CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Atención multimodal y seguimiento robusto del interlocutor en HRI natural

Valencia-Rojas, A.*, Hormigo-Jiménez, P., Quemada-Torres, E., Moreno, F.-A., González-Jiménez, J.

Grupo de Percepción Artificial y Robótica Inteligente (MAPIR), Dept. de Ingeniería de Sistemas y Automática, Instituto Universitario en Ingeniería Mecatrónica y Sistemas Ciberfísicos (IMECH.UMA), Universidad de Málaga, Blvr. Louis Pasteur, 35, 29071 Málaga, España.

To cite this article: Valencia-Rojas, A., Hormigo-Jiménez, P., Quemada-Torres, E., Moreno, F.-A., González-Jiménez, J. 2026. Multimodal attention and robust interlocutor tracking in natural HRI Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13811

Resumen

Este artículo presenta una arquitectura multimodal diseñada para mejorar la naturalidad en la Interacción Humano-Robot. Para superar las ambigüedades espaciales, se integra una matriz de micrófonos que estima la dirección de llegada del sonido de forma omnidireccional. Esto permite al robot orientarse de forma natural mediante la coordinación de su cabeza y su base móvil. Además, se propone un método de reconocimiento visual eficiente que desplaza el peso computacional del análisis facial continuo hacia un seguimiento de cuerpo completo. Este enfoque utiliza una caché de identidades que logra mantener la identidad del interlocutor ante oclusiones temporales o giros del rostro. El sistema se completa con un agente conversacional gestionado por Grandes Modelos de Lenguaje y un módulo de expresividad que sincroniza la boca con el habla y adapta la expresión facial al contexto del diálogo. La arquitectura ha sido validada con éxito en un robot social del grupo de investigación MAPIR-UMA.

Palabras clave: Robótica inteligente, Interacción multimodal, Percepción y sensado, Robótica móvil, Interacción humano-vehículo

Multimodal attention and robust interlocutor tracking in natural HRI

Abstract

This paper presents a multimodal architecture designed to enhance naturalness in Human-Robot Interaction. To overcome spatial ambiguities, a microphone array is integrated to estimate the direction of arrival of sound omnidirectionally. This allows the robot to orient itself naturally through the coordination of its head and mobile base. Furthermore, an efficient visual recognition method is proposed that shifts the computational burden from continuous facial analysis to full-body tracking. This approach utilizes an identity cache that successfully maintains the interlocutor's identity during temporary occlusions or face turns. The system is completed with a conversational agent managed by Large Language Models and an expressivity module that synchronizes the mouth with speech and adapts the facial expression to the dialogue context. The architecture has been successfully validated on a social robot from the MAPIR-UMA research group.

Keywords: Intelligent robotics, Multi-modal interaction, Perception and sensing, Mobile robots, Human and vehicle interaction

1. Introducción

La incorporación de robots sociales a entornos compartidos con humanos demanda sistemas de Interacción Humano-Robot (HRI, por sus siglas en inglés) con características y capacidades que permitan una interacción lo más natural posible (Matheus et al., 2025).

De entre los desafíos que enfrentan estos sistemas, destaca el conseguir que el robot responda de forma fluida a estímulos de su entorno, sin importar de dónde provengan. Para ello, típicamente, se utilizan configuraciones multimodales de sensores (visuales y auditivos, principalmente) para conseguir una amplia área de atención, pero aún es habi-

*Autor para correspondencia: avalenciarojas@uma.es

tual que existan ambigüedades en determinadas situaciones (Cañete et al., 2024). Esta limitación restringe las capacidades del robot para responder a interacciones que se produzcan fuera de su campo de visión, afectando a la fluidez y naturalidad de la interacción. Para superar esta ambigüedad, en los últimos años se está produciendo una transición hacia configuraciones, denominadas multi-aurales, que emplean matrices de tres o más micrófonos en diversas geometrías (como circulares o esféricas). Estas matrices, combinadas con algoritmos de localización como MUSIC, estimación de diferencias de tiempo de llegada (TDOA), GCC-PHAT, o incluso apoyados por modelos basados en Redes Neuronales Convolucionales (CNN), permiten una estimación robusta de la dirección de llegada del sonido (*Direction of Arrival*, DoA) en un rango completo de 360º (Desai and Mehendale, 2022).

Otro pilar fundamental en la interacción natural con humanos es la capacidad de identificar al hablante. La detección de caras ha evolucionado desde algoritmos clásicos como el de Viola-Jones (Wang, 2014) hasta el uso de CNNs (Khan et al., 2024), pero sigue presentando ciertas limitaciones en entornos dinámicos como a los que se suelen enfrentar los robots sociales (Kumar et al., 2019). En general, su eficacia se ve comprometida por la sensibilidad a cambios en la pose del interlocutor o a oclusiones temporales, lo que provoca que el robot pierda el rastro de las identidades en escenarios multi-persona ante cruces o giros momentáneos, incluso si estas ya habían sido verificadas con anterioridad. Además, la ejecución del flujo completo de reconocimiento en cada instante —que abarca desde la detección y preprocesamiento hasta la extracción de características y su clasificación— implica una carga computacional elevada y, en gran medida, redundante. Trabajos recientes en HRI han demostrado que integrar algoritmos de seguimiento (*tracking*) para mantener la identidad del usuario a lo largo del tiempo mitiga estas deficiencias, reduciendo la carga de procesamiento y mejorando la tasa de reconocimiento (Khalifa et al., 2022).

Para solventar estas limitaciones, extendiendo el trabajo previo de (Cañete et al., 2024), este artículo propone una arquitectura multimodal orientada a incrementar la naturalidad en la interacción humano-robot. La propuesta combina un sistema de atención auditiva omnidireccional mediante una matriz de micrófonos y un procesamiento visual eficiente que desplaza el reconocimiento facial continuo hacia un seguimiento robusto de cuerpo completo (*body tracking*). Esta combinación permite determinar de manera unívoca la identidad del hablante y gestionar dinámicamente la atención en escenarios multi-persona. El sistema se completa con un agente conversacional basado en Grandes Modelos de Lenguaje (LLMs) y un módulo de expresividad afectiva que se adapta al contexto de la interacción. Esta arquitectura ha sido integrada y validada en la plataforma robótica Sancho (ver Figura 1).

2. Descripción general del sistema

En este trabajo, se ha utilizado una plataforma robótica con la siguiente configuración *hardware*: (i) un sistema de 4 micrófonos *ReSpeaker XMOS XVF3800* capaz de estimar el valor DoA de forma omnidireccional, (ii) una cámara RGB montada sobre la cabeza del robot, (iii) una base móvil con la capacidad de moverse y rotar sobre sí misma, (iv) una unidad *pan-tilt* para mover la cabeza, y (v) un sistema de altavoces. Además, el robot incorpora pantallas LCD en los ojos y una tira LED en la boca como elementos de retroalimentación visual encargados de la expresión de estados emocionales.

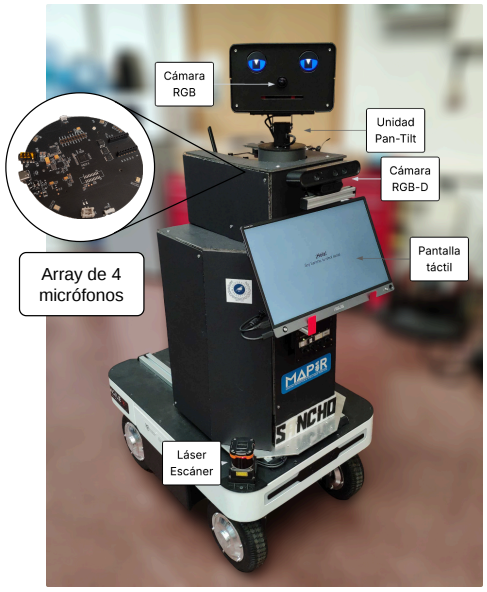

Figura 1: Robot de servicio Sancho del grupo de investigación MAPIR-UMA.

El funcionamiento general del sistema (ver Figura 2) se inicia al detectar una señal auditiva. El robot se orienta hacia el sonido y, tras identificar visualmente al usuario, fija su rostro y lo sigue mediante la unidad *pan-tilt*, gestionando así la atención de forma dinámica en entornos con múltiples personas. El flujo conversacional integra un modelo STT (*Speech-to-Text*) para capturar la petición, un LLM para generar la respuesta y un sistema TTS (*Text-to-Speech*) para sintetizarla vocalmente. Simultáneamente, el robot adapta su expresión facial según su estado (escuchando, pensando o hablando). Al concluir, el sistema retorna al estado de espera.

Aunque el sistema es capaz de identificar al hablante y mantener su seguimiento en el tiempo de manera robusta, el comportamiento del robot hacia la persona no se adapta al hablante. Este aspecto queda fuera del ámbito de este trabajo pero será abordado en futuros desarrollos.

3. Inicio de la interacción

La secuencia de percepción-actuación que se realiza para el inicio de la interacción involucra la detección de una señal sonora seguida de una rotación del robot hacia la dirección de la misma.

3.1. *Detección de la interacción*

La interacción comienza mediante la detección de una palabra de activación (*hotword*). Existe una gran variedad de algoritmos y modelos diseñados específicamente para esta tarea como Porcupine (Picovoice, 2026) u OpenWakeWord (Scripka, 2026). En este trabajo se ha optado por este último, aunque

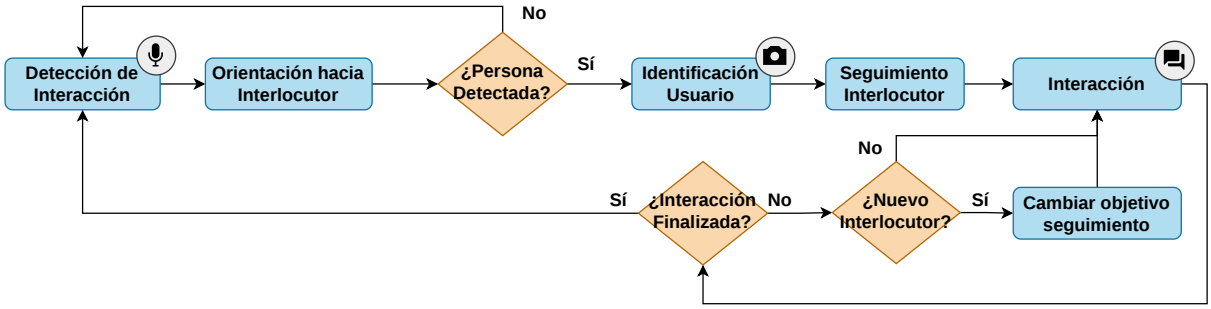

Figura 2: Diagrama de flujo del comportamiento general del sistema

puede ser intercambiado por cualquier otro. Un proceso en segundo plano será el encargado de leer fragmentos de audio captados por el sistema de micrófonos y detectar si en ellos se encuentra la palabra de activación.

3.2. *Orientación hacia el sonido*

Una vez detectado el inicio de la interacción, el sistema utiliza la DoA para responder físicamente al estímulo. Si el ángulo se encuentra dentro de un rango cercano al campo de visión del robot, éste reaccionará moviendo únicamente el cuello. Si por el contrario el estímulo proviene de los laterales o de la parte posterior, coordinará la cabeza y la base móvil para girar y lograr una reacción natural.

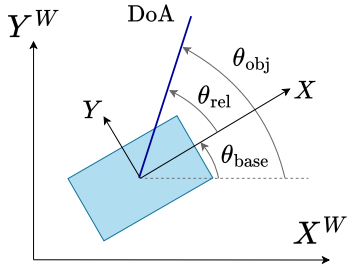

Figura 3: Cálculo del ángulo objetivo relativo al sistema de la base del robot.

Dado que el sistema de micrófonos se encuentra integrado en la base del robot, el ángulo de incidencia detectado (θ_{DoA}) es relativo a la orientación de dicha base. Para adaptar el movimiento deseado a los comandos de control del robot, esta medida local se transforma en un ángulo objetivo θ_{obj} en el sistema de referencia global del entorno, considerando la orientación de la base $\theta_{\text{base}}(t)$ en el instante de detección (t_0):

$$\theta_{\text{obj}} = \theta_{\text{base}}(t_0) + \theta_{\text{DoA}}$$

A partir de este valor absoluto, es posible determinar en todo momento el ángulo relativo $\theta_{\text{rel}}(t)$ hacia el objetivo. Este se obtiene como la diferencia entre θ_{obj} y $\theta_{\text{base}}(t)$ normalizada al intervalo $[-180°, 180°]$ para garantizar la trayectoria más corta (ver Figura 3):

$$\theta_{\text{rel}}(t) = \text{atan2}(\sin(\theta_{\text{obj}} - \theta_{\text{base}}(t)),\ \cos(\theta_{\text{obj}} - \theta_{\text{base}}(t)))$$

La primera reacción del sistema consiste en rotar el cuello hacia $\theta_{\text{rel}}(t_0)$ (equivalente a θ_{DoA}), saturando el resultado para respetar los límites físicos de los actuadores (habitualmente $\pm 90°$). Después de mandar la orden a la unidad *pan-tilt* que sostiene la cabeza, sin esperar a que esta termine de orientarse, el sistema introduce un retardo intencional y aleatorio (típicamente entre 0,3 y 1,0 segundos) antes de girar la base, emulando la latencia natural humana (Zangemeister and Stark, 1982). Pasado este tiempo, se inicia la rotación de la base para alinear el frente del robot con θ_{obj}.

Durante este giro interviene un movimiento de compensación para contrarrestar el giro del cuerpo, análogo al Reflejo Vestíbulo-Ocular (VOR) humano. Un bucle de control recalcula $\theta_{\text{rel}}(t)$ y envía comandos de posición al cuello para que la cabeza permanezca fijada en θ_{obj} mientras la base completa el movimiento. El resultado es un movimiento más natural del conjunto cuerpo-cabeza del robot, lo que redunda en una mejor interacción.

El flujo de toma de decisiones y control para este proceso de orientación se resume en la Figura 4.

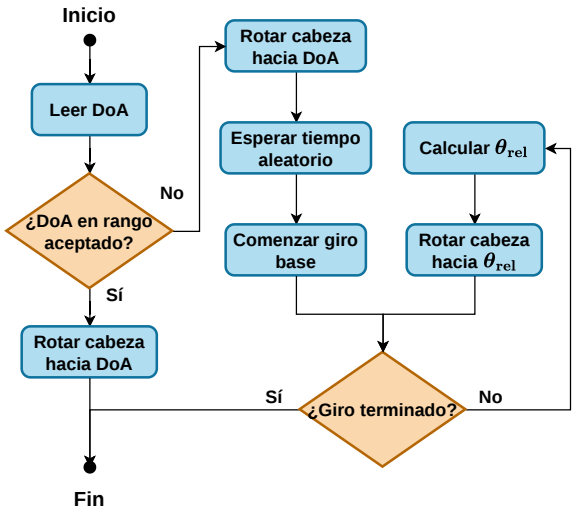

Figura 4: Diagrama de flujo sobre la orientación del robot hacia la dirección del sonido.

4. Reconocimiento eficiente de rostros

La segunda etapa de la interacción involucra el reconocimiento del hablante y el seguimiento eficiente de su cuerpo a lo largo del tiempo. Este procedimiento incluye (i) la detección y seguimiento continuo de los cuerpos de las personas presentes en la escena, (ii) la identificación de sus rostros mediante caché, y (iii) la especificación del hablante.

4.1. Seguimiento de cuerpos

A diferencia de los enfoques tradicionales que realizan un reconocimiento facial continuo, nuestro sistema implementa un seguimiento basado en el cuerpo completo de los interlocutores. Otras propuestas basadas exclusivamente en el rastreo de rostros (Khalifa et al., 2022) presentan deficiencias en escenarios dinámicos: si el usuario gira la cabeza, mira hacia abajo o da la espalda al robot, el detector facial falla, el rastreador pierde el objetivo y la identidad se desvanece, obligando a reiniciar el proceso de reconocimiento cuando el rostro vuelve a ser visible.

Para superar esta limitación, el primer paso de nuestra propuesta consiste en la detección y seguimiento continuo de las personas en la escena mediante el modelo YOLOX-Tiny (Ge et al., 2021) acoplado al algoritmo ByteTrack (Zhang et al., 2022). Esta configuración procesa la imagen para generar y mantener cajas delimitadoras (*bounding boxes*) de los cuerpos presentes, asignando a cada uno un identificador único (*Tracking ID*, TID). Dado que el cuerpo humano es un objetivo visualmente mucho más estable y robusto ante cambios en la orientación que un rostro, el sistema es capaz de mantener el TID de forma persistente incluso si la persona se gira por completo, se sienta de espaldas o sufre oclusiones temporales.

4.2. Reconocimiento puntual mediante caché de identidades

El último paso de la identificación del hablante se realiza de manera selectiva, actuando únicamente sobre las regiones de interés proporcionadas por el rastreador de cuerpos. Para cada persona detectada, el sistema aísla dinámicamente la región superior de su caja delimitadora (aproximadamente el 40-60 % superior, donde se espera que esté la cabeza) y la procesa mediante un modelo de detección facial (MTCNN (Zhang et al., 2016)).

Cuando se detecta un rostro válido y frontal, este se alinea, se extraen sus características mediante un modelo codificador (EfficientFaceV2S (Alansari et al., 2025)) y se determina su identidad. El aspecto clave de nuestra propuesta es que esta identidad se vincula inmediatamente al TID del cuerpo en una caché de identidades. Cuando el sistema vuelve a detectar un cuerpo con un TID que ya ha sido reconocido, revisará que ya existe una entrada en la caché y obtendrá directamente la información de la persona sin necesidad de reconocerlo de nuevo.

4.3. Especificación del interlocutor

Una vez que el sistema ha identificado a las personas presentes en la escena, el siguiente paso es determinar de manera unívoca cuál de ellas es la que está interactuando con el robot. Para ello, es necesario fusionar la información espacial del *array* de micrófonos sobre el DoA, con las detecciones realizadas con la cámara.

En primer lugar, el ángulo de llegada del sonido, que originalmente es relativo a la base del robot, se transforma para ser relativo al sistema de referencia de la cámara. Para evitar valores fuera de rango, este ángulo θ_{cam} se satura en el intervalo $[-89{,}9°, 89{,}9°]$.

Como paso previo al procesamiento, la imagen captada es rectificada y sobre ella se aplica el modelo de cámara estenopeica (*pinhole*) utilizando los parámetros intrínsecos del sensor: la distancia focal en el eje X (f_x) y el centro óptico (c_x). De este modo, se proyecta θ_{cam} sobre el plano de la imagen y se calcula la coordenada horizontal (x_{DoA}) de la fuente sonora (ver Figura 8):

$$x_{\text{DoA}} = c_x - f_x \cdot \tan(\theta_{\text{cam}})$$

Para cada persona detectada, se calcula la coordenada horizontal del centro de su caja delimitadora (x_{det}). La señal de audio se asocia con el individuo cuyo x_{det} minimiza la distancia absoluta con respecto a x_{DoA}. Para reducir asignaciones erróneas, se establece un umbral máximo de desviación, de manera que cualquier persona que lo exceda se descarta como posible interlocutor. Este umbral se fija de forma dinámica en función de la anchura de la caja delimitadora de cada persona (w_{det}), lo que permite al sistema adaptarse a diferentes distancias entre el interlocutor y el robot, exigiendo mayor precisión a medida que el sujeto se aleja.

5. Flujo conversacional y expresividad

Durante la conversación, los ojos y la boca del robot se adaptan para reflejar su estado operativo: (i) escuchando la petición del usuario (*escuchando*), (ii) procesando la información (*pensando*) o (iii) reproduciendo la respuesta (*hablando*). La transición entre estos estados permite al usuario identificar el momento preciso para intervenir, mejorando la fluidez del diálogo.

En el estado *hablando*, se ha implementado un mecanismo de modulación lumínica para dotar de mayor realismo a la síntesis de voz. El sistema analiza en tiempo real el flujo de audio del TTS, extrayendo el valor RMS (*Root Mean Square*) para mapearlo directamente hacia la tira LED de la boca. Este proceso hace que la intensidad y apertura de la boca varíen acorde con el volumen de la voz, simulando el movimiento de esta al hablar.

(a) Alegre (b) Enfadado

Figura 5: Representación visual de los estados emocionales en la interfaz facial.

De forma complementaria, el sistema integra una capa de expresividad afectiva basada en el contenido del diálogo. Cuando el LLM genera la respuesta, analiza su contexto

semántico para extraer una emoción coherente con el mensaje. Esta etiqueta emocional dicta la paleta de colores y el diseño de los patrones en ojos y boca (ver Figura 5); por ejemplo, la emoción *enfadado* prioriza tonos rojizos y formas angulares, mientras que *alegre* emplea tonos cálidos y una mayor apertura ocular.

Para cerrar el ciclo de la interacción, cuando al entrar en el modo de escucha no se detecta ninguna entrada de audio por parte del usuario, el robot vuelve a su estado de reposo.

Para un análisis detallado sobre el diseño de los *prompts* de extracción de emociones y las especificaciones técnicas del *hardware* de visualización, se remite al lector a la memoria técnica del proyecto Quemada-Torres (2025).

6. Pruebas y validación

Para evaluar el funcionamiento del sistema multimodal propuesto, se han diseñado algunos casos de uso ejecutados por el robot móvil Sancho en un entorno real de laboratorio. Estos escenarios de prueba se centran en contrastar el comportamiento de esta nueva arquitectura frente a las limitaciones de sistemas previos basados en audio estéreo y detección facial continua. Todo el código fuente implementado y los vídeos de demostración de estos procesos estarán disponibles en `https://github.com/MAPIRlab/sancho_robot`.

6.1. *Resolución de ambigüedad espacial*

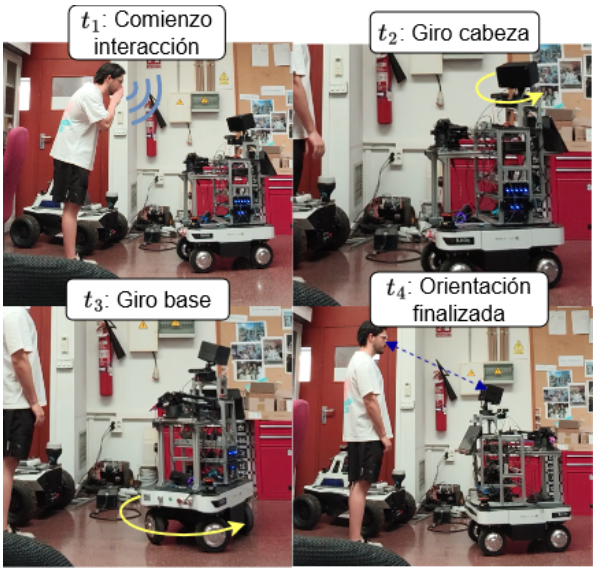

Figura 6: Reorientación del robot ante un estímulo sonoro proveniente de su espalda, coordinando el movimiento de la cabeza y la base móvil.

Este primer escenario evalúa la capacidad del robot para orientarse ante un estímulo sonoro proveniente de su parte posterior, fuera de su campo de visión inicial. En implementaciones previas con micrófonos estéreo (Cañete et al., 2024), el sistema sufría una indeterminación espacial que le impedía discernir si el usuario se encontraba delante o detrás del robot. Al integrar la matriz de cuatro micrófonos, esta ambigüedad desaparece.

La Figura 6 ilustra el instante inicial donde el interlocutor llama al robot desde su espalda. El sistema actual estima correctamente la dirección de llegada y coordina la cinemática: primero rota la cabeza y, tras un breve retardo, gira la base móvil compensando el movimiento del cuello. El resultado es un reposicionamiento exitoso frente a un estímulo que el modelo anterior no habría podido ubicar correctamente.

6.2. *Mantenimiento de identidad ante giros y oclusiones del rostro*

El segundo caso evalúa la robustez del sistema ante la pérdida temporal de visibilidad del rostro. A diferencia de los enfoques basados en detección facial continua, que presentan baja robustez bajo estas condiciones, nuestro sistema de seguimiento de cuerpo completo vincula de forma persistente la identidad al identificador único (TID) en caché tras una única verificación frontal.

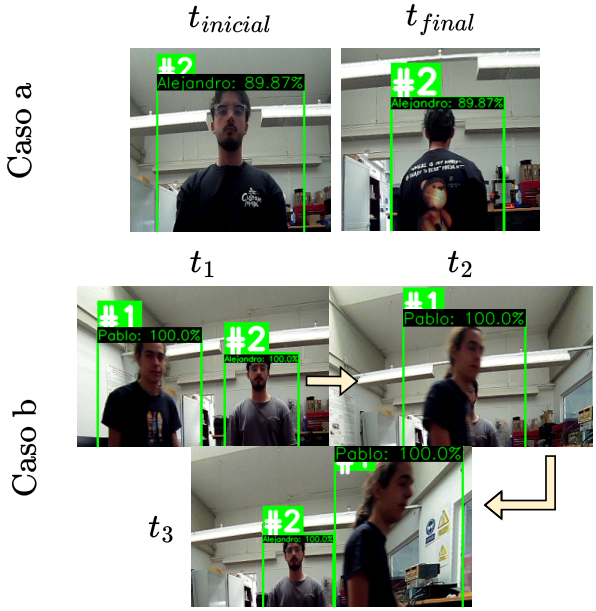

Figura 7: Preservación de la identidad del usuario (TID) ante (a) pérdida de rasgos faciales y (b) cruce temporal en la línea de visión.

La Figura 7 ilustra esta ventaja en dos escenarios: el interlocutor dando la espalda al robot (*Caso a*) y un tercero cruzándose en la línea de visión (*Caso b*). En ambos casos, el algoritmo garantiza la persistencia de la identificación sin asignaciones erróneas, favoreciendo una interacción continua y fluida.

6.3. *Seguimiento del interlocutor*

El tercer escenario de evaluación demuestra la capacidad del sistema para integrar la información acústica y visual con el objetivo de identificar y seguir dinámicamente al interlocutor activo en un entorno con múltiples personas.

Como se ilustra en la Figura 8, en el instante inicial (t_1), el sistema detecta una señal de voz. La estimación del ángulo de llegada (DoA), representada visualmente por la línea vertical azul, se proyecta sobre el plano de la imagen. El sistema

asocia esta coordenada horizontal con la persona ubicada a la izquierda, identificándolo como el interlocutor principal.

Posteriormente, en el instante t_2, la segunda persona presente en la escena interviene en la conversación. El sistema detecta inmediatamente el cambio en la dirección de origen del sonido, asociando la nueva señal acústica con la caja delimitadora del segundo sujeto.

Finalmente, en el instante t_3, una vez verificado el cambio de hablante, el bucle de control actúa sobre los motores de la unidad *pan-tilt*. La cabeza del robot gira automáticamente para centrar al nuevo interlocutor en su campo de visión, lo que demuestra un comportamiento de atención visual natural y dinámico frente a interacciones multipersona.

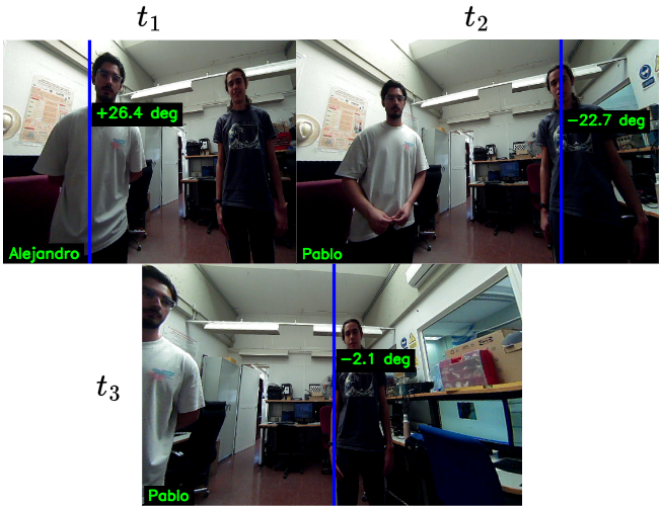

Figura 8: Identificación del interlocutor y seguimiento de su posición utilizando una unidad *pan-tilt*.

7. Conclusiones y trabajos futuros

Este trabajo presenta una arquitectura multimodal de interacción humano-robot sobre la plataforma Sancho, centrada en incrementar su naturalidad. La combinación de una matriz de micrófonos, el seguimiento continuo de cuerpos y un *pipeline* conversacional basado en LLMs permite resolver de forma eficaz las ambigüedades espaciales y de identificación. Asimismo, la coordinación de la respuesta verbal con expresiones visuales (ajustadas tanto a la semántica del mensaje como a la acústica en tiempo real) confiere al robot un comportamiento más natural.

Como trabajo futuro, se plantea optimizar la gestión de turnos en diálogos grupales, integrar memoria a largo plazo para personalizar interacciones con usuarios recurrentes y explorar Modelos de Lenguaje Visual (VLM) que enriquezcan la expresividad durante el diálogo. Adicionalmente, se considera investigar el uso de modelos de incertidumbre para fusionar de forma probabilística la identidad de la voz con la del rostro, aportando información complementaria que incremente la precisión y la tasa de aciertos en la clasificación.

Agradecimientos

Este trabajo ha sido desarrollado en el contexto de los proyectos MINDMAPS (PID2023-148191NB-I00) y Voxeland (PPRO-B1-2023-017, JA.B1-09), financiados por el Ministerio de Ciencia e Innovación y la Universidad de Málaga, respectivamente, así como del Plan Andaluz de Investigación, Desarrollo e Innovación (PPRO-TEP960-G-2023).

Referencias

Alansari, M., Alnuaimi, K., Ganapathi, I., Alansari, S., Javed, S., Shoufan, A., Zweiri, Y., Werghi, N., 2025. Efficientfacev2s: A lightweight model and a benchmarking approach for drone-captured face recognition. Expert Systems with Applications 273, 126786.
DOI: https://doi.org/10.1016/j.eswa.2025.126786

Cañete, A., Quemada-Torres, E., Ruiz-Sarmiento, J.-R., Moreno, F. Á., Gonzalez-Jimenez, J., 2024. Sistema multimodal para la orientación de robots móviles hacia su interlocutor. Jornadas de Automática 45.
DOI: https://doi.org/10.17979/ja-cea.2024.45.10939

Desai, D., Mehendale, N., 2022. A review on sound source localization systems. Archives of Computational Methods in Engineering 29, 4631–4642.
DOI: https://doi.org/10.1007/s11831-022-09747-2

Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430.
DOI: https://doi.org/10.48550/arXiv.2107.08430

Khalifa, A., Abdelrahman, A. A., Strazdas, D., Hintz, J., Hempel, T., Al-Hamadi, A., 2022. Face recognition and tracking framework for human-robot interaction. Applied Sciences 12 (11), 5568.
DOI: https://doi.org/10.3390/app12115568

Khan, S. S., Sengupta, D., Ghosh, A., Chaudhuri, A., 2024. Mtcnn++: A cnn-based face detection algorithm inspired by mtcnn. The Visual Computer 40, 899–917.
DOI: https://doi.org/10.1007/s00371-023-02822-0

Kumar, A., Kaur, A., Kumar, M., 2019. Face detection techniques: a review. Artificial Intelligence Review 52 (2), 927–948.
DOI: https://doi.org/10.1007/s10462-018-9650-2

Matheus, K., Ramnauth, R., Scassellati, B., Salomons, N., 2025. Long-term interactions with social robots: Trends, insights, and recommendations. J. Hum.-Robot Interact. 14 (3).
DOI: https://doi.org/10.1145/3729539

Picovoice, 2026. Porcupine: On-device wake word detection powered by deep learning. https://github.com/Picovoice/porcupine, repositorio de GitHub. Accedido el 21 de abril de 2026.

Quemada-Torres, E., Jun. 2025. Sistema multimodal de interacción humano robot basado en técnicas de ia. Trabajo fin de grado, Universidad de Málaga.
URL: https://hdl.handle.net/10630/40843

Scripka, D., 2026. openwakeword: An open-source audio wake word detection framework. https://github.com/dscripka/openWakeWord, repositorio de GitHub. Accedido el 21 de abril de 2026.

Wang, Y.-Q., 2014. An Analysis of the Viola-Jones Face Detection Algorithm. Image Processing On Line 4, 128–148.
DOI: https://doi.org/10.5201/ipol.2014.104

Zangemeister, W. H., Stark, L., 1982. Gaze latency: Variable interactions of head and eye latency. Experimental Neurology 75 (2), 389–406.
DOI: https://doi.org/10.1016/0014-4886(82)90169-8

Zhang, K., Zhang, Z., Li, Z., Qiao, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23 (10), 1499–1503.
DOI: https://doi.org/10.1109/LSP.2016.2603342

Zhang, Y., Sun, P., et al., 2022. Bytetrack: Multi-object tracking by associating every detection box. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer, pp. 1–21.
DOI: https://doi.org/10.1007/978-3-031-20047-2_1