

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



“CLASIFICACIÓN AUTOMÁTICA DE LA
CALIDAD DE LECTURA ORAL PARA LA
EVALUACIÓN DEL DOMINIO LECTOR EN
ESTUDIANTES DE CHILE DE EDUCACIÓN
BÁSICA: CASO TEACHEAR.”

NICOLÁS BENJAMÍN ARMIJO CALQUÍN

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesora Guía: Dra. Carolina Saavedra
Profesor Correferente: Pedro Godoy Barrera

Junio - 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Clasificación automática de la calidad de lectura oral para la evaluación del dominio lector en estudiantes de Chile de educación básica: Caso TeacHear.

Nombre del candidato(a): Nicolás Benjamín Armijo Calquín

Carrera / Grado: Ingeniería Civil Informática

Campus: Casa Central Valparaíso

Departamento: Departamento de Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Dra. Carolina Saavedra, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA que:**

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 24/06/2026

Firma:

Estudiante o Candidato(a):

Fecha: 23/06/2026

Firma:

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

DEDICATORIA

A mi familia, especialmente a mis padres Jorge y Viviana, junto con mi abuelita Eliana, que me entregaron todo para que mi única preocupación fueran mis estudios.

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a mi equipo de Teachear, por contribuir a que esta experiencia fuera mucho más enriquecedora tanto en lo académico como en lo personal. A Gaspar, Camus, Valentina, Bastián y Javier, gracias por el compromiso, la colaboración y el ambiente humano que construimos a lo largo del proyecto. En especial, quiero agradecer a aquellos Valentina, Camus y Gaspar que se convirtieron en grandes amigos, con quienes compartí innumerables desafíos, trabajos, conversaciones y momentos de alegría. Más allá de los logros académicos, guardo con especial cariño los recuerdos, las risas y la compañía que hicieron de esta etapa una experiencia inolvidable.

Asimismo, agradezco al Colegio GEA y a todos sus estudiantes que participaron voluntariamente en las actividades de evaluación y recolección de datos que hicieron posible el desarrollo de este trabajo. Su disposición y colaboración fueron fundamentales para llevar adelante la investigación.

Deseo realizar un reconocimiento especial a los hermanos de Gaspar Navarro quienes participaron de manera voluntaria en la generación de datos utilizados durante el entrenamiento y evaluación de los modelos desarrollados. Una mención a mi tía Tati, que participó como conejillo de indias en las pruebas de usabilidad de Teachear. Su apoyo contribuyó significativamente al desarrollo de esta memoria.

Finalmente, quiero agradecer a todas las personas que me acompañaron y apoyaron durante este proceso. A mi familia y amigos, gracias por escucharme una y otra vez hablar sobre este proyecto, por su paciencia, interés y apoyo constante, incluso en aquellos momentos en que el camino parecía más complejo. Cada conversación, consejo y palabra de aliento contribuyó de alguna forma a que pudiera llegar hasta aquí.

Y, por último, quiero agradecer de manera especial a Ignacia. Aunque llegó en la etapa final de este proceso, eso no hace que su aporte haya sido menos importante. Su compañía, apoyo y motivación diaria fueron fundamentales para afrontar los desafíos de esta última etapa. Gracias por creer en mí, por alentarme cuando más lo necesitaba y por acompañarme en el cierre de un ciclo tan significativo.

RESUMEN

Resumen—La evaluación de la calidad de lectura oral constituye una herramienta fundamental para monitorear el desarrollo lector de estudiantes de educación básica. Sin embargo, su aplicación requiere tiempo y la participación de evaluadores capacitados. En este contexto, esta memoria aborda el problema de la clasificación automática de la calidad de lectura oral en estudiantes chilenos a partir de registros de audio.

Se propone evaluar la factibilidad de utilizar técnicas de aprendizaje automático para clasificar la calidad lectora según categorías definidas en instrumentos utilizados en el contexto educativo nacional. Para ello, se construyó un conjunto de datos compuesto por grabaciones de estudiantes de educación básica y se compararon distintos enfoques de clasificación basados en características acústicas tradicionales y *embeddings* extraídos mediante modelos preentrenados de procesamiento de voz. La evaluación se realizó mediante validación cruzada estratificada y el análisis de métricas de desempeño y matrices de confusión.

Los resultados muestran que es posible identificar patrones asociados a distintos niveles de calidad lectora pese al tamaño reducido del conjunto de datos. Asimismo, se identificaron limitaciones relacionadas con la variabilidad entre clases y el riesgo de sobreajuste. Estos hallazgos constituyen una base para el desarrollo de herramientas de apoyo a la evaluación lectora.

Palabras Clave—Calidad de lectura oral, clasificación automática, aprendizaje automático, procesamiento de voz, evaluación lectora.

ABSTRACT

Abstract—Oral reading quality assessment is a fundamental tool for monitoring the reading development of elementary school students. However, its application is time-consuming and requires trained evaluators. In this context, this thesis addresses the problem of automatically classifying oral reading quality in Chilean students using audio recordings.

The objective of this work is to evaluate the feasibility of applying machine learning techniques to classify reading quality according to categories defined by assessment instruments used in the Chilean educational context. To this end, a dataset composed of audio recordings from elementary school students was constructed, and different classification approaches based on traditional acoustic features and embeddings extracted from pre-trained speech processing models were compared. The proposed approach was evaluated using stratified cross-validation, performance metrics, and confusion matrix analysis.

The results show that it is possible to identify patterns associated with different levels of reading quality despite the limited size of the dataset. Furthermore, limitations related to class variability and the risk of overfitting were identified. These findings provide a foundation for the development of tools to support reading assessment.

Keywords—Oral reading quality, automatic classification, machine learning, speech processing, reading assessment.

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
ÍNDICE DE FIGURAS	VII
ÍNDICE DE TABLAS	VII
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	3
1.1. Contexto	3
1.2. Proyecto TeacHear	3
1.3. Alcance y objetivos	3
CAPÍTULO 2: MARCO CONCEPTUAL	5
2.1. Calidad de lectura oral	5
2.2. Procesamiento de audio	6
2.3. Extracción de características	6
2.4. Representaciones aprendidas	7
2.5. Modelos de aprendizaje automático	7
2.6. Evaluación y metodologías	8
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN	10
3.1. Descripción general del flujo experimental	10
3.2. Conjunto de datos	10
3.3. Procesamiento de audios	12
3.4. Generación de representaciones	13
3.4.1. Representaciones acústicas clásicas	13
3.4.2. Embeddings obtenidos mediante modelos preentrenados	14
3.5. Modelos de aprendizaje automático	15
3.5.1. Modelos clásicos sobre representaciones fijas	15
3.5.2. Ajuste fino parcial con clasificador neuronal	16
3.5.3. Resumen de enfoques considerados	17
3.6. Diseño experimental	17
3.6.1. Unidad de partición	18
3.6.2. Validación cruzada	18
3.6.3. Control de fuga de información	18
3.6.4. Configuración de los experimentos	18
3.6.5. Reproducibilidad	18
3.6.6. Reporte de resultados	19
3.7. Protocolo de evaluación	19
CAPÍTULO 4: VALIDACIÓN DE LA SOLUCIÓN	20
4.1. Resumen del protocolo de validación	20
4.2. Resultados del enfoque basado en características acústicas	20
4.2.1. Configuración experimental	21
4.2.2. Resultados globales	21
4.2.3. Análisis de resultados	21
4.3. Resultados del enfoque basado en embeddings preentrenados	22

4.3.1.	Configuración experimental	22
4.3.2.	Resultados globales	22
4.3.3.	Análisis de resultados	23
4.4.	Resultados del enfoque basado en transfer learning	24
4.4.1.	Configuración experimental	25
4.4.2.	Resultados globales	25
4.4.3.	Análisis de resultados	25
4.5.	Comparación global de enfoques	26
4.6.	Discusión de resultados	27
CAPÍTULO 5: CONCLUSIONES		31
ANEXOS		36
	Resultados de clasificación por clase de enfoque descriptores acústicos clásicos	36
	Resultados de clasificación por clase de enfoque embeddings preentrenados	36
	Resultados de clasificación por clase de enfoque transfer learning	37

ÍNDICE DE FIGURAS

1.	Transformación de un espectograma a coeficientes MFCC.	6
2.	Aprendizaje auto-supervisado multilingüe en Wav2Vec2.	8
3.	Flujo general de experimentación.	11
4.	Distribución de la calidad de lectura oral en segmentos de audio.	13
5.	Distribución de la duración de los segmentos de audio.	14
6.	Arquitectura del clasificador neuronal.	17
8.	Matrices de confusión para modelos clásicos con representación de embeddings preentrenados.	24
7.	Matrices de confusión para modelos clásicos con representación acústica.	29
9.	Matriz de confusión obtenidas para el ajuste fino de Wav2Vec2 con un clasificador neuronal.	30

ÍNDICE DE TABLAS

1.	Categorías de calidad de lectura oral.	1
2.	Distribución de participantes por curso.	12
3.	Hiperparámetros utilizados en los modelos tradicionales.	16
4.	Resumen de enfoques de modelamiento considerados.	17
5.	Resultados globales del enfoque basado en características acústicas.	21
6.	Resultados globales del enfoque basado en características acústicas.	23
7.	Resultados globales del ajuste fino parcial del encoder.	25
8.	Comparación global de los mejores resultados por enfoque.	26
9.	Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo SVM.	36
10.	Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo Random Forest.	36
11.	Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo XGBoost.	36
12.	Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo SVM.	36
13.	Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo Random Forest.	37
14.	Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo XGBoost.	37
15.	Resultados de clases enfoque transfer learning.	37

INTRODUCCIÓN

La lectura es una de las habilidades más críticas y valiosas que desarrollan los estudiantes dentro del ciclo de educación básica, tanto que no influye exclusivamente en el desarrollo del lenguaje, sino que afecta al aprendizaje de otras áreas, como lo es la matemática, la comprensión del medio, e incluso moldea el autoestima del estudiante [1, 2].

Siendo una habilidad fundamental, su monitoreo es muy importante, ya que, como toda competencia clave, su desarrollo requiere seguimiento constante para identificar oportunamente dificultades y orientar estrategias pedagógicas adecuadas [2, 3]. La evaluación del dominio lector permite obtener o estimar el nivel de competencia lectora, permitiendo conocer cuán bien se decodifican las palabras (calidad de lectura oral) y qué tan rápido se hace (velocidad de lectura oral) [1].

La estandarización de los criterios e instrumentos de evaluación que generalmente se utilizan en los establecimientos chilenos fueron propuestos por la Fundación Educacional Arauco [1]. Las categorías para clasificar la calidad de lectura oral se presentan en la Tabla 1. Sin embargo, aún con la presencia de estos criterios, en la práctica, esta categorización queda sujeta a la atención, la experiencia y criterio profesional del docente que la realiza. Esta condición genera un problema central, el proceso manual introduce un grado significativo de subjetividad, lo que afecta la consistencia y confiabilidad de los resultados. Las causas de esta subjetividad se relacionan a factores humanos, como el agotamiento y sobrecarga laboral de los profesores, también como las diferencias de las trayectorias profesionales de estos, influyendo en la comparabilidad de resultados y las conclusiones del análisis que se puede extraer.

Tabla 1: Categorías de calidad de lectura oral. Se ejemplifica las lecturas de las 5 categorías que establece la Fundación Arauco.

Fuente: Elaboración propia. Información extraída de [1].

Categoría Lectura	Definición	Ejemplo
No Lector	Se entiende un alumno que no sabe leer nada, que sólo reconoce algunas letras aisladamente, pero no es capaz de unirlos en sílabas, o bien, que sólo lee algunas sílabas aisladas.	No aplica.
Silábica	El alumno lee las palabras sílaba a sílaba, no respetando las palabras como unidades.	La – me – sa – es – tá – muy – su – cia – El – lu – nes – la – voy – a – lim – piar.
Palabra a Palabra	El alumno lee el texto palabra a palabra, sin respetar las unidades de sentido.	La – mesa – está – muy – sucia – El – lunes – la – voy – a – limpiar.
Unidades Cortas	El alumno une algunas palabras formando pequeñas unidades de sentido.	La mesa – está – muy sucia – El lunes la – voy – a limpiar.
Fluida	El alumno lee en forma continua. Una buena lectura fluida implica dar una inflexión de voz adecuada al contenido del texto, respetando las unidades de sentido y la puntuación.	La mesa está muy sucia. – El lunes la voy a limpiar.

La automatización de la evaluación del dominio lector se presenta como una alternativa prometedora para reducir la subjetividad inherente al proceso manual y mejorar la consistencia de los resultados, además de ser comparable a la precisión de un educador [4, 5]. Considerando este escenario, el proyecto TeacHear desarrollado en el marco de la Feria de Software 2025 en la Universidad Técnica Federico Santa María, busca enfrentar esta problemática. Su objetivo es automatizar el proceso de la evaluación bajo técnicas de inteligencia artificial [6].

En etapas iniciales del desarrollo del proyecto, los esfuerzos realizados para lograr la automatización de

la calidad de lectura oral consistió en la exploración y utilización de una única línea de modelamiento sobre los datos disponibles. Sin embargo, tal proceso no consideró en profundidad ciertos aspectos del diseño experimental, tales como la correcta separación de datos, la validación estadística de los resultados y el control de sesgos, lo que limita la interpretabilidad y veracidad de los resultados obtenidos.

En este contexto, y considerando la existencia de múltiples enfoques en técnicas de procesamiento de audio y aprendizaje automático, el presente trabajo tiene como objetivo analizar diferentes modelos de aprendizaje automático para clasificar la calidad de lectura oral en estudiantes de educación básica en Chile, con el fin de apoyar la automatización y estandarización de la evaluación del dominio lector.

Para ello, se revisan y caracterizan técnicas de procesamiento de audio y modelos de aprendizaje automático adecuados para esta tarea. Posteriormente, se implementan y entrenan distintos enfoques de clasificación, los cuales son evaluados y comparados en relación con el trabajo del proyecto TeacHear, utilizando métricas de desempeño pertinentes. Finalmente, se proponen líneas de trabajo futuro orientadas a mejorar y extender las herramientas de evaluación automática del dominio lector.

Se exploran 3 enfoques; el enfoque clásico, que busca crear descriptores acústicos del audio para entrenar modelos de clasificación; el enfoque *embedding*, que utiliza un modelo previamente entrenado para generar una representación semántica del contenido del audio para luego entrenar los mismos modelos de clasificación; y por último, el enfoque *transfer learning*, que busca tomar estos modelos previamente entrenados en representaciones del habla para ajustarlos finamente al contexto del problema en particular, en este caso la clasificación de lectura oral en estudiantes de enseñanza básica.

Finalmente, la presente memoria se estructura de la siguiente manera. En el Capítulo 1 se presenta la problemática del estudio y se establecen los objetivos del trabajo. En el Capítulo 2 se desarrolla el marco conceptual, donde se revisan los fundamentos teóricos y trabajos relacionados. En el Capítulo 3 se describe la propuesta de solución, incluyendo las metodologías y modelos utilizados. En el Capítulo 4 se presenta la validación de la solución, junto con el análisis y discusión de los resultados obtenidos. Finalmente, en el Capítulo 5 se exponen las conclusiones del trabajo y se proponen líneas de trabajo futuro.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

1.1. Contexto

En los procesos educativos de nivel básico vinculados a la lectura, es fundamental evaluar dos dimensiones clave para fortalecer y apoyar el aprendizaje de los estudiantes: el dominio lector y la comprensión lectora. El primero se refiere a la capacidad de decodificar correctamente el texto, constituyendo una habilidad base indispensable para alcanzar una adecuada comprensión de lo leído. [1].

Cuando se evalúa el dominio lector se busca medir la calidad de la lectura oral y la velocidad lectora. Al cuantificar estos aspectos los docentes pueden tomar decisiones pedagógicas para apoyar el proceso de aprendizaje a nivel individual y de grupo curso. Actualmente, la evaluación se realiza de manera manual, donde, según la Fundación Educacional Arauco [1], es necesario contar con lo siguiente: Instructivo para la evaluación del dominio lector, criterios de evaluación, hojas de registro, parámetros referenciales, lápiz y cronómetro.

Aunque existen criterios de evaluación claramente definidos, su aplicación en la práctica resulta en un proceso lento y con un grado significativo de subjetividad. La atención del docente debe distribuirse entre múltiples tareas, y factores como su experiencia, fatiga o nivel de concentración influyen directamente en el resultado de la evaluación. Esto implica que, si bien la prueba es adecuada para medir el nivel de competencia lectora del estudiante, se ve afectada la consistencia entre evaluadores. Además, la calidad de la evaluación tiende a deteriorarse cuando se realiza de manera continua durante periodos prolongados, especialmente en cursos con alta cantidad de estudiantes [7, 8, 9].

1.2. Proyecto TeacHear

A raíz de esta problemática general nace TeacHear, un proyecto desarrollado en el marco Feria de Software 2025 de la Universidad Técnica Federico Santa María que busca automatizar la evaluación del dominio lector, con el fin de reducir los tiempos de evaluación y entregar resultados objetivos para apoyar la toma de decisiones pedagógicas [6]. Durante el desarrollo del proyecto, específicamente en el área de modelos de inteligencia artificial, los esfuerzos en la clasificación automática de la calidad de lectura oral no lograron abarcar satisfactoriamente los desafíos presentes, en particular lo relativo a; la disponibilidad de datos, ya que la muestra pequeña, tanto a nivel de cantidad de audios, duración acumulada de hablantes y a cantidad de lectores influyen en la interpretación y robustez de los resultados; y por parte de la evaluación de modelos, solo se siguió una línea experimental sin considerar alternativas a nivel de representación de audios y modelos de clasificación.

En particular, el desarrollo presentó una serie de limitantes que no se abordaron correctamente, lo que compromete la validez estadística, reproducibilidad y la capacidad de generalización de los modelos obtenidos. Principalmente, la falta de rigor experimental pone en jaque las decisiones de diseño, ya que no se consideraron factores como la baja cantidad de muestras disponibles, la ausencia de un protocolo de evaluación riguroso, y la limitada exploración y comparación de enfoques de modelamiento. Estas condiciones necesitan abordarse para obtener resultados confiables y reproducibles, que generen conclusiones sólidas capaces de orientar el desarrollo de herramientas computacionales que apoyen las labores docentes.

1.3. Alcance y objetivos

Por lo tanto, el problema consiste en el protocolo experimental utilizado en el desarrollo del proyecto TeacHear. Bajo este contexto, surge la necesidad de establecer un marco experimental que permita evaluar, de manera rigurosa y reproducible modelos de clasificación de lectura oral en escenarios de datos limitados, priorizando su validez y comparabilidad de resultados. Para ello, se plantean los siguientes objetivos:

Objetivo general – Mejorar el diseño experimental utilizado en el proyecto TeacHear para analizar diferentes modelos de aprendizaje automático que clasifiquen la calidad de lectura oral en estudiantes chilenos de

educación básica, con el fin de apoyar la automatización y estandarización de la evaluación del dominio lector.

Objetivos específicos –

- Revisar y caracterizar técnicas de evaluación en experimentos de aprendizaje automático, tal como lo son validación cruzada, identificación de fuga de información y estimación de desempeño.
- Revisar y caracterizar técnicas de procesamiento de audio y modelos de aprendizaje automático adecuados para la clasificación de la calidad de lectura oral.
- Implementar y entrenar distintos enfoques de modelos seleccionados para la clasificación de la calidad de lectura oral.
- Evaluar y comparar el desempeño de los enfoques propuestos frente al trabajo realizado en el proyecto de feria de software TeachHear, considerando métricas de desempeño, como lo son *accuracy*, *F1* y *recall*.
- Proponer recomendaciones del trabajo futuro para la mejora y extensión de la tarea de clasificación automática de la calidad de lectura oral en estudiantes.

CAPÍTULO 2

MARCO CONCEPTUAL

Para abordar el problema planteado, es necesario establecer un marco conceptual que permita comprender los fundamentos teóricos y metodológicos asociados. Este capítulo presenta los principales conceptos relacionados con la calidad de lectura oral, procesamiento de señales de audio, la extracción de características, modelos de aprendizaje automático y las técnicas de evaluación en escenarios de datos limitados.

2.1. Calidad de lectura oral

La calidad de lectura oral hace referencia al grado de fluidez con la que un estudiante lee. En Chile, este criterio de evaluación fue estandarizado en la propuesta educacional de la Fundación Educacional Arauco publicada el año 2004 [1]. En ella, se establecen 5 niveles de fluidez lectora la que el docente debe determinar al escuchar 1 minuto de lectura en voz alta del estudiante. En particular, se debe de fijar en las pausas que el alumno hace al leer, para así clasificarlo en una de las siguientes categorías:

- **No Lector:** Se entiende un alumno que no sabe leer nada, que sólo reconoce algunas letras aisladamente, pero no es capaz de unirlos en sílabas, o bien, que sólo lee algunas sílabas aisladas.
- **Lectura Silábica:** El alumno lee las palabras sílaba a sílaba, no respetando las palabras como unidades.
- **Lectura Palabra a Palabra:** El alumno lee el texto palabra a palabra, sin respetar las unidades de sentido.
- **Lectura Unidades Cortas:** El alumno une algunas palabras formando pequeñas unidades de sentido.
- **Lectura Fluida:** El alumno lee en forma continua. Una buena lectura fluida implica dar una inflexión de voz adecuada al contenido del texto, respetando las unidades de sentido y la puntuación.

Un breve estado del arte sobre la aplicación de técnicas de inteligencia artificial para la evaluación de fluidez lectora muestra un interés creciente sobre el uso de herramientas computacionales para apoyar y automatizar los procesos docentes, con fines de liberar carga laboral, brindar resultados más detallados y objetivos. En su mayoría, los investigadores han utilizado herramientas de reconocimiento automático del habla (*Automatic Speech Recognition*, ASR), las cuales permiten la transcripción automática de audio a texto, donde posteriormente se extraen métricas objetivas como palabras correctas por minuto, velocidad lectora, tasa de error, pausas y medidas de prosodia. A partir de estas variables, distintos modelos de aprendizaje automático y redes neuronales han sido empleados para clasificar niveles de desempeño lector o predecir puntajes otorgados por evaluadores humanos [4, 5, 10, 11, 12].

Bajo esta línea, la tendencia predominante ha consistido en abordar el problema de manera secuencial, donde primero se transcribe el texto leído y luego se calculan métricas derivadas de dicha transcripción. Si bien se han logrado buenos resultados, se depende fuertemente en la calidad de la transcripción automática y toma mayor peso las variables explícitas presentes en el texto, dejando fuera otros tipos de señales acústicas presente en el audio. Otra línea de investigación busca explorar enfoques de *deep learning* aplicados directamente sobre la señal de la grabación [13].

En el contexto nacional la investigación es escasa. En general los esfuerzos se presentan en diferentes soluciones comerciales para fines educativos o diagnósticos, concentrándose en la obtención de transcripciones y en métricas tradicionales de velocidad, exactitud, y prosodia. Sin embargo, no se presta particular atención a los estándares pedagógicos locales. A su vez, los niveles o categorías utilizados en estudios internacionales suelen responder a escalas propias de cada país, lo que dificulta la aplicación al contexto chileno [14, 15].

De este modo, al momento de la elaboración de esta memoria y según la revisión bibliográfica realizada, no se encontraron más antecedentes que abordaran específicamente la clasificación de la fluidez lectora infantil utilizando directamente la señal de audio como fuente principal de información, ni propuestas ajustadas al estándar de evaluación utilizado en Chile. Esta brecha evidencia una oportunidad de investigación relevante, tanto desde el punto de vista metodológico como aplicado, al buscar soluciones automáticas contextualizadas a la realidad educativa nacional.

2.2. Procesamiento de audio

El habla corresponde a una señal acústica continua que, para su almacenamiento y análisis computacional, debe ser transformada a una representación digital. Este proceso consiste en discretizar la señal de audio mediante muestreos en intervalos regulares en tiempo, obteniéndose una secuencia discreta y cuantizada de la amplitud, obteniendo valores numéricos que puede ser procesados mediante algoritmos computacionales [16]. Así se puede capturar información relevante que define la pronunciación, ritmo, pausas y entonación del hablante.

Una propiedad fundamental de la digitalización es la frecuencia de muestreo, definida como la cantidad de muestras registradas por segundo. En tareas relacionadas con procesamiento de voz, es común utilizar frecuencias de 16 kHz o superiores, ya que permiten representar adecuadamente el rango espectral más relevante para la inteligibilidad del habla [17].

Además de la frecuencia de muestreo, los registros de audio suelen presentar variabilidad en aspectos como duración, nivel de amplitud, cantidad de canales y condiciones de grabación. Por esta razón, una etapa habitual de preprocesamiento consiste en estandarizar el formato de los archivos, convirtiéndolos a canal mono, normalizando amplitudes y unificando la frecuencia de muestreo. Estas transformaciones permiten reducir fuentes de variación no deseadas y facilitan una comparación consistente entre muestras.

2.3. Extracción de características

El uso directo de la señal de audio cruda presenta diversos desafíos, entre ellos la alta cantidad de muestreos resulta en una representación extensa, la variabilidad de su longitud entre distintas muestras y la presencia de información poco relevante para la tarea de interés, en este caso la clasificación automática de la calidad de lectura oral. Por esta razón, diferentes técnicas de extracción de características son utilizadas para transformar la señal en representaciones acordes que faciliten la manipulación de las señales de audio [17]

Entre las representaciones más utilizadas se encuentran los coeficientes cepstrales en frecuencia Mel, Mel-Frequency Cepstral Coefficients en inglés, de ahora en adelante MFCC. Estos son ampliamente adoptados en sistemas tradicionales de reconocimiento de habla y clasificación acústica [18]. Los MFCC buscan resumir la envolvente espectral de la señal utilizando una escala de frecuencias inspirada en la percepción auditiva humana (ver Figura 1). Su cálculo considera, de manera general, la transformación espectral de cada ventana, la aplicación de un banco de filtros distribuidos en escala Mel, una compresión logarítmica de energía y una transformación final que produce un conjunto reducido de coeficientes representativos.

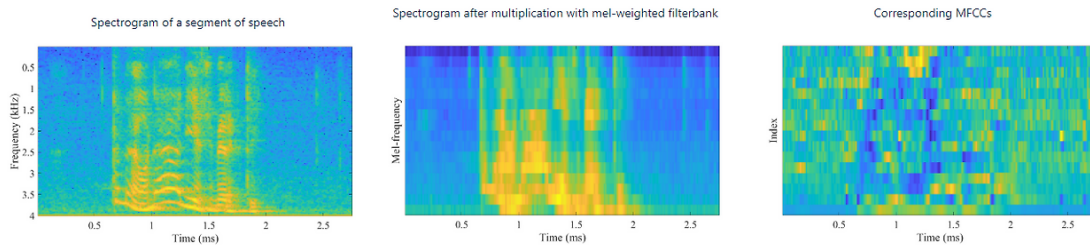


Figura 1: Transformación de un espectrograma a coeficientes MFCC. Se presenta una comparación entre la representación espectral de una señal de audio y su transformación a coeficientes MFCC. Esta transformación permite resumir la información acústica del espectrograma en una representación más compacta, utilizada comúnmente como entrada para modelos de aprendizaje automático.

Fuente: MFCC Feature extraction for Sound Classification [19].

El uso de MFCC presenta diversas ventajas prácticas. En primer lugar, permiten obtener una representación más compacta de la señal de audio cruda, al resumir la información acústica distribuida en una gran cantidad de muestras temporales mediante un conjunto reducido de coeficientes. Además, concentran información acústica relevante para tareas relacionadas con el habla y poseen un costo computacional

relativamente bajo, lo que explica su amplio uso histórico en sistemas ASR [17, 20].

No obstante, estas representaciones también presentan limitaciones. Al depender de un diseño manual, los MFCC incorporan supuestos fijos sobre qué información resulta relevante, lo que puede restringir su capacidad de adaptación a tareas específicas. Asimismo, al resumir localmente la señal, pueden perder información contextual de mayor alcance y verse afectados por condiciones adversas de grabación, como ruido ambiente o variaciones entre dispositivos.

2.4. Representaciones aprendidas

En tareas de procesamiento de audio es posible analizar la señal en distintos niveles de granularidad temporal. Tradicionalmente, la extracción de características como los coeficientes MFCC se realiza a nivel de *frame*, es decir, sobre ventanas cortas de la señal de audio, generalmente entre 20 a 30 milisegundos. Este enfoque permite capturar variaciones locales en la señal, pero genera secuencias de longitud variable, lo que introduce desafíos para modelos que requieren entradas de tamaño fijo.

En este contexto, resulta necesario definir un nivel de representación adecuado para la tarea de interés. Una alternativa consiste en trabajar a nivel de registro, donde cada muestra de audio completa es representada mediante un único vector de características. Este enfoque facilita el uso de modelos de aprendizaje automático tradicionales, tales como máquinas de soporte vectorial (*Support Vector Machine*, SVM), bosques aleatorios (*Random Forest*, RF) y *Extreme Gradient Boosting* (XGBoost), los cuales requieren entradas de dimensionalidad fija. Esta representación encaja con la tarea que se busca abordar en este trabajo, ya que se debe clasificar el audio como unidad completa bajo los niveles de lectura oral.

Para obtener representaciones a nivel de registro, es común calcular medidas estadísticas sobre las características extraídas en cada *frame* [21]. En el caso de los MFCC, una estrategia habitual consiste en calcular estadísticos globales como la media y la desviación estándar de cada coeficiente a lo largo del tiempo, obteniendo así una representación compacta y de tamaño fijo para cada muestra de audio [22].

Si bien esta técnica permite adaptar representaciones a modelos clásicos, presenta limitaciones importantes. En particular, al resumir la información temporal mediante estadísticas simples, se pierde la estructura secuencial del habla, lo que puede afectar la capacidad de capturar patrones complejos asociados a la calidad de la lectura oral.

En respuesta a estas limitaciones, en los últimos años han surgido enfoques basados en aprendizaje de representaciones, donde modelos entrenados sobre grandes volúmenes de datos permiten obtener *embeddings* acústicos directamente desde la señal de audio. Estos *embeddings* buscan capturar información relevante de manera automática, sin depender de un diseño manual de características.

Entre estos enfoques destaca *Wav2Vec2* (ver Figura 2), un modelo basado en aprendizaje auto supervisado que aprende representaciones del habla a partir de audio crudo [23]. Este modelo es capaz de capturar patrones acústicos y fonéticos de alto nivel, incluso en escenarios con disponibilidad limitada de datos etiquetados, lo que lo hace especialmente adecuado para aplicaciones como la clasificación de la calidad de lectura oral.

2.5. Modelos de aprendizaje automático

Una vez obtenidas representaciones de tamaño fijo para cada muestra de audio, el problema de clasificación de la calidad de lectura oral puede abordarse mediante distintos modelos de aprendizaje automático. En este trabajo se consideran tanto modelos clásicos como enfoques basados en redes neuronales, con el objetivo de analizar su desempeño en escenarios con limitada cantidad de datos.

Entre los modelos clásicos, se incluyen las máquinas de soporte vectorial (*Support Vector Machine*, SVM), bosques aleatorios (*Random Forest*, RF) y *Extreme Gradient Boosting* (XGBoost). Estos enfoques han sido ampliamente utilizados en tareas de clasificación debido a su capacidad de generalización y su relativa robustez en contextos donde la cantidad de datos es limitada.

Las SVM corresponden a modelos discriminativos que buscan encontrar un hiperplano que maximice el margen entre clases en el espacio de características [25]. Mediante el uso de funciones kernel, estas permiten modelar relaciones no lineales entre las variables, lo que resulta especialmente útil en escenarios donde los datos no son linealmente separables.

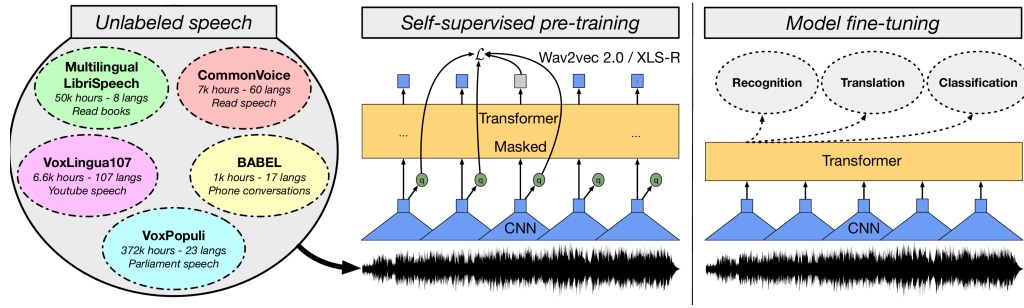


Figura 2: Aprendizaje auto-supervisado multilingüe en Wav2Vec2 usando XLS-R para la obtención de representaciones acústicas del habla a partir de señales de audio.

Fuente: XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale [24].

Por otra parte, los RF corresponden a un método de ensamble basado en la construcción de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos [26]. Este enfoque permite reducir la varianza del modelo y mejorar su capacidad de generalización, siendo particularmente robusto frente a ruido y sobreajuste en comparación con árboles individuales.

En cuanto a XGBoost se basa en la construcción secuencial de modelos débiles que corrigen los errores de los anteriores, optimizando una función de pérdida de manera iterativa [27]. Este tipo de modelos ha mostrado un alto desempeño en diversas tareas de clasificación, aunque su efectividad puede verse limitada en escenarios con alta dimensionalidad y bajo número de muestras si no se realiza un ajuste adecuado de hiperparámetros, ya que es sensible al sobreajuste.

Adicionalmente, se considera el uso de una arquitectura simple de red neuronal, específicamente un perceptrón multicapa, MLP por sus siglas en inglés, utilizado como clasificador sobre los *embeddings* obtenidos. A diferencia de modelos más complejos, este enfoque permite mantener una estructura relativamente simple, reduciendo el riesgo de sobreajuste en contextos con datos limitados, al mismo tiempo que introduce capacidad de modelar relaciones no lineales en el espacio de características.

2.6. Evaluación y metodologías

La evaluación de modelos de aprendizaje automático en escenarios de datos limitados presenta desafíos importantes, particularmente en lo relativo a la estimación confiable de su desempeño y la validez de los resultados obtenidos. En este contexto, resulta fundamental adoptar metodologías experimentales que permitan mitigar problemas como el sobreajuste, la fuga de información y la variabilidad inherente a los procesos de entrenamiento.

Uno de los principales riesgos en este tipo de escenarios es el sobreajuste, el cual ocurre cuando un modelo se ajusta excesivamente a los datos de entrenamiento, capturando patrones específicos que no generalizan a nuevos datos. Este fenómeno se ve acentuado en conjuntos de datos pequeños y en modelos con alta capacidad de representación, lo que hace necesario utilizar estrategias de evaluación que permitan estimar de mejor manera el desempeño esperado en datos no observados.

Una técnica ampliamente utilizada para abordar este problema es la validación cruzada, la cual consiste en dividir el conjunto de datos en múltiples particiones, utilizando iterativamente subconjuntos para entrenamiento y validación. Este enfoque permite aprovechar de mejor forma los datos disponibles y obtener una estimación más robusta del desempeño del modelo en comparación con una única partición fija de entrenamiento y prueba. En particular, en este trabajo se emplea validación cruzada para evaluar de manera consistente los distintos enfoques de modelamiento considerados.

Otro aspecto crítico en el diseño experimental es la prevención de la fuga de información, la cual ocurre cuando información del conjunto de validación o prueba es utilizada, directa o indirectamente, durante el entrenamiento del modelo. Esto puede suceder, por ejemplo, al realizar procesos de normalización, selección de características o extracción de representaciones utilizando la totalidad de los datos antes de la

partición. La presencia de fuga de información puede generar estimaciones de desempeño artificialmente optimistas, comprometiendo la validez de los resultados obtenidos. Por esta razón, es fundamental asegurar que todas las etapas de procesamiento y modelamiento se realicen de manera consistente dentro de cada partición de la validación cruzada.

Adicionalmente, es importante considerar la naturaleza estocástica de muchos algoritmos de aprendizaje automático, los cuales pueden producir resultados distintos en ejecuciones sucesivas debido a factores como la inicialización de parámetros o la selección aleatoria de muestras. Para garantizar la reproducibilidad de los experimentos, es común fijar semillas aleatorias, lo que permite obtener resultados consistentes entre ejecuciones y facilita la comparación entre distintos modelos y configuraciones.

Finalmente, la evaluación del desempeño de los modelos se realiza mediante métricas cuantitativas que permiten comparar de manera objetiva los distintos enfoques considerados. La selección de métricas adecuadas resulta fundamental para interpretar correctamente los resultados, especialmente en escenarios donde pueden existir desbalances entre clases o donde distintos tipos de error tienen implicancias diferentes.

El problema abordado corresponde a una tarea de clasificación multiclase, donde cada muestra de audio es asignada a una de cuatro categorías de calidad de lectura oral. Para evaluar el desempeño de los modelos, se utilizan métricas que permiten capturar distintos aspectos de la capacidad de clasificación, más allá de la proporción global de aciertos.

Una de las métricas consideradas es la exactitud (*accuracy*), que mide la proporción de predicciones correctas sobre el total de muestras evaluadas. Si bien esta métrica entrega una visión general del desempeño, puede resultar insuficiente en escenarios donde existe desbalance entre clases, ya que no distingue entre distintos tipos de error.

Para complementar esta medida, se emplean métricas basadas en precisión (*precision*) y exhaustividad (*recall*), las cuales permiten analizar el desempeño del modelo en cada clase de manera individual. A partir de estas, se calcula la métrica *F1-score*, que corresponde a la media armónica entre *precision* y *recall*, proporcionando una medida balanceada del desempeño del modelo.

Dado que el problema corresponde a una tarea de clasificación multiclase, se consideran distintas formas de agregación del *F1-score*. En particular, el *F1-score* macro calcula el promedio simple del *F1-score* obtenido para cada clase, asignando el mismo peso a todas ellas. Esto permite evaluar el rendimiento del modelo de manera equilibrada entre clases, independientemente de su frecuencia en el conjunto de datos. Por otro lado, el *F1-score* ponderado (*weighted*) incorpora el número de muestras de cada clase en el cálculo del promedio, por lo que entrega una medida más representativa del rendimiento global cuando existe desbalance entre clases.

El uso conjunto de estas métricas permite obtener una evaluación más completa del comportamiento de los modelos, considerando tanto su desempeño global como su capacidad para clasificar adecuadamente cada categoría de interés.

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

A partir de las limitaciones identificadas en la definición del problema, este capítulo presenta la propuesta metodológica desarrollada para abordar la clasificación automática de la calidad de lectura oral en estudiantes de educación básica. La solución se centra en establecer un flujo experimental reproducible que permita comparar distintos enfoques de modelamiento bajo condiciones controladas, considerando especialmente el escenario de datos limitados.

De este modo, el capítulo describe el flujo de trabajo, incluyendo el procesamiento de los audios, la extracción de características, la configuración de los modelos y el diseño experimental utilizado para su posterior evaluación.

3.1. Descripción general del flujo experimental

La propuesta de solución se estructura a partir de un flujo experimental diseñado para evaluar distintos enfoques de clasificación automática de la calidad de lectura oral. Este flujo busca organizar de manera reproducible las etapas necesarias para transformar los registros de audio en representaciones utilizables por modelos de aprendizaje automático, entrenar distintos clasificadores y evaluar su desempeño bajo un protocolo común.

De manera general, el flujo de trabajo comienza con la construcción y caracterización del conjunto de datos, compuesto por registros de lectura oral asociados a una etiqueta de calidad lectora. Posteriormente, los audios son sometidos a una etapa de preprocesamiento, cuyo objetivo es estandarizar su formato y reducir fuentes de variabilidad no relevantes para la tarea de clasificación. Entre estas operaciones se consideran aspectos como la frecuencia de muestreo, el canal de audio y la consistencia de los archivos utilizados.

Una vez preparados los audios, se generan distintas representaciones acústicas. Por una parte, se consideran características tradicionales basadas en MFCC, las cuales buscan representar a nivel de muestra completa mediante estadísticos globales, y concatenadas con descriptores del silencio presente en el audio, con el fin de generar vectores descriptivos de la calidad de lectura oral. Por otra parte, se utilizan representaciones aprendidas obtenidas desde modelos preentrenados de voz, como *Wav2Vec2*, con el objetivo de capturar información acústica de mayor nivel a partir de la señal de audio.

Las representaciones generadas son utilizadas como entrada para distintos modelos de clasificación. En particular, se evalúan modelos clásicos de aprendizaje automático, tales como SVM, RF y XGBoost, junto con un enfoque basado en una red neuronal simple utilizada como clasificador sobre el extractor de los *embeddings*. Esta comparación permite analizar el comportamiento de enfoques de distinta naturaleza bajo un mismo escenario experimental.

Un aspecto central del flujo corresponde al diseño experimental utilizado para la evaluación. Considerando que el conjunto de datos disponible es limitado y que existen múltiples registros asociados a distintos estudiantes o hablantes, se adopta una estrategia de partición que evita la fuga de información. En particular, las divisiones de datos se realizan considerando la separación por locutor, de modo que audios pertenecientes a un mismo hablante no aparezcan simultáneamente en conjuntos de entrenamiento y validación.

Finalmente, el desempeño de los modelos se evalúa mediante un conjunto común de métricas para clasificación multiclase, incluyendo *accuracy*, *recall* y *F1-score* en sus variantes macro y ponderada. Los resultados se reportan de manera agregada sobre las particiones definidas, permitiendo estimar el desempeño promedio de cada enfoque y analizar su variabilidad. De esta forma, el flujo propuesto busca asegurar que la comparación entre modelos sea consistente, reproducible y alineada con las restricciones reales del problema abordado.

3.2. Conjunto de datos

El conjunto de datos utilizado en este trabajo fue recolectado especialmente para el desarrollo del proyecto TeachHear, desarrollado durante la Feria de Software 2025 de la Universidad Técnica Federico Santa María.

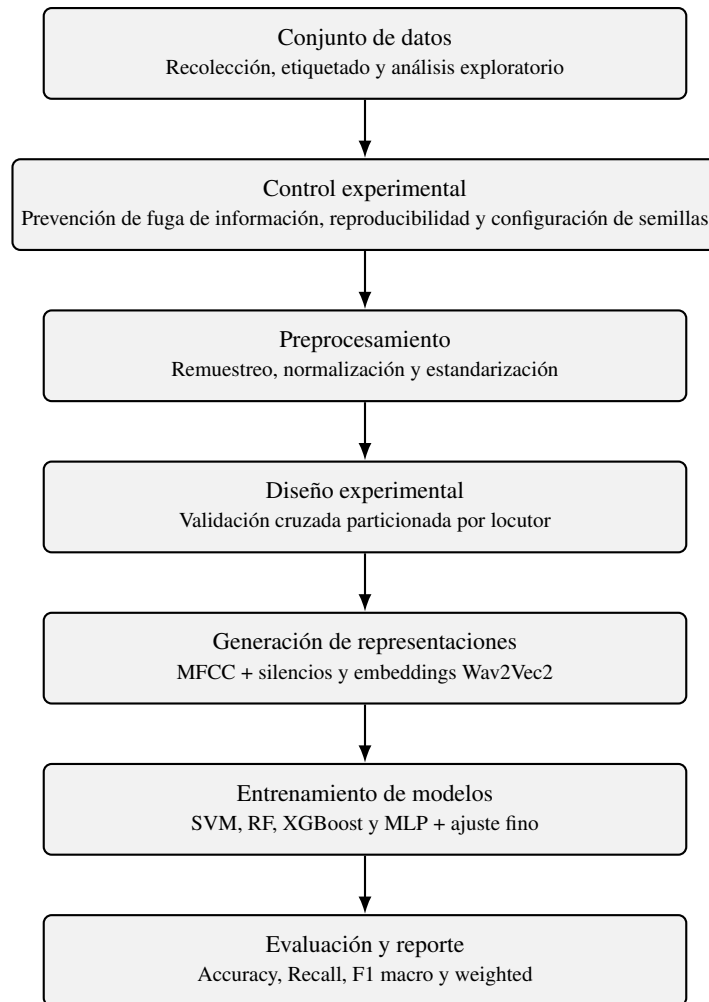


Figura 3: Flujo general de experimentación. La propuesta experimental para la clasificación automática de la calidad de lectura oral.
Fuente: Elaboración propia.

Su construcción tuvo como objetivo disponer de registros reales de lectura oral que permitieran estudiar la clasificación automática de la calidad de lectura en estudiantes de educación básica.

La recolección se realizó en el Colegio GEA de Quilpué, considerando estudiantes pertenecientes a los niveles de tercero, cuarto y quinto básico. Las grabaciones fueron efectuadas durante la instancia de evaluación semestral del dominio lector realizada por la jefa de unidad técnico pedagógica del establecimiento, lo que permitió capturar audios en un contexto educativo real y asociado a una práctica de evaluación ya existente dentro del colegio.

Las grabaciones fueron realizadas utilizando distintos dispositivos móviles, lo que introduce variabilidad en las condiciones de captura, tales como diferencias en la calidad del micrófono, volumen de grabación, distancia al hablante y presencia de ruido ambiente. Esta característica representa un desafío adicional para el modelamiento, pero también permite aproximar el conjunto de datos a condiciones de uso más realistas.

En total, participaron 71 estudiantes, distribuidos entre los cursos de tercero, cuarto y quinto básico. La Tabla 2 presenta la cantidad de participantes por nivel educativo.

Las etiquetas asociadas a cada audio corresponden a las categorías de calidad de lectura oral definidas en la evaluación del dominio lector. En este trabajo se consideran cuatro categorías clasificables: *Lectura*

Tabla 2: Distribución de participantes por curso. En total, se cuenta con 71 estudiantes, distribuidos entre tercero, cuarto y quinto básico. Se observa una mayor cantidad de participantes en cuarto básico, seguido por tercero básico y, en menor proporción, quinto básico.

Fuente: Elaboración propia.

Curso	Cantidad de estudiantes
Tercero básico	24
Cuarto básico	32
Quinto básico	15
Total	71

silábica, *Lectura palabra a palabra*, *Lectura por unidades cortas* y *Lectura fluida*. Estas categorías son las que se pueden identificar a partir de los audios, debido a que los alumnos que poseían un nivel de *No lector* fueron descartados, dado a que sus registros acústicos eran principalmente silencio, decisión que disminuye el costo computacional de la plataforma TeachHear.

A raíz que el conjunto de datos fue recolectado en un contexto educativo específico y con un número limitado de participantes, su uso presenta ciertas restricciones metodológicas. En particular, el tamaño reducido del conjunto de datos, la posible distribución desigual entre categorías y la variabilidad asociada a los dispositivos de grabación deben ser consideradas al momento de diseñar el protocolo experimental y evaluar el desempeño de los modelos. Por esta razón, se recopiló grabaciones extras con 2 voluntarios para suplir la diferencias en la distribuciones de categorías, resultando en un total de 73 locutores dentro del conjunto de datos.

3.3. Procesamiento de audios

Con el fin de obtener un conjunto de datos consistente y adecuado para las etapas posteriores de extracción de características y entrenamiento de modelos, se aplicó una etapa de procesamiento básico sobre las grabaciones recolectadas. Esta etapa tuvo como objetivo principal estandarizar el formato de los audios y definir una unidad de análisis homogénea para el desarrollo experimental.

En primer lugar, cada grabación original fue segmentada secuencialmente en fragmentos de una duración máxima de 30 segundos. Este procedimiento se realizó recorriendo la grabación completa hasta agotar su contenido, generando tantos fragmentos como fuese necesario. Siempre que fue posible, los cortes se realizaron en zonas de silencio, evitando interrumpir palabras o segmentos de lectura. De esta forma, se buscó preservar la continuidad de la señal de voz dentro de cada fragmento y reducir discontinuidades artificiales introducidas por el proceso de procesamiento.

La decisión de limitar la duración de los fragmentos a un máximo de 30 segundos responde tanto a criterios metodológicos como a consideraciones prácticas asociadas al procesamiento automático de habla. En particular, esta duración se encuentra alineada con arquitecturas modernas de reconocimiento automático del habla, como *Whisper*, cuyo procesamiento se basa en dividir el audio de entrada en segmentos de 30 segundos, convertirlos en espectrogramas log-Mel y posteriormente entregarlos al *encoder* del modelo [28]. De esta forma, el uso de fragmentos de duración acotada permite mantener compatibilidad conceptual con flujos modernos de procesamiento de voz.

Por otra parte, la segmentación en fragmentos de hasta 30 segundos también resulta conveniente para la extracción de representaciones acústicas mediante modelos preentrenados como *Wav2Vec2*. Aunque este tipo de modelos opera directamente sobre la forma de onda, el uso de audios de duración acotada permite reducir el costo computacional, limitar el uso de memoria y obtener unidades de análisis más homogéneas para la posterior agregación a nivel de fragmento [23]. En consecuencia, la segmentación aplicada no solo estandariza la duración máxima de las muestras, sino que también favorece un procesamiento más estable y reproducible dentro del flujo experimental.

Posteriormente, cada fragmento fue estandarizado mediante *ffmpeg* para asegurar un formato común entre todas las muestras. En particular, los audios fueron convertidos a formato *.wav*, con una frecuencia de muestreo de 16 kHz, un solo canal de audio (mono) y codificación PCM de 16 bits (*pcm_s16le*). Esta

transformación permite reducir fuentes de variabilidad asociadas al formato original de grabación y asegurar compatibilidad con las herramientas y modelos utilizados en el flujo experimental.

Cabe destacar que este corresponde al único procesamiento aplicado sobre los audios previo a la etapa de extracción de características. No se realizaron procedimientos adicionales de limpieza, eliminación de ruido, detección de voz ni aumento de datos. Esta decisión busca mantener el conjunto de datos lo más cercano posible a las condiciones reales de captura en las que fue recolectado.

Como resultado de este proceso, se obtuvo un conjunto de fragmentos de audio homogéneos en formato y preparados para su uso en las siguientes etapas del flujo experimental. La Figura 4 presenta la distribución de etiquetas en el conjunto de fragmentos procesados, mientras que la Figura 5 muestra la distribución de duraciones de los fragmentos obtenidos tras la segmentación.

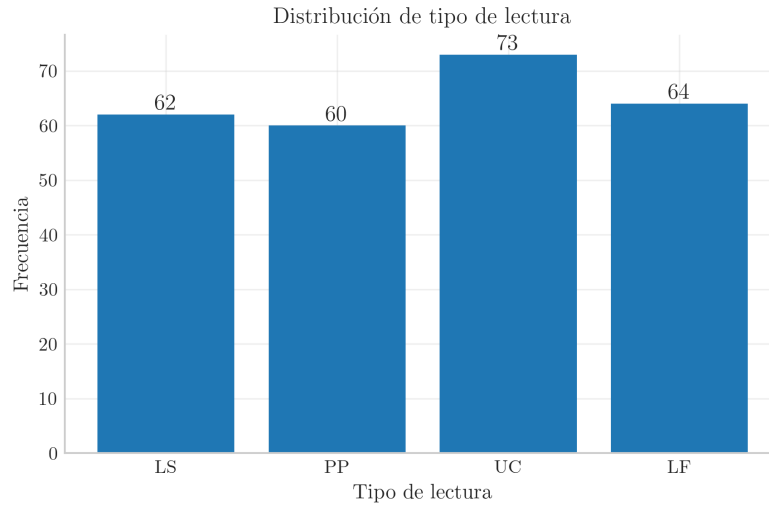


Figura 4: Distribución de las etiquetas de calidad de lectura oral en los 259 segmentos resultantes del procesamiento de los audios. Se observa una mayor presencia de segmentos etiquetados como *Unidades Cortas* (UC), en comparación con las demás categorías: *No Lector* (NL), *Lectura Silábica* (LS), *Lectura Palabra a Palabra* (LPP) y *Lectura Fluida* (LF).

Fuente: Elaboración propia.

3.4. Generación de representaciones

Una vez procesados los audios, la siguiente etapa del flujo consiste en transformar cada fragmento en una representación numérica adecuada para ser utilizada por los modelos de clasificación. En este trabajo se consideran dos familias de representaciones: por una parte, descriptores acústicos clásicos basados en características diseñadas manualmente; y por otra, *embeddings* obtenidos desde modelos profundos preentrenados para procesamiento de habla.

3.4.1. Representaciones acústicas clásicas

Como primera aproximación, se construyó un conjunto de características acústicas clásicas a partir de cada fragmento de audio. Estas características buscan capturar información relacionada con la forma espectral de la señal, la energía del habla, la duración del fragmento y la presencia de pausas. Este enfoque permite obtener vectores de tamaño fijo, interpretables y de bajo costo computacional, adecuados para modelos clásicos de aprendizaje automático.

En primer lugar, se calcularon 13 coeficientes MFCC a nivel de *frame*. Posteriormente, con el objetivo de obtener una representación a nivel de fragmento completo, se aplicó una agregación estadística mediante el cálculo de la media y la desviación estándar de cada coeficiente a lo largo del tiempo. Esta estrategia

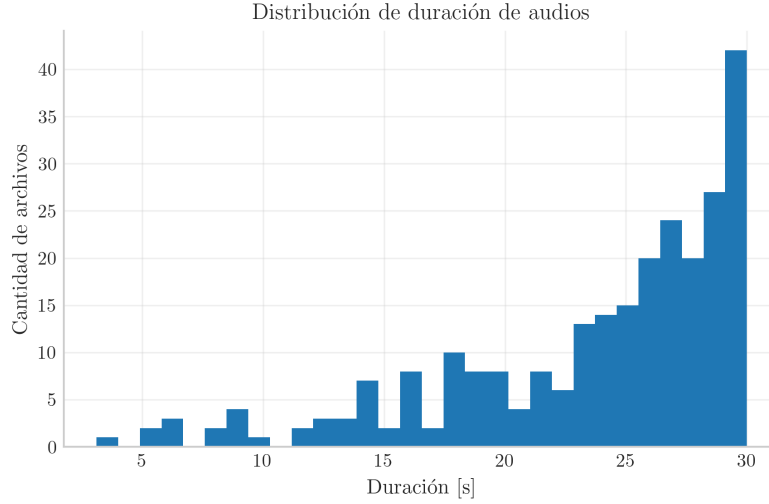


Figura 5: Distribución de la duración, en segundos, de los segmentos resultantes del procesamiento de los audios. La mayoría de los fragmentos se concentra entre los 20 y 30 segundos, con una menor presencia de segmentos de corta duración.

Fuente: Elaboración propia.

permite transformar una secuencia temporal de características en un vector de tamaño fijo.

Adicionalmente, se incorporaron características asociadas a la energía de la señal mediante el cálculo del valor cuadrático medio (*Root Mean Square*, RMS). En particular, se consideró la media y desviación estándar del RMS a lo largo del fragmento, con el fin de capturar información relacionada con la intensidad promedio y la variabilidad energética de la lectura.

Junto con lo anterior, se incluyó la duración total del fragmento como una característica adicional. Si bien los audios fueron segmentados con una duración máxima de 30 segundos, no todos los fragmentos poseen la misma extensión, por lo que esta variable permite representar diferencias temporales relevantes, especialmente en fragmentos finales o segmentos más breves.

Finalmente, se incorporaron características asociadas a pausas mediante un detector de actividad de voz (*Voice Activity Detection*, VAD). En particular, se utilizó WebRTC VAD, el cual clasifica segmentos cortos de audio como voz o no voz, permitiendo identificar regiones de silencio o ausencia de habla [29]. Para este procedimiento, cada audio fue analizado en ventanas de 20 milisegundos, utilizando el modo de mayor agresividad del detector. A partir de los segmentos clasificados como no voz se calcularon cinco características: número de pausas, tiempo total de pausa, duración media de pausa, proporción de pausa respecto de la duración total del fragmento y una bandera binaria que indica si la proporción de pausa supera un umbral mínimo. La decisión de incorporar la bandera binaria corresponde a problemas con el cálculo de las características de silencio debido al considerable ruido que presentaban algunas grabaciones, ruido externo que es natural en ambientes educativos, incluso cuando la prueba se realiza en salas con ruido controlado como lo puede ser la biblioteca.

En conjunto, la representación clásica queda compuesta por 34 características: 26 provenientes de los MFCC agregados mediante media y desviación estándar, 2 características RMS, 1 característica de duración y 5 características asociadas a pausas. Esta representación busca combinar información espectral, energética y temporal relevante para la caracterización de la calidad de lectura oral.

3.4.2. Embeddings obtenidos mediante modelos preentrenados

Como segunda familia de representaciones, se utilizaron *embeddings* acústicos obtenidos desde un modelo profundo preentrenado. A diferencia de las características clásicas, que dependen de transformaciones definidas manualmente, los *embeddings* aprendidos buscan capturar patrones acústicos y lingüísticos de mayor

nivel a partir de grandes volúmenes de datos.

En particular, se utilizó una versión de *Wav2Vec2* basada en XLS-R, una arquitectura de aprendizaje auto-supervisado multilingüe diseñada para aprender representaciones de habla a gran escala. XLS-R extiende el enfoque de *Wav2Vec2* a un escenario multilingüe (ver Figura 2), entrenando modelos con datos de habla provenientes de 128 idiomas y con tamaños de hasta 2 mil millones de parámetros [24]. En este trabajo se empleó la versión *wav2vec2-large-xlsr-53-spanish* de Jonatas Grosman, disponible en *Hugging Face*, ajustada para reconocimiento automático del habla en español [30].

La elección de un modelo XLS-R ajustado para español responde a la necesidad de obtener representaciones más adecuadas al dominio lingüístico del conjunto de datos. Dado que las grabaciones corresponden a estudiantes chilenos leyendo en español, el uso de un modelo preentrenado y posteriormente ajustado en este idioma permite aprovechar conocimiento acústico y fonético más cercano al habla analizada.

Para cada fragmento de audio, el modelo genera una secuencia de representaciones internas a lo largo del tiempo. Dado que los modelos clásicos utilizados posteriormente requieren entradas de tamaño fijo, estas representaciones temporales fueron agregadas a nivel de fragmento mediante una estrategia de *pooling*. De este modo, cada audio queda representado por un único vector de *embeddings*, el cual resume la información acústica capturada por el modelo preentrenado. Las estrategias de *pooling* utilizadas corresponden a la media y la desviación estándar del *last hidden state* del modelo, el cual corresponde a la última capa del *encoder*, generando así un vector fijo de dimensión 2048 para cada registro de audio.

El uso de *embeddings* permite comparar un enfoque basado en representaciones aprendidas con otro basado en descriptores acústicos clásicos. Esta comparación resulta relevante en el contexto del presente trabajo, ya que permite analizar si modelos preentrenados de habla pueden aportar información útil para la clasificación de calidad lectora en un escenario con disponibilidad limitada de datos etiquetados.

3.5. Modelos de aprendizaje automático

Una vez generadas las representaciones descritas en la sección anterior, se definieron distintos enfoques de modelamiento para abordar la clasificación automática de la calidad de lectura oral. La selección de modelos no solo busca comparar clasificadores, sino también analizar la capacidad informativa de las representaciones utilizadas. En particular, se busca determinar si las características acústicas clásicas o los *embeddings* obtenidos desde modelos preentrenados entregan una representación más adecuada para esta tarea.

Para ello, se consideran tres enfoques experimentales. El primero utiliza representaciones acústicas clásicas, compuestas por MFCC agregados, características de energía, duración y pausas. El segundo utiliza *embeddings* obtenidos desde un modelo *Wav2Vec2/XLS-R* preentrenado y ajustado para reconocimiento automático del habla en español. En ambos casos, las representaciones son evaluadas utilizando los mismos modelos clásicos de aprendizaje automático: SVM, RF y XGBoost. Esta estrategia permite comparar ambas familias de características bajo condiciones similares, reduciendo el efecto del clasificador sobre la interpretación de los resultados.

El tercer enfoque corresponde a un modelo basado en ajuste fino parcial de *Wav2Vec2/XLS-R*. A diferencia de los enfoques anteriores, donde las representaciones se extraen previamente y luego son entregadas a un clasificador externo, en este caso se permite adaptar parcialmente las últimas capas del *encoder* junto con un clasificador basado en una red neuronal. Se busca otorgar mayor capacidad de adaptación al modelo, aproximándose a una estrategia de tipo *end-to-end*, pero manteniendo controlado el número de parámetros entrenables debido al tamaño limitado del conjunto de datos.

3.5.1. Modelos clásicos sobre representaciones fijas

Los modelos clásicos de aprendizaje automático se utilizan sobre dos tipos de representaciones fijas: características acústicas clásicas y *embeddings Wav2Vec2/XLS-R* agregados a nivel de fragmento. En ambos casos, cada audio queda representado mediante un vector de tamaño fijo, lo que permite utilizar clasificadores tradicionales y comparar directamente el aporte de cada representación.

En este trabajo se consideran tres modelos clásicos: máquinas de soporte vectorial (SVM), Random Forest y XGBoost. La elección de estos modelos responde a su uso extendido en tareas de clasificación y

a su capacidad para trabajar con conjuntos de datos de tamaño reducido. Además, al aplicar los mismos modelos sobre ambas representaciones, es posible analizar si las diferencias de desempeño se explican principalmente por la riqueza de la representación utilizada.

La elección de hiperparámetros en los modelos clásicos no busca maximizar exhaustivamente el desempeño de cada clasificador, sino establecer configuraciones comparables y reproducibles que permitan analizar el comportamiento relativo de las representaciones evaluadas. Por esta razón, se utilizan configuraciones fijas (ver Figura 3) a lo largo de las particiones de validación, evitando una búsqueda agresiva de hiperparámetros que pudiera ajustar en exceso los modelos al conjunto disponible.

Tabla 3: Hiperparámetros utilizados en los modelos tradicionales de aprendizaje automático. Los parámetros quedan fijos independiente de el enfoque utilizado.

Fuente: Elaboración propia.

Modelo	Hiperparámetro	Valor
SVM	<i>C</i>	1.0
SVM	kernel	linear
SVM	probability	True
RF	n_estimators	300
RF	max_depth	10
RF	n_jobs	Según disponibilidad computacional
XGBoost	n_estimators	200
XGBoost	learning_rate	0.03
XGBoost	max_depth	3
XGBoost	subsample	0.8
XGBoost	colsample_bytree	0.3
XGBoost	min_child_weight	3
XGBoost	reg_lambda	5.0
XGBoost	reg_alpha	0.1
XGBoost	eval_metric	mlogloss
XGBoost	n_jobs	1
XGBoost	tree_method	hist
XGBoost	max_bin	64

3.5.2. Ajuste fino parcial con clasificador neuronal

Como tercer enfoque, se considera un modelo basado en ajuste fino parcial del *encoder* de *Wav2Vec2/XLS-R*. Este enfoque se diferencia de los anteriores en que el modelo preentrenado no se utiliza únicamente como extractor de *embeddings*, sino que parte de sus parámetros son actualizados durante el entrenamiento para la tarea específica de clasificación de calidad lectora.

Dado que el conjunto de datos disponible es limitado, no se realiza un ajuste fino completo del modelo. En su lugar, se congela aproximadamente el 90% del *encoder*, dejando entrenables los parámetros de las últimas capas junto con el clasificador neuronal. Esta decisión busca equilibrar dos objetivos: por una parte, aprovechar las representaciones generales aprendidas durante el preentrenamiento y el ajuste al español; y por otra, permitir que las capas superiores del modelo se adapten a los patrones específicos de la tarea.

El clasificador corresponde a una arquitectura neuronal, encargada de transformar la representación generada por el *encoder* en una predicción sobre las cuatro categorías de calidad de lectura oral consideradas. Al mantener una arquitectura simple y limitar el número de capas entrenables, se busca reducir el riesgo de sobreajuste, manteniendo al mismo tiempo una mayor capacidad de adaptación que en los enfoques basados únicamente en representaciones congeladas.

El clasificador se compone por 3 capas (ver Figura 6); una capa densa, con dimensión de entrada y salida igual al *hidden state size* del *encoder*; una capa de *dropout*, con porcentaje de 30%; y finalmente la capa densa de clasificación, que reduce de la dimensión de la capa anterior a las 4 categorías a clasificar.

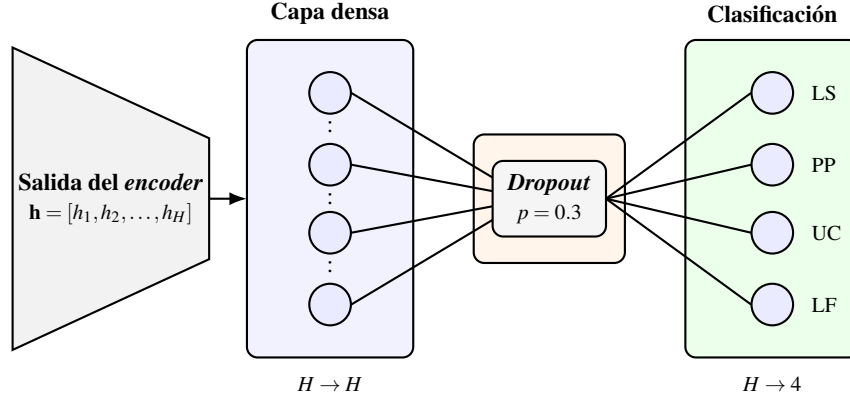


Figura 6: Arquitectura del clasificador utilizada sobre la salida del *encoder*. La representación de entrada corresponde a un vector de dimensión H , seguido por una capa densa que preserva la dimensionalidad ($H \rightarrow H$), una capa de *dropout* con probabilidad de 0.3 y una capa densa final de clasificación que proyecta la salida a las cuatro categorías objetivo.
Fuente: Elaboración propia.

Este tercer enfoque permite evaluar si la adaptación supervisada parcial del modelo preentrenado entrega beneficios adicionales respecto del uso de *embeddings* congelados. De esta forma, la comparación entre los tres enfoques permite orientar futuras líneas de trabajo, distinguiendo si los esfuerzos deberían centrarse en mejorar las representaciones acústicas, en utilizar *embeddings* preentrenados como características fijas o en avanzar hacia estrategias de ajuste fino más cercanas a modelos *end-to-end*.

3.5.3. Resumen de enfoques considerados

La Tabla 4 resume los enfoques de modelamiento considerados en este trabajo, indicando la representación utilizada y los modelos aplicados en cada caso.

Tabla 4: Resumen de enfoques de modelamiento considerados.
Fuente: Elaboración propia.

Enfoque	Representación	Modelos
Representación acústica clásica	MFCC, RMS, duración y pausas	SVM, Random Forest, XGBoost
Embeddings congelados	Embeddings Wav2Vec2/XLS-R	SVM, Random Forest, XGBoost
Ajuste fino parcial	Encoder Wav2Vec2/XLS-R parcialmente entrenable	MLP y cabeza de clasificación

3.6. Diseño experimental

El diseño experimental utilizado busca evaluar los distintos enfoques de modelamiento bajo condiciones controladas y comparables. En este se considera el tamaño limitado del conjunto de datos y la existencia de diferentes fragmentos de audio de un mismo estudiante, con el fin de reducir sesgos, fuga de información y obtener una estimación más confiable del desempeño de los enfoques propuestos.

En el presente trabajo, todos los enfoques son evaluados utilizando las mismas particiones de datos y grupo de métricas. Así, las diferencias en los resultados pueden atribuirse tanto a la representación y modelo utilizado.

3.6.1. Unidad de partición

Es fundamental considerar y determinar lo que se entiende como unidad al momento de particionar los datos. Debido a que una misma grabación puede generar más de un fragmento de audio al momento del procesamiento, la voz de un estudiante puede estar asociada a más de uno de estos. En este escenario, fragmentos provenientes del mismo locutor podrían quedar simultáneamente en ambos conjuntos, permitiendo que el modelo aprenda características particulares del hablante en lugar de patrones generales asociados a la calidad de lectura oral.

Para evitar esta fuga de información, la partición de datos se realiza a nivel de locutor. Esto se traduce en que los fragmentos de un mismo estudiante son asignados exclusivamente a una misma partición. De esta forma, durante la evaluación, el modelo es probado sobre estudiantes no observados en el entrenamiento, lo que pone a prueba la capacidad de generación de modelo.

3.6.2. Validación cruzada

Debido a la cantidad reducida de datos disponibles, se utiliza validación cruzada como estrategia de evaluación. Con esta técnica se aprovecha de mejor manera el conjunto de datos, entrenando y evaluando los modelos en diferentes particiones en lugar de depender de una única partición fija de entrenamiento y prueba.

Para cada iteración de la validación cruzada, un subconjunto de locutores es reservado para validación mientras el resto se utiliza para el entrenamiento. Cada partición corresponde a una división específica del conjunto de locutores entre entrenamiento y validación. Este proceso se repite hasta que todas las particiones hayan sido utilizadas como conjunto de validación. Finalmente, los reportes de desempeño se calculan de forma agregada a partir de los resultados obtenidos en cada partición.

Para el tercer enfoque se emplea una validación cruzada anidada, con el fin de monitorear el entrenamiento de la red neuronal y tomar decisiones de diseño sin caer en resultados optimistas.

3.6.3. Control de fuga de información

El protocolo experimental considera reducir las distintas fuentes de fuga de información. En primer lugar, la separación por locutor evita que fragmentos de un mismo estudiante aparezca en el conjunto de entrenamiento y validación, asegurando que los modelos no incorporen características propias del hablante, como timbre, condiciones de grabación y pronunciación.

En segundo lugar, las transformaciones realizadas a los datos en el momento de entrenamiento se ajustan dentro de cada partición. Así, los modelos que requieren estandarización de características, los parámetros obtenidos para el escalado, se calculan únicamente sobre el conjunto de entrenamiento y luego se aplican sobre el conjunto de validación. Con esto se logra evitar filtrar información estadística del conjunto de validación sobre el conjunto de entrenamiento.

Finalmente, la comparación de modelos se realizan sobre las mismas particiones.

3.6.4. Configuración de los experimentos

El diseño experimental contempla tres enfoques principales: modelos tradicionales entrenados sobre características acústicas clásicas, los mismos modelos entrenados sobre *embeddings* obtenidos desde *Wav2Vec2/XLS-R*, y un enfoque de ajuste fino parcial del *encoder* junto con un clasificador neuronal. Esta comparación permite evaluar el aporte relativo de representaciones acústicas diseñadas manualmente, representaciones preentrenadas congeladas y representaciones adaptadas supervisadamente a la tarea. Para asegurar condiciones comparables y reproducibles, los hiperparámetros de cada modelo se mantienen fijos a través de las particiones, sin buscar una optimización exhaustiva de cada clasificador.

3.6.5. Reproducibilidad

Con el fin de asegurar la reproducibilidad de los experimentos, se fijan semillas aleatorias en las etapas que involucran procesos estocásticos, tales como la generación de particiones, la inicialización de mode-

los y ciertos procedimientos de entrenamiento. Además, se registra la configuración utilizada para cada experimento, incluyendo representación de entrada, modelo, hiperparámetros y partición correspondiente.

Este control resulta necesario debido a que distintos modelos pueden producir resultados variables entre ejecuciones, especialmente en escenarios con pocos datos. La fijación de semillas permite reducir esta variabilidad y facilita la comparación entre enfoques bajo condiciones experimentales equivalentes.

3.6.6. Reporte de resultados

El desempeño de los modelos se reporta a partir de las métricas definidas para clasificación multiclase, incluyendo *accuracy*, *recall* y *F1-score* en sus variantes macro y ponderada. Para cada modelo, estas métricas se calculan en cada partición de validación y posteriormente se agregan mediante promedio y desviación estándar.

3.7. Protocolo de evaluación

Si bien el desempeño de los modelos se estima mediante validación cruzada, se define un protocolo de evaluación común para establecer cómo se calculan, agregan e interpretan los resultados obtenidos en cada partición. La evaluación se realiza a nivel de fragmento de audio, considerando cada segmento procesado como una muestra individual asociada a una etiqueta de calidad lectora. Sin embargo, dado que distintos fragmentos pueden provenir de un mismo estudiante, la partición se realiza por locutor, evitando que segmentos de un mismo participante aparezcan simultáneamente en entrenamiento y validación.

En cada partición, el modelo se entrena únicamente con los datos del conjunto de entrenamiento y se evalúa sobre el conjunto de validación correspondiente. A partir de las predicciones obtenidas, se calculan métricas de clasificación multiclase, incluyendo *accuracy*, *recall* y *F1-score* en sus variantes macro y ponderada. Mientras que la *accuracy* entrega una medida general de predicciones correctas, el *F1-score* macro permite evaluar el desempeño de forma equilibrada entre clases y el *F1-score* ponderado refleja el rendimiento global considerando la cantidad de muestras por categoría.

Adicionalmente, se utilizan matrices de confusión y reportes de clasificación por clase (ver Anexo) para analizar los errores cometidos por los modelos, especialmente en confusiones entre niveles cercanos de calidad lectora. Finalmente, los resultados de cada partición se agregan mediante el promedio y la desviación estándar de las métricas. El promedio resume el desempeño general, mientras que la desviación estándar permite observar la variabilidad entre particiones, aspecto relevante en escenarios con datos limitados y particiones a nivel de locutor que pueden generar conjuntos de validación desbalanceados.

CAPÍTULO 4

VALIDACIÓN DE LA SOLUCIÓN

En este capítulo se presentan y analizan los resultados obtenidos durante la validación experimental de la propuesta de solución. Esta validación tiene como propósito evaluar los diferentes enfoques propuestos con el fin de clasificar automáticamente la calidad de lectura oral en estudiantes de enseñanza básica a partir de registros de audio, considerando los niveles definidos en el contexto chileno para la evaluación del dominio lector.

La evaluación se estructura en torno a tres enfoques experimentales. El primero corresponde al uso de modelos clásicos de aprendizaje automático entrenados sobre características acústicas extraídas manualmente desde los audios. El segundo enfoque utiliza representaciones obtenidas a partir de modelos preentrenados de habla, las cuales son posteriormente clasificadas mediante los mismos modelos. Finalmente, el tercer enfoque considera una estrategia de transferencia de aprendizaje mediante el ajuste parcial de un *encoder* preentrenado, complementado con un clasificador neuronal.

La comparación entre estos enfoques permite analizar no solo el desempeño predictivo de cada alternativa, sino también el aporte de distintas representaciones del audio para la tarea de clasificación. En particular, se busca determinar si las características acústicas tradicionales son suficientes para capturar diferencias relevantes entre niveles de lectura, o si las representaciones aprendidas por modelos preentrenados entregan una base más rica para abordar el problema. Asimismo, se evalúa si el ajuste parcial del modelo permite mejorar el desempeño respecto al uso de *embeddings* congelados, considerando las restricciones propias de un escenario con una cantidad reducida de datos.

A lo largo del capítulo, los resultados se presentan mediante métricas globales y matrices de confusión, con el objetivo de entregar una visión del comportamiento de los modelos. Además, se analiza la variabilidad de los resultados entre particiones de validación cruzada, debido al tamaño acotado del conjunto de datos y a la necesidad de evitar fuga de información entre audios pertenecientes a un mismo estudiante.

El detalle de métricas por clase de calidad de lectura oral para cada modelo evaluado se presenta en el Anexo, con el fin de complementar los resultados globales reportados en esta sección.

4.1. Resumen del protocolo de validación

Los resultados presentados en este capítulo fueron obtenidos siguiendo el protocolo experimental definido en el capítulo anterior. En particular, todos los enfoques fueron evaluados utilizando las mismas particiones de validación cruzada, con el objetivo de asegurar una comparación consistente entre modelos y representaciones.

La evaluación se realizó mediante una estrategia de validación cruzada estratificada y agrupada por locutor. Esta decisión permite mantener, en la medida de lo posible, la distribución de clases en cada partición, evitando al mismo tiempo que registros pertenecientes a un mismo estudiante aparezcan simultáneamente en los conjuntos de entrenamiento y evaluación. De esta forma, los resultados reflejan el desempeño de los modelos frente a locutores no observados durante el entrenamiento.

El desempeño fue evaluado mediante métricas globales y por clase, considerando principalmente *accuracy*, *recall* y *F1-score*. Para las métricas de clases se reportan promedios macro y ponderados. Adicionalmente, se utilizaron matrices de confusión para analizar los principales patrones de error entre niveles de lectura.

Dado el tamaño acotado del conjunto de datos, los resultados se interpretan considerando tanto el promedio como la variabilidad entre folds. Esto permite analizar no solo qué enfoque obtiene mejor desempeño, sino también qué tan estable resulta bajo distintas particiones de evaluación.

4.2. Resultados del enfoque basado en características acústicas

El primer enfoque evaluado corresponde al uso de características acústicas extraídas manualmente desde los registros de audio, combinadas con modelos clásicos de aprendizaje automático. Este enfoque permite

establecer una línea base inicial para la tarea de clasificación, utilizando representaciones tradicionalmente empleadas en problemas de análisis de habla y audio.

El objetivo principal de esta etapa no es únicamente obtener un desempeño competitivo, sino también analizar si una representación acústica de baja dimensionalidad relativa es suficiente para capturar diferencias relevantes entre los niveles de calidad lectora. De esta manera, los resultados obtenidos en este enfoque sirven como punto de comparación frente a representaciones más complejas, como los embeddings generados por modelos preentrenados de habla.

4.2.1. Configuración experimental

Para este enfoque se utilizaron características acústicas extraídas desde cada audio preprocesado, con el objetivo de representar propiedades espectrales y temporales asociadas a la lectura oral. En particular, se consideraron 13 coeficientes MFCC agregados mediante estadísticos sobre el eje temporal, junto con características relacionadas con pausas y segmentos de silencio, permitiendo obtener una representación fija por audio.

Sobre esta representación se evaluaron modelos clásicos de aprendizaje automático, específicamente Support Vector Machine (SVM), Random Forest (RF) y XGBoost (XGB).

Dado el tamaño reducido del conjunto de datos, no se realizó una búsqueda exhaustiva de hiperparámetros, sino que se utilizaron configuraciones razonables y comparables entre modelos. De esta forma, los resultados de este enfoque se interpretan como una línea base controlada para evaluar el poder discriminativo de las características acústicas, manteniendo el mismo protocolo de validación cruzada descrito previamente.

4.2.2. Resultados globales

La Tabla 5 presenta los resultados obtenidos por los modelos clásicos entrenados sobre características acústicas. Se reportan las métricas promedio obtenidas durante la validación cruzada, junto con su desviación estándar entre particiones.

La Figura 7 presenta las matrices de confusión obtenidas para los modelos evaluados bajo el enfoque de características acústicas. Cada matriz representa el desempeño agregado a partir de las predicciones realizadas sobre los conjuntos de prueba de todos las particiones de la validación cruzada.

Tabla 5: Resultados globales del enfoque basado en características acústicas. XGBoost obtiene los mejores resultados promedio en todas las métricas evaluadas, aunque las desviaciones estándar muestran una variabilidad relevante entre particiones.

Fuente: Elaboración propia.

Modelo	Accuracy	F1 macro	F1 ponderado	Recall macro
SVM	0.460 $\pm$ 0.075	0.373 $\pm$ 0.113	0.425 $\pm$ 0.090	0.395 $\pm$ 0.096
Random Forest	0.473 $\pm$ 0.128	0.359 $\pm$ 0.165	0.417 $\pm$ 0.138	0.396 $\pm$ 0.154
XGBoost	0.510 $\pm$ 0.100	0.384 $\pm$ 0.152	0.449 $\pm$ 0.115	0.424 $\pm$ 0.137

4.2.3. Análisis de resultados

Los resultados obtenidos con características acústicas muestran que esta representación permite, en cierta medida, capturar algunos patrones relevantes para la clasificación de la calidad lectora. Esto es esperable, ya que variables asociadas al espectro de la señal, la presencia de pausas y la continuidad del habla pueden reflejar diferencias entre estudiantes con distintos niveles de fluidez lectora.

No obstante, el desempeño alcanzado también evidencia las limitaciones de este enfoque. Al trabajar con estadísticos agregados sobre el audio completo, se pierde parte de la información temporal fina presente en la lectura. Esto puede ser especialmente relevante en una tarea donde la calidad lectora no depende únicamente de propiedades globales de la señal, sino también de la secuencia, el ritmo, las interrupciones y la forma en que el estudiante articula unidades de lectura a lo largo del tiempo.

Además, las clases consideradas poseen una relación progresiva entre sí. Por esta razón, las diferencias entre categorías adyacentes pueden ser sutiles desde el punto de vista acústico. En particular, a este enfoque se le dificulta distinguir entre niveles contiguos, sin embargo, también posee dificultades para separar casos extremos. Esta característica del problema puede limitar el desempeño de modelos que operan sobre descriptores acústicos manuales, debido a que las fronteras entre clases no necesariamente se expresan mediante diferencias simples en las características extraídas.

Estos resultados permiten establecer un punto de referencia, para evaluar si representaciones aprendidas mediante modelos preentrenados de habla entregan información adicional para la tarea. Si los enfoques posteriores muestran mejoras consistentes sobre esta línea base, ello sugeriría que la clasificación de la calidad lectora requiere representaciones más ricas que las características acústicas manuales utilizadas en esta etapa.

4.3. Resultados del enfoque basado en embeddings preentrenados

El segundo enfoque evaluado corresponde al uso de representaciones aprendidas mediante modelos preentrenados de habla. A diferencia del enfoque basado en características acústicas manuales, esta estrategia utiliza *embeddings* extraídos desde un *encoder* previamente entrenado sobre grandes volúmenes de datos de audio. El objetivo de este enfoque es analizar si dichas representaciones contienen información más rica y discriminativa para la clasificación de la calidad lectora.

En este caso, cada registro de audio es procesado por el modelo preentrenado *Wav2Vec2/XLS-R* para obtener una representación vectorial de alto nivel. Posteriormente, estos *embeddings* son utilizados como entrada para distintos clasificadores. De esta manera, se mantiene congelado el modelo base y se evalúa únicamente la capacidad de los clasificadores para separar los niveles de lectura a partir de la representación obtenida.

Este enfoque cumple un rol intermedio dentro de la validación experimental. Por una parte, permite comparar directamente el aporte de los *embeddings* frente a las características acústicas tradicionales. Por otra parte, permite evaluar el desempeño de representaciones preentrenadas sin modificar los pesos del *encoder*, lo que reduce el riesgo de sobreajuste en un escenario de datos reducidos.

4.3.1. Configuración experimental

Para este enfoque se utilizó un modelo preentrenado y ajustado al idioma español como extractor de características, generando *embeddings* a partir de cada audio preprocesado. Durante este proceso, el *encoder* se mantuvo congelado, permitiendo aprovechar el conocimiento previamente aprendido por el modelo sin adaptar sus pesos al conjunto de datos de este trabajo.

Dado que la salida del *encoder* corresponde a una secuencia temporal de representaciones, se aplicaron estrategias de *pooling* mediante la media y la desviación estándar del *last hidden state*. Con ello, cada audio fue representado mediante un vector fijo de dimensión 2048.

Sobre estos *embeddings* se evaluaron nuevamente SVM, RF y XGBoost, manteniendo el mismo protocolo de validación y configuraciones razonables de hiperparámetros, sin realizar una búsqueda exhaustiva. De esta forma, los resultados permiten analizar principalmente el aporte de las representaciones preentrenadas frente al enfoque basado en características acústicas, bajo una comparación controlada.

4.3.2. Resultados globales

La Tabla 6 presenta los resultados obtenidos por los clasificadores entrenados sobre embeddings preentrenados. Se reportan las métricas promedio de validación cruzada junto con su desviación estándar entre particiones.

La Figura 8 presenta las matrices de confusión obtenidas para los modelos evaluados bajo el enfoque de embeddings preentrenados. Cada matriz representa el desempeño agregado a partir de las predicciones realizadas sobre los conjuntos de prueba de todas las particiones de la validación cruzada.

Tabla 6: Resultados globales del enfoque basado en características acústicas. RF obtiene los mejores resultados promedio en todas las métricas evaluadas, aunque las desviaciones estándar muestran una variabilidad relevante entre particiones.

Fuente: Elaboración propia.

Modelo	Accuracy	F1 macro	F1 ponderado	Recall macro
SVM	0.717 ± 0.121	0.663 ± 0.157	0.709 ± 0.131	0.669 ± 0.157
Random Forest	0.737 ± 0.128	0.693 ± 0.148	0.728 ± 0.130	0.702 ± 0.144
XGBoost	0.727 ± 0.155	0.692 ± 0.170	0.721 ± 0.156	0.696 ± 0.166

4.3.3. Análisis de resultados

Los resultados obtenidos con *embeddings* preentrenados permiten evaluar el aporte de utilizar representaciones aprendidas del audio en lugar de características acústicas diseñadas manualmente. Los modelos preentrenados de habla pueden capturar información acústica, fonética y temporal de mayor complejidad que aquella resumida mediante estadísticos tradicionales.

Se observa una mejora respecto al enfoque basado en características acústicas, esta diferencia puede interpretarse como evidencia de que los *embeddings* contienen información más discriminativa para la tarea. Esto sugiere que la calidad lectora no solo se manifiesta en propiedades globales del audio, como pausas o descriptores espectrales agregados, sino también en patrones más complejos asociados a la continuidad, articulación y estructura temporal de la lectura oral.

No obstante, el uso de *embeddings* congelados también presenta limitaciones. Al no ajustar el *encoder* al dominio específico de la tarea, la representación utilizada corresponde a aquella aprendida durante el preentrenamiento general del modelo. Por lo tanto, aunque los *embeddings* puedan contener información útil, no necesariamente se encuentran optimizados para distinguir categorías de calidad lectora definidas bajo un estándar educativo específico. Esta limitación es especialmente relevante considerando que las clases utilizadas en este trabajo responden al contexto chileno de evaluación del dominio lector.

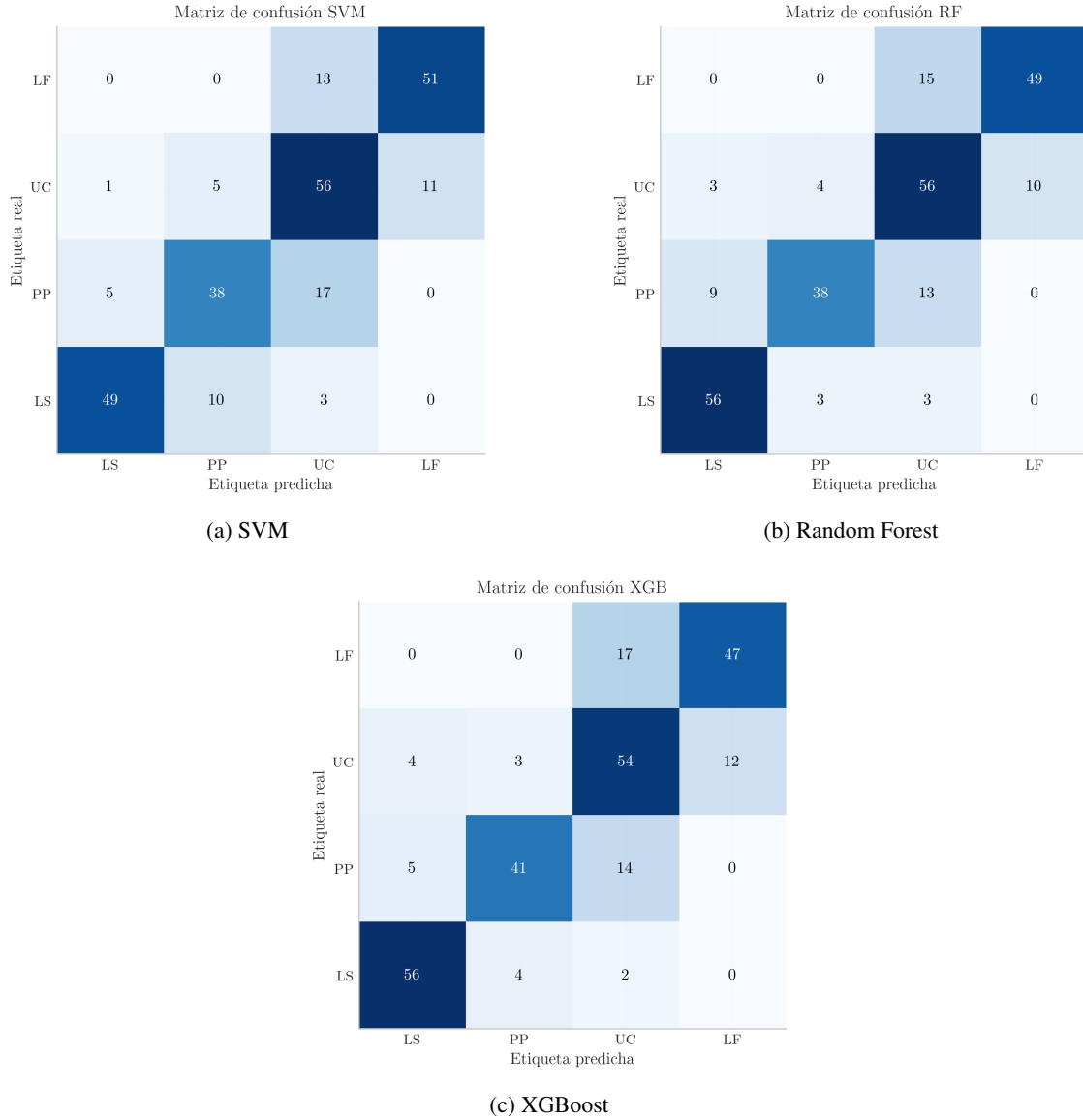


Figura 8: Matrices de confusión obtenidas por los modelos clásicos utilizando la representación de *embedding* de Wav2Vec2. Los resultados evidencian confusiones frecuentes entre categorías contiguas de calidad lectora, sin embargo disminuyen errores entre clases ubicadas en extremos opuestos de la escala.
Fuente: Elaboración propia.

Finalmente, la interpretación de este enfoque debe considerar su menor costo de entrenamiento en comparación con el último enfoque. Al mantener congelado el *encoder*, se reduce la cantidad de parámetros entrenables y se obtiene una alternativa más simple y estable para escenarios de baja disponibilidad de datos. Por ello, este enfoque puede resultar atractivo desde el punto de vista de reproducibilidad, eficiencia computacional y menor riesgo de sobreajuste.

4.4. Resultados del enfoque basado en transfer learning

El tercer enfoque evaluado corresponde a una estrategia de transferencia de aprendizaje mediante el ajuste parcial de un modelo preentrenado de habla. A diferencia del enfoque anterior, donde el *encoder* se utilizó únicamente como extractor de características congelado, en este caso se permite actualizar una parte de sus

parámetros durante el entrenamiento. El objetivo de esta estrategia es adaptar parcialmente la representación aprendida por el modelo al dominio específico de la clasificación de calidad lectora.

Este enfoque representa la alternativa de mayor capacidad dentro de la validación experimental, ya que combina el conocimiento previo adquirido por el modelo durante su preentrenamiento con un proceso de ajuste supervisado sobre los datos disponibles para esta tarea. Sin embargo, debido al tamaño acotado del conjunto de datos, el ajuste del modelo podría aumentar el riesgo de sobreajuste. Para reducir el riesgo, se optó por una estrategia de ajuste fino parcial, manteniendo congelada la mayor parte del *encoder* y entrenando únicamente sus capas superiores junto con una cabeza de clasificación.

La evaluación de este enfoque permite analizar si adaptar parcialmente el modelo preentrenado al problema específico mejora el desempeño respecto al uso de *embeddings* congelados. Asimismo, permite estudiar si el aumento en complejidad del entrenamiento se traduce en una mejora suficientemente consistente como para justificar su uso en un escenario de datos reducidos o es necesario un volumen mayor de datos para observar resultados.

4.4.1. Configuración experimental

Para este enfoque se utilizó un modelo preentrenado de habla como base del sistema de clasificación para ajustarlo a la tarea abordada en este trabajo. Sobre las representaciones internas generadas por el *encoder* se incorporó una cabeza de clasificación neuronal, entrenada de forma supervisada para predecir las categorías de calidad lectora definidas en este trabajo.

Con el objetivo de reducir el riesgo de sobreajuste, se mantuvo congelada aproximadamente el 90% de las capas del *encoder*, permitiendo el ajuste únicamente de sus capas superiores y de la cabeza de clasificación. Esta estrategia busca conservar el conocimiento general aprendido durante el preentrenamiento y adaptar solo las representaciones más cercanas a la salida del modelo al conjunto de datos utilizado.

El entrenamiento se realizó bajo el mismo protocolo de validación cruzada empleado en los enfoques anteriores, manteniendo las particiones para asegurar una comparación consistente. En el conjunto de entrenamiento se vuelve a realizar una partición por locutor de los datos para tener un conjunto de validación.

4.4.2. Resultados globales

La Tabla 7 presenta los resultados obtenidos por el enfoque basado en ajuste parcial del *encoder*. Se reportan las métricas promedio obtenidas durante la validación cruzada, junto con la desviación estándar entre las particiones.

La Figura 7 presenta la matriz de confusión obtenida bajo el enfoque de *transfer learning* a través del ajuste fino del *encoder* de *Wav2Vec2*. La matriz representa el desempeño agregado a partir de los resultados de las particiones.

Tabla 7: Resultados globales del enfoque basado en ajuste fino parcial del *encoder*.

Fuente: Elaboración propia.

Modelo	Accuracy	F1 macro	F1 ponderado	Recall macro
Encoder + cabeza clasificación	0.665 ± 0.164	0.609 ± 0.193	0.609 ± 0.214	0.671 ± 0.139

4.4.3. Análisis de resultados

El enfoque basado en *transfer learning* permite entregar mayor capacidad de adaptación al modelo, ya que no se limita a utilizar representaciones congeladas, sino que ajusta parte del *encoder* a la tarea específica de clasificación de calidad lectora. Esta característica puede ser beneficiosa cuando las representaciones generales del modelo preentrenado no capturan de manera suficiente las diferencias asociadas al dominio particular de la evaluación del dominio lector.

Desde el punto de vista del problema abordado, esta adaptación resulta relevante porque las categorías de calidad lectora no corresponden simplemente a diferencias acústicas generales, sino a niveles definidos según criterios educativos y perceptuales. Por ello, permitir que las capas superiores del *encoder* se ajusten

al conjunto de datos puede ayudar a capturar patrones más específicos relacionados con la continuidad de la lectura, las pausas, la articulación y la estructura temporal del habla.

No obstante, este enfoque también presenta mayores riesgos que las estrategias anteriores. Al aumentar la cantidad de parámetros entrenables, se incrementa la posibilidad de que el modelo se ajuste excesivamente a las particularidades del conjunto de entrenamiento. Este riesgo es especialmente importante en un escenario de pocos datos, donde la cantidad de ejemplos disponibles por clase es limitada y donde las particiones de validación pueden presentar diferencias relevantes en su composición.

Los resultados obtenidos mediante el enfoque muestran un comportamiento inestable entre las distintas particiones de validación cruzada. La alta variabilidad observada en las métricas sugiere que el modelo presenta sensibilidad a las particiones utilizadas y un posible sobreajuste a los datos de entrenamiento.

La matriz de confusión evidencia una leve mejora en la identificación de las clases asociadas a una mayor calidad lectora, en los niveles más fluidos, donde el modelo no presenta una tendencia marcada a cometer errores extremos en estas clases. Sin embargo, en los extractos con menor calidad lectora, el desempeño es considerablemente más débil: el modelo no logra delimitar adecuadamente las fronteras entre categorías cercanas y presenta confusiones frecuentes, incluyendo algunos errores hacia clases alejadas. Esto indica que el ajuste parcial del *encoder* no fue suficiente para capturar con robustez los patrones acústicos asociados a los niveles lectores más bajos.

En comparación con el enfoque basado en *embeddings* congelados, este no entrega una mejora suficientemente consistente como para justificar de manera clara su mayor costo computacional y complejidad de entrenamiento. Si bien el ajuste de las últimas capas del *encoder* permite adaptar parcialmente las representaciones al dominio específico de lectura infantil, los resultados obtenidos muestran que esta estrategia requiere una mayor cantidad de datos para estabilizar el aprendizaje y reducir el riesgo de sobreajuste.

En este sentido, el enfoque debería verse beneficiado por una ampliación sustancial del conjunto de datos, idealmente en el orden de cientos de horas de grabaciones, que permita ajustar con mayor robustez los parámetros del *encoder*. Con una base de datos más amplia y diversa, también sería pertinente explorar distintos grados de congelamiento del modelo preentrenado, evaluando si una adaptación más profunda permite capturar mejor las diferencias entre niveles de calidad lectora sin comprometer la generalización.

4.5. Comparación global de enfoques

A continuación, se comparan los mejores resultados obtenidos por cada enfoque experimental, considerando métricas globales, matrices de confusión y variabilidad entre particiones. El objetivo es identificar las diferencias principales entre las representaciones evaluadas y analizar el alcance de los resultados bajo el tamaño actual del conjunto de datos.

La Tabla 8 resume el mejor modelo de cada enfoque, seleccionado según el *F1-score* macro. Esta métrica se utiliza como referencia principal debido a que entrega el mismo peso a cada clase, lo que resulta relevante en un problema donde la distribución de ejemplos por nivel lector puede no ser uniforme.

Tabla 8: Comparación global de los mejores resultados por enfoque.

Enfoque	Mejor modelo	Accuracy	F1 macro	F1 pond.	Recall macro
Características acústicas	XGB	0.510 ± 0.100	0.384 ± 0.152	0.449 ± 0.115	0.424 ± 0.137
Embeddings preentrenados	RF	0.737 ± 0.128	0.693 ± 0.148	0.728 ± 0.130	0.702 ± 0.144
Fine-tuning parcial encoder	-	0.665 ± 0.164	0.609 ± 0.193	0.609 ± 0.214	0.671 ± 0.139

Los resultados muestran que el mejor desempeño promedio fue obtenido por el enfoque basado en *embeddings* preentrenados, alcanzando un *F1-score* macro de 0.693 ± 0.148 . Este resultado supera al enfoque basado en características acústicas, que obtuvo 0.384 ± 0.152 , y al ajuste parcial del *encoder*, que alcanzó 0.609 ± 0.193 . Esta diferencia sugiere que las representaciones aprendidas por modelos preentrenados capturan información más útil para la clasificación de la calidad lectora que las características acústicas diseñadas manualmente.

Las matrices de confusión permiten observar diferencias relevantes entre los enfoques evaluados. En particular, el enfoque basado en características acústicas sugiere una peor discriminación entre clases lectoras que el enfoque de *embeddings* preentrenados, reflejado en un menor desempeño global en las métricas agregadas. En todos los casos, independiente del enfoque, una parte importante de los errores se concentra en clases adyacentes al valor real, lo que resulta consistente con la naturaleza progresiva de los niveles de calidad lectora. Sin embargo, en el enfoque de ajuste parcial se observa la aparición de errores entre clases extremas que no está presente en los *embeddings* congelados, lo que sugiere un posible sobreajuste al conjunto de entrenamiento y refuerza la necesidad de interpretar sus resultados con cautela.

La variabilidad entre particiones es un elemento importante para interpretar estos resultados. Debido a que la validación cruzada fue agrupada por locutor, cada partición evalúa estudiantes no observados durante el entrenamiento, lo que entrega una estimación más exigente del desempeño. En este contexto, las diferencias entre promedios deben analizarse con cautela cuando están acompañadas de desviaciones estándar elevadas.

En síntesis, la comparación global indica que el enfoque basado en *embeddings* preentrenados corresponde a la alternativa más prometedora bajo el escenario experimental actual, al obtener el mejor desempeño promedio y una variabilidad menor que el ajuste parcial. Sin embargo, el tamaño reducido del conjunto de datos limita el alcance de las conclusiones, por lo que estos resultados deben interpretarse como evidencia inicial. Una validación más robusta requiere ampliar la cantidad de datos y evaluar el comportamiento de los modelos en nuevas cohortes de estudiantes.

4.6. Discusión de resultados

Los resultados obtenidos permiten discutir la factibilidad de abordar la clasificación automática de la calidad de lectura oral a partir del audio, considerando un escenario de datos reducidos y un estándar de evaluación asociado al contexto chileno. Más allá de identificar el modelo con mejor desempeño promedio, la validación experimental permite analizar el aporte de distintas representaciones del audio, la dificultad propia de la tarea y las limitaciones que deben considerarse al utilizar este tipo de solución en contextos educativos reales.

En primer lugar, la comparación entre enfoques evidencia que la representación del audio cumple un rol central en el desempeño de los modelos. El enfoque basado en características acústicas permite establecer una línea base relevante, ya que incorpora información asociada a propiedades espectrales, pausas y continuidad de la señal. Sin embargo, al utilizar estadísticos agregados, esta representación puede perder información temporal fina que resulta importante para caracterizar la lectura oral. Esto es especialmente relevante en una tarea donde la calidad lectora no depende únicamente de atributos globales del audio, sino también de la forma en que el estudiante organiza la lectura, articula las palabras, mantiene la continuidad y realiza pausas a lo largo del tiempo.

En contraste, los enfoques basados en modelos preentrenados permiten aprovechar representaciones aprendidas desde grandes volúmenes de habla, capaces de capturar patrones acústicos y temporales más complejos que aquellos definidos manualmente. Por ello, el mejor desempeño obtenido por los *embeddings* preentrenados sugiere que estos modelos contienen información útil para distinguir niveles de calidad lectora, incluso manteniendo el *encoder* congelado. Esta observación resulta relevante para escenarios con pocos datos etiquetados, donde el uso de modelos de habla como extractores de representación puede ser una alternativa efectiva frente al entrenamiento completo de modelos profundos.

No obstante, el uso de *embeddings* congelados también presenta limitaciones, ya que las representaciones no necesariamente están optimizadas para los criterios específicos de la tarea. Las categorías de calidad lectora responden a una evaluación educativa particular, asociada a la fluidez, continuidad y forma de lectura, por lo que no coinciden necesariamente con los objetivos originales de entrenamiento de los modelos preentrenados. En este sentido, el ajuste parcial del *encoder* busca adaptar parte del modelo al dominio específico; sin embargo, en los resultados obtenidos esta estrategia debe interpretarse con cautela, ya que la aparición de errores entre clases extremas sugiere un posible sobreajuste al conjunto de entrenamiento.

Un aspecto importante de la tarea es que las clases evaluadas no corresponden a categorías completamente independientes, sino a niveles progresivos de calidad lectora. Esto implica que las fronteras entre algunas clases pueden ser difusas tanto para los modelos como para los propios evaluadores humanos. En

particular, las diferencias entre niveles intermedios pueden ser más sutiles que aquellas existentes entre categorías extremas. Por esta razón, resulta esperable que una parte de los errores se concentre entre clases adyacentes, como *Lectura Palabra a Palabra* y *Lectura en Unidades Cortas*, más que entre niveles claramente separados como *Lectura Silábica* y *Lectura Fluida*.

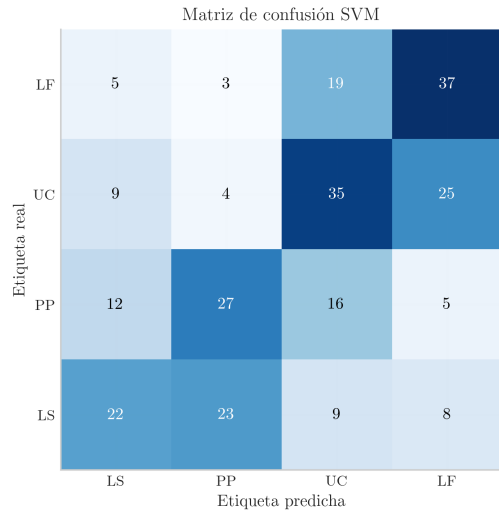
Esta característica del problema tiene implicancias relevantes para la interpretación de los resultados. En una tarea de naturaleza progresiva, no todos los errores poseen la misma severidad desde el punto de vista del dominio. Una confusión entre clases vecinas puede indicar que el modelo identifica parcialmente el nivel lector, aunque no logre establecer con precisión la categoría exacta. En cambio, una confusión entre categorías extremas indicaría una falla más severa en la representación o en la capacidad discriminativa del modelo. Por ello, el análisis de matrices de confusión resulta fundamental para complementar las métricas globales y comprender el comportamiento real de cada enfoque.

Desde esta perspectiva, el mejor enfoque no necesariamente corresponde al que obtiene el mayor valor promedio en una métrica particular. En un escenario de datos acotados, una diferencia pequeña entre modelos puede no ser suficiente para afirmar superioridad si está acompañada de una alta variabilidad. Por el contrario, un enfoque con desempeño levemente menor, pero más estable entre particiones, puede ser más adecuado para futuras iteraciones del sistema. Esta consideración es especialmente importante en aplicaciones educativas, donde la estabilidad, reproducibilidad y confiabilidad del modelo son aspectos tan relevantes como el desempeño promedio.

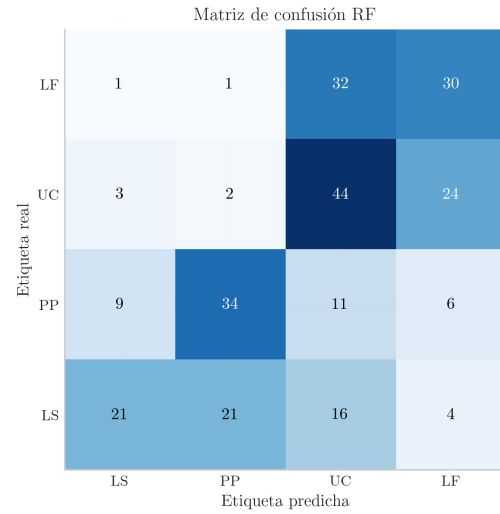
Los resultados también deben interpretarse considerando las condiciones reales de captura de los audios. A diferencia de conjuntos de datos recolectados en ambientes controlados, los registros utilizados en este trabajo provienen de un contexto educativo real, con variaciones asociadas a dispositivos de grabación, ruido ambiental, características individuales de los estudiantes y condiciones propias de la evaluación que afectan la calidad del etiquetado. Estas variaciones introducen complejidad adicional al problema, pero al mismo tiempo hacen que la validación sea más representativa del escenario en el cual una solución de este tipo podría ser utilizada.

En términos de la propuesta de solución, los resultados entregan evidencia inicial sobre la posibilidad de automatizar la evaluación de la calidad lectora a partir del audio. Sin embargo, esta automatización debe entenderse como un apoyo al proceso evaluativo, no como un reemplazo directo del juicio profesional. La evaluación de la lectura oral involucra criterios pedagógicos y contextuales que pueden requerir interpretación humana, especialmente en casos ambiguos o cercanos a los límites entre categorías. Por ello, un sistema automático puede ser especialmente útil como herramienta de apoyo, priorización o preclasificación, manteniendo la supervisión de profesionales del área educativa.

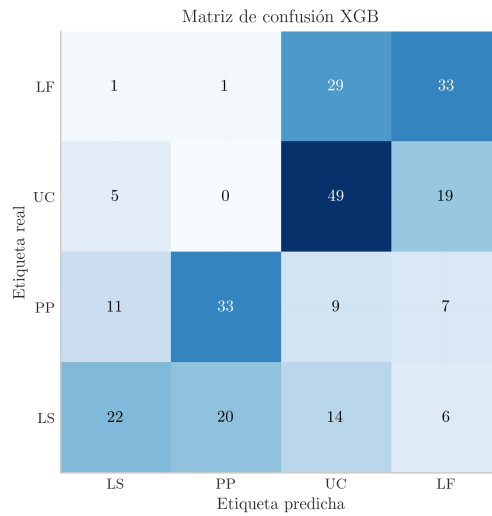
En síntesis, la discusión de los resultados muestra que la clasificación automática de calidad lectora a partir del audio es una tarea factible, pero desafiante. Las principales dificultades se relacionan con el tamaño reducido del conjunto de datos, la naturaleza progresiva de las clases, la variabilidad entre locutores y la necesidad de capturar patrones temporales complejos de la lectura oral. A pesar de estas limitaciones, los resultados obtenidos permiten establecer una base experimental para la propuesta y entregan indicios sobre qué enfoques resultan más prometedores para futuras etapas de desarrollo y validación.



(a) SVM



(b) Random Forest



(c) XGBoost

Figura 7: Matrices de confusión obtenidas por los modelos clásicos utilizando la representación acústica basada en MFCC. Los resultados evidencian confusiones frecuentes entre categorías contiguas de calidad lectora, además de algunos errores entre clases ubicadas en extremos opuestos de la escala.

Fuente: Elaboración propia.

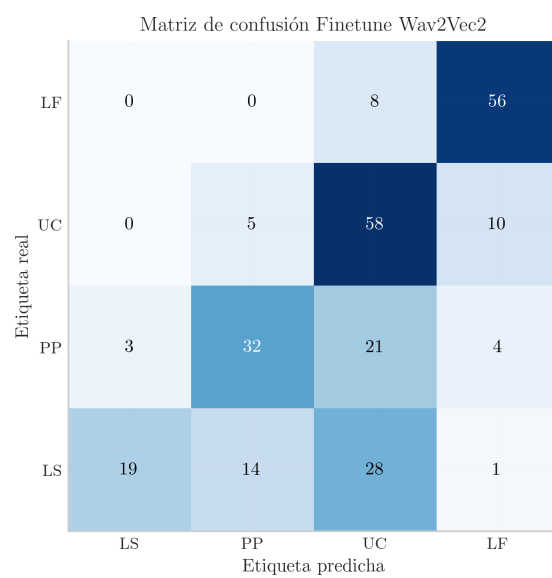


Figura 9: Matriz de confusión obtenidas para el ajuste fino de Wav2Vec2 con un clasificador neuronal. En calidades lectoras más fluidas se observa una leve mejora. Sin embargo, en calidades inferiores la confusión aumenta, tanto en clases adyacentes como extremas.
Fuente: Elaboración propia.

CAPÍTULO 5

CONCLUSIONES

El presente trabajo abordó el problema de la clasificación automática de la calidad de lectura oral en estudiantes de enseñanza básica en Chile, a partir del análisis directo de registros de audio. Para ello, se propuso y evaluó un marco experimental compuesto por tres enfoques: modelos clásicos entrenados sobre características acústicas manuales, clasificadores entrenados sobre *embeddings* de modelos preentrenados de habla, y una estrategia de *transfer learning* mediante ajuste parcial de un *encoder* preentrenado.

A partir de los resultados obtenidos, se puede concluir que la clasificación automática de la calidad lectora desde audio es una tarea factible, pero desafiante. Los experimentos realizados permiten observar que existen patrones en la señal de habla que pueden ser utilizados por modelos de aprendizaje automático para distinguir entre distintos niveles lectores. Sin embargo, la naturaleza progresiva de las clases, la variabilidad entre locutores y el tamaño acotado del conjunto de datos limitan el alcance de las conclusiones y exigen interpretar los resultados con cautela.

Respecto al cumplimiento de los objetivos, el trabajo permitió construir una base experimental para comparar distintas formas de representar el audio en la tarea de clasificación de calidad lectora. En particular, se estableció una línea base mediante características acústicas manuales y modelos clásicos, se evaluó el aporte de representaciones preentrenadas mediante *embeddings*, y se exploró una estrategia de ajuste parcial para adaptar un modelo de habla al dominio específico del problema. Esta comparación permitió analizar no solo el desempeño de los modelos, sino también el efecto que tiene la representación del audio sobre la capacidad de clasificación.

Uno de los principales aprendizajes del trabajo es que la cantidad y calidad de los datos constituyen factores críticos para el desarrollo de soluciones automáticas en este dominio. La lectura oral infantil presenta una alta variabilidad asociada a diferencias individuales entre estudiantes, condiciones de grabación, textos leídos, dispositivos utilizados y criterios de evaluación. Además, durante el proceso experimental se observó que, en la práctica, un mismo estudiante no necesariamente mantiene una única calidad lectora durante toda la lectura. En distintas secciones del texto pueden aparecer pasajes leídos con mayor fluidez, mientras que en otros momentos la lectura puede volverse más entrecortada o menos continua. Esta variabilidad interna dentro de un mismo audio puede explicar parte de las confusiones observadas entre clases contiguas, ya que la etiqueta global asignada al registro no siempre representa de manera uniforme todo el contenido leído.

Esta observación abre una línea de investigación relevante para trabajos futuros. En lugar de abordar la tarea exclusivamente como una clasificación multiclase con una única etiqueta por audio, podría explorarse una formulación de clasificación multietiqueta o de estimación proporcional de calidad lectora. Bajo este enfoque, un registro no sería representado únicamente por una clase dominante, sino por la presencia y proporción de distintos niveles de lectura a lo largo del audio. Esto permitiría cuantificar, por ejemplo, qué porcentaje de la lectura corresponde a segmentos más fluidos y qué porcentaje presenta patrones más silábicos, entrecortados o de menor continuidad. Una aproximación de este tipo podría entregar una evaluación más fina y representativa del desempeño lector real de los estudiantes.

Por esta razón, futuros trabajos deberían considerar la construcción de conjuntos de datos de mayor escala, con diversidad de locutores, textos, cursos, establecimientos e instrumentos de grabación. Esto permitiría obtener resultados más robustos y evaluar con mayor precisión la capacidad de generalización de los modelos. Asimismo, contar con más datos permitiría analizar con mayor detalle la variabilidad intraestudiante e intralectura, favoreciendo el desarrollo de modelos capaces de capturar cambios temporales en la calidad lectora dentro de un mismo registro.

En línea con lo anterior, el etiquetado de los datos representa otro aspecto fundamental. Al tratarse de un enfoque supervisado, la calidad de las etiquetas condiciona directamente el aprendizaje de los modelos y la validez de los resultados obtenidos. En este trabajo, las etiquetas fueron asignadas por una experta en dominio lector, lo que entrega una base relevante para la validación inicial. No obstante, este proceso también se encuentra sujeto a las propias limitaciones del contexto de evaluación, tales como la subjetividad inherente a la clasificación y las condiciones bajo las cuales se realiza la revisión. Por ello, futuros

trabajos deberían considerar mecanismos de etiquetado con múltiples expertos, protocolos de evaluación más controlados y análisis de acuerdo interevaluador, con el objetivo de consolidar una fuente de datos más confiable y reproducible.

Sin una fuente de datos suficientemente amplia, diversa y confiablemente etiquetada, los resultados de cualquier enfoque automático estarán condicionados por la calidad de la información disponible. Esto es especialmente relevante en tareas educativas, donde las decisiones del modelo pueden influir en la interpretación del desempeño de estudiantes. Por lo tanto, el desarrollo de mejores modelos debe ir acompañado de una estrategia rigurosa de construcción, validación y documentación de los datos utilizados.

Otro aspecto relevante corresponde a la posibilidad de construir una fuente de datos abierta para investigación en tecnologías aplicadas a la evaluación de lectura oral. Una iniciativa de este tipo podría favorecer el desarrollo académico y tecnológico en el área, permitiendo comparar enfoques bajo condiciones comunes y acelerar la investigación en modelos de habla aplicados al contexto educativo chileno. Sin embargo, esta posibilidad debe ser analizada cuidadosamente desde el punto de vista legal, ético y de privacidad, dado que involucra registros de voz de menores de edad. En consecuencia, cualquier esfuerzo de apertura de datos debería considerar consentimiento informado, anonimización, restricciones de uso, resguardo de identidad y una finalidad claramente orientada al beneficio educativo y social.

En cuanto a las contribuciones del trabajo, se destaca la formulación de un marco experimental específico para evaluar la clasificación de calidad lectora desde audio en el contexto chileno. A diferencia de enfoques centrados en transcripción automática o cálculo posterior de métricas a partir de texto, este trabajo exploró la clasificación directa desde la señal de audio. Esto permite abrir una línea de investigación complementaria para estudiar si los patrones acústicos y temporales de la lectura oral pueden ser utilizados como indicadores automáticos del nivel lector.

Asimismo, el trabajo contribuye al análisis comparativo entre representaciones acústicas manuales y representaciones aprendidas mediante modelos preentrenados de habla. Esta comparación resulta relevante para orientar futuras investigaciones, ya que permite observar si los esfuerzos deberían concentrarse en mejorar la construcción de los descriptores acústicos utilizados, aprovechar *embeddings* preentrenados o avanzar hacia estrategias de aprendizaje profundo de extremo a extremo. En este sentido, los resultados obtenidos no deben entenderse únicamente como la selección de un modelo final, sino como una base para decidir qué líneas metodológicas resultan más prometedoras.

También resulta relevante contrastar este trabajo con el desarrollo inicial realizado en el marco del proyecto TeacHear. En dicha etapa se siguió tempranamente una línea metodológica basada en el uso de representaciones preentrenadas de habla, lo que coincide con una de las direcciones exploradas en esta memoria. Sin embargo, las limitaciones del diseño experimental original, particularmente en aspectos como la separación de datos, el control de particiones y la validación del modelo, pudieron producir resultados optimistas respecto al desempeño real del sistema. Por ello, aunque los resultados no son directamente comparables debido a las diferencias en el marco de evaluación, esta memoria permitió reformular el problema bajo un protocolo experimental más riguroso y entregar una evaluación más cautelosa de la capacidad de generalización de los modelos.

Respecto al trabajo futuro, una primera línea prioritaria corresponde a profundizar en enfoques de *transfer learning* y soluciones *end-to-end*. Con un conjunto de datos de mayor escala, idealmente a nivel de cientos de horas de grabaciones, sería posible evaluar arquitecturas más especializadas, modificar cabezas de clasificación, explorar mecanismos de *pooling* temporal más sofisticados y adaptar con mayor robustez modelos preentrenados al dominio de lectura infantil chilena. Esta línea resulta especialmente prometedora, ya que los modelos de habla modernos pueden beneficiarse significativamente de una mayor cantidad de datos supervisados.

Una segunda línea de trabajo futuro consiste en mejorar y ampliar las representaciones acústicas manuales. Si bien en este trabajo se utilizaron características acústicas como línea base, no se descarta que representaciones más detalladas, mejor diseñadas o combinadas con modelos neuronales puedan aportar información relevante para la tarea. Por ejemplo, podrían incorporarse descriptores prosódicos, medidas de ritmo, duración de pausas, velocidad de lectura, estabilidad temporal y otros indicadores asociados a fluidez. Un estudio más profundo de estas representaciones permitiría determinar con mayor claridad si los enfoques manuales son insuficientes para el problema o si requieren una formulación más especializada.

También sería relevante explorar enfoques híbridos que combinen características acústicas interpreta-

bles con representaciones aprendidas. Este tipo de aproximación podría entregar un equilibrio entre desempeño y explicabilidad, aspecto importante en contextos educativos donde no basta con obtener una predicción, sino que también es deseable comprender qué señales del audio influyen en la clasificación. La explicabilidad de los modelos representa, por tanto, otra línea relevante para futuras investigaciones.

Finalmente, este trabajo busca relevar la importancia de desarrollar tecnología aplicada a problemas educativos reales del contexto chileno. La evaluación de la lectura oral es un proceso relevante para identificar necesidades de apoyo en etapas tempranas del desarrollo escolar, pero también puede demandar tiempo y esfuerzo significativo por parte de docentes y especialistas. En este contexto, los sistemas automáticos basados en inteligencia artificial pueden constituir una herramienta de apoyo para facilitar procesos de evaluación, priorización y seguimiento, siempre desde una perspectiva complementaria y no sustitutiva del criterio profesional.

En conclusión, la propuesta desarrollada entrega evidencia inicial sobre la factibilidad de clasificar automáticamente la calidad de lectura oral a partir de audio, junto con identificar los principales desafíos técnicos, metodológicos y éticos asociados a esta tarea. Los resultados obtenidos permiten establecer una base experimental para futuras investigaciones y refuerzan la necesidad de avanzar en la construcción de conjuntos de datos más amplios, diversos y confiables. Además, el trabajo permitió reconocer que la calidad lectora puede variar dentro de un mismo registro, lo que abre nuevas posibilidades de modelación más finas que una única etiqueta global por audio. De esta manera, el trabajo contribuye tanto al desarrollo académico de modelos de habla aplicados a educación como a la proyección de soluciones tecnológicas que puedan apoyar la labor docente y aportar al mejoramiento de procesos educativos en Chile.

Referencias

- [1] T. Marchant, I. Recart, B. Cuadrado, and J. Sanhueza, *Pruebas de Dominio Lector: Evaluación del dominio lector en alumnos de 1° a 8° básico*. Chile: Fundación Educacional Arauco, 2024.
- [2] T. Marchant, G. Lucchini, and B. Cuadrado, “¿por qué leer bien es importante?: Asociación del dominio lector con otros aprendizajes,” *Psyke*, vol. 16, no. 2, pp. 3–16, 2007. [Online]. Available: <https://ojs.uc.cl/index.php/psykhe/article/view/19889>
- [3] L. S. Fuchs and D. Fuchs, “What is scientifically-based research on progress monitoring?” National Center on Student Progress Monitoring, Washington, DC, Tech. Rep., 2001, available: <https://eric.ed.gov/?id=ED502460>.
- [4] G. C. da Silva, R. L. Rodrigues, A. N. Amorim, L. Jeon, E. X. Albuquerque, V. C. Silva, V. F. da Silva, A. L. Pinheiro, J. P. Nunes, S. X. de Souza, M. S. Silva, I. Mauro, and A. M. A. Maciel, “Assessing reading fluency in elementary grades: A machine learning approach,” *Computers and Education: Artificial Intelligence*, vol. 8, p. 100411, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2666920X25000517>
- [5] M. van der Velde, W. Harmsen, B. P. Veldkamp *et al.*, “Speech enabled reading fluency assessment: a validation study,” *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 2569–2595, 2025.
- [6] TeacHear, “Sobre teachear,” <https://teachear.feriadesoftware.cl/sobre-teachear/>, 2025, [Online; accessed Oct. 3, 2025].
- [7] J. Fumagalli, J. P. Barreyro, and V. Jaichenco, “La evaluación de la fluidez lectora: Medidas subjetivas y objetivas,” *Cuadernos de Neuropsicología*, vol. 14, no. 1, pp. 78–87, 2020.
- [8] G. Ling, P. Mollaun, and X. Xi, “A study on the impact of fatigue on human raters when scoring speaking responses,” *Language Testing*, vol. 31, no. 4, pp. 479–499, 2014.
- [9] S. E. Stemler, “A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability,” *Practical Assessment, Research, and Evaluation*, vol. 9, no. 4, 2004. [Online]. Available: <https://doi.org/10.7275/96jp-xz07>
- [10] O. Henkel, H. Horne-Robinson, L. Hills, and *et al.*, “Supporting literacy assessment in west africa: Using state-of-the-art speech models to assess oral reading fluency,” *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 282–303, Mar. 2025, accepted: 13 October 2024; Published: 28 January 2025. [Online]. Available: <https://doi.org/10.1007/s40593-024-00435-9>
- [11] B. Molenaar, C. Tejedor-Garcia, C. Cucchiari, and H. Strik, “Automatic assessment of oral reading accuracy for reading diagnostics,” in *INTERSPEECH 2023*, ser. interspeech2023. ISCA, Aug. 2023, p. 5232–5236. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2023-1681>
- [12] M. Yıldız, H. K. Keskin, S. Oyucu, D. K. Hartman, M. Temur, and M. Aydoğmuş, “Can artificial intelligence identify reading fluency and level? comparison of human and machine performance,” *Reading & Writing Quarterly*, vol. 41, no. 1, pp. 66–83, 2025. [Online]. Available: <https://doi.org/10.1080/10573569.2024.2345593>
- [13] M. Vaidya, B. K. Sahoo, and P. Rao, “Deep learning for assessment of oral reading fluency,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.19426>
- [14] Departamento de Ingeniería Industrial, Universidad de Chile. (2025, Apr.) Lecturia: Inteligencia artificial y modelos de lenguaje promueven desarrollo lector en aulas chilenas. Accedido: 25 abril 2026. [Online]. Available: <https://www.dii.uchile.cl/2025/04/25/lecturia-inteligencia-artificial-y-modelos-de-lenguaje-promueven-desarrollo-lector-en-aulas-chilenas/>

- [15] Leo Mejor. (n.d.) Leo mejor. Sitio web institucional. Accedido: 25 abril 2026. [Online]. Available: <https://leomejor.cl>
- [16] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1999.
- [17] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.
- [18] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [19] seriousran. (n.d.) Mfcc feature extraction for sound classification. Kaggle Notebook. Accedido: 17 mayo 2026. [Online]. Available: <https://www.kaggle.com/code/seriousran/mfcc-feature-extraction-for-sound-classification>
- [20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [21] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [22] B. Schuller *et al.*, "The interspeech 2010 paralinguistic challenge," in *Interspeech*, 2010.
- [23] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: <https://arxiv.org/abs/2006.11477>
- [24] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," 2021. [Online]. Available: <https://arxiv.org/abs/2111.09296>
- [25] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, Sep. 1995.
- [26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.
- [27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, p. 785–794. [Online]. Available: <http://dx.doi.org/10.1145/2939672.2939785>
- [28] OpenAI. (2023) Whisper: Robust speech recognition via large-scale weak supervision. [Online]. Available: <https://openai.com/research/whisper>
- [29] J. Wiseman, "py-webrtcvad: Python interface to the webrtc voice activity detector," <https://github.com/wiseman/py-webrtcvad>, 2017, accedido: 2026-05-05.
- [30] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in Spanish," <https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-spanish>, 2021.

ANEXOS

Resultados de clasificación por clase de enfoque descriptores acústicos clásicos

Tabla 9: Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo SVM.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
SVM	LF	0.493	0.578	0.532	64
	UC	0.443	0.479	0.461	73
	PP	0.474	0.450	0.462	60
	LS	0.458	0.355	0.400	62

Tabla 10: Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo Random Forest.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
RF	LF	0.469	0.469	0.469	64
	UC	0.427	0.603	0.500	73
	PP	0.586	0.567	0.576	60
	LS	0.618	0.339	0.438	62

Tabla 11: Métricas por clase obtenidas por el enfoque basado en descriptores acústicos clásicos modelo XGBoost.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
XGBoost	LF	0.508	0.516	0.512	64
	UC	0.485	0.671	0.563	73
	PP	0.611	0.550	0.579	60
	LS	0.564	0.355	0.436	62

Resultados de clasificación por clase de enfoque embeddings preentrenados

Tabla 12: Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo SVM.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
SVM	LF	0.823	0.797	0.810	64
	UC	0.629	0.767	0.691	73
	PP	0.717	0.633	0.673	60
	LS	0.891	0.790	0.838	62

Tabla 13: Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo Random Forest.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
RF	LF	0.831	0.766	0.797	64
	UC	0.644	0.767	0.700	73
	PP	0.844	0.633	0.724	60
	LS	0.824	0.903	0.862	62

Tabla 14: Métricas por clase obtenidas por el enfoque basado en embeddings preentrenados modelo XGBoost.

Fuente: Elaboración propia.

Modelo	Clase	Precision	Recall	F1-score	Support
XGBoost	LF	0.797	0.734	0.764	64
	UC	0.621	0.740	0.675	73
	PP	0.854	0.683	0.759	60
	LS	0.862	0.903	0.882	62

Resultados de clasificación por clase de enfoque transfer learning

Tabla 15: Métricas por clase obtenidas por el enfoque de ajuste fino de Wav2Vec 2.0.

Fuente: Elaboración propia.

Clase	Precision	Recall	F1-score	Support
LF	0.789	0.875	0.830	64
UC	0.504	0.795	0.617	73
PP	0.628	0.533	0.577	60
LS	0.864	0.307	0.452	62