ASISTENTE INTELIGENTE DE COMPRAS CON INTELIGENCIA ARTIFICIAL ESPECIALIZADA

INTELLIGENT SHOPPING ASSISTANT WITH SPECIALIZED ARTIFICIAL INTELLIGENCE

María Elisa Rojas[1]
https://orcid.org/0009-0002-0071-8146

Recibido: 14-05-2026
Aceptado: 23-06-2026

Resumen

El siguiente artículo, cuyo objetivo es desarrollar un asistente de compras con inteligencia artificial especializada, integrado en el microcontrolador ESP32-S3, capaz de consultar información en tiempo real del inventario para automatizar la atención al cliente en establecimientos minoristas. La investigación se enmarcó en la modalidad de proyecto factible con carácter de campo, descriptivo, responde a la necesidad de comercios de pequeña y mediana escala de contar con herramientas accesibles para la consulta autónoma de información. El sistema desarrollado, denominado MIA (Módulo de Inteligencia Artificial), integra hardware embebido compuesto por un microcontrolador ESP32-S3, micrófono digital INMP441, amplificador de audio MAX98357A y pantalla circular GC9A01. La arquitectura de software se estructuró en seis etapas: activación, captura de voz, transcripción, procesamiento conversacional, síntesis de voz y reproducción. La incorporación dinámica del inventario mediante recuperación contextual ligera permitió generar respuestas coherentes y especializadas. Las pruebas arrojaron un índice de coherencia global del 100%, una precisión de transcripción del 90% y tiempos de respuesta de entre 17 y 26 segundos por ciclo. Los resultados confirman la viabilidad técnica y económica de incorporar inteligencia artificial conversacional en dispositivos de bajo costo para la modernización del comercio minorista.

***Palabras clave**:* inteligencia artificial; ESP32-S3; sistemas embebidos; atención al cliente; comercio minorista.

Abstract

This study aimed to develop a specialized artificial intelligence-powered shopping assistant, integrated into an ESP32-S3 microcontroller, capable of querying real-time inventory information to automate customer service in retail establishments. The research was framed within the feasible project modality with a technological focus, addressing the need for small and medium-sized retailers to have accessible tools for autonomous information retrieval. The developed system, named MIA (Artificial Intelligence Module), integrates embedded hardware consisting of an ESP32-S3 microcontroller, an INMP441 digital microphone, a MAX98357A audio amplifier, and a GC9A01 circular display. The software architecture was structured into six stages: activation, voice capture, transcription, conversational processing, speech synthesis, and playback. The dynamic incorporation of inventory data through lightweight contextual retrieval allowed the assistant to generate coherent and specialized responses within the defined commercial domain. Evaluation tests yielded a global coherence index of 100%, a transcription accuracy rate of 90%, and response times between 17 and 26 seconds per cycle. The results confirm the technical and economic feasibility of incorporating conversational artificial intelligence into low-cost embedded devices for the modernization of the retail sector.

Keywords: artificial intelligence; ESP32-S3; embedded systems; customer service, retail.

[1] Universidad Yacambú. Venezuela. Correo: v-28388257@Micorreo.uny.edu.ve

Introducción

En el contexto actual del comercio minorista, la atención al cliente representa uno de los factores más determinantes en la experiencia de compra y, al mismo tiempo, uno de los eslabones con mayor fragilidad operativa. Los establecimientos de pequeña y mediana escala enfrentan una tensión estructural permanente: mientras el consumidor moderno exige respuestas inmediatas, información precisa y trato personalizado, el modelo tradicional de atención depende casi exclusivamente de la disponibilidad y el conocimiento del personal de ventas. Esta dependencia genera tiempos de espera, saturación del personal en horas pico, errores en la transmisión de información y una experiencia fragmentada que contrasta directamente con las expectativas de un consumidor cada vez más informado y exigente.

En la última década, el avance del procesamiento de datos impulsado por la Inteligencia Artificial (IA) ha redefinido las fronteras de la automatización. Investigaciones previas han explorado el potencial de estas tecnologías en entornos comerciales: Vázquez García (2024) analizó la implementación de chatbots basados en IA generativa en organizaciones del mercado español, concluyendo que estos sistemas ofrecen una precisión comparable a la humana en la gestión de servicios al cliente. En el ámbito de los sistemas embebidos, Benito (2019) documentó las capacidades del microcontrolador ESP32 para aplicaciones IoT conectadas a servicios en la nube, demostrando su robustez para prototipos de bajo costo. Por su parte, Lewis et al. (2021) exploraron arquitecturas que combinan modelos de lenguaje con recuperación de información contextual, demostrando que es posible consultar bases de datos específicas en tiempo real reduciendo significativamente los errores de alucinación en las respuestas generadas.

En ese marco, este trabajo propone el desarrollo de MIA (Módulo de Inteligencia Artificial), un asistente inteligente de compras integrado en el microcontrolador ESP32-S3. A diferencia de los asistentes de respuesta preprogramada, MIA emplea modelos de lenguaje de gran tamaño (LLM) a través de la API de OpenAI y una estrategia de recuperación contextual ligera que incorpora dinámicamente el inventario real del comercio como contexto de procesamiento, sin requerir motores de búsqueda semántica ni bases vectoriales. Esto permite al sistema generar respuestas sobre disponibilidad, precios, ubicación y recomendaciones de productos, restringiendo adecuadamente las consultas fuera del dominio comercial definido.

La investigación se enmarca en la línea de Innovación de procesos industriales y productos tecnológicos de la Universidad Yacambú (UNY) y tiene como propósito demostrar que es posible incorporar IA conversacional especializada en dispositivos embebidos de bajo costo como alternativa accesible para la modernización del comercio minorista de pequeña y mediana escala. El presente artículo se organiza en las siguientes secciones: materiales y métodos, donde se describe la arquitectura del

sistema y las fases de desarrollo; resultados, con las pruebas de evaluación del prototipo; discusión, donde se contrastan los hallazgos con la literatura existente; y conclusiones con las líneas de mejora identificadas.

Fundamentos Teóricos

La Inteligencia Artificial (IA) constituye un área de la informática orientada al desarrollo de sistemas capaces de ejecutar tareas que requieren capacidades cognitivas humanas. Según Russell y Norvig (2021), estos sistemas se orientan al razonamiento y la toma de decisiones basada en datos. En el contexto de esta investigación, la IA se emplea de forma especializada para optimizar la interacción en el dominio concreto del comercio minorista, restringiendo el comportamiento del asistente a la información real del inventario del establecimiento.

Una API (Application Programming Interface) es un conjunto de reglas y protocolos que permite que diferentes aplicaciones se comuniquen entre sí, intercambiando datos y funcionalidades mediante solicitudes estructuradas. En este sentido, IBM (2024) señala que una API permite que distintas aplicaciones de software interactúen entre sí, posibilitando la integración y reutilización de servicios dentro de entornos digitales interconectados. En el presente proyecto, las APIs de OpenAI constituyen el núcleo del procesamiento cognitivo del sistema, delegando en la nube las tareas de transcripción de voz, procesamiento conversacional y síntesis de audio que el microcontrolador no puede ejecutar localmente.

Un sistema embebido es un sistema informático diseñado para realizar funciones específicas dentro de un dispositivo mayor, generalmente con restricciones de hardware (Wolf, 2017). Dentro de esta categoría, el módulo ESP32-S3 WROOM N16R8 destaca por integrar conectividad WiFi, una arquitectura de doble núcleo Xtensa LX7 y memoria extendida de 16 MB Flash y 8 MB PSRAM, características que lo convierten en un componente idóneo para prototipos de bajo costo que requieren procesar tareas de complejidad media y mantener conectividad estable con servicios en la nube. Según Maier et al. (2017), estas características hacen del ESP32 una plataforma adecuada para aplicaciones IoT de bajo consumo donde coexisten tareas de captura de señal, comunicación inalámbrica y control de periféricos.

El Internet de las Cosas (IoT) consiste en la interconexión de objetos físicos mediante redes digitales, permitiendo la comunicación entre dispositivos y el procesamiento de información del entorno. Evans (2011) sostiene que la integración de IoT posibilita la interacción desde dispositivos ubicados físicamente en el punto de venta con bases de datos remotas, permitiendo el procesamiento de solicitudes en tiempo real. En el sistema desarrollado, el ESP32-S3 actúa como nodo IoT que conecta el entorno físico del comercio con los servicios de inteligencia artificial en la nube.

La recuperación contextual ligera, estrategia adoptada en este trabajo, consiste en suministrar al modelo de lenguaje información específica del dominio antes de generar la respuesta, sin requerir indexación semántica ni bases vectoriales. A diferencia de los modelos de lenguaje generales, que responden únicamente con base en el conocimiento adquirido durante su entrenamiento, este enfoque permite incorporar información actualizada y específica del entorno operativo. Sus principales ventajas son la actualización dinámica de la información sin necesidad de reentrenamiento del modelo, el control del dominio de respuesta limitando la generación a datos específicos del comercio, y la simplicidad de implementación en dispositivos embebidos con recursos limitados.

Finalmente, la Interfaz Hombre-Máquina (HMI) se refiere a la capa de interacción entre el usuario y el sistema electrónico. Según Shneiderman et al. (2016), en sistemas físicos comerciales es fundamental que la interfaz sea intuitiva y reduzca la curva de aprendizaje del usuario. En MIA, esta interfaz se materializa a través de dos canales complementarios: la interacción por voz mediante el micrófono INMP441 y el altavoz conectado al amplificador MAX98357A, y la retroalimentación visual mediante la pantalla circular GC9A01, que comunica el estado operativo del asistente en tiempo real.

Materiales y Métodos

La investigación se definió bajo la modalidad de Proyecto Factible con carácter de campo descriptivo, orientada a la creación de un prototipo funcional mediante la aplicación de conocimientos de ingeniería, microcontroladores e inteligencia artificial. Según Arias (2012), esta modalidad permite el diseño de soluciones tecnológicas sustentadas en un diagnóstico de necesidades, lo que se corresponde con el enfoque adoptado. El diseño fue de campo, descriptivo, organizado en cuatro fases metodológicas: diagnóstico de necesidades, estudio de factibilidad, diseño, construcción, evaluación y pruebas técnicas.

Fase I: Diagnóstico de Necesidades

En esta fase se llevó a cabo un análisis exhaustivo de la situación actual de la atención al cliente en el sector minorista, con el propósito de identificar las brechas tecnológicas que podían ser atendidas mediante un sistema embebido. El proceso se sustentó en una revisión de fuentes especializadas en inteligencia artificial, sistemas de control basados en microcontroladores y protocolos de comunicación inalámbrica, lo que permitió establecer el estado del arte y definir los criterios de desempeño técnico que el prototipo debía alcanzar. A partir de este diagnóstico se identificaron como requerimientos fundamentales: conectividad inalámbrica estable, procesamiento de voz en lenguaje natural, capacidad de consulta de inventario en tiempo real y retroalimentación visual del estado del sistema.

Tabla 1

Personal involucrado en la investigación

Personal	Oficio/Rol
Estudiante María Elisa Rojas Botello	Realización e implementación del estudio
Tutor Douglas Domínguez	Revisión y asesorías

Nota. Elaboración propia.

Tabla 2

Cronograma de Ejecución

Fase	Semanas
Diagnóstico y planificación	2
Diseño y desarrollo del prototipo	3
Pruebas funcionales del prototipo	2
Ajustes y documentación final	1

Nota. Elaboración propia.

Fase II: Estudio de Factibilidad

El estudio de factibilidad permitió determinar la viabilidad real de la propuesta desde tres dimensiones. En términos de factibilidad técnica, se constató la disponibilidad en el mercado de los componentes electrónicos esenciales y del ecosistema de software necesario, incluyendo entornos de desarrollo como VS Code y la infraestructura de servicios en la nube de OpenAI. En cuanto a la factibilidad operativa, se verificó que el prototipo podía integrarse en escenarios simulados para validar su funcionamiento sin afectar procesos existentes. Respecto a la factibilidad económica, se elaboró un presupuesto detallado que contempló los componentes de hardware y los costos de los servicios de procesamiento de inteligencia artificial, resultando en una inversión total de 46,34 euros, significativamente inferior a las soluciones comerciales equivalentes disponibles en el mercado.

Tabla 3

Presupuesto del Prototipo Tecnológico

Componente	Cantidad	Costo total	Observaciones
ESP32-S3 WROOM N16R8	1	9,46 €	Microcontrolador con WiFi/Bluetooth
Módulo de micrófono INMP441	1	3,27 €	Para entrada de voz del usuario
Amplificador MAX98357 I2S	1	5,95 €	
Altavoz 40mm	1	3,75 €	Para salida de voz de la IA
Saldo de OpenAI inicial	-	5 €	Saldo para pruebas
Placa de fibra de vidrio perforada	1	6,23 €	Placa perforada para soldar los componentes
Pantalla GC9A01 240x240	1	7,68	Pantalla circular
Otros (cables jumper, estaño, etc.)	-	5 € aprox	Materiales adicionales para pruebas
Total Inversión		**46,34 €**	

Nota. Elaboración propia.

Fase III: Diseño y Construcción del Prototipo

Esta fase materializó la arquitectura tecnológica del sistema, integrando hardware de procesamiento de señal digital con servicios de inteligencia artificial en la nube. El diseño comprendió tres componentes principales: la arquitectura de hardware, la arquitectura de software y el flujo de datos entre ambas capas.

Arquitectura de Hardware

Se utiliza el microcontrolador ESP32-S3 WROOM N16R8, seleccionado por su amplia memoria PSRAM y Flash, indispensables para el manejo de buffers de audio. Los periféricos y protocolos de comunicación se configuran de la siguiente manera:

Sistema de Adquisición y Salida de Audio

Se implementa el protocolo I2S (Inter-IC Sound) para garantizar una transferencia de datos de audio digital libre de ruido. El micrófono INMP441 captura la voz del usuario, mientras que el amplificador MAX98357 gestiona la señal hacia un altavoz de 40mm para la reproducción de la respuesta sintetizada. Los pines asignados al bus I2S son: WS (GPIO 11), SD (GPIO 13) y SCK (GPIO 12).

Interfaz Táctil

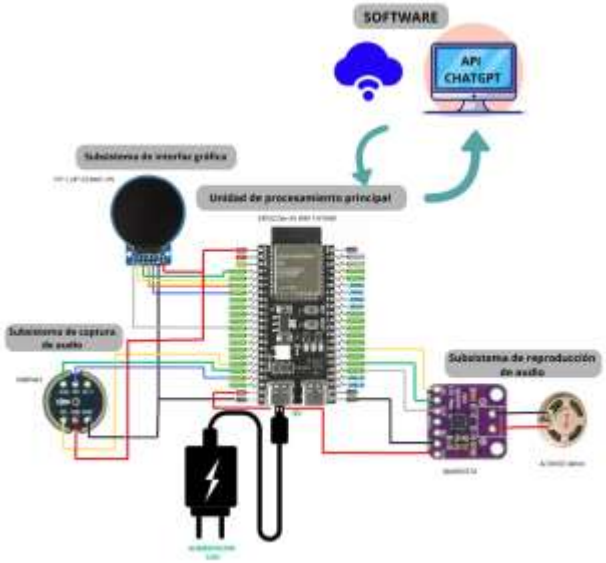
Para el control de interacción, se configura el GPIO 9 como sensor táctil capacitivo (TOUCH_PIN), el cual actúa como disparador de la interrupción que inicia el ciclo de escucha. Este mecanismo permite al usuario activar el asistente de forma intuitiva con un simple toque, sin necesidad de botones físicos.

Interfaz Visual

La retroalimentación visual del sistema se gestiona mediante la pantalla circular LCD GC9A01 de 240×240 píxeles, conectada al microcontrolador a través del protocolo SPI. Esta pantalla presenta al usuario el estado operativo del asistente en tiempo real mediante un sistema de colores, símbolos e iconos de interpretación intuitiva. Los estados representados son: en espera (idle), escuchando, procesando y respondiendo. Adicionalmente, durante la fase de respuesta, la pantalla puede mostrar fragmentos del texto generado o información relevante del producto consultado, ofreciendo así una experiencia multimodal que complementa la interacción por voz.

Figura 1

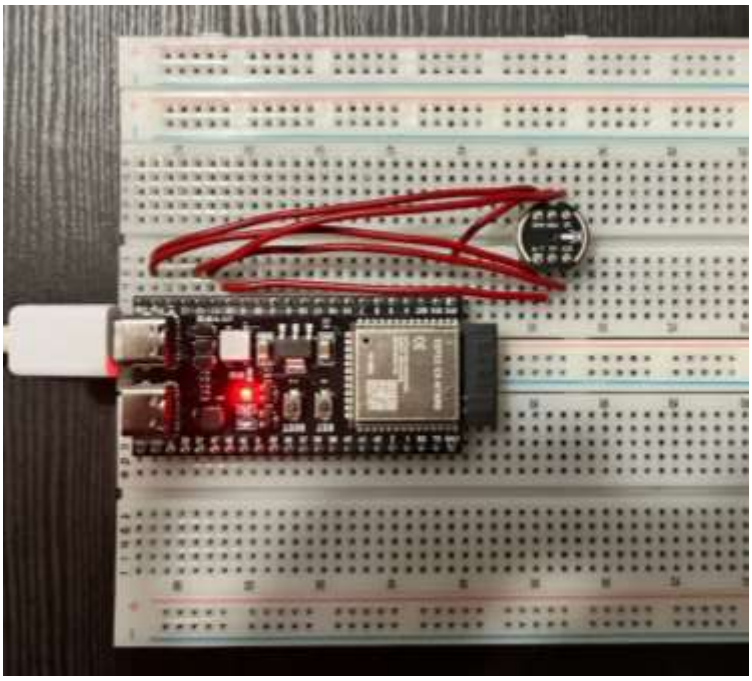
Diagrama de conexiones del sistema.



Nota. Elaboración propia.

Figura 2

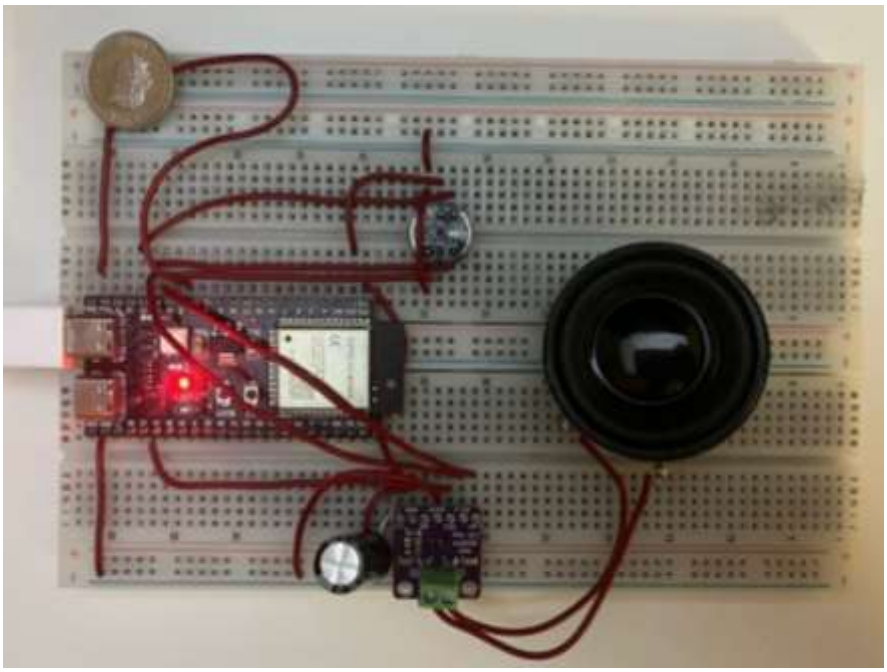
Integración progresiva de los componentes del circuito 1



Nota. Elaboración propia.

Figura 3

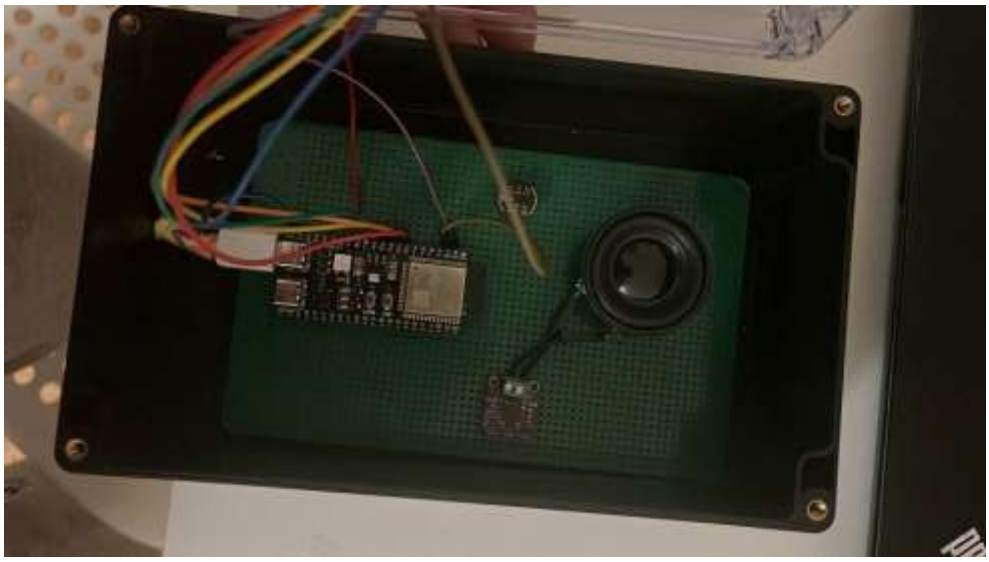
Integración progresiva de los componentes del circuito 3



Nota. Elaboración propia.

Figura 4

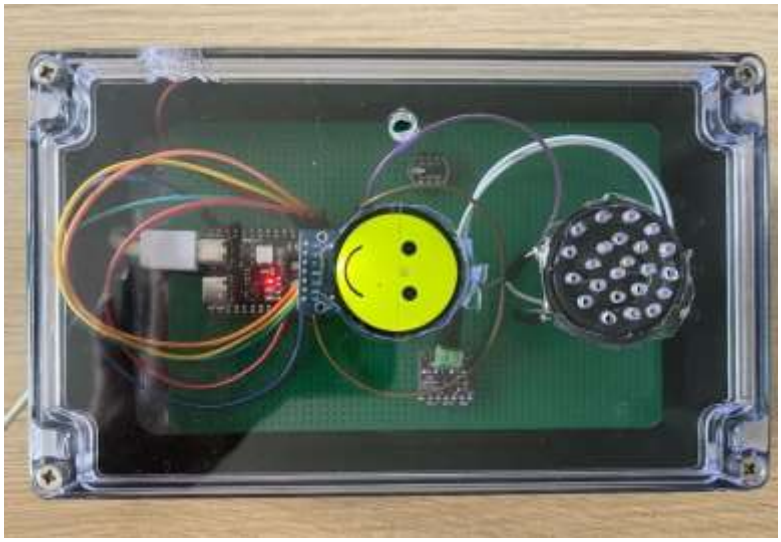
Integración progresiva de los componentes del circuito 4



Nota. Elaboración propia.

Figura 5

Imagen del prototipo ya montado



Nota. Elaboración propia.

Arquitectura de Software y Flujo de Datos

El software del sistema fue diseñado siguiendo un flujo operativo secuencial dividido en etapas funcionales claramente definidas. Cada etapa representa una fase del proceso de interacción entre el usuario y el asistente inteligente, permitiendo transformar una solicitud por voz en una respuesta audible generada mediante inteligencia artificial.

Etapa 1: Activación del Asistente

El sistema permanece en estado de espera hasta que el usuario activa el asistente mediante contacto sobre el sensor táctil capacitivo conectado al ESP32-S3. El microcontrolador supervisa continuamente el valor leído mediante la función touchRead(), comparándolo con un umbral previamente establecido. Cuando el valor detectado supera dicho umbral, el software identifica una activación válida e inicia el proceso completo de interacción. La detección se realiza únicamente ante una transición desde estado inactivo a activo, evitando ejecuciones repetidas durante un mismo contacto.

El almacenamiento local del sistema se implementa mediante el sistema de archivos FFat integrado en la memoria flash del ESP32-S3, lo que permite gestionar información persistente necesaria para el funcionamiento autónomo del asistente.

Etapa 2: Captura de Audio

Una vez activado el sistema, el ESP32-S3 configura el periférico I2S en modo recepción para establecer comunicación directa con el micrófono digital INMP441 y proceder a la captura de la voz del usuario.

El micrófono transmite audio digital en formato de 32 bits a una frecuencia de muestreo de 16 kHz. Dado que los datos útiles se encuentran en los bits más significativos de cada muestra, el software aplica un desplazamiento binario para convertirlas al formato de 16 bits, adecuado para su almacenamiento y procesamiento posterior.

El audio capturado se guarda en la memoria flash interna como un archivo WAV denominado rec.wav, incluyendo su cabecera correspondiente para garantizar compatibilidad con servicios externos de reconocimiento de voz. Este archivo actúa como intermediario entre la captura local y la etapa de transcripción en la nube. La grabación se realiza durante un tiempo previamente definido, asegurando estabilidad en el uso de memoria del sistema embebido.

Etapa 3: Transcripción de Voz (Speech-to-Text)

Finalizada la grabación, el archivo de audio es enviado mediante conexión WiFi a un servicio de reconocimiento automático de voz, con el fin de convertir la señal acústica capturada en texto digital interpretable por el sistema.

El ESP32-S3 construye una solicitud HTTP que incluye el archivo WAV y la transmite hacia el servidor remoto. El servicio procesa el audio mediante modelos de inteligencia artificial especializados en reconocimiento del habla y devuelve una respuesta en formato JSON que contiene la transcripción textual de la consulta del usuario, la cual constituye la entrada lingüística del asistente inteligente.

Etapa 4: Procesamiento Conversacional

Durante esta fase, el software incorpora tanto la consulta del usuario como información contextual relacionada con el comercio, incluyendo el catálogo de productos almacenado localmente. El modelo analiza la solicitud considerando dicho contexto y genera una respuesta coherente dentro del dominio comercial definido.

El catálogo de productos utilizado para especializar el comportamiento del asistente se descarga desde una fuente externa durante la primera ejecución del sistema y permanece almacenado en memoria flash, evitando descargas repetidas. Durante cada consulta, el catálogo es leído y convertido en texto estructurado que se incorpora como contexto al modelo de lenguaje, permitiendo generar respuestas basadas en información real del comercio. Esta etapa representa el núcleo cognitivo del asistente, donde se produce la interpretación semántica y la generación de respuestas inteligentes.

Configuración de APIs. El ESP32-S3 utiliza APIs proporcionadas por OpenAI para delegar tareas de procesamiento avanzado que no pueden ejecutarse localmente debido a las limitaciones de hardware propias de los sistemas embebidos.

El acceso a estos servicios se realiza mediante una clave de autenticación (API Key) vinculada a una cuenta personal con saldo previamente cargado. OpenAI opera bajo un modelo de facturación por uso, en el cual el costo de cada solicitud depende del volumen de datos procesados, específicamente:

- La cantidad de audio procesado (transcripción),
- El número de tokens generados y procesados (procesamiento conversacional),
- La duración del audio sintetizado (síntesis de voz).

Si el saldo disponible se agota, las solicitudes devuelven errores de autenticación o facturación y el sistema deja de operar correctamente. Por razones de seguridad, la clave de autenticación debe mantenerse confidencial y no debe incluirse directamente en el código fuente ni exponerse en repositorios públicos, ya que su uso genera costo económico. En un entorno de producción, esta clave debería gestionarse mediante variables de entorno o mecanismos seguros de almacenamiento.

API de Reconocimiento de Voz. Corresponde al sistema Whisper de OpenAI, cuya función consiste en recibir un archivo de audio y convertirlo en texto mediante modelos entrenados para identificar lenguaje hablado en múltiples idiomas. El ESP32-S3 envía el archivo grabado mediante una solicitud HTTP de tipo multipart/form-data y recibe como respuesta un objeto JSON con el texto reconocido.

API de Procesamiento Conversacional. Este servicio permite interactuar con modelos de lenguaje capaces de comprender contexto y generar respuestas en lenguaje natural. En el sistema desarrollado se emplea el modelo gpt-4o-mini, al cual el microcontrolador envía conjuntamente las instrucciones de comportamiento del asistente, la información del catálogo y la consulta del usuario. El modelo procesa estos elementos y genera una respuesta textual contextualizada.

API de Síntesis de Voz. Esta API convierte texto en audio mediante técnicas de síntesis de voz (Text-to-Speech). El sistema envía la respuesta textual generada y recibe como resultado un archivo de audio en formato WAV producido por el modelo tts-1, cerrando el ciclo conversacional con una respuesta audible.

Personalización del Modelo de Lenguaje. Con el objetivo de adaptar el comportamiento del modelo al entorno del comercio minorista, se implementó un proceso de personalización mediante el uso de un prompt de sistema. Para las pruebas experimentales, se definió la identidad del asistente como MIA (Microprocesador–Inteligencia–Artificial), concebida como asistente virtual de un supermercado, lo que permite simular escenarios reales de asistencia al cliente en un entorno de compras.

El prompt utilizado establece reglas de comportamiento, personalidad conversacional y limitaciones de respuesta:

"Eres Mia, asistente virtual de un supermercado. Eres amable y atenta. Solo puedes responder usando información del catálogo proporcionado. Si un producto o categoría no existe en el catálogo: - indica que no está disponible. - No inventes ubicaciones, precios ni productos. Si la pregunta no está relacionada con compras o productos: - responde amablemente que solo puedes ayudar con productos del supermercado. Nunca respondas preguntas de cultura general, chistes o explicaciones académicas. Los horarios de apertura y cierre son todos los días de 8 horas a 21 horas";

La identidad y las restricciones del asistente pueden modificarse fácilmente para adaptarse a distintos tipos de comercio, lo que demuestra la flexibilidad del sistema desarrollado. La especialización se refuerza mediante la incorporación dinámica del catálogo de inventario almacenado localmente, el cual constituye la fuente de conocimiento utilizada por el modelo para responder consultas sobre productos disponibles, recomendaciones y asistencia durante el proceso de compra. De esta manera, el modelo general de lenguaje se transforma en un asistente especializado capaz de operar dentro de un dominio comercial específico.

Etapa 5: Síntesis de Voz (Text-to-Speech)

La respuesta textual generada es enviada a un servicio de síntesis de voz, el cual la transforma en una señal de audio que simula voz humana. El archivo resultante es recibido por el ESP32-S3 y almacenado como tts.wav en su memoria interna, lo que permite su reproducción inmediata sin requerir transmisión continua de datos.

Etapa 6: Reproducción de Respuesta

Finalmente, el sistema reconfigura el periférico I2S en modo transmisión para enviar la señal digital hacia el amplificador MAX98357A y el altavoz integrado. Antes de la reproducción, el software analiza la cabecera del archivo WAV para sincronizar automáticamente la frecuencia de muestreo del hardware. Durante la emisión se aplica un ajuste digital de volumen junto con un limitador que previene la saturación o distorsión sonora. Con esta etapa se completa el ciclo de interacción voz–procesamiento–respuesta del asistente inteligente.

Flujo de Datos

Desde el punto de vista arquitectónico, el sistema implementado sigue un flujo de datos secuencial que describe cómo la información generada por el usuario es capturada, procesada y transformada en una respuesta audible. El software se organiza mediante funciones independientes que

gestionan cada etapa del procesamiento, permitiendo separar las tareas de conectividad, captura de audio, procesamiento en la nube y reproducción de la respuesta, como se observa en la figura 10.

1. El usuario realiza una consulta por voz. La señal de audio es capturada por el módulo INMP441 mediante la interfaz I2S y almacenada en memoria flash por la función recordAudio().
2. El dispositivo establece conexión con la red mediante la función connectWiFi(), habilitando la comunicación con los servicios externos.
3. El audio capturado es enviado a la API mediante una solicitud HTTP autenticada. La función transcribeAudio() gestiona el envío del archivo y la recepción de la transcripción generada por el servidor.
4. El texto resultante es procesado por el modelo conversacional a través de la función askMIA(), que envía la consulta al servicio de inteligencia artificial y recibe la respuesta generada.
5. El texto de respuesta es convertido en audio mediante la función generateTTS(), que solicita al servicio de síntesis de voz la generación del archivo de audio correspondiente.
6. Finalmente, el audio recibido es reproducido a través del altavoz mediante la función playWav(), completando el ciclo de interacción con el usuario.

De este modo, el sistema transforma progresivamente la información a lo largo del proceso: audio → texto → respuesta textual → audio, permitiendo una interacción conversacional natural entre el usuario y el asistente inteligente.

Figura 6

Captura de pantalla del loop con todas las etapas del programa

```
void loop() {

  vTaskDelay(10 / portTICK_PERIOD_MS);
  static bool lastTouched = false;
  int touchValue = touchRead(TOUCH_PIN);
  bool isTouched = (touchValue > TOUCH_THRESHOLD);

  // --- Si se activa el sistema ---
  if (!lastTouched && isTouched) {

    etapaActual = "TRABAJANDO";

    // --- ETAPA 1: ACTIVACIÓN ---
    startStage("ETAPA 1: ACTIVACIÓN DEL ASISTENTE");

    vTaskDelay(50 / portTICK_PERIOD_MS);
    setupI2S_RX();
    vTaskDelay(100 / portTICK_PERIOD_MS);

    // --- ETAPA 2: CAPTURA DE AUDIO ---
    recordAudio();
    i2s_driver_uninstall(I2S_PORT);
    vTaskDelay(50 / portTICK_PERIOD_MS);

    // --- ETAPA 3: TRANSCRIPCIÓN ---
    String question = transcribeAudio();

    // --- Si hay texto ---
    if (question.length() > 2) {
      Serial.println("Pregunta: " + question);

      // --- ETAPA 4: IA ---
      String response = askMIA(question);

      // --- ETAPA 5: Síntesis de voz ---
      generateTTS(response);
      setupI2S_TX();
      vTaskDelay(100 / portTICK_PERIOD_MS);

      // --- ETAPA 6: REPRODUCCIÓN ---
      playWav(TTS_FILE, response);
      i2s_driver_uninstall(I2S_PORT);
      vTaskDelay(50 / portTICK_PERIOD_MS);
    }

    // --- Volver a la espera
    Serial.println("Sistema en espera...");
    etapaActual = "NADA";
    vTaskDelay(500 / portTICK_PERIOD_MS);
  }
  if (!isTouched && etapaActual != "TRABAJANDO") {
    mostrarEspera();
  }
  lastTouched = isTouched;
}
```

Nota. Elaboración propia.

Fase IV: Evaluación y pruebas técnicas

La fase de evaluación comprendió cuatro pruebas diseñadas para verificar el desempeño, la estabilidad y el cumplimiento de los requisitos establecidos durante el diseño.

La prueba de sensibilidad del micrófono evaluó la capacidad del sensor INMP441 para captar señales de voz a distintas distancias y niveles de intensidad sonora. Se midió la amplitud RMS de la señal capturada y la precisión de la transcripción resultante a 10 cm y 30 cm de distancia, en ambiente silencioso, con el objetivo de determinar el rango de funcionamiento óptimo del micrófono.

Tabla 4

Pruebas de sensibilidad del micrófono

Nº	Distancia (cm)	Voz	Ambiente	Fs (Hz)	RMS	Transcripción	Resultado	Observaciones
1	10	Normal	Silencioso	16000	0.042	"¿Dónde está el arroz?"	Correcta	Señal estable
2	30	Baja	Silencioso	16000	0.015	"donde está el arroz"	Parcial	Baja amplitud

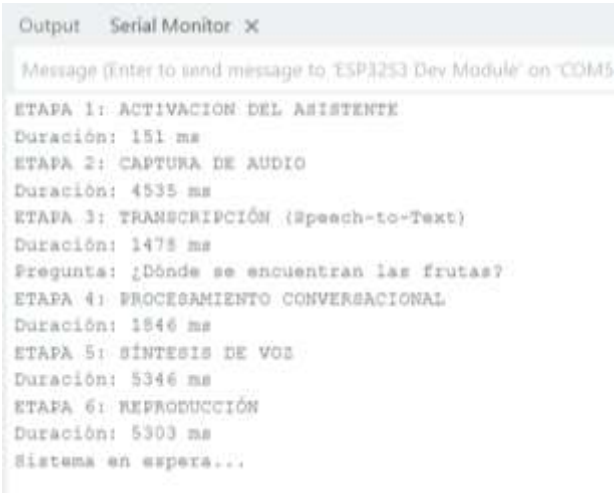
Nota. Elaboración propia.

La prueba de éxito de transcripción consistió en 20 consultas realizadas bajo condiciones de silencio y ruido moderado, comparando el texto esperado con el texto generado por el sistema. Los resultados se clasificaron en tres categorías: correctos (coincidencia semántica completa, aunque faltasen tildes o signos de puntuación menores), parciales (error menor que no alteró completamente el significado) e incorrectos (cambio significativo de sentido o pérdida de palabras clave).

La prueba de latencia midió el tiempo individual de cada una de las seis etapas del ciclo de procesamiento en 10 iteraciones consecutivas, con el fin de identificar los cuellos de botella del sistema y evaluar si el tiempo total de respuesta era adecuado para una interacción fluida en el punto de venta.

Figura 7

Ejemplo de prueba de latencia



Nota. Elaboración propia.

La prueba de coherencia evaluó 20 respuestas generadas por MIA ante consultas de cuatro categorías: disponibilidad de productos, ubicación, precio y recomendaciones, más un grupo de preguntas deliberadamente fuera del dominio comercial (cultura general, chistes, explicaciones académicas) para

evaluar la robustez del sistema. Cada respuesta fue calificada mediante tres criterios: precisión (grado en que la información proporcionada es correcta), relevancia (nivel de relación entre la respuesta y el contexto del sistema) y concordancia (coherencia lógica entre la pregunta y la respuesta generada), cada uno en escala de 1 a 3 puntos, para una puntuación máxima de 9 por respuesta y 180 puntos en total.

Resultados

La prueba de sensibilidad del micrófono mostró que, a 10 cm de distancia en ambiente silencioso, el sistema registró una amplitud RMS de 0,042 con transcripción correcta; a 30 cm, la amplitud descendió a 0,015, produciendo transcripciones parciales por baja señal. La distancia de uso óptima se sitúa entre 10 y 20 cm del usuario.

La prueba de éxito de transcripción sobre 20 consultas arrojó 14 correctas (70%), 4 parciales (20%) y 2 incorrectas (10%), para una precisión global del 90%. Los errores se concentraron en condiciones de ruido moderado con frases cortas o fonéticamente ambiguas. En ambiente silencioso, la precisión fue del 100%.

La prueba de latencia reveló un tiempo promedio total de 22,8 segundos por ciclo completo, distribuidos así: activación 0,15 s, captura de audio 4,6 s, transcripción 1,9 s, procesamiento conversacional 1,8 s, síntesis de voz 7,5 s y reproducción 6,4 s. El tiempo mínimo registrado fue de 17,0 s y el máximo de 26,4 s. Las etapas de síntesis y reproducción concentraron el 60,7% del tiempo total, condicionadas por los servicios en la nube y la duración de las respuestas generadas.

La prueba de coherencia sobre 20 respuestas obtuvo una puntuación total de 180/180, equivalente a un índice de coherencia global del 100%. En todas las categorías —disponibilidad, ubicación, precio, recomendaciones y preguntas fuera del dominio—, el asistente generó respuestas con puntuación máxima en precisión, relevancia y concordancia. Ante preguntas fuera del dominio comercial (cultura general, chistes, explicaciones académicas), MIA declinó correctamente la consulta e indicó que solo podía asistir con productos del establecimiento.

Tabla 5

Prueba de éxito de transcripción

Nº	Ambiente	Texto esperado	Texto transcrito	Resultado	Observaciones
1	Silencioso	“¿Dónde está el arroz?”	¿Dónde está el arroz?	Correcta	

2	Silencioso	“¿Cuánto cuesta la leche entera?”	¿Cuánto cuesta la leche entera?	Correcta	
3	Ruido moderado	“¿Tienen pan integral?”	¡Tienen pan integral!	Parcial	Ofrece la respuesta correcta
4	Ruido moderado	“¿Dónde se encuentran las frutas?	¿Dónde se encuentran las fritas?	Parcial	Ofrece la respuesta correcta
5	Ruido moderado	“¿Cuál es la capital de Nigeria?”	¿Cuál es la capital de Nigeria?	Correcta	
6	Silencioso	“¿Dónde están los huevos?”	¿Dónde están los huevos?	Correcta	
7	Silencioso	“¿Tienen aceite de oliva?”	Tienen aceite de oliva	Correcta	
8	Silencioso	“¿Cuál es el precio del azúcar?”	¡Otra vez el pez sin el azúcar!	Incorrecta	Ofrece respuesta incorrecta
9	Ruido moderado	“Busco pasta para lasaña.”	Busco pasta para lasaña	Correcta	
10	Ruido moderado	“¿Dónde encuentro bebidas sin azúcar?”	donde encuentro bebidas sin azúcar	Parcial	

11	Ruido moderado	“Estoy buscando algo económico para preparar una cena.”	Estoy buscando algo económico para preparar una cena.	Correcta	
12	Ruido moderado	“¿Me puedes recomendar algo para el desayuno que no tenga lactosa?”	¿Me puedes recomendar algo para el desayuno que no tenga lactosa?	Correcta	
13	Ruido moderado	“Necesito ingredientes para hacer una ensalada.”	Necesito ingredientes para hacer una ensalada.	Correcta	
14	Ruido moderado	“¿Tienen pan sin gluten?”	¿Tienen pan sin gluten?	Correcta	
15	Ruido moderado	“¿Hay ofertas hoy?”	Adiós gente soy...	Incorrecta	
16	Ruido moderado	“¿Hay ofertas hoy?”	¿Hay oferta hoy?	Parcial	Se repitió la pregunta anterior que transcribió incorrectamente
17	Ruido moderado	“¿Tienen pollo empanizado?”	¿Tienen pollo empanizado?	Correcta	

18	Ruido moderado	"¿Tienen jamón serrano?"	¿Tienen jamón serrano?	Correcta
19	Ruido moderado	"¿Quién inventó el Internet?"	¿Quién inventó el internet?	Correcta
20	Ruido moderado	"¿Qué hora es?"	¿Qué hora es?	Correcta

Nota. Elaboración propia.

Total de pruebas: 20

Correctas: 14 (70%)

Parciales: 4 (20%)

Incorrectas: 2 (10%)

Precisión global: 90%

Correcta → Coincidencia semántica completa (aunque falten tildes).

Parcial → Error menor que no altera completamente el significado.

Incorrecta → Cambio significativo de sentido o pérdida de palabras clave.

Tabla 6

Tiempos individuales por fase (ms)

Nº	Activación	Captura	Transcripción	Procesamiento	Síntesis	Reproducción	Total
1	151	4535	1478	1846	5346	5303	18659
2	152	4655	2303	2452	7394	8614	25570
3	152	4681	1751	1745	9230	8364	25923
4	152	4551	1676	2254	5620	3175	17428

5	151	4507	1687	1332	6742	6153	20572
6	151	4704	2418	1742	8806	7665	25486
7	152	4525	2088	1535	9064	9028	26392
8	151	4536	1557	2155	8678	7916	24993
9	151	4589	2243	1325	5067	3652	17027
10	151	4676	1739	1636	9251	8428	25881

Nota. Elaboración propia.

Tabla 7

Promedios de latencia

Fase	Promedio (ms)	Promedio (s)
Activación	151	0.15 s
Captura	4596	4.6 s
Transcripción	1894	1.9 s
Procesamiento	1802	1.8 s
Síntesis	7520	7.5 s
Reproducción	6430	6.4 s
TOTAL	22793	22.8 s

Nota. Elaboración propia.

Tabla 8

Resultados prueba de coherencia.

Nº	Categoría	Pregunta	Respuesta MIA	Precisión (1-3)	Relevancia (1-3)	Concordancia (1-3)	Global (1-9)	Observaciones
1	Disponibilidad	¿Tienen arroz blanco de 1 kg?	¡Sí, tenemos arroz blanco de 1 kg de la marca La Campiña a sólo 2!50! Lo puedes encontrar en el pasillo 3.	3	3	3	9	Correcto
2	Disponibilidad	¿Hay leche sin lactosa disponible?	Lo siento, pero no tengo información sobre leche sin lactosa en el catálogo. Pero tenemos leche entera y desnatada disponibles. Si necesitas más información sobre otros productos, estaré encantada de ayudarte.	3	3	3	9	Correcto
3	Disponibilidad	¿Tienen aceite de oliva extra virgen marca La Española?	Lo siento, no tenemos aceite de oliva extra virgen de la marca "La Española" en nuestro catálogo, pero tenemos el aceite de	3	3	3	9	Correcto

			oliva Carbonell a un excelente precio de 6.50. Si deseas, aquí estoy para ayudarte a encontrar otra cosa.					
4	Disponibilidad	¿Está disponible el pan integral hoy?	Lo siento, pero no tengo pan integral en el catálogo. Pero no te preocupes, tengo otras opciones como el pan de molde Bimbo, que es delicioso. ¡Dame un toque si lo necesitas!	3	3	3	9	Correcto
5	Ubicación	¿Dónde está el arroz?	El arroz está en el pasillo 3. ¡Asegúrate de llevar suficiente para tus platos favoritos!	3	3	3	9	Correcto
6	Ubicación	¿En qué pasillo encuentro las bebidas gaseosas?	Las gaseosas se encuentran en el pasillo 9. ¡Elige tu favorito y a disfrutar!	3	3	3	9	Correcto
7	Ubicación	¿Dónde puedo encontrar productos sin gluten?	Lo siento, pero no tengo información sobre productos sin gluten en nuestro catálogo. Puedo ayudarle con otros productos disponibles en el supermercado.	3	3	3	9	Correcto

8	Ubicación	¿Dónde están los productos de limpieza?	Lo siento, pero no tengo información sobre productos de limpieza. ¿Te gustaría saber sobre algún otro tipo de producto?	3	3	3	9	Correcto
9	Precio	¿Cuánto cuesta la leche entera?	La leche entera cuesta 1.20 euros por litro. ¡Perfecta para un buen café!	3	3	3	9	Correcto
10	Precio	¿Cuál es el precio del arroz de 1 kg?	El arroz de un kilo cuesta 2.50. ¡Perfecto para un buen plato de arroz!	3	3	3	9	Correcto
11	Precio	¿Cuánto vale el detergente marca Ariel?	Lo siento, pero no tengo información sobre detergentes. Pero si necesitas algo más, ¡estoy aquí para ayudarte!	3	3	3	9	Correcto
12	Precio	¿Cuál es el producto más barato en la categoría de pastas?	El producto más barato en la categoría de pastas es la Pasta Spaghetti de 500g de Barilla, que cuesta 1.80. ¡Perfecto para una cena rápida y deliciosa!	3	3	3	9	Correcto

13	Recomendaciones y asesoramientos	¿Qué me recomiendas para preparar una pasta rápida?	Te recomiendo la Pasta Spaghetti de Barilla, que está a solo 1.80, y un buen tomate frito de Orlando a 1.70. ¡Listo para un festín rápido!	3	3	3	9	Correcto
14	Recomendaciones y asesoramientos	¿Qué producto es mejor para lavar ropa delicada?	Lo siento, pero solo puedo ayudar con productos del supermercado. Si necesitas información sobre nuestra selección de productos, estaré encantada de ayudarte.	3	3	3	9	Correcto.
15	Recomendaciones y asesoramientos	¿Qué alternativa saludable hay para el azúcar?	Lo siento, pero solo puedo ayudar con productos del supermercado. Si necesitas información sobre algún producto específico, estaré encantada de ayudarte.	3	3	3	9	Correcto.
16	Recomendaciones y asesoramientos	¿Qué puedo comprar para una	Para una cena ligera, te recomendaría una ensalada con tomates frescos, zanahorias y un poco de pollo. ¡Y de	3	3	3	9	Correcto

		cena ligera?	postre, un yogur natural para endulzar la noche!					
17	Fuera de contexto	¿Quién ganó el mundial de 2014?	¡Hola! Me temo que en el supermercado solo manejo productos, pero si necesitas algo para ver el próximo mundial, ¡aquí estoy para ayudarte!	3	3	3	9	Correcto
18	Fuera de contexto	¿Cuál es la capital de Francia?	Lo siento, pero solo puedo ayudar con productos del supermercado. Si necesitas información sobre compras, estaré encantada de ayudarte.	3	3	3	9	Correcto
19	Fuera de contexto	Cuéntame un chiste.	Lo siento, pero solo puedo ayudar con productos del supermercado. Si necesitas información sobre compras, estaré encantada de ayudarte.	3	3	3	9	Correcto
20	Fuera de contexto	¿Puedes explicarme qué es la física cuántica?	Lo siento, pero solo puedo ayudar con productos del supermercado. Si necesitas información sobre compras, estaré	3	3	3	9	Correcto

	encantada de ayudarte.

Nota. Elaboración propia.

Discusión

El índice de coherencia del 100% valida el enfoque de recuperación contextual ligera como mecanismo de especialización del asistente. Al incorporar el catálogo de inventario directamente como contexto en cada solicitud al modelo, el sistema operó exclusivamente sobre información real del comercio sin requerir reentrenamiento ni infraestructura semántica compleja. Este hallazgo es consistente con lo reportado por Lewis et al. (2021), quienes demostraron que la incorporación dinámica de información contextual en modelos de lenguaje reduce significativamente los errores de alucinación y mejora la pertinencia de las respuestas.

La precisión de transcripción del 90% resulta adecuada para el contexto comercial propuesto, donde la mayoría de las consultas son estructuralmente simples y predecibles. Los errores registrados (concentrados en ruido moderado con enunciados cortos) sugieren que el rendimiento puede mejorarse con técnicas de cancelación de ruido o incrementando la duración mínima de grabación.

La latencia promedio de 22,8 segundos representa la limitación operativa más relevante. Las etapas de síntesis de voz (7,5 s) y reproducción (6,4 s) dependen de factores externos como la velocidad de conexión y la carga de los servidores, lo que dificulta su optimización desde el hardware. Reducir la latencia percibida (mediante compresión de audio, modelos TTS más ligeros o conexiones HTTP persistentes) constituye la principal línea de mejora futura. No obstante, en el entorno de un punto de venta físico donde el usuario realiza consultas esporádicas sobre productos concretos, estos tiempos pueden resultar aceptables dependiendo del perfil del usuario.

En términos económicos, la inversión de 46,34 euros del prototipo contrasta favorablemente con las soluciones comerciales de kioscos de autoservicio o sistemas de atención automatizada, que requieren inversiones significativamente mayores.

Conclusiones

El desarrollo del asistente inteligente MIA demostró la viabilidad de integrar inteligencia artificial conversacional especializada en sistemas embebidos de bajo costo orientados al entorno comercial. La estrategia de recuperación contextual ligera (que incorpora el catálogo de inventario como contexto dinámico en cada consulta al modelo de lenguaje) permitió alcanzar un 100% de coherencia en las

respuestas sin requerir infraestructura semántica compleja, reentrenamiento del modelo ni grandes inversiones en hardware.

La arquitectura modular propuesta (ESP32-S3, INMP441, MAX98357A y GC9A01) demostró ser funcionalmente completa para el caso de uso propuesto, con una inversión total de 46,34 euros. La precisión de transcripción del 90% confirmó la adecuación del micrófono INMP441 en condiciones normales de uso, con degradación esperada en entornos de ruido moderado.

Las limitaciones identificadas (dependencia de conectividad a internet, restricciones de memoria del microcontrolador y latencia promedio de 22,8 s condicionada por los servicios en la nube) no comprometen la funcionalidad general del prototipo. La línea de mejora más relevante reside en la reducción de la latencia de síntesis de voz, etapa que concentra el mayor tiempo del ciclo.

Este trabajo aporta una referencia práctica sobre la integración de IA conversacional con sistemas embebidos de bajo costo en el sector comercial, alineándose con las iniciativas de transformación digital del contexto nacional y contribuyendo a la línea de investigación de innovación de procesos industriales y productos tecnológicos de la Universidad Yacambú.

Referencias

Arias, F. (2012). El proyecto de investigación. Introducción a la Metodología Científica. (6ª ed.). Editorial Episteme. Caracas. Venezuela.

Benito, Á. (2019). Desarrollo de aplicaciones para IoT con el módulo ESP32 [Trabajo de grado, Universidad de Alcalá Escuela Politécnica Superior]. Repositorio institucional. https://ebuah.uah.es/dspace/bitstream/handle/10017/35420/TFG_Benito_Herranz_2019.pdf

Evans, D. (2011). *The Internet of Things: How the Next Evolution of the Internet Is Changing Everything*. Cisco Internet Business Solutions Group. https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

IBM. (2024). *What Is an API (Application Programming Interface)?* IBM Cloud Topics. Recuperado de https://www.ibm.com/think/topics/api

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., y Kiela, D. (2021). R*etrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. https://arxiv.org/abs/2005.11401

Maier, A., Sharp, A., & Vagapov, Y. (2017). Comparative analysis and practical implementation of the ESP32 microcontroller module for the internet of things. In *2017 7th International Conference on*

Internet Technologies and Applications (ITA) (pp. 143-148). IEEE. https://doi.org/10.1109/ITECHA.2017.8101926

Russell, S., y Norvig, P. (2021). *Artificial Intelligence*: *A Modern Approach* (4.a ed.). Pearson. http://lib.ysu.am/disciplines_bk/efdd4d1d4c2087fe1cbe03d9ced67f34.pdf

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., y Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* (6.a ed.). Pearson.

Vázquez García, A. (2024). Chatbots basados en IA generativa para optimizar el servicio al cliente en el contexto del mercado español [Tesis doctoral, Universidad de Alcalá]. Biblioteca Digital Universidad de Alcalá. https://ebuah.uah.es/dspace/handle/10017/63000

Wolf, W. (2017). *Computers as Components: Principles of Embedded Computing System Design* (3.a ed.). Morgan Kaufmann.