ORIOLUS
Revista de divulgación científica de la UTN

Enero - Junio 2026
Vol. 1 No. 1

DOI 10.47633/oriolus1.1-7

DAVID ALONSO FLORES CASTRO, JOEL ADRIÁN GARCÍA BRAVO, JORGE GUDIÑO-LAU, PEDRO C. SANTANA-MANCILLA*

Universidad de Colima

Autor de correspondencia: *Pedro C. Santana-Mancilla (psantana@ucol.mx).

Imagen ilustrativa / Creada con Inteligencia Artificial. ChatGPT, OpenAI, 2026.

ISSN:2215-6100

ORIOLUS
Revista de divulgación científica de la UTN
Enero - Junio 2026
Vol. 1 No. 1

¿POR QUÉ IMPORTA QUE UN ROBOT CONVERSE?

Hablar con un robot ya no es solo cosa de la ciencia ficción. En laboratorios, aulas y espacios de demostración, la voz se ha convertido en una forma natural de pedir ayuda, explicar una tarea o corregir una acción. Esta evolución es importante porque muchas personas pueden tener interés en la robótica, pero no necesariamente saben programar ni redactar instrucciones técnicas.

En los últimos años, la Interacción Humano-Robot (HRI, por sus siglas en inglés) ha comenzado a abandonar la rigidez de los comandos codificados para entrar en un terreno mucho más natural: el de la conversación (Wang et al., 2025). Recientemente, los grandes modelos de lenguaje, denominados LLMs (Ramos-Rivera et al., 2025), han abierto una ruta prometedora: utilizar el lenguaje cotidiano como puente entre la intención humana y la acción robótica (Ahn et al., 2022; Zhang et al., 2023).

Estos modelos pueden interpretar frases incompletas, expresiones coloquiales y peticiones que tengan cierto grado de ambigüedad. Esta transición marca un punto importante: ya no basta con que los robots sean precisos; ahora deben ser comprensivos, accesibles y capaces de responder a instrucciones emitidas de la manera en que las personas realmente hablan, con matices, informalidad, acentos, dudas y frases incompletas (Hancock et al., 2011; Sheridan, 2016).

Con este objetivo, en este estudio se desarrolló y evaluó un prototipo basado en un robot educativo. El prototipo no se fabricó desde cero. Se utilizó una plataforma educativa comercial, el Yahboom DOFBOT , capaz de seguir instrucciones habladas mediante un sistema que combina reconocimiento de voz, interpretación semántica con un LLM y control cinemático ejecutado en una Jetson Nano (Gudiño-Lau et al., 2025). La meta no era solo hacer que el robot se moviera, sino explorar qué tan natural, intuitiva y confiable podía ser la interacción con estudiantes universitarios.

DEL OÍDO AL MOVIMIENTO: ¿CÓMO FUNCIONA EL SISTEMA?

La HRI ha avanzado notablemente con robots colaborativos, sensores hápticos y sistemas de visión, pero el verdadero avance se produce cuando el lenguaje humano se convierte en la interfaz principal. Los LLMs han demostrado una capacidad sobresaliente para procesar intenciones expresadas de manera flexible (Li et al., 2025), incluso cuando la instrucción no es ni directa ni técnica.

Para lograr que un robot no solo oiga, sino que también entienda, se diseñó una arquitectura (ver **Figura 1**) basada en el concepto de Interfaz de Lenguaje Natural para la Robótica (NLIR). El sistema imita un proceso cognitivo simplificado en tres etapas: oído, cerebro y cuerpo. En un contexto educativo, esta capacidad resulta especialmente valiosa: los estudiantes no hablan como programadores, sino como personas. Usan referencias ambiguas, comparaciones relativas y expresiones coloquiales que normalmente confundirían a un sistema computacional. Integrar un LLM en el flujo de interacción permite que el robot escuche y piense antes de actuar, acercando la experiencia a una conversación más que a una instrucción mecánica.

Oído: La primera capa del sistema utiliza Vosk, una herramienta de reconocimiento de voz que opera localmente en la computadora del usuario (Alpha Cephei, 2026). Vosk transcribe la voz a texto sin necesidad de enviar el audio a la nube en esta etapa, actuando como un primer filtro rápido y seguro.

Cerebro: El texto transcrito se envía a Gemini, el modelo de lenguaje multimodal de Google (Gemini Team et al., 2023). Mediante una instrucción diseñada para esta tarea, el modelo actúa como traductor de intenciones: recibe frases como “ve más a la izquierda” o “saluda” y las convierte en una categoría de acción que el robot puede ejecutar.

Cuerpo: Una vez que la IA decide qué hacer, la instrucción viaja a través de los protocolos TCP/IP hacia una Jetson Nano, una pequeña pero potente computadora que controla los servomotores del brazo robótico DOFBOT. Este controlador traduce la intención abstracta en señales eléctricas precisas para mover los grados de libertad del robot mediante la ejecución en el borde (Edge computing).

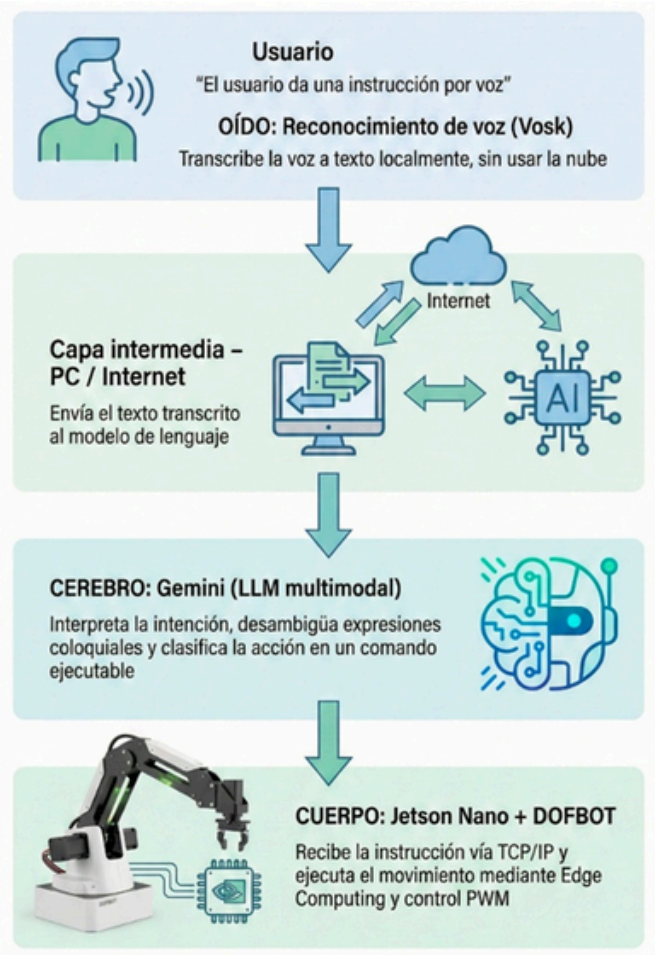

Figura 1. Arquitectura del sistema NLIR implementado: del usuario al robot.

Además de interpretar instrucciones en lenguaje natural, el sistema incorpora un componente de retroalimentación verbal que resultó fundamental para la experiencia del usuario. Mientras ejecuta cada movimiento, el robot responde con frases breves (como saludar, despedirse o anunciar la acción que está por realizar); esto ayuda a confirmar que la instrucción fue entendida y refuerza la percepción de claridad y confianza

¿CÓMO SE EVALUÓ LA INTERACCIÓN?

Para evaluar qué tan natural y efectiva podía ser esta interacción conversacional, se diseñó un protocolo en el que el robot debía ejecutar siete acciones básicas: saludar, despedirse y moverse a cinco posiciones espaciales (derecha, extrema derecha, centro, izquierda, extrema izquierda). Estas acciones se representaron en un tablero impreso que sirvió de guía visual para los participantes.

Participaron 21 miembros del estudiantado de la Universidad de Colima en México, con edades entre 18 y 23 años. Respecto al sexo, predominó el masculino (76%) frente al femenino (24%). El 57% de los usuarios no tenía experiencia previa significativa con robots, lo cual resultó útil para validar la facilidad de uso del sistema para usuarios novatos. El 95% ya había usado asistentes como Alexa o Siri, lo que sugiere que la barrera de entrada para la interacción por voz era baja.

Del total de estudiantes, nueve pertenecían a áreas no tecnológicas y doce a áreas de tecnologías de la información. Esta decisión intencional aseguró que la evaluación no quedara condicionada por la familiaridad técnica y que los estilos de habla fueran lo más heterogéneos posible.

Cada participante debía lograr que el DOFBOT ejecutara cada acción dos veces (ver **Figura 2**), pero había un elemento crucial: podía pedirlas en el orden que quisiera y con cualquier expresión natural. Antes de hablar, el usuario marcaba en silencio en un tablero la acción deseada; después, pronunciaba la instrucción. De esta manera, fue posible comparar objetivamente la intención del usuario con la acción efectivamente realizada por el robot.

Las frases pronunciadas resultaron tan variadas como cabría esperar en una interacción no restringida. Hubo frases directas ("*Muévete hacia la derecha*"), instrucciones comparativas ("*Ve más a la izquierda*"), expresiones vagas ("*Muévete poquito*") y órdenes conversacionales ("*Ve todo lo que puedas a la derecha*"). Esta diversidad permitió poner a prueba no solo la capacidad técnica del robot, sino, sobre todo, su tolerancia al lenguaje humano tal como se usa realmente.

Figura 2. Estudiante interactuando con el robot.

RESULTADOS

Se midieron los resultados en tres dimensiones: aciertos, errores y la percepción de los estudiantes.

En total se registraron 220 aciertos, de las dos ejecuciones buscadas de cada movimiento por parte de los participantes. El sistema logró una tasa general de acierto del 74.8%, un resultado notable, dado que los comandos eran completamente libres y no existía un diccionario oficial de instrucciones. Las acciones más exitosas fueron centro, izquierda, saludo y despedida, todas con valores superiores a 85% de precisión.

Las acciones más difíciles fueron las extremas: la extrema derecha y la extrema izquierda (ver **Figura 3**). En el primer caso, la precisión bajó a 38%. Esto ocurrió principalmente cuando los participantes usaron comparativos ("*un poco menos*") o expresiones graduales ("*ve todo lo que puedas*"). Para un robot que solo puede elegir entre posiciones discretas, estos matices resultaron confusos incluso cuando el modelo entendía la dirección general. La distinción entre "derecha" y "extrema derecha", por ejemplo, sigue siendo un reto cuando el lenguaje humano no la expresa con claridad.

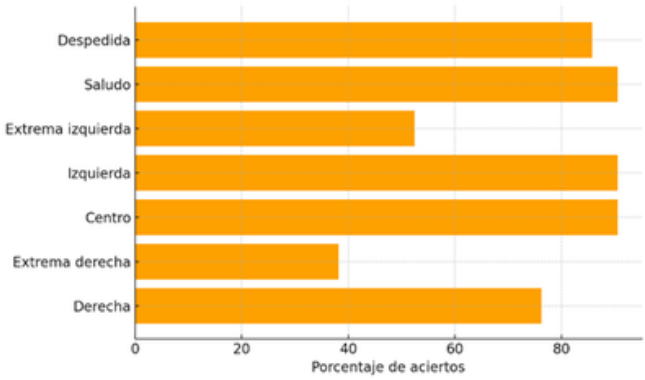

Figura 3. Porcentaje de aciertos por acción.

¿DÓNDE SE EQUIVOCÓ EL SISTEMA Y POR QUÉ?

Para comprender mejor el comportamiento del sistema, cada error se clasificó en una de cuatro categorías (véase la **Tabla 1**).

Tabla 1. Catálogo de errores.

Tipo de error	Descripción
Comando de voz	Ocurre cuando el usuario titubea, pronuncia de forma poco clara, hace una pausa prolongada o deja la instrucción incompleta.
Interpretación (LLM)	Ocurre cuando la frase fue transcrita correctamente, pero el modelo la clasifica como una acción distinta de la esperada.
Transcripción	Ocurre cuando Vosk no reconoce correctamente una palabra importante debido al ruido, al acento o a una variación fonética.
Falta de contexto	Ocurre cuando la instrucción no contiene suficiente información para elegir entre posiciones discretas.

Como se aprecia en la **Tabla 2** y la **Figura 4**, se registró un total de 74 errores (25.2% de las instrucciones). El 43.2% de los errores se debió a la falta de contexto. En estos casos, el usuario pedía un movimiento, pero la frase no contenía suficientes elementos para decidir entre posiciones discretas. Expresiones como "*muévete un poquito*" o "*ve más para allá*" tienen sentido para una persona que comparte el espacio visual, pero no para un robot que no puede inferir referencias espaciales implícitas.

Tabla 2. Porcentaje de errores por tipo.

Tipo de error	Número de errores	Porcentaje
Comando de voz	5	6.8%
Interpretación (LLM)	24	32.4%
Transcripción	13	17.6%
Falta de contexto	32	43.2%
Suma total	**74**	**100%**

La segunda categoría más frecuente fue la de los errores de interpretación del LLM (32.4%). El modelo comprendía la frase, pero devolvía una acción distinta de la esperada. Esto ocurrió cuando la instrucción era semánticamente ambigua o cuando el matiz del usuario podía corresponder a más de una categoría posible; por ejemplo, "*ve a la izquierda*" podía interpretarse como la acción "*ver*", en lugar de "*ir*".

Los errores de transcripción representaron el 17.6% y se debieron a ruidos, acentos o variaciones fonéticas que llevaron a Vosk a confundir palabras relevantes. Por último, un 6.8% correspondió a errores en los propios comandos humanos: frases incompletas ("*muévete a…*") o referencias físicas del propio robot ("a tu lado derecho").

Aunque estos errores pudieran parecer limitaciones, en realidad representan un hallazgo fundamental. La interacción natural no es perfecta, pero eso no la hace menos valiosa. Comprender dónde fallan los sistemas permite construir robots más robustos y diseñar experiencias más fluidas para usuarios no especializados.

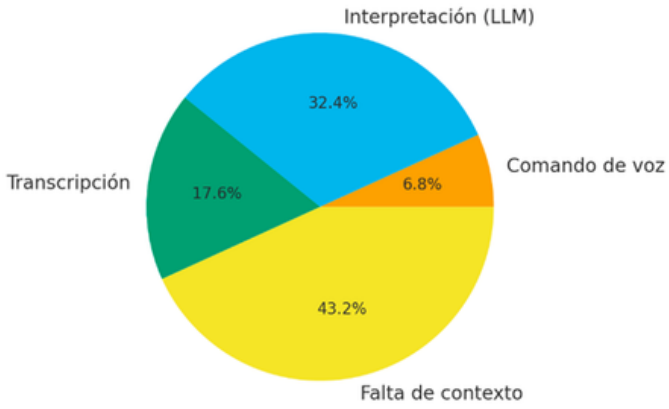

Figura 4. Distribución de errores por tipo.

¿CÓMO PERCIBIERON LA EXPERIENCIA LOS USUARIOS?

Además de medir aciertos y errores, se evaluó la percepción de los estudiantes mediante un cuestionario tipo Likert de 1 a 5, donde 1 significó "*totalmente en desacuerdo*" y 5 "*totalmente de acuerdo*". A partir de las respuestas de los 21 participantes se construyó un índice global de interacción, cuyo promedio fue de 4.35 puntos, lo que indica una experiencia mayoritariamente positiva. Los detalles más relevantes se presentan en la **Tabla 3.**

Tabla 3. Desempeño en el análisis de la percepción.

Dimensión evaluada	Media (1 a 5)	Interpretación
Éxito general	4.76	Los usuarios percibieron que el robot realizó los movimientos solicitados.
Entendimiento	4.76	Se consideró que el sistema comprendió correctamente las instrucciones.
Comodidad	4.71	Hablar con el robot no generó incomodidad significativa.
Claridad	4.62	Las respuestas del robot fueron fáciles de comprender.
Confianza	4.62	Los estudiantes confiaron en que el robot interpretaría sus órdenes.
Naturalidad	4.38	La interacción se sintió conversacional, aunque con margen de mejora.
Velocidad	4.29	La latencia fue el aspecto más débil en términos relativos.
Tono de voz	4.62	La voz del robot fue bien recibida.

Los datos sugieren que la percepción de éxito global se asocia fuertemente con dos factores: que la interacción haya sido agradable y sin frustraciones, y que el robot haya realizado correctamente los movimientos solicitados.

Es decir, los estudiantes tienden a considerar exitosa la experiencia cuando el robot se mueve como esperan y el proceso se siente fluido y poco frustrante; la rapidez de respuesta o la voz agradable, aunque valoradas, parecen jugar un papel secundario frente a estos dos elementos.

El análisis por subgrupos muestra que estas percepciones son relativamente consistentes entre los perfiles. Los estudiantes con experiencia previa en el uso de robots presentaron un índice global de interacción muy similar al de quienes no la tienen, y lo mismo ocurre al comparar hombres y mujeres. Esto indica que la interfaz basada en lenguaje natural logra nivelar la experiencia, de modo que no se requieren conocimientos técnicos avanzados para sentirse cómodo e interactuar con el sistema.

Las respuestas abiertas ayudan a matizar estas cifras. Entre lo que más les gustó, los estudiantes mencionan que el robot "interpreta lenguaje natural, no te ves forzado a usar comandos predefinidos", que la interacción se siente "amistosa" y que la voz resulta "agradable". Varios subrayan la curiosidad por observar cómo el robot interpreta sus instrucciones y la posibilidad de "hablarle como a un compañero".

En cuanto a áreas de mejora, aparecen de forma recurrente la necesidad de contar con más retroalimentación "dar información sobre cuándo inicia o termina la acción", una reducción del tiempo de reacción y una mayor cobertura de frases coloquiales. Pese a estas observaciones, casi todos responden que dar instrucciones por voz fue fácil, porque "son frases muy sencillas y las entiende muy bien" o porque "es más rápido que escribir las instrucciones".

DISCUSIÓN Y CONCLUSIONES

Los resultados muestran que un LLM conectado a un robot físico puede interpretar una amplia variedad de instrucciones naturales y, en la mayoría de los casos, ejecutar movimientos coherentes. Alcanzar un 74.8% de precisión con lenguaje libre constituye una base sólida para futuras iteraciones, especialmente porque la percepción de los participantes fue ampliamente positiva. El análisis también evidencia el desafío central de la interacción natural: el lenguaje humano es ambiguo y contextual, y suele expresar matices espaciales difíciles de traducir en categorías discretas de acción física.

La **Figura 5** permite observar de manera integrada este comportamiento y muestra un desempeño consistentemente alto en dimensiones como el éxito general, la comprensión, la claridad y la comodidad. El ligero hundimiento en la dimensión de Velocidad revela el principal punto de fricción: la transcripción de Vosk fue la etapa más lenta del flujo, lo que afectó la fluidez percibida. Aun así, la experiencia se mantuvo estable y bien valorada.

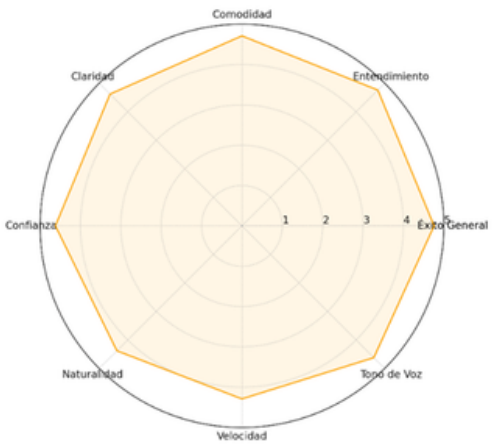

Figura 5. Percepción por dimensión.

En conjunto, el sistema desarrollado demuestra que es posible construir una interfaz de lenguaje natural para la robótica que funcione de manera intuitiva para usuarios no especializados. El DOFBOT, apoyado por un LLM multimodal, no solo ejecuta tareas: participa en una interacción más cercana a la forma en que las personas realmente se expresan.

Los desafíos identificados (en particular en matices espaciales, ambigüedad y latencia) representan oportunidades para mejorar la naturalidad y la inmediatez del sistema. Este experimento abre una ventana a una nueva generación de robots que conversan, interpretan y se ajustan al lenguaje humano, lo que hace que la robótica educativa sea más accesible, comprensible y humana.

DECLARACIÓN DEL USO DE LA IA GENERATIVA

Este estudio utilizó IA generativa para: (1) generar imágenes (ChatGPT 5.1 y Gemini 3.0), (2) interpretar semánticamente los comandos del usuario (API Gemini 2.5 Flash) durante todos los experimentos con el robot, y (3) realizar la revisión gramatical del manuscrito, sin alterar el contenido científico ni las conclusiones.

REFERENCIAS

- Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., … Zeng, A. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances (Versión 2). arXiv. https://doi.org/10.48550/ARXIV.2204.01691
- Alpha Cephei. (2026). Vosk Speech Recognition Toolkit. https://alphacephei.com/vosk/
- Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2023). Gemini: A Family of Highly Capable Multimodal Models (Versión 5). arXiv. https://doi.org/10.48550/ARXIV.2312.11805
- Gudiño-Lau, J., Durán-Fonseca, M., Anido-Rifón, L. E., & Santana-Mancilla, P. C. (2025). A Symmetry-Informed Multimodal LLM-Driven Approach to Robotic Object Manipulation: Lowering Entry Barriers in Mechatronics Education. Symmetry, 17(10), 1756. https://doi.org/10.3390/sym17101756
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors: The Journal of the Human Factors and Ergonomics Society, 53(5), 517-527. https://doi.org/10.1177/0018720811417254
- Li, Z., Zhang, H., Peng, C., & Peiris, R. (2025). Exploring Large Language Model-Driven Agents for Environment-Aware Spatial Interactions and Conversations in Virtual Reality Role-Play Scenarios. 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR), 1-11. https://doi.org/10.1109/VR59515.2025.00025
- Ramos-Rivera, R. E., Santana Mancilla, P. C., Garcia-Mancilla, J., & Gaytán-Lugo, L. S. (2025). Modelos de lenguaje en educación: Inteligencia Artificial Generativa para optimizar el análisis del desempeño docente. INNOVACADEMIA, 1(2), 70-81. https://doi.org/10.29105/innoacad.v1i2.36
- Sheridan, T. B. (2016). Human–Robot Interaction: Status and Challenges. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(4), 525-532. https://doi.org/10.1177/0018720816644364
- Wang, J., Shi, E., Hu, H., Ma, C., Liu, Y., Wang, X., Yao, Y., Liu, X., Ge, B., & Zhang, S. (2025). Large language models for robotics: Opportunities, challenges, and perspectives. Journal of Automation and Intelligence, 4(1), 52-64. https://doi.org/10.1016/j.jai.2024.12.003
- Zhang, C., Chen, J., Li, J., Peng, Y., & Mao, Z. (2023). Large language models for human–robot interaction: A review. Biomimetic Intelligence and Robotics, 3(4), 100131. https://doi.org/10.1016/j.birob.2023.100131