

Percepción multimodal para navegación terrestre mediante reconocimiento de sonidos y objetos

Multimodal perception for ground vehicle navigation using sound and object recognition

L. Carmona-Vieyra ^{a,*}, G.J. Aguilar-Belmonte ^a, G. Ramírez-Villa ^a, M. Torres-Rivera ^a

^aUniversidad Aeronáutica en Querétaro, Colón 76270 Querétaro, México.

Resumen

Este proyecto presenta el desarrollo de un vehículo terrestre autónomo con capacidades de percepción multimodal, cuyo principal objetivo es la localización e identificación de fuentes sonoras y objetos visuales. Para ello, el sistema integra un arreglo de micrófonos que ayudan a estimar la dirección de llegada del sonido, así como una cámara OAK-D Lite con capacidad de percepción de profundidad para la detección de objetos usando redes neuronales. A través de la fusión de información auditiva y visual, el vehículo puede desplazarse hacia la fuente de un sonido detectado o hacia un objeto identificado en su entorno. La plataforma utiliza ruedas omnidireccionales que permiten maniobras ágiles en cualquier dirección, y una Raspberry Pi encargada del procesamiento de señales, ejecución de inferencias y control del movimiento. Este proyecto sienta las bases para sistemas robóticos inteligentes capaces de responder a estímulos complejos del entorno mediante integración audiovisual.

Palabras Clave: Percepción multimodal, Redes neuronales, Integración sensorial, Vehículos autónomos.

Abstract

This project presents the development of an autonomous ground vehicle with multimodal perception capabilities, whose main objective is the localization and identification of sound sources and visual objects. To achieve this, the system integrates a microphone array to estimate the direction of arrival of sounds, as well as an OAK-D Lite camera with depth perception capabilities for object detection using neural networks. Through the fusion of auditory and visual information, the vehicle can move toward a detected sound source or an identified object in its environment. The platform uses omnidirectional wheels that enable agile manoeuvres in any direction, and a Raspberry Pi is responsible for signal processing, inference execution, and motion control. This project lays the foundation for intelligent robotic systems capable of responding to complex environmental stimuli through audiovisual integration.

Keywords: Multimodal perception, Neural networks, Sensor integration, Autonomous vehicles.

1. Introducción

El desarrollo de vehículos autónomos que sean capaces de interactuar de manera inteligente con el entorno que los rodea ha cobrado gran relevancia durante los últimos años, esto debido al amplio campo de aplicación que estos vehículos tienen y a los avances en diferentes áreas como la inteligencia artificial y la creación de sistemas cada vez más potentes para el desarrollo de estos vehículos. Este tipo de sistemas tienen aplicaciones en contextos como la vigilancia, la automatización de almacenes, los operativos de búsqueda y rescate, e incluso en escenarios de

robótica interactiva.

Como parte de la navegación autónoma de los vehículos, actualmente se utilizan diferentes sensores para el reconocimiento del entorno, lo que permite a los vehículos desplazarse de manera más segura y eficiente. Entre las tecnologías más utilizadas se encuentra la visión artificial, la cual permite detectar objetivos y patrones que guían la trayectoria del vehículo. Un ejemplo representativo es el sistema de detección de peatones desarrollado para vehículos autónomos, el cual emplea técnicas de procesamiento de imágenes para identificar personas

*Autor para correspondencia: lucycarmonavieyra@outlook.com

Correo electrónico: lucycarmonavieyra@outlook.com (Lucy Carmona-Vieyra) jairab17@hotmail.com (Gerardo Jair Aguilar-Belmonte), goretti.ramirez@unaq.mx (Goretti Ramírez-Villa), moises.torres@unaq.mx (Moisés Torres-Rivera)

Historial del manuscrito: recibido el 16/07/2025, última versión-revisada recibida el 20/10/2025, aceptado el 14/10/2025, publicado el 25/03/2026. **DOI:** <https://doi.org/10.29057/icbi.v14iEspecial.15536>



en el entorno y adaptar la navegación en consecuencia (Galarza Sánchez, 2019). De igual forma, se ha implementado un sistema de visión artificial de apoyo al comando de vehículos no tripulados, orientado a la detección de patrones específicos que sirven como referencia para la navegación (González Sierra y Mafla Meneses, 2018). Ambos trabajos destacan el papel de la visión artificial como herramienta clave para lograr una navegación autónoma más precisa y segura.

La localización precisa de fuentes sonoras también ha sido utilizada en diversos sistemas autónomos que interactúan con su entorno, especialmente en aplicaciones de robótica y redes de sensores acústicos. Técnicas como la correlación cruzada generalizada con transformada de fase (GCC-PHAT) se han estudiado y aplicado como métodos eficientes para estimar la dirección de llegada del sonido debido a su robustez y precisión, sin sacrificar la velocidad de inferencia respecto a otros métodos más complejos (Van Den Broeck *et al.*, 2012). Esta capacidad permite que sistemas autónomos puedan realizar tareas avanzadas de reconocimiento y seguimiento de fuentes sonoras en tiempo real, facilitando la identificación y clasificación de sonidos en entornos complejos.

Una de las últimas tendencias es la implementación de percepción multimodal, que permite combinar múltiples fuentes de información sensorial para la toma de decisiones. Esta capacidad aumenta la robustez de la navegación autónoma y la interacción con el entorno, lo que permite que los sistemas robóticos no dependan únicamente de un solo sensor.

La percepción multimodal ha sido empleada exitosamente en aplicaciones como robots guías en museos, hoteles y aeropuertos, así como en contextos de cuidado asistencial, demostrando su efectividad para enriquecer la interacción humano-robot mediante la fusión de visión y sonido, identificando personas, fuentes de sonido y frases con el uso de inteligencia artificial (Duncan *et al.*, 2024). Lo cual genera un precedente importante para el desarrollo de este tipo de tecnologías.

En el desarrollo de sistemas autónomos, es importante analizar cómo se combinan los distintos sensores y cómo es que hacen que el sistema pueda tomar decisiones con base en algoritmos de navegación autónoma. Usar varios tipos de sensores a la vez, como cámaras y micrófonos, puede ofrecer ventajas claras, pero también plantea desafíos. Por eso, evaluar el desempeño real de sistemas con percepción multimodal no solo nos ayuda a validar su correcto funcionamiento, sino también en qué situaciones pueden fallar realizando un análisis de fallas de nuestro sistema. Al observar cómo se complementan las distintas fuentes de información, es posible ajustar mejor los componentes del sistema según la tarea. Esto también abre la puerta a diseños más balanceados, que aprovechen bien los recursos disponibles sin sobrecargar al sistema.

El objetivo de este proyecto es diseñar un sistema de navegación terrestre capaz de navegar de forma autónoma en entornos controlados, utilizando percepción multimodal para identificar y desplazarse hacia fuentes sonoras o seguir objetos visuales en movimiento. Este robot está equipado con ruedas

omnidireccionales, e integra un sistema de percepción multimodal mediante el uso de una cámara OAK-D Lite S2 con capacidades de detección de objetos y un arreglo de micrófonos para la estimación de la dirección de llegada de sonidos. A través de la implementación de redes neuronales para el reconocimiento visual y algoritmos como el GCC-PHAT para el análisis de señales acústicas, se busca permitir el desplazamiento del vehículo hacia fuentes sonoras o seguir objetos específicos de su entorno.

Para validar el funcionamiento del sistema en cuanto a percepción visual y percepción auditiva, se espera llegar a una precisión en las detecciones visuales de al menos 80 %, y en cuanto a la estimación de dirección del sonido se espera alcanzar un error promedio menor a 10.

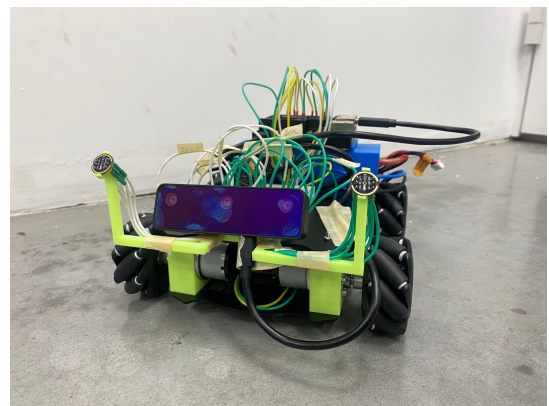


Figura 1: Prototipo de vehículo con micrófonos y cámara

2. Marco Teórico

Para continuar con el desarrollo de este proyecto requerimos de realizar una ardua investigación previa, dentro de la cual nos enfocamos en los siguientes puntos centrales: percepción auditiva (localización de fuentes sonoras y DOA), percepción visual (detección de objetos mediante redes neuronales), robótica móvil y navegación. A continuación, condensamos la información más importante encontrada en nuestro proceso de investigación, en la cual, debido a la naturaleza del sistema de percepción multimodal, nos enfocamos también para combinar estos conocimientos mediante la fusión sensorial para lograr un sistema con integración de percepción multimodal de manera adecuada y satisfactoria.

2.1. Percepción auditiva: localización de fuentes sonoras

La percepción auditiva en robótica permite a los sistemas identificar y localizar fuentes sonoras en su entorno, una capacidad esencial para tareas como búsqueda de víctimas o interacción humano-robot. La estimación de la dirección de llegada (DOA, por sus siglas en inglés) se basa en la diferencia temporal y de amplitud que presenta una señal acústica al llegar a un arreglo de micrófonos.

Métodos clásicos como GCC-PHAT (Generalized Cross-Correlation with Phase Transform) han sido ampliamente utilizados debido a su robustez en entornos con ruido y reverberación (Brandstein y Ward, 2001; Benesty *et al.*, 2008). La incorporación de microcontroladores como el ESP32, junto con micrófonos digitales como el INMP441, facilita la adquisición y procesamiento en tiempo real de señales acústicas, habilitando la localización activa de fuentes sonoras y mejorando la respuesta autónoma de los vehículos.

Sin embargo, en trabajos recientes sobre la identificación de fuentes sonoras utilizando únicamente arreglos de micrófonos y en ambientes controlados como una habitación con dimensiones conocidas, se han reportado limitaciones en la precisión de la localización. En particular, suelen presentarse errores tanto en el ángulo como en la estimación de la distancia, lo que puede generar desviaciones de hasta un metro respecto a la posición real de la fuente. Por esta razón, se ha recomendado complementar este enfoque con métodos adicionales de identificación, con el fin de reducir el error en la estimación de la posición (Pekár, 2023).

Uno de los puntos importantes para la transmisión digital de audio es el uso del protocolo I2S (Inter-IC Sound), el cual está incorporado en microcontroladores como la ESP32 y permite transferir datos de audio de alta fidelidad de forma síncrona y eficiente entre dispositivos (Espressif Systems, 2023). Este protocolo resulta una pieza fundamental en aplicaciones donde es necesaria una alta frecuencia de muestre (como lo es en el sonido) así como aplicaciones de bajo retardo y alta calidad en las señales transmitidas, necesarias para aplicar algoritmos de estimación de la dirección de llegada.

2.2. Percepción visual: detección de objetos con redes neuronales

La percepción visual basada en redes neuronales ha tenido un gran auge dentro del campo de la robótica móvil, permitiendo no solo el uso de visión por computadora como los filtros en los diferentes espacios de color (como RGB, HSV, etc.), sino que permite el reconocimiento preciso y en tiempo real de objetos, personas, señalamientos, entre otras cosas.

Actualmente existen algoritmos como YOLO (You Only Look Once) que han demostrado gran eficacia debido a su arquitectura que ha mejorado con las versiones más actuales, brindando mayor capacidad de reconocimiento sin sacrificar el tiempo de inferencia de cada frame que se procesa. Versiones más recientes, como YOLOv10, optimizan aún más el equilibrio entre precisión y velocidad, lo cual facilita y propicia su implementación en dispositivos embebidos que no tengan alta capacidad de procesamiento (Jocher *et al.*, 2023).

El uso de diferentes tipos de cámaras depende de la aplicación que se requiera utilizar, en este caso, cámaras especializadas como la OAK-D S2, proporciona una gran cantidad de información del entorno, al ser una cámara dual permite la estimación de profundidad de los diferentes objetos, así mismo es una cámara especializada para ejecutar redes neuronales, en

un pipeline interno que disminuye el costo de procesamiento a nuestro microcontrolador o microprocesador (Luxonis, 2023).

2.3. Vehículo terrestre y movilidad del sistema

La movilidad es un componente crítico en vehículos autónomos, ya que determina su capacidad de desplazarse en entornos complejos, existen diferentes modelos estudiados para sistemas de vehículos terrestres, dentro de los cuales, se encuentra el vehículo omnidireccional, este sistema debido a sus ruedas ofrecen una ventaja significativa al permitir movimientos en cualquier dirección sin necesidad de cambiar la orientación del chasis ni de ejecutar movimientos traslacionales para lograr movimientos rotacionales, lo que resulta de suma utilidad en aplicaciones de seguimiento o búsqueda activa (Siegwart *et al.*, 2011).

En el desarrollo de robots móviles con ruedas omnidireccionales, diversas investigaciones señalan que aún existen áreas de mejora tanto en el diseño mecánico como en el control. Una de las principales limitaciones identificadas es la ausencia de sistemas de suspensión, lo cual incrementa los errores en pruebas debido a las irregularidades del terreno. Se ha indicado que la inclusión de una suspensión adecuada permitiría absorber estas perturbaciones, mejorar la estabilidad del robot y reducir errores en la estimación de posición, siempre que los motores sean capaces de compensar el incremento de peso asociado (Campos Siurana, 2018).

2.4. Fusión sensorial, percepción multimodal y control

Dentro de la rama del control para vehículos terrestres con disposición de ruedas omnidireccionales, se facilita el modelado de estos, debido a que reduce su número de variables a únicamente la velocidad y dirección de rotación de cada una de las ruedas, lo cual puede verse reducido al tomar asunciones como igualar cada una de las velocidades aplicadas a cada rueda. Lo cual se profundizará en la sección de desarrollo.

Para este sistema, al tener lecturas multimodales se debe de definir un método de jerarquización de cada uno de los modos de percepción, para validar que el apartado visual o el sonoro tienen una mayor prioridad en cuanto a diferentes aspectos, como lo puede ser la fidelidad de las mediciones de ángulos de desviación respecto a fuentes sonoras u objetos detectados mediante una red neuronal. Para lograr una correcta fusión sensorial, con el objetivo de construir una representación más robusta y precisa del entorno (Khaleghi *et al.*, 2013).

Para sistemas de esta naturaleza se debe de definir un tiempo de muestreo fijo que dependerá de nuestra capacidad de determinación de ángulos mediante la red neuronal y mediante el algoritmo de detección de ángulo de llegada del sonido, cumpliendo con el teorema de Nyquist para posteriormente ejecutar movimientos rotativos hasta lograr sintonizar nuestra referencia (visual o auditiva).

3. Metodología

3.1. Ensamblaje del vehículo terrestre

La estructura mecánica principal está conformada por dos placas paralelas fabricadas en material rígido y liviano, tales

como acrílico o aluminio, dispuestas de forma horizontal. Estas placas forman un compartimiento central que sirve como soporte estructural y alojamiento para los componentes internos. Entre estos elementos internos se encuentran los motores DC tipo GM37-520, seleccionados debido a su alta eficiencia, torque adecuado para aplicaciones móviles y compatibilidad con controladores de velocidad PWM. Dichos motores permiten el accionamiento de un sistema de ruedas omnidireccionales, brindando la capacidad de realizar desplazamientos en cualquier dirección y rotaciones precisas sin necesidad de reorientar todo el chasis.

El compartimiento interno también alberga los módulos de alimentación eléctrica, los cuales consisten en baterías recargables y circuitos de regulación de voltaje. Estos componentes son cruciales para garantizar la autonomía energética del sistema y evitar fluctuaciones de tensión que puedan comprometer el desempeño de los módulos de procesamiento y control. Asimismo, en el interior se ubica el dispositivo ESP32, encargado específicamente del procesamiento de audio proveniente de los micrófonos digitales. El ESP32 se seleccionó por su versatilidad, capacidad de procesamiento y soporte nativo para el protocolo I2S, esencial para la transmisión digital de señales de audio con alta fidelidad y bajo retardo.

En la parte superior de la estructura se dispusieron los módulos sensoriales principales, compuestos por un arreglo de micrófonos y la cámara OAK-D S2. La ubicación elevada y centrada de estos dispositivos se definió estratégicamente para maximizar el campo de visión y optimizar la captación de señales acústicas provenientes de distintas direcciones. La cámara OAK-D S2, equipada con capacidades de percepción de profundidad y procesamiento interno de redes neuronales, se instala en un ángulo frontal que facilita la identificación y seguimiento de objetos en el entorno inmediato del vehículo.

Por su parte, el arreglo de micrófonos, basado en sensores digitales INMP441, se encuentra distribuido de forma equidistante en la superficie superior para mejorar la precisión en la estimación de la dirección de llegada (DOA) de las fuentes sonoras. Este diseño favorece la separación espacial de las señales y permite realizar algoritmos de formación de haces o filtrado espacial en futuras extensiones del sistema.

El ensamblaje general busca una distribución funcional y ergonómica de todos los componentes, considerando tanto el equilibrio dinámico como la facilidad de mantenimiento y expansión. La disposición de placas paralelas ofrece una estructura rígida y modular que facilita el acceso a los módulos internos para tareas de inspección, sustitución de baterías o ajustes en los circuitos electrónicos. Además, este enfoque modular permite incorporar o sustituir sensores y actuadores adicionales, adaptando el vehículo a nuevas tareas o escenarios operativos con mínima reconfiguración mecánica.

3.2. Modelado de la dinámica del vehículo

Para obtener un modelado adecuado de las ruedas omnidireccionales de acuerdo con su dirección de rotación y la

velocidad procederemos con el análisis físico y matemático correspondiente. En la figura 2 podemos observar un diagrama de cuerpo libre para la velocidad en una rueda omnidireccional, así como sus componentes que van una en dirección paralela a la rueda y una componente a 45° entre la dirección de rotación y el exterior del vehículo que se produce por los mecanismos rotativos en la misma.

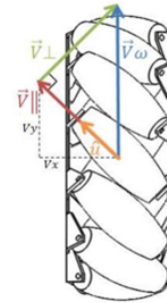


Figura 2: Diagrama de cuerpo libre para la velocidad en una rueda omnidireccional

Al considerar la distribución de ruedas elegida, es necesario pensar en la suma vectorial que pueden tener estas fuerzas y, por lo tanto, la dirección de vector de aceleración que tiene el sistema para ciertos esfuerzos. Al tener una distribución de cuatro ruedas distribuidas en forma de cuadrilátero, cada una con su respectivo motor DC, se pueden comenzar a realizar combinaciones de movimientos, que, por medio de un juego vectorial, se avanza de diferentes formas.

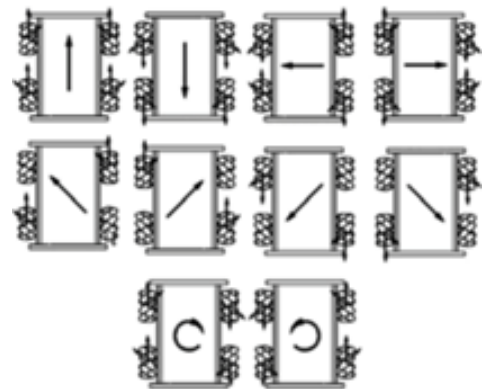


Figura 3: Movimiento resultante según combinaciones de giros en ruedas omnidireccionales

En la figura 3 se observan las diversas combinaciones de movimientos posibles para las ruedas, indicando en la parte inferior de cada combinación, el movimiento general que tiene el sistema. Sin embargo, cabe aclarar que la imagen sólo muestra las imágenes para una velocidad constante en todas las ruedas (en todos los motores DC) en movimiento. Lo que se busca es la elección de una combinación de movimiento en las ruedas, para después permitir la variación de la velocidad de estas, de tal forma que se permita la variación del ángulo de respuesta entre los 45° y 90° respecto del eje x.

Para el desarrollo del movimiento, lo que se busca es un desplazamiento frontal con ángulos de inclinación, el cual se puede lograr con un análisis vectorial sencillo.

$$T_{1,y} = T_1 \sin(\alpha), T_{1,x} = T_1 \cos(\alpha) \quad (1)$$

$$T_{4,y} = T_4 \sin(\alpha), T_{4,x} = T_4 \cos(\alpha) \quad (2)$$

$$T_{2,y} = T_2 \sin(\beta), T_{2,x} = T_2 \cos(\beta) \quad (3)$$

$$T_{3,y} = T_3 \sin(\beta), T_{3,x} = T_3 \cos(\beta) \quad (4)$$

Donde:

- α es el ángulo del vector T_n (con $n = 1, 4$) respecto del eje x
- β es el ángulo del vector T_n (con $n = 2, 3$) respecto del eje x

Para una fuerza en los motores mayor que cero, sentido de giro antihorario y $\alpha = \beta = 45^\circ$:

$$T_{s,x} = T_{1,x} - T_{2,x} - T_{3,y} + T_{4,x} \quad (5)$$

$$T_{s,x} = (T_1 - T_2 - T_3 + T_4) * \cos(45^\circ) \quad (6)$$

$$T_{s,y} = T_{1,y} + T_{2,y} + T_{3,y} + T_{4,y} \quad (7)$$

$$T_{s,y} = (T_1 + T_2 + T_3 + T_4) * \sin(45^\circ) \quad (8)$$

El ángulo de dirección del vector resultante de fuerza φ está dado por:

$$\varphi = \tan^{-1} \left(\frac{(T_1 + T_2 + T_3 + T_4) * \sin(45^\circ)}{(T_1 - T_2 - T_3 + T_4) * \cos(45^\circ)} \right) \quad (9)$$

Si $T_1 = T_4$ y $T_2 = T_3$:

$$\varphi = \tan^{-1} \left(\frac{(T_1 + T_2)}{(T_1 - T_2)} * \tan(45^\circ) \right) \quad (10)$$

Con la ecuación (10) se establece el comportamiento del vehículo, de tal forma que se busca que el movimiento de la rueda 1, sea replicado por la rueda 4 (en relación con la velocidad angular); lo mismo ocurre con las ruedas 2 y 3. Esto permite simplificar el modelo del sistema y, en la práctica, hacer más simple los movimientos a seguir del móvil, ya que se asegura que su ángulo de dirección varíe desde -90° hasta -45° (cuando $T_1 < T_2$) y desde los -45° hasta los 90° (cuando $T_2 < T_1$), respecto del eje x . Tomando en cuenta que cuando $T_1 = T_2$, el ángulo de dirección es a 90° . Considerando que esto es únicamente la aportación de los rodillos de las ruedas se pueden generar movimientos que combinando la velocidad de rotación (que siempre irá a 90° o a -90°) se pueden generar los diferentes movimientos que se pueden realizar con esta disposición.

3.3. Estimación de la dirección de llegada del sonido (DOA) mediante GCC-PHAT

Para estimar la dirección de llegada del sonido (Direction Of Arrival) a partir de dos señales acústicas adquiridas simultáneamente por dos micrófonos, se implementó un algoritmo basado en la correlación cruzada generalizada con transformación de fase, conocido como GCC-PHAT (Generalized Cross-Correlation – Phase Transform). Este método es utilizado en sistemas de localización de fuentes sonoras debido a su capacidad para detectar el retardo temporal entre señales (Time Difference of Arrival, TDOA), incluso en presencia de ruido.

Cada que una fuente sonora emite una señal en el espacio, esta llega a ambos micrófonos con una ligera diferencia de tiempo que depende del ángulo de incidencia y de la distancia entre los sensores. Denotando las señales adquiridas como $x_1(t)$ y $x_2(t)$, se puede calcular el retardo temporal τ que maximiza la similitud entre ambas señales. En condiciones ideales, este retardo se relaciona con el ángulo de llegada θ mediante la siguiente expresión:

$$\theta = \sin^{-1} \left(\frac{c \cdot \tau}{d} \right) \quad (11)$$

Donde:

- c es la velocidad del sonido
- τ es el retardo
- d es la distancia entre micrófonos

Para estimar este retardo, se emplea la correlación cruzada generalizada, una extensión del enfoque clásico de correlación cruzada. En el dominio de la frecuencia, la función de correlación cruzada $R_{12}(\tau)$ entre $x_1(t)$ y $x_2(t)$ se calcula como:

$$R_{12}(\tau) = \mathcal{F}^{-1} \left\{ \frac{X_1(f) \cdot X_2^*(f)}{|X_1(f) \cdot X_2^*(f)|} \right\} \quad (12)$$

Donde:

- $X_1(f)$ y $X_2^*(f)$ son las transformadas de Fourier de las señales $x_1(f)$ y $x_2(f)$
- $*$ denota el conjugado complejo
- $\mathcal{F}^{-1}$ es la transformada de Fourier

La fracción en la expresión representa la fase normalizada de la señal, lo que constituye el paso más importante del algoritmo: la Transformada de Fase (PHAT). Esta operación elimina la influencia de la amplitud de las frecuencias, haciendo que el algoritmo sea más sensible a las diferencias de fase entre las señales, que son las que contienen la información del retardo.

Una vez obtenida la función de correlación $R_{12}(\tau)$ que indica el grado de alineación entre las dos señales para distintos posibles retardos τ . Esta función suele presentarse como una secuencia de valores (un arreglo), en la que cada posición representa un valor de retardo (positivo o negativo) entre las señales.

La estimación del retardo se realiza buscando el valor de τ que maximiza esta correlación. Es decir, se busca el índice donde la función $R_{12}(\tau)$ alcanza su valor máximo:

$$\hat{\tau} = \operatorname{argmax} R_{12}(\tau) \quad (13)$$

En la práctica, como la función se obtiene en forma discreta (un arreglo de puntos), el proceso es:

1. Se calcula $R_{12}(\tau)$ como la transformada inversa de la multiplicación compleja de las FFT de las señales normalizadas.
2. El resultado se desplaza para centrar el retardo cero en la mitad del arreglo.
3. Se localiza el índice del valor máximo en ese arreglo. Ese índice representa el número de muestras de desfase entre las señales.
4. Si se ha aplicado interpolación (upsampling), se ajusta el retardo dividiendo por la frecuencia de muestreo multiplicada por el factor de interpolación.

Entonces, al final el valor de τ se calcula como:

$$\tau = \frac{\text{desplazamiento}}{f_s \cdot F}, \quad (14)$$

Donde:

- f_s es la frecuencia de muestreo
- F es el factor de interpolación

Este valor de τ se utiliza después en (11) para estimar el ángulo de llegada del sonido.

Este algoritmo fue implementado en Python y ejecutado en una Raspberry Pi5, en la que se reciben datos acústicos provenientes de dos micrófonos digitales INMP441, conectados a una ESP32 que transmite por comunicación serial los datos en bloques temporales. Para mejorar la resolución de retardo, se aplicó una interpolación temporal de las señales usando técnicas de resampleo, lo cual permitió estimar ángulos con mayor precisión incluso cuando el retardo es inferior a un periodo de muestreo.

A pesar de su robustez, el algoritmo presenta ciertas limitaciones que deben de ser consideradas:

- Con dos micrófonos, el sistema únicamente puede estimar ángulos en un plano bidimensional, por lo que no puede distinguir si la fuente está delante o detrás del eje de los micrófonos
- Se asume una única fuente dominante, por lo que no es confiable en presencia de múltiples emisores simultáneos
- Si el retardo estimado excede el máximo permitido por la geometría, el resultado de la estimación es inválido. El límite está dado por:

$$|\tau| > \frac{d}{c} \quad (15)$$

3.4. Percepción visual mediante redes neuronales implementadas en una OAK-D S2

En esta sección, se utilizó un modelo personalizado basado en YOLOv10n, una de las últimas variantes de la familia YOLO (You Only Look Once), caracterizada por su arquitectura optimizada para lograr un equilibrio entre precisión y velocidad de inferencia. Este modelo fue previamente entrenado en PyTorch para identificar un objeto específico de interés, con el objetivo de facilitar tareas de búsqueda y seguimiento dentro del entorno del vehículo.

Una vez completado el proceso de entrenamiento, el modelo fue adaptado y convertido a un formato compatible con la OAK-D S2 utilizando las herramientas en línea provistas por Luxonis. Estas herramientas permiten transformar modelos en formatos estándar, como PyTorch o ONNX, a un formato optimizado para el pipeline de inferencia de la OAK-D S2, asegurando compatibilidad y eficiencia computacional. Gracias a esta transformación, la cámara puede ejecutar el modelo directamente en su procesador interno, liberando recursos de la unidad de cómputo principal y reduciendo significativamente la latencia del sistema.

El sistema de visión artificial implementado tiene la capacidad de identificar el objeto entrenado y proporcionar información detallada sobre el ítem detectado (etiqueta de clase) y su posición dentro del plano de la imagen. La cámara genera datos que incluyen coordenadas relativas a la posición del objeto en la pantalla (por ejemplo, centro del bounding box) y el nivel de confianza asociado a cada detección. Esta información resulta esencial para la estrategia de navegación autónoma, ya que permite determinar la dirección y magnitud de los movimientos necesarios para aproximarse al objeto o mantenerlo centrado en el campo visual.

Una vez procesados, los resultados generados por la OAK-D S2 son transmitidos a la Raspberry Pi 5, la cual actúa como unidad central de control. La Raspberry Pi recibe las coordenadas y la información de identificación del objeto, y a partir de estos datos ejecuta el algoritmo de control de movimiento del vehículo. El sistema de control implementado analiza en tiempo real la desviación lateral del objeto respecto al centro del campo de visión, generando comandos de corrección para cada motor con el fin de alinear el vehículo y optimizar la trayectoria hacia el objetivo. Esto principalmente con movimientos puramente rotatorios en la fase de alineación con el objetivo, aprovechando la arquitectura del vehículo.

Este esquema de procesamiento distribuido ofrece mayores capacidades de rendimiento y velocidad, las cuales son capaces de mantener un rendimiento elevado incluso en escenarios dinámicos o con múltiples objetos en el entorno. Asimismo, el uso de la OAK-D S2 permite reducir aún más el trabajo de procesamiento de nuestra Raspberry Pi 5, así como provee varias funcionalidades que pueden ser explotadas en futuras implementaciones como la capacidad de realizar algoritmos de tipo SLAM para una navegación más segura de nuestro vehículo de forma autónoma.

4. Resultados

Como parte de nuestros resultados destacamos la integración de todos los sensores, la alimentación eléctrica y el uso de la Raspberry Pi 5 y la ESP32 dentro del espacio destinado para el vehículo, el cual nos dio la suficiente capacidad de resguardo de la parte electrónica, así como las capacidades estructurales necesarias para lograr tener un sistema robusto y estable mecánicamente hablando.

4.1. Resultados de percepción sonora

A continuación se muestran algunos de los resultados obtenidos en la estimación del ángulo de dirección de la fuente sonora. Se realizaron varias pruebas en las que se colocó la fuente en diferentes direcciones y a distintas distancias del vehículo.

Tabla 1: Comparación entre ángulo real y ángulo estimado

Real angle (°)	Estimated angle (°)
3	4.9
32	37.6
74	72.2
21	25.5
38	33.2
19	13.9
70	64.9
69	67.5
7	9.2
13	16.5

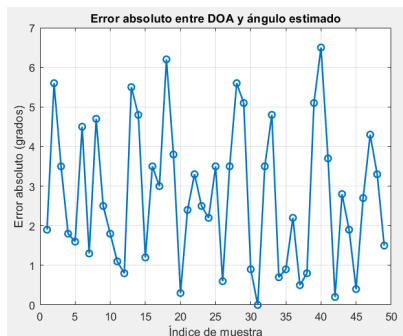


Figura 4: Error absoluto en grados de la estimación mediante DOA

Como se mencionó previamente, la estimación de la dirección de llegada del sonido se encuentra limitada por el hardware disponible. En este proyecto se utilizan únicamente dos módulos INMP441, lo que implica una restricción importante, ya que con esta configuración no es posible determinar si la fuente sonora se ubica al frente o detrás del eje formado por los micrófonos. Esta limitación podría resolverse mediante la implementación de un hardware diferente, por ejemplo, sustituyendo los módulos INMP441 por un arreglo con más de tres micrófonos, como el ReSpeaker USB Mic Array V3.1, que integra cuatro micrófonos. Esta alternativa disminuiría la carga de procesamiento de la Raspberry Pi 5 y haría innecesario el uso de la ESP32, al contar con un componente que procesa internamente las señales recibidas y estima la dirección de la

fuente.

El error absoluto entre las estimaciones y los ángulos reales utilizados en las pruebas no supera los 7°, lo cual cumple con la métrica establecida dentro de los 10° que se especificaron como límite aceptable. Además, recordando que la propuesta actual plantea incorporar la cámara OAK-D para compensar las limitaciones de detección de los micrófonos, se implementa no solo un sistema de estimación de la dirección de llegada del sonido, sino también un sistema de percepción visual. Ambas técnicas se complementan para conformar un sistema más robusto, capaz de guiar al vehículo hacia puntos de interés con gran precisión.

4.2. Resultados de percepción visual

Para validar de forma experimental el sistema propuesto se realizaron una serie de pruebas controladas que permitieran valorar de forma integral el desempeño del sistema en cuanto a la percepción visual. Para ello, se evaluó que la precisión en las detecciones de la red neuronal entrenada fuera mayor al 80 %.

Para probar la red neuronal, se utilizó una cámara OAK-D S2 y se obtuvieron inferencias a la velocidad de procesamiento de la cámara dando como resultados entre 40 y 60 fps, lo cual nos deja un tiempo promedio de obtención de datos de 20 ms, por lo cual nuestro sistema debía de ser adecuado para soportar este tiempo de muestreo, de igual forma se registraron resultados del entrenamiento con precisión mayor al 80 %, siendo exactamente de 83,6 %, lo cual representa un nivel de precisión aceptable en las inferencias realizadas.

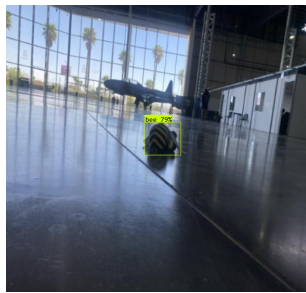
La matriz de confusión resultante del entrenamiento de la red, se presenta en la Figura 5. En esta matriz se puede observar una concentración significativa de aciertos sobre la diagonal principal, lo que indica que las predicciones de la red presentan un nivel de precisión adecuado.

	anger	bee	happy	sad	false neg.
anger	15	0	3	2	4
bee	0	56	0	0	1
happy	1	0	76	3	2
sad	3	0	1	22	11
False Pos.	9	0	8	10	0

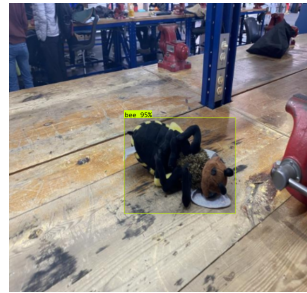
Figura 5: Matriz de confusión de la red neuronal entrenada

Los resultados obtenidos muestran un alto grado de fiabilidad en el reconocimiento de las clases “bee” y “happy”, mientras que demuestra deficiencias en las clases restantes, por lo cual se debe tomar en cuenta dichos resultados para futuros entrenamientos para mejorar el dataset propuesto o el método de entrenamiento utilizado.

De igual forma se incluyen como resultados algunas de las inferencias realizadas por el modelo de la red neuronal, donde podemos observar resultados donde las bounding boxes se acercan mucho a los propuestos como dataset, así como resultados donde no se puede ajustar con la misma fiabilidad, esto dependiendo de las condiciones en las cuales se encuentre el objeto a identificar, por ejemplo, en la inferencia de la Figura 6a se presenta una precisión en la detección del 79 % debido a la baja iluminación, mientras que en la inferencia de la Figura 6b se presentó una precisión mayor (del 95 %) al ser un entorno con una mejor iluminación.



(a) Inferencia con iluminación baja



(b) Inferencia con iluminación alta

Figura 6: Inferencias del modelo bajo diferentes condiciones de iluminación.

5. Conclusiones

El desarrollo e implementación del vehículo terrestre autónomo con percepción multimodal descrito en este trabajo permitió demostrar la viabilidad de integrar módulos avanzados de visión y audición en un único sistema móvil, capaz de interactuar y adaptarse dinámicamente a su entorno. En el caso de este sistema, la utilización de la cámara OAK-D S2, junto con un modelo personalizado basado en YOLOv10n, proporcionó una solución robusta y eficiente para la detección y el seguimiento de objetos en tiempo real, mientras que el módulo auditivo, implementado mediante el ESP32 y el protocolo I2S, permitió estimar la dirección de llegada (DOA) de fuentes sonoras con un grado de precisión adecuado para aplicaciones prácticas por la mezcla de percepción sonora y visual.

En comparativa con proyectos que están sobre la misma línea revisados en el marco teórico se obtienen mejores resultados al aplicar un sistema de percepción multimodal, por ejemplo, se obtienen resultados similares a los presentados por J. Pekár en la parte sonora, sin embargo, al incluir identificación visual se elimina ese error y se ignora la necesidad de tener un ambiente controlado, ampliando sus aplicaciones a diferentes áreas de desarrollo.

A pesar de lograr buenos resultados, consideramos que el presente desarrollo abre la puerta a múltiples mejoras y líneas de investigación futuras. Una de las principales áreas de oportunidad radica en optimizar la precisión del módulo auditivo, mediante el rediseño del arreglo geométrico de micrófonos, la implementación de técnicas avanzadas de filtrado espacial

(beamforming) y la reducción del impacto del ruido ambiental, como lo puede ser la aplicación de filtros, redes neuronales o algoritmos DOA más avanzados como MUSIC.

En el ámbito de la percepción visual, se podrían incorporar modelos más complejos que permitan reconocer múltiples clases de objetos y distinguir entre diferentes escenarios dinámicos, además de la implementación de las herramientas extras de la OAK-D S2 como la detección de profundidad y la creación de nubes de puntos para algoritmos de tipo SLAM.

Por otra parte, la arquitectura del vehículo puede verse mejorada mediante la implementación de mecanismos que permitan la movilidad más sencilla en rutas menos amistosas con el vehículo, para validar su uso en diferentes espacios y pueda utilizarse para aplicaciones al aire libre.

Desde una perspectiva de aplicación, el prototipo presentado podría adaptarse para tareas de búsqueda y rescate, patrullaje autónomo, asistencia en entornos industriales y escenarios de interacción social avanzada, entre otros. Estas posibles extensiones demuestran el potencial del sistema como base para el desarrollo de robots móviles inteligentes, capaces de actuar en entornos complejos y cambiantes de manera autónoma y segura.

En conclusión, el trabajo realizado sienta las bases para el diseño de plataformas robóticas más versátiles, resilientes y adaptativas, al tiempo que abre un amplio abanico de oportunidades para investigaciones futuras y desarrollos tecnológicos que contribuyan a la evolución de la robótica autónoma con percepción multimodal.

Referencias

- Benesty, J., Chen, J., y Huang, Y. (2008). *Microphone array signal processing*. Springer.
- Brandstein, M. y Ward, D. (2001). *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media.
- Campos Siurana, A. (2018). *Diseño e implementación de un vehículo de cuatro ruedas omnidireccionales*. Tesis doctoral, Universitat Politècnica de València.
- Duncan, J. A., Alambeigi, F., y Pryor, M. W. (2024). A survey of multimodal perception methods for human–robot interaction in social environments. *ACM Transactions on Human-Robot Interaction*, 13(4):1–50.
- Espressif Systems (2023). *ESP32 Technical Reference Manual*. Espressif Systems. Available online: https://www.espressif.com/sites/default/files/documentation/esp32_technical_reference_manual_en.pdf.
- Galarza Sánchez, F. J. (2019). Sistema de detección de peatones para vehículos autónomos.
- González Sierra, R. S. y Mafla Meneses, J. M. (2018). Sistema de visión artificial de apoyo para el comando de vehículos no tripulados.
- Jocher, G., Stokast, A., Abhishek, B. J., Mishra, N., Ojha, U., Wang, Z., Sandler, M., y Howard, A. (2023). YOLOv8: State-of-the-art object detection. *Ultralytics Documentation*.
- Khaleghi, B., Khamis, A., Karray, F. O., y Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44.
- Luxonis (2023). *OAK-D S2 Documentation*. Luxonis. Incluye especificaciones y guía de uso del modelo OAK-D S2 (visión robótica con percepción de profundidad) disponible online: documentación oficial de Luxonis.
- Pekár, J. (2023). Sound source localization from a microphone array.
- Siegwart, R., Nourbakhsh, I. R., y Scaramuzza, D. (2011). *Introduction to autonomous mobile robots*. MIT press.

Van Den Broeck, B., Bertrand, A., Karsmakers, P., Vanrumste, B., van Hamme, H., y Moonen, M. (2012). Time-domain gcc-phat sound source localization for small microphone arrays. En *Proceedings of 5th European DSP*

Education and Research Conference (EDERC), pp. 13–14. European DSP Education and Research Conference.