

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA E INFORMÁTICA
CONCEPCION - CHILE



**“SISTEMA DE RECONOCIMIENTO DE VOZ A TEXTO DE
INGREDIENTES Y RECETAS DE COCINA: AUDIO SENSE”**

ALEJANDRO MANUEL VILLALOBOS MORA

**MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO EN INFORMÁTICA**

Profesor Guía: Rhoddy Angel Viveros Muñoz

Marzo – 2024

DEDICATORIA

En este proyecto, quiero dedicar unas palabras a los dos pilares más importantes de mi vida. Mamá, gracias por ser mi fuerza motivadora, mi fan número uno y mi fuerza en cada subida y bajada durante estos 4 años. Tus palabras alentadoras y tus abrazos en los momentos difíciles han sido mi mayor apoyo.

Papá, tu sabiduría y tus consejos siempre han guiado mi camino. Desde las charlas nocturnas sobre la vida hasta las risas compartidas durante las cenas, cada momento contigo ha sido invaluable. Gracias por enseñarme sobre el esfuerzo y la determinación, gracias a ti esos conceptos son pilares en mi vida.

También quiero dedicar esto a mi novia, quien ha estado conmigo en estos años, apoyándome siempre que lo he necesitado, nunca dejándome bajar los brazos y animándome a seguir adelante. Sin su apoyo, nada de esto hubiera sido posible.

Este proyecto no solo representa mi esfuerzo, sino también el suyo. Cada sacrificio, cada esfuerzo extra, ha contribuido a este logro conjunto. A través de estas líneas, quiero expresar mi profunda gratitud por ser los padres increíbles que son.

Gracias por darme la oportunidad de estudiar y por llenar mi vida con valores que van más allá de cualquier aula. Este logro es tan suyo como mío, y lo celebro con ustedes en mi corazón.

AGRADECIMIENTOS

Quiero expresar mi profundo agradecimiento a todas las personas que fueron parte fundamental de mi trayecto académico y que hicieron posible la culminación de este proyecto.

Agradezco sinceramente a mi tutor por su dedicación y paciencia incansables. Sin sus orientaciones precisas y correcciones detalladas, no habría alcanzado este logro tan anhelado. Gracias por guiarme a lo largo de este camino y por brindarme consejos que llevaré grabados en mi memoria, orientándome en mi futuro profesional.

A todos mis docentes, Son numerosos los docentes que contribuyeron a mi formación universitaria, y quiero expresar mi agradecimiento a cada uno de ellos. Gracias por transmitirme los conocimientos esenciales que hoy me permiten estar aquí.

Además, a mis compañeros, Quiero agradecer a todos mis compañeros, quienes no solo fueron colegas de clase, sino también amigos, cómplices y hermanos. Gracias por las horas compartidas, los trabajos en conjunto y las inolvidables historias vividas durante nuestro recorrido académico.

RESUMEN

Resumen—El presente trabajo aborda la falta de un sistema de reconocimiento de voz especializado en el contexto culinario, enfocado en ingredientes y recetas. Mediante técnicas de *deep learning*, específicamente redes neuronales, se desarrolla una solución para transcribir instrucciones de voz a texto preciso. Se investigan modelos de vanguardia como Wav2Vec 2.0 y *Transformer*. Luego se diseña y entrena una arquitectura neuronal sobre un conjunto de datos en español. Finalmente, se evalúa con métricas estándar. Los resultados sentarán bases para una incorporación efectiva en la app *FoodieSpal*, mejorando la experiencia gastronómica de los usuarios.

Palabras Clave—Reconocimiento de voz, Deep learning, Redes neuronales, ingredientes, Ajuste fino

ABSTRACT

Abstract— This paper addresses the lack of a speech recognition system specialized in the culinary context, focused on ingredients and recipes. Using deep learning techniques, specifically neural networks, a solution is developed to transcribe voice instructions into accurate text. State-of-the-art models such as Wav2Vec 2.0 and Transformer are investigated. Then, a neural architecture is designed and trained on a Spanish data set. Finally, it is evaluated with standard metrics. The results will lay the groundwork for an effective incorporation into the FoodieSpal app, improving the users' dining experience.

Keywords— Speech recognition, Deep learning, Neural networks, ingredients, Fine tuning

GLOSARIO

Aquí se deben colocar las siglas mencionadas en el trabajo y su explicación, por orden alfabético. Por ejemplo:

ASR: Automatic Speech Recognition (Reconocimiento automático del habla).

RNN: Recurrent Neural Networks (Redes neuronales recurrentes).

CNN: Convolutional Neural Networks (Redes neuronales convolucionales).

HMM: Hidden Markov Models (Modelos ocultos de Markov).

LSTM: Long Short-Term Memory (Memoria de largo y corto plazo).

MFCC: Mel Frequency Cepstrum Coeficients (Coeficientes cepstrales de frecuencia Mel).

WER: Word Error Rate (Tasa de error de palabras).

API: Application Programming Interface (Interfaz de programación de aplicaciones).

CTC: Connectionist Temporal Classification (Clasificación temporal conexionista)

INDICE DE CONTENIDO

RESUMEN.....	4
ABSTRACT	4
INDICE DE FIGURAS.....	8
INDICE DE TABLAS	9
INTRODUCCIÓN.....	10
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	11
1.1 La problemática central.....	11
1.2 Identificación de Stakeholders	12
1.3 Estado actual de la situación a mejorar	12
1.4 Desafíos actuales de los actores.....	12
1.5 Escenarios futuros sin solución a los desafíos.....	14
1.6 Competencia en la actualidad	15
1.7 Alcance	15
1.8 Objetivo general	16
1.9 Objetivos específicos.....	16
CAPÍTULO 2: MARCO CONCEPTUAL.....	17
2.1 Reconocimiento de voz.....	17
2.1.1 <i>Arquitectura</i> general de un sistema de reconocimiento de voz	17
2.1.2 Variables en los sistemas	19
2.1.3 Dependencia del hablante	19
2.1.4 Aplicaciones del ASR en la actualidad	20
2.2 Modelo oculto de Markov	22
2.2.1 Importancia del modelo oculto de Markov en el ASR	23
2.3 Redes neuronales.....	24
2.3.1 Redes neuronales recurrentes (RNN)	25
2.3.2 Redes convolucionales.....	25
2.3.4 Redes LSTM	26

CAPÍTULO 3: PROPUESTA DE SOLUCION	27
3.1 METODOLOGÍA.....	27
3.2 CONJUNTO DE DATOS.....	29
3.2.1 Mozilla Common Voice:	29
3.3 MODELO.....	34
3.3.1 WHISPER	35
3.3.2 AJUSTE FINO.....	38
CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN	41
4.1 Validación por parte de usuarios.....	41
CAPÍTULO 5: CONCLUSIONES.....	45
5.1 Trabajos futuros.....	45
REFERENCIAS BIBLIOGRÁFICAS	47

INDICE DE FIGURAS

FIGURA 1. ARQUITECTURA GENERAL DE UN SISTEMA DE RECONOCIMIENTO DE VOZ.....	17
FIGURA 2. ALGORITMO DE TRELLIS	23
FIGURA 3. MODELO DE RED NEURONAL ARTIFICIAL.....	24
FIGURA 4. HISTOGRAMA DE LA DURACIÓN DE CLIPS POR SUS FRECUENCIAS.....	31
FIGURA 5. PORCENTAJE DEL GÉNERO DE DONDE PROVIENEN LAS VOCES.....	33
FIGURA 6. ARQUITECTURA MODELO WHISPER	36
FIGURA 7. PUNTOS DE CONTROL Y CARACTERÍSTICAS DE LOS MODELOS WHISPER.....	37
FIGURA 8. CONVERSIÓN DE MATRIZ DE AUDIO MUESTREADA A ESPECTROGRAMA LOG-MEL. IZQ: SEÑAL DE AUDIO UNIDIMENSIONAL MUESTREADA. A LA DERECHA: ESPECTROGRAMA LOG-MEL CORRESPONDIENTE.	39
FIGURA 9. CALIFICACIONES PREGUNTA 1.....	41
FIGURA 10. CALIFICACIONES PREGUNTA 2.....	42
FIGURA 11. CALIFICACIONES PREGUNTA 3.....	42
FIGURA 12. CALIFICACIONES PREGUNTA 4.....	43

INDICE DE TABLAS

TABLA 1. MUESTRA TODAS LAS CARACTERÍSTICAS DEL CONJUNTO DE DATOS**Error! Bookmark not defined.**

TABLA 2. CANTIDAD TOTAL DE DATOS 32

TABLA 3. CANTIDAD DE DATOS UTILIZADA PARA EL AJUSTE FINO 34

INTRODUCCIÓN

La problemática central que aborda este proyecto se basa en la carencia actual de los sistemas de reconocimiento de voz especializado y preciso para ingredientes y recetas de comida en el idioma español. La falta de soluciones efectivas ha generado una necesidad insatisfecha entre los usuarios que buscan una comunicación fluida con sistemas de este tipo para expresar sus necesidades culinarias y recibir transcripciones precisas de sus instrucciones [Harjanto, 2021]

La metodología utilizada para desarrollar una solución ante estos desafíos consiste, en primer lugar, en una investigación profunda de técnicas de *deep learning* para el reconocimiento automático del habla, enfocándose en modelos basados en redes neuronales. Posteriormente, se diseñará y entrenará una arquitectura de red neuronal sobre un conjunto de datos en español. Finalmente, se evaluará el desempeño del modelo resultante con métricas estándar para determinar la precisión en la transcripción de voz a texto.

La elección de esta metodología se sustenta en la eficacia demostrada en estudios anteriores [Chandra et al., 2016]. En particular, los estudios de [Hinton et al., 2012] y [LeCun et al., 2015] han demostrado que los modelos de *deep learning* basados en redes neuronales pueden alcanzar altas tasas de precisión en el reconocimiento de voz.

Este documento está estructurado en cinco capítulos. El Capítulo 1 presenta en detalle la definición del problema, identificando a los *stakeholders* clave, el estado actual de la situación, desafíos existentes y escenarios a futuro. El Capítulo 2 expone el marco conceptual, estudiando modelos de *deep learning* relevantes al área del procesamiento de lenguaje natural y reconocimiento de voz. El Capítulo 3 explica la propuesta de solución, detallando la metodología de implementación, el conjunto de datos utilizado y el modelo elegido. Luego, el Capítulo 4 aborda la validación de la solución mediante la evaluación de métricas de precisión y una encuesta a usuarios potenciales. Finalmente, el Capítulo 5 presenta las conclusiones del trabajo, reflexionando sobre los logros alcanzados y las oportunidades de mejora futura.

CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA

1. Definición del Problema

1.1 La problemática central

El problema central que aborda mi proyecto es la falta de un sistema de reconocimiento de voz especializado y preciso para ingredientes y recetas de comida. Actualmente, no existen soluciones efectivas que permitan a los usuarios comunicarse de manera fluida y natural con un sistema de este tipo, expresando sus necesidades culinarias y recibir una transcripción precisa y comprensible de sus instrucciones relacionadas con la comida. La mayoría de las soluciones de reconocimiento de voz existentes no están específicamente diseñadas para comprender las complejidades y especificidades del lenguaje culinario, lo que lleva a errores en la transcripción y, en última instancia, a una experiencia culinaria insatisfactoria. Por lo tanto, el problema es la falta de un sistema de reconocimiento de voz altamente especializado y preciso que pueda entender y transcribir de manera confiable las instrucciones relacionadas con ingredientes y recetas de comida, lo que limita la eficiencia y la satisfacción en la cocina.

El objetivo principal de este proyecto es resolver esta deficiencia y mejorar significativamente la comunicación entre los usuarios y el sistema en el contexto de la gastronomía, así como abordar e implementar de manera fundamental la inclusión de personas con discapacidad visual, quienes enfrentan barreras adicionales para acceder a los sistemas tradicionales de recetas, ya que no pueden leerlas. Por lo tanto, la implementación de la asistencia por voz en esta aplicación no solo facilitará el acceso a las recetas para estas personas, sino que también su autonomía en la cocina. Adicionalmente, es importante destacar la relevancia de adaptarse a las necesidades y tecnologías del mundo actual. A medida que la sociedad avanza hacia sistemas más digitales e interconectados, el tradicional sistema de recetas impresas o consultadas a través de libros o sitios puede significar una complicación a la hora de cocinar. Por lo tanto, esta aplicación no busca solo satisfacer las demandas de los usuarios actuales en términos de eficiencia y accesibilidad, sino también mantenerse al día con las necesidades de las personas en relación con las tendencias tecnológicas emergentes. En resumen, la problemática abordada por este proyecto se centra en la necesidad de un sistema de reconocimiento de la voz altamente especializado y para ser precisos facilitar la búsqueda y seguimiento de recetas, la inclusión de personas discapacitadas visuales y adaptar a las demandas y tecnologías del mundo real.

1.2 Identificación de Stakeholders

Los usuarios clave que se beneficiarán de mi proyecto son diversos y abarcan desde cocineros casuales y entusiastas de la cocina hasta chefs profesionales y, de manera importante, personas con discapacidades visuales y/o físicas. Estos usuarios desean explorar y cocinar nuevas recetas con facilidad y precisión, independientemente de su nivel de habilidad o capacidades.

1.3 Estado actual de la situación a mejorar

El principal obstáculo en la forma en que se accede y se sigue las recetas. Tanto en los motores de búsqueda como en los libros de receta tradicionales, puesto que los usuarios generalmente deben leer las recetas en pantalla, lo que se puede hacer complicado y poco práctico, especialmente cuando están en medio de la preparación de una receta. Este método no solo requiere que el usuario tenga una pantalla visible, sino que puede hacer difícil de manejar si las manos están ocupadas con los ingredientes y utensilios de cocina.

Por otro lado, las aplicaciones móviles son cada vez más habituales en la vida cotidiana de las personas. Hemos observado en el mercado que una aplicación móvil orientada a tareas que no consigue satisfacer las necesidades no instrumentales de la gente, como la diversión y el placer, sigue teniendo probabilidades de fracasar en el mercado [Feng et al., 2019]. La satisfacción está estrechamente relacionada a la "Experiencia de usuario" concepto que incluye todos los aspectos de cómo interactúan los individuos con un producto, que depende en gran medida del contexto y es dinámico. Es por esto por lo que la implementación de los asistentes de voz significaría una importante solución frente a esta problemática por su capacidad para facilitar las interacciones entre el ser humano y el ordenador o celular de forma natural e intuitiva, similar a las conversaciones entre seres humanos. Una encuesta reciente indica que el 52% de la gente prefiere usar un asistente de voz a un sitio web para buscar información, porque les parece más cómodos. Otro 48% afirma que utilizan los asistentes de voz porque les permiten realizar varias tareas a la vez y trabajar con las manos libres y trabajar con las manos libres [Statista, 2018]. Paralelamente, con la situación actual existe complejidad para desarrollar las actividades culinarias producto de la falta de fluidez que genera tomar un libro o dispositivo para seguir las instrucciones de la receta que se desea realizar.

1.4 Desafíos actuales de los actores

La interacción con el texto durante la preparación de comidas representa un desafío práctico y cognitivo para todos los usuarios, ya que puede causar distracciones y afectar la precisión en la ejecución de las recetas. Además, la naturaleza práctica y activa de la cocina complica aún más la tarea de leer o escribir simultáneamente, especialmente cuando las manos están ocupadas con la manipulación de ingredientes y utensilios. Esta dificultad puede generar

frustración y afectar la experiencia culinaria en general. Este contexto se vuelve aún más complejo para las personas con discapacidades visuales y/o físicas, quienes enfrentan desafíos adicionales al acceder a la información culinaria de manera convencional. La dependencia de la lectura visual o la escritura puede limitar o imposibilitar su participación en la cocina, disminuyendo su autonomía y capacidad culinaria.

La Ley Núm. 20.422, establece normas sobre igualdad de oportunidades e inclusión social de personas con discapacidad, define a una persona con discapacidad como: "aquella que teniendo una o más deficiencias físicas, mentales, sea por causa psíquica o intelectual, o sensoriales, de carácter temporal o permanente, al interactuar con diversas barreras presentes en el entorno, ve impedida o restringida su participación plena y efectiva en la sociedad, en igualdad de condiciones con las demás [Ley Chile, 2010]. Por consiguiente, la discapacidad es un fenómeno complejo, donde la interacción entre las características del individuo y las de su entorno juega un papel crucial.

Cocinar es una actividad vital pero difícil para las personas con discapacidad visual ya que implica varias tareas relacionadas con la visión, como localizar objetos, verter, medir ingredientes, preparar, limpiar e interactuar con los electrodomésticos de la cocina [Bilyk et al., 2009]. Algunas tareas de cocina pueden incluso plantear riesgos de seguridad, por ejemplo, la manipulación de utensilios afilados y electrodomésticos calientes [Kashyap, 2020]. Para simplificar la tarea de cocinar, muchas personas con discapacidad visual preparan alimentos congelados o preelaborados ricos en calorías o eliminan por completo la cocina comiendo con más frecuencia en restaurantes. Como resultado, tienden a tener un estado nutricional deficiente [Jones et al., 2019] [Montero, 2005] o a tener sobrepeso debido a un uso de los restaurantes superior a la media.

Por la dificultad que presenta la cocina para las personas con discapacidad y la insuficiencia de recursos que generen oportunidades de inclusión en este ámbito se ha prestado poca atención a la cocina, una actividad importante pero difícil para la vida básica. Se ha demostrado que cocinar y preparar la comida es uno de los principales retos a los que se enfrentan las personas con discapacidad visual en su vida diaria [Cimarolli et al., 2012]. Salminen y Karhula [Salminen et al., 2014] entrevistaron a 14 jóvenes con discapacidad visual y a 22 padres para investigar sus desafíos y descubrieron que la preparación de la comida era uno de los principales desafíos y que la preparación de la comida era uno de los principales desafíos. Esta situación refleja la necesidad de abordar las barreras que enfrentan en su entorno para lograr una participación plena y efectiva en la sociedad, así como para mejorar su calidad de vida y bienestar general.

En este contexto, la implementación de una solución tecnológica que permita el acceso a recetas e ingredientes a través de comandos de voz se presenta como una opción innovadora y prometedora. Esta solución eliminaría la barrera de la lectura y escritura, proporcionando una experiencia más fluida y accesible para todos los usuarios, aquellos con discapacidades visuales y/o físicas. De esta manera, se promovería la inclusión y se facilitaría

la participación en la cocina, mejorando la autonomía y la satisfacción de todos los involucrados.

1.5 Escenarios futuros sin solución a los desafíos

La falta de soluciones efectivas para abordar los desafíos actuales en la cocina podría tener una serie de consecuencias negativas tanto para las personas con discapacidad como para los usuarios comunes sin discapacidad. Esta carencia de medidas innovadoras podría intensificar las dificultades existentes y limitar la calidad de vida y la inclusión social de ambos grupos. Para las personas con discapacidad visual, la persistencia de barreras en el acceso a la información culinaria podría tener un impacto significativo en su salud y bienestar. Según Salminen y Karhula (2015) [Salminen et al., 2015], la preparación de alimentos es uno de los principales desafíos que enfrentan las personas con discapacidad visual en su vida diaria. La incapacidad para participar plenamente en la preparación de alimentos podría aumentar el riesgo de dependencia de opciones alimentarias poco saludables, como alimentos congelados o preelaborados, lo que podría contribuir a problemas de salud física y mental. Además, la falta de acceso equitativo a opciones alimentarias saludables podría exacerbar las disparidades de salud existentes. Según datos del Centers for Disease Control and Prevention [Cdc, 2019], las personas con discapacidad tienen un mayor riesgo de sufrir obesidad en comparación con la población general. La dependencia de opciones alimentarias poco saludables podría aumentar aún más este riesgo y contribuir a problemas de salud crónicos, como la diabetes y las enfermedades cardíacas. Por otro lado, para los usuarios comunes sin discapacidad, la falta de soluciones innovadoras para mejorar la accesibilidad y la eficiencia en la cocina también podría tener consecuencias negativas en su salud y bienestar.

Según [Dahl et al., 2019], la experiencia culinaria desempeña un papel fundamental en la vida diaria de las personas, y la incapacidad para participar plenamente en actividades culinarias puede afectar su sentido de identidad y satisfacción general. Además, la falta de acceso a opciones alimentarias saludables podría contribuir a la creciente epidemia de obesidad en la población general. Según la Organización Mundial de la Salud [OMS, 2020], la obesidad es un importante factor de riesgo para una serie de enfermedades crónicas, como la diabetes tipo 2, las enfermedades cardíacas y ciertos tipos de cáncer. La dependencia de alimentos procesados y poco saludables podría aumentar el riesgo de desarrollar estas enfermedades y afectar negativamente la salud y el bienestar de la población en general. Para abordar estos desafíos, es fundamental desarrollar soluciones innovadoras y accesibles que permitan a todas las personas, independientemente de sus capacidades, participar plenamente en la experiencia culinaria y llevar un estilo de vida saludable.

1.6 Competencia en la actualidad

Si bien existen algunas soluciones de reconocimiento de voz para recetas o ingredientes en el mercado, estas soluciones a menudo no son lo suficientemente precisas ya que no son destinadas específicamente a la gastronomía. Mi proyecto busca superar estas limitaciones al proporcionar una solución superior que combine precisión, facilidad de uso y accesibilidad. La mía será especializada en ASR para ingredientes de cocina y receta en español

- Tasty: Esta aplicación ofrece una interfaz atractiva con vídeos cortos y recetas paso a paso. Si bien incorpora reconocimiento de voz, este se limita a búsquedas generales de recetas, no a la identificación de ingredientes o instrucciones específicas [Tasty, S.F].
- Cookpad: Similar a Tasty, Cookpad destaca por su comunidad de usuarios que comparten recetas e ideas. El reconocimiento de voz se limita a la búsqueda de recetas por nombre o ingrediente principal [Cookpad, S.F].
- Yummly: Con un enfoque en la personalización, Yummly permite crear colecciones de recetas y personalizar recomendaciones. El reconocimiento de voz se utiliza para buscar recetas, pero no para interactuar con las instrucciones durante la cocción [Yummly, S.F].
- Voicipe: Esta aplicación se centra en recetas de igual forma. Esta es una de las que más se parece, ya que como su nombre lo dice su enfoque es con reconocimiento de voz. El reconocimiento de voz se limita a la búsqueda de recetas por nombre o ingrediente principal en inglés [Voicipe, S.F].
- Google: Por último, y no menos importante, Google, el buscador más conocido de todos, también ofrece la posibilidad de realizar búsquedas por voz. Sin embargo, una vez realizada la búsqueda, es posible que tengas que navegar entre varios resultados para encontrar la información de la receta [Google trends, S.F].

1.7 Alcance

La misión de la aplicación es mejorar la accesibilidad y eficiencia en la cocina, especialmente para dos grupos clave de usuarios: personas con discapacidad visual y aquellos interesados en seguir recetas de cocina. Diseñada para abordar las dificultades actuales en la preparación de alimentos, la meta es proporcionar una herramienta accesible y útil para

todos los usuarios, promoviendo la inclusión de personas con discapacidad visual. Esto se traduce en una mejora en la calidad de vida y el bienestar al facilitar el acceso a opciones alimentarias y promover un estilo de vida activo y saludable.

1.8 Objetivo general

Desarrollar un sistema de reconocimiento de voz mediante el uso de inteligencia artificial que permita al usuario dictar los ingredientes y parámetros necesarios para la búsqueda y recomendación de recetas.

1.9 Objetivos específicos

- 1) Investigar las diferentes técnicas y modelos de *Deep learning* disponibles para el reconocimiento automático del habla, enfocándose en redes neuronales recurrentes (RNN), redes neuronales convolucionales (CNN), y arquitecturas basadas en Transformers como Wav2Vec2.0
- 2) Diseñar una arquitectura de red neuronal, que sea capaz de procesar eficientemente la señal de audio como entrada y transcribirla a texto.
- 3) Entrenar el modelo diseñado realizando ajuste fino en un conjunto de datos masivo en inglés.
- 4) Evaluar métricas como precisión y tasa de error de palabra.

CAPÍTULO 2: MARCO CONCEPTUAL

2.1 Reconocimiento de voz

Un sistema de reconocimiento de voz es una herramienta computacional capaz de procesar la señal de voz emitida por el ser humano y reconocer la información contenida en esta, convirtiéndola en texto o emitiendo órdenes que actúan sobre un proceso. En su desarrollo intervienen diversas disciplinas, tales como: la fisiología, la acústica, la lingüística, el procesamiento de señales, la inteligencia artificial y la ciencia de la computación.

2.1.1 Arquitectura general de un sistema de reconocimiento de voz

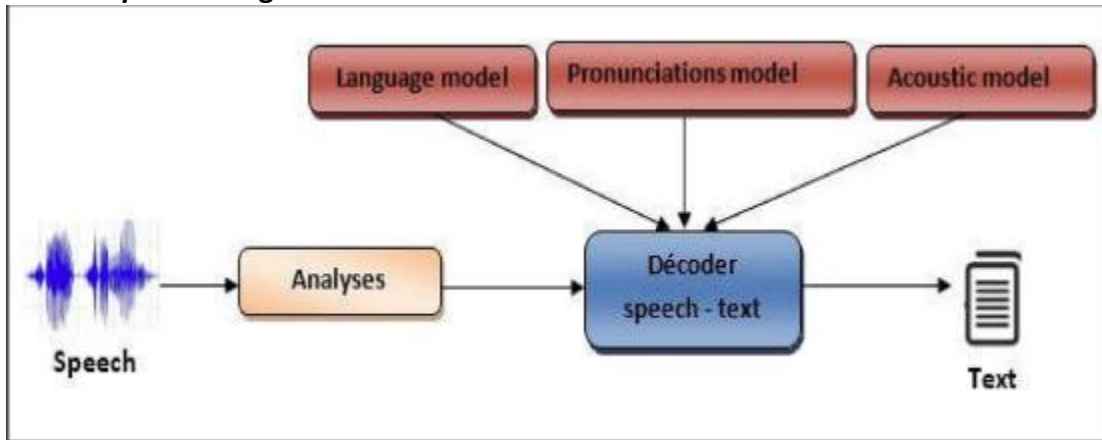


Figura 1. Arquitectura general de un sistema de reconocimiento de voz

Fuente: [Benkerzaz et al., 2019]

Las principales componentes del reconocimiento automático del habla son:

1. Análisis del habla: El ASR utiliza algoritmos y técnicas para analizar las señales de audio del habla y extraer características relevantes, como la frecuencia y la duración de los sonidos.

El análisis del habla se puede realizar de tres modos según: [Gulzar et al., 2014]

- Análisis de segmentación: Es el proceso de dividir una secuencia de habla continua en segmentos más pequeños y significativos, como palabras, sílabas o fonemas. Este análisis es esencial para el reconocimiento del habla, ya que permite identificar los límites entre las unidades lingüísticas y facilita la transcripción y comprensión del habla. El propósito de este paso es extraer la información del hablante mediante el uso de fotogramas de un tamaño de 10 a 30. [Gulzar et al., 2014]

- **Análisis sub segmental:** Se refiere a un nivel de análisis muy detallado dentro del procesamiento de señales de voz. En este contexto, se refiere al análisis de características y detalles a nivel de fonemas, sub-fonemas o incluso trazas de audio extremadamente pequeñas en un esfuerzo por mejorar la precisión del reconocimiento del habla. Esta técnica se utiliza para extraer y analizar las características del estado de excitación y para ello se utiliza una trama de tamaño de 3 a 5 ms.
 - **Análisis supra segmental:** Este último, se enfoca en aspectos prosódicos del habla, como la entonación, el ritmo, la duración de las sílabas, la velocidad del habla y la acentuación de las palabras y frases. En este paso se analiza y caracteriza el carácter conductual del hablante, ya que todos los aspectos que se analizan en este tipo pueden proporcionar información sobre el estado emocional, la intención comunicativa y otros aspectos del habla que están relacionados con el comportamiento y la comunicación del hablante. [Gulzar et al., 2014]
2. **Modelos de lenguaje:** Los modelos de lenguaje se utilizan para predecir la probabilidad de ocurrencia de palabras o frases en un determinado contexto. Estos modelos ayudan al ASR a mejorar la precisión y la comprensión del habla. [Aurora et al., 2012]
 3. **Modelos acústicos:** Los modelos acústicos se utilizan para mapear las características del habla analizadas a unidades fonéticas o palabras. Estos modelos se entrenan utilizando grandes conjuntos de datos de habla para mejorar la capacidad de reconocimiento del ASR. [Aurora et al., 2012]
 4. **Modelo de pronunciación:** Es una parte esencial de la arquitectura de reconocimiento del habla que se encarga de representar y entender cómo se pronuncian las palabras en un determinado idioma o dialecto [Aurora et al., 2012]. Este componente utiliza una variedad de reglas fonéticas y acústicas para mapear las secuencias de sonidos en el habla humana a unidades fonéticas o fonemas. Estos modelos permiten al sistema de reconocimiento de voz interpretar las palabras habladas por los usuarios y convertirlas en texto escrito. Un ejemplo de cómo se utiliza este modelo, generalmente hablando, sería el siguiente: La palabra "anillo" se descompone en unidades fonéticas o fonemas de la siguiente manera:

"a" se descompone en [a].

"ni" se descompone en [n] y [i].

"llo" se descompone en [y] y [o]

Este es solo un ejemplo simplificado, ya que, en la realidad, los modelos de pronunciación abordan muchas más variaciones y reglas fonéticas para garantizar una transcripción precisa del habla en diferentes contextos y acentos.

2.1.2 Variables en los sistemas

Existen muchas variables y características en los sistemas de reconocimiento del habla, se mencionan en [Benkerzaz et al., 2019]:

Palabras aisladas: En este tipo de habla, las palabras se pronuncian de manera individual y se separan por pausas. Es común en sistemas de reconocimiento de voz que requieren comandos específicos, como los asistentes virtuales.

Palabras conectadas: En este tipo de habla, las palabras se pronuncian de manera continua, sin pausas entre ellas. Es más natural y se asemeja al habla cotidiana. Los sistemas de reconocimiento del habla deben ser capaces de reconocer y separar las palabras en este tipo de habla.

Habla continua: Este tipo de habla implica una conversación fluida y continua, sin pausas significativas entre las palabras. Es más desafiante para los sistemas de reconocimiento del habla, ya que deben ser capaces de reconocer y transcribir el habla en tiempo real.

Habla espontánea: El habla espontánea se refiere a la forma natural de hablar, sin guiones o preparación previa. Puede incluir repeticiones, interrupciones y cambios en el ritmo y la entonación. Los sistemas de reconocimiento del habla deben ser capaces de lidiar con la variabilidad y la falta de estructura en este tipo de habla.

2.1.3 Dependencia del hablante

Como bien es sabido, el reconocimiento de voz se ha convertido en una herramienta fundamental en la interacción natural entre humanos y máquinas. Esta tecnología, que permite a los dispositivos interpretar y responder al lenguaje hablado, ha abierto un sinfín de posibilidades en diversos campos, desde la asistencia personal hasta el control de dispositivos domóticos.

Los algoritmos que se utilizan para el reconocimiento de voz tienen diferentes necesidades y características dependiendo lo que se quiera lograr, por lo tanto, existen 3 tipos de dependencia del hablante según [Benkerzaz et al., 2019]:

1. Los sistemas dependientes del hablante requieren que el usuario entrene el sistema en función de su voz. Ofrecen mayor precisión en el reconocimiento de voz y pueden adaptarse a las características específicas del habla del usuario [Huang et al., 1993]. Sin embargo, necesitan un proceso de entrenamiento previo y no son tan flexibles como los sistemas independientes del hablante.

2. Por otro lado, los sistemas independientes del hablante están diseñados para cualquier hablante, aunque normalmente se centran en un solo acento/dialecto. No requieren entrenamiento previo de cada hablante y son más flexibles. Sin embargo, su precisión puede ser menor que la de los sistemas dependientes del hablante, y no pueden adaptarse a las características específicas del habla del usuario.
3. Los sistemas adaptables al hablante ofrecen un equilibrio entre precisión y flexibilidad. Desarrollados para ajustarse a las características de nuevos hablantes, pueden adaptarse a las particularidades del habla del usuario. Aunque pueden ser más complejos que los sistemas dependientes e independientes del hablante [Huang et al., 1993].

La elección entre estos tipos de sistemas depende de las necesidades específicas del algoritmo. Si se valora la precisión y se está dispuesto a invertir tiempo en el entrenamiento, los sistemas dependientes del hablante son la mejor opción. Para aquellos que necesitan una solución rápida y no les importa una precisión ligeramente menor, los sistemas independientes del hablante pueden ser adecuados. Mientras tanto, los sistemas adaptables al hablante ofrecen una opción intermedia que combina precisión y flexibilidad

2.1.4 Aplicaciones del ASR en la actualidad

1. Asistentes Virtuales:

- Los asistentes virtuales, como Siri de Apple, Google Assistant y Amazon Alexa, utilizan el reconocimiento automático de voz para comprender las órdenes y preguntas habladas por los usuarios. Esto permite a las personas realizar una amplia variedad de tareas, como configurar alarmas, enviar mensajes y controlar dispositivos domésticos, simplemente mediante el habla. [Dxwand, s.f]

2. Sistemas de Dictado:

- En el ámbito profesional y personal, el ASR se utiliza en sistemas de dictado que permiten a las personas convertir su voz en texto de manera eficiente. Esto es especialmente útil para la transcripción de notas, documentos y correos electrónicos, lo que ahorra tiempo y esfuerzo en la escritura manual [Microsoft, 2019].

3. Servicios de Voz en la Nube:

- Varias empresas ofrecen servicios de reconocimiento de voz en la nube que permiten a los desarrolladores integrar esta tecnología en aplicaciones y servicios. Esto es valioso para aplicaciones de atención al cliente, transcripción automática, traducción de voz y más [Kimura et al., 2019].

4. Control de Dispositivos:

- El ASR se utiliza en sistemas de control de dispositivos habilitados para voz, como televisores, automóviles y electrodomésticos inteligentes. Los usuarios pueden dar comandos de voz para cambiar canales, ajustar la temperatura o realizar otras acciones sin necesidad de usar un control remoto o una interfaz táctil [Microsoft, 2019].

5. Accesibilidad y Ayuda a Personas con Discapacidades:

- Para personas con discapacidades visuales o motoras, el ASR proporciona una forma fundamental de interactuar con la tecnología. Permite la navegación en línea, la escritura de mensajes y el control de dispositivos con solo la voz, lo que mejora significativamente la calidad de vida y la inclusión [Benito, 2022], [Chaparro, 2021].

6. Transcripción y Análisis de Datos de Audio:

- El ASR es esencial en aplicaciones de transcripción automática de grabaciones de audio, como llamadas de servicio al cliente, entrevistas, conferencias y más. También se utiliza en el análisis de datos de voz para obtener información valiosa en áreas como la atención médica, la atención al cliente y la investigación. [Bokhove et al., 2018]

En resumen, su capacidad para convertir el habla en texto ha abierto nuevas puertas en la interacción entre humanos y máquinas, brindando un enfoque más natural y accesible para una variedad de tareas y servicios.

2.2 Modelo oculto de Markov

El Modelo Oculto de Markov (HMM) es un modelo estadístico utilizado para describir sistemas estocásticos que evolucionan a lo largo del tiempo, donde el estado actual depende principalmente del estado anterior. La tarea fundamental consiste en determinar los parámetros ocultos a partir de los parámetros observados. Según [Blunsom, 2004] la definición formal de un HMM es la siguiente:

$$\lambda = (A, B, \pi) \quad (1)$$

S es nuestro conjunto alfabético de estados, y V es el conjunto alfabético de observaciones:

$$S = (s_1, s_2, \dots, s_N) \quad (2)$$

$$V = (v_1, v_2, \dots, v_M) \quad (3)$$

Definimos Q como una secuencia de estados fija de longitud T, y las observaciones correspondientes O:

$$Q = q_1, q_2, \dots, q_T \quad (4)$$

$$O = o_1, o_2, \dots, o_T \quad (5)$$

A es una matriz de transición que almacena la probabilidad de que el estado j siga al estado i. Obsérvese que las probabilidades de transición de estado son independientes del tiempo:

$$A = [a_{ij}] , a_{ij} = P(q_t = s_j | q_{t-1} = s_i) . \quad (6)$$

B es la matriz de observación, que almacena la probabilidad de que la observación k se produzca a partir del estado j, independientemente de t:

$$B = [b_i(k)] , b_i(k) = P(x_t = v_k | q_t = s_i) . \quad (7)$$

π es la matriz de probabilidad inicial:

$$\pi = [\pi_i] , \pi_i = P(q_1 = s_i) . \quad (8)$$

El modelo parte de dos supuestos. La primera, denominada hipótesis de Markov, establece que el estado actual depende únicamente del estado anterior, esto representa la memoria del modelo:

$$P(q_t | q_{t-1} \dots q_1) = P(q_t | q_{t-1}) \quad (9)$$

La hipótesis de independencia establece que la observación de salida en el tiempo t depende sólo del estado actual, es independiente de observaciones y estados anteriores:

$$P(o_t | o_{t-1} \dots o_1, q_1 \dots q_t) = P(o_t | q_t) \quad (10)$$

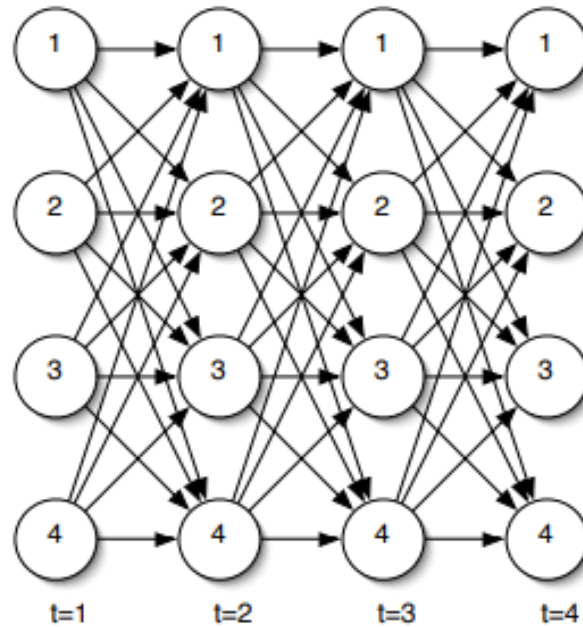


Figura 2. Algoritmo de *trellis*

Fuente: [Blunsom, 2004]

2.2.1 Importancia del modelo oculto de Markov en el ASR

El HMM desempeña un papel fundamental en el reconocimiento de voz por varias razones clave:

- **Modelado de la Variabilidad en el Habla:** El habla humana es altamente variable debido a diferencias en la pronunciación, el ritmo y el acento entre las personas. El HMM es eficaz para modelar esta variabilidad, ya que puede representar la probabilidad de que un sonido o palabra específica ocurra en un contexto particular.
- **Secuencias de Sonidos:** El habla se compone de secuencias de sonidos que evolucionan a lo largo del tiempo. Los HMM son especialmente adecuados para modelar secuencias temporales, lo que los convierte en una herramienta valiosa para capturar cómo las palabras y los sonidos se relacionan en el tiempo.
- **Adaptación a Acentos y Estilos de Habla:** Los HMM se pueden entrenar para adaptarse a diferentes acentos y estilos de habla. Esto es crucial en aplicaciones de reconocimiento de voz donde los usuarios pueden hablar de manera variada.

- **Capacidad para Tratar Datos Incompletos:** En situaciones de reconocimiento de voz en tiempo real, es común que no se disponga de toda la información al principio de una oración. Los HMM son capaces de manejar datos incompletos y actualizar continuamente las probabilidades a medida que se escucha más habla.
- **Historia en la Tecnología de ASR:** Los HMM han sido una técnica central en el desarrollo histórico del reconocimiento automático de voz (ASR, por sus siglas en inglés). Aunque se han desarrollado otros enfoques, los HMM siguen siendo una parte integral de muchas soluciones de ASR efectivas.

2.3 Redes neuronales

Una red neuronal es una estructura de procesamiento de información inspirada en la biología, diseñada para realizar tareas de aprendizaje automático y reconocimiento de patrones. Está compuesta por neuronas artificiales interconectadas que toman decisiones basadas en la información que reciben de sus vecinas. Se caracterizan por tener una capa de entrada, cero o más capas ocultas, y una capa de salida.

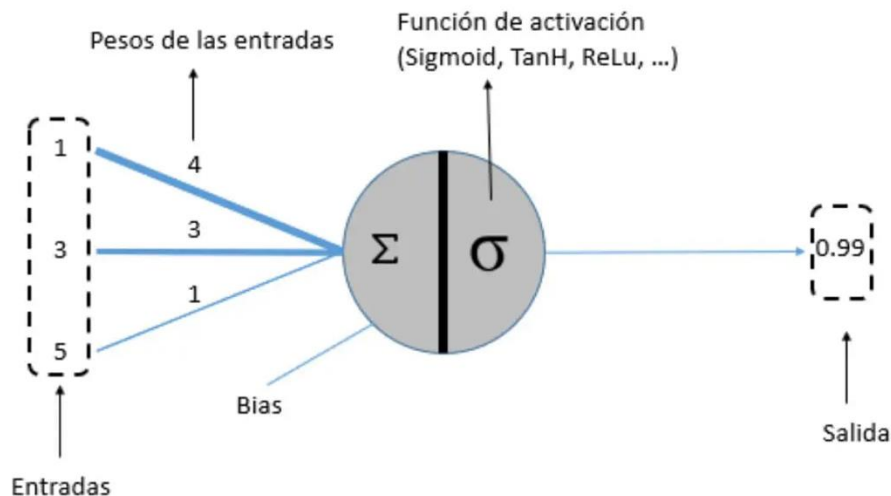


Figura 3. Modelo de red neuronal artificial

Fuente: <https://www.xeridia.com>

2.3.1 Redes neuronales recurrentes (RNN)

Redes neuronales recurrente (RNN): Una RNN es un tipo especializado de red neuronal que se destaca por su capacidad para manejar secuencias de datos y modelar relaciones temporales [IBM]. A diferencia de las redes neuronales convencionales, las RNN tienen conexiones retroactivas que les permiten mantener y utilizar información anterior en su procesamiento. Esto las hace excepcionalmente adecuadas para tareas que involucran datos secuenciales, como el procesamiento de lenguaje natural, el reconocimiento de voz y la predicción de series temporales.

Uno de los desafíos que enfrentan las RNN es el problema conocido como "*short-term memory*" o "memoria a corto plazo" [Hochreiter y Schmidhuber, 1997]. Este problema se deriva del fenómeno bien conocido en el campo del aprendizaje profundo, que es el "desvanecimiento del gradiente".

Desvanecimiento del Gradiente en RNN:

El desvanecimiento del gradiente es una limitación que afecta a las RNN durante el entrenamiento. Ocurre cuando las actualizaciones de los pesos en las capas ocultas de la red son tan pequeñas que los valores de los gradientes se vuelven insignificantes. Esto sucede especialmente en RNN largas o profundas, donde la información debe propagarse a través de múltiples pasos de tiempo.

Cuando los gradientes se desvanecen, la red tiene dificultades para aprender y retener relaciones a largo plazo en los datos secuenciales. Esto resulta en una "memoria a corto plazo", donde la red se enfoca en información reciente y pierde la capacidad de recordar eventos o patrones importantes que ocurrieron en el pasado distante de la secuencia.

2.3.2 Redes convolucionales

Son una categoría de redes neuronales profundas diseñadas específicamente para procesar y analizar datos de tipo gráfico, como imágenes y videos. Estas redes se destacan por su capacidad para aprender automáticamente características y patrones relevantes en los datos, lo que las hace ideales para tareas de visión por computadora, reconocimiento de imágenes y análisis de contenido visual.

La convolución se basa en el concepto de filtros o núcleos convolucionales que se aplican de manera repetida a regiones pequeñas de la imagen de entrada.

Estas consisten en tomar "grupos de píxeles cercanos" de la imagen de entrada e ir operando matemáticamente (producto escalar) contra una pequeña matriz que se llama *kernel*. Ese *kernel* supongamos de tamaño 3x3 píxeles "recorre" todas las neuronas de

entrada (de izquierda-derecha, de arriba-abajo) y genera una nueva matriz de salida, que en definitiva será nuestra nueva capa de neuronas ocultas [AprendeMachineLearning, s.f]

2.3.4 Redes LSTM

Son un tipo especializado de red neuronal recurrente (RNN) diseñado para abordar el desafío de capturar y aprender relaciones a largo plazo en secuencias de datos. Las LSTM se destacan por su capacidad para retener y gestionar información de manera efectiva a lo largo del tiempo, lo que las hace ideales para tareas que involucren secuencias, como el procesamiento de lenguaje natural, la traducción automática, el reconocimiento de voz y la predicción de series temporales.

Lo que hace que las LSTM sean únicas en comparación con las RNN tradicionales es su incorporación de un mecanismo especializado que incluye celdas de memoria y puertas de control. Este mecanismo les permite recordar información de manera selectiva y actualizarla conforme avanza una secuencia de datos. Este enfoque resuelve el desafío conocido como "desvanecimiento del gradiente", lo cual significa que las LSTM pueden aprender y reconocer relaciones a largo plazo en los datos [Hochreiter et al., 1997]. Esto, en última instancia, convierte a las LSTM en una herramienta excepcionalmente poderosa en el ámbito del aprendizaje automático y la inteligencia artificial.

CAPÍTULO 3: PROPUESTA DE SOLUCION

La propuesta plantea la utilización de un modelo pre-entrenado adaptado a las particularidades del problema, con el fin de llevar a cabo un ajuste fino (*fine-tuning*) utilizando datos en español. Se pretende que este modelo sea de tamaño reducido y eficiente, buscando un equilibrio entre la estructura base del modelo y las modificaciones resultantes del proceso de *fine-tuning*.

El ajuste fino implica entrenar el modelo con un conjunto de datos específicos, en este caso, un conjunto que contenga palabras y frases relacionadas con el español. Este proceso posibilita que el modelo se adapte a las características particulares de dicho conjunto de datos, mejorando así su rendimiento en la tarea específica deseada.

Tanto la elección del modelo como el proceso de entrenamiento mediante *fine-tuning* se busca a que ambas instancias sean gestionadas a través de la arquitectura 'Transformers'. Esta última es una forma de modelo de aprendizaje profundo, específicamente una arquitectura de red neuronal diseñada principalmente para abordar tareas de procesamiento del lenguaje natural

La implementación del sistema se estructura en las siguientes etapas:

3.1 METODOLOGÍA

Metodología de Implementación:

1. **Colección de Datos:** La primera fase implica la recopilación de un conjunto de datos que contenga palabras y frases asociadas al idioma español. Estos datos pueden ser obtenidos de diversas fuentes, como pueden ser *Mozilla-Foundation common voice* [Mozilla, s.f] u *OpenSLR* [OpenSLR, s.f] que son base de datos de audios en diferentes idiomas.
2. **Preparación de Datos:** La segunda etapa se centra en la preparación del conjunto de datos para el entrenamiento del modelo. Este proceso engloba actividades como la limpieza de datos, *tokenización* y normalización, con el propósito de asegurar la coherencia y calidad de los datos.
3. **Modelo:** Esta fase se encarga de elegir el modelo base, teniendo en cuentas sus características y parámetros con los que fue entrenado anteriormente, como también la facilidad de acoplarse a los nuevos datos de entrenamiento a través del *fine-tuning*.

4. **Entrenamiento del Modelo:** Esta fase se dedica al entrenamiento del modelo base con el conjunto de datos previamente preparado. Este entrenamiento se lleva a cabo mediante un *framework* de aprendizaje automático como *PyTorch*, adaptado a las necesidades específicas del problema.
5. **Evaluación del Modelo:** La quinta etapa implica la evaluación del rendimiento del modelo en un conjunto de datos de prueba. Esta evaluación proporciona información crucial para determinar los resultados del modelo para su implementación en un entorno de producción.

En este quinto paso, la evaluación se hará con la siguiente métrica:

Wer: La métrica WER (Word Error Rate) es una medida de la precisión de un sistema de reconocimiento de voz. Se define como la proporción de palabras que se reconocen incorrectamente, dividida por el número total de palabras en el texto original [Aurora et al., 2012].

La investigación en reconocimiento del habla suele evaluar y comparar los sistemas basándose en la métrica de la tasa de error de palabra (WER)

Un WER bajo indica que el sistema de reconocimiento de voz tiene una alta precisión. Por ejemplo, un WER del 5% significa que el sistema reconoce incorrectamente el 5% de las palabras. Su fórmula matemática sería la siguiente:

$$W E R = S + D + I / N * 100.$$

Donde:

S es el número de sustituciones,

B es el número de borrados,

I es el número de inserciones,

N es el número de palabras que tiene la frase de referencia.

Un ejemplo simple sería el siguiente

Transcripción de referencia: "El cielo es azul y despejado"

Transcripción generada: "Cielo azul despejado"

Inserciones: Ninguna.

Eliminaciones: "El" (1 palabra eliminada).

Sustituciones: Ninguna.

La cantidad total de errores es 1, y la longitud total de la transcripción de referencia es 6 palabras.

$$\text{WER} = 1/6 \times 100\% \approx 16.67\%$$

3.2 CONJUNTO DE DATOS

3.2.1 Mozilla Common Voice:

Common Voice es un conjunto de datos de voz a disposición del público, alimentado por las voces de colaboradores voluntarios de todo el mundo. Quienes deseen crear aplicaciones de voz pueden utilizar el conjunto de datos para entrenar modelos de aprendizaje automático, ya que es una base de datos *Open Source*, por lo cual cuenta con muchos idiomas disponibles, hasta diferentes versiones de esta base de datos. Una explicación simple de cómo se recopilan los datos en la página es:

El proceso de recopilación de datos en *Mozilla Common Voice* es transparente y colaborativo. Los pasos clave son los siguientes:

1. **Contribución Voluntaria:** Los usuarios contribuyen voluntariamente grabaciones de su voz a través de la plataforma en línea, leen frases proporcionadas por el sistema.
2. **Verificación de Calidad:** El sistema incorpora un mecanismo de verificación por pares para garantizar la calidad de las grabaciones. Los colaboradores también pueden validar y verificar las contribuciones de otros.
3. **Licencia de Datos:** Todas las grabaciones se comparten bajo licencias abiertas y accesibles, fomentando la investigación y el desarrollo de tecnologías de reconocimiento de voz en la comunidad.

En este caso se decidió trabajar con la versión 11 del conjunto de datos, una versión que tiene algunos años ya y sus datos han sido más validados por los usuarios. A continuación, se presenta una tabla detallando el nombre y descripción de cada una de las columnas que componen el conjunto de datos de *Common Voice* versión 11 para el idioma español. Conocer el significado de cada columna nos permite comprender mejor el contenido y aprovechar de forma adecuada los datos.

Tabla 1. Características del conjunto de datos

Columna	Descripción
path	La ruta al archivo de audio mp3 que contiene la grabación
client_id	Un id único que tiene cada persona que realizó una grabación
sentence	La transcripción textual de la frase hablada
up_votes	El número de votos "positivos" que los voluntarios dieron a la grabación para indicar si la transcripción era precisa
down_votes	El número de votos "negativos" que los voluntarios dieron a la grabación para indicar si la transcripción era imprecisa
age	La edad aproximada del locutor que hizo la grabación
gender	El género del locutor (masculino o femenino)
accent	El acento con el que habla el locutor
locale	Localidad del orador
segment	Normalmente es vacío, no significa nada

Fuente: elaboración propia

Antes de analizar en detalle las características del conjunto de datos de audio, es necesario e interesante estudiar uno de los aspectos más básicos: la duración de las grabaciones de voz. En particular, queríamos entender la variabilidad y distribución de la longitud de los clips, ya que esto podría tener implicaciones en el preprocesamiento y compatibilidad con el modelo en el que se realizará el ajuste fino.

La Figura 4 muestra un histograma que resume visualmente este análisis. El eje horizontal representa rangos de duración en segundos, mientras que el eje vertical indica la frecuencia o cantidad de muestras que caen en cada rango.

Como podemos observar, la mayoría de las grabaciones tienen una extensión de entre 3 y 6 segundos. Sin embargo, también existe una variabilidad significativa, con algunos clips de menos de 1 segundo y otros superando los 10. Esto se ve reflejado en la forma relativamente ancha de la distribución.

Entender estos patrones de duración permite tomar mejores decisiones respecto al tamaño de los segmentos de audio para la entrada al modelo, velocidad de muestreo y otras configuraciones relevantes. Además, nos indica que deberemos implementar técnicas para lidiar con variabilidad en la extensión de las muestras de entrenamiento.

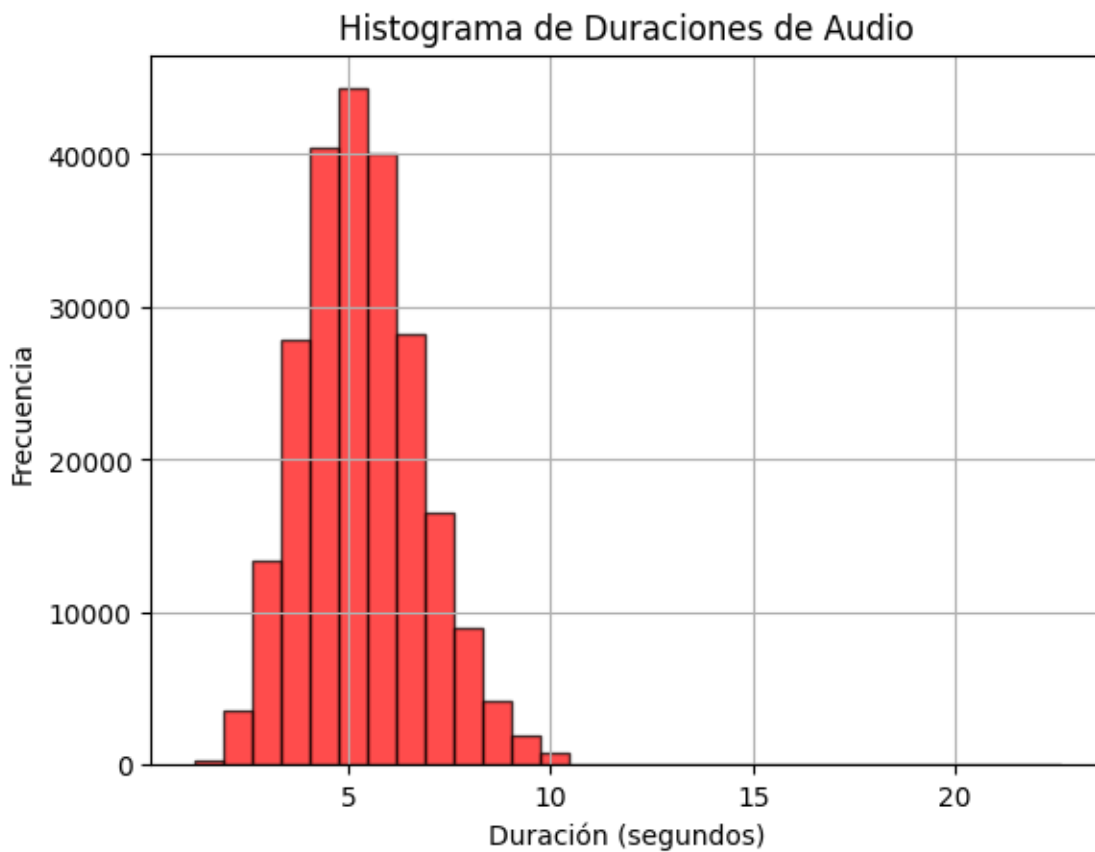


Figura 4. Histograma de la duración de clips por sus frecuencias

Fuente: Elaboración propia

Divisiones de los datos:

Mozilla Common Voice proporciona conjuntos de datos separados para entrenamiento (*train*) y evaluación (*test*) en cada idioma, para facilitar la experimentación y comparación de resultados en tareas de procesamiento de voz.

En total, para español la versión 11 contiene:

Tabla 2. Cantidad total de datos

Conjunto	Muestras
Entrenamiento	230.467
Evaluación	15.520

Fuente: Elaboración propia

Una de la característica de interés en el conjunto de datos de voz es el género de las personas que contribuyeron con las muestras. Esta información nos permite comprender la diversidad de hablantes capturada.

La Figura 5 muestra la proporción de grabaciones realizadas por locutores masculinos, femeninos y de género no especificado mediante un gráfico circular. Como se observa, aproximadamente un 59.5% de las muestras provienen de colaboradores hombres, mientras que cerca de un 24.7% fueron contribuidas por mujeres. Además, el conjunto de datos se caracteriza por una variada contribución de diferentes "*client_id*", es decir, diferentes locutores que grabaron las muestras. En total son 4275 personas únicas.

Asimismo, un pequeño porcentaje (15.8%) de las muestras corresponde a casos donde no se pudo identificar el género del hablante. Si bien no es la mayoría, esto podría indicar margen para mejorar la recolección de metadatos. En general, la distribución encontrada brinda un panorama inicial para caracterizar la variedad de voces cubiertas en nuestro conjunto de datos de audio para el entrenamiento de modelos de reconocimiento de voz.

Esta diversidad facilitará que el sistema desarrolle una representación integral del habla humana, y no específicamente la de un grupo demográfico particular. En consecuencia, el modelo entrenado estará mejor capacitado para generalizar en la etapa de implementación, extrayendo patrones robustos aplicables para transcribir precisa y equitativamente la voz de nuevos usuarios.

En contraste, una falta de variedad en los datos de entrenamiento podría sesgar al modelo dificultando su adaptación a diferentes timbres, tonos o acentos. Por ello, la distribución de géneros hallada sienta bases prometedoras para un modelo inclusivo.

Porcentaje de voces hombres y mujeres

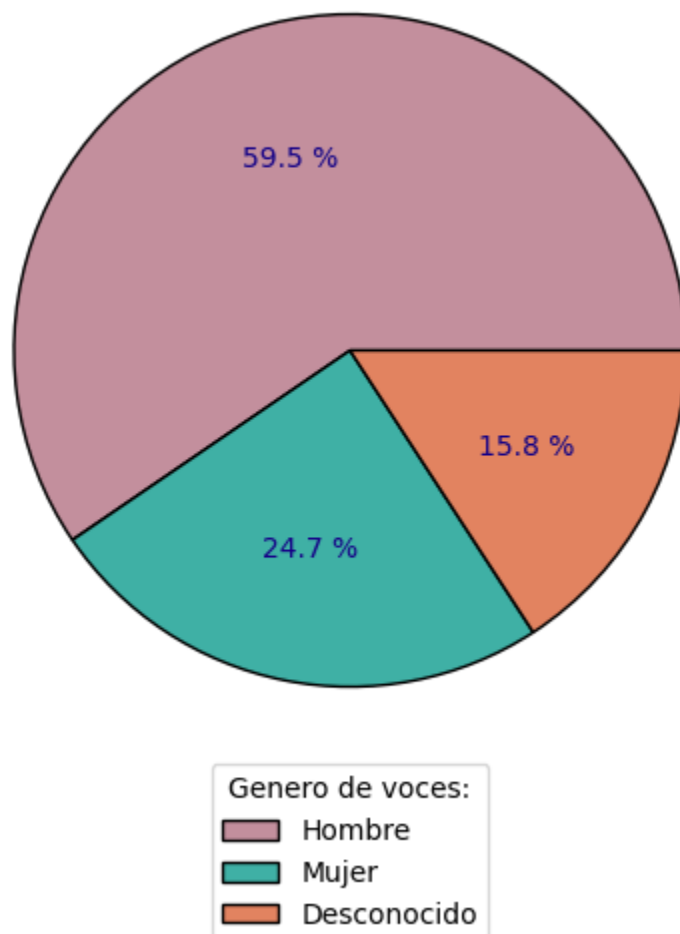


Figura 5. Porcentaje del género de donde provienen las voces

Fuente: Elaboración propia

Debido a limitaciones de tiempo y recursos computacionales, en este trabajo sólo se podrá utilizar un subconjunto de los datos totales disponibles:

Tabla 3. Cantidad de datos utilizada para el ajuste fino

Conjunto	% del Total	Muestras
Entrenamiento	15%	34.570
Evaluación	26%	4.035

Fuente: Elaboración propia

3.3 MODELO

En el corazón de la propuesta para el sistema de reconocimiento automático de voz a texto, la elección del modelo de aprendizaje profundo es un componente crítico. Este aspecto desempeña un papel esencial, ya que su responsabilidad principal radica en transformar la señal de audio de voz en una representación de texto coherente.

En esta sección se profundizará específicamente en el modelo elegido, la arquitectura *Transformer Whisper (Small)* de *OpenAI*. Se abordarán sus características, capacidades, ventajas respecto a alternativas, y el razonamiento detrás de esta elección.

Asimismo, se cubrirá cómo se entrenará *Whisper* mediante un proceso de ajuste fino sobre un *dataset* en español (Mencionado anteriormente en el punto 3.2). Se detallarán los hiperparámetros involucrados, la función de costo optimizada, y las técnicas aplicadas para mejorar el aprendizaje sobre nuestros datos específico

Por lo tanto, esta sección no solo desglosará la elección de nuestro modelo, sino que también proporcionará una visión detallada de cómo prepararemos y mejoraremos su rendimiento para que se adapte a las particularidades del idioma español y a nuestro conjunto de datos específico.

3.3.1 WHISPER

Publicado en septiembre de 2022 por OpenAI [Radford et al., 2022], Whisper es un modelo pre-entrenado revolucionario para el reconocimiento automático del habla (ASR). A diferencia de modelos previos como Wav2Vec 2.0 que se pre-entrenan solo con datos de audio sin etiquetar, Whisper aprovecha 680,000 horas de datos de transcripción de audio etiquetados para su preentrenamiento, incluyendo 117,000 horas de datos ASR multilingües. Esto permite entrenar Whisper desde el inicio en la tarea supervisada de reconocimiento de voz a texto, generando puntos de control para más de 96 idiomas.

Esta cantidad de datos etiquetados permite *pre-entrenar* a *Whisper* directamente en la tarea supervisada de reconocimiento del habla, aprendiendo un mapeo de habla a texto a partir de los datos de preentrenamiento de audio-transcripción etiquetados. En consecuencia, Whisper requiere pocos ajustes adicionales para obtener un modelo ASR eficaz.

Whisper es un modelo codificador-decodificador basado en *Transformer*, también conocido como modelo secuencia a secuencia. Asigna una secuencia de características de espectrograma de audio a una secuencia de tokens de texto. En primer lugar, las entradas de audio en bruto se convierten en un espectrograma log-Mel mediante la acción del extractor de características. El espectrograma log-Mel es una representación visual y matemática de las características del sonido en función del tiempo y la frecuencia. Es una herramienta comúnmente utilizada en el procesamiento de señales de audio y en el reconocimiento automático del habla [Xu et al., 2004].

A continuación, el codificador *Transformer* codifica el espectrograma para formar una secuencia de estados ocultos del codificador. Por último, el decodificador predice auto regresivamente los tokens de texto en función de los tokens anteriores y de los estados ocultos del codificador.

Hablando un poco más sobre el decodificador, podemos decir que este desempeña el papel de un modelo lingüístico, procesando las representaciones de estado oculto y generando las transcripciones de texto correspondientes como se mencionó en el párrafo anterior. La incorporación de un modelo lingüístico interno en la arquitectura del sistema se denomina fusión profunda. Esto contrasta con la fusión superficial, en la que un modelo de lenguaje se combina externamente con un codificador, como ocurre con *CTC*. Con la fusión profunda, todo el sistema puede entrenarse de extremo a extremo con los mismos datos de entrenamiento y la misma función de pérdida, lo que proporciona una mayor flexibilidad y un rendimiento generalmente superior. La figura 6 resume el modelo *Whisper*.

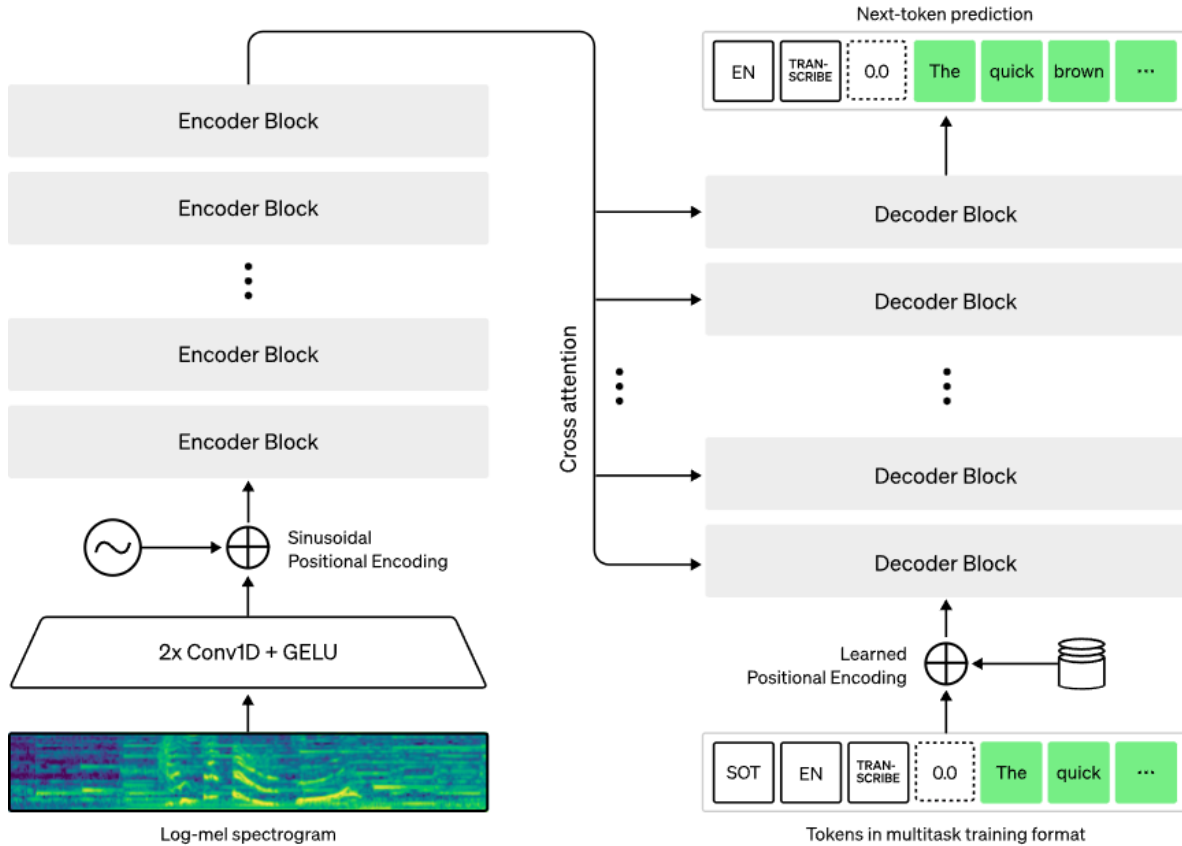


Figura 6. Arquitectura Modelo *Whisper*

Fuente: OpenAI Whisper Blog

Whisper se pre-entrena y se ajusta utilizando la función objetivo de entropía cruzada, una función objetivo estándar para entrenar sistemas de secuencia a secuencia en tareas de clasificación. En este caso, el sistema se entrena para clasificar correctamente el token de texto objetivo a partir de un vocabulario predefinido de tokens de texto.

Los puntos de control *Whisper* se presentan en cinco configuraciones de distintos tamaños de modelo. Los puntos de control se resumen en la figura número 9:

El punto de control que se eligió fue el *Whisper-small* porque es una variante más compacta que ofrece una solución de compromiso entre precisión en ASR y requerimientos computacionales, lo cual se ajusta al problema.

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Figura 7. Puntos de control y características de los modelos *whisper*

Fuente: [Radford et al., 2022]

Algunas de las razones y ventajas porque elegir un modelo de *Whisper*, es debido a su estado del arte actual y por ser de código abierto, por lo tanto, cada desarrollador podría ajustar sus datos y su problemática a el modelo, también hay ventajas de *Whisper* sobre otros enfoques existentes. A continuación, se presentará un resumen de algunas de sus principales virtudes:

1. Robustez: *Whisper* demuestra una mejor generalización y robustez fuera de distribución en comparación a otros modelos de voz cuando es evaluado sin entrenamiento fino (*fine-tuning*). Esto probablemente se deba a su extenso y diverso conjunto de datos de entrenamiento.
2. Capacidad de *Zero-shot*: A diferencia de otros modelos que requieren de un entrenamiento fino específico a la tarea, *Whisper* funciona bien en reconocimiento de voz, traducción y otras tareas sin necesidad de entrenamiento adicional en otros conjuntos de datos.
3. Escala y Diversidad: Entrenado en 680,000 horas de audio etiquetado en 117 idiomas, *Whisper* ha visto más diversidad de datos que modelos previos. Los experimentos muestran que su desempeño en varias tareas sigue mejorando conforme aumenta su escala.
4. Simplicidad: *Whisper* usa una sencilla arquitectura *Transformer* y procedimiento de entrenamiento estándar

5. Multitarea: Un solo modelo *Whisper* puede realizar reconocimiento de voz, traducción, identificación de idiomas y otras tareas de procesamiento de voz simplemente condicionando en una indicación específica de tarea.
6. Transcripción Extensa - Aprovechando la predicción precisa de marcas de tiempo y heurísticas adicionales de decodificación, *Whisper* puede transcribir grabaciones extensas de varias horas con precisión sin degradación.

3.3.2 AJUSTE FINO

En el ámbito del aprendizaje automático, el ajuste fino se utiliza para refinar un modelo pre-entrenado en conjuntos de datos generales, adaptándolo a conjuntos de datos más específicos y mejorando su capacidad de generalización. Este proceso implica modificar cuidadosamente ciertos hiper parámetros, como tasas de aprendizaje o estructuras de capas, para adaptar el modelo a las características particulares de los datos de interés, evitando así el sobreajuste o la suboptimización.

En resumen, el ajuste fino es una técnica de aprendizaje profundo que consiste en reutilizar los pesos de un modelo previamente entrenado en una nueva tarea. Esto se traduce en un ahorro de recursos computacionales y tiempo [Devlin et al., 2018]. También algunas ventajas del ajuste fino son:

- Permite aprovechar el conocimiento adquirido por el modelo en una tarea general para resolver una tarea específica
- Permite obtener mejores resultados que el aprendizaje por transferencia, ya que se adapta mejor a los datos nuevos
- Permite entrenar modelos precisos con conjuntos de datos limitados, ya que se beneficia de la regularización implícita del modelo pre-entrenado.

Teniendo más claro lo que es un ajuste fino de manera general, podemos continuar explicando cómo se creará el modelo con esta técnica de aprendizaje profundo. Se realizará de la siguiente forma:

1. Lo primero que se hará es sacar las muestras precisas del conjunto de datos con las cuales se trabajarán, como se mencionó anteriormente, se seleccionaron 34.570 de para el entrenamiento y 4.035 muestras para la evaluación.

2. Para trabajar de manera más limpia y mantener todo de la forma más general posible, eliminaremos las columnas que no nos sirven para entrenar el modelo, así quedándonos solo con la columna que contiene el audio y la frase correspondiente a ese audio
3. Con un extractor de características se va a preprocesar las entradas de audios sin procesar. El modelo que usamos es el que se encargara de que se realice el mapeo de secuencia a secuencia y por último un *tokenizador* post procesa los resultados del modelo en formato texto.

Siguiendo con la extracción de características, tenemos que el habla se representa mediante una serie unidimensional que varía en el tiempo, siendo el valor en cada paso de tiempo la amplitud de la señal en ese punto. Para abordar la naturaleza continua del habla, discretizamos la señal mediante la toma de valores en intervalos de tiempo fijos. Es crucial que coincida la tasa de muestreo del audio con la esperada por el modelo, ya que diferentes tasas de muestreo tienen distribuciones muy diferentes. *Whisper* espera tasas de muestreos de 16 kHz.

también el extractor de características *Whisper* realiza dos operaciones clave. Primero, ajusta lotes de muestras de audio para que todas tengan una longitud de entrada de 30 segundos, ya sea agregando ceros al final para muestras más cortas o truncándolas para las más largas. Segundo, convierte las matrices de audio ajustadas en espectrogramas log-Mel, representaciones visuales de las frecuencias en la señal de habla. Un ejemplo de la transformación del habla a *log-mel* espectrograma en la figura 8.

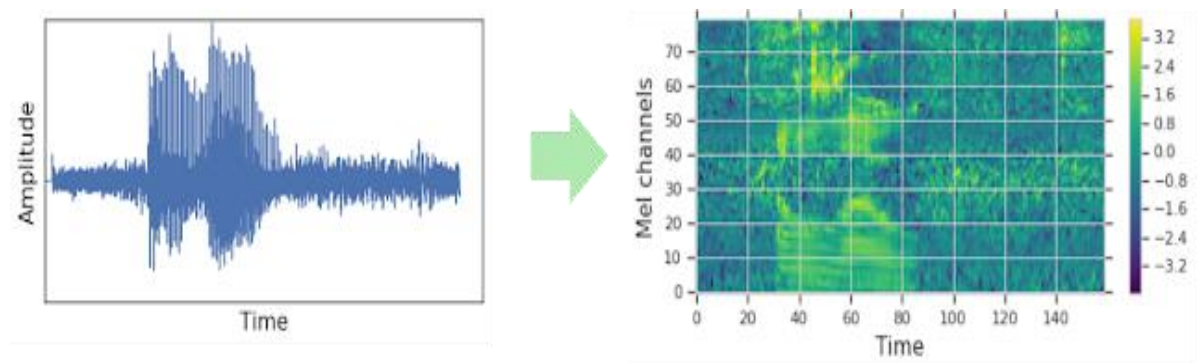


Figura 8. Conversión de matriz de audio muestreada a espectrograma *log-mel*. Izq: señal de audio unidimensional muestreada. A la derecha: espectrograma log-Mel correspondiente.

Fuente: Google SpecAugment Blog.

Luego, cargamos el *tokenizador* que tiene el propio modelo y este produce tokens de texto que indican el índice del texto predicho en el diccionario de elementos de vocabularios. El *tokenizador* asigna una secuencia de tokens de texto a la cadena de texto real (por ejemplo, [1169, 3797, 3332] -> "el gato se sentó").

4. Como sabemos, los datos de audio del conjunto de datos están muestreados en una matriz de forma unidimensional y su correspondiente transcripción. Como ya hablamos de la importancia de la frecuencia de muestreo y del hecho que tienen que coincidir el del modelo. En el conjunto de datos los datos están muestreados a 48 kHz, entonces tenemos que reducirlo a 16 kHz antes de pasarlo al extractor de características.
5. Ahora, en este punto se aplicó el *tokenizador* y el extractor de características a los datos, por lo que nuestras entradas de audio se convertirán en espectrogramas de *log-mel* y sus textos se convertirán en tokens de identificadores de etiquetas
6. Antes del ajuste fino se crearon los parámetros para llevar a cabo el entrenamiento
7. Teniendo nuestros datos transformados y los parámetros seteados, aplicamos el ajuste fino, cargando el modelo base, en este caso *Whisper-Small* donde se obtienen sus capas, parámetros y pesos.
8. Luego de haber realizado el entrenamiento, se procede a evaluar el modelo final con la métrica *WER*.

CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN

4.1 Validación por parte de usuarios

Este proyecto es diseñado para implementarlo en un proyecto más grande, llamado *FoodieSpal*. Es una aplicación que se encargara de ayudar a las personas que no saben que cocinar y recomendar recetas personalizadas para cada tipo de persona.

Con este contexto, el pasado viernes 17 de noviembre, durante la participación en la Feria de Software, se llevó a cabo una valiosa fase de validación de nuestro proyecto. Los usuarios que tuvieron la oportunidad de interactuar con nuestra aplicación compartieron sus opiniones y experiencias a través de una encuesta diseñada para evaluar diversos aspectos clave de nuestro producto.

A los usuarios se les hizo las siguientes preguntas:

1. ¿Consideras que la accesibilidad para personas con discapacidades es un aspecto crucial en un sistema de reconocimiento de voz para ingredientes y recetas?

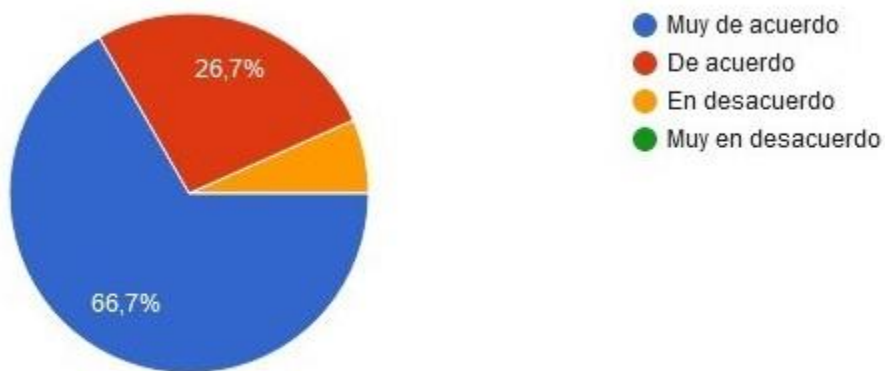


Figura 9. Calificaciones pregunta 1

Fuente: Elaboración propia

2. ¿Crees que un sistema de reconocimiento de voz basada en inteligencia artificial mejoraría significativamente tu capacidad para buscar y cocinar recetas de manera eficiente?

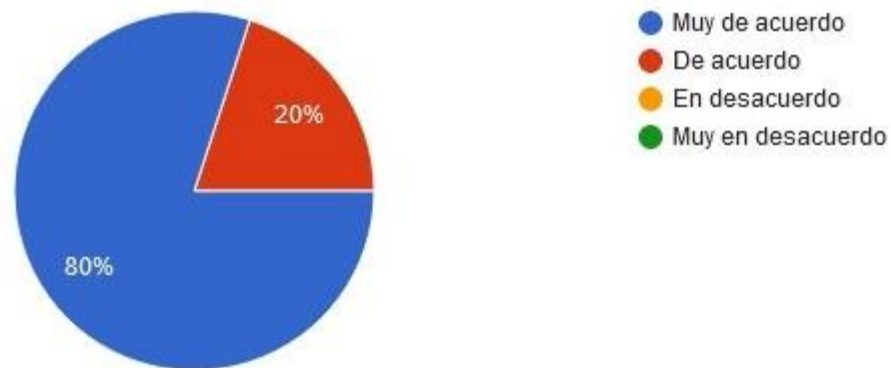


Figura 10. Calificaciones pregunta 2

Fuente: Elaboración propia

3. ¿Consideras que la falta de un sistema de reconocimiento de voz especializado y preciso para ingredientes y recetas de comida afecta negativamente tu experiencia culinaria?

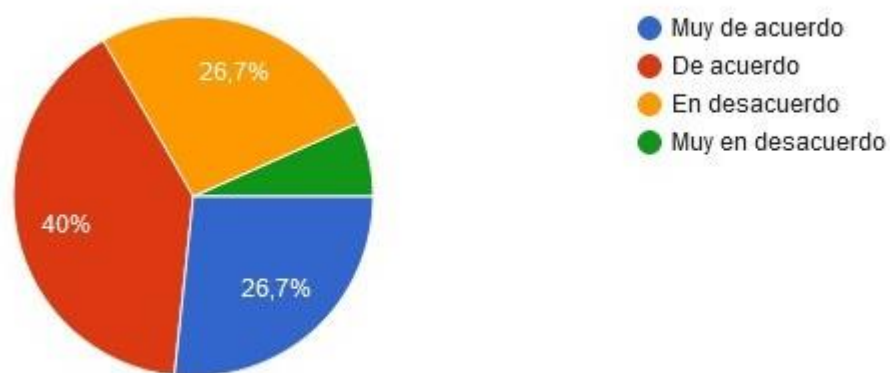


Figura 11. Calificaciones pregunta 3

Fuente: Elaboración propia

4. ¿Consideras que la necesidad de escribir mientras cocinas representa un desafío significativo para los cocineros actuales?

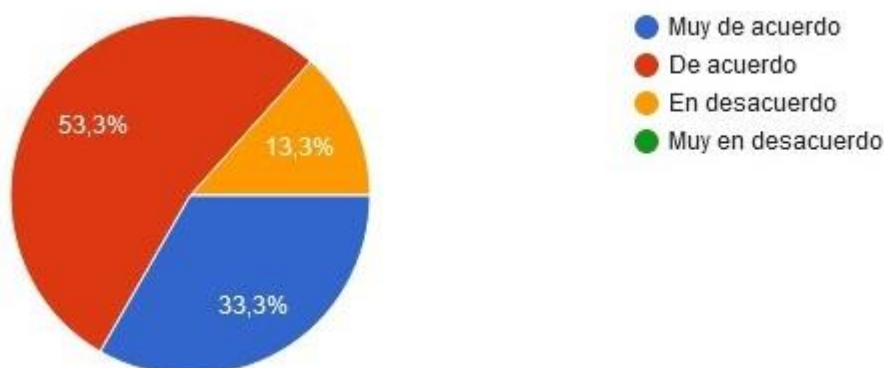


Figura 12. Calificaciones pregunta 4

Fuente: Elaboración propia

Después de realizar una encuesta a personas sobre la necesidad y el potencial de un sistema de reconocimiento de voz para ingredientes y recetas, los resultados obtenidos son altamente prometedores, teniendo en cuenta los límites de esta encuesta que fue realizada en una feria de software y que puede no ser generalizada a toda la población. La gran mayoría de los encuestados reconocen la importancia de la accesibilidad para personas con discapacidades en este tipo de sistemas, con un 93.4% estando de acuerdo o muy de acuerdo en que es un aspecto crucial.

Un hallazgo clave es que el 100% de los participantes consideran que un sistema de reconocimiento de voz basado en inteligencia artificial mejoraría significativamente su capacidad para buscar y cocinar recetas de manera eficiente. Esto respalda fuertemente la idea de que una herramienta así sería de gran valor para los cocineros.

Si bien existe una división de opiniones sobre si la falta de un sistema de reconocimiento de voz especializado afecta negativamente la experiencia culinaria actual, con un 66.7% estando de acuerdo o muy de acuerdo y un 33.3% en desacuerdo, es evidente que una parte sustancial de los encuestados ven esta carencia como un problema.

Además, el 86.6% de los participantes están de acuerdo o muy de acuerdo en que la necesidad de escribir mientras se cocina representa un desafío significativo para los cocineros actuales. Esto sugiere que un sistema de reconocimiento de voz podría abordar una necesidad importante en el ámbito culinario.

En resumen, los resultados de la encuesta respaldan firmemente el desarrollo de un sistema de reconocimiento de voz para ingredientes y recetas, destacando su potencial para mejorar la eficiencia, la accesibilidad y la facilidad de uso en la cocina. Estos hallazgos cuantitativos proporcionan una base sólida para continuar explorando e implementando esta solución en el campo culinario.

CAPÍTULO 5: CONCLUSIONES

En el desarrollo de este proyecto de reconocimiento automático de voz para ingredientes y recetas, se lograron importantes avances que sentaron las bases para un sistema funcional. Sin embargo, debido a limitaciones de tiempo y recursos, quedaron algunos puntos pendientes.

En cuanto a lo realizado, el modelo *Whisper* demostró ser una elección acertada como núcleo del sistema de reconocimiento de voz. Mediante un proceso de ajuste fino sobre el conjunto de datos en español de *Mozilla Common Voice*, se adaptó exitosamente a las particularidades de nuestro idioma. Específicamente, se realizó la selección, limpieza y preprocesamiento de un subconjunto de las muestras disponibles, las cuales luego se utilizaron en la etapa de entrenamiento.

Asimismo, técnicas como la extracción de características de audio en forma de espectrogramas log-Mel, *tokenización* y empaquetado en lotes de entrenamiento, permitieron acondicionar los datos para su procesamiento por el modelo *Transformer*. Se investigó en profundidad este tipo de arquitecturas, entendiendo conceptos clave como codificadores, decodificadores y su funcionamiento seq2seq.

De igual forma, se expandió significativamente el conocimiento sobre distintos enfoques en ASR mediante aprendizaje profundo, específicamente redes neuronales recurrentes, LSTM y técnicas de ajuste fino. Trabajar con grandes conjuntos de datos de voz en español, preprocesarlos y adaptarlos al modelo pre entrenado, brindó una experiencia práctica invaluable para aplicar los conceptos teóricos estudiados.

En cuanto a las tareas pendientes, no se llegó a implementar completamente la evaluación final del desempeño. Si bien se realizó un planteamiento inicial de la función de pérdida CTC, métrica WER para medir precisión, no se concretó esta fase por temas de Hardware. Completar el ciclo de evaluación y analizar cuantitativamente los resultados sobre el *dataset* de pruebas es esencial para terminar de validar el sistema desarrollado.

5.1 Trabajos futuros

Otro punto importante es extender las pruebas cualitativas a un grupo más diverso y numeroso de usuarios potenciales. Hasta ahora solo se realizó una evaluación preliminar durante una feria tecnológica con una pequeña muestra enfocada en el contexto de la aplicación *FoodieSpal*. Se requieren pruebas más grandes y variadas para recopilar retroalimentación desde otras perspectivas.

De cara al futuro, integrar el modelo ya entrenado y optimizado en la aplicación *FoodieSpal* sería el siguiente paso lógico. Allí podría escalarse a más usuarios y recetas, probándolo en un entorno real de uso cotidiano. Asimismo, se podrían incorporar capacidades adicionales como identificación de idioma.

Por otro lado, aplicar técnicas de aumentación de datos, compresión y normalización harían al modelo *Transformer* más eficiente, tanto en tamaño como en velocidad de inferencia. Esto facilitaría su implementación incluso en dispositivos móviles con limitantes de memoria y capacidad de cómputo. De esta manera, se cumpliría la meta de llevar la solución directamente a los cocineros en sus hogares.

En conclusión, este proyecto sentó bases sólidas y prometedoras para un sistema de ASR enfocado en el contexto culinario. Si bien quedan pendientes aspectos importantes por abordar, también se obtuvieron logros significativos en cuanto al modelo, los datos y la adaptabilidad a características lingüísticas específicas. Sobre esta plataforma ya establecida se puede seguir construyendo y expandiendo las capacidades. Las contribuciones hasta ahora apuntan a una futura integración efectiva con la aplicación *FoodieSpal*, brindando una herramienta de reconocimiento de voz para una mejor experiencia gastronómica.

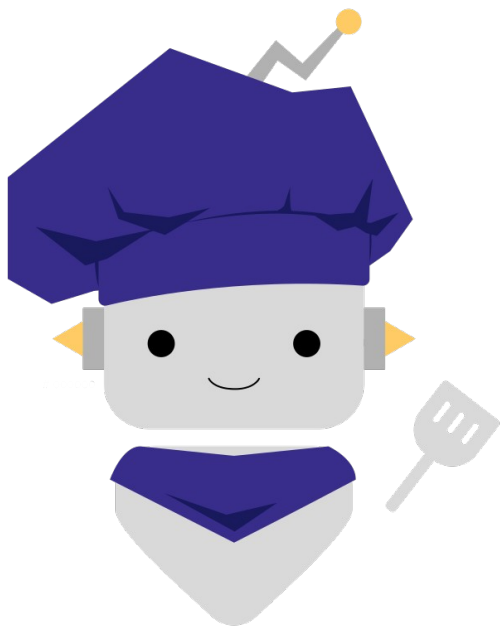
REFERENCIAS BIBLIOGRÁFICAS

1. [Benkerzaz et al., 2019] Benkerzaz, S., Elmir, Y., & Dennai, A. (2019). A Study on Automatic Speech Recognition. *Journal of Information Technology Review*.
2. [Svendsen05] Svendsen, Torbjørn. (2005). Pronunciation modeling for speech technology. En *Proceedings of the 9th International Conference on Signal Processing* (págs. 11-16). doi: 10.1109/SPCOM.2004.1458347.
3. [Gulzar et al., 2014] Gulzar, Taabish, Singh, Anand, Kumar, Dinesh, & Farooq, Najma. (2014). A Systematic Analysis of Automatic Speech Recognition: An Overview.
4. [Morales et al., 2005] Morales, Nicolas, Hansen, John, & Toledano, Doroteo. (2005). MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech. En *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vol. 1, págs. 521-524). doi: 10.1109/ICASSP.2005.1415165.
5. [Blunsom, 2004] Blunsom, P. (2004). Hidden Markov Models.
6. [IBM] IBM. (s.f.). Recurrent Neural Networks. Recuperado de [https://www.ibm.com/mx-es/topics/recurrent-neural-networks#:~:text=Una%20red%20neuronal%20recurrente%20\(RNN,voz%20y%20subt%C3%ADtulos%20de%20im%C3%A1genes](https://www.ibm.com/mx-es/topics/recurrent-neural-networks#:~:text=Una%20red%20neuronal%20recurrente%20(RNN,voz%20y%20subt%C3%ADtulos%20de%20im%C3%A1genes).
7. [Hochreiter y Schmidhuber, 1997] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.
8. [AprendeMachineLearning, s.f.] AprendeMachineLearning. (s.f.). Cómo funcionan las Convolutional Neural Networks (CNN) en visión por computadora. Recuperado de <https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/>.
9. [IBM, s.f.] IBM. (s.f.). Speech Recognition. Recuperado de <https://www.ibm.com/es-es/topics/speech-recognition>.
10. [Devlin et al., 2018] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), 4171-4186.
11. [Radford et al., 2022] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

12. [Harjanto, 2021] Harjanto. A (2021). Can voice technology make you a better cook?. Medium: 55 minutes. Recuperado de <https://medium.com/55-minutes/can-voice-technology-make-you-a-better-cook-82521755a541>
13. [Chandra et al., 2016] Chandra, E. H., & Karpagavalli, S. (2016). A review on automatic speech recognition architecture and approaches.
14. [Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. y Kingsbury, B. (2012). Deep Neural Networks for acoustic modeling in speech recognition.
15. [LeCun et al., 2015] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning
16. [Aurora et al., 2012] Shipra J. Aurora, Rishi Pal Singh. (2012). Automatic Speech Recognition: A Review
17. [Huang et al., 1993] Xuedong Huang, Kai-Fu Lee. (1993). On Speaker-Independent, Speaker-Dependent, and Speaker- Adaptive Speech Recognition
18. [Dxwand, s.f] Dxwand (s.f). What is a virtual assistant and why is it important? Rucuperado de [What Is A Virtual Assistant And Why Is It Important? \(dxwand.com\)](https://dxwand.com/What-Is-A-Virtual-Assistant-And-Why-Is-It-Important/)
19. [Microsoft, 2019] Microsoft. (2019). Voice Report: From answers to action: customer adoption of voice technology and digital assistants
20. [Kimura et al., 2019] Kimura, T., Nose, T., Hirooka, S., Chiba, Y., Ito, A. (2019). Comparison of Speech Recognition Performance Between Kaldi and Google Cloud Speech API.
21. [Benito, 2022] Benito, P. (2022). Hábitos en el uso de los asistentes virtuales de voz entre la población de personas con discapacidad visual.
22. [Chaparro, 2021] David C. M. (2021) Desarrollo de una aplicación móvil para la comunicación de personas con discapacidad con asistentes virtuales inteligentes de dispositivos IoT.
23. [Bokhove et al., 2018] Christian, B. Christopher D. (2018). Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data.
24. [Hochreiter et al., 1997] Hochreiter, S. Schmidhuber, J. (1997). LONG SHORTTERM MEMORY

25. [Mozilla, s.f] Mozilla Common Voice. (s.f). Recuperado de [Common Voice \(mozilla.org\)](https://commonvoice.mozilla.org/)
26. [OpenSLR, s.f] OpenSLR. (s.f). Recuperado de openslr.org
27. [Xu et al., 2004] Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C., & Tian, Q. (2004). HMM-Based Audio Keyword Generation.
28. [Tasty, S.F] Tasty. (S.F). Recuperado de [Tasty - Food videos and recipes](https://tasty.co/)
29. [Cookpad, S.F] Cookpad. (S.F). Recuperado de [Cookpad - Cocina casera más divertida](https://cookpad.com/)
30. [Yummly, S.F] Yummly. (S.F). Recuperado de [Yummly: Personalized Recipe Recommendations and Search](https://yummly.com/)
31. [Voicipe, S.F] Voicipe. (S.F). Recuperado de [Voicipe | Voice recipe reader app](https://voicipe.com/)
32. [Google trends, S.F] Google trends. (S.f) Recuperado de [Recetas de comida - Explorar - Google Trends](https://trends.google.com/trends/explore?q=recetas+de+comida&hl=es)
33. [Feng et al., 2019] Feng L., Wei W. (2019). *An empirical study on user experience evaluation and identification of critical UX issues*
34. [Statista, 2018] Statista. (2018). Recuperado de [Global rationale for choice of voice assistants over websites and apps 2017 | Statista](https://www.statista.com/statistics/542814/global-rationale-for-choice-of-voice-assistants-over-websites-and-apps-2017/)
35. [Ley Chile, 2010] Ley Chile 20.422 Art 5 Recuperado de [Ley Chile - Ley 20422 - Biblioteca del Congreso Nacional \(bcn.cl\)](http://www.bcn.cl/Biblioteca/LeyChile/ley-20422)
36. [Bilyk et al., 2009] Bilyk, M. C., Sontrop, J. M., Chapman, G. E., Barr, S. I., & Mamer, L. (2009). Food experiences and eating patterns of visually impaired and blind people.
37. [Kashyap, 2020] Kashyap, A. R. (2020). Behaviors, Problems and Strategies of Visually Impaired Persons During Meal Preparation in the Indian Context: Challenges and Opportunities for Design.
38. [Jones et al., 2019] Jones, N., Bartlett, H. E., & Cooke, R. (2019). An analysis of the impact of visual impairment on activities of daily living and vision-related quality of life in a visually impaired adult population.
39. [Montero, 2005] Montero, P. (2005). Nutritional assessment and diet quality of visually impaired Spanish children.

40. [Cimarolli et al., 2012] Cimarolli, V. R., Boerner, K., Brennan-Ing, M., Reinhardt, J. P., & Horowitz, A. (2012). Challenges faced by older adults with vision loss: a qualitative study with implications for rehabilitation.
41. [Salminen et al., 2014] Salminen, A-L., & Karhula, M. E. (2014). Young persons with visual impairment: Challenges of participation.
42. [Salminen et al., 2015] Salminen, A-L., & Karhula, M. E. (2015). *The challenges of young visually impaired adults in daily activities and in self-determination*.
43. [Cdc, 2019] *Centers for Disease Control and Prevention. (2019). Recuperado de [Disability and Obesity | CDC](#)*
44. [Dahl et al., 2019] Dahl, J., & Moreau, N. (2019). The impact of cooking and the importance of meaningful participation in daily life.
45. [OMS, 2020] *Organización Mundial de la Salud. (2020). World report on vision. Geneva: WHO Press*



foodies
Pal

Integrantes: Brian Spielman

Diego Lucero

Cristóbal Carrasco

Alejandro Villalobos

Claudio Nuñez

Introducción.

En el año 2023, la feria de software, un evento emblemático organizado por la Universidad Federico Santa María marcará un hito al ser llevada por primera vez en sus 31 años de historia a la Sede Concepción. Coincidiendo con este acontecimiento significativo, la primera generación de estudiantes de la carrera de Ingeniería en Informática en la Sede Concepción asumirá el desafío de presentar sus proyectos de software. En el contexto de nuestro equipo, el cual se encuentra dentro de la categoría de **Inteligencia Artificial y Ciencia de Datos** surgió la idea de desarrollar un software dirigido a personas que tienen una agenda laboral ocupada y pasan la mayor parte del día fuera de su hogar, para quienes por uno u otro motivo no pueden responder la pregunta de **el que van a cocinar el día de hoy**.

La propuesta consiste en ofrecer una solución que simplifique este proceso. El software no solo proporcionará sugerencias de comidas equilibradas y nutritivas, sino que también considerará las preferencias personales y las restricciones alimenticias. De esta manera, al llegar a casa, estas personas encontrarán una guía amigable y práctica que les permitirá tomar decisiones informadas sobre sus comidas sin complicaciones ni estrés adicional.

Con todo esto en mente nace **FoodiesPal**. Más allá de simplemente sugerir recetas basadas en los ingredientes que ya posees en casa, cuenta con un asistente el cual está diseñado para ser tu compañero en la cocina. Además de brindarte recomendaciones culinarias, guiará paso a paso durante el proceso de preparación. Pero eso no es todo, **FoodiesPal** también ofrece la posibilidad de compartir tus recetas favoritas con amigos y familiares, fomentando el intercambio de deliciosas creaciones culinarias. Y si deseas recrear una receta que te haya cautivado, el asistente te permite acceder a las recetas compartidas por otros usuarios para replicarlas con facilidad en la comodidad de tu hogar. Con **FoodiesPal**, la cocina se convierte en una experiencia colaborativa, simple y emocionante.

• **Modelo de Negocios**

En este informe , emplearemos el Lean Canvas como una herramienta gráfica en la presentación para exponer de manera efectiva y resumida nuestra propuesta de valor y estrategia de negocio, en esta sección la explicaremos punto por punto.

- **Propuesta de valor**

Nuestra propuesta de valor reside en proporcionar a los usuarios una selección rápida, diversa y adaptable de recetas culinarias a través de la integración de inteligencia artificial y ciencia de datos en una sola herramienta. Esto posibilita que los usuarios descubran recetas de manera sencilla y las compartan con sus pares, fomentando la creación de una red de compañeros culinarios.

- **Segmentos de cliente**

- Personas que buscan mejorar sus hábitos alimenticios y adoptar opciones más saludables en su dieta diaria.
- Personas con limitados conocimientos culinarios que desean adentrarse en la preparación de platos.
- Aventureros culinarios que tienen interés en explorar y experimentar con una amplia variedad de recetas.
- Personas que enfrentan el dilema diario de qué cocinar y buscan soluciones prácticas y creativas.
- Chefs y profesionales de la cocina que desean promocionar y compartir sus creaciones culinarias a través de nuestra plataforma
- Estudiantes universitarios con poco tiempo, habilidades limitadas en la cocina y pocos ingredientes.

- **Canales**

- Distribución a través de tiendas de aplicaciones (App Stores) para acceder a usuarios móviles y dispositivos inteligentes.
- Colaboración con acuerdos gubernamentales o privados para promover la adopción de hábitos alimenticios saludables.
- Estrategias de publicidad en redes sociales para alcanzar audiencias específicas y generar conciencia sobre nuestras soluciones culinarias.
- Uso de publicidad impresa en medios selectos para llegar a audiencias que prefieren canales tradicionales de comunicación.
- Participación en ferias de software y eventos relacionados para presentar directamente nuestra plataforma a una audiencia interesada.

- **Ventaja injusta**

- **Sistema de Reconocimiento OCR:** La aplicación cuenta con una implementación avanzada de reconocimiento óptico de caracteres (OCR) que permite la extracción precisa de ingredientes e instrucciones de recetas presentes en imágenes. Esto ofrece a los usuarios una experiencia intuitiva y efectiva en la búsqueda de recetas.
- **Sistema de Recomendación Híbrido Personalizado:** Se ha desarrollado un sistema de recomendación híbrido que combina algoritmos de filtrado colaborativo y basado en contenido. Esta técnica proporciona a los usuarios recomendaciones efectivas y eficientes, además de personalizadas, adaptadas a sus preferencias, restricciones y patrones culinarios individuales.

- **Tecnología Propia de Text-to-Speech y Speech-to-Text:** La aplicación presenta una tecnología interna de conversión de texto a voz (TTS) y de voz a texto (STT). Este desarrollo posibilita a los usuarios recibir instrucciones y brindar comandos por medio de la voz, mejorando la interacción y ofreciendo una experiencia manos libres en la cocina.
- **Fuentes de Ingreso**
 - **Regalías por Publicidad:** Generamos ingresos al mostrar anuncios pertinentes en la plataforma a través de asociaciones publicitarias, obteniendo ganancias por cada interacción que los usuarios tengan con dichos anuncios.
 - **Venta de Utensilios y Equipamiento:** Establecemos alianzas con fabricantes y proveedores de utensilios de cocina, como cuchillos, hornos y otros equipos.
 - **Suscripciones a Creadores de Contenido:** Permitimos a los usuarios suscribirse a sus chefs o creadores de contenido culinario preferidos. Los suscriptores obtienen acceso a contenido exclusivo, tutoriales avanzados y experiencias interactivas, generando ingresos recurrentes para los chefs y FoodiesPal.
 - **Membresía Premium:** Ofrecemos funciones premium como recetas exclusivas, asesoramiento personalizado y características avanzadas. Esta membresía ofrece un valor diferenciado y una fuente constante de ingresos.
- **Métricas clave**
 - **Registro y uso de aplicación:** Mide la proporción de usuarios que se registran y autentican exitosamente en la aplicación, evaluando el uso de esta, los módulos accedidos y tiempo general en ella.

- **Búsqueda de Recetas:** Rastrea la cantidad de búsquedas de recetas que resultan en listas de recetas las cuales los usuarios acceden, realizan o guardan, demostrando la eficacia del sistema de búsqueda.
 - **Numero de usuarios seguidores:** Evalúa cuántos usuarios optan por seguir a chefs específicos, indicando el atractivo del contenido y la funcionalidad de seguimiento.
 - **Tasa de Conversión de Usuarios Gratuitos a Premium:** Analiza cuántos usuarios gratuitos actualizan a membresía premium, indicando el valor percibido de las funcionalidades premium.
 - **Tasa de Abandono de Usuarios en la Aplicación:** Evalúa el porcentaje de usuarios que abandonan la aplicación en cada fase, identificando posibles puntos débiles en la experiencia.
- **Problema**
 - **Baja Experiencia de Usuario y Limitada Tecnología Culinaria:** Los recetarios impresos resultan incómodos y poco prácticos en una cocina cotidiana, especialmente debido a la dificultad de manipulación con las manos ocupadas y sucias. De manera similar, buscar recetas en un dispositivo móvil conlleva el riesgo de contaminación cruzada. La compleja dinámica de la vida moderna dificulta incluso decidir qué cocinar, lo que complica aún más la experiencia culinaria.
 - **Calidad de Recomendaciones :** Los métodos tradicionales de búsqueda de recetas suelen generar una avalancha de opciones que abruman a los usuarios, ya que se basan en palabras clave en lugar de considerar preferencias personales, desmotivando a los usuarios a probar nuevas recetas. Incluso en libros, la abundancia de opciones complica encontrar recetas adecuadas, llevando a un uso poco frecuente de estos recursos.

- **Escasa Comunidad Dedicada y Compartida:** Establecer una comunidad culinaria en plataformas generales resulta complejo debido a la diversidad de contenidos. Compartir recetas en redes sociales convencionales dificulta la retroalimentación efectiva y la captura de recetas para futuras ocasiones. Las interacciones con recetas son fugaces y carecen de un espacio específico para el intercambio de conocimientos, lo que limita la creación de una comunidad comprometida y colaborativa.

- **Solucion**

- **Búsqueda Inteligente de Recetas:** Implementamos un sistema de búsqueda machine learning para que los usuarios encuentren recetas adaptadas a los ingredientes disponibles en sus cocinas, en la boleta de compras, mediante el análisis de su perfil, etc. facilitando la planificación y minimizando el desperdicio alimentario.
- **Ayudante de Cocina Avanzado:** Integramos una tecnología de reconocimiento de voz para proporcionar instrucciones y asistencia en tiempo real mientras cocinas, eliminando la necesidad de tocar la pantalla y garantizando una preparación más fluida y entretenida.

Red Social Culinaria: Creamos una comunidad en la que los entusiastas de la cocina pueden compartir sus creaciones, intercambiar consejos y explorar nuevas ideas culinarias, fomentando la colaboración y la creatividad.

Alcance del proyecto

Buscamos desarrollar una aplicación que elimine la carga de pensar constantemente en que cocinar el día a día, proponiendo recetas en base a las restricciones alimentarias del usuario, sus tipos de alergias o la dieta que se quiera llevar a cabo.

Las características y funcionalidades claves son las siguiente:

1. Registrar usuario:

- El usuario puede crear una cuenta personal.
- Se recopilará información sobre el usuario acerca de sus restricciones alimentarias, alergias y preferencias dietéticas.

2. Buscar recetas:

- Se le permitirá al usuario buscar recetas según los ingredientes que tiene disponible.
- Se considerarán las restricciones alimenticias, alergias y preferencias dietéticas en la búsqueda del usuario.

3. Filtrar y personalizar:

- Se le ofrecerá al usuario filtrar sus preferencias para adaptar la recomendación en base a las necesidades específicas del usuario.

4. Perfil de usuario:

- Se le permitirá al usuario ver y editar su perfil.

5. Generación de recetas:

- La aplicación generará recomendaciones de recetas en base a la información del perfil del usuario y sus preferencias seleccionadas.

6. Recetas detalladas:

- Se le proporcionará al usuario información detallada sobre la receta seleccionada, incluyendo los ingredientes, las instrucciones paso a paso de la receta, el valor nutricional y posibles sustituciones en caso de que el usuario desee otra cosa.

7. Listas de compras:

- Se le permitirá al usuario crear una lista de compras en caso de que el usuario necesite ciertos ingredientes o le haga falta algún ingrediente.

Anexo B

8. Interacción social:

- Se le permitirá al usuario compartir recetas y comentar sobre estas en recetas que el desee. Se le permitirá al usuario seguir a otros usuarios y recibir notificaciones sobre sus actividades.

9. Notificaciones:

- Se le notificará al usuario acerca de sus recordatorios como poner el pollo en el horno, actualizaciones de recetas o actividad de usuarios que sigue.

Alcances excluidos:

- La aplicación no tendrá la capacidad para que el usuario realice compras de alimentos a directamente a través de la aplicación.
- No se podrá hacer pedidos a cadenas de comida rápida o restaurantes.

Desglose del desarrollo

Día 1	Descripción
Desarrollo del Back-end	Desarrollo de OCR. Desarrollo del sistema de recomendación. Desarrollo de TTS y STT.

Día 2	Descripción
Continuación del desarrollo	Se continuará desarrollando el Back-end y se comenzará a integrar con lo que ya está desarrollado.

Día 3	Descripción
Pruebas unitarias	Se realizarán los modelos de entrenamiento para verificar que funcionen correctamente.
Desarrollo Front-end	Comienza el desarrollo de la aplicación móvil con lo anterior testado para su uso.
Integración de todos los Back-end	Se realizará la integración de todos los Back-end desarrollados en el día 1.

Día 4	Descripción
Pruebas unitarias	Se le realizarán las pruebas correspondientes a la aplicación en busca de errores.
Mejora del UI	Se harán las mejoras de la interfaz de usuario correspondientes para el lanzamiento de la aplicación final.

Día 5	Descripción
Verificación final	Se verificará la aplicación por última vez para el lanzamiento final.