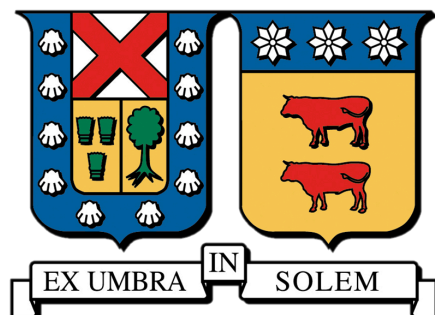


**UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA**  
**DEPARTAMENTO DE ELECTRÓNICA**  
**VALPARAÍSO – CHILE**



# **Diseño y Desarrollo de Plataforma de Apreciación Artística para Personas con Discapacidad Visual**

**Alexey Nikolay Mitjaew Hupat**

**Memoria de Titulación para optar al título de Ingeniero Civil Telemático**

**Profesor Guía:**

Patricio Olivares Roncagliolo

**Profesor Correferente:**

Nicolás Torres Rudloff

**Diciembre 2025**

## Agradecimientos

A mi familia,  
mi perro Elvis  
y el buen café.

## Resumen

Este documento expone el desarrollo de una plataforma accesible de apreciación artística para público con discapacidad visual. La propuesta consiste en transformar obras en representaciones sonoras multifacéticas, que integran descripciones narrativas detalladas, contexto histórico-artístico y atmósferas auditivas generadas mediante modelos multimodales de lenguaje, sistemas de síntesis de voz de alta y modelos de generación sonora basada en texto.

La solución expuesta propone facilitar el acceso al arte mediante una experiencia adaptada, con el objetivo de ofrecer alternativas de apreciación sensorial para personas con discapacidad visual.

**Palabras clave:** *Accesibilidad / Arte inclusivo / Inteligencia artificial / Sonificación*

## **Lista de Figuras**

Figura 1 Multimodal LLMs .....	9
Figura 2 Tacotron 2 Architecture .....	10
Figura 3 AudioLDM .....	11
Figura 4 Arquitectura General de Sistema .....	12
Figura 5 Flujo de Navegación Frontend .....	14

## **Lista de Tablas**

# Contenidos

Introducción .....	5
Objetivo General .....	6
Objetivos Particulares .....	6
Estructura .....	6
Marco Teórico .....	7
Modelos de Lenguaje Multimodales .....	7
Modelos Generativos de Audio .....	7
Estándares de Accesibilidad .....	8
Estado del Arte .....	9
Modelos de Lenguaje Multimodales .....	9
Modelos de Texto a Voz .....	10
Modelos Generativos de Ambiente Sonoro .....	11
Desarrollo de la Plataforma .....	12
API Gateway .....	13
Narrador de Contexto .....	13
Generador de Audio Descriptivo .....	13
Generador de Sonidos Ambientales .....	13
Catálogo de Obras/Imagenes .....	14
Frontend .....	14
Limitaciones .....	14
Resultados .....	15
Conclusiones .....	15
Bibliografía .....	15

## Introducción

El acceso equitativo al arte constituye un componente esencial del desarrollo cultural y social, sin embargo, las personas con discapacidad visual continúan enfrentando barreras significativas para disfrutar de obras visuales en condiciones comparables al resto del público. Aunque las tecnologías de apoyo han avanzado, aún existe una brecha entre la experiencia estética que ofrecen los museos y plataformas digitales tradicionales y las necesidades sensoriales de quienes no pueden percibir elementos visuales directamente.

En este contexto, la inteligencia artificial emerge como una oportunidad para reimaginar la forma en que se transmite el contenido artístico, permitiendo crear experiencias inmersivas que integren narrativa, sonido y contextualización cultural. Este proyecto desarrolla una plataforma que transforma obras de arte en representaciones sonoras enriquecidas, combinando modelos multimodales capaces de interpretar imágenes, sistemas de síntesis de voz de alta naturalidad y tecnologías generativas orientadas a la creación de paisajes sonoros coherentes con la obra original.

La presente investigación se enmarca en los principios de accesibilidad universal y diseño inclusivo, procurando que la experiencia resultante no solo sea funcional, sino también significativa desde una perspectiva estética. La plataforma busca responder a la necesidad de ofrecer alternativas sensoriales que amplíen el acceso al patrimonio artístico y cultural, fortaleciendo la inclusión mediante herramientas tecnológicas avanzadas.

## Objetivo General

Mejorar la experiencia estética y promover la accesibilidad universal a las obras de arte para personas con discapacidad visual mediante el diseño e implementación de una plataforma accesible que utilice inteligencia artificial para transformar automáticamente imágenes de obras artísticas en paisajes sonoros narrativos, contextuales y ambientales.

## Objetivos Particulares

- **Proveer un catálogo curado de obras de dominio público**, basado en fuentes autoritativas y verificadas, que permita a los usuarios acceder a contenido confiable, culturalmente riguroso y legalmente seguro.
- **Generar descripciones auditivas de alta calidad** utilizando modelos de lenguaje multimodales capaces de interpretar intuitivamente los elementos visuales presentes en cada obra, garantizando una representación objetiva y comprensible de su contenido.
- **Crear narraciones históricas y biográficas** fundamentadas en fuentes reconocidas, con el fin de contextualizar la obra, su autor(a) y su relevancia artística dentro de un marco cultural accesible.
- **Producir ambientes sonoros inmersivos** que recreen atmósferas coherentes con la escena o época sugerida por la obra, mediante modelos generativos capaces de transformar imágenes en audio evocativo y sensorialmente enriquecido.
- **Diseñar una interfaz centrada en accesibilidad**, que permita la navegación autónoma de personas en situación de discapacidad visual, incorporando criterios y normativas vigentes en materia de accesibilidad universal y diseño inclusivo.

## Estructura

El presente análisis incluye:

- *Revisión del estado del arte sobre modelos generativos y estándares de accesibilidad.*
- *Evaluación del marco teórico utilizado para el desarrollo de la plataforma.*
- *Descripción de la plataforma desarrollada a nivel de componentes.*
- *Revisión de resultados obtenidos y conclusiones al respecto.*

## **Marco Teórico**

### **Modelos de Lenguaje Multimodales**

Los modelos de lenguaje (LLMs, por sus siglas en inglés) representan sistemas computacionales diseñados para procesar, comprender y generar texto en lenguaje natural mediante arquitecturas basadas en aprendizaje profundo. Estos modelos, entrenados en vastos corpus textuales, capturan patrones lingüísticos, semánticos y contextuales, permitiendo tareas como la traducción automática, la generación de respuestas coherentes o el análisis de sentimientos. Su evolución ha dado paso a los modelos de lenguaje multimodales, los cuales extienden estas capacidades al integrar múltiples formas de información, como imágenes y audio.

Estos modelos multimodales no solo interpretan contenido visual o auditivo, sino que también generan descripciones detalladas de escenas, objetos o composiciones artísticas, traduciendo elementos visuales en narrativas estructuradas. Además, su capacidad multimodal les permite procesar entradas de voz y participar en flujos conversacionales voz a voz, facilitando interacciones más naturales y accesibles. Esta convergencia de modalidades (texto, imagen y sonido) facilita la creación de sistemas más ricos y adaptables, capaces de ofrecer interpretaciones contextualizadas y personalizadas según las necesidades de percepción del usuario.

### **Modelos Generativos de Audio**

Los modelos generativos de audio representan sistemas capaces de crear sonidos, música y voces artificiales a partir de descripciones semánticas o patrones aprendidos. Estos modelos, basados en arquitecturas de aprendizaje profundo, pueden sintetizar desde elementos literales —como el canto de pájaros, el sonido de una tormenta o pasos sobre grava— hasta composiciones musicales complejas, ajustándose a estilos, géneros o emociones específicas. Su funcionamiento se sustenta en la interpretación de instrucciones textuales o parámetros acústicos, permitiendo generar audio coherente y contextualizado

sin necesidad de muestras preexistentes. Esta capacidad no solo amplía las posibilidades creativas en producción musical y diseño sonoro, sino que también facilita la creación de entornos auditivos personalizados, como paisajes sonoros para aplicaciones de accesibilidad o asistentes de voz con tonos y matices más naturales.

## **Estándares de Accesibilidad**

Los estándares de accesibilidad proporcionan el marco normativo que guía el diseño de interfaces inclusivas, asegurando que los contenidos digitales sean utilizables por personas con diversas capacidades sensoriales, motoras o cognitivas. Entre los más relevantes se encuentran las Pautas de Accesibilidad para el Contenido Web (WCAG), que establecen criterios relacionados con la perceptibilidad, operabilidad, comprensibilidad y robustez de los sistemas web. En el contexto chileno, estas directrices se integran mediante la Ley N.º 20.422, que regula la igualdad de oportunidades e incorpora exigencias específicas para tecnologías asistivas. La plataforma desarrollada en este proyecto adopta estos lineamientos para garantizar compatibilidad con lectores de pantalla, navegación por teclado, descripciones alternativas, controles auditivos accesibles y una estructura de interacción que facilite la exploración autónoma.

El cumplimiento de estos estándares es esencial para asegurar que la solución propuesta no solo sea técnicamente avanzada, sino también verdaderamente inclusiva y centrada en las necesidades de la comunidad usuaria.



# Estado del Arte

## Modelos de Lenguaje Multimodales

Durante los últimos años los modelos multimodales de lenguaje han experimentado avances significativos, integrando capacidades de procesamiento de texto, imagen, audio y video. Este progreso ha sido impulsado por arquitecturas unificadas basadas en Transformers que permiten compartir un espacio de representación común entre modalidades. Avances como los *Vision-Language Encoders (VLMs)*, los *Audio-Text aligners* y la *Tokenización Unificada* han permitido integrar imágenes, audio y texto como secuencias compatibles. Además, técnicas como el *cross-attention multimodal*, la *late fusion optimizada* y el *modality routing* han mejorado la coherencia entre modalidades, habilitando tareas complejas de razonamiento y generación conjunta en múltiples formatos.

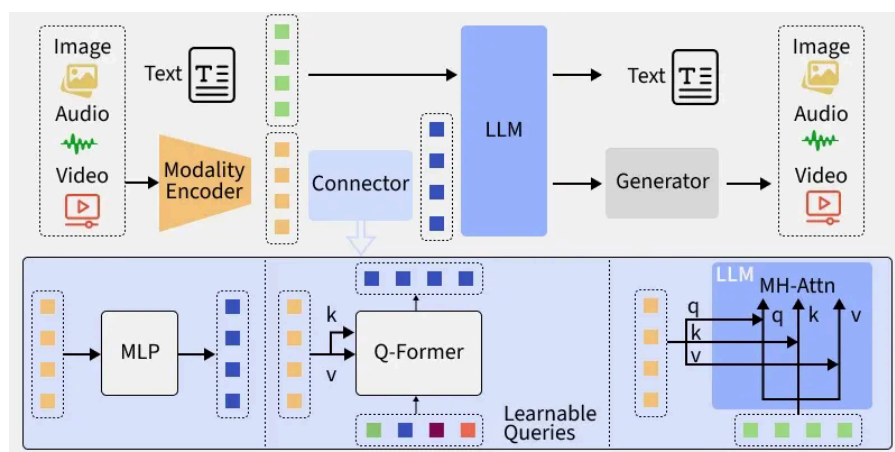


Figura 1: Multimodal LLMs

Entre los avances destacados hasta la fecha (2025) se encuentran:

- **Llama v4 Maverick (Meta):** Modelo multimodal con descripción semántica detallada de imágenes y análisis de escenas. Pesos liberados para investigación.
- **GPT-4o (OpenAI) y Gemini 1.5 (Google):** Modelos con integración nativa de texto, imagen y audio. Sin pesos liberados.
- **Voice-to-Voice (V2V):** Sistemas como Whisper v3 (OpenAI) y SeamlessM4T (Meta) para transcripción y traducción en tiempo real. Voicebox (Meta) genera voz sintética con control de emociones. Whisper v3 tiene pesos liberados; SeamlessM4T y Voicebox no.

## Modelos de Texto a Voz

En los últimos años, arquitecturas como Tacotron 2, VITS, Kokoro TTS y modelos basados en difusión/transformers han ganado relevancia. Estos modelos, generalmente ligeros en cantidad de parámetros, son razonables de autohostear con GPUs de grado consumidor e incluso en CPU en algunos casos. Generan voces naturales con control sobre prosodia, entonación y emociones. Kokoro TTS, de código abierto, destaca por su expresividad y ajuste de voces mediante manipulación del espacio latente, evitando reentrenamiento.

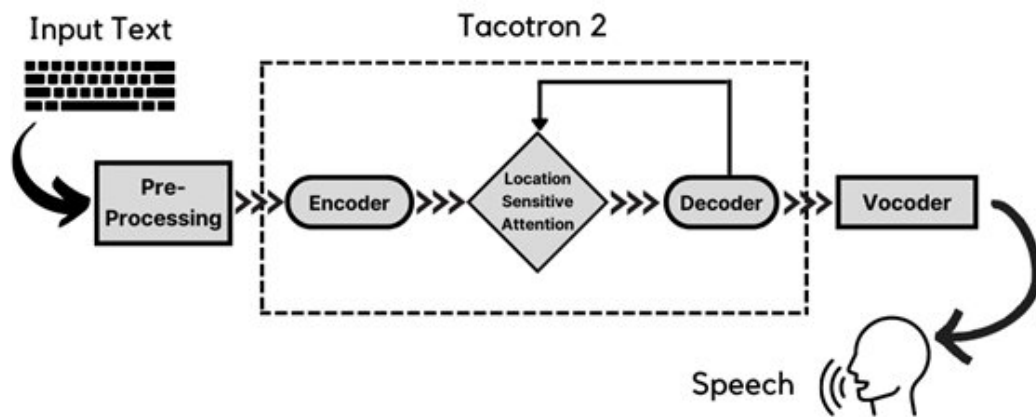


Figura 2: Tacotron 2 Architecture

Además de los modelos de código abierto, proveedores comerciales como Google Cloud, Amazon Polly, Microsoft Azure, ElevenLabs y OpenAI ofrecen soluciones TTS de alta calidad. Estas plataformas incorporan voces multilingües, opciones de personalización avanzada y capacidades de clonación vocal.

No obstante los avances observados, subsisten desafíos técnicos comunes en las alternativas investigadas, como la generación de expresividad controlable, la adaptación a entornos acústicos adversos y la síntesis eficiente de secuencias de larga duración.

## Modelos Generativos de Ambiente Sonoro

La generación de ambientes sonoros de alta calidad sigue en fase experimental, sin soluciones comerciales consolidadas. Modelos como CLAP permiten mapear texto y audio en un espacio semántico compartido, aunque no generan audio directamente.

Entre los modelos evaluados para la plataforma se encuentran:

- **I Hear Your True Colors:** La arquitectura propuesta en este paper combina el uso de un VQVAE, transformers y CLIP. El VQVAE extrae representaciones jerárquicas como secuencias discretas, los transformers las modelan de forma autorregresiva y CLIP alinea el audio con lo visual. Aunque el código es abierto, no existen fuentes públicas de pesos preentrenados. Sin versiones comerciales.
- **AudioLDM:** Utiliza difusión para generar espectrogramas condicionados por texto, incorporando embeddings de CLAP para mejorar la alineación semántica entre descripciones textuales y contenido sonoro. Presenta resultados prometedores en coherencia semántica y diversidad acústica. El modelo preentrenado se puede encontrar con pesos abiertos.

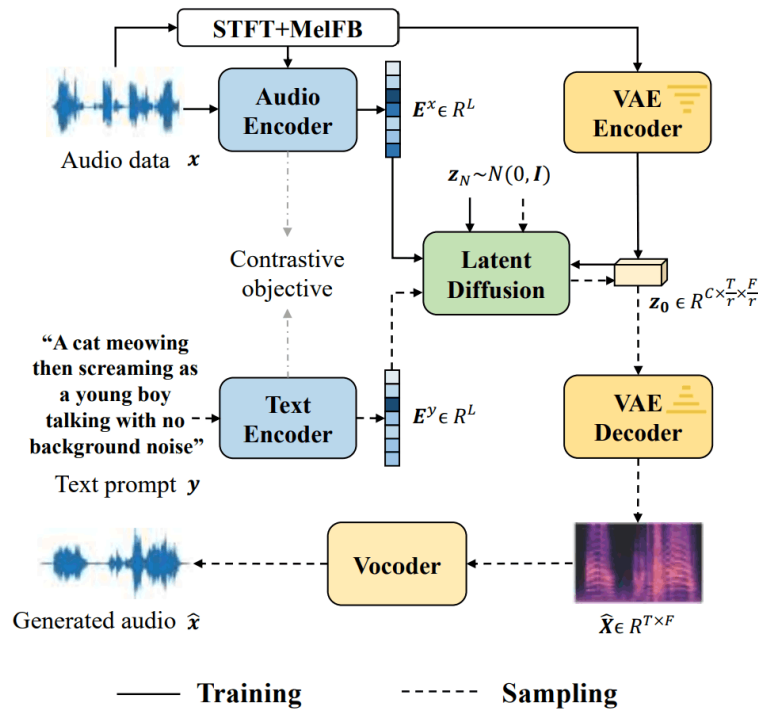


Figura 3: AudioLDM

## Desarrollo de la Plataforma

La arquitectura propuesta sigue un modelo cliente-servidor e incorpora un *API Gateway* como componente central. Este elemento cumple dos funciones principales: por un lado, centraliza las operaciones computacionalmente intensivas (principalmente inferencia de modelos de aprendizaje automático y las solicitudes a APIs externas), y por otro, gestiona un catálogo centralizado de recursos. Además, se desarrolla un *frontend* que aplica principios de accesibilidad web, con especial atención a las necesidades de usuarios con discapacidad visual.

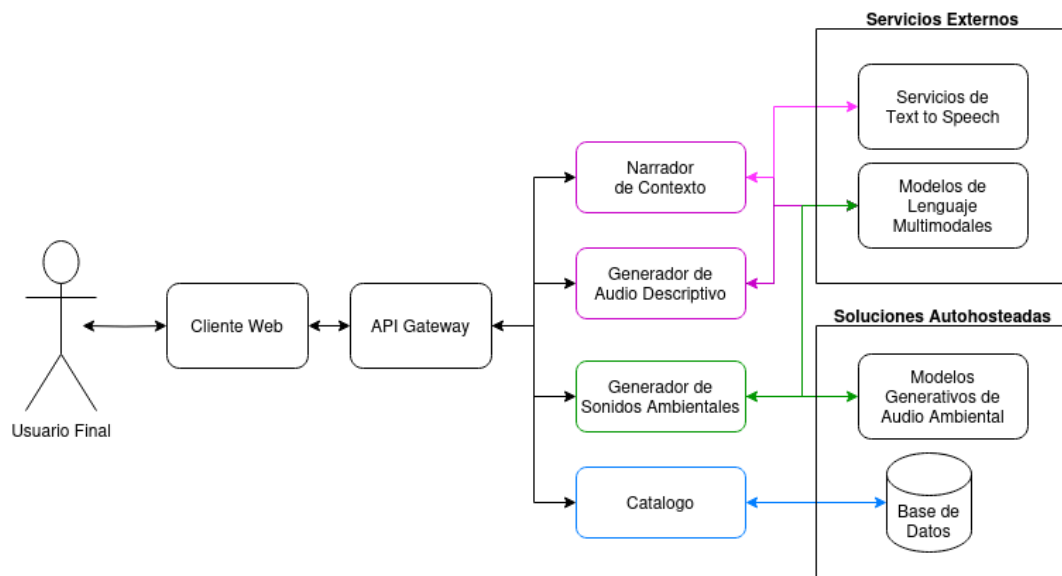


Figura 4: Arquitectura General de Sistema

## **API Gateway**

La API Gateway consiste en un servidor de Django, encargado de controlar 4 servicios:

### **Narrador de Contexto**

Este módulo se encarga de producir descripciones textuales detalladas y estructuradas de los contenidos visuales presentes en las obras. Para ello, emplea modelos de lenguaje multimodales, los cuales procesan tanto elementos gráficos como metadatos asociados mediante prompts especializados. Dichos prompts están diseñados para extraer información semántica relevante, garantizando una representación fiel y contextualizada del material original (**ANEXO PROMPT AQUI**). Posteriormente, las descripciones generadas son convertidas a formato de audio mediante sistemas de síntesis de voz (TTS) (**ANEXO TTS AQUI**), optimizados para ofrecer una experiencia auditiva clara y accesible.

### **Generador de Audio Descriptivo**

Este módulo genera narraciones auditivas que contextualizan históricamente cada obra. Para ello, se recopila información de fuentes autoritativas (en este caso, se procesó manualmente un conjunto de datos con artículos de Wikipedia vinculados a cada pieza) (**ANEXO DATASET AQUI**). A partir de estos datos, se elabora un relato contextualizado y adaptado al caso de uso (**ANEXO PROMPT AQUI**), que posteriormente se convierte en audio mediante modelos de síntesis de voz (TTS) (**ANEXO TTS AQUI**), asegurando una experiencia auditiva coherente y accesible.

### **Generador de Sonidos Ambientales**

Este módulo se encarga de la generación de ambientes sonoros que representan los contenidos literales (no abstractos) de sus imágenes de entrada. La heurística utilizada consiste en:

- Dividir el espacio en cuadrantes (en nuestro caso 9) (**ANEXO**)
- Procesar cada cuadrante con LLMs multimodales, obteniendo salidas estructuradas en JSON (**ANEXO**) representando cada elemento detectado.
- Por cada cuadrante generar una mezcla de sonido integrando todos los elementos detectados en cada respectiva sección. (**ANEXO**)

## Catálogo de Obras/Imagenes

Se elaboró un conjunto de datos en formato JSON que incluye 30 obras artísticas, conteniendo todo el material requerido para el funcionamiento de los sistemas previamente descritos. Este dataset incorpora, además, enlaces a los artículos correspondientes de Wikipedia asociados a cada obra artística. **(ANEXO)**

## Frontend

El frontend se implementó utilizando NextJS para proporcionar acceso al contenido gestionado por la API Gateway. En su diseño, se incorporaron textos alternativos y etiquetas HTML semánticas con el objetivo de garantizar una integración completa con herramientas de asistencia para usuarios con discapacidad visual. Asimismo, se desarrolló una interfaz de usuario que facilita la navegación mediante atajos de teclado, optimizando la accesibilidad y usabilidad del sistema.

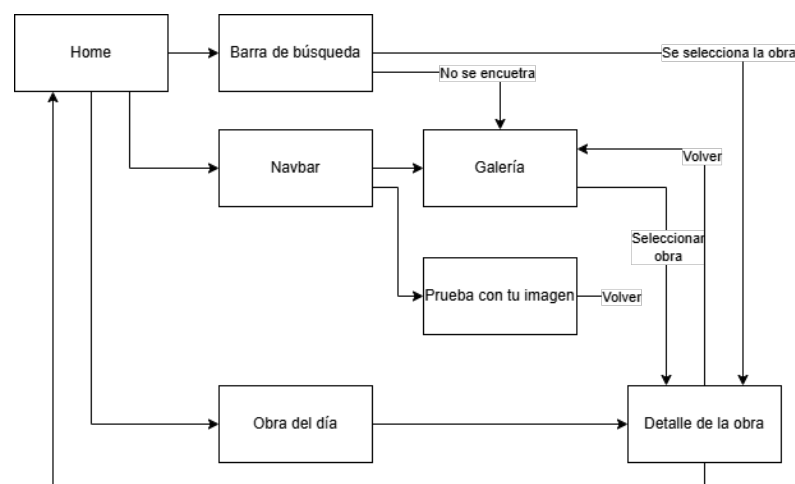


Figura 5: Flujo de Navegación Frontend

## Limitaciones

Por limitaciones computacionales durante el desarrollo, los módulos no se integraron directamente en la API Gateway. En su lugar, se implementó un prototipo funcional en un cuadernillo Jupyter (<https://www.kaggle.com/code/alexeymitjaew/sound-generation-pipeline>) para validar el diseño y flujo de procesamiento. Esto permitió generar y preprocesar los conjuntos de datos necesarios, que fueron incorporados a la base de datos para su uso en la versión final.

## Resultados

## Conclusiones

[1] said [2] said

## Bibliografía

- [1] Ö. Aksin *et al.*, «Effect of immobilization on catalytic characteristics of saturated Pd-N-heterocyclic carbenes in Mizoroki-Heck reactions», *J.~Organomet. Chem.*, vol. 691, n.º 13, pp. 3027-3036, 2006.
- [2] G. Westfahl, «The True Frontier». pp. 55-65.