INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO
SECCIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

Modelo de detección, clasificación y notificación de errores en la pronunciación de fonemas vocálicos del francés emitidos por alumnos mexicanos, empleando Cómputo Móvil y Machine Learning

TESIS

QUE PARA OBTENER EL GRADO DE:

MAESTRO EN CIENCIAS EN SISTEMAS COMPUTACIONALES MÓVILES

PRESENTA:

TIMOTEO ENRIQUE LATISNERE MEJIA

Directores de tesis:
Dra. Elena Fabiola Ruiz Ledesma
Dra. Laura Méndez Segundo

Ciudad de México Junio 2025

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi familia, en especial a mi madre, por su apoyo incondicional y, sobre todo, por su paciencia durante los momentos más difíciles de este camino. Agradezco también a mi colega Irving Guerra, por compartir conmigo su experiencia en desarrollo móvil multiplataforma y por su valiosa asesoría, lo cual ayudó enormemente para lograr la ágil creación de la aplicación 'PhonessaAI'.

Mi gratitud se extiende al jefe del Departamento de Lenguas Indoeuropeas y Orientales del Centro de Lenguas Extranjeras del IPN, Unidad Santo Tomás, el Mtro. Miguel Ángel Guzmán Medina, así como al director del plantel, el Mtro. Agustín Domínguez Flores, por haberme abierto las puertas de sus aulas y permitir la colaboración para la recolección de muestras de audio, fundamentales para la construcción del *dataset* utilizado en este trabajo.

De igual forma, agradezco a la maestra de francés Floriane Koechler, por dedicarme parte de su tiempo para orientarme sobre los métodos existentes de corrección de pronunciación en francés, lo cual fue clave para enfocar adecuadamente el proyecto.

Finalmente, quiero reconocer y agradecer a mis directoras de tesis, la Dra. Elena Fabiola Ruiz Ledesma y la Dra. Laura Méndez Segundo, así como a todas las personas del departamento de posgrado de la ESCOM, por el respaldo y apoyo brindados a lo largo de todo este proceso de investigación.

Dedicatoria

A la vida, por permitirme recorrer esta senda y haber puesto en el camino a numerosas personas muy especiales, quienes me ayudaron a aprender valiosas lecciones para seguir adelante.

Gracias.

SIP-13
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARIA DE INVESTIGACIÓN Y POSGRADO

Dirección de Posgrado

ACTA DE REGISTRO DE TEMA DE TESIS
Y DESIGNACIÓN DE DIRECTOR DE TESIS

Ciudad de México, a 14 de marzo del 2025

El Colegio de Profesores de Posgrado de Escuela Superior de Cómputo en su Sesión
(Unidad Académica)

Ordinaria No. 3 celebrada el día 14 del mes marzo de 2025 conoció la solicitud presentada por el (la) alumno (a):

Apellido Paterno:	Latisnere	Apellido Materno:	Mejia	Nombre (s):	Timoteo Enrique

Número de boleta: B 2 3 0 6 2 4

del Programa Académico de Posgrado: Maestría en Ciencias en Sistemas Computacionales Móviles

Referente al registro de su tema de tesis

1.- Se acordó aprobar el tema de tesis:

Modelo de detección, clasificación y notificación de errores en la pronunciación de fonemas vocálicos del francés emitidos por alumnos mexicanos, empleando Cómputo Móvil y Machine Learning

Objetivo general del trabajo de tesis:

Proponer un modelo de detección y clasificación de errores en la pronunciación de los fonemas vocálicos del idioma francés, empleando machine learning para identificar errores comunes de la pronunciación en alumnos hispanohablantes mexicanos de francés como lengua extranjera, presentándole retroalimentación sobre estos.

2.- Se designa como Directores de Tesis a los profesores:

Director: Dra. Elena Fabiola Ruiz Ledesma Director: Dra. Laura Méndez Segundo

No aplica: ☐

3.- El Trabajo de investigación base para el desarrollo de la tesis será elaborado por el alumno en:

SEPI-ESCOM

que cuenta con los recursos e infraestructura necesarios.

4.- El interesado deberá asistir a los seminarios desarrollados en el área de adscripción del trabajo desde la fecha en que se suscribe la presente, hasta la aprobación de la versión completa de la tesis por parte de la Comisión Revisora correspondiente.

Director(a) de Tesis
Dra. Elena Fabiola Ruiz Ledesma

Director de Tesis (en su caso)
Dra. Laura Méndez Segundo

Alumno
Timoteo Enrique Latisnere Mejia

Presidente del Colegio
M. en C. Iván Giovanny Mosso

S.E.P.
INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE COMPUTO

SIP-14
REP 2017

INSTITUTO POLITÉCNICO NACIONAL

SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

Dirección de Posgrado

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México siendo las 13:00 horas del día 27 del mes de mayo del 2025 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Posgrado de: Escuela Superior de Cómputo para examinar la tesis titulada:

Modelo de detección, clasificación y notificación de errores en la pronunciación de fonemas vocálicos del francés emitidos por alumnos mexicanos, empleando Cómputo Móvil y Machine Learning

del (la) alumno (a):

Apellido Paterno:	Latisnere	Apellido Materno:	Mejia	Nombre (s):	Timoteo Enrique

Número de boleta: B 2 3 0 6 2 4

Alumno del Programa Académico de Posgrado: Maestría en Ciencias en Sistemas Computacionales Móviles

Una vez que se realizó un análisis de similitud de texto, utilizando el software antiplagio, se encontró que el trabajo de tesis tiene 1 % de similitud. Se **adjunta reporte de software utilizado**.

Después que esta Comisión revisó exhaustivamente el contenido, estructura, intención y ubicación de los textos de la tesis identificados como coincidentes con otros documentos, concluyó que en el presente trabajo **SI** [ ] **NO** [X] **SE CONSTITUYE UN POSIBLE PLAGIO.**

JUSTIFICACIÓN DE LA CONCLUSIÓN: *(Por ejemplo, el % de similitud se localiza en metodologías adecuadamente referidas a fuente original)*

Principalmente a nombres de institución, escuela y términos de literatura.

Finalmente y posterior a la lectura, revisión individual, así como el análisis e intercambio de opiniones, los miembros de la Comisión manifestaron **APROBAR** [X] **SUSPENDER** [ ] **NO APROBAR** [ ] la tesis por **UNANIMIDAD** [X] o **MAYORÍA** [ ] en virtud de los motivos siguientes:

Las pruebas y resultados presentados permiten visualizar el buen término de la tesis

COMISIÓN REVISORA DE TESIS

Dra. Elena Fabiola Ruiz Ledesma
Director de Tesis
Nombre completo y firma

Dra. Lorena Chavarría Báez
Nombre completo y firma

M. en C. Jesús Alfredo Martínez Nuño
Nombre completo y firma

Dra. Laura Méndez Segundo
2° Director de Tesis (en su caso)
Nombre completo y firma

M. en C. Erika Hernández Rubio
Nombre completo y firma

M. en C. Iván Giovanny Mosso García
Nombre completo y firma
PRESIDENTE DEL COLEGIO DE PROFESORES

S.E.P.
INSTITUTO POLITECNICO NACIONAL
ESCUELA SUPERIOR DE COMPUTO

Carta de autorización de uso de obra para difusión

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

CARTA DE AUTORIZACIÓN DE USO DE OBRA PARA DIFUSIÓN

En la Ciudad de México el día 04 del mes de junio del año 2025, el (la) que suscribe TIMOTEO ENRIQUE LATISNERE MEJIA alumno(a) del programa Maestría en Ciencias en Sistemas Computacionales Móviles con número de registro B230624, adscrito(a) a Escuela Superior de Cómputo manifiesta que es autor(a) intelectual del presente trabajo de tesis bajo la dirección de Dra. Elena Fabiola Ruiz Ledesma y Dra. Laura Méndez Segundo y cede los derechos del trabajo intitulado: "Modelo de detección, clasificación y notificación de errores en la pronunciación de fonemas vocálicos del francés emitidos por alumnos mexicanos, empleando Cómputo Móvil y Machine Learning", al Instituto Politécnico Nacional, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expresado del autor y/o director(es). Este puede ser obtenido escribiendo a las siguiente(s) dirección(es) de correo. tlatisnerem1000@alumno.ipn.mx, lmendezs@ipn.mx, eruizl@ipn.mx. Si el permiso se otorga, al usuario deberá dar agradecimiento correspondiente y citar la fuente de este.

TIMOTEO ENRIQUE LATISNERE MEJIA

Nombre completo y firma autografá del (de la) estudiante

Índice general

Índice de figuras

Índice de tablas

Resumen

Este trabajo presenta un modelo para la detección, clasificación y notificación de errores en la pronunciación de fonemas vocálicos del francés emitidos por alumnos mexicanos, combinando técnicas de cómputo móvil y aprendizaje automático. El modelo integra una metodología basada en parámetros acústicos (formantes F1 y F2) y un enfoque de clasificación secuencial mediante Redes Neuronales Recurrentes (RNN) de Unidad Recurrente Cerrada (GRU), focalizado en el eje tonal de vocales complejas para hispanohablantes ([y], [œ], [o]). La recolección de datos se realizó mediante una aplicación móvil, y el análisis se sustentó en alineación forzada, normalización estadística y etiquetado fonético automático con base en distancias Z-score. Con un total de 35 participantes, el modelo alcanzó una precisión de clasificación superior al 63% en condiciones de datos limitados. La propuesta demuestra ser una contribución viable para el diagnóstico automatizado y la retroalimentación personalizada en el aprendizaje de lenguas extranjeras desde una perspectiva fonético-computacional.

Palabras clave: inteligencia artificial, fonética, francés, pronunciación, machine learning

Abstract

This work presents a model for the detection, classification, and notification of pronunciation errors in French vowel phonemes produced by Mexican students, combining mobile computing and machine learning techniques. The model integrates a methodology based on acoustic parameters (formants F1 and F2) and a sequential classification approach using Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU), focusing on the tonal axis of vowels that are complex for Spanish speakers ([y], [œ], [o]). Data collection was carried out through a mobile application, and the analysis relied on forced alignment, statistical normalization, and automatic phonetic labeling based on Z-score distances. With a total of 35 participants, the model achieved classification accuracy exceeding 63% under limited-data conditions. The proposal proves to be a viable contribution to automated diagnosis and personalized feedback in foreign language learning from a phonetic-computational perspective.

Keywords: artificial intelligence, phonetics, French, pronunciation, machine learning

Résumé

Ce travail présente un modèle de détection, de classification et de notification des erreurs de prononciation des phonèmes vocaliques du français produits par des étudiants mexicains, en combinant des techniques de calcul mobile et d'apprentissage automatique. Le modèle intègre une méthodologie fondée sur des paramètres acoustiques (formants F1 et F2) et une approche de classification séquentielle à l'aide de Réseaux Neuronaux Récurrents (RNN) avec Unités Récurrentes Fermées (GRU), en se concentrant sur l'axe tonal des voyelles complexes pour les hispanophones ([y], [œ], [o]). La collecte de données a été réalisée au moyen d'une application mobile, et l'analyse s'est appuyée sur l'alignement forcé, la normalisation statistique et l'étiquetage phonétique automatique basé sur des distances Z-score. Avec un total de 35 participants, le modèle a atteint une précision de classification supérieure à 63 % dans des conditions de données limitées. Cette proposition constitue une contribution pertinente au diagnostic automatisé et à la rétroaction personnalisée dans l'apprentissage des langues étrangères, selon une approche phonétique-informatique.

Mots clés: intelligence artificielle, phonétique, français, prononciation, machine learning

CAPÍTULO 1: Introducción

Este capítulo revisa la importancia del idioma francés en el contexto de la educación en México, y las dificultades fonéticas a las que se enfrentan los estudiantes mexicanos que estudian este idioma. Se presenta la problemática a tratar que conduce a la pregunta de investigación planteada, y se enumeran los objetivos y justificación de esta investigación. Además, se presenta el estado del arte. Finalmente, se introducen conceptos fundamentales para el desarrollo de esta tesis, que se detallan en el marco teórico conceptual. Cabe mencionar que todas las referencias que se hacen al idioma francés en este documento se refieren al francés estándar, que es el que se enseña en los programas de las instituciones oficiales de enseñanza de este idioma en el país.

1.1 Antecedentes

El idioma francés es hablado por 321 millones de personas en todos los continentes, posicionándose como la 5ª lengua más hablada del mundo y la 3ª más utilizada en los negocios internacionales [1].

En el año 2022, el Observatorio demográfico y estadístico del espacio francófono de la Universidad Laval de Québec [2], estimó – con bajo grado de calidad, debido al hecho de que los censos de población y vivienda en el país no recolectan datos sobre los idiomas extranjeros hablados [3] – que en México vivían aproximadamente 30 183 mexicanos francófonos, lo que representa el 0.02% de la población total del país. A pesar de este bajo porcentaje, según datos de la Embajada Francesa en México [4], se estima en 250 000 el número de estudiantes de francés como lengua extranjera (FLE) en el país, los cuales, al lado del sistema de enseñanza superior en México, se encuentran distribuidos en su red de centros de enseñanza del idioma, compuesta del IFAL y las Alianzas Francesas, lo que coloca al francés como la segunda lengua extranjera viva más enseñada en el país.

Si bien cada alumno es distinto y aprende a su propio ritmo, existen varios retos implícitos en la enseñanza de idiomas y, en el caso particular del francés, la pronunciación es uno de ellos. Esto por ciertas características del idioma que le dan a la vez su particular toque distintivo, así como su inherente complejidad y profundidad, que se discutirán brevemente.

Por un lado, están las dificultades provenientes de la componente escrita del idioma, como que tenga muchas reglas gramaticales – e igual de muchas excepciones a éstas – entre las que destacan: el género de los sustantivos, la concordancia de adjetivos con respecto a género y número, el uso de acentos ortográficos agudos, graves y circunflejos, la conjugación con tiempos y modos verbales, las reglas de elisión y liaison, ente otras. Además, la lengua de Molière cuenta con una gran cantidad de homónimos, es decir, palabras que se escriben igual o suenan igual, pero cuyo significado es diferente, lo cual hace que el contexto de lo previamente introducido en una frase, resulte vital para la correcta comprensión del resto del mensaje a transmitir [5].

Por otra parte, existe también el factor sonoro, ya que, como Huerta Espinosa argumenta en [6]: “el idioma francés es considerado por muchos estudiantes hispanohablantes de francés como lengua extranjera (FLE), como un idioma muy difícil de pronunciar.” Esto debido en parte a que, comparado con los 24 fonemas del español: 5 vocálicos y 19 consonánticos [7], el francés cuenta con 37 fonemas en total, de los cuales, 16 son vocálicos y 3 son semivocálicos o nasales; en otras palabras, el idioma francés tiene más sonidos que el español.

Además, existe una dificultad marcada en la asociación grafema-fonema, o sea, la unidad mínima distintiva en el plano gráfico (grafema), y la mínima distintiva en el plano fónico (fonema) [8]. Como Blin expone en [9]: "… por las características lingüísticas comunes que tienen ambos idiomas, un aprendiente de francés en México puede entender muy rápidamente el sentido de un texto de nivel A1, pero no puede leerlo en voz alta de manera adecuada, lo que representa un obstáculo en su proceso de aprendizaje."

Como respuesta ante los retos intrínsecos de la enseñanza de idiomas, en las últimas décadas se han desarrollado herramientas para la enseñanza de idiomas asistida por computadora, dando lugar a la creación de múltiples plataformas y aplicaciones móviles, lo que provee al alumno de diversos recursos para aprender la gramática, practicar la pronunciación, y valerse de otras funcionalidades, que han sido pensadas para complementar el aprendizaje tradicional. Por ejemplo, hay aplicaciones destinadas a la enseñanza de los fonemas en particular, las cuales muestran al alumno esquemas del aparato articulatorio a manera de imágenes o inclusive modelos tridimensionales de la boca, paladar, lengua y laringe, con el fin de que este tenga una referencia sobre la correcta posición de estos componentes y pueda asociarlos con los sonidos que les corresponden. [10]

Aunque ya existen estas herramientas, suelen dirigirse a un público muy diverso, por lo que es poco común que se consideren las deficiencias particulares del alumno —a menos que incluyan la asesoría de un maestro calificado—que se deben corregir o complementar. Como se menciona en el estudio realizado por la universidad de Toulouse en 2017 [11], sobre el análisis de las herramientas disponibles para el apoyo de los estudiantes de FLE en Japón, "… históricamente, estas herramientas se han creado con base en las limitaciones técnicas de su época y no con el ideal didáctico en mente, ni considerando otras variables como el perfil del alumno y sus antecedentes de aprendizaje en el idioma."

Es en este contexto, se encuentra un área de oportunidad para la integración de tecnologías, que permitan a los profesores de francés, automatizar la detección y clasificación de los errores de pronunciación del alumno, así como la generación de retroalimentación sobre estos, pensada desde un enfoque más didáctico, que le puedan ser útiles al aprendiz para aislar los errores en su pronunciación y poder atenderlos efectivamente. En este aspecto, la inteligencia artificial surge como una opción, debido a las técnicas que existen para la clasificación de datos. Por lo tanto, en este trabajo se realizará una clasificación de errores de pronunciación utilizando estas técnicas.

1.2 Contexto del problema

En el proceso de aprendizaje del idioma francés, hay dificultades que conducen a producir errores, por ejemplo, la identificación de los sonidos, su reproducción, y la prosodia, es decir, el ritmo de la frase y su entonación. En el caso de la pronunciación, los sonidos vocálicos plantean un desafío común para los estudiantes de una segunda lengua (denominada L2), especialmente cuando esos sonidos no existen en su idioma natal (L1).

En el caso de los hablantes hispanos que aprenden francés, a menudo se tienen dificultades con la pronunciación de las vocales [o], [y] y [oe], siendo las dos últimas inexistentes en el español. Los errores típicos en hispanohablantes relacionados a estos fonemas son la pronunciación de la "o" abierta [ɔ] en lugar de "o" cerrada [o], pronunciar [y] como [i] o [u], y pronunciar [oe] como [o] o [e] [12].

En general, existen dos métodos – o enfoques, según el autor - de corrección fonética para solucionar errores de pronunciación como los mencionados: el Método Articulatorio y el Método Verbo-Tonal (MVT).

El Método Articulatorio se basa en la idea de que se debe tener un conocimiento explícito de la articulación de un sonido para poder pronunciarlo correctamente, y ha prevalecido desde incluso antes de la invención del MVT. En este enfoque, se enseña al alumno a reconocer las partes que componenen el órgano fonatorio; principalmente integrado por el conducto nasal, el bucal y la laringe, como se observa en la Figura 1.1.

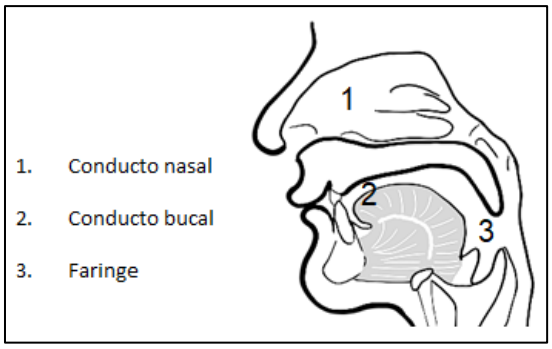


Figura 1.1 Corte transversal del aparato fonatorio.

Fuente: Elaboración propia.

Billières [13] destaca que este método también busca otorgar "bases articulatorias" al alumno que sirvan para generar comparaciones entre su lengua L1 y L2. Lo anterior por medio de la "gimnasia articulatoria", valiéndose del canal visual-kinestésico, y de ejercicios de discriminación con aumento de dificultad progresiva que van desde sonidos aislados, pasando por sílabas y palabras aisladas, para finalmente llegar a sonidos integrados en frases y en frases completas. El canal auditivo no tiene el énfasis en este método, ya que parte de la premisa de que una mejor producción oral conduce a una mejor percepción de los sonidos.

Por otro lado, el Método Verbo-Tonal es más reciente y da prioridad al ritmo y entonación en la pronunciación para la corrección de errores, así como a mejorar incrementalmente la pronunciación del alumno a lo largo del tiempo [14]. Para aplicar el proceso correctivo correspondiente en este método, el diagnóstico preciso de la mala pronunciación es crucial y suele partir de una clasificación

del error basada en los formantes F1 y F2; frecuencias fundamentales de la voz que se relacionan estrechamente con la tensión y claridad tonal [15]. Este método es el que será utilizado como referencia para el desarrollo de este trabajo, por lo cual se explica un poco más a detalle en el Marco Teórico.

En el contexto de la enseñanza en México, la técnica más común para solucionar errores de pronunciación es una mezcla del método articulatorio con la corrección uno a uno. Primero, el maestro escucha cuidadosamente la pronunciación del alumno para detectar el error, y después brinda retroalimentación correctiva. Sin embargo, con este enfoque surgen desafíos importantes que limitan la efectividad del aprendizaje. En primer lugar, en una enseñanza grupal, el tiempo disponible para atender las necesidades individuales de cada estudiante es limitado, lo que dificulta el progreso personalizado. En segundo lugar, la enseñanza de la pronunciación depende en gran medida de un oído experto, lo cual representa un problema cuando los docentes no cuentan con una formación especializada en fonética.

A pesar de los avances recientes en sistemas automáticos de detección de errores de pronunciación, subsisten limitaciones importantes que dificultan su aplicación efectiva en contextos educativos reales. Modelos *end-to-end* como los basados en RNN-T, CNN o SincNet han demostrado mejoras en tareas de diagnóstico fonético [16], [17], pero su desempeño depende en gran medida de grandes volúmenes de datos etiquetados manualmente y presentan escasa adaptabilidad a fonemas no presentes en los conjuntos fonológicos estándar.

Por otra parte, enfoques como el método verbo-tonal enfatizan la importancia del eje de claridad tonal en el diagnóstico y corrección de errores vocálicos. No obstante, no existen modelos computacionales accesibles que permitan automatizar esta evaluación desde dicho enfoque, ni herramientas integradas en entornos móviles que faciliten la retroalimentación en tiempo real sin intervención docente especializada.

1.3 Delimitación del problema

Aunque existen modelos basados en redes neuronales que procesan espectrogramas o audio crudo para detectar errores de pronunciación, estos no los clasifican en términos ligados a la causa del error.

1.4 Pregunta de investigación

¿Cómo construir un modelo de *machine learning*, utilizando redes neuronales, que permita detectar y clasificar errores de pronunciación en fonemas vocálicos entre aprendices mexicanos de francés, ofreciendo retroalimentación desde una aplicación móvil?

1.5 Propuesta de Solución

Por lo que se propone construir un modelo que, además de detectar errores, los clasifique con base en los formantes de la voz, los cuales se pueden utilizar como parámetro de diagnóstico del error de pronunciación, de acuerdo con el método de corrección verbo-tonal.

1.6 Justificación

El diagnóstico fonético en la enseñanza de lenguas extranjeras, tradicionalmente depende de docentes con formación especializada y sesiones individualizadas, lo cual es inviable en contextos grupales. En respuesta, diversos estudios han explorado la automatización del diagnóstico fonético mediante aprendizaje profundo y procesamiento del habla [17], [18].

Técnicas como las redes convolucionales o recurrentes aplicadas a características acústicas como MFCCs, espectrogramas - o incluso señales crudas - han mostrado mejoras relevantes sobre métodos clásicos como el Goodness of Pronunciation (GoP), alcanzando incrementos de más de 10 puntos porcentuales en F1-score [17], [18], [19]. Sin embargo, la mayoría de estos modelos carecen de interpretabilidad didáctica, es decir, detectan los errores de pronunciación, pero no la causa de este, por lo cual presentan dificultades de implementación fuera del laboratorio.

Algunos estudios han buscado mitigar la falta de datos anotados utilizando errores simulados o corpus sintéticos, obteniendo resultados prometedores (precisión del 77.1 %) [11]. Asimismo, enfoques recientes han propuesto la personalización de filtros acústicos para adaptarse al perfil L1 del aprendiz [6], reforzando la viabilidad de soluciones adaptativas.

Este trabajo propone un modelo integral basado en formantes y claridad tonal, entrenado con datos de hablantes mexicanos, desplegado en la nube y accesible desde una aplicación móvil. La solución busca combinar rigurosidad fonética con portabilidad tecnológica, ofreciendo así una herramienta práctica y escalable para la enseñanza del francés como lengua extranjera.

1.7 Objetivos

1.7.1 Objetivo general

Proponer un modelo de detección y clasificación de errores en la pronunciación de los fonemas vocálicos del idioma francés, empleando *machine learning,* para identificar errores de pronunciación comunes en alumnos hispanohablantes mexicanos, que hablan francés como lengua extranjera, presentando retroalimentación sobre estos.

1.7.2 Objetivos específicos

- Definir el esquema de medición en la pronunciación del alumno para proporcionarle retroalimentación.
- Definir las métricas de evaluación de rendimiento del modelo.
- Construir el *dataset* con base en la recolección de muestras de audio de pronunciaciones de alumnos mexicanos aprendices de francés.
- Evaluar el modelo propuesto, a partir de las métricas seleccionadas.
- Construir un prototipo de aplicación móvil para la presentación del diagnóstico de pronunciación, a partir del modelo.

1.8 Estado del Arte

En esta sección se presenta el estado del arte como resultado de la información obtenida en la investigación literaria realizada para el desarrollo del presente trabajo, de acuerdo con artículos, revistas y trabajos de tesis, los cuales se concentran en la Tabla 2, al final de esta sección.

En [20], Zhang et al. observan un gran crecimiento de las herramientas para el entrenamiento de pronunciación asistido por computadora (CAPT por sus siglas en inglés), las cuales emplean diversas tecnologías, incluyendo una que ellos consideraron como clave: la detección y diagnosis de mala pronunciación (MDD por sus siglas en inglés), la cual da retroalimentación correctiva para guiar a estudiantes de un idioma no nativo (L2). Así mismo, los autores proponen la siguiente clasificación para herramientas de MDD, según el método que estas utilizan:

- *Force-alignment*: compara un texto de referencia con la pronunciación y le da un puntaje mediante una función heurística, por ejemplo, con el algoritmo *Goodness of Pronunciation* (GOP).
- *Phoneme recognition*: Convierte audio L2 en fonemas reconocibles y da retroalimentación basado en fonemas de textos de referencia.
- *End-to-end (E2E):* predice los errores de pronunciación sin datos intermedios.

También se observó en este estudio que los modelos no-autorregresivos como el Clasificador Temporal Conexionista (CTC) son los más populares, así como que la mayoría de los modelos MDD de nivel fonético son no-autorregresivos entrenados con pérdida CTC, lo cual puede causar una falta de fidelidad en sus predicciones, debido a que pueden llegar a predecir fonemas consecutivos idénticos, o bien predecir secuencias de fonemas impronunciables. Los estudiantes L2 suelen arrastrar los fonemas de sus lenguas nativas al pronunciar palabras extranjeras. La mayoría de los métodos MDD solo usan datos de entrenamiento de los fonemas del lenguaje por aprender. La mayoría de los modelos MDD son entrenados con *datasets* que no capturan acertadamente variaciones de pronunciación de lenguaje nativo o acentos del estudiante.

Finalmente, los autores encontraron que, en comparación con los enfoques CTC existentes que ignoran la historia previa durante el reconocimiento de fonemas, el enfoque fonético RNN-T autorregresivo propuesto puede aprender y aprovechar patrones de pronunciación errónea de secuencias de fonemas L2, utilizando un conjunto de fonemas extendido y un corpus de entrenamiento débilmente supervisado, compuesto por corpus masivos L1, L2 y codificados, lo cual condujo a reducir significativamente la tasa de aceptación falsa, al tiempo que supera a los modelos basados en CTC en F1score y PER para estudiantes de inglés como lengua extranjera, nativos del idioma español.

En otro estudio [19], los autores argumentaron que un modelo integral para los fenómenos fonológicos y fonéticos, debe basarse en redes neuronales porque estos modelos pueden manejar tanto la creación gradual de categorías, como la dispersión auditiva; integrando el comportamiento lingüístico discreto de apariencia simbólica a partir de datos de entrada graduales en los niños, lo que los hace más plausibles biológicamente y prometedores para la lingüística teórica. El propósito de los autores fue demostrar la viabilidad y necesidad de los modelos de redes neuronales para proporcionar un marco unificado que explicara diversos datos lingüísticos y hallazgos experimentales en el contexto de la fonología y la fonética.

Su investigación encontró que los modelos de redes neuronales pueden representar eficazmente fenómenos lingüísticos claves, como el efecto del imán perceptual y la dispersión auditiva, que son cruciales para entender la adquisición y evolución del lenguaje. Al simular estos modelos con datos sintéticos, mostraron que las redes neuronales pueden cerrar la brecha entre las representaciones continuas y discretas en el procesamiento del lenguaje, ofreciendo un enfoque más integral y biológicamente plausible para modelar el conocimiento fonológico y fonético.

En el estudio [18], los investigadores propusieron un método para la detección automática de errores de pronunciación comunes cometidos por hablantes no nativos de inglés mediante redes neuronales convolucionales y recurrentes, debido a que estos modelos pueden predecir con alta precisión la presencia de errores típicos en palabras específicas, basándose en características acústicas extraídas de grabaciones de audio. La intención de los autores fue demostrar la eficacia de este método para proporcionar una herramienta de aprendizaje asistido por computadora que ofrezca retroalimentación individualizada y correcciones en tiempo real a los aprendices de inglés.

Entre sus hallazgos, se encontró que el sistema propuesto es capaz de detectar errores de pronunciación con alta precisión, en la mayoría de las palabras, logrando un incremento medio precisión de 12.21 puntos porcentuales, sobre la regla cero en un conjunto representativo del *dataset*. El estudio utilizó una combinación de redes neuronales convolucionales, redes neuronales recurrentes y su combinación (CRNN) para optimizar la topología y los hiperparámetros, lo cual se validó mediante un proceso de optimización bayesiana. Los resultados muestran que el sistema supera significativamente las técnicas previas y ofrece una metodología robusta para la detección de errores sin necesidad de anotaciones fonémicas detalladas.

En el estudio reportado en [11], Fontan realizó un análisis de las herramientas automatizadas existentes hasta la época para la enseñanza de FLE en estudiantes japoneses, en el cual se determinó que el más utilizado y emblemático era *Goodness of Pronunciation* – lo que coincide con lo reportado en [20] – que se utiliza cuando se conoce de antemano el enunciado objetivo a ser pronunciado y se usa en tareas de lectura o repetición. En dicho artículo, se determinó que los sistemas de retroalimentación eran creados de acuerdo con limitaciones técnicas, más que con la idoneidad didáctica. Desde entonces se ha buscado cada vez más crear sistemas que den retroalimentación concreta al estudiante sobre cómo mejorar su pronunciación.

En su trabajo de tesis [6], Huerta Espinoza llevó a cabo un análisis de la fonética del Español y el Francés para entender de dónde provienen las dificultades en el aprendizaje del segundo entre alumnos mexicanos. La autora del estudio, quien también es profesora de FLE, elaboró también una encuesta para evaluar la percepción de la dificultad de pronunciación de cada fonema, la discriminación auditiva (encuesta oral), y la relación entidad sonora-grafía (asociar palabras con símbolos del fonema), la cual mostró que el 78% de los encuestados percibían el francés como una lengua difícil de aprender. La encuesta se aplicó en 2013 a 709 alumnos de FLE hispanófonos (con 21 años como media de edad y que en promedio tendrían alrededor de 1 año estudiando el FLE), en la CDMX en instituciones como la UNAM, la Alianza Francesa, el CUC, el Olinca, el TAE, el INHUMYC, la UIC, el Liceo Franco Mexicano y en varios Institutos Gastronómicos.

La investigación presentada en [17], presenta un enfoque innovador para la detección y diagnóstico de errores de pronunciación (MDD) al procesar directamente formas de onda crudas en lugar de depender de características acústicas diseñadas manualmente. El modelo utiliza el módulo SincNet, que emplea filtros de paso de banda parametrizados basados en funciones seno cardinal (sinc), optimizados para capturar información espectral clave como los formantes. Estos elementos

son cruciales en la pronunciación, ya que reflejan las frecuencias de resonancia de las vocales. El enfoque propuesto no solo permite al modelo aprender representaciones acústicas específicas de las desviaciones fonéticas, sino que también mejora la interpretabilidad al alinearse con propiedades perceptivas del sistema auditivo humano. En experimentos con el dataset L2-ARCTIC, el modelo demostró una mayor precisión en la detección de errores relacionados con formantes, alcanzando resultados competitivos o superiores a los métodos tradicionales en métricas, como la tasa de error por fonema (PER) y el diagnóstico de pronunciación. Esto subraya la capacidad de SincNet para capturar y modelar características acústicas críticas para guiar el aprendizaje de la pronunciación, de una forma más precisa en hablantes no nativos.

En cuanto a aplicaciones actualmente disponibles en el mercado, para las tareas de detección y diagnóstico de la pronunciación, algunos de los principales exponentes encontrados durante la revisión literaria son los siguientes: ELSA Speak [21], la cual es una aplicación móvil de pago que utiliza la inteligencia artificial para proveer retroalimentación sobre la pronunciación, entonación y fluidez, mostrándolos como porcentajes y dando el detalle hasta el nivel de las sílabas donde hubo errores. A la fecha de redacción de este documento, esta aplicación se encuentra solo disponible para el idioma inglés, pero el soporte para otros idiomas se encuentra en desarrollo; Speech Ace [22] es una plataforma en línea de cursos de enseñanza de varios idiomas, incluido el francés, la cual cuenta también con una API de pago, capaz de proveer retroalimentación sobre la pronunciación a nivel de fonemas; Azure AI Services [23] es el servicio en la Nube de Microsoft que cuenta con APIs para la conversión de audio a texto, siendo la evaluación de la pronunciación una de las muchas funcionalidades que ofrece. Este servicio también es capaz de reconocer fonemas y entregar un diagnóstico con porcentajes de precisión con base a un texto de referencia.

De igual forma, también existen herramientas que permiten analizar señales de audio y producir mapas acústicos a partir de estas, para así detectar (manualmente) los fonemas contenidos en ellas. Por ejemplo, Praat [24] es un software especializado en este campo, el cual cuenta con múltiples funcionalidades, como la extracción de formantes – frecuencias fundamentales de una señal de audio cuyo análisis puede emplearse para caracterizar a los fonemas [25] – y generación de espectrogramas, cálculo de la energía y la media cuadrática de la señal, entre muchas otras. Sin embargo, este es un proceso un tanto complicado y la herramienta y requiere de una PC para realizar el análisis.

En la Tabla 1.1 se incluye la información de los estudios revisados.

Tabla 1.1. Clasificación de la literatura consultada.

Título	Autor	Año	Tipo
Phonetic RNN-Transducer for Mispronunciation Diagnosis	D. Y. Zhang, S. Saha y S. Campbell	2023	Articulo
Neural network models for phonology and phonetics	Paul Boersma, Titia Benders y Klaas Seinhorst	2020	Artículo
Detection of Typical Pronunciation Errors in Non-native English Speech Using Convolutional Recurrent Neural Networks	Aleksandr Diment, Eemi Fagerlund, Adrian Benfield y Tuomas Virtanen	2019	Artículo
Évaluer la parole des apprenants de FLE: approches et outils automatiques	Lionel Fontan	2017	Revista
Fonética comparada, español-francés, francés como segunda lengua para hispanohablantes, los fonemas complicados: contraste fónico de una lengua extranjera	Huerta Espinosa, Ericka	2013	Tesis
End-to-End Mispronunciation Detection and Diagnosis from Raw Waveforms	Yan y B. Chen	2021	Artículo

Fuente: Elaboración propia.

1.9 Clasificación de la investigación

La investigación desarrollada se encuentra ubicada en la línea de Desarrollo de Sistemas para el Cómputo Móvil de la SEPI ESCOM y, de acuerdo con la lista de grupos de interés especializados de la *Association for Computing Machinery*, se clasifica en el de inteligencia artificial.

En este capítulo se estableció la problemática a estudiar y su objetivo, adicionalmente, se mencionaron algunas investigaciones que se relacionan con lo planteado. Para el desarrollo de los objetivos enlistados, se requieren varios conocimientos teóricos que permitan realizar la investigación, los cuales se abordan en el Marco Teórico y componen la base sobre la cual se sustenta la tesis.

CAPÍTULO 2: Marco Teórico

Este capítulo se divide en dos secciones principalmente, en la primera se aborda lo referente a la fonética del idioma francés, incluyendo qué es un fonema, el origen de estos desde un punto de vista fisiológico, y los métodos de enseñanza que existen para la corrección de su pronunciación. La segunda parte se centra en revisar algunas técnicas de reconocimiento y análisis de fonemas, así como en explicar brevemente los principios sobre los cuales se basan los dos tipos de redes neuronales artificiales consideradas para el desarrollo del modelo de clasificación de errores en la pronunciación a proponer.

2.1 Fonemas vocálicos del francés

2.1.1 ¿Qué es un fonema?

Antes de hablar de fonema, se debe hablar de fonología y fonética: dos conceptos comúnmente utilizados de manera intercambiable. Mientras ambas son ciencias cuyo objetivo es el estudio de los sonidos del lenguaje, lo que las diferencia es el enfoque con el que lo abordan. La fonética se preocupa de la manera como los sonidos se producen, se transmiten y son percibidos por los interlocutores. La fonología trata de descubrir cómo es que estos sonidos contribuyen al funcionamiento de una lengua en el acto del habla y cómo aseguran su codificación [25].

Como se explica en [6], la unidad mínima de la fonología son los fonemas, mientras que en la fonética son los fonos: los fonemas se escriben entre líneas, por ejemplo /y/, y los fonos entre corchetes, como [b].

Los fonemas son representaciones abstractas de sonidos establecidos entre sí según su función dentro de un todo organizado (el sistema fonológico). Por ejemplo, el sistema fonológico francés tiene 36 fonemas: 16 vocales – 14 dependiendo del autor, debido a que los fonemas /ɑ/ y /œ̃ / tienden a ser asimilados debido a la falta de palabras o a regionalismos – 17 consonantes y 3 semivocales (también llamadas semiconsonantes) [26].

En el Anexo A se provee una lista completa de los fonemas definidos por la Asociación Internacional de Fonética, entre ellos, los del idioma francés.

2.1.2 Formantes

De manera muy simple, y citando a Astésano [27], “los formantes corresponden a las frecuencias de resonancia naturales de las cavidades supraglóticas, en las cuales el tamaño y la forma son determinados por la configuración articulatoria de la vocal. Sobre el espectro [acústico], los formantes corresponden a picos locales de amplitud”.

Según CALLIOPE[1] [25], la estructura acústica de las vocales se caracteriza principalmente por la presencia de máximos espectrales, es decir, zonas de frecuencia donde los armónicos tienen mayor intensidad, los cuales se denominan formantes.

[1] Nombre colectivo que se dio al conjunto de 36 autores cuyas obras se concentraron en el libro “La parole et son traitement automatique” en 1989, como producto de las sesiones organizadas por el Grupo de la Comunicación Hablada de la Sociedad Francesa de Acústica (SFA por sus siglas en francés).

Los sonidos de los fonemas sonoros, tanto vocales como de algunas consonantes como [b], [d] y [g], se producen por medio de vibraciones de las cuerdas vocales, denominadas F_l. La frecuencia fundamental (referida como F_0 en el contexto del análisis de Fourier) de una señal del habla, puede obtenerse como una estimación de la frecuencia de vibración laríngea. Las variaciones del valor de F_0 en el tiempo componen la curva melódica de la frase.

La medición antes mencionada puede realizarse a partir de una señal en el dominio del tiempo, después del filtrado de la señal, así como en el dominio espectral, por medio del análisis del espectro de un sonido hablado; representación que incluye no solo las frecuencias de una señal, sino también la distribución de la energía o potencia de la señal en dichas frecuencias. La frecuencia laríngea también puede ser estimada mediante la variación de impedancia eléctrica a nivel de la glotis, así como por la observación de datos fisiológicos ligados a la vibración de las cuerdas vocales, en cuyo caso el cálculo de la frecuencia fundamental está dado por la ecuación ((1).

$$F_l = \frac{1}{(t_1 - t_2)}, para\ t_1 < t \leq\ t_2 \qquad (1)$$

donde $t_1\ y\ t_2$ designan el inicio de dos ciclos de vibración glotal[2] consecutivos.

La frecuencia laríngea puede variar considerablemente durante la fonación. En casos extremos, es posible observar transiciones que van de 100 a 400 Hz (cambio del régimen de fonación normal al falsete) durante el intervalo de dos o tres ciclos. Por otro lado, los ciclos sucesivos presentan variaciones de varios puntos porcentuales en torno al valor medio, dependiendo, entre otras cosas, del estado fisiológico. Más adelante veremos que esta frecuencia fundamental del habla no se considera en el análisis del espectro acústico de las vocales, ya que los cambios melódicos en la voz no afectan la construcción de estas.

Por otra parte, un aumento de la apertura articulatoria da origen al siguiente formante, denominado F1, del mismo modo, una anteriorización de la articulación, corresponderá a un aumento de la siguiente frecuencia fundamental, F2. Para las vocales anteriores, un aumento en la labialización (redondeo de los labios) resulta en una disminución en F2, al igual que de la tercera frecuencia fundamental – denominada F3 – con una articulación lingual constante. La caída de F3 es un índice menos ambiguo, porque no está ligado a una posible posteriorización de la articulación. Para las vocales posteriores, labializadas en francés (u, o, ɔ), observamos la misma caída de F3. Pero este formante es muy débil y su medición puede resultar difícil en espectrografía.

En la Figura 2.1, se muestra la relación entre la apertura articulatoria y la posición de los labios, con los formantes que estos producen.

[2] Relativo a la glotis, que es el orificio o abertura anterior de la laringe.

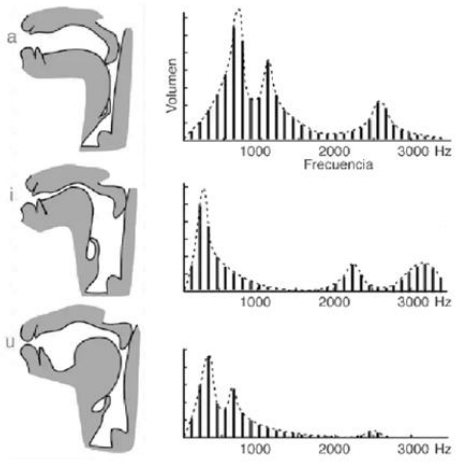


Figura 2.1. Corte transversal del aparato fonatorio correspondientes a los fonemas [a], [i] y [u], y sus frecuencias fundamentales.

Fuente: [28]

Conviene representar una vocal en un plano con F1 y F2 como ejes, donde las vocales "extremos" [i], [y] y [u] están dispuestas en los extremos de un triángulo que apunta hacia abajo. Este triángulo articulatorio representa de forma muy aproximada la posición media de la lengua en la cavidad bucal según dos ejes denominados "anterior-posterior" y "abierto-cerrado", dependiendo de si la lengua se masajea hacia delante y hacia la zona dentaria durante [i], bajo y extendido lejos del paladar para [a] (abierto), o con masa posterior hacia el tejido blando para [u]. En la Figura 2.2 se muestra una representación de dicho plano.

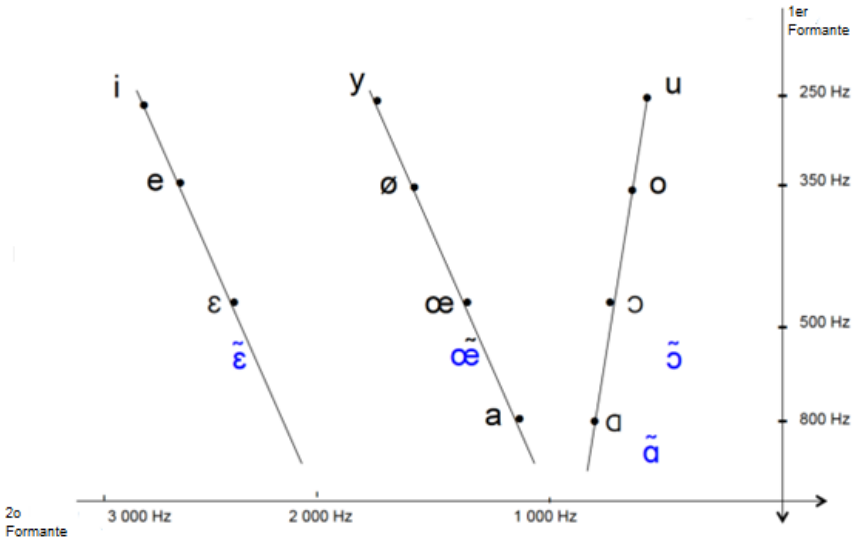


Figura 2.2. Clasificación de las vocales del francés, según sus formantes F1 y F2.

Fuente: Traducido de [15]. C-> F2 y T -> F1

De lo explicado en [25], se destaca que el proceso del habla depende de varios parámetros, como los músculos relacionados a las cuerdas vocales, la presión subglótica, etc. Por lo tanto, hay una interacción compleja de movimientos de la glotis y las variaciones temporales de la forma del conducto vocal. En consecuencia, al no ser un fenómeno estacionario, la medición de la frecuencia laríngea requiere de técnicas de análisis que se adapten a este tipo de fenómenos y que superen las limitantes de las herramientas matemáticas tradicionalmente usadas en el tratamiento de señales periódicas.

2.1.3 Áreas de dispersión e inferencia

Como se mencionó en el subtema anterior, el fenómeno de la voz es complicado y depende de múltiples factores, lo cual produce mediciones únicas toda vez que incluso el mismo locutor pronuncie un fonema determinado. Si bien diversos estudios han establecido las frecuencias "canónicas" para los fonemas vocálicos, en la práctica el cerebro es capaz de interpretar los fonemas correctamente aun cuando existan estas variaciones.

Si se trazan los valores de F1 y F2 de estas vocales en un espacio biplano, se obtiene un "mapa" como el de la Figura 2.3.

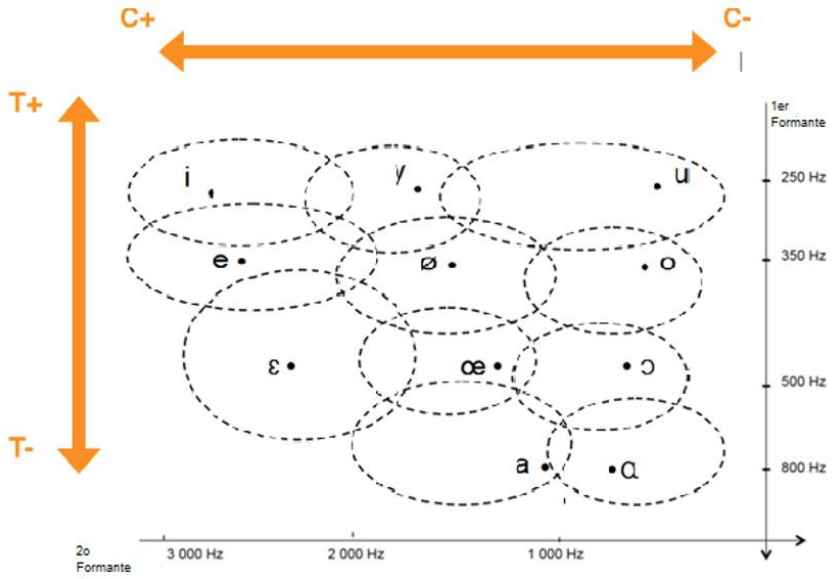


Figura 2.3. Áreas de dispersión de las vocales (orales) del francés.

Fuente: Traducido de [29].

Luego entonces, partiendo de dicho plano, se le llama área de dispersión al espacio radial alrededor de un punto canónico que representa un fonema, donde la comprensión de este se puede suscitar por el interlocutor. Sin embargo, cabe señalar la existencia de zonas de traslape entre las áreas de dispersión de distintos fonemas. Estas zonas superpuestas corresponden a zonas de interferencia, donde los sonidos producidos tienen un timbre particular que no corresponde a los patrones sonoros canónicos, los cuales son habituales en el habla espontánea y, la mayoría de las veces, no son graves, ya que el cerebro procesa la información en contexto, compensando y restaurando a la forma "correcta" a la que se está acostumbrado [30]. Más allá de dichas zonas, no será ya posible la comprensión del mensaje original.

2.2 Métodos de corrección de la pronunciación

Para Billières [13], en las décadas de 1970 y 1980, los manuales de fonética correctiva se basaban en una serie de instrucciones como "Escuchar", "Repetir", "Comparar" y "Leer", organizadas de diferentes formas. En contraste, desde finales de la década de 1980, los libros de fonética correctiva mejoraron notablemente en contenido y presentación, ofreciendo una variedad de actividades que se pueden agrupar en cinco categorías principales:

- Ejercicios para mejorar la percepción y discriminación auditiva.
- Prácticas de repetición e integración que incluyen:
  - Palabras individuales
  - Frases breves
  - Ejercicios estructurales
  - Mini diálogos
  - Rimas y poemas
- Actividades para desarrollar la sensibilidad al ritmo, la entonación y las características del francés hablado, como los fenómenos de enlace, elisión y asimilación.
- Ejercicios que relacionan la pronunciación con la escritura (grafema-fonema), con explicaciones detalladas.
- Diversas actividades lúdicas para el aprendizaje.

A continuación, se explican los fundamentos del método verbo-tonal en materia de fonética correctiva, aplicada a la didáctica de lenguas, utilizado por profesores de francés como lengua extranjera.

2.2.1 Método verbo-tonal

Tal y como expone Billières [31], el método verbo-tonal tiene su origen en la década de 1950, cuando Petar Guberina (1913-2005) realizaba trabajos de investigación en Croacia para mejorar la pronunciación de estudiantes de francés en ese país, así como para la reeducación de personas con problemas de sordera a consecuencia de los bombardeos de la 2ª Guerra Mundial. En este periodo, Guberina realizó varios estudios de audiometría – medición de la audición mediante aparatos para determinar las frecuencias que son percibidas por la oreja de un paciente – que lo llevaron a desarrollar la audiometría verbo-tonal: utilización de estímulos del habla (verbo) con el fin de evaluar la sensibilidad auditiva a diferentes frecuencias (tonal).

Como parte de dicha labor, Guberina descubrió, mediante la utilización de aparatos para la detección de la sordera llamados SUVAG, que la oreja es muy sensible a los cambios de “altura” de un sonido o tonalidad, lo cual provoca que un sonido del habla (consonante o vocal) pueda ser percibido y comprendido mejor en cierta octava[3] de este, obteniendo así una lista de frecuencias “óptimas” en las cuales el cerebro percibe las vocales del francés de mejor manera.

[3] Se habla de octava cuando, partiendo de un sonido a una frecuencia base, se aumenta o disminuye al doble dicha frecuencia, permitiendo percibirlo de manera muy similar, pero en un tono distinto, es decir, más agudo o grave.

A continuación, se muestra en la Tabla 2.1 las frecuencias óptimas en Hertz (Hz) para los fonemas del francés

.

Tabla 2.1. Frecuencias óptimas de algunos fonemas del francés.

Fonema	Frecuencia (Hz)
i	3200 - 6400
e ɛ	1600 - 3200 2400 - 4800
u	150 - 300
ɔ o	400 - 800 300 - 600
a ɑ	1200 - 2400 600 - 1200
t	1600 - 3200
p	300 - 600
s	6400 - 12800
l	800 - 1600

Fuente: Traducido de [32].

En este método, la intelectualización del funcionamiento del sistema articulatorio no se recomienda, ya que se parte de la idea de que una mejor percepción conduce a una mejor producción, para lo cual, en principio, no es necesario el conocimiento detallado de la articulación: se parte de situaciones prácticas durante las sesiones de enseñanza.

Cabe destacar, que en este método también se le da gran importancia a la prosodia, indicando al alumno con movimientos corporales el ritmo que debe llevar la frase. Aunado a lo anterior, los profesores que usan este método también emplean los entornos facilitadores: alteraciones a la palabra o sílaba donde el alumno muestra dificultades, cambiando partes de estas por otras que ayuden a incrementar/decrementar la atención y aumentar/bajar la altura del tono, con el fin de llevar al alumno a la pronunciación correcta de manera progresiva.

En la Tabla 2.2, se muestra un ejemplo del procedimiento de corrección de este método a partir de su diagnóstico.

Tabla 2.2. Ejemplo de técnicas de corrección de la pronunciación en la MVT, según el eje de claridad tonal.

	Sonido objetivo	[y]	
	Sonido realizado	[i]	[u]
	Diagnóstico	Tono muy claro, hay que oscurecerlo	Tono muy sombrío, hay que aclararlo
Procedimientos de corrección	Entonación	Descendente	Ascendente
		hueco de entonación	pico de entonación
	Pronunciación matizada	[y] -> [u]	[y] -> [i]
	Entorno de facilitación	consonantes oscurecidas	consonantes aligerantes
		f v p b m	s z t d
	Gestos	Descendentes, con mano/cabeza acompañando el movimiento entonativo	Ascendentes, con mano / cabeza acompañando el movimiento entonativo

Fuente: Fonetix: Formación de instructores de corrección fonética [33].

2.3 Reconocimiento de fonemas

Existen varios métodos para el reconocimiento de fonemas. En el caso de la herramienta Praat, para el análisis fonético [24], el sonido se re-muestrea a una frecuencia de muestreo del doble del valor del techo de formante (máxima frecuencia a considerar para la búsqueda de formantes). Después de esto, se aplica una etapa de preénfasis α, calculada como se observa en la ecuación ((2):

$$\alpha = \exp(-2\pi F \Delta t) \tag{2}$$

donde Δt es el periodo de muestreo del sonido.

Cada muestra x_i del sonido, excepto x_1, es entonces cambiada, bajando de la muestra anterior, como se aprecia en la ecuación ((3):

$$x_i = x_i - \alpha x_{i-1} \tag{3}$$

Para cada ventana de análisis, Praat aplica una ventana de tipo gaussiano y calcula los coeficientes LPC con el algoritmo de Burg. El número de "polos" que calcula este algoritmo es el doble del número máximo de formantes.

Inicialmente, el algoritmo encontrará el número máximo de formantes en todo el rango entre 0 Hz y el techo de formantes. Por consecuencia, los formantes encontrados inicialmente pueden tener a veces frecuencias muy bajas (cerca de 0 Hz) o frecuencias muy altas (cerca del techo de formantes). Estos "formantes" bajos o altos tienden a ser artefactos del algoritmo LPC, es decir, el algoritmo tiende a usarlos para igualar la pendiente espectral, si esa pendiente difiere de la suposición de 6 dB/octava. Por lo tanto, estos "formantes" bajos o altos normalmente no pueden asociarse con las

resonancias del tracto vocal buscadas. Para que se pueda identificar los tradicionales F1 y F2, se eliminan todos los formantes por debajo de 50 Hz y todos los formantes por encima del techo de formantes menos 50 Hz.

2.3.1 Análisis del espectro acústico

De acuerdo con [25], se pueden ubicar los formantes dentro de un espectrograma en forma de bandas negras más o menos paralelas al eje del tiempo. En el rango de validez del modelo de producción, es decir para frecuencias inferiores a aproximadamente 5 kHz – 5.5kHz para una voz femenina – las vocales de un sujeto masculino tienen 5 formantes. El primer formante (F1) es el pico espectral con la frecuencia más baja, el segundo formante (F2) es el siguiente pico, etc. No obstante, no se toma en consideración el pico de frecuencia muy baja (formante glotal), en torno a 200-300 Hz, que a veces puede aparecer para vocales "abiertas" (tipo ɛ, ɑ,a,ɔ,ə) de baja intensidad. Algunos experimentos han mostrado que la posición de frecuencia de los primeros 3 formantes caracteriza el timbre de las vocales. Ocasionalmente se observa un formante adicional en la región de 800 a 1200 Hz.

En la Figura 2.4 se presenta la señal de voz correspondiente a las vocales [u] y [y], junto con su espectrograma – elaborado con la herramienta Praat – donde se pueden apreciar los armónicos intensos de estas. También se muestran los formantes correspondientes en el dominio de la frecuencia.

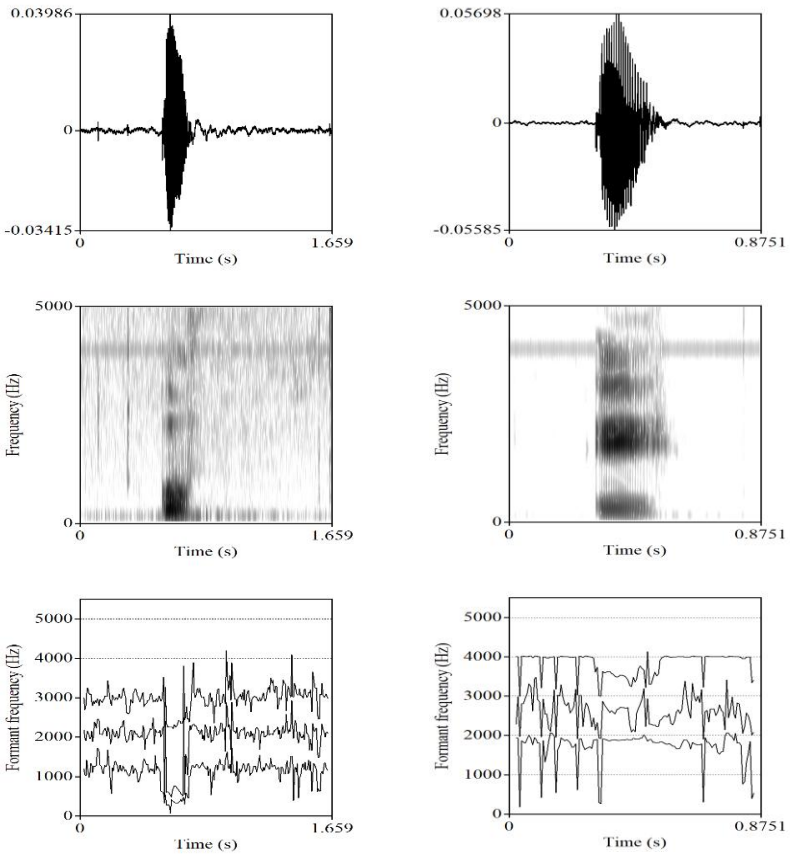


Figura 2.4. En orden descendente: señal de audio, espectrograma y formantes de los fonemas /u/ y /y/.

Fuente: Elaboración propia.

2.3.2 Distancia euclidiana

La distancia euclidiana es la distancia en línea recta entre dos puntos [34], y es una medida utilizada para cuantificar la similitud entre dos puntos en un espacio multidimensional. En el contexto del análisis acústico de vocales, esta métrica permite evaluar cuán distante se encuentra una vocal producida por un hablante no nativo respecto a su contraparte de referencia, basada en formantes típicos de hablantes nativos.

Matemáticamente, si se considera un vector de características acústicas x = [F1, F2] de una muestra de voz, y otro vector c = [F1*, F2*] correspondiente al centroide de un fonema de referencia, la distancia euclidiana se define como se muestra en la ecuación (4):

$$d(x, c) = \sqrt{\{(F_1 - F_1^*)^2 + (F_2 - F_2^*)^2\}} \quad (4)$$

Esta medida proporciona una forma intuitiva de estimar desviaciones fonéticas en el plano acústico, permitiendo clasificar errores de pronunciación cuando la distancia a un fonema objetivo supera a la de otros centros vocálicos.

2.3.3 Distancia Z-score

Las puntuaciones *Z (Z-score* en inglés) miden la distancia de un punto de datos desde la media en términos de la desviación estándar [35]. La distancia basada en Z-score constituye una alternativa a la distancia euclidiana que incorpora normalización estadística de las características acústicas, tomando en cuenta la variabilidad propia de cada dimensión. Esta métrica resulta útil cuando las dimensiones del espacio tienen escalas distintas, o cuando se desea ajustar las distancias en función de la dispersión de los datos.

En este enfoque, cada dimensión del vector acústico se normaliza utilizando su media y desviación estándar, generando un vector estandarizado z = [z_1, z_2], empleando la fórmula de la ecuación ((5):

$$z_i = (x_i - \mu_i)/\sigma_i \quad (5)$$

Luego, la distancia entre la muestra y el centroide se calcula en este espacio estandarizado como una distancia euclidiana sobre los valores Z. Este procedimiento permite comparar muestras independientemente de su escala original y resulta particularmente útil para compensar diferencias entre géneros o condiciones de grabación [36], como se hace al aplicar techos de frecuencia distintos para hombres y mujeres.

2.4 Inteligencia artificial para la tarea de clasificación

Para Goodfellow et al. [37], en la actualidad, la inteligencia artificial tiene como principal reto resolver tareas que son fáciles de realizar para un humano, pero que son difíciles de describir formalmente, es decir, problemas que se pueden solucionar de manera intuitiva o casi automática por una persona, como por ejemplo, el reconocimiento de rostros en una imagen o de las palabras habladas – siendo el segundo de especial interés para el desarrollo de esta tesis.

Una posible solución para este reto – la que resulta más relevante para el caso de estudio de esta tesis – es permitir que las computadoras aprendan a partir de la experimentación y entiendan el mundo mediante la jerarquía de conceptos, definiendo cada uno a través de relaciones con otros más simples. Es en este contexto donde se acuña el término de *AI deep learning* (inteligencia artificial con aprendizaje profundo), ya que, si se representaran en un grafo las relaciones que se forman entre estos conceptos, que se van construyendo a partir de otros, se notaría que el grafo tiene gran profundidad e implica múltiples capas.

Aun cuando puede no resultar tan evidente, el resolver tareas que parecen triviales para un humano, como la detección del habla, implica una basta cantidad de conocimiento sobre el mundo, dando lugar a que las computadoras se vean ante la necesidad de modelar y adquirir tal conocimiento. Para resolver esta tarea, han surgido diferentes propuestas, por ejemplo, la construcción de bases de conocimiento, que son enunciados en lenguajes formales que representan al mundo, a partir de los cuales una computadora puede razonar usando reglas lógicas de inferencia. Otro enfoque es el de *machine learning* (aprendizaje automático), con el cual se pretende superar las limitaciones del anterior, dando a la IA la capacidad de adquirir su propio conocimiento extrayendo patrones en datos "crudos". En la Figura 2.5 se muestra la relación de jerarquía entre estos conceptos.

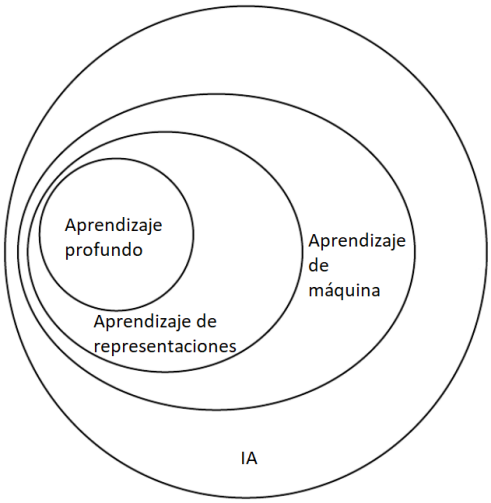


Figura 2.5. Diagrama de Venn de la Inteligencia Artificial.

Fuente: Traducido de [37].

Muchas tareas pueden ser resueltas con el aprendizaje automático, teles como la predicción, la detección de anomalías y la clasificación – entre otras. En esencia, en programas para la clasificación, se le pide a la computadora que determine a cuáles k categorías pertenece alguna entrada. Para resolver este problema, el algoritmo de aprendizaje produce una función $f: R^n \rightarrow \{1, \ldots, k\}$, de tal forma que cuando $y = f(x)$, el modelo asigna una entrada descrita por un vector x a una categoría identificada por un código numérico y. También existe la variante de esta tarea de clasificación, donde $f(x)$ regresa una probabilidad de distribución sobre las clases [37].

Durante el proceso de aprendizaje – también conocido como entrenamiento – se utiliza un conjunto de datos llamado *dataset,* el cual contiene muchos ejemplos – también llamados *data points.*

Es importante mencionar que los algoritmos de aprendizaje automático pueden categorizarse de manera general en supervisados y no-supervisados, según el tipo de proceso de aprendizaje que

tienen. A grandes rasgos, un algoritmo de aprendizaje supervisado opera con *datasets* que contienen características junto con etiquetas u objetivos – se dice que los datos están "anotados" – que, en el caso de la tarea de clasificación, le indican de antemano al algoritmo de aprendizaje a qué clase corresponde el *data point* proporcionado. Por otra parte, el aprendizaje no-supervisado implica la observación de múltiples vectores x aleatorios y tratar de aprender o determinar propiedades de interés sobre estos; no hay una guía o "maestro". También existen otras variantes, como el aprendizaje semi-supervisado, donde algunos elementos del *dataset* están anotados y otros no.

En párrafos previos, se ha comentado la forma en cómo se puede abordad la inteligencia artificial, desde un enfoque que trabaja con conceptos relacionados entre sí para modelar al mundo – asemejando redes – para encontrar propiedades interesantes de los datos. Una forma de abordar el aprendizaje automático es imitando las neuronas del cerebro humano, mediante el concepto de redes neuronales artificiales.

De acuerdo con Hagan et al. [38], aun cuando las redes neuronales artificiales no se acercan a la complejidad del cerebro humano, existen dos similitudes clave entre lo biológico y las redes neuronales artificiales. Primero, los componentes básicos de ambas redes son simples: dispositivos computacionales que están altamente interconectados. En segundo lugar, las conexiones entre neuronas determinan la función de la red.

A continuación, se explica brevemente el tipo de red neuronal artificial considerada en el desarrollo de la tesis.

2.4.1 Redes Neuronales Recurrentes

Como se explica en [37], las redes neuronales recurrentes, o RNN, son una familia de redes neuronales para procesar una secuencia de valores $x^1, \ldots, x^\tau$. Estas redes pueden acomodar secuencias largas, mucho más de lo que es realista para redes sin especialización basada en cadenas. La mayoría de las redes recurrentes también pueden manejar secuencias de longitud variable.

Esto se logra, en gran medida, gracias al concepto de compartir parámetros a través de diferentes partes del modelo, lo cual es de gran importancia cuando una pieza específica de información puede ocurrir en múltiples posiciones dentro de la longitud de la secuencia.

Las RNN también se pueden aplicar en dos dimensiones a datos espaciales como imágenes, e incluso cuando se aplica a datos relacionados con el tiempo, la red puede tener conexiones que se remontan al pasado, siempre que toda la secuencia sea observada antes de ser proporcionada a la red.

Un grafo computacional es una forma de formalizar la estructura de un conjunto de cálculos, como aquellos que implican mapear entradas y parámetros con salidas y pérdidas. Al desplegar este gráfico, se obtienen parámetros compartidos en una estructura de red profunda. Por ejemplo, si se considera la forma clásica de un sistema dinámico, dada por la ecuación ((6):

$$s^{(t)} = f\left(s^{(t-1)}; \theta\right) \tag{6}$$

donde $s^{(t)}$ se denomina el estado del sistema.

Esta ecuación es recurrente porque la definición de s en el tiempo t referencia a la misma definición en el tiempo t-1.

Para un número finito de cambios en el tiempo τ, el grafo puede desplegarse aplicando la definición de τ – 1 veces. Por ejemplo, si se despliega la Ecuación (6 para el cambio en el tiempo τ = 3, se obtiene ((7):

$$s^{(3)} = f\left(s^{(2)}; \theta\right) = f(f\left(s^{(1)}; \theta\right); \theta) \tag{7}$$

Al desarrollar la ecuación ((7) aplicando de forma reiterada la definición, se obtiene una expresión que no implica ninguna repetición. Esta expresión ahora se puede representar mediante un gráfico de cálculo acíclico dirigido tradicional.

2.4.2 Algoritmo de Alineación Forzada

La alineación forzada (*forced alignment*) es una técnica utilizada ampliamente en el procesamiento automático del habla para determinar los límites temporales de fonemas, sílabas o palabras dentro de una señal de audio, dada su transcripción textual correspondiente. Este procedimiento resulta especialmente útil en contextos en los que se requiere etiquetar grandes volúmenes de datos, sin intervención manual directa.

El objetivo principal de la alineación forzada es sincronizar la secuencia textual con la señal acústica, produciendo marcas temporales (*timestamps*), que indican el inicio y fin estimado de cada unidad lingüística en la señal de voz. Para lograrlo, estos sistemas, generalmente, utilizan un modelo acústico entrenado previamente —como los basados en Hidden Markov Models (HMMs) o Deep Neural Networks (DNNs)— en conjunto con un diccionario fonético y un modelo del lenguaje [39]. En la Figura 2.6, se muestra un ejemplo de emisiones de probabilidad utilizadas por el algoritmo Forced Alignment de Pytorch Audio para identificar fonemas.

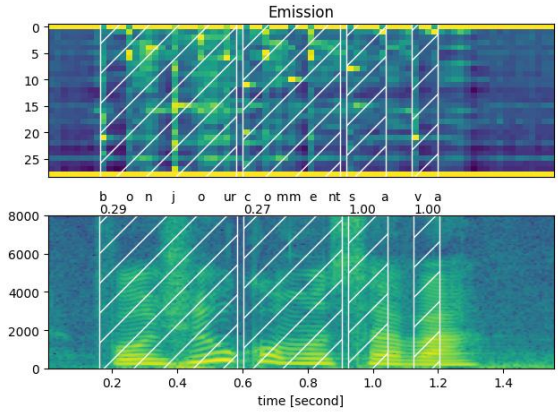


Figura 2.6. Emisiones de probabilidad generadas con Pytorch Audio.

Fuente: Elaboración propia.

En la práctica, uno de los sistemas más empleados para este propósito, es Montreal Forced Aligner (MFA), una herramienta de código abierto que automatiza el proceso de alineación, utilizando

el toolkit Kaldi como *backend* [40]. MFA toma como entrada una grabación de audio junto con su transcripción textual, y produce una alineación detallada a nivel fonémico, lo que permite extraer automáticamente las regiones temporales, en las que se encuentra cada fonema dentro del flujo de habla.

El funcionamiento interno de MFA combina modelos acústicos entrenados con representaciones fonéticas, utilizando Kaldi como *backend* para llevar a cabo la decodificación. Los modelos acústicos utilizados por MFA se entrenan a partir de *corpus* etiquetados fonéticamente. Están diseñados para capturar la relación estadística entre características espectrales del habla y secuencias fonémicas esperadas. Durante la alineación, MFA genera representaciones acústicas de las señales de entrada (por ejemplo, MFCCs) y las compara con los modelos acústicos preentrenados, para determinar la secuencia de fonemas más probable en función de la transcripción proporcionada.

El proceso de alineación incluye también el uso de un diccionario fonético, lo que permite mapear las palabras del texto a sus transcripciones fonémicas. Además, ajusta automáticamente los límites temporales cuando se utilizan modelos adaptados al dominio específico o al tipo de hablante. En la Figura 2.7, se observa un ejemplo del resultado de MFA, a partir de una señal de audio y su correspondiente transcripción textual. En este, se pueden apreciar los fonemas determinados por MFA, junto con sus marcas temporales.

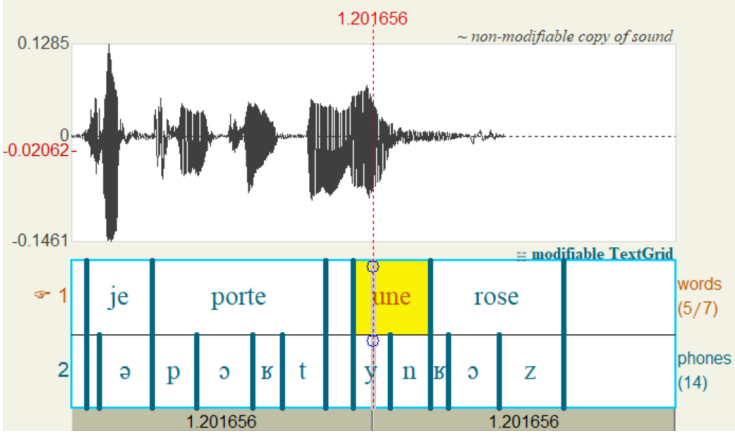


Figura 2.7. Visualización en Praat de una señal de audio anotada con MFA.

Fuente: Elaboración propia.

La precisión de la alineación forzada depende tanto de la calidad del modelo acústico como de la fidelidad de la transcripción. Aunque estos sistemas no son infalibles, han demostrado ser suficientemente robustos para tareas fonéticas como el análisis de segmentación vocálica, especialmente en estudios de aprendizaje de segundas lenguas (L2), donde el alineamiento preciso permite estudiar desviaciones acústicas en la producción del hablante.

En este capítulo se han explicado algunos métodos para la corrección de la pronunciación del idioma francés, así como las técnicas para el reconocimiento de los fonemas revisados durante el análisis del problema de investigación, incluyendo también algunas definiciones esenciales para el entendimiento de varios conceptos mencionados brevemente en el Introducción.

CAPÍTULO 3: Metodología de la investigación

En este capítulo se presentan los métodos, técnicas y procedimientos que se emplearon en el desarrollo de la investigación, en la que se evaluó la viabilidad de un sistema de diagnóstico automático de errores vocálicos en aprendices mexicanos de francés.

3.1 Enfoque metodológico

Según Sampieri [41], la metodología representa el conjunto de procesos lógicos mediante los cuales se abordan los problemas de investigación, articulando métodos y técnicas para garantizar resultados válidos y confiables. En este estudio, se adoptó un enfoque mixto, entendido como la combinación de estrategias cuantitativas, orientadas a la medición objetiva de variables y al análisis estadístico, y cualitativas, centradas en la comprensión profunda de fenómenos complejos en contextos reales.

El enfoque cuantitativo permitió estructurar y validar un modelo de diagnóstico fonético con base en datos acústicos medibles, mientras que el enfoque cualitativo brindó los elementos necesarios para interpretar el fenómeno desde una perspectiva pedagógica y fonética.

Asimismo, se trata de una investigación aplicada, ya que busca resolver un problema práctico —la identificación automatizada de errores vocálicos en aprendices de francés— mediante la implementación de un sistema funcional, más allá del interés puramente teórico. Esta orientación metodológica responde a la necesidad de soluciones tecnológicas contextualizadas en entornos reales de aprendizaje y permite una mayor transferencia de conocimiento hacia la práctica docente.

3.2 Alcance

Sampieri [41] también menciona que los estudios correlacionales buscan identificar el grado de asociación entre dos o más variables. Dado que a través del entrenamiento de modelos supervisados y la observación de métricas de desempeño por clase se exploran correlaciones implícitas entre variables acústicas y tipos de error, el alcance de la investigación aquí presentada se determina como descriptivo-correlacional.

El enfoque descriptivo se aplicó en la caracterización detallada de los fenómenos acústicos asociados a la pronunciación de vocales francesas en aprendices hispanohablantes, mientras que el enfoque correlacional permite examinar posibles asociaciones entre variables independientes —como los valores de F1 y F2, el fonema objetivo y el género del hablante— y la variable dependiente, correspondiente a la clase de error fonético (C, C+, C–).

Aunque no se establece una relación causal directa, el análisis de estas correlaciones proporciona una base empírica para futuras investigaciones experimentales más profundas y sugiere patrones relevantes que podrían ser útiles tanto en el diseño de herramientas pedagógicas como en la adaptación de modelos de inteligencia artificial al contexto del aprendizaje de lenguas extranjeras.

3.3 Etapas del proyecto

La estructura general del trabajo se organizó tomando como referencia el marco CRISP-DM (Cross Industry Standard Process for Data Mining) [42], adaptado para integrar las particularidades de un estudio aplicado al campo de la fonética y del aprendizaje automático en un contexto educativo. CRISP-DM propone una secuencia de seis fases iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. En la Figura 3.1, se muestran las fases del modelo de referencia CRISP-DM.

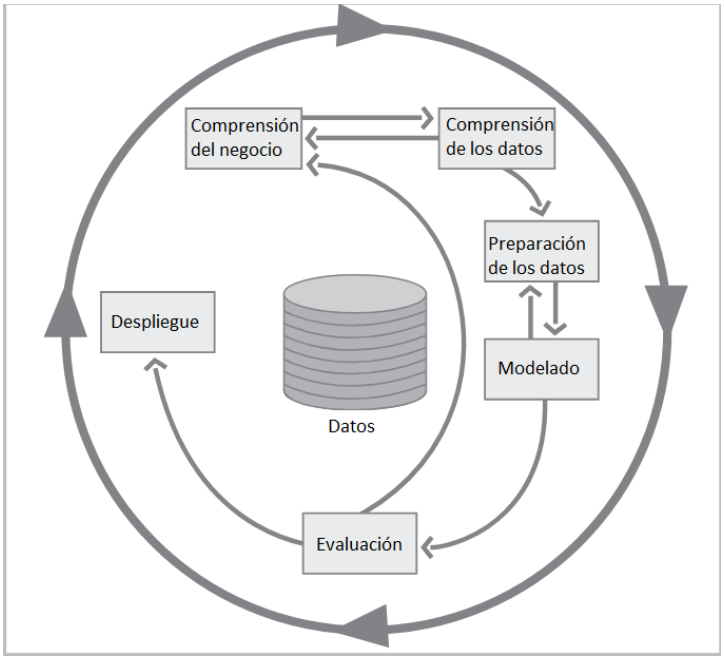


Figura 3.1. Fases del modelo de referencia CRISP-DM.

Fuente: Traducido de [43].

Sin embargo, este enfoque —diseñado principalmente para proyectos industriales centrados en minería de datos— no contempla explícitamente etapas fundamentales en entornos educativos o investigaciones aplicadas con personas, tales como la recolección empírica con humanos, la interacción pedagógica o el diseño y validación de aplicaciones.

Por el contrario, el presente estudio incluyó etapas adicionales que permiten capturar dichas dimensiones: el diseño y ejecución de un estudio piloto exploratorio, el desarrollo iterativo de una aplicación móvil y la incorporación de cuestionarios contextualizados, entre otras. En contraste, CRISP-DM incluye la fase de "comprensión del negocio", que fue reinterpretada aquí como "comprensión del problema" en el contexto de la adquisición del francés como lengua extranjera.

Esta adaptación del marco metodológico permitió enriquecerlo y adecuarlo a un entorno multidisciplinario, integrando componentes de la lingüística aplicada, la ingeniería de software y el aprendizaje automático, dando lugar al diagrama metodológico que se muestra en la Figura 3.2.

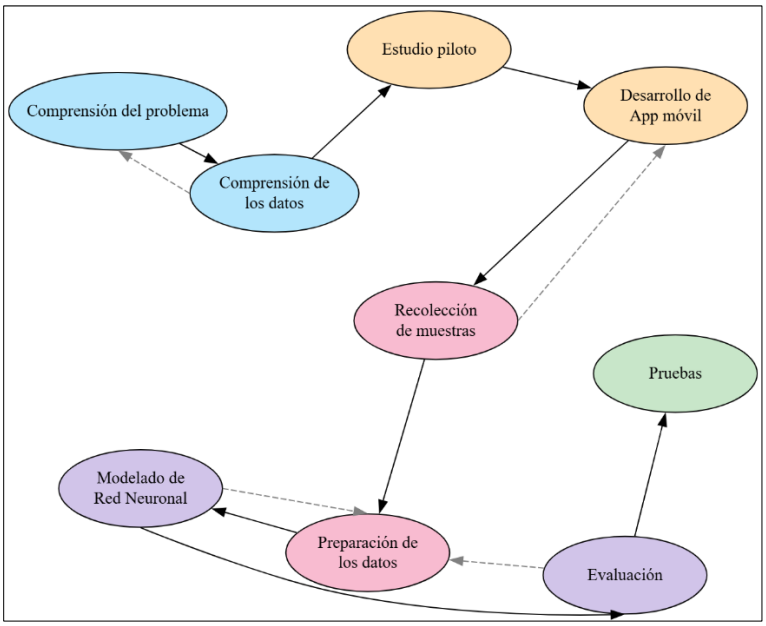


Figura 3.2. Diagrama metodológico general del proyecto.

Fuente: Elaboración propia

En conjunto, la estrategia metodológica de esta investigación combinó enfoques cualitativos y cuantitativos. El enfoque cualitativo fue esencial para la comprensión inicial del problema y el diseño de la solución, mientras que el enfoque cuantitativo permitió validar la precisión y robustez del modelo. Esta integración metodológica, conocida como enfoque mixto, se caracteriza por la complementariedad entre la exploración de significados y la medición objetiva [4]. Fue clave para alcanzar los objetivos del estudio y asegurar tanto la pertinencia pedagógica como la solidez técnica del sistema propuesto.

A continuación, se expande un poco más a detalle cada una de las etapas mostradas en el diagrama metodológico.

3.3.1 Comprensión del problema

En la fase de comprensión del problema, se optó por un enfoque cualitativo, entendido como aquel que busca comprender fenómenos desde la perspectiva de los participantes, explorando sus significados y contextos [4]. Esta elección metodológica permitió captar la complejidad del fenómeno de la pronunciación de fonemas vocálicos en aprendices hispanohablantes de francés. Se implementaron tres técnicas complementarias: la revisión documental sistemática, entrevistas semiestructuradas a especialistas, y observación participante mediante la inscripción a un curso de fonética. La revisión documental sistemática permite identificar y sintetizar el conocimiento existente sobre un fenómeno específico, asegurando rigurosidad y exhaustividad en la identificación de fuentes relevantes [4]. Las entrevistas semiestructuradas, por su parte, proporcionan flexibilidad para explorar experiencias y conocimientos de los informantes, siendo especialmente útiles cuando se busca

profundidad sin imponer estructuras rígidas. Finalmente, la observación participante en un curso especializado brindó una comprensión empírica del proceso de enseñanza de la pronunciación, contextualizando el problema de investigación desde una perspectiva práctica.

3.3.2 Comprensión de los datos

La fase de comprensión de los datos siguió un enfoque exploratorio, que se caracteriza por analizar fenómenos poco estudiados o en los que se requiere familiarización con la información disponible [4]. Este tipo de enfoque es común en estudios iniciales que buscan identificar patrones, tendencias o relaciones sin formular hipótesis definitivas. Se utilizaron herramientas especializadas como Praat y Montreal Forced Aligner (MFA), lo cual permitió analizar distribuciones espectrales, relaciones entre fonemas y formantes, y evaluar la calidad de las alineaciones fonéticas automáticas. Estas herramientas técnicas facilitaron el análisis acústico y fonético necesario para establecer criterios robustos de procesamiento posterior.

3.3.3 Estudio piloto

En el estudio piloto se adoptó un diseño no experimental de tipo exploratorio, es decir, un diseño en el que no se manipulan variables independientes ni se establecen grupos de control, sino que se observa el fenómeno en condiciones naturales [41]. Este enfoque fue adecuado dado que el objetivo principal era evaluar la viabilidad técnica del pipeline de segmentación y análisis acústico, así como afinar los instrumentos y procedimientos para la recolección futura. El muestreo por conveniencia, técnica que selecciona participantes disponibles y accesibles para el investigador, aunque no representativos del total de la población, se consideró adecuado para esta fase exploratoria, donde se privilegia la prueba de instrumentos sobre la generalización estadística [41].

3.3.4 Desarrollo de la aplicación móvil

Para el desarrollo de la aplicación móvil, se empleó una metodología de desarrollo iterativa e incremental, la cual consiste en construir el sistema en ciclos breves, donde se entregan versiones parciales pero funcionales del producto y se recibe retroalimentación temprana para realizar mejoras sucesivas [44]. A diferencia de metodologías rígidas como “cascada”, esta estrategia permite adaptarse a cambios en los requerimientos o descubrimientos durante el desarrollo. Si bien no se implementó un marco completo como SCRUM, se siguieron principios de iteración y validación continua con usuarios reales en cada versión.

3.3.5 Recolección de muestras

Durante la recolección de muestras, se aplicó un muestreo intencional, técnica que implica seleccionar a los participantes con base en criterios específicos preestablecidos [41]. En este caso, se eligieron estudiantes de nivel A1 en francés, pues representan el perfil más susceptible a errores en la pronunciación vocálica. El muestreo intencional se utilizó en tres ciclos consecutivos, cada uno de ellos ajustando funcionalidades y recopilando retroalimentación que enriqueció la siguiente iteración. Se emplearon cuestionarios de metadatos embebidos en la aplicación para caracterizar adecuadamente a los participantes y contextualizar los resultados del modelo.

3.3.6 Preparación de los datos

La preparación de los datos combinó alineación forzada —una técnica automática que segmenta el audio a nivel fonémico utilizando modelos acústicos entrenados [40]— con métodos estadísticos de normalización, particularmente la transformación Z-score. Esta última permite

estandarizar los valores de los formantes F1 y F2 con base en su media y desviación estándar, facilitando la comparación entre diferentes hablantes y reduciendo el sesgo derivado de diferencias fisiológicas o de entorno [35].

3.3.7 Modelado de Red Neuronal

El modelado de los datos siguió un enfoque cuantitativo de tipo experimental. Este tipo de enfoque busca establecer relaciones medibles entre variables mediante la manipulación controlada y el uso de técnicas estadísticas [4]. En este caso, se entrenó un modelo supervisado de red neuronal GRU (Gated Recurrent Unit), el cual recibió como entradas secuencias acústicas normalizadas junto con el género del hablante. Se utilizó una partición del dataset en 80% para entrenamiento y 20% para evaluación, balanceando también un subconjunto interno para validación. La estrategia de entrenamiento supervisado se basa en la disponibilidad de etiquetas correctas, permitiendo al modelo aprender a predecir dichas etiquetas a partir de ejemplos anotados [37].

3.3.8 Evaluación

La evaluación del modelo se realizó con base en métricas estándar en problemas de clasificación multiclase: precisión, F1-score macro, recall y matriz de confusión. Estas métricas permiten cuantificar la calidad del modelo no solo a nivel global, sino en la capacidad de identificar correctamente cada clase individual [37]. La precisión indica el porcentaje de aciertos sobre las predicciones positivas, el F1-score balancea precisión y exhaustividad, mientras que la matriz de confusión permite visualizar errores específicos de clasificación.

3.3.9 Pruebas

La última etapa consistió en pruebas técnicas de integración y validación del sistema en condiciones reales de uso. Estas pruebas permitieron verificar la interoperabilidad de los componentes – tanto la aplicación móvil, el *backend*, como la infraestructura - la protección de datos sensibles, y la estabilidad general del sistema, asegurando así su preparación para futuras pruebas de usuario.

En este capítulo se detalló la metodología seguida para el desarrollo del modelo de detección y clasificación y notificación de errores vocálicos en aprendices hispanohablantes de francés. A partir del marco de referencia, se abordaron de forma estructurada las etapas de comprensión del problema y de los datos, así como la validación temprana mediante un estudio piloto.

CAPÍTULO 4: Desarrollo

Este capítulo describe en detalle cómo se llevó a cabo el desarrollo del proyecto, etapa por etapa, dando seguimiento a la estructura metodológica establecida en el capítulo anterior. El propósito de este capítulo es explicar cómo se implementó técnicamente cada fase, desde la definición del problema hasta las pruebas del sistema final. La organización de las secciones permite dar cuenta de las decisiones técnicas adoptadas en cada fase y su alineación con los objetivos generales del proyecto.

4.1 Comprensión del problema

En la fase de comprensión del problema, se realizó un análisis del fenómeno fonético, así como de los errores típicos en la pronunciación de vocales del francés, por parte de hablantes hispanos, lo que permitió enfocar la investigación en los fonemas que presentan mayores desafíos. Como parte de esta fase, también se llevaron a cabo entrevistas con una docente de francés, quien es originaria de Francia, y fue profesora en la Alianza Francesa de México. Las entrevistas se enfocaron en revisar los métodos de corrección de pronunciación empleados por el profesorado de dicha institución.

Además, también se revisaron aspectos teóricos implicados en la identificación automatizada de los fonemas y, se tuvieron también algunas sesiones con profesores de nivel doctorado, especialistas en redes neuronales artificiales, procedentes de distintas unidades académicas del IPN, con quienes se discutió la propuesta inicial de la problemática para refinarla hasta obtener el planteamiento del problema presentado en el Introducción.

4.2 Comprensión de los datos

En la etapa de comprensión de los datos, se examinó el *dataset* de *Mozilla Common Voice* [45], con muestras de hablantes nativos (L1) y se hicieron pruebas de concepto con la herramienta Praat sobre estos datos, enfocándose en la distribución de los formantes vocales y su correspondencia con patrones perceptivos conocidos.

Luego, se hicieron pruebas de concepto para la extracción de marcas temporales, para lo cual, primero se utilizó la implementación del algoritmo de alineación forzada de la librería TorchAudio de PyTorch [46]. Sin embargo, aunque esta herramienta soporta múltiples lenguas – incluido el francés – y las pruebas con ella resultaron prometedoras, se decidió no utilizarla por dos razones principales; requiere la romanización manual de las transcripciones de entrada para idiomas distintos al inglés (aumentando así la complejidad del proceso de etiquetado), y no soporta el formato TextGrid como formato de salida, el cual es requerido por Praat. Es por esto por lo que se eligió la herramienta Montreal Forced Aligner (MFA) sobre TorchAudio, dada su capacidad de romanizar automáticamente la transcripción del audio y, aún más importante, es capaz de producir la anotación del audio en formato TextGrid; listo para ser utilizado por Praat directamente.

Durante la primera iteración de esta fase, se propuso el flujo ilustrado en la Figura 4.1 para la anotación de los datos.

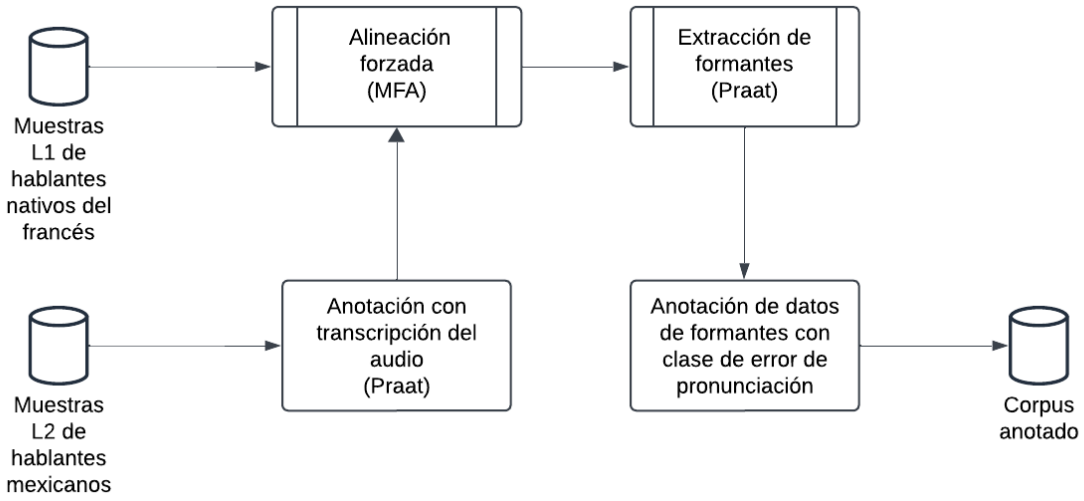


Figura 4.1. Primera versión del flujo de datos propuesto para etiquetado automático.

Fuente: Elaboración propia

No obstante, aunque MFA permitió segmentar automáticamente las grabaciones de voz y asignar fonemas a cada segmento, los resultados no fueron adecuados para usarse directamente como etiquetas en el modelo de clasificación. Al revisar los datos etiquetados por MFA, se observó que la mayoría de las muestras eran clasificadas como el fonema objetivo, sin reflejar los errores comunes que se esperaban encontrar en la pronunciación de estudiantes. Esto ocurre porque MFA no evalúa cómo suena realmente la vocal, sino que simplemente alinea el audio con la transcripción escrita que se le proporciona. Como resultado, tiende a "asumir" que el hablante dijo correctamente la vocal esperada, aunque no sea así. Esta limitación impidió capturar la variabilidad real en la producción de los participantes, y, por lo tanto, se decidió aplicar un esquema de etiquetado alternativo más sensible a las características acústicas de cada muestra.

4.3 Estudio piloto y recolección de muestras

En la primera iteración del estudio, se recopiló un conjunto de grabaciones de voz mediante procedimientos manuales, sin el uso de una aplicación móvil, utilizando un micrófono de condensador para estudio "Samson C01" y una interfaz de audio "Focusrite Scarlet 2i2". La grabación se realizó a una frecuencia de muestreo de 44100Hz. La muestra de participantes estuvo conformada por 3 mujeres y 2 hombres, quienes eran hablantes hispanos con niveles de competencia en francés entre A1 y A2, según el Marco Común Europeo de Referencia para las Lenguas (MCER).

El objetivo principal de esta etapa fue validar la viabilidad técnica de las herramientas de alineación y extracción de características acústicas. Para esto, las grabaciones se remuestrearon a una frecuencia de 16000Hz para poder ser procesadas con la herramienta Montreal Forced Aligner [40], y así conseguir una segmentación fonética automática. Así mismo, se usó Praat [24] para la extracción de los formantes vocales.

A cada participante, se le solicitó grabar una lista de palabras y frases - en total 20 para cada fonema - en 3 intentos, diseñada para incluir ocurrencias de los fonemas objetivo [y], [œ] y [o], reconocidos por su complejidad articulatoria en aprendices hispanohablantes. Esta lista fue validada

por docentes de francés con experiencia en la enseñanza de la fonética. En total, se recopilaron 1800 muestras de audio en esta etapa. En la Tabla 4.1 se listan las frases y palabras empleadas.

Tabla 4.1. Lista de frases y palabras recopiladas durante la 1a iteración.

Fonema	Palabras		Frases	
[o]	mot	côte	Je porte une rose.	Il y a une croix sur la carte.
	rose	hôtel	Le robot est tombé.	Son nom est facile à retenir.
	autre	gros	Il fait très chaud aujourd'hui.	Ce pot est en verre.
	bonjour	votre	Bonjour, comment ça va?	C'est votre choix.
	chaud	trop	Il est le maître du corps et de l'esprit.	Il a un corps en pleine forme.
	porte	propre	Cette côte est magnifique.	Je loue une chambre d'hôtel.
	somme	pot	J'ai oublié mon nom.	Cette croix est très ancienne.
	tomber	croix	Nous avons réservé un hôtel.	Le mot est difficile à prononcer.
	corps	tôt	Ce gâteau est trop gros.	Son robot est cassé.
	robot		Il a loué une voiture.	Le soleil est chaud en été.
[y]	brume	lune	Son but est de gagner.	La lune est belle ce soir.
	inutile	fumer	L'usine est fermée aujourd'hui.	Il a fumé dans la rue.
	jupe	cru	Ce livre est inutile.	J'ai vu un mur blanc.
	dur	mur	Le futur est imprévisible.	Elle a hurlé très fort.
	subir	sur	Elle a utilisé une ruse.	Une lune brillante illumine la nuit.
	jus	hurlé	Le soleil brille sur la lune.	Elle a mis sa jupe rouge.
	futur	une	Ils ont subi de lourdes pertes.	Le mur est trop dur.
	usine	rue	J'ai vu une jupe bleue dans la vitrine.	Ce jus est frais.
	ruse	brûler	Il a écrit sur le mur.	Elle a brûlé les papiers.
	but	vu	Elle boit un jus d'orange.	Il a vu une brume épaisse.
[œ]	œil	sœur	Les fleurs sont magnifiques.	Ma sœur a un grand cœur.
	seule	cœur	Il a fait un vœu sous la lune.	J'ai peur du noir.
	œuf	neuf	Il a peur des hauteurs.	Un bœuf court dans le champ.
	meuble	œuvre	Le fauteuil est très confortable.	Il a neuf ans aujourd'hui.
	meurtri	meurtre	Elle a des meubles anciens.	Elle a heurté la porte.
	nœud	peur	L'œuvre est exposée au musée.	Il a brisé un œuf.
	vœux	bœuf	Le vendeur a souri.	Le meurtre a eu lieu hier.
	fleur	beurre	Mon frère est veuf.	J'ai acheté du beurre.
	fauteuil	veuf	Elle a reçu de nombreux vœux.	Le vendeur a donné un bon prix.
	vendeur	heurter	J'ai peur de marcher seule.	Elle a noué un nœud serré.

Fuente: Elaboración propia

La muestra fue por conveniencia debido a que se hizo la invitación a estudiantes de francés, y los 5 participantes fueron los que de manera voluntaria quisieron participar. Se les explicó en qué consistiría el experimento y estuvieron de acuerdo.

Como parte de la recolección, se aplicó también un breve cuestionario de metadatos a cada participante. Este cuestionario incluyó información relativa a la edad, sexo, lugar de origen, nivel

estimado de francés y tipo de dispositivo utilizado para grabar. En particular, el campo correspondiente al género fue utilizado posteriormente para ajustar los límites superiores de frecuencia durante la normalización de formantes, diferenciando entre 5000 Hz para hombres y 5500 Hz para mujeres [47]. En el Anexo B: Cuestionario para grabación de audio, se puede encontrar el cuestionario completo.

Con la aplicación ya funcional, se realizó una nueva ronda de recolección. Se eliminaron las frases largas y se utilizaron palabras aisladas para facilitar el alineamiento. Se identificaron mejoras necesarias en la interfaz y en la estabilidad del envío de muestras.

En la Tabla 4.2 se muestran las palabras seleccionadas, las cuales se determinaron con base en la lista de las 300 palabras más utilizadas del francés, que contuvieran a los fonemas objetivo [y], [œ] y [o], y que incluyeran diversidad de contextos de pronunciación, derivado de las consonantes que los preceden.

Tabla 4.2. Lista de palabras revisada para la etapa de recolección de muestras.

[y]	[œ]	[o]
tu	cœur	eau
lune	sœur	haut
vue	œuf	beau
plus	chœur	mot
rue	fleur	tôt
une	peur	trop
jus	leur	moto
du	heure	faux
cru	beurre	peau
pu	douleur	château
but	couleur	cadeau
lu	chaleur	niveau
vu	humeur	tableau
nu	rumeur	vélo
su	lenteur	zéro
mûr	ampleur	héros
brut	hauteur	piano
fus	largeur	photo
jusque	profondeur	radio
musique	ardeur	vidéo

Fuente: Elaboración propia

Como primera iteración en esta fase, se reunieron muestras de 10 alumnos, pertenecientes a un grupo de estudiantes principiantes de francés, por medio de la App móvil, la cual fue instalada en dispositivos Android.

Posteriormente, en una segunda iteración, se colaboró con el CENLEX Unidad Santo Tomás, para recolectar muestras de alumnos de 2 grupos de nivel básico (A1), logrando así la recolección de audios de 22 personas más.

4.4 Modelo de Red Neuronal

Con la formulación de este modelo se exploró la viabilidad de un enfoque de clasificación de errores de pronunciación vocálica en aprendices mexicanos de francés, basado en representaciones fonético-acústicas intermedias (formantes F1 y F2) y redes neuronales recurrentes. A partir de principios fonéticos perceptivos, el sistema clasifica las producciones vocálicas según dos ejes: tensión (relacionado con F1) y claridad tonal (relacionado con F2), los cuales reflejan propiedades articulatorias clave en la producción de vocales del francés.

Este esquema permite identificar desviaciones en la producción del estudiante respecto a los valores prototípicos de hablantes nativos, empleando etiquetas de error interpretables pedagógicamente:

- Para el eje de tensión: T+ (exceso de tensión), T (pronunciación óptima), T– (falta de tensión),
- Para el eje de tono: C+ (demasiado claro), C (óptimo), C– (demasiado oscuro).

Estas etiquetas se asignan mediante un modelo RNN entrenado sobre un corpus propio, compuesto por muestras de pronunciación de estudiantes mexicanos de francés. Para esto, se empleó alineación forzada para segmentar fonemas en el habla continua, extracción de formantes con Praat, y normalización Z-score en función de distribuciones obtenidas de hablantes L1. Con este enfoque, se buscó evaluar si una arquitectura ligera basada en formantes y aprendizaje secuencial es capaz de capturar patrones fonéticos relevantes en contextos con datos limitados, y si puede integrarse, en etapas futuras, a herramientas móviles de retroalimentación automática.

A continuación, se detallan cada una de las etapas seguidas para llegar al modelo final.

4.4.1 Construcción del dataset

El conjunto de datos utilizado para el entrenamiento del modelo se organizó como una colección de muestras individuales, donde cada muestra contiene tres componentes principales: la secuencia acústica (X), la clase objetivo (Y) y un bloque de metadatos asociados al hablante (*metadata*). El campo X está conformado por una serie temporal de vectores de tres dimensiones que representan los valores de F1, F2 y el identificador del fonema objetivo en cada fotograma. Este diseño permite capturar la evolución articulatoria de la vocal a lo largo del tiempo. Por su parte, el campo Y corresponde a la etiqueta de clase tonal asociada a la muestra —correcta (C), clara (C+) u oscura (C−)— definida a partir del análisis estadístico de los formantes mediante distancia Z-score. Finalmente, el bloque *metadata* almacena información declarativa del participante (como género, edad, nivel de exposición al idioma, tipo de dispositivo de grabación, entre otros), de la cual se extrae el género como entrada auxiliar del modelo. Esta estructura permitió construir un *pipeline* flexible que combina señales acústicas y metadatos lingüísticos para alimentar el modelo de clasificación. La presenta la estructura de cada entrada del *dataset* producido.

Tabla 4.3. Estructura del conjunto de datos (dataset).

Campo	Tipo de dato	Descripción
X	Lista de listas [F1, F2, ID_fonema]	Secuencia de vectores acústicos correspondientes a una muestra fonémica. Cada sublista contiene los valores de F1 y F2, normalizados, y el ID del fonema objetivo.
Y	Entero (int)	Etiqueta de clase asociada al eje tonal de la vocal. Valores posibles: 0 = C, 1 = C+, 2 = C−.
metadata	Objeto (dict)	Información adicional del participante que produjo la muestra, utilizada en análisis complementarios. Contiene campos como edad, género, lengua materna, nivel del alumno, tipo de ambiente de grabación, entre otros.

Fuente: Elaboración propia

4.4.2 Etiquetado de datos

Durante el desarrollo del modelo, se evaluaron diferentes esquemas de etiquetado – basado solo en MFA, por distancia euclidiana, y mediante Z-score - concluyendo que el etiquetado basado en distancias Z-score en el plano F1-F2 ofrecía una mejor agrupación de clases, así como un rendimiento más estable en las métricas de evaluación.

Finalmente, se optó por reducir las clases posibles de ocho a tres, enfocándose exclusivamente en el eje de tono (F2), y clasificando las muestras como pertenecientes a una de tres categorías: C (tono correcto), C+ (tono claro) o C– (tono oscuro). Esta simplificación permitió mejorar la precisión del modelo en condiciones de escasez de datos, manteniendo la relevancia diagnóstica de la retroalimentación fonética (ver la Tabla 4.4).

Tabla 4.4. Clasificación final propuesta para errores de pronunciación.

Clase de error	Formante relacionado	Interpretación
C	F2	Tono correcto
C-	F2	Tono muy claro
C+	F2	Tono muy oscuro

Fuente: Elaboración propia

4.4.3 Modelado de red neuronal

El proceso de modelado consistió en entrenar una red neuronal recurrente (RNN) del tipo GRU (Gated Recurrent Unit) para la tarea de clasificación fonética, tomando como entrada, secuencias de características acústicas extraídas de muestras de voz. Las características empleadas incluyeron los formantes F1 y F2 normalizados mediante Z-score, el fonema objetivo codificado mediante one-hot encoding y una variable binaria correspondiente al género del hablante, utilizada como entrada auxiliar para contextualizar los límites fisiológicos en la producción vocálica.

La red fue optimizada utilizando el algoritmo Adam y entrenada durante 100 épocas, con un tamaño de lote de 32 muestras. El conjunto de datos fue dividido en un 80% para entrenamiento y un 20% para validación. Para mitigar el desbalance entre clases, se aplicó una técnica de sobre-muestreo (oversampling) mediante el uso de WeightedRandomSampler, aunque en iteraciones posteriores se optó por eliminar dicho muestreo al obtener mejores resultados con una distribución no manipulada.

Derivado de lo anterior, se diseñó un modelo de Red Neuronal Recurrente (RNN) del tipo GRU (Gated Recurrent Unit), para procesar secuencias de características acústicas extraídas de fonemas vocálicos. Cada muestra de entrada consiste en una secuencia de vectores que representan los formantes F1 y F2, junto con el identificador del fonema objetivo.

Esta secuencia es procesada por una capa GRU que resume la información temporal en un vector de estado oculto, el cual se concatena con una variable adicional que codifica el género del hablante. Esta combinación se introduce en un perceptrón multicapa compuesto por dos capas lineales con una función de activación ReLU intermedia.

La salida del modelo son tres valores (logits), que representan las puntuaciones no normalizadas para cada clase de error tonal: pronunciación correcta (C), tono demasiado claro (C+) y tono demasiado oscuro (C−). La clase final se determina aplicando una operación de máxima probabilidad (argmax) sobre estos logits. La Figura 4.2 muestra las capas de la Red Neuronal propuesta.

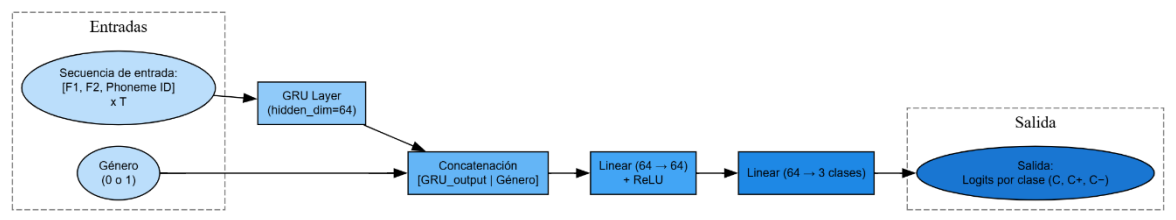


Figura 4.2. Modelo de Red Neuronal Recurrente propuesto.

Fuente: Elaboración propia

4.5 Evaluación

La evaluación del modelo se llevó a cabo mediante un conjunto de métricas comúnmente utilizadas en tareas de clasificación multiclase, como son la exactitud (*accuracy*), la precisión, el recall y la puntuación F1, tanto por clase como en promedio macro. La exactitud indica el porcentaje global de predicciones correctas, mientras que la precisión mide la proporción de verdaderos positivos entre todas las predicciones positivas realizadas.

Por su parte, el recall representa la proporción de verdaderos positivos sobre el total de elementos positivos reales. La métrica F1 proporciona un balance entre precisión y recall, siendo especialmente útil cuando existen desbalances en la distribución de clases, como es el caso de este estudio. El promedio macro permite observar el desempeño del modelo tratando todas las clases por igual, sin importar su frecuencia en el conjunto de datos. Finalmente, se utilizó una matriz de confusión para visualizar con mayor claridad los errores de clasificación entre categorías fonéticas adyacentes y detectar patrones sistemáticos de confusión [48]. La matriz de confusión, generada a partir del entrenamiento y validación del modelo final, se muestra en el apartado de resultados.

4.6 Pruebas

Para las pruebas, se desplegó el *backend* del sistema como una REST API utilizando servicios en la nube pública, principalmente Amazon Web Services (AWS) y Google Cloud Platform (GCP), lo cual permitió garantizar escalabilidad y disponibilidad. En esta etapa se realizaron validaciones funcionales y de seguridad, implementando mecanismos como el protocolo HTTPS para proteger las comunicaciones, Firebase para la autenticación de usuarios y el control de acceso basado en roles mediante Identity and Access Management (IAM).

Las pruebas de extremo a extremo verificaron la robustez del flujo completo desde la grabación hasta la clasificación. En la Figura 4.3, se muestran los subprocesos ejecutados en el *backend* que se validaron durante esta fase.

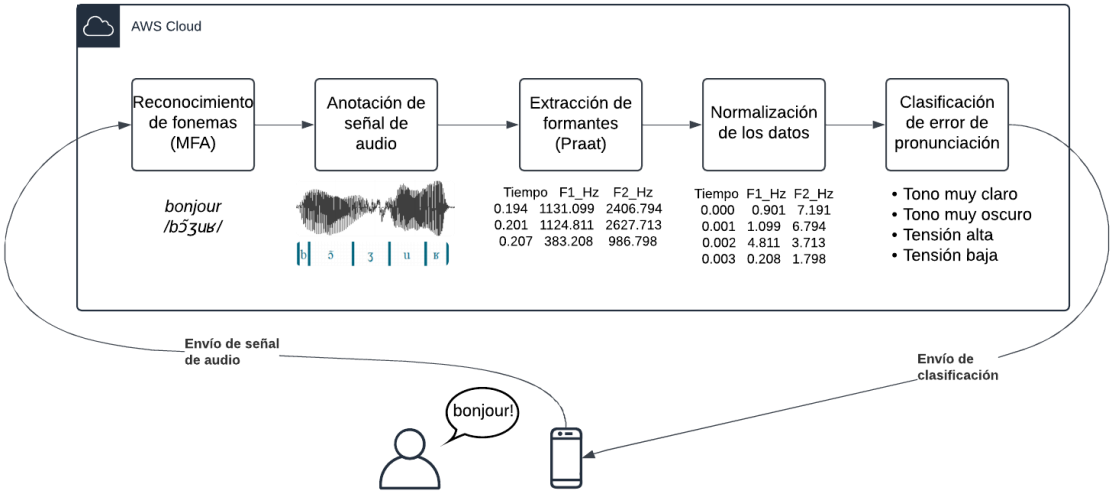


Figura 4.3 Procesos validados durante las pruebas.

Fuente: Elaboración propia.

En este capítulo se mostró como el desarrollo del proyecto integró distintas herramientas, técnicas y metodologías que permitieron materializar un sistema funcional para la detección de errores fonéticos en estudiantes de francés. Cada etapa, desde la exploración inicial hasta las pruebas de integración, fue diseñada para abordar de forma rigurosa los desafíos técnicos y metodológicos del proyecto, asegurando una base sólida para su implementación y posibles aplicaciones futuras.

CAPÍTULO 5: Desarrollo de “PhonessaAI”

Este capítulo describe cómo fue construida la aplicación móvil - a la que se le dio el nombre de “PhonessaAI” - desde la concepción de sus requerimientos hasta su integración con servicios en la nube. A diferencia del capítulo anterior, que se centró en la planeación y diseño metodológico, aquí se detalla el proceso técnico y práctico detrás del desarrollo del sistema que permite recopilar muestras de audio, analizarlas y brindar retroalimentación personalizada sobre la pronunciación de los usuarios.

5.1 Aplicación móvil

Tanto para la recolección de muestras de audio para entrenamiento, como para la presentación de la clasificación entregada por el modelo entrenado, se desarrolló una aplicación móvil a la que se le llamó PhonessaAI. En las secciones subsecuentes, se explican los aspectos más relevantes del desarrollo de esta aplicación.

5.1.1 Requerimientos funcionales

Los requerimientos funcionales de la aplicación móvil se definieron con base en el objetivo principal del sistema: permitir la recolección de muestras de voz por parte de los usuarios y brindar retroalimentación sobre su pronunciación. La aplicación permite a cada usuario registrarse o autenticarse mediante un sistema confiable y rápido, que garantiza la identificación única de cada participante.

Una vez autenticado, el usuario puede grabar su pronunciación palabra por palabra, según la lista configurada desde el *backend*. Tras enviar la muestra al servidor, la aplicación recupera el resultado del diagnóstico emitido por el modelo entrenado y lo presenta de forma comprensible en pantalla. Estas funcionalidades representan el mínimo necesario para validar al modelo presentado y permitir la interacción efectiva entre el usuario final y el sistema de inferencia fonética.

En la Figura 5.1 se muestra el diagrama de casos de uso para el registro y log in de usuarios en la aplicación.

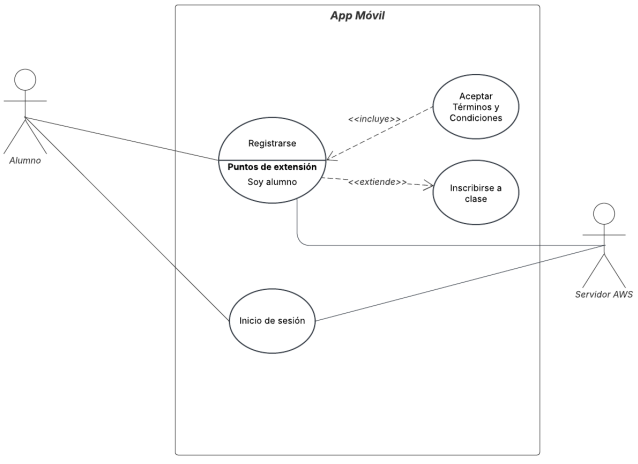


Figura 5.1. Diagrama de casos de uso de inicio de sesión y registro.

Fuente: Elaboración propia.

En la Figura 5.2, se muestra el diagrama de casos de uso para el alumno, incluyendo, principalmente, el registro de muestras para el entrenamiento del modelo y el diagnóstico de pronunciación con un modelo ya entrenado.

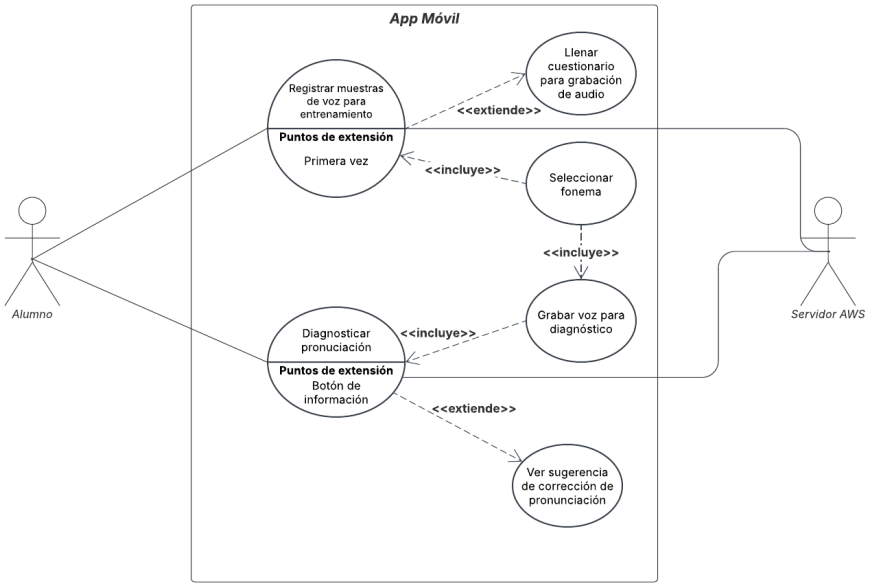


Figura 5.2. Diagrama de casos de uso del alumno.

Fuente: Elaboración propia.

5.1.2 Requerimientos no funcionales

Se consideraron ciertos aspectos clave para asegurar el funcionamiento práctico de la aplicación. Uno de ellos es la disponibilidad de micrófono y bocinas en el dispositivo, condición necesaria para grabar la pronunciación del usuario y reproducir el modelo de pronunciación esperado. Además, se requirió conexión a internet, dado que la inferencia fonética se realizó de forma remota en un servidor.

Desde el punto de vista de compatibilidad, la aplicación se desarrolló en Flutter con la intención de funcionar tanto en Android como en iOS. Sin embargo, por restricciones logísticas y de tiempo relacionadas con el uso de TestFlight, las pruebas se realizaron únicamente en dispositivos Android.

En términos de seguridad, se implementó cifrado de extremo a extremo. La comunicación con el servidor se realizó mediante HTTPS, utilizando un certificado digital emitido por AWS Certificate Manager (ACM), para lo cual se registró un nombre de dominio específico para la aplicación. Asimismo, los archivos de audio almacenados en S3 se encriptaron, y su acceso se controla mediante políticas de identidad y acceso (IAM), así como reglas definidas en Firebase Authentication.

5.1.3 Interfaz

La interfaz de usuario de PhonessaAI fue diseñada para guiar al estudiante en cada una de las etapas del proceso: desde el registro inicial hasta el diagnóstico de su pronunciación. Al iniciar la aplicación, se le solicita un identificador de clase para agrupar sus resultados de diagnóstico. En la Figura 5.3, se muestra este flujo.

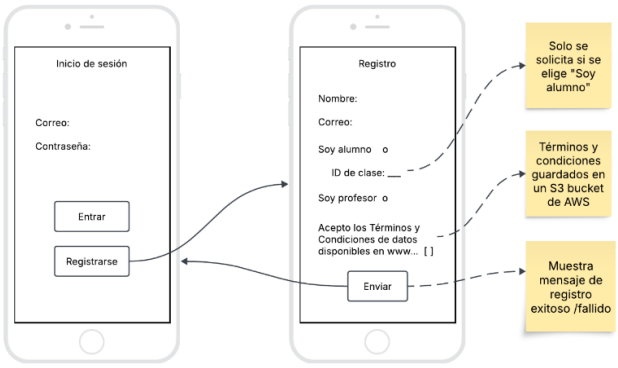


Figura 5.3. Mockups de inicio de sesión y registro.

Fuente: Elaboración propia.

Una vez registrado e iniciado sesión, el estudiante accede al menú principal, donde puede elegir entre dos funcionalidades principales: grabación de muestras de voz para entrenamiento del modelo o diagnóstico de su pronunciación. Al seleccionar la opción de entrenamiento, la aplicación solicita un cuestionario inicial de metadatos (edad, sexo, país de origen, etc.), que se muestra únicamente la primera vez. Posteriormente, el estudiante selecciona el fonema a practicar ([o], [y] u [œ]) y se le presentan palabras asociadas a dicho fonema. El alumno puede reproducir la pronunciación del modelo para escucharla antes de grabar su muestra de voz y se le permite grabar tres intentos por palabra, enviando los archivos al servidor después de cada uno. La Figura 5.4 contiene los *mockups* de estas pantallas.

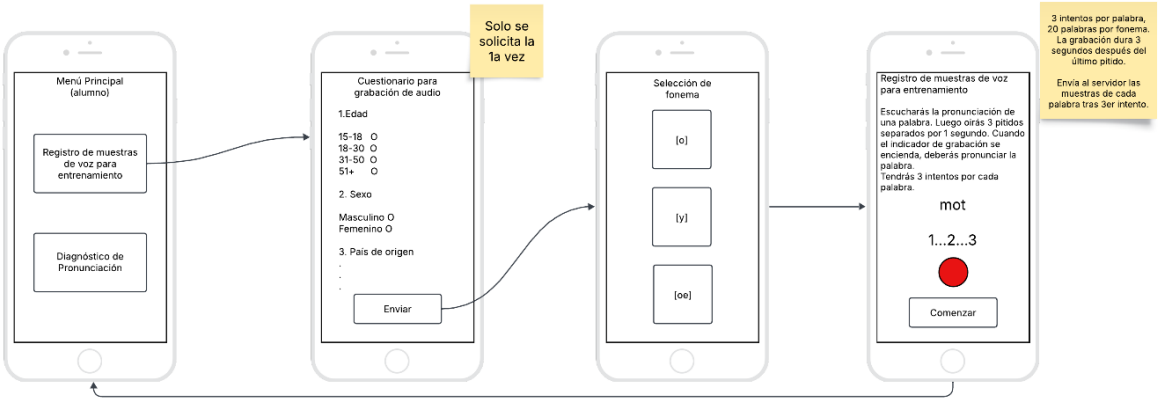


Figura 5.4. Mockups de pantallas del alumno para registro de muestras de voz para entrenamiento.

Fuente: Elaboración propia.

En el módulo de diagnóstico, el estudiante repite un proceso similar: selecciona el fonema, graba una serie de palabras sugeridas y, tras el envío de las grabaciones al servidor, recibe una retroalimentación visual con la clasificación obtenida (correcta, clara o sombría). La interfaz le permite además consultar sugerencias específicas de corrección, incluyendo imágenes de apoyo fonético. El flujo ha sido diseñado para facilitar tanto la recolección de datos de entrenamiento como la entrega de diagnósticos personalizados. En la Figura 5.5, se pueden apreciar los mockups respectivos.

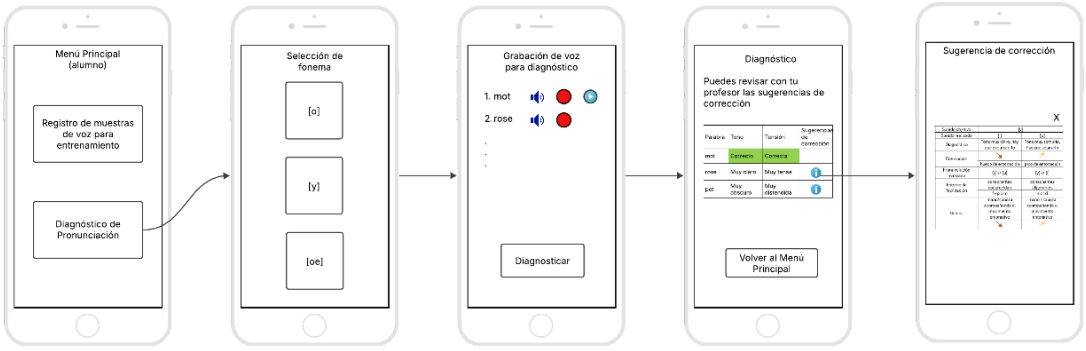


Figura 5.5. Mockups de pantallas del alumno para el diagnóstico de la pronunciación.

Fuente: Elaboración propia.

5.2 API

La API de PhonessaAI fue desarrollada utilizando el *framework* FastAPI, que permite la construcción de servicios web de alto rendimiento con una sintaxis moderna y clara basada en Python. Esta API actúa como el núcleo del sistema, facilitando la comunicación entre la app móvil, la base de datos y los modelos de diagnóstico fonético. Todas las operaciones críticas del sistema —desde la autenticación hasta el análisis acústico— se orquestan a través de sus distintos endpoints.

Uno de los aspectos fundamentales de la API es la integración con Firebase, lo que permite autenticar de forma segura a los usuarios mediante tokens JWT. Esta autenticación protege los *endpoints* sensibles, como el diagnóstico de pronunciación o la gestión de usuarios. La API también está conectada a una base de datos relacional (MySQL), accedida mediante SQLAlchemy, desde la cual se recuperan encuestas, preguntas, respuestas y configuraciones personalizadas.

En términos funcionales, la API permite enviar respuestas a encuestas de diagnóstico lingüístico, consultar si un usuario ha completado una encuesta, y obtener tanto las preguntas como las palabras asociadas a ciertos fonemas. También permite registrar nuevos usuarios y actualizar su información, así como subir grabaciones de audio en diferentes intentos y contextos (entrenamiento o diagnóstico). A partir de estas grabaciones, la API procesa automáticamente los archivos para extraer los formantes F1 y F2 y predecir el tipo de error fonético mediante un modelo de red neuronal entrenado. Finalmente, también ofrece sugerencias de retroalimentación personalizadas en función del fonema y el tipo de error detectado.

Además, se implementaron medidas de robustez como validación de campos, control de errores HTTP, manejo de excepciones en consultas a base de datos, y almacenamiento de logs para

trazabilidad del sistema. La API también permite realizar pruebas de salud del sistema (endpoint /health) que verifican tanto la conectividad con la base de datos como el estado del servicio web. Esta arquitectura modular y extensible garantiza que la API pueda escalar y adaptarse a nuevos requerimientos sin comprometer la seguridad ni la eficiencia.

En la Tabla 5.1, se presenta un resumen con la firma de los principales *endpoints* implementados.

Tabla 5.1 Lista de endpoints de la API de PhonessaAI.

Método	**Endpoint**	**Descripción**
GET	/health	Verifica la conectividad del sistema
POST	/submit-survey	Guarda las respuestas de una encuesta
GET	/survey/{survey_id}/completed	Verifica si un usuario completó la encuesta
GET	/surveys/{survey_id}/questions	Obtiene preguntas de una encuesta
GET	/words/by-phoneme/{symbol}	Devuelve palabras asociadas a un fonema
GET	/phonemes	Lista los fonemas disponibles
POST	/upload-audio	Sube un archivo de audio del usuario
POST	/user	Registra un nuevo usuario
PATCH	/user/{id}	Actualiza la información de un usuario
POST	/diagnose	Realiza el diagnóstico fonético con red neuronal
GET	/suggestion/{phoneme}/{error_class}	Ofrece una sugerencia de corrección según error detectado

Fuente: Elaboración propia

5.3 Infraestructura en Nube

El sistema PhonessaAI fue desplegado en la nube utilizando una arquitectura híbrida que combina servicios de Amazon Web Services (AWS) y Firebase de Google Cloud Platform (GCP). Esta decisión se tomó para aprovechar la escalabilidad, seguridad y modularidad que ofrecen ambas plataformas en sus respectivos dominios.

En la Figura 5.6, se presenta una vista de alto nivel de los elementos que componen esta arquitectura.

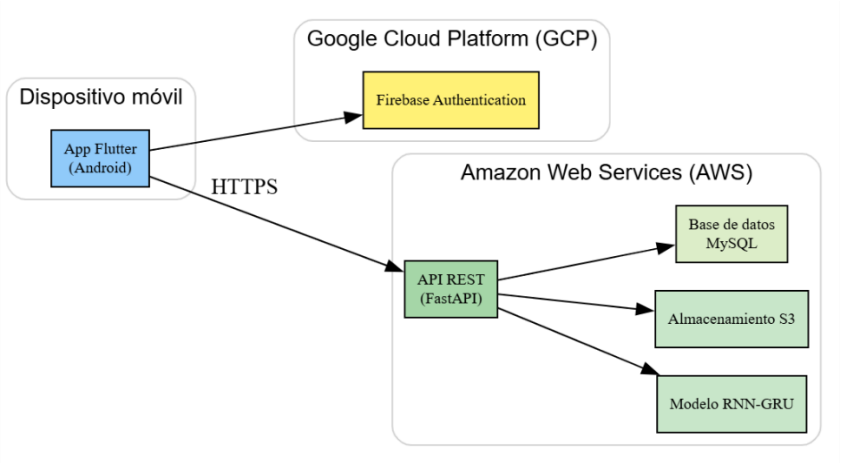


Figura 5.6. Arquitectura de alto nivel.

Fuente: Elaboración propia

En el entorno de AWS, se utilizó Terraform como herramienta de infraestructura como código (IaC) para automatizar el aprovisionamiento de los recursos necesarios. El backend, desarrollado con FastAPI [49], se desplegó en una instancia de EC2 protegida mediante un balanceador de carga de aplicación (ALB) que gestiona el tráfico HTTPS. Para habilitar este canal seguro, se registró el dominio phonessa-ai.com y se generó un certificado TLS a través de AWS Certificate Manager (ACM), el cual fue vinculado al ALB.

Las grabaciones de audio enviadas por los usuarios se almacenan en Amazon S3, en *buckets* protegidos con políticas de acceso específicas y cifrado automático del contenido, asegurando así la confidencialidad e integridad de los datos sensibles. Adicionalmente, se utilizó un contenedor de MySQL alojado en la misma instancia EC2 para persistir la información relativa a usuarios, encuestas, respuestas y metadatos asociados al proceso de diagnóstico.

Complementariamente, Firebase se empleó exclusivamente para la autenticación de usuarios mediante su módulo de Authentication, permitiendo emitir y verificar tokens JWT. Esta integración garantiza que solo usuarios autenticados puedan acceder a los recursos protegidos de la API. Asimismo, se hizo uso de Firebase Firestore, una base de datos NoSQL en la que se almacenaron detalles técnicos de los dispositivos móviles usados, como el sistema operativo, versión y el ID de Firebase correspondiente. Para asegurar estos datos, se configuraron reglas de seguridad que restringen el acceso a los documentos únicamente a los usuarios autenticados.

Esta configuración híbrida, aunque sencilla en su estructura, permitió maximizar la robustez, escalabilidad y seguridad del sistema, habilitando la integración eficiente entre la aplicación móvil, la API y el almacenamiento distribuido de datos. De esta manera, se logró un entorno de despliegue confiable y adaptable para las distintas etapas del proyecto.

En este capítulo, se explicó la integración entre plataformas que ofreció un entorno robusto y eficiente para desplegar todos los componentes necesarios de PhonessaAI, desde la aplicación móvil hasta el procesamiento en servidor. La infraestructura empleada permitió mantener la seguridad de los datos, facilitar el mantenimiento del sistema y escalar su funcionamiento según las necesidades del proyecto.

CAPÍTULO 6: Resultados y análisis

Este capítulo presenta los resultados obtenidos tras el entrenamiento y validación del modelo de clasificación de errores vocálicos, así como el análisis de las métricas derivadas de las distintas configuraciones evaluadas durante el proceso experimental. Se destacan las decisiones metodológicas que impactaron de forma significativa en el desempeño del modelo y se justifican los ajustes aplicados en función de la evidencia empírica obtenida.

6.1 Evaluación del etiquetado automático

6.1.1 Etiquetado con alineación forzada (MFA)

Inicialmente, las etiquetas de pronunciación se derivaron directamente del alineamiento fonético proporcionado por Montreal Forced Aligner. Sin embargo, se observó una distribución altamente desbalanceada, con más del 90% de las muestras etiquetadas en una sola clase (C_T), lo cual sugiere una baja capacidad de discriminación del modelo fonético alineado. El modelo entrenado sobre este dataset mostró una precisión sospechosamente alta desde la segunda época, lo que apuntaba a sobreajuste debido al sesgo de clase, tal y como se observa en la Figura 6.1. Esta falta de variabilidad aparece en la Figura 6.2, donde puede notarse una concentración de muestras dominante en una única clase.

Figura 6.1. Precisión en entrenamiento y validación sobre conjunto de entrenamiento anotado con MFA.

Fuente: Elaboración propia.

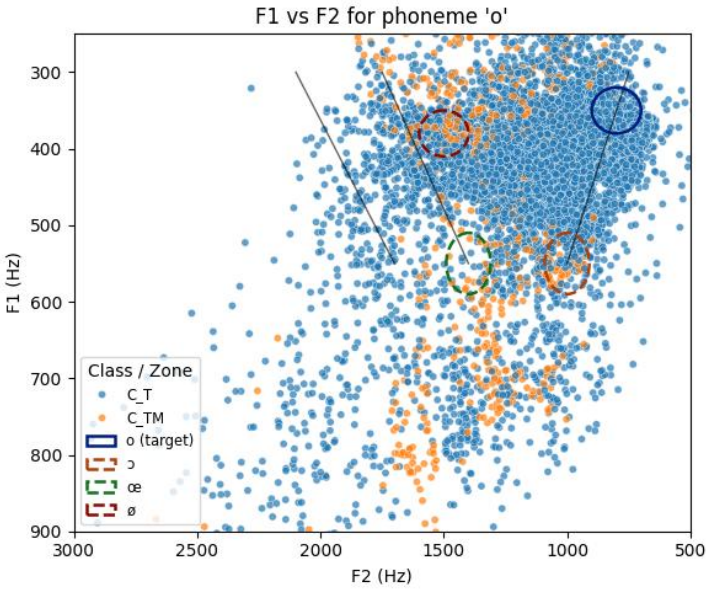


Figura 6.2. Distribución de clases generadas por alineación MFA

Fuente: Elaboración propia

6.1.2 Etiquetado acústico por distancia euclidiana

Se desarrolló un proceso alternativo de etiquetado automático basado en la distancia euclidiana de los formantes F1 y F2 a los centroides de cada fonema, definidos a partir de hablantes nativos. Esta estrategia generó una distribución más diversa de clases y mostró una mayor coherencia fonética al observarse una mejor diferenciación entre fonemas relacionados. La distribución de clases mostrado en la Figura 6.3 confirmaron una distribución más realista y útil para el entrenamiento, donde se aprecia una mayor diversidad y separación entre grupos fonéticos.

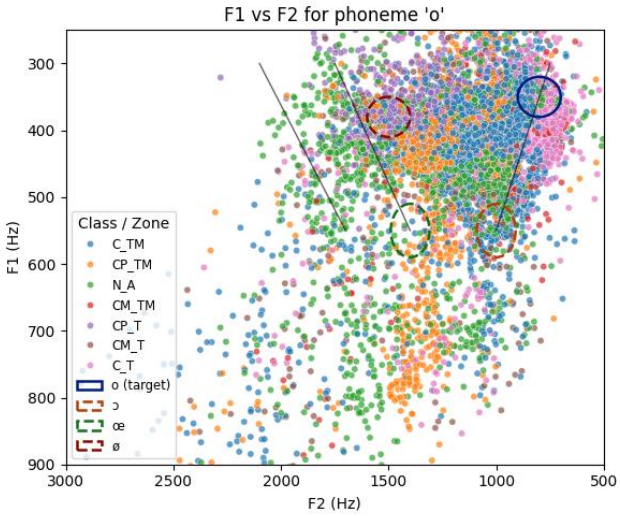


Figura 6.3. Distribución de clases generadas por distancia euclidiana.

Fuente: Elaboración propia

6.1.3 Etiquetado fonético con normalización por Z-score

Posteriormente, se aplicó una normalización Z-score de los formantes con respecto al género del hablante (techo de frecuencia: 5000 Hz para hombres, 5500 Hz para mujeres), con el objetivo de ajustar la escala perceptiva de los valores fonéticos. Esta estrategia logró generar agrupaciones más claras en los *scatter plots* de la Figura 6.4, donde se aprecia una mayor coherencia y formación de los *clusters*. Lo anterior permitió definir clases C, C+ y C– con mayor base acústica.

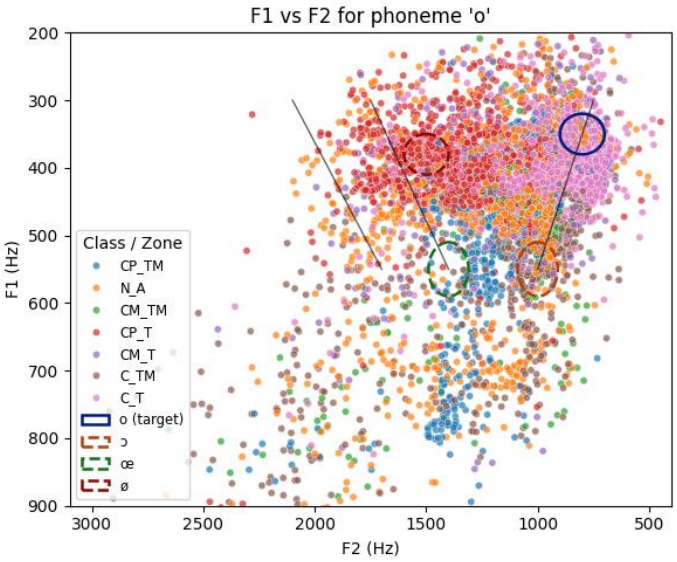


Figura 6.4. Agrupación de clases tras normalización Z-score por género.

Fuente: Elaboración propia

6.2 Estructura y normalización del dataset

El dataset final adoptó la estructura X = [[[F1_t, F2_t, target_t], ...]] y Y = class, para capturar explícitamente la dimensión temporal de cada muestra y facilitar su procesamiento por una red neuronal recurrente. Sin embargo, se detectó una reducción sustancial en el número de muestras viables debido a filtrados estrictos del pipeline, quedando aproximadamente 700 muestras disponibles. Esto condicionó el desempeño general del modelo.

6.3 Estrategias de muestreo

Con el objetivo de mitigar el desbalance de clases, se probaron diferentes esquemas de muestreo durante el entrenamiento. El uso de WeightedRandomSampler no logró mejoras significativas, y generó un comportamiento más inestable. Por el contrario, un muestreo aleatorio estratificado sin pesos mostró mejores resultados en precisión y estabilidad entre épocas.

6.4 Reducción de ruido

Durante las primeras iteraciones del procesamiento acústico se probó la aplicación de técnicas automáticas de reducción de ruido previo a la extracción de formantes. Sin embargo, se observó que los algoritmos de reducción alteraban significativamente los valores de F1 y F2, introduciendo distorsiones que afectaban el alineamiento y generaban errores en la segmentación. Además, el proceso de reducción de ruido provocó una disminución aún mayor en el número de muestras útiles,

lo cual impactó negativamente en la calidad del entrenamiento. Por estas razones, se optó por no aplicar reducción de ruido en el pipeline. En la Tabla 6.1, se muestran los resultados de este entrenamiento, donde se observa una disminución en la precisión general y mayor dispersión de clases.

Tabla 6.1. Métricas de desempeño tras aplicar reducción de ruido.

Clase	Precisión	Recall	F1-score	Soporte
0	0.9000	0.3000	0.4500	30
1	0.0000	0.0000	0.0000	3
2	0.2143	0.8571	0.3429	7
3	0.2593	0.5385	0.3500	13
4	0.0000	0.0000	0.0000	1
5	0.0000	0.0000	0.0000	10
6	0.0000	0.0000	0.0000	4
Accuracy				0.3235
Macro avg	0.1962	0.2422	0.1633	68
Weighted avg	0.4687	0.3235	0.3007	68

Fuente: Elaboración propia

6.5 Reducción del número de clases

El esquema original contemplaba 8 clases combinando los ejes de tono y tensión. Sin embargo, dada la escasez de datos y la alta confusión entre clases relacionadas, se optó por reducir el problema a 3 clases centradas en el eje de tono (C+, C, C–). Esta simplificación incrementó la precisión general del modelo, mejoró la estabilidad del entrenamiento y permitió interpretar los resultados con mayor claridad.

6.6 Métricas de desempeño del modelo final

Con el dataset ampliado y la reducción a 3 clases, los resultados más representativos del último experimento, sin reducción de ruido, fueron:

- Accuracy: 63.49%
- F1-score macro: 0.5108
- Precision macro: 0.5047
- Recall macro: 0.5568

Desempeño por clase:

- Clase C: precision = 0.4706, recall = 0.6400, f1-score = 0.5424
- Clase C+: precision = 0.8696, recall = 0.6667, f1-score = 0.7547
- Clase C–: precision = 0.1739, recall = 0.3636, f1-score = 0.2353

Sin embargo, durante experimentos intermedios, con datasets más reducidos y etiquetados menos refinados, se observaron rendimientos por debajo del 45% de precisión, e incluso caídas hasta el 32% al incorporar reducción de ruido.

En uno de los últimos experimentos con muestreo ponderado, se obtuvo:

- Accuracy: 42.06%
- Macro F1: 0.3995

Y en el experimento con reducción de ruido:

- Accuracy: 32.35%
- Macro F1: 0.1633

Estos resultados confirman que el etiquetado por Z-score y la reducción a 3 clases ofrecieron el mejor equilibrio entre rendimiento y estabilidad del modelo. La matriz de confusión del modelo final puede verse en la Figura 6.5, donde se aprecia que la clase C– presenta mayor dificultad de clasificación, y el resumen de métricas se presenta en la Tabla 6.2.

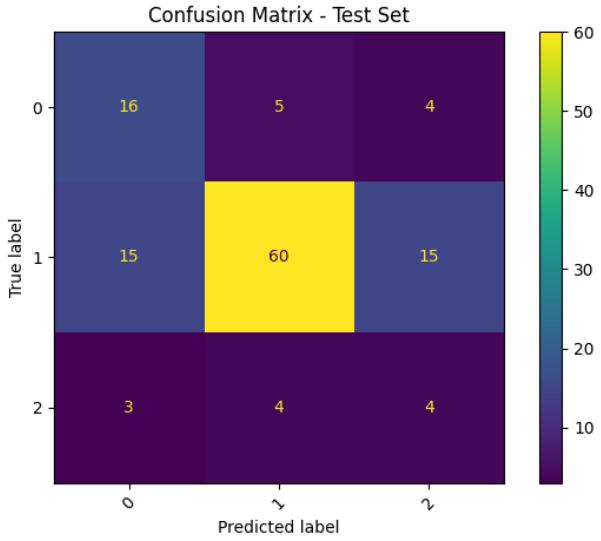


Figura 6.5. Matriz de confusión del modelo final entrenado con clases C, C+ y C–

Fuente: Elaboración propia

Tabla 6.2. Métricas de desempeño del modelo final por clase y valores macro.

Clase	Precisión	Recall	F1-score	Soporte
C	0.4706	0.6400	0.5424	25
C+	0.8696	0.6667	0.7547	90
C–	0.1739	0.3636	0.2353	11
Macro avg	0.5047	0.5568	0.5108	126
Weighted avg	0.7297	0.6349	0.6672	126

Fuente: Elaboración propia

El análisis de errores revela que las muestras clasificadas como C– fueron las más difíciles de distinguir, con alta confusión respecto a la clase C. Esto podría deberse a la variabilidad individual en la producción de vocales como [y] y [œ], y a las limitaciones del entorno de grabación (micrófonos de celular, ruido ambiente). A pesar de esto, se logró detectar agrupamientos acústicos útiles incluso con una arquitectura simple como una GRU y un conjunto de datos reducido.

En resumen, el enfoque propuesto, centrado en formantes, normalización perceptual y reducción controlada de clases, demostró ser viable para detectar patrones de error tonal en un contexto de enseñanza de francés como lengua extranjera, y sienta las bases para mejoras futuras en cobertura de fonemas, volumen de datos y precisión del diagnóstico.

Cabe destacar que la distribución de las clases no es simétrica, ni se espera que lo sea desde un punto de vista fonético. Fonemas como [y] y [œ], al encontrarse en posiciones intermedias dentro del plano F1-F2, permiten errores tanto hacia valores más altos como más bajos en F2, dando lugar a más muestras etiquetadas como Clase 1 (C+) y Clase 2 (C−).

Por el contrario, fonemas como [o], ubicados en la región más posterior del plano vocal, tienen menor espacio para errores hacia F2 más bajos. Esto se traduce en menos casos etiquetables como Clase 2 (C−). Esta asimetría no representa un sesgo del modelo, sino una propiedad intrínseca del espacio articulatorio del francés, y debe ser tenida en cuenta al interpretar métricas de desempeño por clase.

Finalmente, se observa que la Clase 0 (C) —correspondiente a una pronunciación correcta— presenta una frecuencia intermedia entre la Clase 1 (C+) y la Clase 2 (C−). Este comportamiento es esperable y coherente con el contexto de aprendizaje, dado que los participantes son hablantes no nativos que intentan articular correctamente los fonemas objetivo, pero no siempre lo logran. Así, es natural que haya una proporción significativa de aciertos (Clase 0), aunque superada por los errores que tienden a desplazarse hacia un eje más anterior (Clase 1), especialmente en fonemas de posición intermedia como [y] y [œ].

CAPÍTULO 7: Conclusiones

Este trabajo tuvo como objetivo general proponer un modelo de detección y clasificación de errores en la pronunciación de fonemas vocálicos del idioma francés, empleando técnicas de *machine learning*, con el fin de identificar errores comunes en aprendices hispanohablantes mexicanos y brindar retroalimentación automatizada. A lo largo del proyecto se diseñó, implementó y evaluó un modelo computacional que integra un enfoque acústico-formántico con redes neuronales recurrentes (GRU), desplegado en una aplicación móvil funcional. Los resultados obtenidos permiten afirmar que el objetivo general fue alcanzado, mostrando que es posible automatizar el diagnóstico de pronunciación vocálica y generar retroalimentación basada en ejes perceptivos como el tono.

En cuanto al primer objetivo específico, se llevó a cabo una revisión detallada de los algoritmos existentes para la detección y medición de la pronunciación de fonemas vocálicos. Se analizaron enfoques basados en alineación forzada, como el Montreal Forced Aligner (MFA), y métodos acústicos que emplean distancias fonéticas entre formantes vocales. Esta revisión permitió seleccionar un conjunto de herramientas adecuadas para segmentar las muestras de voz y extraer características relevantes, sentando las bases para las siguientes etapas del trabajo.

Respecto al segundo objetivo específico, se identificaron diversos modelos de *machine learning* aplicables a tareas de clasificación de errores fonéticos, entre ellos redes convolucionales, redes recurrentes y modelos end-to-end. Se optó por una red neuronal recurrente con arquitectura GRU, debido a su capacidad para capturar relaciones temporales en secuencias acústicas. Esta elección fue validada mediante experimentación progresiva con diferentes configuraciones y esquemas de etiquetado.

Para el tercer objetivo, se compararon distintas estrategias de medición de pronunciación, incluyendo el alineamiento fonético supervisado, la distancia euclidiana entre formantes y la normalización estadística mediante Z-score. Se concluyó que el etiquetado basado en distancias Z-score ofrecía una representación más robusta y sensible a los desvíos en la pronunciación de los alumnos, permitiendo una clasificación más consistente en los ejes tonales definidos.

Con relación al cuarto objetivo específico, se formuló e implementó el modelo final de clasificación, combinando entradas acústicas secuenciales con una variable auxiliar de género, y entrenándolo con un conjunto de datos propio recolectado a través de tres iteraciones. El modelo fue capaz de identificar errores de pronunciación en tres categorías tonales: correcta (C), demasiado clara (C+) y demasiado oscura (C−), alcanzando una precisión superior al 63% bajo condiciones de datos limitados.

Finalmente, para dar cumplimiento al quinto objetivo, se desarrolló un prototipo funcional de aplicación móvil que permite a los estudiantes grabar su voz, enviarla al servidor para análisis, y recibir retroalimentación automática sobre su pronunciación. Este prototipo integró la infraestructura necesaria para realizar inferencias en tiempo real, gestionar usuarios mediante Firebase y almacenar datos en servicios de AWS. Las pruebas realizadas con usuarios reales demostraron la viabilidad técnica del sistema y su potencial como herramienta de apoyo en el aprendizaje de francés como lengua extranjera.

En síntesis, este trabajo constituye una contribución en el campo de la fonética computacional y el aprendizaje asistido por tecnología, al proponer un modelo funcional que integra teoría acústica, aprendizaje automático y diseño de experiencia de usuario para abordar un desafío frecuente en la

enseñanza de lenguas extranjeras: la detección y corrección personalizada de errores de pronunciación.

7.1 Trabajo a futuro

Si bien el modelo desarrollado en este estudio ha demostrado resultados prometedores en condiciones controladas y con un conjunto de datos limitado, existen múltiples líneas de trabajo a futuro que permitirían fortalecer tanto su aplicabilidad pedagógica como su desempeño técnico. Estas líneas abarcan desde la evolución de la aplicación hacia un entorno educativo más completo, hasta mejoras sustantivas en el modelo de clasificación y en el diseño del corpus de entrenamiento.

En primer lugar, se propone el desarrollo de un módulo dirigido a profesores, que permita visualizar de manera clara los resultados de los diagnósticos fonéticos obtenidos por sus estudiantes. Este módulo podría incluir gráficos individuales y grupales, indicadores de progreso por fonema, y reportes descargables para su uso en contextos formales de enseñanza. La disponibilidad de estos datos permitiría al docente tomar decisiones pedagógicas más informadas, intervenir de forma más personalizada y monitorear la evolución de la pronunciación a lo largo del tiempo. Esta funcionalidad ampliaría el impacto del sistema, cerrando el ciclo entre evaluación automática y acción didáctica.

Desde el punto de vista técnico, una prioridad es la ampliación del corpus de entrenamiento, incorporando una mayor cantidad y diversidad de muestras. Esto no solo permitiría mejorar la robustez del modelo, sino también explorar generalizaciones hacia otros grupos de hablantes L2, incluyendo distintas edades, niveles de competencia y acentos regionales. Asimismo, la inclusión de más fonemas del inventario vocálico francés podría permitir escalar el sistema hacia una cobertura más completa del idioma.

Otra línea de mejora está en el enriquecimiento de las representaciones acústicas utilizadas como entrada del modelo. Aunque el presente estudio se centró en los formantes F1 y F2, podrían incorporarse otras características como la duración de los fonemas, la energía, o incluso representaciones derivadas de espectrogramas o coeficientes cepstrales (MFCCs). Estas características podrían ser especialmente útiles en casos de pronunciación ambigua, donde la información articulatoria por sí sola no sea suficiente para clasificar el error con precisión.

Finalmente, se propone investigar arquitecturas más sofisticadas que puedan capturar mejor la complejidad temporal y acústica de los datos. Entre estas opciones se encuentran redes bidireccionales, modelos basados en transformadores o arquitecturas preentrenadas para tareas de procesamiento del habla. Complementariamente, sería relevante incorporar un proceso de evaluación con usuarios finales (profesores y estudiantes), para validar el sistema en condiciones reales de uso y recoger retroalimentación cualitativa que oriente su evolución futura.

Referencias

[1] P. de Sinety, M. Fort, y J. Pécheur, “Rapport au Parlement sur la langue française 2024”. Ministère de la culture - Délégation générale à la langue française et aux langues de France, 2024. Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://www.culture.gouv.fr/Media/Medias-creation-rapide-Ne-pas-supprimer/Rapport-Parlement-langue-francaise-2024-HD.pdf

[2] R. Marcoux, L. Richard, y A. Wolff, *Estimation des populations francophones dans le monde en 2022. Sources et démarches méthodologiques*, Note de recherche de l’ODSEF, Québec., 2022. Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://www.odsef.fss.ulaval.ca/sites/odsef.fss.ulaval.ca/files/uploads/odsef-lfdm-2022.pdf

[3] “2020 censo de Población y Vivienda. Cuestionario Ampliado”. INEGI, 2020. Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://www.inegi.org.mx/contenidos/programas/ccpv/2020/doc/Censo2020_cuest_ampliado.pdf

[4] “Cursos de francés”, La France au Mexique - Francia en México. Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://mx.ambafrance.org/Cursos-de-frances,6569

[5] “Congreso de la Asociación Mexicana de Estudios Internacionales (13 de octubre de 2022)”, La France au Mexique - Francia en México. Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://mx.ambafrance.org/Congreso-de-la-Asociacion-Mexicana-de-Estudios-Internacionales-2022

[6] E. Huerta Espinosa y A. Hernández Bravo, “Fonética comparada, español-francés, francés como segunda lengua para hispanohablantes, los fonemas complicados: contraste fónico de una lengua extranjera”, 2013.

[7] RAE, “Los fonemas del español | Ortografía de la lengua española”, «Ortografía de la lengua española (2010)». Consultado: el 13 de mayo de 2024. [En línea]. Disponible en: https://www.rae.es/ortografía/los-fonemas-del-español

[8] RAE, “Fonemas y grafemas | Ortografía de la lengua española”, «Ortografía de la lengua española (2010)». Consultado: el 12 de mayo de 2024. [En línea]. Disponible en: https://www.rae.es/ortografía/fonemas-y-grafemas

[9] B. Blin, R. Olmedo Yúdico Becerril, y V. Martínez De Badereaux, “Discurso gramatical y contextualización: descripciones interlingüísticas de docentes de francés en México”, *Lenguaje*, vol. 48, núm. 2, pp. 241–260, jul. 2020, doi: 10.25100/lenguaje.v48i2.8144.

[10] B. J. Kröger, V. Graf-Borttscheller, y A. Lowit, “Two- and Three-Dimensional Visual Articulatory Models for Pronunciation Training and for Treatment of Speech Disorders”, presentado en interspeech 2008, Brisbane Australia, sep. 2008. Consultado: el 13 de mayo de 2025. [En línea]. Disponible en: https://www.isca-archive.org/interspeech_2008/kroger08_interspeech.pdf

[11] L. Fontan, “Évaluer la parole des apprenants de FLE : approches et outils automatiques”, *Revue Japonaise de Didactique du Français*, vol. 12, pp. 157–168, ene. 2017.

[12] E. Alonso, M. Bruña, y M. Muñoz, *La lingüística francesa: gramática, historia, epistemología*. Sevilla: Grupo andaluz de pragmática, 1996.

[13] M. Billières, “MÉTHODE ARTICULATOIRE ET MÉTHODE VERBO-TONALE”, PHONÉTIQUE CORRECTIVE EN FLE MÉTHODE VERBO-TONALE. Consultado: el 10 de junio de 2024. [En línea]. Disponible en: https://mvt-uoh.univ-tlse2.fr/seq04P0101.html

[14] “Méthode verbo tonale: origine et fondements – Au son du fle – Michel Billières”. Consultado: el 19 de mayo de 2024. [En línea]. Disponible en: https://www.verbotonale-phonetique.com/origines-fondements/

[15] "Classement des voyelles françaises sur l'axe clair / sombre et sur l'axe de la tension". Consultado: el 14 de junio de 2024. [En línea]. Disponible en: https://mvt-uoh.univ-tlse2.fr/DOCS/DOC07.pdf
[16] D. Y. Zhang, S. Saha, y S. Campbell, "Phonetic RNN-Transducer for Mispronunciation Diagnosis", en *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10094945.
[17] B.-C. Yan y B. Chen, "End-to-End Mispronunciation Detection and Diagnosis From Raw Waveforms", en *2021 29th European Signal Processing Conference (EUSIPCO)*, ago. 2021, pp. 61–65. doi: 10.23919/EUSIPCO54536.2021.9615987.
[18] A. Diment, E. Fagerlund, A. Benfield, y T. Virtanen, "Detection of Typical Pronunciation Errors in Non-native English Speech Using Convolutional Recurrent Neural Networks", en *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary: IEEE, jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8851963.
[19] P. Boersma, T. Benders, y K. Seinhorst, "Neural network models for phonology and phonetics", *JLM*, vol. 8, núm. 1, oct. 2020, doi: 10.15398/jlm.v8i1.224.
[20] S. S. S. C. D. Y. Zhang, "Phonetic RNN-Transducer for Mispronunciation Diagnosis", en *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), no 10.1109/ICASSP49357.2023.10094945*, 2023, pp. 1-5,.
[21] "Your Personal English Coach | Elsaspeak". Consultado: el 16 de junio de 2024. [En línea]. Disponible en: https://elsaspeak.com/en/product
[22] "Speechace API", Speechace API. Consultado: el 11 de mayo de 2024. [En línea]. Disponible en: https://docs.speechace.com
[23] eric-urban, "Use pronunciation assessment - Azure AI services". Consultado: el 16 de junio de 2024. [En línea]. Disponible en: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-pronunciation-assessment
[24] P. Boersma y D. Weenink, *Praat: doing phonetics by computer.* (2024 de 1992).
[25] CALLIOPE, *La Parole et son traitement automatique*. en Collection technique et scientifique des Télécommunications. Paris Milan Barcelone: Masson, 1989.
[26] "Phones et phonèmes : les distinguer". Consultado: el 16 de junio de 2024. [En línea]. Disponible en: https://vitrinelinguistique.oqlf.gouv.qc.ca/24507/la-prononciation/notions-de-base-en-phonetique/les-phones-et-les-phonemes
[27] A. Corine, *Liens articulatoire-acoustique - l'exemple des voyelles du français* -. Consultado: el 16 de junio de 2024. [Mp4]. Disponible en: https://mvt-uoh.univ-tlse2.fr/seq07P0101.html
[28] *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. Berlin: De Gruyter, 1971.
[29] "Le principe des aires de dispersion des voyelles (orales) françaises". Consultado: el 14 de junio de 2024. [En línea]. Disponible en: https://mvt-uoh.univ-tlse2.fr/DOCS/DOC08.pdf
[30] M. Billières, "LA CORRECTION DES ÉLÉMENTS SEGMENTAUX : VOYELLES ET CONSONNES", PHONÉTIQUE CORRECTIVE EN FLE MÉTHODE VERBO-TONALE. Consultado: el 19 de mayo de 2024. [En línea]. Disponible en: https://mvt-uoh.univ-tlse2.fr/seq05P0302.html
[31] M. Billières, "LES FONDEMENTS DE LA MÉTHODE VERBO-TONALE D'INTÉGRATION PHONÉTIQUE", PHONÉTIQUE CORRECTIVE EN FLE MÉTHODE VERBO-TONALE. Consultado: el 19 de mayo de 2024. [En línea]. Disponible en: https://mvt-uoh.univ-tlse2.fr/seq03P0101.html
[32] J. CUREAU y B. VULETIC, "Enseignement de la prononciation. Le système verbo-tonal (SGAV)". 1976.
[33] "Formation de formateurs en correction phonétique", Fonetix. Consultado: el 7 de enero de 2025. [En línea]. Disponible en: https://www.fonetix.fr/formation-de-formateurs-en-correction-phonetique/

[34] P. E. Black, “Euclidean Distance”, *Dictionary of Algorithms and Data Structures*. diciembre de 2004. [En línea]. Disponible en: https://www.nist.gov/dads/HTML/euclidndstnc.html
[35] “zscore - Puntuaciones z estandarizadas - MATLAB”. Consultado: el 13 de mayo de 2025. [En línea]. Disponible en: https://la.mathworks.com/help/stats/zscore.html
[36] J. Hillenbrand, L. A. Getty, M. J. Clark, y K. Wheeler, “Acoustic characteristics of American English vowels”, *The Journal of the Acoustical Society of America*, vol. 97, núm. 5, pp. 3099–3111, may 1995, doi: 10.1121/1.411872.
[37] I. Goodfellow, Y. Bengio, y A. Courville, *Deep Learning*. MIT Press, 2016.
[38] M. T. Hagan, H. B. Demuth, M. H. Beale, y O. De Jésus, *Neural network design*, 2nd edition. s.L: Martin T. Hagan, 2014.
[39] P. C. Woodland y D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition”, *Computer Speech & Language*, vol. 16, núm. 1, pp. 25–47, ene. 2002, doi: 10.1006/csla.2001.0182.
[40] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, y M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi”, en *Proc. Interspeech 2017*, 2017, pp. 498–502. doi: 10.21437/Interspeech.2017-1386.
[41] R. Hernández Sampieri y C. P. Mendoza Torres, *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*, Segunda edición. Ciudad de México: McGraw-Hill Interamericana Editores, 2023.
[42] N. Hotz, “What is CRISP DM?”, Data Science Process Alliance. Consultado: el 19 de junio de 2024. [En línea]. Disponible en: https://www.datascience-pm.com/crisp-dm-2/
[43] P. Chapman, J. Clinton, y R. Kerber, “CRISP-DM 1.0”. Consultado: el 19 de junio de 2024. [En línea]. Disponible en: https://web.archive.org/web/20220401041957/https://www.the-modeling-agency.com/crisp-dm.pdf
[44] “What is scrum | Guide to the most popular agile framework”. Consultado: el 19 de junio de 2024. [En línea]. Disponible en: https://www.scrumalliance.org/about-scrum
[45] “Mozilla Common Voice”. Consultado: el 12 de mayo de 2025. [En línea]. Disponible en: https://commonvoice.mozilla.org/
[46] “Torchaudio Documentation — Torchaudio 2.5.0 documentation”. Consultado: el 1 de diciembre de 2024. [En línea]. Disponible en: https://pytorch.org/audio/stable/index.html
[47] “Sound: To Formant (burg)...” Consultado: el 7 de enero de 2025. [En línea]. Disponible en: https://www.fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html
[48] M. Sokolova y G. Lapalme, “A systematic analysis of performance measures for classification tasks”, *Information Processing & Management*, vol. 45, núm. 4, pp. 427–437, jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[49] “FastAPI”. Consultado: el 13 de mayo de 2025. [En línea]. Disponible en: https://fastapi.tiangolo.com/

Anexo A: Hoja de especificación del Alfabeto Fonético Internacional

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	ʼ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

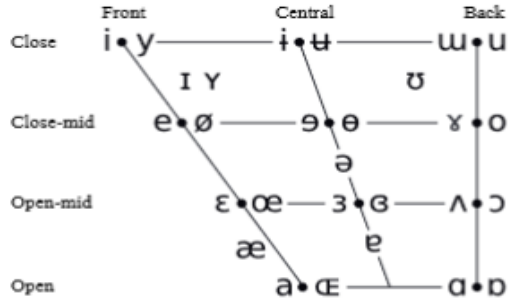
ʍ Voiceless labial-velar fricative
w Voiced labial-velar approximant
ɥ Voiced labial-palatal approximant
ʜ Voiceless epiglottal fricative
ʢ Voiced epiglottal fricative
ʡ Epiglottal plosive

ɕ ʑ Alveolo-palatal fricatives
ɺ Voiced alveolar lateral flap
ʧ Simultaneous ʃ and x

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

t͡s k͡p

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən
ˌ Secondary stress
ː Long eː
ˑ Half-long eˑ
˘ Extra-short ĕ
| Minor (foot) group
‖ Major (intonation) group
. Syllable break ɹi.ækt
‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or ˥	Extra high	ě or ˧˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷌ ˦˥	High rising
è ˨	Low	e᷄ ˧˩	Low rising
ȅ ˩	Extra low	e᷅ ˥˧˩	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤	◌̪	Dental	t̪ d̪
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰	◌̺	Apical	t̺ d̺
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼	◌̻	Laminal	t̻ d̻
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ	◌̃	Nasalized	ẽ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ	◌ⁿ	Nasal release	dⁿ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ	◌ˡ	Lateral release	dˡ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ	◌̚	No audible release	d̚
◌̈	Centralized	ë	~	Velarized or pharyngealized	ɫ			
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝	(ɹ̝ = voiced alveolar fricative)		
◌̩	Syllabic	n̩	◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)		
◌̯	Non-syllabic	e̯	◌̟	Advanced Tongue Root	e̟			
◌˞	Rhoticity	ə˞ a˞	◌̠	Retracted Tongue Root	e̠			

Anexo B: Cuestionario para grabación de audio

1.Edad

Marca solo un óvalo.

15-18

18-30

31-50

51+

2. Sexo

Marca solo un óvalo.

Masculino

Femenino

3. País de origen

_____

4. Lengua materna

Marca solo un óvalo.

Otros:

Español

Otro (Especificar)

5. ¿Cuánto tiempo llevas estudiando francés?

Marca solo un óvalo.

Menos de 1 año

1-2 años

3-5 años

Más de 6 años

6. ¿Cómo calificarías tu nivel de francés?

Marca solo un óvalo.

Principiante

Intermedio

Avanzado

Fluido

7. ¿Has vivido en un país francófono?

Marca solo un óvalo.

Sí

No

8. ¿Si es así, por cuánto tiempo?

Marca solo un óvalo.

Menos de 6 meses

6 meses a 1 año

1-3 años

Más de 3 años

9. ¿Hablas otros idiomas con fluidez (además de español y francés)?

Marca solo un óvalo.

Sí

No

10. Si es así, por favor, indícalos

11. ¿Qué tan confiado te sientes al hablar en francés?

Marca solo un óvalo.

Nada confiado

Algo confiado

Confiado

Muy confiado

12. ¿Con qué frecuencia hablas francés en tu vida diaria?

Marca solo un óvalo.

Raramente

A veces

A menudo

Diariamente

13. ¿Has estado expuesto a diferentes acentos del francés (ej. Quebecois, parisino,francés africano)?

Marca solo un óvalo.

Sí

No

14. Si es así, ¿a qué acentos has estado más expuesto?

Selecciona todas las opciones que correspondan.

Quebecois

Parisino

Francés africano

Otro (Especificar)

15. ¿Tienes alguna dificultad para hablar o problemas auditivos?

Marca solo un óvalo.

Sí

No

16. Si es así, describe brevemente

17. ¿Dónde realizaste tu grabación?

Marca solo un óvalo.

Otros:

Habitación tranquila (mínimo ruido de fondo)

Entorno ruidoso

Otro (Especificar)

18. ¿Qué dispositivo utilizaste para grabar?

Marca solo un óvalo.

Otros:

Micrófono de alta calidad

Micrófono de teléfono móvil

Micrófono de computadora/portátil

Otro (Especificar)