización [9].

Probablemente la implementación más efectiva de filtros de partículas para robots de servicio rastreando múltiples fuentes de voz en movimiento sea la descrita en [50] como parte del proyecto *ManyEars*. Sin embargo, uno de sus propios autores reconoce en [9] que el algoritmo aún ocupa demasiados recursos de cómputo, y propone reemplazarlo por un filtro de Kalman, con el cual obtiene una precisión mejor y con un ahorro en el costo computacional promedio de entre un 70 % y un 95 %, dependiendo del número de fuentes a rastrear de forma simultánea.

2.6. Resumen

En el presente Capítulo hemos abordado el problema de localización de múltiples fuentes de voz, partiendo de cómo se manifiesta en la naturaleza, y luego cómo se modela en aplicaciones tecnológicas. Específicamente, nos hemos centrado en su aplicación en robótica de servicio. Dentro del Estado del Arte, hemos identificado tres categorías principales en las cuales se pueden clasificar los métodos aplicados en audición robótica. Algunos de los más populares son los basados en algoritmos formadores de haz, gracias a su bajo costo computacional. Sin embargo, su aplicación a señales de voz suele estar limitada debido a que la resolución del localizador se ve reducida para las componentes de baja frecuencia de la voz humana. El desempeño resulta especialmente pobre si se usa un reducido número de micrófonos. Las variantes más efectivas, como GSC, requieren aumentar considerablemente el costo computacional y la latencia del algoritmo, por lo cual su uso en robótica de servicio puede ser cuestionable. Por otro lado, los métodos de análisis espectral de alta resolución, una especie de formadores de haz más sofisticados, implican un volumen de cómputo aún más elevado, de forma tal que hasta las versiones más optimizadas son un reto para aplicaciones de audición robótica de servicio. Hemos identificado los métodos basados en diferencias de tiempos de arribo, y específicamente mediante correlación cruzada de las observaciones de los distintos canales de audio como la alternativa que mejor compromiso proporciona entre desempeño y costo computacional.

Hemos abordado también el problema de rastreo de múltiples hablantes por su voz. Luego de una revisión detallada del Estado de Arte, hemos identificado dos algoritmos principales: filtro de Kalman para modelos Gaussianos y lineales de estimación y de movimiento del objetivo, así como filtro de partículas para modelos más complejos con distribución estadística arbitraria y posiblemente no lineal. Aunque en escenarios reales un modelo Gaussiano y lineal no necesariamente es el que representa con mayor fidelidad el espacio de las observaciones, varios trabajos recientes han reportado excelentes resultados con esa simplificación. La ventaja adicional que presenta el filtro de Kalman está en su costo computacional, notablemente inferior al filtrado de partículas. Además, no se ha observado en la literatura una ganancia significativa en el desempeño por emplear modelos más complejos que justifiquen el uso de algoritmos más

costosos que el filtro de Kalman. De hecho, algunos de los proyectos más exitosos de audición robótica, como *ManyEars*, recientemente han reemplazado sus filtros de partículas por filtros de Kalman para el rastreo de múltiples fuentes de voz en movimiento.

Capítulo 3

Metodología y Evaluación del Sistema Propuesto

En el presente Capítulo, describimos nuestra propuesta de sistema de localización y rastreo de múltiples fuentes de voz para un robot de servicio. En la sección 3.1 partimos de analizar algunas de las implementaciones más exitosas en el Estado del Arte según la literatura científica consultada, y posteriormente contrastamos estas con la implementación propuesta en la presente investigación. En la sección 3.2 se realiza un análisis de los resultados obtenidos al emplear una de las implementaciones del Estado del Arte y nuestra implementación propuesta, respectivamente, a partir de un *corpus* diseñado específicamente para este tipo de evaluaciones. En la sección 3.3 se analiza el costo computacional que proporcionan ambas soluciones. Finalmente, en la sección 3.4, finalizamos con un resumen del Capítulo.

3.1. Implementación del sistema de localización y rastreo de múltiples fuentes de voz

La popular plataforma de audición robótica de código abierto *ManyEars* ha sido ampliamente usada en robots de servicio para localización de fuentes de voz, siendo los robots humanoides *Spartacus* [82], *SIG2* [83] y *ASIMO* [84] los ejemplos más relevantes. Usando un arreglo de ocho micrófonos, *ManyEars* es capaz de localizar y rastrear en tiempo real hasta cuatro fuentes móviles en escenarios prácticos. Los autores reconocen en [79] que su complejidad crece a razón $\mathcal{O}\left(M(M-1)/2\right)$, siendo M el número total de micrófonos, de forma tal que la etapa de localización en *ManyEars* resulta la de mayor costo computacional.

Dos de los principales autores del proyecto *ManyEars* han publicado recientemente en [9] un método de localización basado en SRP-PHAT con rastreo de fuentes móviles usando filtro de Kalman, que ocupa menos recursos de procesamiento que métodos basados en MUSIC con filtrado de partículas. La principal novedad de su propuesta radica en realizar una localización jerárquica, comenzando por un análisis de baja resolución espacial, y gradualmente aumentar la resolución para una localización más precisa. Los mejores resultados son obtenidos con un arreglo cúbico de 16 micrófonos, en la localización de hasta 9 fuentes estáticas y 4 en movimiento. Para ilustrar la significativa reducción en el costo computacional, los autores comparan el uso

de recursos en un Raspberry Pi 3 con procesador ARM Cortex-A53 de cuatro núcleos a 1.2 GHz. Solamente el algoritmo de localización propuesto ocupa para 8 micrófonos alrededor del 13 % de la CPU, y para 16 micrófonos alrededor del 50 %. Esto representa la tercera parte de la carga computacional de SRP-PHAT antes de introducir la localización jerárquica. Con respecto al algoritmo de rastreo con filtro de Kalman, ocupa el 64 % de la CPU para 8 micrófonos, lo cual representa casi la cuarta parte del costo del filtro de partículas en iguales condiciones. Sin embargo, para un arreglo de 16 micrófonos, ambas soluciones superan ampliamente la capacidad de esa arquitectura.

Como lo evidencia dicho proyecto del Estado del Arte, resulta un verdadero reto el poder implementar un sistema de localización y rastreo de fuentes de voz ligero en cuanto a consumo de recursos. En el presente Capítulo, nos proponemos describir una implementación más ligera, sin comprometer demasiado el desempeño resultante en escenarios prácticos.

3.1.1. Implementación del sistema de localización

En nuestra implementación, para cada par de micrófonos del arreglo, se obtienen las diferencias de tiempo de arribo de las ondas sonoras que provienen de las distintas fuentes sonoras a su alrededor. La estimación de dichas diferencias de tiempo se realiza a partir del coeficiente de correlación de Pearson calculado como se indica en la Ecuación 2.16. Para cada segmento de audio procesado, la fuente sonora detectada con mayor energía corresponde al coeficiente de correlación más alto. Esto se desprende de la suposición de que durante el período correspondiente al segmento de audio procesado solamente hay una voz humana activa, y en caso de haber más de una voz activa de forma simultánea, la contribución de una de ellas es más importante que el resto, dado que en general no se solapan completamente. En particular, hemos verificado que dicha suposición es válida para la voz humana si el intervalo de procesamiento es de alrededor de 20 ms. Para una tasa de muestreo de 48 mil muestras por segundo, la cual es una tasa de muestreo común en muchas tarjetas de audio, este intervalo de tiempo corresponde a 960 muestras de audio. Para optimizar el cómputo de los coeficientes de Pearson, tal y como se indica en el apartado 2.3.3.1 del presente trabajo, se hace uso del algoritmo FFT, y se escogió la potencia de 2 más cercana para la elección del intervalo de procesamiento. De esta forma, ventanas de audio de 1024 muestras corresponden a 21.3 ms aproximadamente con la tasa de muestreo seleccionada.

Una ventaja adicional de obtener los coeficientes de Pearson mediante FFT, es la posibilidad de aplicar alguna de las funciones de peso en el dominio de la frecuencia de la Tabla 2.1. En nuestra implementación hacemos uso específicamente de la transformación de fase, porque comprobamos que mejora la robustez del estimador ante reverberación acústica, manteniendo un bajo costo computacional y sin necesidad de estimar previamente el ruido como se requiere por ejemplo, en el Filtro de Eckart. El trabajo en el dominio de la frecuencia nos permite también cancelar las componentes de frecuencias menos importantes en la voz humana, para descartar falsos positivos producidos por ruidos más graves o más agudos de los que son de interés. Se forzaron a cero las componentes de frecuencias discretas entre 1000 Hz y 4000 Hz. No se observaron efectos perjudiciales por haber aplicado este filtrado no lineal, en cambio, el sistema es de esta forma robusto antes ciertos tipos de ruidos fuera de esa banda de frecuencias.

Los ángulos de acimut relativos a cada par de micrófonos se obtienen a partir de las diferencias de tiempos de arribo, y haciendo uso de la Ecuación 2.17. Adicionalmente a las soluciones del primer y del segundo cuadrante: $-90^{\text{o}} < \theta \leq 90^{\text{o}}$, que son ángulos agudos, se deben tener en cuenta las soluciones del tercer y del cuarto cuadrante: los ángulos obtusos en los intervalos $-180^{\text{o}} < \theta \leq -90^{\text{o}}$ y $90^{\text{o}} < \theta \leq 180^{\text{o}}$, respectivamente. Es decir, las soluciones en el primer cuadrante tienen asociada otra solución de la Ecuación 2.17 en el cuarto cuadrante, y las soluciones

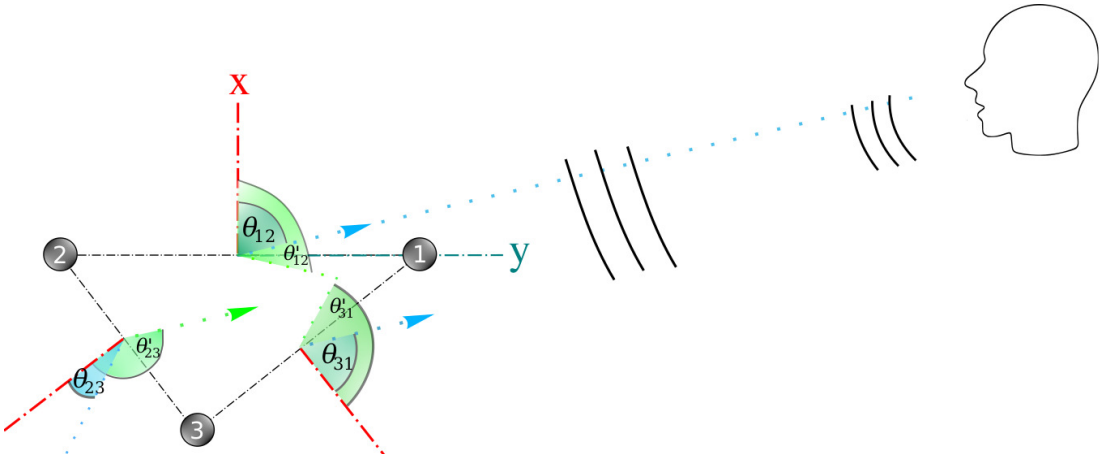


Figura 3.1: Representación geométrica de un arreglo triangular de dimensiones escogidas arbitrariamente, y los 3 pares de ángulos asociados a la fuente de voz presente.

en el segundo cuadrante tienen asociada otra solución en el tercer cuadrante. Es importante hacer esta salvedad, porque en este trabajo también nos interesa detectar y saber diferenciar hablantes que estarían detrás del robot, y no solo los que están enfrente. Para obtener las soluciones de los terceros y cuartos cuadrantes, aplicamos la regla siguiente: $\theta' = 180^o - \theta$, donde θ es la solución de la Ecuación 2.17 en el primer o el segundo cuadrante, y θ' es su complemento en el tercer o el cuarto cuadrante, respectivamente. El resultado siempre se lleva al intervalo $-180^o < \theta' \leq 180^o$. Luego de este paso, contamos con 3 pares de ángulos para realizar la localización de la fuente de voz.

En la Figura 3.1 se representa un arreglo triangular de dimensiones arbitrarias, con tres micrófonos en total, uno en cada vértice del triángulo. Ante la presencia de cierto hablante, para cada par de micrófonos (i, j), se ha señalado el ángulo de acimut θ_{ij} y su complemento θ'_{ij}, relativos al eje transversal a la línea que une los dos micrófonos. Para este ejemplo en particular, de los 6 ángulos de acimut relativos, los 3 que apuntan consistentemente hacia una misma dirección son θ_{12}, θ'_{23}, y θ_{31}. Sin embargo, cada par de ángulos relativos no es de mucha utilidad sin un marco de referencia común. En este ejemplo, $\theta_{12} = 77^o$, $\theta'_{23} = -155^o$, y $\theta_{31} = -65^o$. No es evidente a primera vista que estos tres ángulos se refieran a la misma fuente de voz. Por eso es útil usar un marco de referencia común. Si se conocen los ángulos internos del triángulo, es fácil encontrar las transformaciones que permiten tener el eje de los micrófonos 1 y 2 dos como referencia común. A los ángulos de los pares de micrófonos 2-3 y 3-1 se les deben sumar las siguientes correcciones:

$$\Delta_{23} = \measuredangle 2 - 180^o \tag{3.1}$$

$$\Delta_{31} = 180^o - \measuredangle 1 \tag{3.2}$$

siendo $\measuredangle 1$ y $\measuredangle 2$ los ángulos interiores del triángulo en los dos vértices de referencia. Para el ejemplo de la Figura 3.1 tenemos $\measuredangle 1 = 38^o$ y $\measuredangle 2 = 52^o$, por tanto: $\theta'_{23} + \Delta_{23} = -283^o = 77^o$, y $\theta_{31} + \Delta_{31} = 77^o$, los cuales coinciden con el acimut de referencia $\theta_{12} = 77^o$. Con esta transformación se hace evidente que los tres ángulos de acimut apuntan en la misma dirección.

Para detectar el acimut que corresponde a la fuente de voz activa en cierto instante, basta que el algoritmo de localización detecte de entre los 6 ángulos de acimut, cuáles 3 son iguales entre sí luego de haber aplicado las correcciones 3.1 y 3.2. En la práctica, sin embargo, no

suelen encontrarse tres ángulos que coincidan exactamente entre sí, aunque en efecto se estén refiriendo a una misma fuente de voz. Una de las causas comunes de esta desigualdad es el error que se introduce por la presencia de reverberación. Cada micrófono percibe la señal de interés con un distinto efecto de reverberación debido a que se ubican en distintas posiciones respecto a la fuente y su entorno, y esto distorsiona en cierta medida la estimación de acimut. Otra causa es que en general la resolución angular que percibe cada par de micrófonos es distinta, en parte porque la separación entre los micrófonos no necesariamente es la misma, y en parte porque la Ecuación 2.17 es no lineal respecto a la diferencia de tiempos de arribo.

Para resolver esta desigualdad, el sistema de localización debe tener definida cierta tolerancia de variación de acimut. De esta forma, el objetivo es encontrar cuáles 3 ángulos de los 6 disponibles son más parecidos entre sí, con una margen de error de hasta dicha tolerancia. Si no hay 3, o al menos 2 ángulos que cumplan este requisito, se descarta la presencia de alguna fuente de voz en el entorno.

3.1.2. Implementación del sistema de rastreo

Una vez que finaliza la estimación del acimut de cada fuente de voz detectada, se alimenta este vector de ángulos al módulo que se encarga de interpretar esa información y darle un significado definitivo. No se trata ya de tener valores aislados de una iteración a otra, sino de identificarlos, agruparlos y darles seguimiento a lo largo del tiempo.

En nuestra implementación, se lleva un registro histórico de las últimas N_{hist} direcciones de arribo que han sido detectas, como se ilustra en la Figura 3.2a para $N_{\text{hist}} = 4$. Cuando se llena el registro por primera vez, se promedian los valores de acimut, y se toma dicho promedio $\hat{\theta}_0$ como referencia para comenzar a ubicar a los hablantes que rodean al robot. Inicialmente se supone que hay como máximo K personas hablando alrededor del robot. Entonces, se divide el escenario que rodea al robot en K regiones de igual amplitud, como se ilustra en la Figura 3.2b para el caso de $K = 8$.

De esta forma, se tiene un vector de centroides de dichas regiones, cuyas K componentes son:

$$\hat{\theta}_i = \hat{\theta}_0 + 360^{\text{o}}\frac{i}{K}; \qquad i = 0, 1, ..., K-1. \tag{3.3}$$

El vector obtenido por medio de 3.3 sirve para inicializar el algoritmo de agrupamiento k-*means*, el popular algoritmo de aprendizaje no supervisado que permitirá agrupar y clasificar cada nuevo acimut en alguna de las K regiones definidas. Con cada nuevo acimut que se alimente este algoritmo (Figura 3.2c), se actualizan los centroides, de forma tal que la clasificación de la Figura 3.2 que inicialmente distribuye equitativamente el espacio, comenzará a ajustarse de acuerdo a la densidad de las direcciones de arribo, como en la Figura 3.2d.

Por cada centroide luego de aplicar k-*means*, se generan 4 vectores de estado con los cuales inicializar el filtro de Kalman asociado a él, haciendo uso de las Ecuaciones 2.32, 2.33, 2.34, y 2.35. De ese momento en adelante, cada nuevo acimut que entra a la etapa de rastreo, es nuevamente clasificado, se actualizan los centroides para identificar cada acimut con su fuente de voz, y esta clasificación sirve para saber a cuál de los filtros de Kalman se debe asociar dicha mediciones $\mathbf{z}_i$ de la Ecuación 2.27.

Dentro del filtro de Kalman, en la etapa de predicción se estima el estado siguiente de la forma tradicional:

$$\hat{\boldsymbol{\xi}}_i^-(t) = \boldsymbol{F}\hat{\boldsymbol{\xi}}_i(t-\tau) \tag{3.4}$$

y se estima igualmente la matriz de covarianza del error:

$$\boldsymbol{P}_i^-(t) = \boldsymbol{F}\boldsymbol{P}_i(t-\tau)\boldsymbol{F}^{\text{T}} + \boldsymbol{Q} \tag{3.5}$$

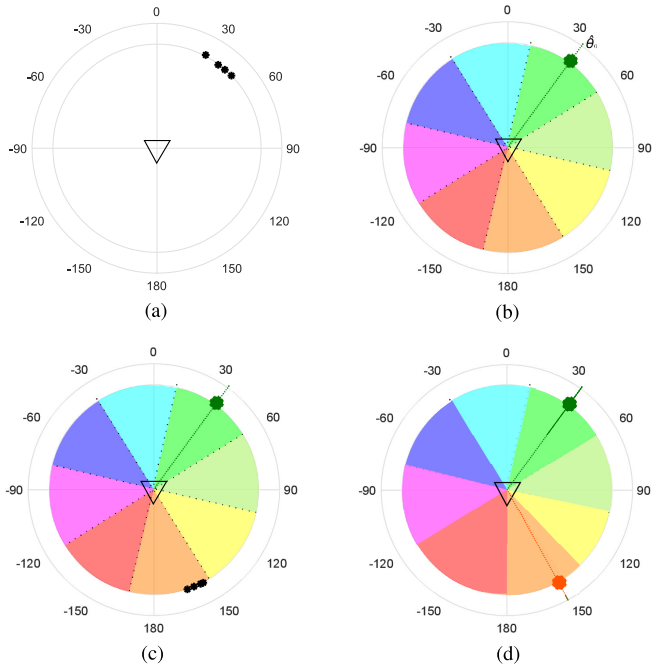


Figura 3.2: Diagrama que ilustra la clasificación de fuentes de voz previo a al rastreo con filtro de Kalman.

con lo cual a su vez se actualiza la ganancia de Kalman:

$$\boldsymbol{K}_i(t) = (\boldsymbol{P}_i^-(t)\boldsymbol{H}^{\mathrm{T}}) \times (\boldsymbol{H}\boldsymbol{P}_i^-(t)\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1} \tag{3.6}$$

Hemos decidido modelar aquí las matrices de transición (**F**) y de medición (**H**) como invariantes en el tiempo, coincidiendo con las mostradas en la sección 2.5.1.1. Sin embargo, a diferencia del modelo analizado en la sección 2.5.1.1, aquí representamos un movimiento circular acelerado, similar al de la Ecuación 2.38. Es decir, se supone que cada hablante puede moverse de una iteración a otra con una velocidad variable, con aceleración distribuida normalmente. Como consecuencia, proponemos las siguientes matrices de control del modelo y de covarianza del ruido del modelo respectivamente:

$$\mathbf{G} = \begin{bmatrix} \frac{\tau^2}{2} & 0 \\ 0 & \frac{\tau^2}{2} \\ \tau & 0 \\ 0 & \tau \end{bmatrix} \tag{3.7}$$

y

$$\mathbf{Q} = \sigma_a^2 \mathbf{G}\mathbf{G}^T = \sigma_a^2 \begin{bmatrix} \tau^4/4 & 0 & \tau^3/2 & 0 \\ 0 & \tau^4/4 & 0 & \tau^3/2 \\ \tau^3/2 & 0 & \tau^2 & 0 \\ 0 & \tau^3/2 & 0 & \tau^2 \end{bmatrix} \tag{3.8}$$

donde τ es el intervalo de tiempo de una iteración a otra del filtro de Kalman, y σ_a^2 es la varianza de la aceleración, un parámetro ajustado empíricamente hasta observar que el modelo se ajusta

mejor al comportamiento real de los hablantes. Respecto a la matriz $\mathbf{R}$ de covarianza del ruido, mantenemos la Ecuación 2.29.

Finalmente, se actualiza la estimación de los estados de cada fuente de voz a partir de la nueva ubicación $\mathbf{z}_i$:

$$\hat{\boldsymbol{\xi}}_i(t) = \hat{\boldsymbol{\xi}}_i^-(t) + \boldsymbol{K}_i(t) \times (\mathbf{z}_i(t) - \boldsymbol{H}\hat{\boldsymbol{\xi}}_i^-(t)) \tag{3.9}$$

y se actualiza igualmente la matriz de covarianza del error:

$$\boldsymbol{P}_i(t) = (\boldsymbol{I} - \boldsymbol{K}_i(t)\boldsymbol{H}) \times \boldsymbol{P}_i^-(t). \tag{3.10}$$

Como consecuencia, el nuevo estado dado por la Ecuación 3.9 es la salida de nuestro algoritmo de rastreo y corresponde a la trayectoria del i-ésimo hablante.

La implementación del algoritmo de rastreo descrito, nos permitió pasar de estimaciones aisladas y ruidosas del acimut donde se ubicaba cada una de las fuentes de voz (por ejemplo, Figura 3.3), a estimaciones clasificadas y con menor dispersión (Figura 3.4).

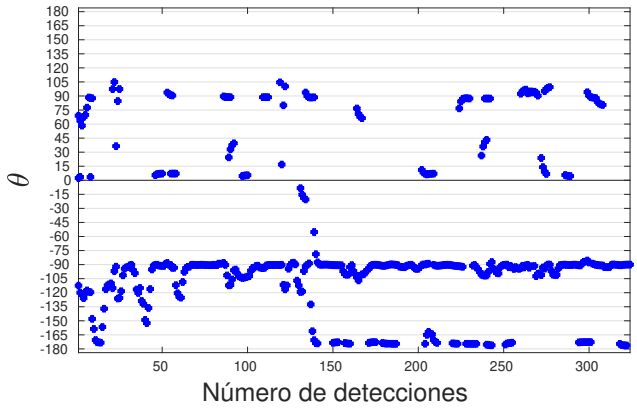


Figura 3.3: Salida del sistema de localización aplicado a un audio del corpus AIRA (ver sección 3.2.1). Escenario reverberante y ruidoso con 4 hablantes, en 0^o, 90^o, -90^o y 180^o, respectivamente.

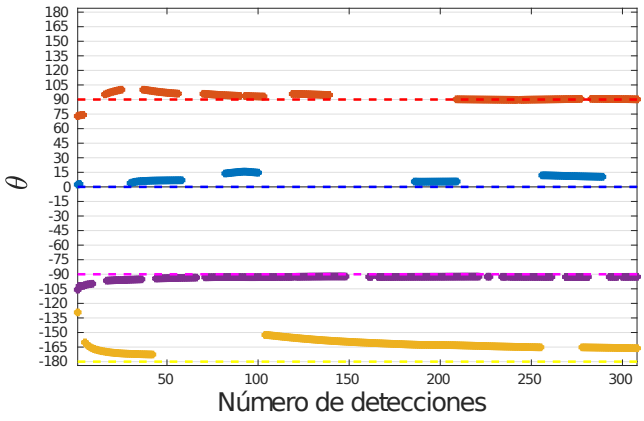


Figura 3.4: Salida del sistema de rastreo tomando como entrada las estimaciones de la Fig. 3.3.

También con hablantes móviles, el filtrado de Kalman nos permitió pasar de estimaciones aisladas y ruidosas como en la Figura 3.5, a estimaciones menor dispersión y menos sesgo como se muestra en la Figura 3.6.

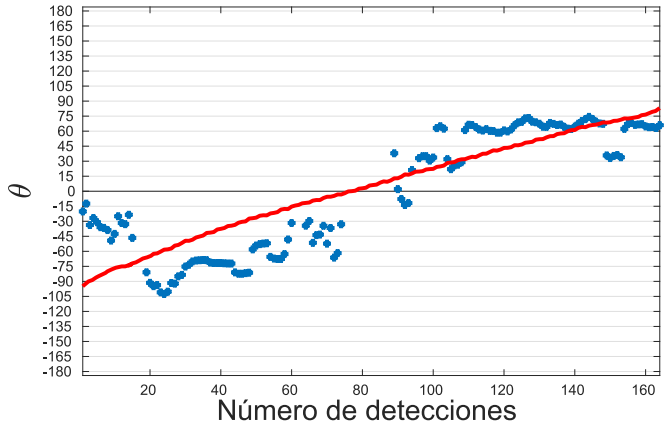


Figura 3.5: Salida del sistema de localización aplicado a un audio del corpus AIRA (ver sección 3.2.1). Escenario reverberante y ruidoso con un hablante en movimiento. La línea roja indica la trayectoria aproximada del hablante, obtenida por interpolación.

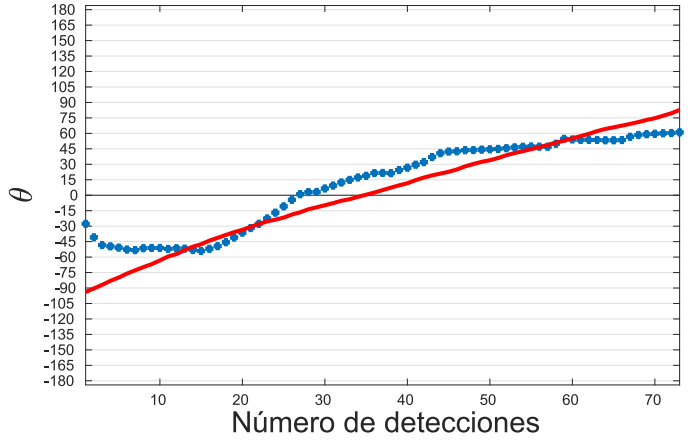


Figura 3.6: Salida del sistema de rastreo tomando como entrada las estimaciones de la Fig. 3.5. La línea roja indica la trayectoria aproximada del hablante, obtenida por interpolación.

Sin dudas el rastreo de fuentes móviles, o cuando el propio robot está en movimiento, es una tarea más compleja que detectar hablantes que están en reposo. En la Figura 3.6 se aprecia que el sesgo del estimador es todavía importante incluso después de haber pasado las primeras iteraciones de convergencia del filtro de Kalman, llegando a ser de 15° e incluso un poco más. A pesar de esto, y aunque no es objetivo de este trabajo centrarse en hablantes móviles, los resultados son relativamente buenos, ya que en lo esencial, el sistema es capaz de detectar con cierto margen de error donde se ubica el hablante en cada instante de tiempo. Un análisis más detallado de la precisión del sistema propuesto se da más adelante, en las secciones 3.2.2 y 3.2.3.

3.1.3. Descripción general de los sistema localización y rastreo trabajando juntos

En la Figura 3.7 se muestra un diagrama en bloques del sistema implementado, conformado por la etapa de localización y la de rastreo. El arreglo de tres micrófonos en configuración triangular capturan el audio del entorno, que puede contener la voz de múltiples hablantes. Una tarjeta de audio genérica se encarga de digitalizar las grabaciones de audio, para que puedan ser procesadas por una computadora. La salida de datos crudos de la interfaz de audio tiene que ser manejada por un programa que se encargue de entregar al módulo de audición robótica las muestras de audio por paquetes sincronizados y con baja latencia. Si el sistema va a correrse en tiempo real, un gestor de tramas de audio muy recomendable es *JACK Audio Connection Kit*[1]. También JACK puede configurarse para aplicar el sistema de localización a audios previamente grabados, aunque para procesar un *corpus* de audio voluminoso, puede ser más efectivo hacer un programa simple que automatice la lectura de tramas de audio y las pase al módulo de audición robótica. Para la presente investigación, ambas variantes fueron implementadas, y pueden ser consultadas en el repositorio correspondiente en GitHub (`https://github.com/lmiguelgato/DAP_project`).

Una vez dentro del módulo de audición robótica, las tramas de audio se rellenan con ceros (*zero-padding*), y se llevan al dominio de la frecuencia con la transformada rápida de Fourier implementada en FFTW. El relleno con ceros es necesario si se quiere obtener en la siguiente etapa la correlación aperiódica, y no una correlación con solapamiento debido a la respuesta periódica que produce la transformada inversa [85].

Los tres canales de audio en el dominio de la frecuencia se alimentan por pares a una etapa de cálculo de la correlación cruzada con transformación de fase como se indicó en la Ecuación 2.19, y anulando las componentes de frecuencias fuera del intervalo de interés $[f_{\min}, f_{\max}]$. De manera empírica, nuestras pruebas en escenarios reales nos indicaron que seleccionar $f_{\min} = 1$ kHz y $f_{\max} = 4$ kHz era un intervalo apropiado para localizar fuentes de voz.

También se usa la implementación de FFTW para regresar al dominio del tiempo. La salida de la transformada inversa de Fourier contiene los coeficientes de correlación de Pearson, así que lo siguiente es detectar la ubicación y la magnitud de los picos de correlación para cada par de canales. Las magnitudes se emplean para verificar si hay alguien hablando, o al menos si hay una nivel de energía lo suficientemente alto en la banda $[f_{\min}, f_{\max}]$. Si no hay actividad vocal, se ignora la trama de audio actual. Empíricamente, comprobamos que un umbral de $\frac{120}{N_{\text{FFT}}}$, proporciona el mejor desempeño en la mayoría de los escenarios reales en donde se hicieron pruebas. En este caso el tamaño de la FFT es $N_{\text{FFT}} = 2048$. Si hay actividad vocal, entonces los índices de los picos de correlación para la trama de audio actual se mapean a ángulos de acimut, haciendo uso de la Ecuación 2.17. Como esta tiene dos soluciones dentro del intervalo $[0, \pi]$, entonces se obtienen tres pares de ángulos. Se deben descartar aquellos ángulos que no son consistentes con el resto, tomando un marco común de referencia a partir de las Ecuaciones 3.1 y 3.2, y verificando que se encuentren dentro del rango de tolerancia definido. De acuerdo a nuestras pruebas prácticas, de forma empírica verificamos que una tolerancia de $\pm 15^{\text{o}}$ es suficiente para los entornos con una reverberación moderada. La media de los dos o tres ángulos que cumplan esas restricciones son el resultado final del sistema de localización.

Como se aprecia en la Figura 3.7, la salida del sistema de localización $\theta(t)$ alimenta a un

[1] *JACK Audio Connection Kit* es un software libre multiplataforma para gestionar tramas de audio sincronizadas y con baja latencia, que permite conectar varias aplicaciones a un dispositivo de audio, y les permite intercambiar tramas de audio entre sí (`https://jackaudio.org`).

registro de desplazamiento que tiene un tamaño de memoria $N_{\text{hist}} \times K$. El algoritmo k-*means* encargado de clasificar y agrupar cada nuevo acimut con su correspondiente hablante está inicializado con el conjunto de centroides dado por la Ecuación 3.3, y con cada nueva observación que entra al registro de desplazamiento, se recalculan los centroides con el contenido del registro de desplazamiento que tenga observaciones pasadas. La salida de la etapa de clasificación es un número que indica a qué hablante corresponde el acimut actual. Ese identificador sirve además para seleccionar el filtro de Kalman correspondiente a ese hablante, dentro del cual se implementan las Ecuaciones 3.4 a 3.10. A la salida de los filtros de Kalman se obtienen estimaciones menos ruidosas que la que se obtienen a la salida del localizador, y gracias a la clasificación previa, se conoce a qué hablante pertenece cada estimación. En la Figura 3.7 se ha representado la salida del sistema de rastreo como una secuencia del tipo $\hat{\theta}_1(t'), \hat{\theta}_2(t'), ..., \hat{\theta}_K(t')$, aunque la cantidad de elementos a la salida dependerá de cuántos filtros de Kalman se hayan seleccionado de acuerdo al algoritmo de clasificación k-*means*. Si hay un solo hablante, el sistema de rastreo es capaz de notarlo, y selecciona un solo filtro, y por tanto hay un solo acimut salida en cada iteración. Los tiempos en que se obtiene una salida los hemos nombrado t' para diferenciarlos del tiempo t en que entró un ángulo desde el sistema de localización. En general $t' \geq t$, porque aunque la localización $\theta(t)$ se calcula en el mismo intervalo de tiempo que dura una trama de audio, el sistema de rastreo introduce un retardo mientras se vayan llenando las N_{hist} primeras muestras del registro de desplazamiento para el filtro de Kalman correspondiente.

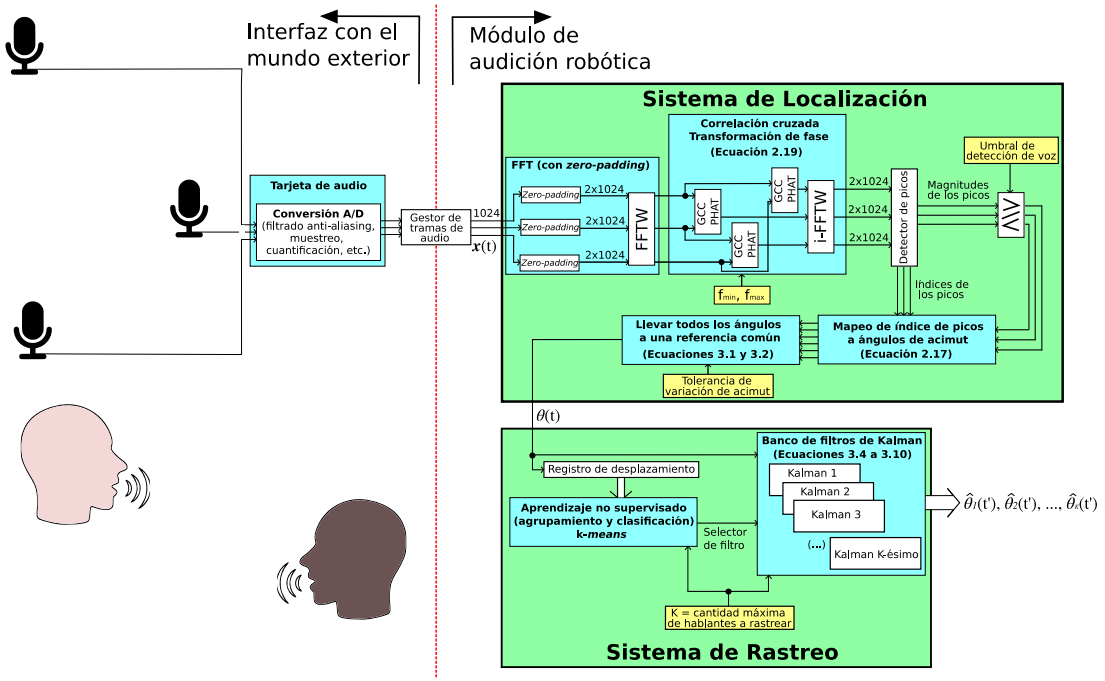


Figura 3.7: Diagrama en bloques del sistema implementado, conformado por la etapa de localización y la de rastreo.

3.2. Evaluación y análisis de resultados

Para evaluar el desempeño del sistema de localización y rastreo propuesto en este trabajo, y realizar la comparación con otros algoritmos del Estado del Arte, es importante poder obtener resultados repetibles, y realizar siempre las comparaciones bajo las mismas condiciones para que sea una comparación justa e ilustre verdaderamente los méritos o desventajas de nuestra propuesta.

Por estos motivos, se decidió evaluar el sistema de localización y rastreo propuesto, así como un competidor en el Estado del Arte, basándose en un *corpus* de grabaciones de audio obtenidas con arreglos de micrófonos en múltiples escenarios y ante la presencia de un variado número de hablantes activos.

Se emplearon las mismas métricas de desempeño para evaluar ambos algoritmos, y tener así una comparación justa entre ambos. Dichas métricas son las más usadas en el Estado del Arte, y son las descritas en la sub-sección 2.4.1 del presente trabajo.

3.2.1. Selección del *corpus* de audio para la evaluación

Para la presente investigación se identificaron varios *corpus* de audio para tareas de localización y rastreo de fuentes sonoras, y específicamente personas hablando. Por las características de este trabajo, y dadas las métricas de desempeño que se desean evaluar, son de nuestro interés únicamente aquellos *corpus* que cumplen los siguientes requisitos:

1. proporcionan grabaciones de audio multi-canal, que contengan como mínimo 3 canales asociados a 3 micrófonos no alineados, de forma tal que puedan formar un triángulo sobre un plano horizontal;
2. que correspondan a grabaciones en múltiples escenarios prácticos con cierta presencia de reverberación y ruido acústico;
3. que se encuentren presentes en las grabaciones un variado número de hablantes activos;
4. preferiblemente, que dispongan también de grabaciones con hablantes en movimiento, para evaluar el algoritmo de rastreo en estas condiciones.

Teniendo en cuenta estos requisitos, se identificaron varios *corpus* compatibles. Sin embargo, en algunos casos no fue posible acceder a algunos de estos, ya sea por que no eran públicos, o porque fueron retirados de Internet previo a nuestra investigación. A continuación describimos tres a los cuales sí pudimos acceder:

1. "AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking" [86]. Este contiene grabaciones de audio realizadas con dos arreglos circulares de 8 micrófonos cada uno. Las grabaciones se realizaron en un escenario práctico (un salón de clases), ante la presencia de un número variable de hablantes activos: desde uno hasta tres, en algunos casos estáticos y en otros en movimiento. Cuando hay más de un hablante, se manifiestan algunos segmentos de audio con oclusión, es decir, con más de un hablante activo de forma simultánea. En total, consta de 42 grabaciones, de duración entre 14 segundos y 9 minutos cada una, para un total de 85 minutos de audio.
2. "The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking" [67]. Es el corpus de audio usado por el Comité Técnico de Audio y Procesamiento de

Señales Acústicas de IEEE durante el reto para competir en las áreas de localización y rastreo de fuentes de voz. Este contiene grabaciones de audio realizadas con cuatro arreglos diferentes de micrófonos: un arreglo plano compuesto por 15 micrófonos, un arreglo de 2 micrófonos colocados en los oídos de la cabeza de un maniquí, un arreglo cuasi-esférico de 12 micrófonos, y un arreglo esférico de 32 micrófonos. Este último sí cumple con los requisitos del sistema propuesto en la presente investigación, y dispone de alrededor de 10 minutos de grabaciones. Las grabaciones se realizaron en un escenario práctico (un laboratorio de computación con tiempo de reverberación de 500 ms aproximadamente y cierto nivel de ruido acústico causado, por ejemplo, por el sonido del tráfico en las afueras del local). En las grabaciones hay uno o varios hablantes activos, en algunos casos estáticos y en otros en movimiento. Cuando hay más de un hablante, se manifiestan algunos segmentos de audio con oclusión, es decir, con más de un hablante activo de forma simultánea.

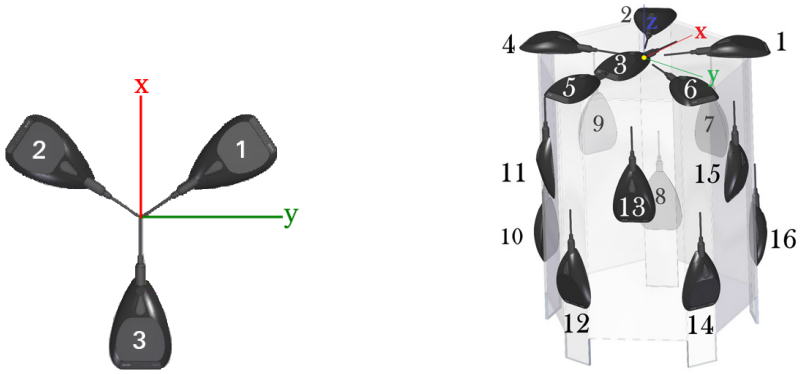
3. AIRA: "Acoustic interactions for robot audition: A corpus of real auditory scenes" [1]. Este fue grabado en seis escenarios prácticos distintos, con niveles de ruido y tiempos de reverberación conocidos y bastante diferentes entre sí. Adicionalmente, hay grabaciones en la cámara anecoica con tiempo de reverberación inferior a 10 ms descrita en [87]. Para cada escenario hay grabaciones con entre una y cuatro personas hablando simultáneamente. Contiene grabaciones obtenidas con dos arreglos de micrófonos distintos: un arreglo de 3 micrófonos en los vértices de un triángulo equilátero como se muestra en la Figura 3.8a, y un arreglo de 16 micrófonos con la distribución mostrada en la Figura 3.8b, tres de dichos 16 micrófonos forman un triángulo isósceles. Por tanto, ambos arreglos son compatibles con el sistema propuesto en la presente investigación. En el caso de las fuentes de voz estáticas, las voces de dichas personas se reproducen usando hardware profesional de alta calidad, y las voces son de personas reales, en idioma español, tomadas del corpus "DIMEx100" [88]. En el caso de las fuentes de voz móviles, se realizaron grabaciones de voces de personas reales en movimiento, cuyas posiciones en cada instante de tiempo fueron obtenidas con un sistema de rastreo por láser, o en algunos casos haciendo una estimación menos confiable a partir de conocer la posición inicial y la posición final. Cada grabación dura entre 30 y 32 segundos aproximadamente, para un total de 790 grabaciones con 16 micrófonos y 439 con 3 micrófonos.

   Como este corpus fue creado por algunos de los miembros del grupo de investigación donde se desarrolló este trabajo, disponemos de información más detallada sobre sus características. A continuación describimos los siete escenarios en que se realizaron las grabaciones. Más detalles pueden encontrarse en [1].

   a) **Cámara anecoica del Laboratorio de Acústica y Vibraciones del ICAT [1].** Descrita con mayor nivel de detalles en [87]. Se realizaron grabaciones con ambos arreglos. Foto del escenario de pruebas mostrada en la Figura 3.9.

      - Nivel de presión sonora del ruido promedio: 0.13 dB (ruido de fondo)
      - Tiempo de reverberación promedio: inferior a 10 ms
      - Relación señal a ruido promedio: 43 dB

[1] ICAT: Instituto de Ciencias Aplicadas y Tecnología, es una entidad académica perteneciente a la Coordinación de la Investigación Científica de la UNAM (https://www.icat.unam.mx).



(a) Arreglo triangular de 3 micrófonos.

(b) Arreglo tridimensional de 16 micrófonos.

Figura 3.8: Arreglos de micrófonos usados para obtener el corpus AIRA [1].

Figura 3.9: Escenario de pruebas dentro de la cámara anecoica del Laboratorio de Acústica y Vibraciones del ICAT. Foto tomada de [2].

Escenarios prácticos:

b) **Cafetería estudiantil dentro del campus central de la UNAM.** Grabaciones realizadas con el arreglo tridimensional durante un horario de alta concurrencia de estudiantes. Foto del escenario de pruebas mostrada en la Figura 3.10.

- Nivel de presión sonora del ruido promedio: 71 dB (personas hablando alrededor, y otros ruidos asociados a la actividad normal de la cafetería)
- Tiempo de reverberación promedio: 270 ms
- Relación señal a ruido promedio: 16 dB

Figura 3.10: Escenario de pruebas dentro de una cafetería estudiantil del campus central de la UNAM. Foto tomada de [2].

c) **Tienda UNAM.** Grabaciones realizadas con el arreglo tridimensional durante un horario de concurrencia habitual de clientes. Foto del escenario de pruebas mostrada en la Figura 3.11.

- Nivel de presión sonora del ruido promedio: 63 dB (personas hablando alrededor, anuncios publicitarios ocasionales, y otros ruidos asociados a la actividad normal de la tienda)
- Tiempo de reverberación promedio: 160 ms
- Relación señal a ruido promedio: 17 dB

d) **Pasillo entre oficinas, al interior del edificio principal del IIMAS.** Grabaciones realizadas con el arreglo triangular durante un horario de concurrencia habitual de trabajadores y estudiantes.

- Nivel de presión sonora del ruido promedio: 48 dB (personas hablando dentro de las oficinas, ruido propio del motor del robot Golem-II [89] usado en las grabaciones)
- Tiempo de reverberación promedio: 210 ms
- Relación señal a ruido promedio: 10 dB

e) **Oficina A, al interior del edificio principal del IIMAS.** Grabaciones realizadas con ambos arreglos durante un horario de concurrencia habitual de trabajadores y estudiantes. Foto del escenario de pruebas mostrada en la Figura 3.12. Las fuentes de voz son altavoces en reposo.

- Nivel de presión sonora del ruido promedio: 52 dB (personas hablando dentro de oficinas vecinas, ruido de ventiladores)
- Tiempo de reverberación promedio: 200 ms
- Relación señal a ruido promedio: 21 dB

Figura 3.11: Escenario de pruebas dentro de Tienda UNAM. Foto tomada de [2].

Figura 3.12: Escenario de pruebas dentro de la Oficina A. Foto tomada de [2].

f) **Oficina B, al interior del edificio principal del IIMAS.** Grabaciones realizadas con el arreglo triangular durante un horario de concurrencia habitual de trabajadores y estudiantes. Las fuentes de voz son altavoces en reposo.

- Nivel de presión sonora del ruido promedio: 42 dB (personas hablando dentro de oficinas vecinas, ruido de ventiladores)
- Tiempo de reverberación promedio: 420 ms
- Relación señal a ruido promedio: 20 dB

g) **Oficina C, al interior del edificio principal del IIMAS.** La misma Oficina A, pero ahora las fuentes de voz son personas reales en movimiento. Grabaciones realizadas con el arreglo triangular solamente.

- Nivel de presión sonora del ruido promedio: 52 dB (personas hablando dentro de oficinas vecinas, ruido de ventiladores)

- Tiempo de reverberación promedio: 200 ms
- Relación señal a ruido promedio: 14 dB

Como se puede apreciar, este último *corpus* es el que proporciona un conjunto más variado de escenarios, incluyendo niveles de relación señal a ruido de entre 10 dB y 43 dB, y tiempos de reverberación desde menos de 10 ms hasta 420 ms. Precisamente esa variedad nos permite garantizar que las evaluaciones que se realicen con este *corpus* son más fiables para generalizar los resultados, que si se restringiera nuestro análisis exclusivamente a ciertos escenarios en particular, como sucede con los corpus AV16.3 y el del LOCATA. Es por ese motivo que nos enfocamos en este Capítulo en la evaluación sobre el *corpus* AIRA.

3.2.2. Resultados obtenidos dentro de la cámara anecoica

Los resultados que se muestran a continuación se obtuvieron fijando el umbral de detección que proporcionaba una tasa de inserción (I) igual o inferior a 10 %, tanto para ODAS[1] como para el sistema propuesto [2]. Esto se logró, para cada sistema por separado, ajustando el umbral de detección (umbral de energía de las señales detectadas) de forma empírica, hasta que la tasa de inserción promedio era la deseada. El objetivo de este procedimiento es estandarizar los resultados, pues en ambos sistemas de localización, a un menor umbral aumenta la tasa de detección, pero también aumenta la tasa de inserción o falsas alarmas. Por el contrario, aumentar el umbral reduce la tasa de inserción o falsas alarmas con el costo de reducir también la tasa de detección. Para que las comparaciones fueran justas y referidas a un mismo criterio de desempeño, se tomó para cada una de las tablas a continuación el umbral que proporcionaba la misma tasa de inserción. En este trabajo se considera que ha ocurrido una inserción de una detección falsa cuando el error absoluto del acimut es superior a 15°.

Comenzando por las mediciones en cámara anecoica, en la Figura 3.13 se muestra, para uno, dos, tres y cuatro hablantes activos, la tasa de detección de hablantes correspondiente al algoritmo de localización propuesto en este trabajo, y haciendo uso de un arreglo de 3 micrófonos en los vértices de un triángulo (en el caso del arreglo tridimensional, se usaron los micrófonos 1 al 3 de la Figura 3.8b). Se considera una detección como válida si el algoritmo detecta un hablante en un intervalo de acimut de ±15° alrededor del acimut en que realmente se ubicaba una persona hablando. Las detecciones fuera de este intervalo son consideradas falsas alarmas y entran en la tasa de inserción fijada, y por tanto dichas detecciones no han sido incluidas en los gráficos de pastel mostrados a en la presente sección y la siguiente.

En el 100 % de las grabaciones donde había un hablante activo, este fue detectado correctamente. En ningún caso este hablante pasó desapercibido por el sistema de localización. Por otro lado, en el 94 % de las grabaciones donde había dos personas hablando, los dos hablantes fueron detectados, y solamente en el 6 % restante se detectó uno de los dos. En ningún caso se dejaron de detectar los dos hablantes simultáneamente. Por otro lado, en el 80 % de las grabaciones donde había tres personas hablando, los tres hablantes fueron detectados, y en el 20 % restante

[1] Todas las evaluaciones de ODAS son basadas en su versión correspondiente al último *commit* disponible a la hora de escribir esta tesis: 2ed307b, en su página de GitHub (https://github.com/introlab/odas). Versiones anteriores o posteriores podrían proporcionar resultados diferentes a los mostrados en el presente Capítulo.

[2] De igual forma, todos los resultados obtenidos con nuestra propuesta corresponden a la versión del último *commit* realizado hasta el momento de escribir esta tesis: cb82836, en nuestra página de GitHub (https://github.com/lmiguelgato/DAP_project). Versiones anteriores o posteriores podrían proporcionar resultados diferentes a los mostrados en el presente Capítulo.

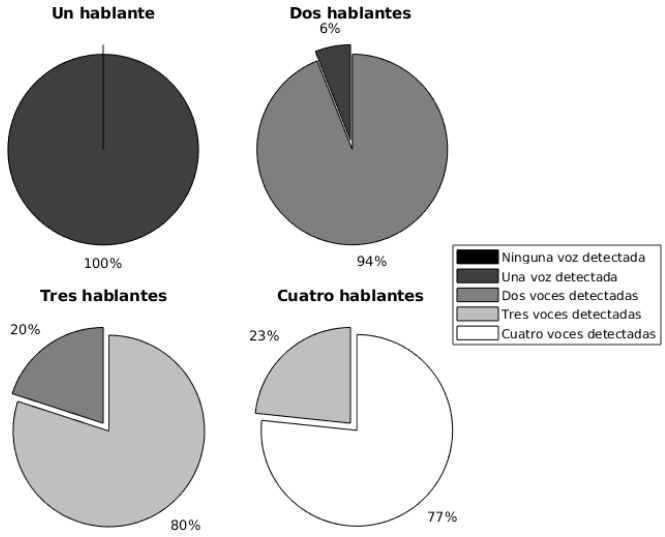


Figura 3.13: Tasa de detección de hablantes en cámara anecoica, usando el algoritmo de localización propuesto.

se detectaron dos de los tres. En ningún caso se detectó solamente uno o ningún hablante. En el caso más complejo, con cuatro personas hablando simultáneamente, en el 77 % de las grabaciones se detectaron correctamente los cuatro hablantes, y en el 23 % restante se detectaron tres de los cuatro hablantes. En ningún caso se detectó un número menor de hablantes.

Veamos ahora en la Figura 3.14 estas mismas métricas de desempeño, exactamente bajo las mismas condiciones, pero esta vez aplicando el algoritmo de localización de ODAS usando los mismos tres micrófonos que fueron usados por el algoritmo propuesto. Es decir, usando las mismas grabaciones en cámara anecoica, y con el mismo arreglo de 3 micrófonos, vamos a comparar las tasas de detección de hablantes.

En el 20 % de las grabaciones donde había un hablante activo, este fue detectado correctamente. En el 80 % restante, ningún hablante fue detectado dentro de un intervalo de $\pm 15^{o}$ alrededor del acimut en que realmente se ubicaba la persona hablando. Por otro lado, en el 80 % de las grabaciones donde había dos personas hablando, solamente un hablante fue detectado, y en el 20 % restante no se detectó ningún hablante. En ningún caso se detectaron simultáneamente los dos hablantes que estaban activos. Por otro lado, en el 40 % de las grabaciones donde había tres personas hablando, los tres hablantes fueron detectados. En el 48 % de las grabaciones se detectó un hablante solamente, y en el 12 % restante, se detectaron dos hablantes. En ningún caso hablaron las tres personas sin que fuera detectada al menos una de ellas. En el caso más complejo, con cuatro personas hablando simultáneamente, solamente en el 13 % de las grabaciones se detectaron correctamente los cuatro hablantes. En el 57 % de las grabaciones se detectaron tres de los cuatro hablantes, y en el 30 % restante se detectaron solamente dos personas hablando. En ningún caso se detectó uno o ningún hablante.

Por simple inspección de estos resultados, podemos concluir que bajo exactamente las mismas condiciones, nuestra propuesta supera ampliamente el desempeño que proporciona ODAS al usar únicamente tres micrófonos. En particular, resalta el hecho de que nuestra propuesta no dejó pasar ningún hablante sin que fuera detectado. A diferencia de ODAS, que ante uno o dos hablantes, determinó en una parte importante de las grabaciones que no había ningún

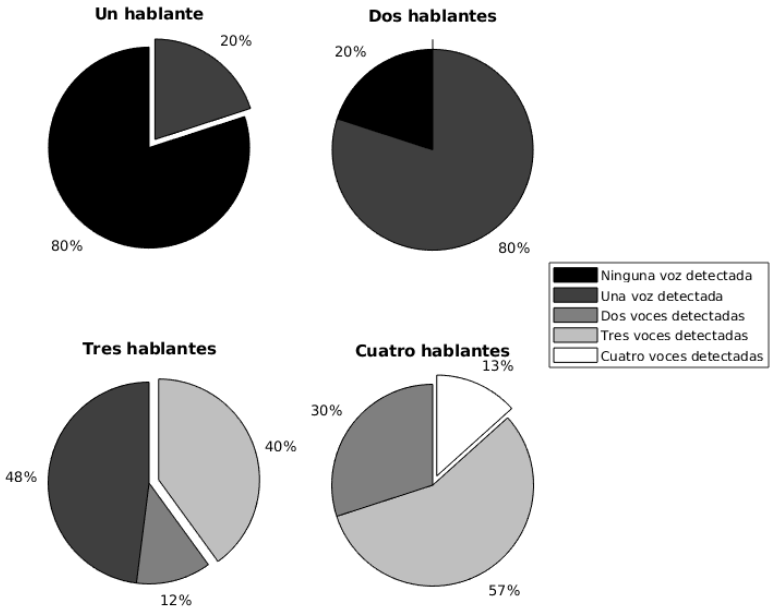


Figura 3.14: Tasa de detección de hablantes en cámara anecoica, usando el algoritmo de localización de ODAS con 3 micrófonos.

hablante activo. En general, nuestra propuesta detecta un mayor número de personas hablando simultáneamente que ODAS, y supera con creces a ODAS en la detección del total de hablantes activos.

Para que la tasa de detección de hablantes que proporciona ODAS sea superior a la de nuestra propuesta, necesita emplear un número significativamente mayor de micrófonos. Por ejemplo, considerando la variante recomendada por sus autores, que es usando 8 micrófonos, llegamos a los resultados que se muestra en la Figura 3.15 al operar en una cámara anecoica.

Consideremos ahora métricas de desempeño que evalúan la precisión de los algoritmos de localización. En la Tabla 3.1 se recogen 5 métricas de desempeño que han sido usadas para evaluar el algoritmo de localización que se emplea en ODAS usando 8 micrófonos. Son las mismas métricas que se recogen en la Tabla 3.2 para comparar contra el algoritmo de localización propuesto.

Los errores absoluto y cuadrático medio usando el sistema propuesto son inferiores a 5° ante la presencia de uno o dos personas hablando de forma simultánea. De hecho, la tasa de detecciones con un error absoluto inferior a 5° (la métrica `E5`) es prácticamente igual a la que se logra con ODAS en esos casos. Incluso las tasas de detección `E10` y `E15` indican que, en promedio, el algoritmo propuesto proporciona una mayor precisión al estimar el acimut de los hablantes. Sin embargo, para tres o cuatros hablantes, el desempeño del sistema propuesto es notablemente inferior. ODAS proporciona en esos casos una mayor precisión en la localización de los hablantes.

Podemos entonces llegar a una conclusión parcial hasta este punto. A pesar de que estamos comparando con ODAS, que contiene algunos de los mejores algoritmos del Estado del Arte, y opera con 8 micrófonos, nuestra propuesta está logrando un desempeño similar, e incluso ligeramente superior para hasta dos hablantes, y usando solamente 3 micrófonos.

Ahora bien, si comparamos nuestra propuesta con el algoritmo de localización de ODAS

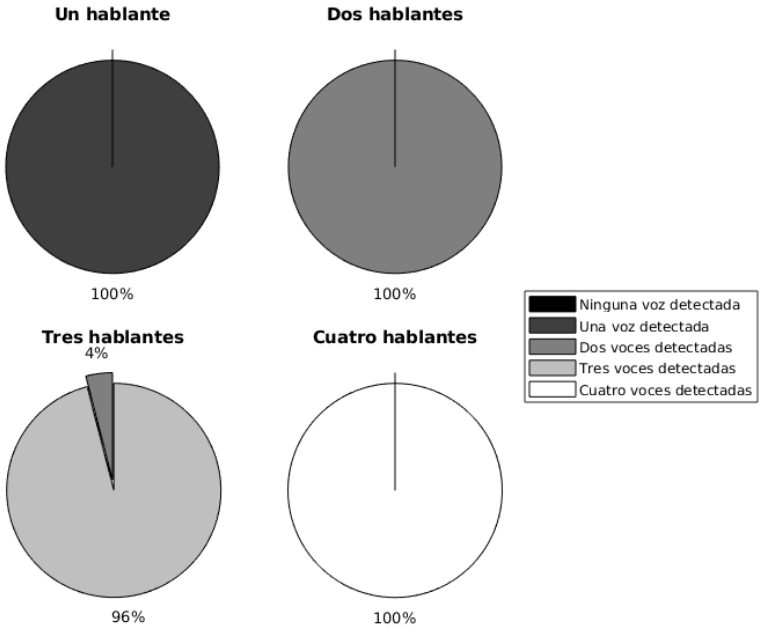


Figura 3.15: Tasa de detección de hablantes en cámara anecoica, usando el algoritmo de localización de ODAS con 8 micrófonos.

usando solamente 3 micrófonos en arreglo triangular, para que estén en igualdad de condiciones, obtenemos unos resultados interesantes. Manteniendo igualmente una tasa de inserción igual o inferior a 10 %, a continuación vamos a analizar cómo se comportan las métricas de desempeño que estamos considerando. En la Tabla 3.3 se muestran los resultados al evaluar ODAS usando los mismos 3 micrófonos en arreglo triangular que hemos usado con nuestra propuesta. A primera vista, comparando directamente con nuestros resultados de la Tabla 3.2, podemos notar que ODAS muestra un desempeño similar, e incluso en algunos casos superior a nuestra propuesta. Pero incluso son resultados superiores a los de el propio ODAS usando 8 micrófonos en la Tabla 3.1. Por ejemplo, para uno y dos hablantes, los errores absolutos y cuadráticos medios son inferiores, y la tasa de detección E5 es incluso del 100 % en ambos casos. Estos números indican que usar ODAS con 3 micrófonos es mucho mejor que usar ODAS con 8 micrófonos. Evidentemente estamos ante una contradicción que no podemos ignorar.

Luego de analizar minuciosamente cada una de las localizaciones que se obtuvieron usando ODAS con 3 micrófonos, detectamos una anomalía del algoritmo de localización que explica los resultados contradictorios de la Tabla 3.3. En efecto, identificamos un sesgo en la localización usando solo 3 micrófonos, que siempre arrojaba la presencia permanente de una fuente de voz alrededor de $\theta = 0^{o}$. Esto sucedía en todos los escenarios, independientemente del número de fuentes realmente presentes, y de sus ubicaciones. O sea, ODAS siempre detecta una fuente de voz "fantasma" justo enfrente del arreglo de micrófonos triangular. Casualmente, en el corpus de audio de AIRA, usado para evaluar algoritmos de localización, en cada recinto donde se realizaron las grabaciones, y para cada número de fuentes, siempre hay dos o más escenarios en donde hay efectivamente una fuente en $\theta = 0^{o}$. Como en la evaluación de las métricas de desempeño se están analizando únicamente las detecciones válidas (y no las inserciones o falsas alarmas), entonces ODAS con 3 micrófonos siempre fue capaz de adivinar por azar que había

alguna fuente enfrente, en alguno de los escenarios bajo los cuales se puso a prueba. Cuando había 3 o 4 hablantes, los valores de acimut estimados por ODAS tienen mayor dispersión alrededor del sesgo de 0°, pero aún así, los resultados siguen siendo relativamente buenos gracias a dicho sesgo.

Tabla 3.1: Métricas de desempeño del algoritmo de localización de ODAS operando en una cámara anecoica con 8 micrófonos (micrófonos 1 al 8 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (E5):	Tasa de detección (E10):	Tasa de detección (E15):
1	3.5°	3.9°	88.5 %	11.4 %	0.1 %
2	4.7°	5.5°	71.9 %	16.0 %	12.1 %
3	4.1°	4.8°	75.7 %	20.1 %	4.2 %
4	3.8°	4.6°	81.3 %	12.9 %	5.8 %

Tabla 3.2: Métricas de desempeño del algoritmo de localización propuesto operando en una cámara anecoica con 3 micrófonos en arreglo triangular (micrófonos 1 al 3 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (E5):	Tasa de detección (E10):	Tasa de detección (E15):
1	2.4°	2.8°	88.0 %	10.5 %	1.5 %
2	3.7°	4.3°	72.1 %	25.1 %	2.8 %
3	6.5°	6.9°	42.9 %	23.7 %	33.4 %
4	5.3°	6.0°	57.1 %	18.6 %	24.3 %

Tabla 3.3: Métricas de desempeño del algoritmo de localización de ODAS operando en una cámara anecoica con 3 micrófonos en arreglo triangular (micrófonos 1 al 3 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (E5):	Tasa de detección (E10):	Tasa de detección (E15):
1	2.5°	2.5°	100.0 %	0.0 %	0.0 %
2	3.9°	4.0°	100.0 %	0.0 %	0.0 %
3	6.0°	6.0°	51.6 %	32.2 %	16.2 %
4	6.0°	6.1°	55.6 %	18.4 %	26.0 %

Para probar nuestra hipótesis, evaluamos nuevamente a ODAS con 3 micrófonos esta vez ignorando las fuentes que había en $\theta = 0^o$. Es decir, no tuvimos en cuenta que había una fuente de voz enfrente suyo en los casos en que realmente sí la había. De esta forma, el sesgo del algoritmo que detecta fuentes fantasmas iba a arrojar únicamente inserciones que no entran dentro de las métricas de desempeño. Al hacer esto, obtuvimos efectivamente resultados pésimos, ya que ODAS no fue capaz de detectar ninguna fuente cuando había una o dos. Cuando había más de dos fuentes, los mejores resultados que se pudieron obtener fueron con una tasa de inserción de 20 %, puesto que fijándola a 10 % para estar en igualdad de condiciones con las pruebas anteriores, no se detectaba ninguna fuente en absoluto. Estos resultados se resumen en la Tabla 3.4.

Aún los resultados de la Tabla 3.4 tienen un comportamiento bastante sospechoso, con los errores absolutos tan concentrados en el intervalo `E10`, probablemente por otro sesgo interno del algoritmo de localización. En cualquier caso, como ya la tasa de inserción es mayor a nuestro objetivo inicial, descartamos por completo la solución de ODAS con 3 micrófonos. Hemos comprobado que ODAS no está pensado para operar con tan pocos micrófonos.

Finalmente, podemos concluir que ante la ausencia de reverberación, usar ODAS con 8 micrófonos logra un desempeño bastante similar a nuestra propuesta, solamente superándolo cuando hay 3 o más hablantes. Sin embargo, la ganancia en precisión en estos casos no parece justificar un aumento tan significativo de micrófonos y de costo computacional. Por otro lado, si reducimos el número de micrófonos de ODAS a 3, para estar en igualdad de condiciones que nuestra propuesta, vemos que su desempeño es notablemente inferior al nuestro en todos los casos analizados.

Tabla 3.4: Luego de ignorar el sesgo en 0^o de acimut, métricas de desempeño del algoritmo de localización de ODAS operando en una cámara anecoica con 3 micrófonos en arreglo triangular (micrófonos 1 al 3 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (`E5`):	Tasa de detección (`E10`):	Tasa de detección (`E15`):
1	-	-	0.0 %	0.0 %	0.0 %
2	-	-	0.0 %	0.0 %	0.0 %
3	7.3^o	7.3^o	0.0 %	99.2 %	0.8 %
4	5.9^o	5.9^o	0.0 %	100.0 %	0.0 %

3.2.3. Resultados obtenidos en los escenarios prácticos

Al igual que en la sección anterior, los resultados que se muestran a continuación se obtuvieron fijando el umbral de detección que proporcionaba una tasa de inserción (`I`) igual o inferior a 10 %, tanto para ODAS como para el sistema propuesto. Se replicaron los mismos experimentos realizados en cámara anecoica, pero esta vez el arreglo de micrófonos se encontraba dentro de los escenarios prácticos descritos en la sección 3.2.1. La principal utilidad de esta comparación es verificar que nuestra propuesta también supera a ODAS ante la presencia de reverberación y ruido acústico típicos de escenarios reales.

En la Figura 3.16 se muestra, para uno, dos, tres y cuatro hablantes activos, la tasa de detección de hablantes correspondiente al algoritmo de localización propuesto en este trabajo.

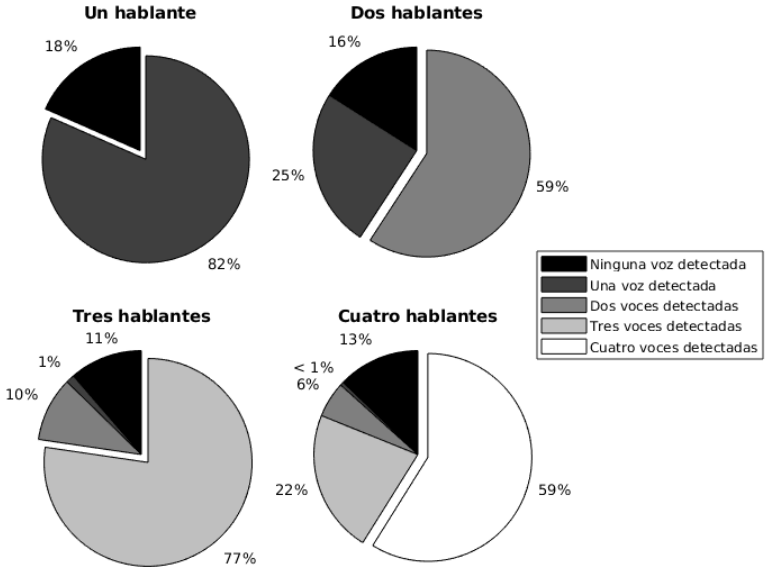


Figura 3.16: Tasa de detección de hablantes en escenarios prácticos, usando el algoritmo de localización propuesto.

Al igual que en la sección anterior, se considera una detección como válida si el algoritmo detecta un hablante en un intervalo de acimut de $\pm 15^{o}$ alrededor del acimut en que realmente se ubicaba una persona hablando. Las detecciones fuera de este intervalo son consideradas falsas alarmas y entran en la tasa de inserción fijada, y por tanto dichas detecciones no han sido incluidas en los gráficos de pastel mostrados a continuación.

En el 82 % de las grabaciones donde había un hablante activo, este fue detectado correctamente. En el 18 % restante de las grabaciones, el algoritmo propuesto no fue capaz de detectar dentro del intervalo de tolerancia la presencia de la persona que hablaba. Ya vemos aquí los efectos de la reverberación y el ruido acústico sobre los resultados. Anteriormente, en cámara anecoica, no hubo ni un solo caso en que el hablante no fuese detectado dentro del intervalo de acimut esperado.

Por otro lado, en el 59 % de las grabaciones donde había dos personas hablando, los dos hablantes fueron detectados. Aquí también se ve una reducción significativa en la tasa de detección, atribuible a los efectos de reverberación y ruidos presentes en los escenarios prácticos. En el 25 % de las grabaciones se detectó uno de los dos, y en el 16 % restante no se detectó ninguna voz que correspondiera con alguno de los dos hablantes presentes.

Por otro lado, en el 77 % de las grabaciones donde había tres personas hablando, los tres hablantes fueron detectados. Esta tasa de detección es similar a la obtenida en cámara anecoica. Sin embargo, en el 10 % de las grabaciones se detectaron solamente dos de los tres, en el 1 % se detectó la voz voz de una única persona, y en el 11 % restante el algoritmo propuesto no detectó ningún hablante dentro del rango de tolerancia en la estimación de acimut. Ya estos resultados son bastante diferentes de los obtenidos en cámara anecoica.

En el caso más complejo, con cuatro personas hablando simultáneamente, en el 59 % de las grabaciones se detectaron correctamente los cuatro hablantes, en el 22 % se detectaron tres hablantes solamente, en el 6 % se detectaron correctamente dos de las personas hablando, y en menos del 1 % de los casos se detectó una única persona. En el 13 % restante de las grabaciones

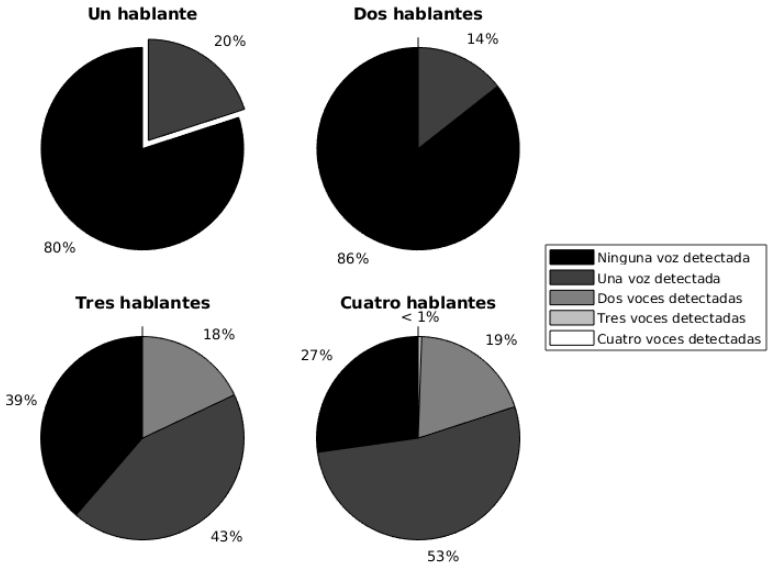


Figura 3.17: Tasa de detección de hablantes en escenarios prácticos, usando el algoritmo de localización de ODAS con 3 micrófonos.

al algoritmo propuesto no fue capaz de detectar dentro del margen de tolerancia permitido la presencia de ningún hablante. Aquí vemos nuevamente un deterioro notable en la tasa de detección de hablantes respecto a los resultados obtenidos dentro de la cámara anecoica. Basta comparar visualmente las Figuras 3.13 y 3.16 para notar la diferencia notable que existe al existir reverberación y ruido en las grabaciones.

Veamos ahora en la Figura 3.17 estas mismas métricas de desempeño, exactamente bajo las mismas condiciones, pero esta vez aplicando el algoritmo de localización de ODAS usando los mismos tres micrófonos que fueron usados por el algoritmo propuesto. Es decir, usando las mismas grabaciones en escenarios prácticos, con el mismo arreglo de 3 micrófonos en los vértices de un triángulo, vamos a comparar las tasas de detección de hablantes.

En el 20 % de las grabaciones donde había un hablante activo, este fue detectado correctamente. En el 80 % restante, ningún hablante fue detectado dentro de un intervalo de $\pm 15^{o}$ alrededor del acimut en que realmente se ubicaba la persona hablando. Este resultado coincide con el obtenido dentro de la cámara anecoica.

Por otro lado, en el 14 % de las grabaciones donde había dos personas hablando, solamente un hablante fue detectado, y en el 86 % restante no se detectó ningún hablante. En ningún caso se detectaron simultáneamente los dos hablantes que estaban activos. Aquí ya vemos una diferencia sustancial respecto a las mismas métricas en cámara anecoica. El desempeño del algoritmo se ha deteriorado significativamente ante la presencia de dos personas hablando simultáneamente.

Por otro lado, en el 18 % de las grabaciones donde había tres personas hablando, solamente dos fueron detectadas correctamente. En el 43 % de las grabaciones se detectó un hablante solamente, y en el 39 % restante no se detectó ningún hablante dentro del margen de tolerancia permitido. En ningún caso el algoritmo fue capaz de localizar a las tres personas simultáneamente. También aquí la diferencia es significativa en comparación con los resultados que se obtuvieron sin reverberación ni ruido.

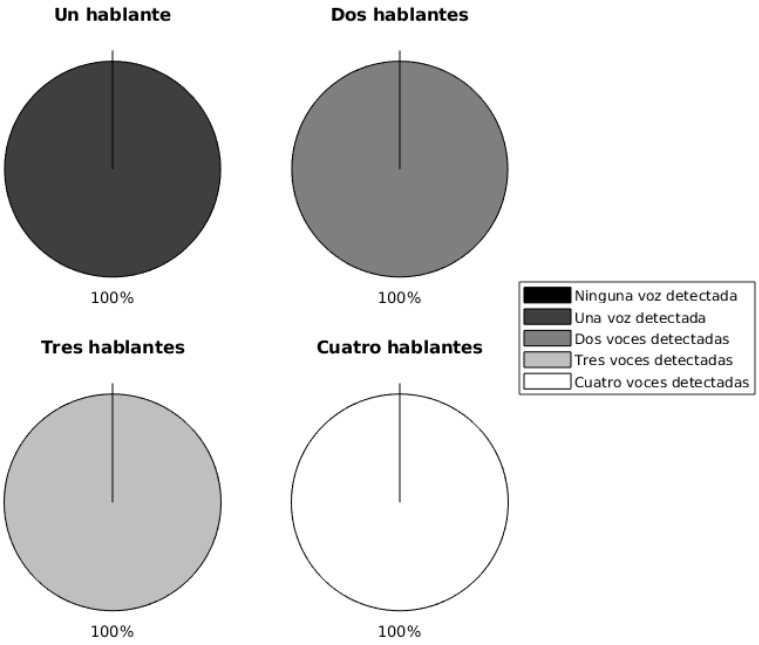


Figura 3.18: Tasa de detección de hablantes en escenarios prácticos, usando el algoritmo de localización de ODAS con 8 micrófonos.

En el caso más complejo, con cuatro personas hablando simultáneamente, en menos del 1 % de las grabaciones se detectaron correctamente tres de los cuatro hablantes. En el 19 % de las grabaciones se detectaron dos hablantes y en el 53 % se detectó una sola persona. En el 27 % de casos restantes, el algoritmo no fue capaz de localizar correctamente a ninguno de los hablantes. En ningún caso el algoritmo fue capaz de localizar a las cuatro personas simultáneamente. Nuevamente vemos un deterioro importante en el desempeño del algoritmo operando en escenarios prácticos.

Si comparamos las Figuras 3.14 y 3.17, y las Figuras 3.13 y 3.16 podemos verificar que, aunque ambos algoritmos han deteriorado su desempeño en escenarios prácticos, el algoritmo de localización propuesto en el presente trabajo continúa superando con creces el desempeño que proporciona ODAS en exactamente las mismas condiciones. En particular, resalta el hecho de que en la mayoría de los casos nuestra propuesta detectó correctamente la presencia de todos los hablantes que había presente. A diferencia de ODAS, que ante más de un hablante nunca consiguió detectarlos a todos correctamente. Entonces, podemos confirmar que en escenarios prácticos, con niveles de ruido y de reverberación típicos, nuestro algoritmo no solamente es menos costoso en recursos de hardware, sino que además proporciona un desempeño muy superior al de ODAS bajo las mismas condiciones. De la única forma que el algoritmo de localización de ODAS puede ser comparable o superior en desempeño a nuestra propuesta, es usando una cantidad bastante mayor de micrófonos. Por ejemplo, considerando la variante recomendada por sus autores, que es usando 8 micrófonos, llegamos a los resultados que se muestra en la Figura 3.18 al operar en los mismos escenarios prácticos.

Consideremos nuevamente las métricas de desempeño que evalúan la precisión de los algoritmos de localización. En la Tabla 3.5 se recogen 5 métricas de desempeño que han sido usadas para evaluar el algoritmo de localización que se emplea en ODAS usando 8 micrófonos.

Tabla 3.6: Métricas de desempeño del algoritmo de localización propuesto operando en escenarios prácticos con 3 micrófonos en arreglo triangular (micrófonos 1 al 3 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (E5):	Tasa de detección (E10):	Tasa de detección (E15):
1	3.6°	3.9°	77.9 %	21.8 %	0.3 %
2	5.0°	5.4°	71.0 %	24.1 %	4.9 %
3	6.1°	6.8°	46.6 %	15.8 %	37.6 %
4	5.6°	6.2°	47.4 %	24.5 %	28.1 %

Son las mismas métricas que se recogen en la Tabla 3.6 para comparar contra el algoritmo de localización propuesto.

Los errores absoluto y cuadrático medio usando el sistema propuesto son inferiores a 5° ante la presencia de una sola persona hablando de forma simultánea. Ante múltiples hablantes, se nota un deterioro en la precisión, tanto en comparación con el mismo sistema en cámara anecoica, como en comparación con ODAS usando 8 micrófonos en los mismos escenarios prácticos. También se nota una mayor diferencia en las tasas de detecciones con un error absoluto inferior a 5° (la métrica E5), de forma tal que ya se hace más evidente la superioridad que le aportan esos 5 micrófonos adicionales en el algoritmo de localización de ODAS. Nuestra propuesta que usa 3 micrófonos en escenarios con reverberación y ruidos moderados presenta una mayor dispersión al estimar el acimut que la solución de ODAS usando 8 micrófonos, sobre todo ante tres o más hablantes.

Podemos entonces llegar a una conclusión parcial hasta este punto. En escenarios prácticos, con niveles de ruido y de reverberación moderados, se nota más claramente la superioridad de la solución de ODAS con 8 micrófonos frente a nuestra solución que únicamente usa 3 micrófonos. Esta diferencia es más evidente en presencia de más de 2 hablantes. Ante uno o dos hablantes, los desempeños son bastante similares. Este es un resultado satisfactorio, teniendo en cuenta el ahorro en términos de hardware y costo computacional que representa nuestra solución frente a la de ODAS.

Tabla 3.5: Métricas de desempeño del algoritmo de localización de ODAS operando en escenarios prácticos con 8 micrófonos (micrófonos 1 al 8 de la Figura 3.8b). N: Número de hablantes.

N	Error absoluto:	Error cuadrático medio:	Tasa de detección (E5):	Tasa de detección (E10):	Tasa de detección (E15):
1	3.6°	4.4°	75.0 %	20.1 %	4.9 %
2	3.3°	4.1°	81.7 %	15.6 %	2.7 %
3	3.0°	3.8°	85.4 %	11.9 %	2.7 %
4	3.9°	4.7°	73.4 %	21.9 %	4.7 %

3.3. Comparación de costo computacional y tiempo de procesamiento

Con el objetivo de profundizar más en nuestra comparación entre las distintas implementaciones consideradas en el presente trabajo, también analizamos a continuación el costo computacional que representa cada una. Las tres implementaciones analizadas son: nuestra propuesta basada en un arreglo triangular de 3 micrófonos, la implementación de ODAS usando el mismo arreglo triangular de 3 micrófonos, y la implementación de ODAS usando un total de 8 micrófonos.

Realizamos un conjunto de pruebas en varios escenarios, en la localización y rastreo de 1, 2, 3 y hasta 4 hablantes simultáneamente. Todas las mediciones de carga computacional se realizaron en una computadora[1] con procesador de cuatro núcleos Intel® Core™ i5-7200U a 2.50 GHz.

Durante un total de una hora y 52 minutos de grabaciones tomadas del corpus AIRA, se registró en intervalos de 100 ms el uso de la CPU que demandaba cada sistema de localización y rastreo. Se tienen en cuenta 3 métricas de la carga computacional asociada a cada una de las implementaciones, procesando exactamente las mismas grabaciones: la carga promedio, la carga RMS, y la carga máxima. La Figura 3.19 muestra el uso de la CPU por cada una de las implementaciones analizadas, realizando todo el procesamiento en un único núcleo.

Aquí hay que resaltar que se podrían utilizar otras métricas para medir el costo computacional, pero hacer un análisis detallado de la complejidad computacional de los algoritmos está fuera del alcance y objetivos del presente trabajo. Por ese motivo se optó por medir el costo computacional de una manera práctica, con el entendido de que es puramente como una métrica comparativa. De hecho, los propios autores de ODAS realizan una comparación similar a la nuestra en su más reciente artículo [9]. Si bien es cierto que esta métrica es dependiente de la CPU en cuestión, la que hemos usado para la comparación en este caso es de una computadora con prestaciones moderadas, la cual es representativa del tipo de computadora que se utilizaría en un robot de servicio.

Los resultados demuestran que nuestra implementación reduce considerablemente la carga computacional. El uso de la CPU se reduce a aproximadamente la mitad del uso que requiere ODAS, tanto en las configuraciones de 3 como de 8 micrófonos.

En cuanto a tiempos de procesamiento, ni siquiera durante los picos de carga máxima de ODAS con 8 micrófonos, se saturó la CPU a su máxima capacidad. Por tanto los tres algoritmos operaron sin dificultades en tiempo real. Teniendo en cuenta la descripción de nuestra propuesta en la sección 3.1.3, la etapa de localización arroja resultados instantáneamente. Es decir, se obtiene un valor de acimut para cada trama de audio procesada, siempre y cuando se detecte actividad vocal. Todo el procesamiento del localizador se realiza durante el período de tiempo que comprende una trama de audio: 1024 muestras, unos 21 ms a una tasa de muestreo de 48 kHz. El rastreo, sin embargo, sí introduce cierto retardo, proporcional al tamaño de registro que alimenta al algoritmo k-*means*. Es decir, el acimut que entrega finalmente el sistema propuesto tiene un retardo de al menos N_{hist} tramas de audio, como se vio en la en la sección 3.1.3. Por ejemplo, si se usa $N_{\text{hist}} = 5$, entonces cada acimut a la salida se obtendrá con retardo de al menos 5×1024 muestras, unos 105 ms a una tasa de muestreo de 48 kHz. Este retardo, bastante inferior a un segundo, no es significativo como para entorpecer la toma de decisiones rápidas por

[1] Más precisamente, se trata de una laptop marca Dell, modelo Inspiron 5567 (https://www.dell.com/si/business/p/inspiron-15-5567-laptop/pd).

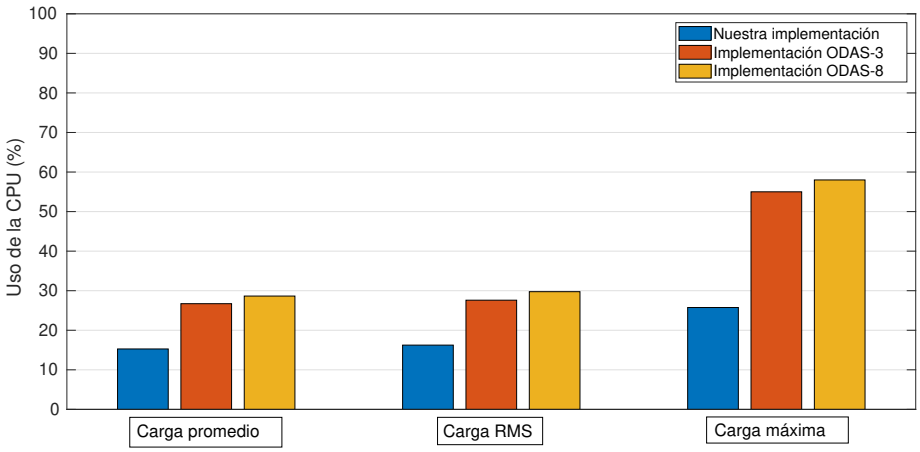


Figura 3.19: Uso de la CPU en un único núcleo Intel® Core™ i5-7200U a 2.50 GHz.

parte de un robot de servicio en un escenario real, así que no lo consideramos un impedimento. Por otro lado, el algoritmo de ODAS, hay que decirlo, no introduce ningún tipo de retardo, el sistema de localización y rastreo arroja un resultado para cada trama de audio donde se detecte actividad vocal.

3.4. Resumen

En el presente Capítulo hemos descrito la implementación de nuestra propuesta de sistema de localización y rastreo de múltiples fuentes de voz para un robot de servicio. Se trata de una variante que únicamente estima el acimut en donde se ubican los hablantes, dado que dispone únicamente de tres micrófonos dispuestos en un plano horizontal formando un triángulo. El algoritmo más costoso computacionalmente en la etapa de localización es la FFT, que se usa varias veces en cada iteración para medir el coeficiente de correlación de Pearson que existe entre cada par de canales de audio. En la etapa de rastreo, se optó por el uso de un filtro de Kalman, y la agrupación de direcciones de arribo de acuerdo a una variante del algoritmo de *k-means*. Esta combinación resultó ser de un reducido costo computacional respecto a otras variantes disponibles en el Estado del Arte, específicamente, al compararlo con la implementación que se encuentra en ODAS.

Realizamos varias pruebas para medir y comparar los desempeños que se alcanzan con nuestra implementación en múltiples escenarios. En el escenario ideal de casi absoluta ausencia de reverberación y ruido, encontramos que nuestra propuesta logra una tasa de detección muy similar a la implementación de ODAS usando un arreglo de 8 micrófonos. Incluso la incertidumbre en la estimación del acimut de los hablantes es prácticamente la misma para uno o dos hablantes. Para un número mayor de interlocutores, se nota la ventaja de usar un mayor número de micrófonos, de forma tal que nuestra propuesta va quedando rezagada en cuanto a precisión y a tasa de detección. En escenarios reales, con presencia de niveles de ruido y reverberación moderados, nuestra solución es un poco menos precisa y detecta generalmente menos hablantes que ODAS usando 8 micrófonos. Para una tasa máxima de falsas alarmas de 10 %,

ODAS siempre fue capaz de detectar a todos los hablantes presentes en el recinto, mientras que nuestra propuesta lo lograba entre el 60 % y el 80 % de las ocasiones para un número de hablantes de entre uno y cuatro.

Por otro lado, en todos los casos, nuestra propuesta supera a ODAS cuando se ponen a competir en igualdad de condiciones. A saber, cuando ambas implementaciones hacen uso de los mismos 3 micrófonos en arreglo triangular. Otro resultado importante que analizamos en este Capítulo fue el costo computacional que representan las 3 variantes analizadas. Entre nuestra propuesta, ODAS con 3 micrófonos, y ODAS con 8 micrófonos, resultó que la implementación nuestra es bastante más ligera en cuanto a uso de la CPU. Tanto en uso promedio como en uso máximo, demostramos que se ahorra por lo menos un 50 % de uso de la CPU. Este resultado es sumamente importante a la hora de tomar la decisión acerca de cuál sistema de localización y rastreo usar en la práctica para un robot de servicio.

Capítulo 4

Conclusiones y trabajo futuro

4.1. Conclusiones

En la naturaleza, con el desarrollo evolutivo de las distintas especies de animales y los seres humanos, podemos encontrar innumerables ejemplos de cómo la audición es un proceso complejo y su imitación para robots en la era moderna resulta un reto importante. Tareas aparentemente simples, como la de localizar la dirección de donde proviene un sonido sin más información que la percepción acústica de nuestro entorno, no resultan triviales de resolver con la tecnología de que disponemos actualmente. Aunque en el Estado del Arte de audición robótica que hemos estudiado en nuestra investigación se encuentran soluciones bastante eficaces, sus implementaciones en la práctica implican un costo computacional elevado y un número de sensores mucho mayor que el que usan los seres vivos en la naturaleza. Lo que los seres humanos y los vertebrados en general somos capaces de hacer con nuestros dos oídos, la tecnología más avanzada en el tema no lo puede resolver en escenarios no controlados si no es usando un arreglo aparatoso de micrófonos, y sobrecargando el módulo de audición robótica con algoritmos sumamente pesados.

En este trabajo de tesis hemos intentado contribuir modestamente al entendimiento de este fenómeno y abordar una solución relativamente menos compleja que la mayoría de las propuestas encontradas en la literatura científica actual. Tomando como referencia la experiencia de trabajos previos en nuestro grupo de investigación, hemos implementado un sistema de localización y rastreo de múltiples hablantes que, con únicamente tres micrófonos distribuidos en los vértices de un triángulo, permite localizar al menos hasta cuatro personas hablando alrededor del robot. A diferencia de trabajos previos de nuestro grupo de investigación, el triángulo en cuestión no necesariamente tiene que se equilátero, sino que puede tener una geometría más o menos arbitraria, siempre que sus dimensiones estén en correspondencia con las frecuencia de las señales a detectar. Específicamente para la voz, estas dimensiones son de aproximadamente las dimensiones de una cabeza humana.

Luego de un análisis comparativo de los métodos disponibles en el Estado del Arte, nuestra investigación permitió seleccionar para la implementación práctica la combinación que prometía el mejor compromiso entre desempeño y costo computacional. Siempre pensando en integrarlo en el módulo de audición de un robot de servicio, nuestro enfoque fue evaluar el desempeño de nuestra implementación, y compararlo con uno de los competidores que más promete en la comunidad científica.

Se realizaron múltiples evaluaciones del sistema implementado, así como de un sistema del

Estado del Arte: ODAS, en una amplia variedad de escenarios. Desde una cámara anecoica, como ejemplo de escenario controlado, hasta recintos no controlados dentro del campus central de la UNAM, con niveles de ruido y de reverberación notables. Para obtener resultados que puedan ser replicados y comparados objetivamente contra otros algoritmos competidores en el Estado del Arte, se decidió usar un *corpus* de audio que dispone de varias horas de grabaciones con arreglos de hasta 16 micrófonos, y ante la presencia de una, dos, tres y hasta cuatros personas hablando o altavoces reproduciendo la voz de personas. Para todos los escenarios se dispone de la localización real de cada hablante respecto al arreglo, de forma tal que pueda medirse el error de localización de cada sistema estudiado.

Para satisfacción nuestra, los resultados obtenidos indican que logramos una solución alternativa que en muchos casos se desempeña de forma similar a soluciones más complejas. En efecto, luego de comparar los resultados obtenidos en igualdad de condiciones, la precisión lograda es similar a la de ODAS ante uno o dos hablantes simultáneos, y aunque la precisión se deteriora con un número mayor de hablantes, sigue siendo bastante buena, en términos de error absoluto y error cuadrático medio del acimut. El resultado más importante de nuestra investigación fue que logramos la precisión mencionada manteniendo un costo computacional significativamente más bajo que los competidores analizados, y usando un número significativamente menor de micrófonos. Haciendo un balance entre precisión a la hora de localizar a los interlocutores, costo de implementación en cuanto a número de micrófonos y costo computacional, podemos concluir que con la presente investigación logramos un sistema de localización más apropiado para su integración en robots de servicio.

4.2. Trabajo futuro

Aunque consideramos que los objetivos del presente trabajo han sido satisfechos, al llegar a este punto de la investigación, hemos identificado varios cauces por donde se puede continuar como trabajo futuro.

- Proponemos integrar el sistema de localización y rastreo al módulo de audición robótica de un robot de servicio. Esto a su vez permitiría evaluar en concreto la utilidad que aporta nuestra investigación en comparación con trabajos similares desarrollados previamente en el grupo de investigación, como [8]. Además, al formar parte de un módulo integral de audición robótica, que incluye separación de fuentes de voz, reconocimiento del locutor y reconocimiento de las instrucciones de voz ante la presencia de varios interlocutores, sería factible evaluar cómo el desempeño del sistema de localización propuesto afecta los resultados de esas otras tareas de audición robótica. En este sentido, proponemos tener en cuenta las métricas descritas en la sección 2.4.2 del presente trabajo: SDR, SIR, SNR, SAR, y WER.

- Aunque consideramos que la evaluación que se realizó del sistema propuesto y un competidor del Estado del Arte fue bastante detallada y minuciosa gracias al uso del *corpus* AIRA, que ya de por sí contiene hasta el momento que se escribe esta tesis más de 10 horas de audio en una amplia variedad de condiciones y escenarios bien descritos en cuanto a los niveles de ruido y reverberación, proponemos continuar la evaluación con otras bases de datos, para generalizar aún más nuestros resultados. En este sentido, proponemos tener en cuenta las descritas en la sub-sección 3.2.1 del presente trabajo: AV16.3 y LOCATA.

- Consideramos que la implementación del sistema de localización propuesto puede ser enriquecida al incluir uno o más micrófonos adicionales que permitan no solamente localizar

a los hablantes en términos de acimut, sino además en cuanto a elevación. La implementación actual únicamente puede estimar acimut porque sus tres sensores están distribuidos sobre un plano horizontal. Al considerar una configuración que incluya sensores en un espacio tridimensional, en principio se puede extender el algoritmo de estimación de direcciones de arribo para que estime además la altura relativa de los interlocutores con respecto al robot.

- El presente trabajo no se enfocó en analizar casos más específicos, de múltiples hablantes en movimiento, y en particular cuando los hablantes se cruzan entre sí, o realizan movimientos más complejos. Si bien se pudiera pensar que se trata de escenarios menos comunes en los cuales operaría un robot de servicio, no dejan de representar retos importantes que ya han sido resueltos en implementaciones de mayor costo computacional, como la versión de ODAS analizada en nuestro análisis comparativo.
- Un proyecto futuro más ambicioso puede concebir al módulo de audición robótica como un sistema interdependiente, y retroalimentar la salida de las etapas de separación, identificación del locutor, y reconocimiento de voz, entre otras, para reforzar la robustez de la localización de los hablantes. Teniendo en cuenta que la audición en la naturaleza no se concibe como una secuencia unidireccional de etapas, sino como un todo, este enfoque más ambicioso podría ser más eficaz. En particular, proponemos considerar el uso de inteligencia artificial: un sistema de aprendizaje automatizado que tome decisiones a partir de las salidas de las distintas tareas de audición robótica, inspirándonos nuevamente en cómo funcionan los sistemas auditivos observados en la naturaleza.

Glosario

CPU *Central Processing Unit* (unidad central de procesamiento). 15, 30, 53, 54, 55

E10 tasa de detecciones con error absoluto entre 5º y 10º. 19, 45, 47, 48, 52

E15 tasa de detecciones con error absoluto entre 10º y 15º. 19, 45, 47, 48, 52

E30 tasa de detecciones con error absoluto entre 15º y 30º. 19

E5 tasa de detecciones con error absoluto inferior a 5º. 19, 45, 46, 47, 48, 52

FFT *Fast Fourier Transform* (transformada rápida de Fourier). 16, 30, 36, 54

FFTW *the Fastest Fourier Transform in the West"* (la transformada de Fourier más rápida en el Oeste). 16, 36

GEVD *Generalized Eigenvalue Decomposition* (descomposición generalizada por valor propio). 14

GNU-GPL *GNU General Public License* (licencia pública general de GNU). 20

GSC *Generalized Sidelobe Canceller* (anulador de lóbulos laterales generalizado). 11, 12, 26

GSVD *Generalized Singular Value Decomposition* (descomposición generalizada por valor singular). 15

I tasa de inserción de una fuente de voz. 19, 43, 48

IEEE *Institute of Electrical and Electronics Engineers* (Instituto de Ingenieros Eléctricos y Electrónicos). 21, 24, 39

MUSIC *Multiple Signal Classification* (clasificación de múltiples señales). 14, 15, 19, 25, 29

MVDR *Minimum Variance Distortionless Response* (respuesta sin distorsión de mínima varianza). 13

ODAS *Open embeddeD Audition System* (sistema abierto de audición embebida). 43, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 58, 59

RMS *Root Mean Square* (raíz de la media cuadrática). 53

RW-PHAT *Reliability Weighted Phase Transform* (transformada de fase compensada de acuerdo a la fiabilidad). 18

SAR *Signal-to-Artifacts Ratio* (relación señal a ruido artificial). 58

SDR *Signal-to-Distortion Ratio* (relación señal a distorsión). 58

SIR *Signal-to-Interference Ratio* (relación señal a interferencia). 58

SNR *Signal-to-Noise Ratio* (relación señal a ruido). 58

SRP-PHAT *Steered Response Power Phase Transform* (transformada de fase de la respuesta de potencia en cierta dirección). 18, 25, 29, 30

WER *Word Error Rate* (tasa de palabras erróneas). 20, 58

Bibliografía

[1] C. Rascon, I. V. Meza, A. Millan-Gonzalez, I. Velez, G. Fuentes, D. Mendoza, y O. Ruiz-Espitia, "Acoustic interactions for robot audition: A corpus of real auditory scenes," *The Journal of the Acoustical Society of America*, vol. 144, pp. EL399–EL403, Noviembre 2018. 39, 40

[2] C. Rascon, I. V. Meza, A. Millan-Gonzalez, I. Velez, G. Fuentes, D. Mendoza, y O. Ruiz-Espitia, "Documentation of the AIRA Corpus," p. 12, Universidad Nacional Autónoma de México, Octubre 2018. 40, 41, 42

[3] A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000. 1, 2

[4] N. Wood y N. Cowan, "The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel?," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 21, no. 1, p. 255, 1995. 1

[5] M. Pedersen, "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Communication*, 2007. 1, 2, 10

[6] W. A. Yost y G. Gourevitch, *Directional hearing*. Springer, 1987. 2, 6, 7

[7] S. Haykin, "Array signal processing," *Englewood Cliffs, NJ, Prentice-Hall, Inc.*, 1985. 2

[8] C. Rascon, G. Fuentes, y I. Meza, "Lightweight multi-DOA tracking of mobile speech sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, p. 11, Mayo 2015. 2, 3, 4, 18, 19, 22, 23, 58

[9] F. Grondin y F. Michaud, "Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations," *Robotics and Autonomous Systems*, vol. 113, pp. 63–80, Marzo 2019. 3, 19, 24, 26, 29, 53

[10] G. G. Rosenthal y M. J. Ryan, "Visual and acoustic communication in non-human animals: a comparison," *Journal of biosciences*, vol. 25, no. 3, pp. 285–290, 2000. 5

[11] H. Haas, "Über den Einfluß eines Einfachechos auf die Hörsamkeit von Sprache," *Acta Acustica united with Acustica*, vol. 1, pp. 49–58, Enero 1951. 6

[12] J. L. Cranford, M. Boose, y C. A. Moore, "Effects of aging on the precedence effect in sound localization," *Journal of Speech, Language, and Hearing Research*, vol. 33, no. 4, pp. 654–659, 1990. 6

[13] M. L. Dent y R. J. Dooling, "Investigations of the precedence effect in budgerigars: The perceived location of auditory images," *The Journal of the Acoustical Society of America*, vol. 113, no. 4, pp. 2159–2169, 2003. 6, 7

[14] J. Huang, N. Ohnishi, y N. Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, pp. 842–846, Agosto 1997. 7

[15] C. Hummersone, R. Mason, y T. Brookes, "A Comparison of Computational Precedence Models for Source Separation in Reverberant Environments," *Journal of the Audio Engineering Society*, vol. 61, pp. 508–520, Agosto 2013. 7

[16] S. Theodoridis, R. Chellappa, M. Viberg, y A. Zoubir, *Academic Press Library in Signal Processing: Array and Statistical Signal Processing.* Academic Press Library in Signal Processing, Elsevier Science, 2013. 8, 12, 13, 14

[17] C. Rascon y I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, Octubre 2017. 11, 12, 15, 16

[18] J. H. DiBiase, H. F. Silverman, y M. S. Brandstein, "Robust Localization in Reverberant Rooms," en *Microphone Arrays: Signal Processing Techniques and Applications* (M. Brandstein and D. Ward, eds.), Digital Signal Processing, pp. 157–180, Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. 10, 14, 18, 19

[19] B. M. Radich y K. M. Buckley, "Single-snapshot DOA estimation and source number detection," *IEEE Signal processing letters*, vol. 4, no. 4, pp. 109–111, 1997. 11, 18

[20] S. Fortunati, R. Grasso, F. Gini, M. S. Greco, y K. LePage, "Single-snapshot DOA estimation by using compressed sensing," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, p. 120, Julio 2014. 11

[21] R. E. Bethel y K. L. Bell, "Maximum likelihood approach to joint array detection/estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 40, pp. 1060–1072, Julio 2004. 11

[22] A. Bazzi, *Techniques d'estimation de paramètres pour la localisation à l'intérieur via WiFi.* PhD. thesis, Paris, ENST, Octubre 2017. 11

[23] E. Aboutanios, A. Hassanien, A. El-Keyi, Y. Nasser, y S. A. Vorobyov, "Advances in DOA estimation and source localization," *International Journal of Antennas and Propagation*, vol. 2017, 2017. 11

[24] L. Griffiths y C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982. 11

[25] H. Lim, I.-C. Yoo, Y. Cho, y D. Yook, "Speaker localization in noisy environments using steered response voice power," *IEEE Transactions on Consumer Electronics*, vol. 61, pp. 112–118, Febrero 2015. 11

[26] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, y N. Otsu, "Integrated natural spoken dialogue system of "Jijo-2"mobile robot for office services," en *Proceedings of the National Conference on Artificial Intelligence and the Innovative Applications of Artificial Intelligence*, pp. 621–627, American Association for Artificial Intelligence, 1999. 12, 20

[27] J.-M. Valin, F. Michaud, B. Hadjou, y J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach," en *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, vol. 1, pp. 1033–1038, Abril 2004. 12

[28] L. Mattos y E. Grant, "Passive sonar applications: target tracking and navigation of an autonomous robot," en *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, vol. 5, pp. 4265–4270 Vol.5, Abril 2004. 12

[29] Y. Tamai, Y. Sasaki, S. Kagami, y H. Mizoguchi, "Three ring microphone array for 3d sound localization and separation for mobile robot audition," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4172–4177, Agosto 2005. 12

[30] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, y H. G. Okuno, "Multiple moving speaker tracking by microphone array on mobile robot," en *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 249–252, 2005. 12

[31] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, y K. Oro, "Spherical microphone array for spatial sound localization for a mobile robot," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 713–718, Octubre 2012. 12

[32] Y. Sasaki, S. Kagami, y H. Mizoguchi, "Multiple Sound Source Mapping for a Mobile Robot by Self-motion Triangulation," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 380–385, Octubre 2006. 12

[33] Y. Sasaki, S. Masunaga, S. Thompson, S. Kagami, y H. Mizoguchi, "Sound localization and separation for mobile robot tele-operation by tri-concentric microphone array," *Journal of Robotics and Mechatronics*, vol. 19, no. 3, p. 281, 2007. 12

[34] S. Argentieri, P. Danès, y P. Souères, "Prototyping filter-sum beamformers for sound source localization in mobile robotics," en *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3551–3556, IEEE, 2005. 13

[35] S. Argentieri, P. Danès, y P. Soueres, "Modal Analysis Based Beamforming for Nearfield or Farfield Speaker Localization in Robotics," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 866–871, Octubre 2006. 13

[36] S. Argentieri y P. Danès, "Convex optimization and modal analysis for beamforming in robotics: Theoretical and implementation issues," en *Proceedings of European Signal Processing Conference (EUSIPCO)*, pp. 773–777, Septiembre 2007. 13

[37] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Marzo 1986. 14

[38] T. Otsuka, K. Nakadai, T. Ogata, y H. G. Okuno, "Bayesian Extension of MUSIC for Sound Source Localization and Tracking," en *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3109–3112, Agosto 2011. 14, 25

[39] M. A. Flores Gómez, *Estimación de la dirección del ángulo de arribo de una fuente de voz.* Tesis de Maestría, Universidad Nacional Autónoma de México, Ciudad de México, 2017. 14

[40] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, y H. Tsujino, "Intelligent sound source localization for dynamic environments," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 664–669, Octubre 2009. 14

[41] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, y H. Tsujino, "Design and implementation of robot audition system 'HARK'. Open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010. 15, 20

[42] E. Vincent, A. Sini, y F. Charpille, "Audio source localization by optimal control of a mobile robot," en *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 5630–5634, Abril 2015. 15

[43] S. Pourmehr, J. Bruce, J. Wawerla, y R. T. Vaughan, "A Sensor Fusion Framework for Finding an HRI Partner in Crowd," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–6, 2015. 15, 24

[44] A. V. Oppenheim y R. W. Schafer, *Discrete-time signal processing.* Pearson Education, 2014. 16

[45] C. Knapp y G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, Agosto 1976. 16, 17

[46] E. Al-Hussaini y S. Kassam, "Robust eckart filters for time delay estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1052–1063, Octubre 1984. 17

[47] F. Grondin y F. Michaud, "Noise Mask for TDOA Sound Source Localization of Speech on Mobile Robots in Noisy Environments," en *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–6, 2016. 18

[48] C. Rascon, I. Meza, G. Fuentes, L. Salinas, y L. Pineda, "Integration of the Multi-DOA Estimation Functionality to Human-Robot Interaction," *International Journal of Advanced Robotic Systems*, vol. 12, no. 8, 2015. 18

[49] J. M. Valin, F. Michaud, y J. Rouat, "Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering," en *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. IV–841–IV–844, Mayo 2006. 18, 25

[50] J.-M. Valin, F. Michaud, y J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007. 18, 24, 25, 26

[51] A. Johansson y S. Nordholm, "Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a real-time implementation," en *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 4, pp. iv/933–iv/936 Vol. 4, Marzo 2005. 18

[52] O. Nadiri y B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 1494–1505, Octubre 2014. 18

[53] B. Rafaely y D. Kolossa, "Speaker localization in reverberant rooms based on direct path dominance test statistics," en *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6120–6124, Marzo 2017. 18

[54] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, y H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," en *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2814–2818, Abril 2015. 18

[55] S. Delikaris-Manias, D. Pavlidi, V. Pulkki, y A. Mouchtaris, "3D localization of multiple audio sources utilizing 2D DOA histograms," en *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1473–1477, Agosto 2016. 19

[56] D. Pavlidi, S. Delikaris-Manias, V. Pulkki, y A. Mouchtaris, "3D localization of multiple sound sources with intensity vector estimates in single source zones," en *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1556–1560, Agosto 2015. 19

[57] X. Li, L. Girin, F. Badeig, y R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," en *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2819–2826, Octubre 2016. 19

[58] N. Ma, T. May, y G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, pp. 2444–2453, Diciembre 2017. 19

[59] T. Suzuki, H. Otsuka, W. Akahori, Y. Bando, y H. G. Okuno, "Influence of Different Impulse Response Measurement Signals on MUSIC-Based Sound Source Localization," *Journal of robotics and mechatronics*, vol. 29, no. 1, pp. 72–82, 2017. 19

[60] T.Takahshi, K. Nakadai, C. T. Ishi y H. G. Okuno, "Investigation of sound source localization and separation under a real environment," en *Proceedings of 29th Annual Meeting of Robotics Society of Japan*, 2011 (en Japonés). 19

[61] E. Vincent, R. Gribonval, y C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006. 20

[62] C. Févotte, R. Gribonval, y E. Vincent, "BSS_EVAL toolbox user guide–Revision 2.0," *Technical report 1706, IRISA*, 2005. 20

[63] Ye-Yi Wang, A. Acero, y C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy?," en *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pp. 577–582, Noviembre 2003. 20

[64] F. Asano, M. Goto, K. Itou, y H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition.," en *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1013–1016, 2001. 20

[65] H. Okuno y K. Nakadai, "Computational Auditory Scene Analysis and its Application to Robot Audition," en *Proceedings of Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 124–127, Mayo 2008. 20

[66] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, vol. 82, pp. 35–45, Marzo 1960. 21, 22

[67] "The LOCATA Challenge Data Corpus for Acoustic Source Localization and Tracking," pp. 410–414, Julio 2018. ISSN: 2151-870X. 21, 24, 25, 38

[68] X. Anguera, C. Wooters, y J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2011–2022, Septiembre 2007. 21

[69] B. Laufer-Goldshtein, R. Talmon, y S. Gannot, "Semi-Supervised Source Localization on Multiple Manifolds With Distributed Microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1477–1491, Julio 2017. 22

[70] B. Laufer-Goldshtein, R. Talmon, y S. Gannot, "Speaker Tracking on Multiple-Manifolds with Distributed Microphones," en *Latent Variable Analysis and Signal Separation* (P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, eds.), Lecture Notes in Computer Science, pp. 59–67, Springer International Publishing, 2017. 22

[71] B. Laufer-Goldshtein, R. Talmon, y S. Gannot, "A Hybrid Approach for Speaker Tracking Based on TDOA and Data-Driven Models," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, pp. 725–735, Abril 2018. 22

[72] D. Salvati, C. Drioli, y G. L. Foresti, "Localization and Tracking of an Acoustic Source using a Diagonal Unloading Beamforming and a Kalman Filter," en *Proceedings of LOCATA Challenge Workshop - a satellite event of IWAENC 2018*, Diciembre 2018. 24

[73] L. D. Mosgaard, D. Pelegrin-Garcia, T. B. Elmedyb, M. J. Pihl, y P. Mowlaee, "Circular Statistics-based low complexity DOA estimation for hearing aid application," en *Proceedings of LOCATA Challenge Workshop - a satellite event of IWAENC 2018*, Diciembre 2018. 24

[74] X. Li, Y. Ban, L. Girin, X. Alameda-Pineda, y R. Horaud, "A cascaded multiple-speaker localization and tracking system," en *Proceedings of LOCATA Challenge Workshop - a satellite event of IWAENC 2018*, Diciembre 2018. 24

[75] S. Ba, X. Alameda-Pineda, A. Xompero, y R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, Diciembre 2016. 24

[76] P. M. Djurić y M. F. Bugallo, "Particle Filtering," en *Adaptive Signal Processing*, pp. 271–331, John Wiley & Sons, 2010. 25, 26

[77] M. Fréchette, D. Létourneau, J. M. Valin, y F. Michaud, "Integration of sound source localization and separation to improve Dialogue Management on a robot," en *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2358–2363, Octubre 2012. 25

[78] A. Reveleau, F. Ferland, M. L. a. D. Letourneau, y F. Michaud, "Visual Representation of Interaction Force and Sound Source in a Teleoperation User Interface for a Mobile Robot," *Journal of Human-Robot Interaction*, vol. 4, no. 2, pp. 1–23, 2015. 25

[79] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, y F. Michaud, "The ManyEars open framework," *Autonomous Robots*, vol. 34, no. 3, pp. 217–232, 2013. 25, 29

[80] X. Qian, A. Cavallaro, A. Brutti, y M. Omologo, "LOCATA challenge: speaker localization with a planar array," en *Proceedings of LOCATA Challenge Workshop - a satellite event of IWAENC 2018*, Septiembre 2018. 25

[81] Y. Liu, W. Wang, y V. Kilic, "Intensity Particle Flow SMC-PHD Filter For Audio Speaker Tracking," en *Proceedings of LOCATA Challenge Workshop - a satellite event of IWAENC 2018*, Diciembre 2018. 25

[82] F. e. a. Michaud, "Spartacus attending the 2005 AAAI conference," *Autonomous Robots*, vol. 22, pp. 369–383, Mayo 2007. 29

[83] S. Yamamoto, K. Nakadai, J. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, y H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," en *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 4040–4045, Agosto 2005. 29

[84] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. Valin, K. Komatani, T. Ogata, y H. G. Okuno, "Real-Time Robot Audition System That Recognizes Simultaneous Speech in The Real World," en *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5333–5338, Octubre 2006. 29

[85] A. V. Oppenheim, J. R. Buck, and R. W. Schafer, *Discrete-time signal processing*. Upper Saddle River, NJ: Prentice Hall, 2001. 36

[86] G. Lathoud, J.-M. Odobez, y D. Gatica-Perez, "AV16.3: An Audio-Visual Corpus for Speaker Localization and Tracking," en *Machine Learning for Multimodal Interaction* (S. Bengio and H. Bourlard, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 182–195, Springer, 2005. 38

[87] R. R. Boullosa y A. Pérez López, "Some acoustical properties of the anechoic chamber at the Centro de Instrumentos, Universidad Nacional Autónoma de México," *Applied Acoustics*, vol. 56, pp. 199–207, Marzo 1999. 39

[88] L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, y L. Villaseñor, "The Corpus DIMEx100: transcription and evaluation," *Language Resources and Evaluation*, vol. 44, Diciembre 2010. 39

[89] L. A. Pineda, A. Rodríguez, G. Fuentes, C. Rascon, y I. V. Meza, "Concept and Functional Structure of a Service Robot," *International Journal of Advanced Robotic Systems*, vol. 12, p. 6, Febrero 2015. Publisher: SAGE Publications. 41