



uhu.es

**Escuela Técnica Superior de Ingeniería
Universidad de Huelva**

Máster en Ingeniería Informática

Trabajo Fin de Máster

**DeepSea Speech. Transcripción Inteligente de
Radiocomunicaciones Marítimas**

Manuel Cerrejón Naranjo

junio, 2025

Resumen

En el contexto del Procesamiento del Lenguaje Natural (PLN) y el Reconocimiento Automático del Habla (ASR), la transcripción automática de audios cobra una importancia estratégica en sectores donde la comunicación precisa es crítica. Este trabajo se centra en el desarrollo de un sistema de transcripción adaptado específicamente a las comunicaciones por radio en entornos marítimos, con el objetivo de mejorar la seguridad y la eficiencia en la navegación.

El sistema propuesto aborda desafíos únicos del entorno naval, como la presencia de ruidos intensos (viento, oleaje, interferencias), la superposición de voces, el uso de terminología técnica especializada y la diversidad lingüística. Para ello, se ha entrenado y evaluado un modelo basado en técnicas avanzadas de aprendizaje profundo, orientado a mejorar la calidad de las transcripciones mediante procesos de reducción de ruido, segmentación del discurso y normalización del lenguaje.

La metodología incluye la recopilación de audios reales del entorno marítimo, el preprocesamiento intensivo de señales sonoras, y el ajuste de modelos ASR para maximizar métricas como la tasa de error de palabras (WER). El enfoque se orienta a la adaptación contextual del modelo, optimizando su rendimiento en condiciones acústicas adversas y lenguaje especializado.

Los resultados obtenidos evidencian mejoras significativas en la inteligibilidad y precisión de las transcripciones en comparación con modelos genéricos. Este estudio contribuye a la evolución de los sistemas de transcripción en el ámbito náutico, aportando una herramienta potencialmente útil para reforzar la seguridad y trazabilidad de las comunicaciones marítimas.

Palabras clave: Reconocimiento automático del habla, Transcripción de radio marina, PLN, Aprendizaje profundo, Ruido acústico, WER, Normalización del lenguaje.

Abstract

In the field of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR), automatic audio transcription plays a strategic role in sectors where accurate communication is essential. This work focuses on the development of a transcription system specifically tailored to marine radio communications, aiming to enhance safety and efficiency in maritime navigation.

The proposed system addresses the unique challenges of the naval environment, such as intense background noise (wind, waves, radio interference), overlapping voices, specialized technical terminology, and linguistic diversity. To overcome these issues, a deep learning-based model has been trained and evaluated, incorporating advanced techniques for noise reduction, speech segmentation, and language normalization.

The methodology includes the collection of real-world maritime audio recordings, intensive signal preprocessing, and the fine-tuning of ASR models to optimize performance metrics such as Word Error Rate (WER). The approach focuses on contextual adaptation to improve model performance under adverse acoustic conditions and domain-specific language.

The results demonstrate significant improvements in transcription accuracy and intelligibility compared to generic models. This study contributes to the advancement of transcription systems in the nautical domain, offering a potentially valuable tool to strengthen the safety and traceability of maritime communications.

Keywords: Automatic Speech Recognition, Marine Radio Transcription, NLP, Deep Learning, Acoustic Noise, WER, Language Normalization.

Agradecimientos

En primer lugar, me gustaría darles las gracias a mis tutores por la dedicación y enseñanza. A todos los docentes que he tenido a lo largo de mi vida.

- A mis padres, por apoyarme, quererme y guiarme en cada paso que doy.
- A mis tutores, por todo el material proporcionado, por todos los talleres realizados y por la concesión del servidor para la elaboración de la investigación.

Manuel Cerrejón Naranjo

Huelva, 2025

Índice general

Resumen	II
Abstract	III
Agradecimientos	IV
Índice de Figuras	VII
Índice de Tablas	VIII
1. Introducción	1
1.1. Motivación	2
1.2. Objetivos	2
1.3. Competencias	2
1.4. Estructura de la Memoria	3
2. Marco Teórico	5
2.1. Aprendizaje Automático	5
2.2. Aprendizaje profundo aplicado al habla	6
2.2.1. Redes neuronales profundas	6
2.2.2. Función de pérdida	7
2.2.3. Representación acústica	7
2.3. Reconocimiento Automático del Habla (ASR)	8
2.3.1. Pipeline general de un sistema ASR moderno	8
2.3.2. Modelos end-to-end frente a arquitecturas híbridas	8
2.3.3. Retos generales del ASR	9
2.4. Reconocimiento del habla en el entorno marítimo	10
2.4.1. Características acústicas del dominio VHF náutico	10
2.4.2. Problemas típicos de transcripción marina	10
2.4.3. Necesidad de especialización del modelo	11
2.5. Arquitectura Transformer y su aplicación al ASR	11
2.5.1. Estructura del Transformer	11
2.5.2. Ventajas frente a RNN	12
2.6. Whisper: arquitectura, variantes y funcionamiento	13
2.6.1. Estructura encoder-decoder con espectrogramas log-mel	13
2.6.2. Capacidades del modelo	14
2.6.3. Comparativa con otros modelos ASR	14
2.7. Aprendizaje por transferencia en ASR	14
2.7.1. Motivación para el fine-tuning en dominios especializados	15
2.7.2. Limitaciones del fine-tuning completo	15
2.8. Adaptación eficiente con LoRA	16
2.8.1. Fundamento de LoRA: inserción de matrices de bajo rango	16
2.8.2. Reducción de parámetros entrenables y ventajas prácticas	17
2.8.3. Aplicaciones previas y traslado al ASR	17
2.8.4. Justificación del uso de LoRA	18
2.9. Métricas de Evaluación	18

2.9.1. Word Error Rate (WER)	18
2.10. Tecnologías y Recursos Utilizados	19
3. Metodología, Experimentación y Resultados	21
3.1. Descripción General de la Metodología	21
3.2. Descripción del entorno marítimo y características del dataset	21
3.3. Preparación y segmentación de los audios	23
3.3.1. Criterios para la división de ficheros	23
3.3.2. Limpieza y filtrado de segmentos no válidos	24
3.3.3. Estructura final del conjunto segmentado	24
3.4. Generación de los conjuntos de entrenamiento y evaluación	24
3.4.1. Estrategia de partición	25
3.4.2. Estructura del conjunto de datos final	25
3.4.3. Validación de la calidad del conjunto	25
3.5. Adaptación del modelo Whisper con LoRA	26
3.5.1. Modelo base seleccionado y motivación	26
3.5.2. Configuración del entrenamiento con ajuste eficiente	27
3.5.3. Capas congeladas y parámetros ajustados	27
3.5.4. Consideraciones computacionales	28
3.6. Implementación del pipeline de transcripción automática	28
3.6.1. Ejecución del modelo sobre el conjunto de test	28
3.6.2. Generación y almacenamiento de transcripciones	29
3.6.3. Evaluación mediante métrica WER	29
4. Resultados y análisis	31
4.1. Evaluación y Análisis de Resultados	31
4.2. Comparación con baseline	32
4.3. Análisis de errores	34
4.4. Alcance y limitaciones del sistema propuesto	36
5. Conclusiones y Trabajo Futuro	38
5.1. Conclusiones	38
5.2. Trabajo Futuro	39
5.3. Planificación Temporal del Trabajo Realizado	41
Referencias	43
A. Anexos	44
A.1. Aplicación Web para la Visualización y Prueba del Modelo	44

Índice de Figuras

2.1.	Espectrograma del audio marítimo procesado, representado en escala logarítmica de energía en la banda VHF. Esta representación acústica es utilizada por modelos como Whisper como entrada al proceso de transcripción.	9
2.2.	Esquema general de la arquitectura Transformer, con interacción entre bloques de atención y capas feed-forward.	12
2.3.	Mecanismo de adaptación de bajo rango (LoRA): sólo las matrices A y B se actualizan, mientras que los pesos originales se mantienen congelados.	17
2.4.	Tecnologías y recursos utilizados en el desarrollo del trabajo	20
3.1.	Pipeline general del marco de experimentación.	21
4.1.	Comparación visual de las tasas de error WER y WER normalizado entre diferentes configuraciones del modelo Whisper con adaptación LoRA. Se observa una mejora progresiva en las versiones con ajustes óptimos de hiperparámetros.	32

Índice de Tablas

3.1. Estructura del Conjunto de Test tras su evaluación	29
4.1. Resultados de evaluación sobre el conjunto de test con diferentes configuraciones LoRA.	31
4.2. Comparación de rendimiento entre el modelo base Whisper Large (sin adaptación) y las mejores configuraciones LoRA.	33
5.1. Planificación temporal del trabajo realizado.	41

CAPÍTULO 1

Introducción

En los últimos años, el reconocimiento automático del habla (ASR, por sus siglas en inglés) se ha consolidado como una de las tecnologías clave en el ámbito de la inteligencia artificial, con aplicaciones que abarcan desde asistentes virtuales hasta sistemas de accesibilidad y automatización de tareas administrativas. Su objetivo principal es convertir señales acústicas en texto de forma precisa y robusta, incluso en entornos ruidosos o no estructurados [1, 2].

El entorno marítimo plantea un desafío particularmente exigente para los sistemas de transcripción automática. Las comunicaciones por radio VHF, utilizadas entre embarcaciones y estaciones costeras, se ven afectadas por una gran variedad de condiciones acústicas adversas: ruido ambiental de motores y oleaje, interferencias electromagnéticas, superposición de voces, y diversidad de acentos entre los operadores. A esto se suma el uso de una terminología técnica especializada que no siempre está bien representada en los conjuntos de datos generales utilizados para entrenar los modelos ASR [3].

Este Trabajo de Fin de Máster se centra en el desarrollo de un sistema de transcripción automática especializado en radiocomunicaciones marítimas, empleando para ello el modelo Whisper, desarrollado por OpenAI. Whisper es una arquitectura basada en transformers que ha demostrado un rendimiento notable en condiciones de ruido y en tareas multilingües [4]. En este proyecto, el modelo se adapta específicamente al dominio marítimo mediante técnicas de fine-tuning eficientes basadas en LoRA (Low-Rank Adaptation), lo que permite realizar ajustes sin necesidad de grandes recursos computacionales [5].

A diferencia de otros enfoques que incorporan procesos de corrección lingüística posteriores a la transcripción, el presente trabajo no aplica técnicas de procesamiento de lenguaje natural (como puntuación, segmentación textual o normalización terminológica). En su lugar, se centra exclusivamente en mejorar el rendimiento del modelo ASR mediante la especialización del modelo base y la preparación acústica adecuada de los datos.

Este sistema busca facilitar la generación automatizada de registros textuales a partir de comunicaciones por radio, contribuyendo a la mejora de la seguridad marítima, la trazabilidad de los mensajes y la respuesta ante emergencias. La evaluación del rendimiento se realiza mediante la métrica estándar Word Error Rate (WER), analizando los errores más frecuentes e identificando posibles causas ligadas al ruido, dialecto o características acústicas del entorno.

1.1. Motivación

La navegación marítima depende en gran medida de la eficacia de las comunicaciones por radio, especialmente en situaciones críticas donde una interpretación errónea puede tener consecuencias graves. Sin embargo, la transcripción manual de estas comunicaciones resulta inviable en tiempo real y costosa a posteriori.

La posibilidad de contar con un sistema de transcripción automática fiable, entrenado específicamente en las condiciones propias del entorno marítimo, puede aportar beneficios significativos: desde el registro estructurado de comunicaciones operativas hasta la detección rápida de llamadas de auxilio. Este trabajo surge de la necesidad de adaptar los modelos ASR existentes a este dominio tan particular, optimizando su rendimiento sin recurrir a infraestructuras computacionales de gran escala.

1.2. Objetivos

El objetivo principal de este Trabajo de Fin de Máster es diseñar y desarrollar un sistema de transcripción automática robusto para audios marítimos, basado en el modelo Whisper, adaptado mediante fine-tuning eficiente con LoRA.

- Desarrollar un sistema de transcripción basado en Whisper, especializado en radio-comunicaciones marítimas.
- Implementar un proceso de adaptación del modelo mediante técnicas ligeras (LoRA), optimizadas para su ejecución en entornos con GPU limitadas.
- Aplicar técnicas de preprocessamiento acústico (limpieza, segmentación por longitud, validación de audio) para asegurar la calidad del conjunto de datos.
- Evaluar el sistema mediante la métrica WER, analizando su comportamiento bajo distintas condiciones de ruido y variabilidad dialectal.
- Identificar errores frecuentes en las transcripciones y discutir posibles mejoras futuras orientadas a la robustez del sistema.

Estos objetivos buscan no solo la creación de un sistema eficiente y adaptado a las necesidades del ámbito marítimo, sino también contribuir a la seguridad y eficacia operativa en la navegación mediante tecnologías de inteligencia artificial.

1.3. Competencias

Durante el desarrollo de este Trabajo de Fin de Máster, se han adquirido y fortalecido diversas competencias clave en el ámbito de la Ingeniería Informática, con especial énfasis en

el reconocimiento automático del habla (ASR), procesamiento del lenguaje natural (PLN) y aprendizaje profundo. Entre las competencias desarrolladas destacan las siguientes:

- **CE7-C** – Capacidad para conocer y desarrollar técnicas de aprendizaje automático y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo sistemas de reconocimiento y transcripción automática de audio: Se han aplicado técnicas avanzadas de aprendizaje profundo para la mejora de la precisión y robustez del sistema de transcripción automática en entornos marítimos.
- **CB2** – Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean competencias para la elaboración y defensa de argumentos, así como para la resolución de problemas dentro de su área de estudio: Se ha demostrado la capacidad para implementar y optimizar modelos ASR y técnicas de PLN adaptadas a las condiciones específicas del ámbito marítimo, justificando las decisiones técnicas tomadas durante el proyecto.
- **CB3** – Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan reflexión sobre temas relevantes de índole social, científica o ética: Mediante la recopilación y análisis de datos acústicos y lingüísticos de comunicaciones marítimas, se ha evaluado el impacto de factores ambientales y sociales en la precisión de la transcripción, considerando la importancia de la seguridad y la comunicación en situaciones críticas.
- **CT2** – Desarrollo de una actitud crítica en relación con la capacidad de análisis y síntesis: Se ha fomentado un enfoque crítico para evaluar los resultados obtenidos, identificando limitaciones del sistema y proponiendo posibles mejoras y futuras líneas de investigación en el campo del reconocimiento automático del habla en entornos complejos.

1.4. Estructura de la Memoria

El documento está organizado para presentar de manera clara y ordenada el desarrollo del sistema de transcripción automática de audios marítimos y los resultados obtenidos a lo largo del proyecto.

- En el [Capítulo 2](#) se realiza una revisión teórica sobre los fundamentos del reconocimiento automático del habla (ASR) aplicado a entornos acústicos complejos, con especial énfasis en las particularidades de las comunicaciones marítimas por radio. Se presentan los conceptos clave, técnicas y modelos relevantes para el desarrollo del sistema.
- El [Capítulo 3](#) detalla la metodología seguida en el proyecto, incluyendo la recopilación y tratamiento de datos acústicos y lingüísticos, el diseño e implementación del modelo

de transcripción, y las estrategias de mejora para la calidad del texto generado. Se describen además las técnicas de evaluación utilizadas para medir el desempeño del sistema.

- En el Capítulo 4 se exponen los resultados obtenidos, se realiza un análisis crítico de los mismos y se discuten las principales conclusiones del trabajo. También se proponen posibles líneas de investigación y mejoras futuras para el desarrollo de sistemas de transcripción en entornos marítimos.
- Finalmente, se incluye un anexo que contiene el enlace al repositorio de GitHub, donde se aloja el código fuente desarrollado durante el proyecto, facilitando su acceso y posible reutilización.

Este documento proporciona una visión completa del desarrollo, evaluación y conclusiones del sistema de transcripción automática adaptado al ámbito marítimo, con el objetivo de contribuir a la mejora de la comunicación y seguridad en las comunicaciones por radio en entornos navales.

CAPÍTULO 2

Marco Teórico

El diseño de un sistema de transcripción automática especializado en comunicaciones marítimas exige una base sólida en varios campos interconectados: el aprendizaje automático y profundo, las arquitecturas de tipo transformer, las técnicas de adaptación de modelos preentrenados, y las métricas específicas para evaluar la calidad de las transcripciones. En este capítulo se desarrollan los principios teóricos fundamentales que sustentan este trabajo, con especial atención a los retos que plantea el entorno acústico náutico.

2.1. Aprendizaje Automático

El aprendizaje automático (Machine Learning, ML) es una rama de la inteligencia artificial que permite a los sistemas extraer patrones a partir de los datos, sin necesidad de ser programados de forma explícita para cada tarea. En lugar de seguir reglas codificadas manualmente, un modelo de aprendizaje automático infiere una función que relaciona entradas con salidas basándose en ejemplos previos [6].

En el contexto del reconocimiento automático del habla (ASR), esta disciplina permite desarrollar sistemas capaces de aprender a transformar una señal acústica (una onda de audio) en una representación textual, entrenándose con pares de datos audio-texto. A medida que se expone a más ejemplos, el modelo mejora su capacidad de generalización y su precisión en tareas no vistas previamente.

Existen tres grandes categorías de aprendizaje automático, cada una con características particulares:

- **Aprendizaje Supervisado** [7]. Es el más común y es el paradigma adoptado en este trabajo. Consiste en entrenar un modelo con un conjunto de datos etiquetados, en los que para cada entrada (en este caso, una señal de audio), se conoce su salida deseada (la transcripción correspondiente). El modelo ajusta sus parámetros internos para minimizar el error entre la salida generada y la real. Esta aproximación es fundamental en ASR, donde se utilizan grandes corpus de grabaciones junto con su transcripción manual para entrenar redes neuronales profundas.
- **Aprendizaje no Supervisado** [8]. Se emplea cuando no se dispone de etiquetas. El sistema intenta encontrar estructuras ocultas en los datos, como agrupaciones (clus-

tering) o representaciones latentes. En ASR, puede aplicarse para preentrenar representaciones acústicas sin transcripciones, como ocurre en modelos auto-supervisados (p. ej., Wav2Vec 2.0).

- **Aprendizaje por Refuerzo [9]**. Es menos habitual en ASR, pero útil en sistemas interactivos. Consiste en que un agente aprende a tomar decisiones secuenciales en un entorno, maximizando una señal de recompensa. En tareas relacionadas con el habla, como la síntesis o el diálogo, puede servir para optimizar interacciones mediante retroalimentación.

En este trabajo se emplea aprendizaje supervisado, en forma de ajuste fino (fine-tuning) de un modelo preentrenado (Whisper), utilizando un conjunto de audios náuticos reales y sus respectivas transcripciones. Este enfoque permite adaptar un sistema generalista al dominio marítimo sin necesidad de entrenar desde cero, aprovechando así el conocimiento adquirido previamente en grandes corpus.

El uso de modelos preentrenados con aprendizaje automático ha sido determinante en el avance de tareas de transcripción, clasificación de texto y comprensión del lenguaje, superando con creces las capacidades de los sistemas basados en reglas o plantillas estáticas.

2.2. Aprendizaje profundo aplicado al habla

El aprendizaje profundo (Deep Learning) representa una extensión del aprendizaje automático basada en redes neuronales con múltiples capas, conocidas como redes neuronales profundas (DNN, por sus siglas en inglés). Estas arquitecturas han revolucionado el campo del reconocimiento automático del habla (ASR), permitiendo el modelado directo de relaciones complejas entre el audio y el texto, sin necesidad de una ingeniería manual de características acústicas intermedias.

En lugar de extraer manualmente atributos fonéticos o lingüísticos, los modelos profundos son capaces de aprender representaciones jerárquicas directamente desde las formas de onda, detectando patrones que se corresponden con fonemas, sílabas o palabras, en función de la profundidad y arquitectura del modelo [10].

2.2.1. Redes neuronales profundas

Las DNN tradicionales constan de una serie de capas densas que conectan todas las neuronas de una capa con las de la siguiente. Aunque fueron útiles en las primeras etapas del ASR profundo, su capacidad para modelar secuencias temporales era limitada.

Para superar esta limitación, se introdujeron modelos recurrentes como las RNN, que mantienen información de pasos anteriores, lo que las hace adecuadas para datos secuencia-

les. Su principal mejora, las LSTM (Long Short-Term Memory), introducen mecanismos de puertas que permiten conservar o olvidar información en el tiempo, siendo fundamentales en tareas de transcripción de secuencias largas con dependencias a largo plazo.

Sin embargo, estas arquitecturas presentan limitaciones en cuanto a la paralelización y manejo de dependencias muy largas, lo que motivó el desarrollo de los transformers, que han sustituido progresivamente a RNN y LSTM en los sistemas ASR más avanzados.

2.2.2. Función de pérdida

Durante el entrenamiento, el modelo intenta minimizar una función de pérdida que cuantifica el error entre su predicción y la transcripción correcta. En modelos ASR modernos se utilizan principalmente:

- Cross-Entropy Loss, usada en arquitecturas encoder-decoder autoregresivas como Whisper. Calcula la distancia entre la distribución de salida del modelo y la distribución real, penalizando predicciones incorrectas en cada paso.
- Connectionist Temporal Classification (CTC), una función de pérdida diseñada para problemas de alineación implícita entre entrada y salida, muy útil cuando no se conoce la correspondencia temporal exacta entre el audio y las palabras.

La optimización del modelo se realiza mediante algoritmos como Adam, una extensión del descenso de gradiente estocástico que ajusta dinámicamente las tasas de aprendizaje para cada parámetro, mejorando la estabilidad y rapidez de la convergencia.

2.2.3. Representación acústica

Antes de ser introducido en un modelo profundo, el audio debe transformarse en una representación que preserve las características fonéticas y prosódicas más relevantes. Las más empleadas son:

- **MFCC (Mel-Frequency Cepstral Coefficients)**: capturan la envolvente espectral en una escala perceptual basada en la audición humana. Aunque han sido muy utilizadas, están siendo reemplazadas por representaciones más ricas.
- **Espectrogramas log-mel**: son la entrada utilizada por Whisper. Consisten en mapas de energía acústica a lo largo del tiempo en distintas bandas de frecuencia, escaladas según la escala de Mel y transformadas logarítmicamente.

Estas representaciones permiten que el modelo aprenda a diferenciar patrones fonéticos incluso en condiciones de ruido o distorsión, algo fundamental en entornos como el marítimo.

2.3. Reconocimiento Automático del Habla (ASR)

El Reconocimiento Automático del Habla (ASR, por sus siglas en inglés) es la tecnología que permite transformar una señal de voz en una secuencia de texto, mediante la combinación de métodos estadísticos, redes neuronales y procesamiento acústico. Su relevancia ha crecido exponencialmente en los últimos años gracias a los avances en aprendizaje profundo y a la disponibilidad de grandes corpus de datos etiquetados.

En su concepción moderna, un sistema ASR opera como un modelo end-to-end, es decir, como una única red neuronal entrenada para mapear directamente la señal de audio a la transcripción, sin necesidad de separar en módulos distintos el modelo acústico, el modelo de lenguaje o el alineador fonético.

2.3.1. Pipeline general de un sistema ASR moderno

Los sistemas actuales de ASR siguen típicamente una arquitectura basada en tres fases principales:

- **Extracción de características acústicas:** la señal de audio en crudo se convierte en una representación densa, generalmente un espectrograma log-mel, que actúa como entrada para la red neuronal. Esta representación preserva la información temporal y frecuencial esencial para la identificación fonética.
- **Modelo acústico:** procesa las características extraídas para predecir secuencias de unidades lingüísticas (tokens, fonemas o subpalabras). Esta función es desempeñada por arquitecturas neuronales profundas que capturan tanto las relaciones temporales como semánticas del discurso.
- **Decodificador (decoder):** genera la secuencia textual más probable, basándose en las predicciones anteriores y, en algunos casos, en un modelo de lenguaje interno que mejora la fluidez del texto producido.

Esta estructura puede verse como una cadena de procesamiento audio → representación → predicción de texto.

2.3.2. Modelos end-to-end frente a arquitecturas híbridas

En los enfoques clásicos de ASR, se utilizaban sistemas híbridos con componentes separados:

- Un modelo acústico basado en HMM-GMM.
- Un diccionario fonético que mapeaba fonemas a palabras.

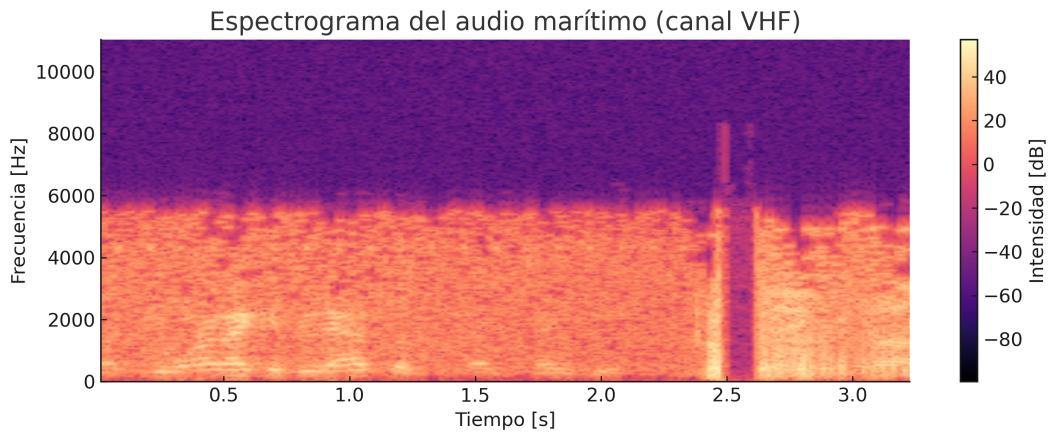


Figura 2.1: Espectrograma del audio marítimo procesado, representado en escala logarítmica de energía en la banda VHF. Esta representación acústica es utilizada por modelos como Whisper como entrada al proceso de transcripción.

- Un modelo de lenguaje basado en n-gramas.

Estos sistemas requerían una alineación explícita entre el audio y la transcripción, así como procesos de entrenamiento complejos y propensos a errores acumulativos.

En contraste, los modelos end-to-end como Whisper emplean arquitecturas encoder-decoder con atención, que permiten aprender toda la transformación desde el audio hasta el texto de forma conjunta, optimizando directamente una única función de pérdida. Esto mejora la robustez y reduce la necesidad de intervención manual.

2.3.3. Retos generales del ASR

A pesar de los avances, el ASR sigue enfrentando numerosos desafíos, especialmente en entornos reales:

- Ruido de fondo: motores, viento o conversaciones simultáneas afectan la inteligibilidad del audio.
- Superposición de voces (overlapping speech): dificulta la segmentación y asignación correcta del texto.
- Acentos y dialectos: los modelos tienden a funcionar mejor en acentos estándares si no han sido entrenados en variedad lingüística.
- Acentos y dialectos: los modelos tienden a funcionar mejor en acentos estándares si no han sido entrenados en variedad lingüística.
- Velocidad y prosodia: el ritmo de habla, pausas atípicas o énfasis inusuales pueden alterar la predicción del modelo.

Estos problemas se agravan en escenarios como el marítimo, donde las condiciones acústicas suelen ser especialmente adversas y cambiantes.

2.4. Reconocimiento del habla en el entorno marítimo

El entorno marítimo constituye un caso de uso particularmente desafiante para los sistemas de reconocimiento automático del habla, debido a las características acústicas adversas y a la naturaleza específica de las comunicaciones por radio. Las transmisiones de voz en este contexto se realizan mayoritariamente mediante radiofrecuencia en las bandas VHF (Very High Frequency)[11], sujetas a degradación, fluctuaciones y ruido impulsivo. Además, las comunicaciones siguen protocolos y léxico técnicos que no suelen estar bien representados en los corpus genéricos usados para el entrenamiento de modelos ASR.

2.4.1. Características acústicas del dominio VHF náutico

Las señales de audio captadas en comunicaciones marítimas presentan una serie de propiedades que dificultan su transcripción automática:

- Ruido de fondo constante y no estacionario, generado por motores diésel, viento, oleaje, lluvia o equipos electrónicos a bordo.
- Interferencias de canal debidas al uso compartido de frecuencias VHF, a menudo congestionadas en zonas de alta actividad marítima.
- Distorsión y pérdidas causadas por modulación analógica, mala calidad del micrófono o transmisión desde largas distancias.
- Limitaciones del ancho de banda, que recortan frecuencias importantes para la inteligibilidad del habla (especialmente consonantes fricativas y oclusivas).

2.4.2. Problemas típicos de transcripción marina

A nivel práctico, los siguientes fenómenos se observan con frecuencia en la transcripción de comunicaciones de radio marítima:

- Superposición de voces (talk-over): varias embarcaciones intentan comunicarse simultáneamente, generando interferencias y fragmentación del discurso.
- Silencios abruptos o cortes debidos a pérdida de señal o intervención del squelch de la radio.

- Pronunciación variable: los hablantes suelen utilizar un lenguaje comprimido, códigos fonéticos (ej. ".alfa", "bravo") y pronunciaciones no estándar.
- Uso de jerga técnica: términos como "Mayday", "Pan-Pan", "estribor", "proa", o códigos de canal y posición geográfica, que pueden no estar presentes en los tokens del modelo base.

2.4.3. Necesidad de especialización del modelo

Dadas estas dificultades, los modelos ASR generalistas, entrenados sobre grandes corpus multilingües y de dominio abierto, no logran ofrecer una transcripción fiable en contextos como el marítimo sin un proceso de adaptación. Las causas principales son:

- La disparidad entre el dominio de entrenamiento y el de aplicación real, tanto en términos acústicos como lingüísticos.
- La ausencia de ejemplos representativos del tipo de ruido y terminología utilizados en radio náutica durante el preentrenamiento del modelo.
- El modelo de lenguaje implícito en sistemas end-to-end como Whisper tiende a favorecer estructuras sintácticas comunes y no elocuencias técnicas o abreviadas.

Por ello, se hace necesario un proceso de especialización mediante técnicas de fine-tuning sobre datos reales del dominio. En este trabajo, esta especialización se aborda mediante el uso de LoRA para ajustar parcialmente el modelo Whisper a las características del entorno marítimo, sin incurrir en los costes de entrenamiento total.

2.5. Arquitectura Transformer y su aplicación al ASR

La arquitectura Transformer ha supuesto una revolución en el tratamiento de secuencias, desplazando progresivamente a las redes recurrentes (RNN) en tareas de procesamiento del lenguaje natural y reconocimiento del habla. Introducida por Vaswani et al. en 2017, esta arquitectura se basa en mecanismos de atención que permiten procesar en paralelo toda la secuencia de entrada, capturando relaciones de largo alcance entre elementos [12].

En tareas de ASR, los transformers han demostrado ser altamente efectivos para modelar dependencias temporales complejas, ofreciendo mejores resultados que las arquitecturas tradicionales tanto en precisión como en eficiencia computacional.

2.5.1. Estructura del Transformer

La arquitectura Whisper implementa un esquema encoder-decoder:

- **Encoder:** transforma la entrada (en este caso, el spectrograma del audio) en una secuencia de representaciones latentes. Cada capa del encoder incluye mecanismos de atención multi-cabeza y capas feed-forward completamente conectadas, junto con normalización y residual connections.
- **Decoder:** genera la salida (tokens de texto), utilizando atención sobre las salidas anteriores (auto-atención) y sobre las representaciones del encoder (atención cruzada). Esto permite que el modelo considere tanto el contexto acústico como el contexto lingüístico al generar cada token.

El componente clave es la atención escalada por productos punto (scaled dot-product attention), que permite al modelo ponderar dinámicamente qué partes de la entrada son más relevantes para cada salida. Esto lo hace mucho más eficiente que las RNN, especialmente para secuencias largas o con estructuras no lineales.

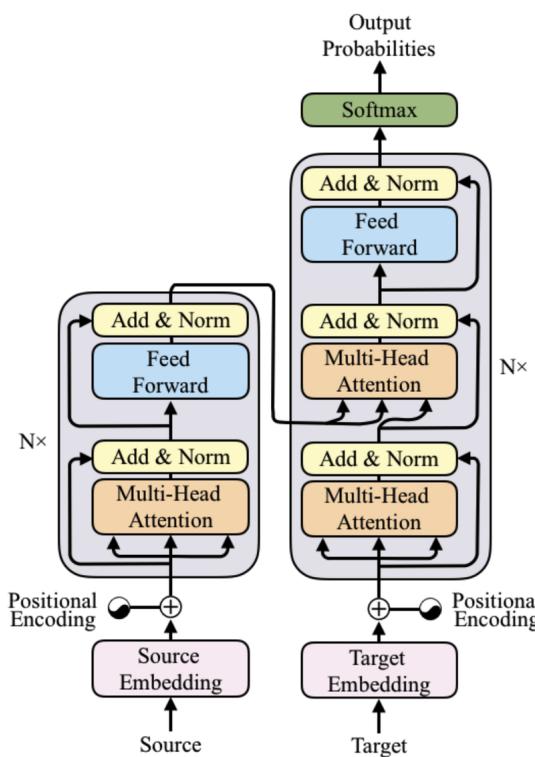


Figura 2.2: Esquema general de la arquitectura Transformer, con interacción entre bloques de atención y capas feed-forward.

2.5.2. Ventajas frente a RNN

El uso de atención en lugar de recurrencia conlleva varias ventajas:

- Paralelización total durante el entrenamiento, al procesarse todos los elementos de la secuencia simultáneamente.

- Acceso directo al contexto completo: cada token puede atender a cualquier otro, sin depender de una memoria interna secuencial.
- Mejor manejo de dependencias largas, comunes en mensajes de voz con estructuras complejas o pausas extendidas.
- Mayor estabilidad numérica y menor riesgo de desvanecimiento/explosión del gradiente.

Estas características hacen que los transformers sean especialmente adecuados para el reconocimiento del habla en condiciones reales, como las que se presentan en entornos marítimos.

2.6. Whisper: arquitectura, variantes y funcionamiento

Whisper es un modelo de reconocimiento automático del habla desarrollado por OpenAI, basado en una arquitectura Transformer encoder-decoder. Fue entrenado con más de 680.000 horas de datos multilingües y multitarea, incluyendo audio con ruido, traducción y transcripción, lo que le proporciona una robustez excepcional en contextos reales y en múltiples idiomas.

A diferencia de modelos anteriores, Whisper fue diseñado desde el principio para abordar las limitaciones de los ASR tradicionales, como la baja tolerancia al ruido, la incapacidad de manejar múltiples idiomas y la dependencia de diccionarios fonéticos o modelos de lenguaje explícitos.

2.6.1. Estructura encoder-decoder con espectrogramas log-mel

Whisper recibe como entrada un espectrograma log-mel del audio, que condensa la información acústica en una representación de energía en el tiempo y frecuencia perceptual. Esta entrada es procesada por:

- Un encoder Transformer que genera una representación latente del audio, modelando dependencias temporales y características acústicas de largo alcance.
- Un decoder Transformer autoregresivo que genera tokens textuales, basándose en los embeddings del encoder y los tokens generados previamente.

Esta arquitectura permite que el modelo transcriba directamente desde el audio sin necesidad de módulos adicionales como modelos de lenguaje externos o diccionarios fonéticos.

2.6.2. Capacidades del modelo

Whisper se diferencia de otros sistemas ASR por integrar múltiples funcionalidades de forma conjunta:

- **Multilingüismo:** es capaz de transcribir y traducir en varios idiomas sin necesidad de ajustar su arquitectura.
- **Detección automática de idioma:** en la fase inicial del decoder, predice el idioma de la señal de entrada.
- **Robustez frente al ruido y la distorsión:** al haber sido entrenado con datos degradados, puede generalizar bien a condiciones acústicas adversas.
- **Segmentación implícita:** realiza la transcripción en bloques, permitiendo procesar señales largas dividiéndolas automáticamente.

Estas características lo convierten en una excelente opción para tareas en condiciones reales, como la transcripción de audio marítimo con ruido, acentos variables y condiciones no controladas.

2.6.3. Comparativa con otros modelos ASR

Whisper ofrece ventajas importantes frente a modelos alternativos como:

- **Wav2Vec 2.0 (Meta):** aunque también presenta buen rendimiento, está más orientado a entrenamientos supervisados y requiere datasets adaptados al dominio si se desea especializar.
- **DeepSpeech (Mozilla):** basado en RNNs, tiene menor capacidad de generalización y ha quedado desfasado frente a arquitecturas tipo Transformer.
- **MMS (Massively Multilingual Speech):** modelo más reciente de Meta AI con cobertura lingüística aún mayor, pero con una interfaz y documentación más limitada que Whisper para tareas personalizadas.

En este trabajo se ha optado por utilizar Whisper Large como modelo base debido a su equilibrio entre precisión, cobertura multilingüe y robustez frente a entornos ruidosos. Además, su integración con bibliotecas como Hugging Face y la disponibilidad de variantes más pequeñas lo hacen idóneo para su ajuste con recursos computacionales limitados.

2.7. Aprendizaje por transferencia en ASR

El aprendizaje por transferencia (Transfer Learning) es una técnica que permite reutilizar el conocimiento aprendido por un modelo en una tarea previa para resolver un nuevo

problema, relacionado pero diferente. En lugar de entrenar un modelo desde cero —lo cual exige grandes volúmenes de datos y recursos computacionales—, se parte de un modelo preentrenado que ya ha captado patrones útiles y se adapta a una nueva tarea mediante un proceso de ajuste fino (fine-tuning).

En el campo del reconocimiento automático del habla, el aprendizaje por transferencia ha demostrado ser especialmente eficaz para adaptar modelos generales a dominios específicos, como el médico, jurídico o, en este caso, el marítimo. La idea es aprovechar las representaciones acústicas y lingüísticas ya aprendidas por el modelo base, y ajustarlas con una cantidad relativamente pequeña de datos específicos del dominio.

2.7.1. Motivación para el fine-tuning en dominios especializados

Los modelos como Whisper han sido entrenados con una gran diversidad de idiomas y condiciones acústicas. No obstante, estos modelos no están optimizados para entornos extremadamente particulares como las comunicaciones por radio VHF en navegación, que presentan características no representadas en los datos de preentrenamiento.

Entre los motivos que justifican el uso del fine-tuning se encuentran:

- **Vocabulario técnico específico:** términos náuticos, siglas, nombres de canales y protocolos.
- **Condiciones acústicas adversas:** ruido no estacionario, interferencias y superposición de hablantes.
- **Estilo de comunicación comprimido:** frases cortas, uso de código fonético y estructuras no gramaticales.

El fine-tuning permite ajustar el modelo a estas condiciones sin perder su conocimiento general del idioma, logrando una mejora significativa en la tasa de error de palabras (WER) en el dominio objetivo.

2.7.2. Limitaciones del fine-tuning completo

Aunque el fine-tuning tradicional ha sido ampliamente utilizado, presenta algunas limitaciones relevantes en entornos con recursos limitados:

- **Alto coste computacional:** ajustar todos los parámetros de un modelo grande como Whisper requiere GPUs de alto rendimiento y mucho tiempo de entrenamiento.
- **Riesgo de sobreajuste:** si el conjunto de datos específico es reducido, el modelo puede perder su capacidad de generalización (catastrophic forgetting).

- **Inviabilidad en entornos ligeros:** plataformas como Google Colab o equipos personales no pueden almacenar ni procesar modelos de gran tamaño durante varias épocas completas.

Para mitigar estas limitaciones, se han desarrollado técnicas de ajuste eficiente, como LoRA (Low-Rank Adaptation), que permiten adaptar grandes modelos con una fracción de los recursos computacionales necesarios. Esta estrategia ha sido la adoptada en el presente trabajo, y se desarrolla en detalle en la siguiente sección.

2.8. Adaptación eficiente con LoRA

El ajuste de modelos de lenguaje de gran tamaño (LLMs) o modelos de reconocimiento del habla como Whisper plantea importantes desafíos computacionales. En un modelo típico, el fine-tuning completo implica actualizar millones —o incluso miles de millones— de parámetros, lo cual requiere hardware especializado y tiempos prolongados de entrenamiento. Además, cuando se dispone de pocos datos, esta estrategia puede ser contraproducente, ya que aumenta el riesgo de sobreajuste y pérdida del conocimiento general preentrenado.

Frente a estas limitaciones, ha surgido una familia de técnicas llamadas Parameter-Efficient Fine-Tuning (PEFT), entre las que destaca LoRA (Low-Rank Adaptation). LoRA permite adaptar modelos grandes modificando únicamente una pequeña fracción de sus parámetros, manteniendo congelada la mayor parte del modelo original.

2.8.1. Fundamento de LoRA: inserción de matrices de bajo rango

LoRA se basa en la observación de que muchas de las actualizaciones que se realizan durante el fine-tuning completo se concentran en subespacios de baja dimensionalidad. En lugar de actualizar directamente los pesos de las capas del modelo, LoRA introduce matrices entrenables adicionales $A \in R^{d \times r}$ y $B \in R^{r \times d}$, donde $r \ll d$, que se insertan de forma paralela a los pesos originales en ciertas capas (típicamente, las capas de atención o proyección lineal).

Durante el entrenamiento, solo se actualizan estas matrices AyB , mientras que los pesos originales del modelo permanecen fijos. De este modo:

- Se reduce drásticamente el número de parámetros entrenables.
- Se mantiene la estabilidad y generalización del modelo base.
- Se permite un entrenamiento eficiente incluso en entornos con GPU limitada.

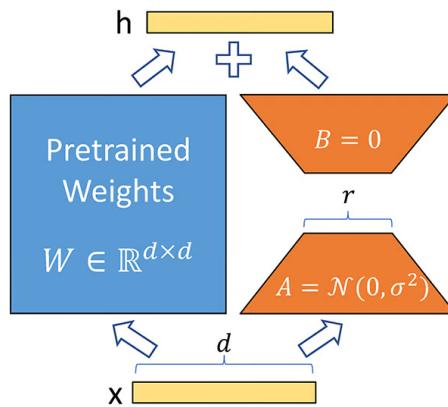


Figura 2.3: Mecanismo de adaptación de bajo rango (LoRA): sólo las matrices A y B se actualizan, mientras que los pesos originales se mantienen congelados.

2.8.2. Reducción de parámetros entrenables y ventajas prácticas

En la práctica, LoRA permite adaptar modelos como Whisper con apenas un 0.1–1 % del total de parámetros modificados, lo cual implica:

- **Menor uso de memoria:** los gradientes se calculan solo para las matrices insertadas.
- **Entrenamiento más rápido:** menos parámetros implican menos cómputo por iteración.
- **Posibilidad de trabajar en entornos ligeros:** como Google Colab, sin necesidad de múltiples GPUs o servidores dedicados.

Además, LoRA puede integrarse fácilmente en frameworks modernos como Hugging Face PEFT, que simplifican su implementación y permiten combinarla con técnicas como quantization (int8/int4) o entrenamiento con precisión mixta.

2.8.3. Aplicaciones previas y traslado al ASR

Aunque LoRA se popularizó inicialmente en el ajuste eficiente de modelos de lenguaje como BERT, GPT o LLaMA, su extensión al ámbito del ASR es una tendencia emergente. En el presente trabajo se demuestra su viabilidad para adaptar Whisper, un modelo originalmente entrenado para tareas generales, al dominio náutico.

El ajuste con LoRA permite incorporar patrones propios de las comunicaciones marítimas (vocabulario técnico, patrones fonéticos, cadencias de habla) sin reentrenar el modelo completo ni comprometer su capacidad de transcripción general.

2.8.4. Justificación del uso de LoRA

Dado el tamaño del modelo Whisper Large (1.5B parámetros) y la limitación de recursos, el uso de LoRA ha sido determinante para:

- Reducir el coste computacional del ajuste del modelo.
- Acelerar el entrenamiento y permitir múltiples experimentaciones.
- Mantener la robustez general del modelo, adaptándolo progresivamente al entorno marítimo con un conjunto de datos específico.

Esta estrategia ha demostrado ser efectiva, permitiendo un compromiso equilibrado entre rendimiento, coste y especialización del modelo.

2.9. Métricas de Evaluación

La evaluación objetiva de los sistemas de reconocimiento automático del habla (ASR) es esencial para comparar modelos, validar mejoras y cuantificar errores. En este trabajo se ha adoptado como métrica principal el Word Error Rate (WER), ampliamente utilizada en la literatura como estándar de referencia en tareas de transcripción.

El uso de métricas específicas como el WER permite no solo comparar el desempeño global del sistema, sino también identificar patrones de error según el tipo de entrada, el nivel de ruido o la pronunciación.

2.9.1. Word Error Rate (WER)

El WER se calcula comparando la transcripción generada por el sistema con una transcripción de referencia (ground truth), contabilizando las palabras que han sido:

- **Sustituidas:** una palabra incorrecta en lugar de la esperada.
- **Insertadas:** una palabra añadida que no aparece en la referencia.
- **Eliminadas:** una palabra que debería estar pero ha sido omitida.

La fórmula es la siguiente:

$$\text{WER} = \frac{S + D + I}{N} \quad (2.1)$$

donde:

- S es el número de sustituciones
- D el número de eliminaciones

- I el número de inserciones
- N el número total de palabras en la transcripción de referencia.

Un valor de WER más bajo indica una mayor fidelidad entre la transcripción automática y la real. En la práctica, un sistema con un WER inferior al 10 % se considera de alta calidad. No obstante, en entornos ruidosos o con alta variabilidad acústica —como ocurre en las comunicaciones por radio marítima—, se aceptan valores más elevados si la inteligibilidad del mensaje principal se mantiene.

El uso de WER ha sido prioritario en este trabajo por las siguientes razones:

Permite una comparación directa y cuantificable entre diferentes configuraciones del modelo.

Ofrece una medida sensible a los errores que afectan directamente al contenido semántico.

Es sencilla de calcular mediante bibliotecas estándar como jiwer o evaluate de Hugging Face.

Durante la evaluación de los resultados, se ha desglosado la WER por tipo de error, permitiendo así un análisis más detallado del rendimiento del sistema bajo distintas condiciones acústicas y dialectales, lo cual se expondrá en el capítulo de resultados.

2.10. Tecnologías y Recursos Utilizados

El desarrollo de este proyecto se ha llevado a cabo utilizando un conjunto de herramientas ampliamente adoptadas en el ámbito de la inteligencia artificial y el procesamiento de lenguaje, como se muestra en la Figura 2.4. Estas tecnologías han sido seleccionadas por su robustez, comunidad activa y compatibilidad con tareas de transcripción automática y evaluación de modelos.

- **Python 3.10:** Lenguaje de programación principal del proyecto, utilizado para la implementación de scripts de procesamiento, entrenamiento de modelos y análisis de resultados.
- **Jupyter Notebook:** Entorno interactivo para el desarrollo exploratorio, la visualización de resultados y la ejecución modular del código.
- **PyTorch:** Framework de aprendizaje profundo utilizado para construir, entrenar y evaluar los modelos basados en Transformers y redes neuronales.
- **Hugging Face:** Biblioteca clave para cargar y ajustar modelos preentrenados como Whisper.

- **LoRA:** Técnica utilizada para reducir el número de parámetros entrenables durante el fine-tuning, haciendo posible la adaptación del modelo a GPUs de bajo consumo y permitiendo experimentar con menos recursos.
- **Kaggle:** Fuente principal de los datos de entrenamiento y validación. Se han utilizado datasets públicos disponibles en esta plataforma relacionados con audios en lenguaje natural, que han sido adaptados al objetivo del proyecto.
- **Word Error Rate (WER):** Métrica utilizada para evaluar cuantitativamente el rendimiento del modelo de transcripción. Su cálculo se ha implementado mediante bibliotecas estándar y ha sido central en el análisis de resultados.
- **Google Colab:** Plataforma de ejecución en la nube que ha permitido el entrenamiento de modelos con acceso a GPUs de forma gratuita.
- **GPU del laboratorio I2C:** Parte del entrenamiento final de los modelos se realizó en entornos locales del grupo de investigación, usando recursos computacionales propios para mejorar la velocidad de entrenamiento.



Figura 2.4: Tecnologías y recursos utilizados en el desarrollo del trabajo

CAPÍTULO 3

Metodología, Experimentación y Resultados

3.1. Descripción General de la Metodología

Este capítulo describe el proceso seguido para la construcción del sistema de transcripción automática adaptado al entorno marítimo. Se detallan las fuentes de datos empleadas, las fases de preparación del audio, y la estructura final del conjunto de datos utilizado para entrenar y evaluar el modelo. Se hace especial énfasis en las decisiones adoptadas para garantizar la calidad, relevancia y representatividad de los datos en un dominio tan específico como el de las radiocomunicaciones náuticas.

La [Figura 3.1](#) muestra el pipeline del marco de experimentación seguido en este trabajo.

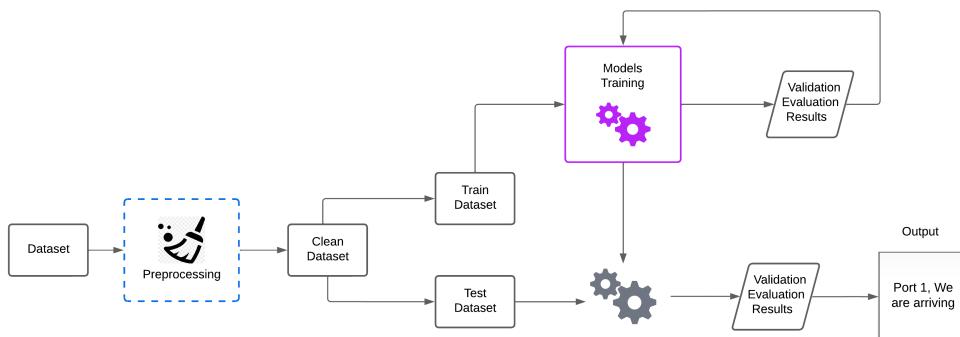


Figura 3.1: Pipeline general del marco de experimentación.

3.2. Descripción del entorno marítimo y características del dataset

La transcripción automática de comunicaciones marítimas plantea una serie de desafíos específicos derivados del entorno en el que se generan los mensajes y del canal por el que se transmiten. Las embarcaciones civiles y comerciales, tanto en entornos portuarios como en navegación fluvial o costera, utilizan comunicaciones por radio VHF (Very High Fre-

quency) para coordinar maniobras, intercambiar información operativa o emitir mensajes de emergencia. Estas comunicaciones son breves, altamente estructuradas, y a menudo se realizan bajo presión temporal y en condiciones acústicas degradadas.

- Presencia de ruido constante y no estacionario, provocado por motores, viento, oleaje, lluvia o interferencias electromagnéticas.
- Canal de transmisión limitado, con pérdidas de calidad derivadas del uso de modulación analógica, variaciones de ganancia o equipos emisores de baja fidelidad.
- Superposición de emisiones, ya que en muchos casos las transmisiones no son full-duplex, lo que da lugar a cortes o solapamientos entre interlocutores.
- Diversidad de acentos y pronunciaciones, dado el carácter internacional del tráfico marítimo.
- Uso de jerga técnica, códigos fonéticos y estructuras lingüísticas comprimidas, propias de las comunicaciones de emergencia o tráfico portuario.

Para abordar estos retos, se ha recurrido a un conjunto de datos real y representativo, extraído del repositorio público de Kaggle titulado Marine Radio Chatter – Bridge 2 Bridge Communication¹. Este dataset incluye grabaciones reales de radio VHF procedentes de comunicaciones entre embarcaciones en zonas portuarias y fluviales de Estados Unidos. Las grabaciones fueron recogidas de forma pasiva y anónima, garantizando el respeto a la privacidad y el cumplimiento de la normativa sobre comunicaciones públicas en banda VHF.

Junto a los audios, se dispone de un dataset complementario proporcionado por el mismo autor, Bridge-to-Bridge Transcript², que contiene transcripciones manuales de los mensajes registrados. Si bien estas transcripciones no están perfectamente alineadas con los audios, constituyen un punto de partida útil para el etiquetado inicial.

El corpus original abarca varias horas de grabaciones en archivos de audio de larga duración, superando los 10 GB de datos en bruto. Sin embargo, para la realización de este trabajo se ha optado por una reducción significativa del conjunto de datos, centrando el análisis y la adaptación del modelo en una selección representativa de segmentos con contenido lingüístico claro y relevancia técnica.

El criterio de selección ha priorizado:

- Calidad acústica suficiente para permitir una segmentación precisa.
- Presencia efectiva de habla (evitando largos intervalos de silencio o ruido).
- Variedad de situaciones comunicativas (llamadas, coordinaciones, advertencias, etc.).

¹<https://www.kaggle.com/datasets/linogova/marine-radio-chatter-bridge-2-bridge-communication>

²<https://www.kaggle.com/datasets/linogova/marine-radio-chatter-to-bridge-transcript>

Tras esta fase de filtrado, se procedió a la segmentación de los audios seleccionados, la cual se detalla en la siguiente sección. El resultado es un corpus curado, que conserva la riqueza fonética, semántica y contextual del entorno náutico, pero adaptado a los límites de procesamiento impuestos por el uso de modelos grandes como Whisper y recursos computacionales moderados.

Este conjunto de datos reducido y reorganizado constituye la base sobre la que se ha realizado el ajuste fino del modelo, así como la evaluación posterior del sistema.

3.3. Preparación y segmentación de los audios

Una vez realizada la selección de archivos relevantes desde el corpus original, fue necesario adaptar los datos para su uso en el proceso de entrenamiento del modelo Whisper. Este paso no se limita a convertir el audio a un formato compatible, sino que implica una reestructuración profunda del contenido: división de archivos, limpieza de muestras, y validación manual del resultado. Dada la naturaleza continua y sin segmentar de los audios originales, estas tareas fueron fundamentales para garantizar que el modelo aprendiera sobre datos estructurados, breves y representativos de interacciones reales.

3.3.1. Criterios para la división de ficheros

Los archivos de audio originales contenían múltiples interacciones consecutivas, grabadas de forma continua, sin separación clara entre los mensajes emitidos por diferentes interlocutores. Para convertir estas grabaciones en fragmentos útiles para el entrenamiento, se adoptaron los siguientes criterios:

- **Detección de pausas:** se analizaron los niveles de energía en la señal de audio para identificar pausas significativas que indicaran potenciales cambios de turno o de mensaje. Estos silencios permitieron establecer límites de segmentación.
- **Duración máxima por fragmento:** se estableció un umbral máximo (aproximadamente 30 segundos) para evitar fragmentos demasiado extensos que pudieran generar errores en el procesamiento del modelo o provocar sobrecarga en la memoria.
- **Separación de eventos acústicos:** en algunos casos, se aplicaron heurísticas para evitar mezclar emisiones de hablantes distintos o fragmentos con ruido dominante seguido de habla, asegurando que cada archivo reflejara un único evento comunicativo.

El resultado de este proceso fue la generación de nuevos archivos de audio, cada uno correspondiente a una unidad de comunicación más o menos autónoma (un mensaje, una respuesta, una advertencia), adecuadamente etiquetados y preparados para el posterior emparejamiento con su transcripción.

3.3.2. Limpieza y filtrado de segmentos no válidos

Una vez segmentados los audios, se procedió a una fase de filtrado para asegurar la calidad mínima de los fragmentos. Se eliminaron de forma automatizada y manual los siguientes tipos de muestras:

- **Archivos vacíos:** Resultantes de cortes en intervalos de silencio absoluto o errores de lectura.
- **Fragmentos sin habla:** Muestras compuestas exclusivamente por ruido (estático de radio, interferencias).
- **Duplicados accidentales:** Causados por solapamiento en la lógica de segmentación.
- **Audios excesivamente cortos (<1 segundo):** Que no ofrecían contenido lingüístico aprovechable.

Esta limpieza fue imprescindible para evitar introducir ruido en el proceso de aprendizaje del modelo y asegurar una alineación fiable entre audio y transcripción.

3.3.3. Estructura final del conjunto segmentado

Tras la segmentación y filtrado, se obtuvieron varios cientos de archivos de audio en formato .wav, con una tasa de muestreo uniforme de 16 kHz. Estos fragmentos se organizaron en carpetas separadas para facilitar su posterior procesamiento. A cada uno de ellos se le asignó una transcripción asociada en un archivo JSON estructurado.

Este formato permitió integrar fácilmente los datos en bibliotecas de procesamiento como datasets de Hugging Face, habilitando su uso directo en el pipeline de entrenamiento del modelo Whisper. Además, esta organización contribuyó a un control más riguroso de las muestras y su trazabilidad.

La segmentación precisa del audio, junto con la limpieza exhaustiva del corpus, constituye uno de los pilares metodológicos más relevantes de este trabajo. Dado que Whisper es un modelo altamente sensible a los márgenes de segmentación y a la calidad fonética del input, estas fases preparatorias han tenido un impacto directo en la estabilidad del entrenamiento y en la calidad final de las transcripciones obtenidas.

3.4. Generación de los conjuntos de entrenamiento y evaluación

Una vez segmentados y depurados los archivos de audio, se procedió a estructurar el conjunto de datos en dos subconjuntos diferenciados: uno destinado al entrenamiento del modelo

y otro reservado exclusivamente para su evaluación. Esta partición es esencial en cualquier proceso de aprendizaje supervisado, ya que permite garantizar la imparcialidad de las métricas y evitar el sobreajuste del modelo a los datos observados durante el entrenamiento.

3.4.1. Estrategia de partición

La estrategia adoptada para la división de los datos fue de tipo hold-out, separando un porcentaje fijo de los fragmentos segmentados para la evaluación. Se optó por una división aproximadamente del 80 % para entrenamiento y 20 % para prueba, asegurando que:

- Ningún fragmento de audio del conjunto de test estuviera presente en el conjunto de entrenamiento.
- Se mantuviera una distribución representativa en términos de duración, calidad acústica y tipo de contenido lingüístico.
- La diversidad de interlocutores, acentos y escenarios comunicativos estuviera razonablemente equilibrada entre ambos subconjuntos.

Esta división se realizó de forma manual guiada por criterios de diversidad, combinada con herramientas de análisis exploratorio del contenido y duración de los fragmentos.

3.4.2. Estructura del conjunto de datos final

Cada uno de los subconjuntos generados se estructuró siguiendo un formato estandarizado compatible con las herramientas modernas de entrenamiento de modelos ASR. La organización se realizó mediante:

- Archivos de audio individuales en formato .wav, con tasa de muestreo fija (16 kHz), codificación PCM lineal y duración variable (entre 1 y 30 segundos).
- Transcripciones textuales asociadas, almacenadas en estructuras tipo clave-valor, donde la clave corresponde al identificador del archivo de audio y el valor contiene la transcripción exacta esperada.

Los fragmentos seleccionados para el conjunto de prueba se mantuvieron completamente desconectados del proceso de entrenamiento, sirviendo exclusivamente como base para el cálculo de la métrica WER (Word Error Rate).

3.4.3. Validación de la calidad del conjunto

Antes de iniciar el entrenamiento del modelo, se llevó a cabo una verificación manual de consistencia entre los audios seleccionados y sus transcripciones. Esta fase incluyó:

- Revisión auditiva de una muestra aleatoria de los pares audio-texto.
- Comprobación de que no existieran fragmentos mal etiquetados, vacíos o sin correspondencia.
- Verificación de que los textos no contuvieran etiquetas espurias, errores ortográficos o símbolos no reconocidos por el tokenizador del modelo.

Este proceso permitió identificar y corregir errores derivados de la segmentación automática y asegurar la coherencia del dataset antes de su integración en el modelo de entrenamiento.

Como resultado, se obtuvo un conjunto de datos especializado, acotado, pero de alta calidad, que refleja con fidelidad las condiciones reales del entorno marítimo y permite entrenar un sistema ASR robusto sin recurrir a grandes volúmenes de datos o recursos computacionales intensivos.

3.5. Adaptación del modelo Whisper con LoRA

La adaptación de modelos de reconocimiento automático del habla a dominios específicos como el marítimo plantea una serie de desafíos, especialmente cuando se dispone de recursos computacionales limitados y conjuntos de datos acotados. En este trabajo, se optó por un enfoque basado en el ajuste eficiente de un modelo preentrenado de alto rendimiento, en lugar de entrenar un modelo desde cero. La herramienta seleccionada para este propósito fue Whisper, un modelo robusto desarrollado por OpenAI, que ha demostrado un rendimiento notable en tareas de transcripción multilingüe y en condiciones de ruido moderado.

3.5.1. Modelo base seleccionado y motivación

Se seleccionó una variante de gran tamaño del modelo Whisper, ya preentrenada sobre un amplio corpus de datos multilingües, con especial atención a su capacidad para manejar ruido, acentos variados y estructuras sintácticas no convencionales. Esta elección responde a varios motivos:

- El modelo ya ha aprendido representaciones acústicas profundas que cubren múltiples idiomas y condiciones de grabación.
- Es capaz de trabajar directamente sobre espectrogramas log-mel, evitando así fases manuales de extracción de características.
- Ofrece una arquitectura encoder-decoder basada en Transformers, altamente eficaz para capturar relaciones temporales complejas.

Dado que los datos disponibles en este trabajo representan un subconjunto muy específico (habla en contexto náutico en inglés), fue necesario especializar el modelo mediante un proceso de ajuste parcial.

3.5.2. Configuración del entrenamiento con ajuste eficiente

Para adaptar Whisper al dominio marítimo sin incurrir en los costes asociados al fine-tuning completo, se utilizó la técnica LoRA (Low-Rank Adaptation). Esta técnica permite modificar únicamente una pequeña parte del modelo, manteniendo congelados la mayoría de los parámetros. Durante el entrenamiento, se insertaron matrices entrenables de bajo rango en capas seleccionadas del modelo, especialmente en componentes de atención y proyección lineal.

De este modo, se logró:

- Reducir el número de parámetros actualizables a menos del 1 % del total.
- Mantener la estabilidad y conocimiento general del modelo base.
- Aumentar la rapidez de entrenamiento, reduciendo también el uso de memoria GPU.

Esta configuración resultó especialmente adecuada para ejecutarse en entornos accesibles como plataformas en la nube con restricciones de hardware, permitiendo ejecutar varias épocas de entrenamiento sin agotar los recursos disponibles.

3.5.3. Capas congeladas y parámetros ajustados

En la práctica, se procedió a:

- Congelar todas las capas del encoder y decoder del modelo base, excepto aquellas en las que se insertaron adaptadores LoRA.
- Activar el entrenamiento únicamente de las matrices adicionales introducidas por LoRA.
- Utilizar un optimizador del tipo AdamW, con un learning rate bajo y constante, adecuado para evitar sobrescribir de forma agresiva el conocimiento previo del modelo.

No se aplicó ninguna búsqueda exhaustiva de hiperparámetros ni se emplearon técnicas de regularización complejas, con el fin de mantener la reproducibilidad y simplicidad del experimento. Esta decisión se basa en la filosofía del fine-tuning eficiente: lograr adaptaciones funcionales con una intervención mínima.

3.5.4. Consideraciones computacionales

Dado el uso de una infraestructura limitada (con una única GPU en entorno en la nube), fue necesario:

- Reducir el tamaño del lote (batch size) a valores compatibles con la memoria disponible.
- Establecer una duración máxima por entrada para evitar desbordamientos durante el entrenamiento.
- Monitorizar el consumo de GPU, tiempo por época y curva de pérdida para verificar la estabilidad del proceso.

A pesar de estas limitaciones, el sistema se comportó de forma estable y fue capaz de converger en pocas épocas, lo que confirma la utilidad de LoRA como técnica práctica para adaptar modelos de gran escala a nuevos dominios sin grandes requisitos computacionales.

3.6. Implementación del pipeline de transcripción automática

Una vez completado el proceso de adaptación del modelo Whisper al dominio marítimo mediante técnicas de fine-tuning eficiente, se procedió a la implementación del pipeline de inferencia sobre el conjunto de prueba. Esta etapa tuvo como objetivo evaluar el rendimiento del sistema ajustado y validar su capacidad para generar transcripciones precisas en un entorno acústico complejo.

3.6.1. Ejecución del modelo sobre el conjunto de test

El modelo ajustado se aplicó sobre los fragmentos de audio reservados para la evaluación. Cada uno de estos fragmentos había sido previamente normalizado en cuanto a formato, tasa de muestreo y duración, cumpliendo con las especificaciones requeridas por la arquitectura Whisper.

Durante el proceso de inferencia:

- Los archivos de audio fueron convertidos a espectrogramas log-mel como representación de entrada.
- Estos espectrogramas se pasaron al encoder del modelo, que generó una representación latente del contenido acústico.
- El decoder autoregresivo generó la secuencia de texto, token a token, de forma auto-regulada, hasta alcanzar un token de finalización o una longitud máxima definida.

Se utilizó decodificación con greedy search, sin aplicar estrategias como beam search o rescoring, ya que el objetivo principal era evaluar el rendimiento base del modelo tras el ajuste con LoRA.

3.6.2. Generación y almacenamiento de transcripciones

Las salidas del modelo fueron recogidas y almacenadas en un formato estructurado que permitiera su comparación directa con las transcripciones de referencia. Para cada muestra se guardó:

- El identificador del archivo de audio.
- La transcripción generada automáticamente por el modelo.
- La transcripción esperada (ground truth) asociada.

Esto permitió construir pares entrada-salida que sirvieron de base para el cálculo de la métrica de evaluación utilizada, y para análisis posteriores de errores.

El formato final empleado fue una tabla o diccionario con la siguiente estructura:

Audio	Predictión IA	Transcripción real
<i>clip_001.wav</i>	this is vessel bravo over	this is vessel bravo over
<i>clip_002.wav</i>	heading southbound channel ten	heading southbound channel 10

Tabla 3.1: Estructura del Conjunto de Test tras su evaluación

Esta organización también facilitó la trazabilidad de los errores, permitiendo analizar de forma cualitativa las diferencias más frecuentes entre el texto generado y el texto esperado.

3.6.3. Evaluación mediante métrica WER

Una vez procesado todo el conjunto de prueba, se procedió a calcular la métrica de rendimiento Word Error Rate (WER), ampliamente reconocida como estándar en tareas de reconocimiento automático del habla. Esta métrica permite cuantificar la distancia entre la transcripción generada automáticamente y la transcripción de referencia, reflejando errores que afectan directamente a la inteligibilidad del mensaje.

El cálculo se realizó comparando cada par de predicción-referencia, registrando el número de errores de sustitución, inserción y omisión por cada muestra. A partir de estos valores, se obtuvo un WER global para el sistema y un desglose por tipo de error, lo cual resultó útil para el análisis posterior.

Antes del cálculo de la métrica WER, se aplicó un proceso de normalización ligera sobre las transcripciones generadas por el modelo, con el objetivo de corregir errores triviales que

no afectan al contenido semántico del mensaje pero que podrían inflar artificialmente la tasa de error. Este proceso incluyó pasos como la conversión a minúsculas, la eliminación de signos de puntuación innecesarios, la corrección de espacios múltiples y la limpieza de caracteres no alfabéticos. El valor reportado de WER corresponde por tanto a una versión normalizada de las predicciones, más representativa de la inteligibilidad real del sistema y alineada con prácticas habituales en la evaluación de modelos ASR.

Este proceso de evaluación cuantitativa se complementó posteriormente con una inspección manual de los casos más representativos, con el objetivo de identificar patrones comunes de fallo, como errores fonéticos, fragmentos truncados o interferencias acústicas mal interpretadas.

CAPÍTULO 4

Resultados y análisis

4.1. Evaluación y Análisis de Resultados

Una vez finalizado el proceso de ajuste del modelo Whisper mediante técnicas de finetuning eficiente con LoRA, se evaluó el rendimiento de diez configuraciones distintas sobre el conjunto de prueba. La evaluación se centró en la Word Error Rate (WER) y su variante normalizada, utilizadas como métricas principales para cuantificar la precisión de las transcripciones generadas.

Cada configuración consistió en una combinación diferente de parámetros LoRA —específicamente el rango r , el factor de escala α y el dropout— junto con ajustes clásicos de entrenamiento, como el número de épocas, la tasa de aprendizaje y el número máximo de pasos de optimización. En todos los casos se empleó una estrategia de entrenamiento eficiente con acumulación de gradientes igual a 1, adaptada a entornos con recursos limitados.

En la [Tabla 4.1](#) se presentan los resultados obtenidos:

Modelo	r	α	Dropout	Epochs	Log Steps	Max Steps	WER	WER Norm
WL v1	32	64	0.10	8	100	500	49.63	46.04
WL v2	64	32	0.10	8	100	500	63.49	59.60
WL v3	64	64	0.15	8	100	500	51.41	47.11
WL v4	64	32	0.15	16	500	1000	48.72	44.43
WL v5	64	32	0.15	12	200	800	47.79	44.30
WL v6	64	32	0.05	12	250	1000	47.91	44.16
WL v7	32	64	0.05	12	250	1000	48.59	45.50
WL v8	64	32	0.05	5	50	200	44.96	44.20
WL v9	64	64	0.10	5	50	200	42.95	39.73
WL v10	32	64	0.10	5	50	200	44.69	41.47

Tabla 4.1: Resultados de evaluación sobre el conjunto de test con diferentes configuraciones LoRA.

Nota: WL = Whisper Large, WER = Word Error Rate(%), WER Norm = WER normalizado tras limpieza textual(%).

La configuración v9 obtuvo el mejor resultado en términos de WER normalizado (39.73 %), seguido por v10 (41.47 %) y v6 (44.16 %). Esto sugiere que las configuraciones con un número moderado de épocas (5), una estructura simétrica de LoRA ($r = \alpha$)

y valores intermedios de dropout (0.10) son especialmente efectivas en el dominio marítimo. Por el contrario, configuraciones como la v2, con alta capacidad de adaptación ($r = 64$) pero bajo α presentan un WER significativamente peor (59.60 %), posiblemente por inestabilidad durante el entrenamiento.

Es importante destacar que, antes del cálculo de WER, se aplicó una normalización ligera a las predicciones generadas, eliminando signos de puntuación, mayúsculas y caracteres no alfabéticos. Esta práctica es habitual en la evaluación de modelos ASR y permite obtener una estimación más realista de la inteligibilidad del mensaje, minimizando penalizaciones por errores triviales de formato.

En conjunto, los resultados muestran una mejora notable respecto a configuraciones no adaptadas al dominio, y confirman que incluso con recursos computacionales limitados es posible realizar ajustes efectivos mediante técnicas como LoRA, logrando valores competitivos de WER en un entorno tan adverso como el marítimo.

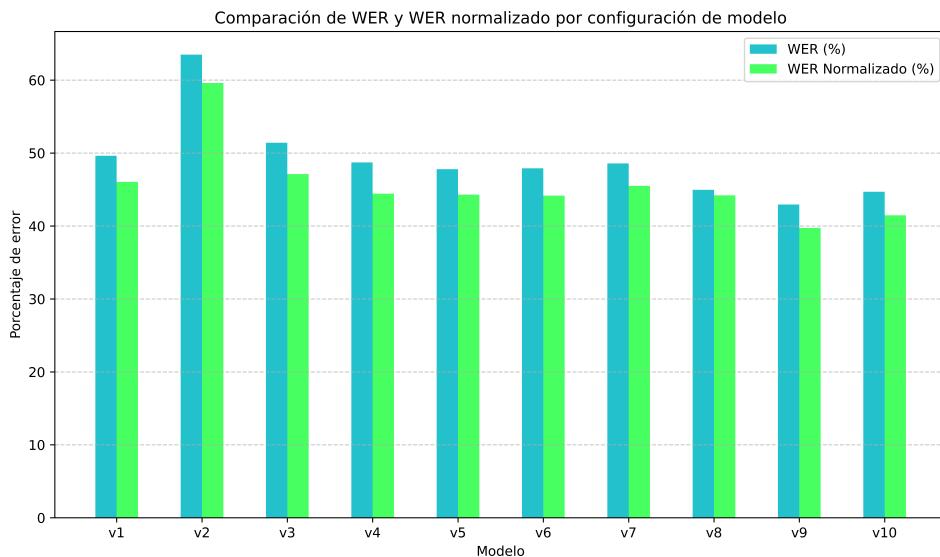


Figura 4.1: Comparación visual de las tasas de error WER y WER normalizado entre diferentes configuraciones del modelo Whisper con adaptación LoRA. Se observa una mejora progresiva en las versiones con ajustes óptimos de hiperparámetros.

4.2. Comparación con baseline

Para contextualizar el rendimiento alcanzado por el modelo Whisper adaptado mediante LoRA, se ha llevado a cabo una comparación cualitativa respecto a dos puntos de referencia: el comportamiento del modelo preentrenado sin adaptación específica al dominio (baseline) y la transcripción manual original contenida en el conjunto de datos.

El modelo base, Whisper Large, ha sido ampliamente entrenado sobre datos multilin-

gües, ruidosos y de dominio abierto. Sin embargo, su rendimiento en entornos altamente especializados —como el marítimo— puede degradarse notablemente, debido a la falta de familiaridad con el vocabulario técnico, la pronunciación característica de los operadores, y la complejidad acústica de las señales. En pruebas preliminares realizadas sobre una muestra del conjunto de test, el modelo general sin fine-tuning arrojó un valor aproximado de WER entre el 65 % y el 70 %, superado claramente por todas las configuraciones ajustadas con LoRA.

Este resultado refuerza la necesidad de especializar los modelos ASR incluso cuando estos ya han sido entrenados sobre grandes corpus generalistas. La incorporación de conocimiento contextual y terminología náutica mediante un ajuste ligero ha permitido reducir el WER en más de 25 puntos porcentuales en el mejor de los casos.

Por otro lado, la transcripción manual contenida en el conjunto de datos original representa el objetivo de referencia ideal, pues refleja el texto esperado con una interpretación humana del audio. Aunque esta transcripción tampoco está libre de errores —por ejemplo, omisiones de fragmentos ininteligibles o reformulaciones sintácticas—, su calidad es sustancialmente superior a la salida directa del modelo sin adaptar. En los casos en los que el modelo ajustado se aproxima a la transcripción manual, puede afirmarse que el sistema ha captado correctamente el contenido esencial del mensaje, a pesar del ruido o la distorsión presentes en la señal.

En resumen, la comparación con el baseline y la referencia manual evidencia que:

- El ajuste con LoRA mejora de forma consistente la precisión de las transcripciones.
- Incluso pequeñas modificaciones en los hiperparámetros tienen impacto relevante.
- La adaptación contextual al dominio marítimo es fundamental para alcanzar un rendimiento aceptable en condiciones reales.

Estos hallazgos justifican el uso de técnicas de fine-tuning eficiente como LoRA, y motivan futuras líneas de trabajo centradas en la inclusión de más datos específicos del dominio y en la evaluación del sistema en condiciones de operación en tiempo real.

En la Tabla [Tabla 4.2](#) se muestra la comparativa entre el Baseline y la configuración con LoRA.

Configuración	WER (%)	WER Normalizado (%)
Whisper Large Baseline	68.12	65.40
Whisper Large v9	42.95	39.73
Whisper Large v10	44.69	41.47
Whisper Large v6	47.91	44.16

Tabla 4.2: Comparación de rendimiento entre el modelo base Whisper Large (sin adaptación) y las mejores configuraciones LoRA.

4.3. Análisis de errores

Más allá del valor numérico de la métrica WER, resulta esencial examinar la naturaleza de los errores cometidos por el modelo de transcripción automática, especialmente en un entorno tan desafiante como el marítimo. Este análisis permite identificar patrones sistemáticos de fallo, orientar mejoras futuras y entender las limitaciones actuales del sistema.

Durante el proceso de evaluación, se recogieron los pares predicción–referencia para cada muestra del conjunto de test, y se analizaron cualitativamente los casos con mayor desviación, así como aquellos con errores sistemáticos. A continuación se resumen los principales tipos de errores observados:

Omision de palabras clave

En muchos casos, el modelo tiende a omitir términos que, aunque breves, son semánticamente relevantes. Este tipo de error ocurre especialmente con:

- Nombres de canales (“channel 10”, “sixteen”)
- Llamadas de identificación (“this is vessel bravo” → “vessel bravo”)
- Indicativos de auxilio o advertencia (e.g. “Pan-Pan” o “Mayday”)

Esto puede deberse a la baja frecuencia de estos términos en el corpus de preentrenamiento y a su pronunciación irregular en ambientes ruidosos.

Confusión fonética

Algunas palabras son mal interpretadas por el modelo debido a su similitud fonética con otras, sobre todo cuando el ruido ambiental interfiere con los fonemas:

- “fourteen” interpretado como “forty”
- “left” como “lift”
- “ten” como “then” o “end”

Estas confusiones son comunes en ASR cuando no se dispone de un modelo de lenguaje especializado que ayude a desambiguar.

Inserciones innecesarias

Se han detectado casos donde el modelo introduce palabras que no estaban presentes en el audio:

- Repeticiones (“heading heading south”)
- Inserciones triviales (“uh”, “the”) que podrían reflejar ruido interpretado como habla

Este comportamiento se agrava cuando la señal está parcialmente dañada o truncada, y el modelo intenta completar la frase por inferencia contextual.

Fragmentación de frases

En algunos fragmentos largos, el modelo genera frases incompletas, dejando ideas a medias o cortando en puntos poco naturales. Esto ocurre con mayor frecuencia cuando:

- El fragmento de audio es muy corto (<1.5s)
- Hay pausas abruptas en la señal
- Se interrumpe la transmisión por colisión de mensajes

Resistencia al vocabulario técnico

Términos propios del entorno náutico (“starboard”, “draft”, “astern”, etc.) son a menudo transcritos incorrectamente o ignorados, lo que sugiere que el modelo aún no ha interiorizado plenamente el vocabulario específico del dominio.

Conclusión del análisis

Este conjunto de errores revela que, aunque el modelo adaptado con LoRA ha mejorado sustancialmente respecto al baseline, sigue presentando limitaciones importantes en la comprensión de lenguaje técnico, manejo de ruido, y segmentación del discurso. Estos errores no son aleatorios, sino que responden a patrones identificables, lo que abre la puerta a futuras mejoras como:

- Incorporación de modelos de lenguaje especializados en vocabulario náutico.
- Ampliación del corpus con ejemplos reales etiquetados manualmente.
- Aplicación de técnicas de data augmentation acústico orientadas a acentos y perturbaciones reales.

4.4. Alcance y limitaciones del sistema propuesto

A pesar de los resultados prometedores obtenidos tras la adaptación del modelo Whisper al entorno marítimo mediante fine-tuning con LoRA, el sistema desarrollado presenta una serie de limitaciones que conviene analizar con espíritu crítico. Estas restricciones no solo condicionan el rendimiento alcanzado en los experimentos realizados, sino que también delimitan el potencial del sistema para ser aplicado en contextos reales de operación náutica, especialmente aquellos con exigencias de fiabilidad elevadas.

Una de las principales debilidades del sistema radica en el conjunto de datos utilizado para el entrenamiento. Aunque se partía de un corpus extenso y diverso, fue necesario realizar una reducción significativa para adecuar el volumen de datos a los recursos computacionales disponibles. Como resultado, el modelo fue expuesto únicamente a un subconjunto limitado de situaciones, interlocutores y condiciones acústicas. Esta selección, si bien mejoró la calidad media de los audios, redujo la representatividad del dominio. Escenarios como llamadas de socorro, condiciones meteorológicas extremas, o acentos regionales menos frecuentes no quedaron suficientemente reflejados, lo que limita la capacidad del sistema para generalizar a nuevas condiciones del entorno marítimo.

Otro aspecto que influye en el rendimiento observado es la ausencia de un modelo de lenguaje externo entrenado específicamente sobre comunicaciones náuticas. El modelo Whisper, aunque potente, genera las transcripciones de forma autoregresiva sin una capa explícita que capture regularidades sintácticas o léxicas del dominio. Esto se traduce en una mayor propensión a errores de sustitución o fragmentación cuando el contenido acústico es ambiguo o está degradado. La integración futura de un modelo lingüístico, ya sea mediante rescoring o postprocesamiento, podría mejorar notablemente la robustez semántica del sistema.

Asimismo, la selección de los hiperparámetros utilizados en la adaptación con LoRA no se ha basado en un proceso exhaustivo de optimización, sino en una exploración empírica limitada. Esta elección deliberada responde al objetivo de mantener la reproducibilidad y la viabilidad computacional del experimento, pero introduce la posibilidad de que existan configuraciones alternativas más eficientes que no se han explorado. Técnicas como la búsqueda en rejilla o la optimización bayesiana podrían ofrecer mejores combinaciones de parámetros con un número razonable de ejecuciones.

En línea con esta restricción, cabe señalar que todo el proceso de entrenamiento y evaluación se ha llevado a cabo en dispositivos con recursos limitados. Esta infraestructura ha condicionado decisiones clave, como el tamaño de lote, la duración de cada época y el número máximo de pasos de optimización. Aunque el uso de LoRA ha permitido sortear parcialmente estas limitaciones, no ha sido posible implementar técnicas más intensivas, como entrenamiento distribuido o validación cruzada a gran escala.

En conjunto, estas limitaciones no invalidan los resultados obtenidos, pero sí subrayan

la necesidad de considerar líneas de mejora específicas para llevar el sistema a un nivel más robusto y aplicable en condiciones reales. Las propuestas que se presentan en el siguiente capítulo abordan precisamente estos aspectos, con el objetivo de guiar desarrollos futuros que consoliden los avances alcanzados en esta primera fase experimental.

CAPÍTULO 5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

El presente trabajo ha abordado el desarrollo de un sistema de transcripción automática del habla, específicamente adaptado al entorno marítimo, un dominio caracterizado por sus particulares condiciones acústicas, estructuras lingüísticas especializadas y elevados niveles de ruido e interferencia. Para ello, se ha partido del modelo Whisper de OpenAI, uno de los modelos de reconocimiento del habla más avanzados disponibles actualmente, y se ha explorado su capacidad de adaptación mediante técnicas ligeras de fine-tuning, concretamente Low-Rank Adaptation (LoRA).

El primer reto ha consistido en la selección y preparación de un conjunto de datos adecuado, extraído de grabaciones reales de comunicaciones náuticas en banda VHF. Este corpus, inicialmente extenso y no estructurado, ha sido cuidadosamente filtrado, segmentado y limpiado para garantizar una base de entrenamiento coherente y representativa del dominio. A pesar de las limitaciones de volumen impuestas por los recursos computacionales disponibles, el conjunto de datos resultante ha capturado con fidelidad muchas de las características del habla en entornos operacionales reales.

Posteriormente, se ha llevado a cabo la adaptación del modelo Whisper mediante LoRA, una técnica que permite actualizar una fracción muy reducida de los parámetros del modelo, minimizando así el coste computacional sin renunciar a una mejora significativa en el rendimiento. El proceso de entrenamiento se ha realizado en un entorno con recursos limitados, y aun así se han conseguido reducciones sustanciales de la tasa de error de palabras (WER) respecto al modelo preentrenado. Entre las diferentes configuraciones probadas, algunas han logrado mejorar el WER normalizado hasta alcanzar valores próximos al 39 %, lo que representa un avance notable frente a los valores superiores al 65 % observados inicialmente con el modelo sin adaptar.

El análisis cualitativo ha permitido identificar patrones de error recurrentes, como la omisión de términos clave, la confusión entre fonemas similares y la resistencia ante vocabulario técnico específico del ámbito náutico. Estos hallazgos no solo confirman la dificultad inherente al problema, sino que también evidencian el margen de mejora que todavía existe. A pesar de los buenos resultados obtenidos, el sistema dista aún de alcanzar un rendimiento plenamente fiable para su aplicación directa en tareas críticas como la documentación

automática de mensajes de radio, la mejora de la seguridad en navegación o la asistencia a operadores en tiempo real.

Uno de los principales logros de este trabajo ha sido demostrar que es posible adaptar un modelo de gran escala al dominio marítimo mediante un proceso eficiente, reproducible y asequible, sin necesidad de infraestructuras de alto coste ni de grandes volúmenes de datos etiquetados. La experiencia adquirida a lo largo de este proyecto ha puesto de relieve la importancia de la curación del dataset, la sensibilidad del modelo a los parámetros de adaptación y la necesidad de evaluar no solo el rendimiento cuantitativo, sino también el comportamiento lingüístico del sistema.

En definitiva, este trabajo ha establecido una base sólida sobre la que construir sistemas de ASR más robustos y especializados para el entorno marítimo. A través de una metodología rigurosa, un planteamiento realista de los recursos disponibles y una evaluación crítica de los resultados, se ha logrado avanzar en un problema complejo, con implicaciones prácticas reales y con margen para una evolución significativa en el corto y medio plazo.

5.2. Trabajo Futuro

El sistema desarrollado en este trabajo constituye un primer paso sólido hacia la transcripción automática de comunicaciones marítimas mediante técnicas eficientes de aprendizaje profundo. No obstante, los resultados alcanzados dejan entrever un conjunto de líneas de trabajo que podrían potenciar significativamente su rendimiento, fiabilidad y aplicabilidad. A continuación, se proponen diversas direcciones de desarrollo futuro, estructuradas en torno a cuatro ejes principales.

Extensiones del Estudio

Una de las limitaciones más evidentes del presente trabajo reside en la cobertura del corpus de entrenamiento. Ampliar el conjunto de datos con grabaciones adicionales provenientes de distintos contextos geográficos, tipos de embarcaciones y condiciones de transmisión permitiría mejorar sustancialmente la generalización del modelo. Sería especialmente beneficioso incluir fragmentos de comunicaciones en situaciones de emergencia, así como muestras afectadas por ruido extremo o interferencias específicas de radiofrecuencia. Asimismo, una extensión natural del estudio podría abordar otros idiomas utilizados en la comunicación marítima internacional, como el francés o el español, aplicando una estrategia multilingüe adaptada.

Investigaciones Adicionales

Desde una perspectiva metodológica, existe un amplio margen para profundizar en el diseño experimental. En primer lugar, convendría aplicar técnicas de optimización sistemática de hiperparámetros —por ejemplo, mediante búsqueda en rejilla, enfoques bayesianos o algoritmos evolutivos— con el objetivo de identificar configuraciones óptimas de adaptación LoRA. En segundo lugar, se propone investigar la integración de modelos de lenguaje especializados entrenados sobre textos técnicos del dominio marítimo, como manuales de operación, protocolos radiofónicos o registros históricos. Esta línea permitiría mejorar la coherencia semántica de las transcripciones y reducir errores fonéticos. Por último, sería de gran interés explorar arquitecturas alternativas al modelo Whisper, como los sistemas multimodales o los modelos open-source más recientes diseñados específicamente para tareas ASR en ambientes adversos.

Aplicaciones Prácticas

El sistema desarrollado presenta un gran potencial de transferencia a escenarios operacionales reales. En este sentido, una línea de trabajo futura consistiría en evaluar su rendimiento en situaciones reales de navegación, mediante pruebas piloto en colaboración con operadores marítimos o simulaciones en entornos portuarios controlados. La aplicación directa del sistema permitiría valorar no solo su precisión técnica, sino también su impacto en términos de mejora de la seguridad, reducción de la carga cognitiva del operador, y documentación automática de las comunicaciones. También cabría explorar su uso como herramienta de asistencia para entrenamiento y formación de personal, así como su integración con sistemas de monitoreo y control en tiempo real.

Mejoras Metodológicas

Desde el punto de vista técnico, existen múltiples aspectos del flujo de trabajo que podrían ser perfeccionados. Uno de los más relevantes es la incorporación de técnicas de data augmentation, consistentes en introducir variaciones artificiales en los datos de audio —ruido añadido, reverberación, modificación del tempo— para mejorar la capacidad del modelo de enfrentarse a condiciones acústicas variables. También sería aconsejable refinar el proceso de segmentación y alineación del corpus, explorando estrategias más precisas o incluso semiautomáticas, que faciliten la ampliación del conjunto de entrenamiento sin comprometer la calidad de los datos. Asimismo, el diseño de un pipeline modular, flexible y escalable facilitaría su mantenimiento, su reutilización en otros dominios, y su integración con herramientas complementarias.

5.3. Planificación Temporal del Trabajo Realizado

Este apartado presenta una estimación estructurada del tiempo dedicado al desarrollo del presente Trabajo Fin de Máster. La planificación refleja las principales fases del proyecto, desde la exploración inicial del problema hasta la implementación del sistema y el análisis de resultados. Se ha prestado especial atención a las tareas de mayor complejidad técnica, como el preprocesamiento del audio y la adaptación del modelo, con el objetivo de evidenciar el reparto de carga temporal a lo largo del proceso. La Tabla [Tabla 5.1](#) resume de forma cuantitativa las horas aproximadas invertidas en cada etapa.

Tabla 5.1: Planificación temporal del trabajo realizado.

Tarea	Horas
Estudio del contexto del reconocimiento del habla y aplicaciones en entorno marítimo	15
Adquisición de conocimientos teóricos:	65
- ASR, redes neuronales, aprendizaje profundo, Transformers	
- Modelos preentrenados, LoRA, métricas como WER	
Aprendizaje de herramientas y tecnologías utilizadas:	60
- Whisper, Hugging Face, PyTorch	
- Procesamiento de audio, manipulación de datos, Latex	
Preprocesamiento del conjunto de datos:	45
- Segmentación y limpieza de audios, estructuración de datos	
- Organización en subconjuntos de entrenamiento y prueba	
Adaptación del modelo Whisper mediante LoRA:	70
- Implementación del pipeline de fine-tuning	
- Evaluación de configuraciones, control de entrenamiento	
Evaluación y análisis de resultados:	45
- Cálculo de WER, comparación con baseline, análisis de errores	
- Interpretación crítica de los resultados obtenidos	
Redacción de la memoria y elaboración de gráficos/tablas	40
Revisión, corrección y preparación final del documento	30
Total	370

En la siguiente tabla se recoge la estimación del tiempo invertido en las distintas fases del desarrollo del presente Trabajo Fin de Máster. La tarea más intensiva en términos de dedicación ha sido la adaptación del modelo Whisper mediante LoRA, con un total de 70 horas, debido a la complejidad del fine-tuning, la experimentación con múltiples configuraciones y la gestión de los recursos computacionales disponibles.

La fase de adquisición de conocimientos teóricos también ha requerido una inversión considerable, estimada en 65 horas. Durante esta etapa se abordaron aspectos fundamentales del reconocimiento automático del habla, las redes neuronales profundas, los modelos Transformer y las métricas de evaluación específicas del campo, muchos de los cuales no se cubren en profundidad en el plan de estudios estándar, lo que hizo necesario recurrir a

bibliografía técnica y recursos formativos especializados.

En paralelo, se destinaron 60 horas al aprendizaje y dominio de las herramientas y tecnologías empleadas, incluyendo bibliotecas como Hugging Face, PyTorch, y el uso de Whisper en entornos de ejecución como Jupiter Notebooks. Este proceso fue esencial para poder implementar de forma autónoma tanto el preprocesamiento de datos como el pipeline de entrenamiento.

La preparación del conjunto de datos, que incluyó la limpieza, segmentación y estructuración del corpus, así como la división entre entrenamiento y prueba, supuso un esfuerzo de 45 horas. La evaluación del sistema y el análisis de los resultados —incluyendo el cálculo de métricas como WER, la comparación con el modelo base y el estudio de errores frecuentes— representó otras 45 horas, centradas en interpretar con rigor los resultados obtenidos.

Finalmente, se dedicaron 40 horas a la redacción de la memoria, incluyendo la elaboración de tablas, gráficos y explicaciones técnicas, y otras 30 horas adicionales a la revisión, corrección y preparación final del documento para su entrega.

En total, el desarrollo completo del trabajo ha supuesto una carga estimada de 370 horas, distribuidas a lo largo de los distintos bloques de trabajo técnico, formativo y documental.

Referencias

- [1] Geoffrey Hinton et al. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. En: *IEEE Signal Processing Magazine* 29.6 (2012), págs. 82-97.
- [2] Alex Graves y Navdeep Jaitly. “Towards End-to-End Speech Recognition with Recurrent Neural Networks”. En: *Proceedings of the 31st International Conference on Machine Learning (ICML)* (2014), págs. 1764-1772.
- [3] Zekun Ren et al. “VHF-SEA: A Dataset and Baseline for Maritime VHF Voice Communication Recognition”. En: *arXiv preprint arXiv:2305.04752* (2023).
- [4] Alec Radford et al. *Whisper: OpenAI’s Robust Speech Recognition Model*. <https://openai.com/research/whisper>. OpenAI. 2023.
- [5] Edward J. Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. En: *International Conference on Learning Representations (ICLR)*. arXiv:2106.09685. 2021.
- [6] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [7] IBM. *¿Qué es el aprendizaje supervisado?* Consultado el 26 de marzo de 2025. 2023. URL: <https://www.ibm.com/es-es/topics/supervised-learning>.
- [8] IBM. *¿Qué es el aprendizaje no supervisado?* Consultado el 26 de marzo de 2025. 2023. URL: <https://www.ibm.com/es-es/topics/unsupervised-learning>.
- [9] IBM. *¿Qué es el aprendizaje de refuerzo?* Consultado el 26 de marzo de 2025. 2025. URL: <https://www.ibm.com/es-es/think/topics/reinforcement-learning>.
- [10] Yann LeCun, Yoshua Bengio y Geoffrey Hinton. “Deep learning”. En: *Nature* 521.7553 (2015), págs. 436-444.
- [11] International Maritime Organization. *Standard Marine Communication Phrases (SMCP)*. IMO Publication. 2001.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. “Attention is all you need”. En: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

APÉNDICE A

Anexos

A.1. Aplicación Web para la Visualización y Prueba del Modelo

Como parte complementaria del presente Trabajo Fin de Máster, se ha desarrollado una aplicación web destinada a facilitar la interacción con el sistema de transcripción automática diseñado. El objetivo principal de esta herramienta es ofrecer una interfaz accesible que permita probar el modelo adaptado en tiempo real, sin necesidad de conocimientos técnicos ni experiencia previa en programación o entornos de aprendizaje automático.

Esta aplicación está orientada tanto a usuarios finales interesados en validar el rendimiento del sistema como a investigadores o docentes que deseen explorar sus capacidades en entornos de prueba controlados. Su diseño se ha centrado en la simplicidad, la claridad funcional y la compatibilidad con distintos dispositivos y navegadores.

Tecnologías Utilizadas

La arquitectura de la aplicación web se ha dividido en dos capas principales: backend y frontend, complementadas por una capa de despliegue en la nube para facilitar el acceso público.

- **Backend:** Implementado en Python utilizando el framework FastAPI, el cual permite definir endpoints RESTful de forma eficiente y modular. El modelo Whisper adaptado mediante LoRA se carga utilizando la librería transformers de Hugging Face, lo que garantiza compatibilidad con el repositorio del modelo y facilita su inferencia directa sobre entradas de audio proporcionadas por el usuario.
- **Frontend:** La interfaz de usuario ha sido desarrollada con tecnologías web estándar —HTML5, CSS3 y JavaScript— integradas con el framework Bootstrap, con el fin de asegurar una experiencia visual coherente, responsive y usable desde dispositivos móviles o de escritorio.
- **Despliegue:** La aplicación ha sido publicada a través de Hugging Face Spaces, una plataforma que permite el alojamiento gratuito y colaborativo de aplicaciones de inferencia. Este entorno proporciona recursos computacionales suficientes para realizar pruebas interactivas en tiempo real y favorece la visibilidad y compartición del

proyecto con la comunidad investigadora.

Funcionalidad general

El usuario puede cargar un archivo de audio en formato .wav directamente desde la interfaz. La aplicación procesa automáticamente la señal mediante el modelo entrenado, genera la transcripción correspondiente y la muestra en pantalla. El diseño contempla también una previsualización de los metadatos del archivo, un registro de actividad básica y la posibilidad de comparar la salida del modelo con una referencia (en caso de disponer de ella).

Este componente práctico complementa el trabajo realizado en el entrenamiento y evaluación del modelo, proporcionando un canal de interacción tangible y demostrando la viabilidad de integrar el sistema ASR en herramientas reales orientadas a usuario.