Software de reconocimiento de voz para el desarrollo de la pronunciación de inglés: Una revisión sistemática

Speech recognition software for english pronunciation development: A systematic review

Eliud Alberto García Pazos
Universidad Veracruzana, Xalapa Veracruz, México
garciapazos61@gmail.com
ORCID: 0009-0002-5724-3030

Juana Elisa Escalante Vega
Universidad Veracruzana, Xalapa Veracruz, México
jescalante@uv.mx
ORCID: 0000-0001-8192-6267

Oscar Alonso Ramírez
Universidad Veracruzana, Xalapa Veracruz, México
oalonso@uv.mx
ORCID: 0009-0007-6476-5781

Fredy Castañeda Sánchez
Universidad Veracruzana, Xalapa Veracruz, México
fcastaneda@uv.mx
ORCID: 0000-0003-1675-6438

doi: https://doi.org/10.36825/RITI.13.29.014

Recibido: Marzo 04, 2025
Aceptado: Junio 28, 2025

Resumen: Este artículo presenta una revisión sistemática de la literatura enfocada en herramientas de software diseñadas para mejorar la pronunciación del vocabulario en inglés mediante el reconocimiento de voz. El objetivo principal es identificar y analizar las características, desafíos y efectividad de estas herramientas. Se realizó un análisis comparativo de diversas soluciones existentes, evaluando su impacto en la mejora de la pronunciación, sus funcionalidades específicas para el aprendizaje de vocabulario en inglés y su capacidad para adaptarse a diferentes contextos de enseñanza. El propósito de este estudio es proporcionar una visión integral del estado actual de las herramientas de reconocimiento de voz en este campo, destacando áreas para mejorar y posibles direcciones futuras para la investigación y el desarrollo.

Palabras clave: *Reconocimiento de Voz, Software, Calidad de la Pronunciación, Evaluación de la Pronunciación, Análisis de Voz.*

Abstract: This article presents a systematic review of the literature focusing on software tools designed to improve English vocabulary pronunciation through speech recognition. The main objective is to identify and analyze the features, challenges, and effectiveness of these tools. A comparative analysis of various existing solutions will be carried out, evaluating their impact on pronunciation improvement, their specific functionalities for English vocabulary learning and their ability to adapt to different teaching contexts. The purpose of this study is to provide a comprehensive overview of the current state of the art of speech recognition tools in this field, highlighting areas for improvement and possible future directions for research and development.

Keywords: *Speech Recognition, Software, Pronunciation Quality, Pronunciation Assessment, Voice Analysis.*

1. Introducción

Una de las dificultades en la enseñanza del inglés para los estudiantes como una lengua extranjera es el desarrollo de la habilidad oral, de la pronunciación en particular. Con el desarrollo de las tecnologías de procesamiento de voz, el reconocimiento automático del habla (ASR) se ha convertido en una forma de ayudar a los estudiantes a practicar su pronunciación mediante herramientas de aprendizaje personalizadas y accesibles. Estas tecnologías realizan detección de errores de una manera instantánea, identificando errores fonéticos y se adaptan a estrategias diferentes dependiendo del nivel del usuario, lo que resulta especialmente útil en contextos educativos donde el acceso a los maestros de inglés es limitado o nulo.

En los últimos años, se ha desarrollado software educativo con tecnología de reconocimiento de voz, donde se incorporan algoritmos tradicionales y algoritmos avanzados de aprendizaje automático implementación de métricas especializadas, de evaluación fonética y elementos interactivos. Sin embargo, existe una gran variabilidad en estos enfoques que se utilizan, sus funcionalidades, y niveles de precisión que alcanzan. Esta variabilidad plantea la necesidad de realizar una revisión sistemática para identificar, clasificar y analizar las herramientas disponibles.

El objetivo de este artículo es identificar las características de las técnicas, los algoritmos y métricas empleadas en herramientas de software que hacen uso del reconocimiento de voz para evaluar la pronunciación del inglés. Entre las principales conclusiones encontradas se destaca que los enfoques que se basan en el aprendizaje profundo superan en la precisión y adaptabilidad a los métodos tradicionales. Además, técnicas como gamificación e interfaces interactivas surgen como factores que fomentan la práctica continua en los estudiantes de inglés. La organización de este artículo está estructurada de la siguiente manera: en la sección 2 se describe la metodología PRISMA utilizada para la sección de los estudios; en la sección 3 se presenta un análisis cuantitativo llevado a cabo para la selección de los artículos; en la sección 4 se presenta un análisis cualitativo de las técnicas, algoritmos y métricas identificadas; la sección 5 sintetiza los hallazgos en función de las preguntas de investigación; en la sección 6 se discute los resultados y finalmente en la sección 7 se presentan las conclusiones y propuestas para futuras investigaciones.

2. Metodología

La metodología propuesta en el desarrollo de esta revisión sistemática fue PRISMA [1], esta metodología está diseñada para ayudar a documentar de manera transparente el porqué de la revisión, cómo fue hecha la búsqueda por parte de los autores, así como todo lo que se encontró al respecto. La metodología PRISMA describe una serie de pasos para realizar una revisión sistemática, los cuales son:

- Generar preguntas de investigación: se define el objetivo de la revisión y se formulan preguntas específicas.
- Elaborar un protocolo: se describen los criterios de inclusión y exclusión para el método de búsqueda, realizar una síntesis de la información recabada y definir los criterios de calidad de los estudios que se incluyan.
- Realizar búsqueda: se realiza una cadena de búsqueda para identificar en bases de datos bibliográficas ensayos y repositorios todos los estudios que intenten responder las preguntas de investigación planteadas.

- Extraer información: se realiza un análisis cualitativo y cuantitativo donde se analiza la información relevante de los estudios incluidos, como características y resultados.
- Interpretar los resultados: se discuten las implicaciones de los hallazgos y su relevancia para las preguntas de investigación.

Al seguir estos pasos de la metodología, se realiza una revisión sistemática de la literatura de una forma más rigurosa y transparente y de esta manera contribuir a generar una evidencia confiable y útil para tomar decisiones en la investigación.

2.1. Preguntas de investigación

Se desarrollaron cuatro preguntas de investigación para guiar el análisis como parte integral de esta revisión sistemática:

1. ¿Qué herramientas de software basadas en el reconocimiento de voz han sido desarrolladas para mejorar la pronunciación del vocabulario en inglés?
2. ¿Qué técnicas utilizan las herramientas para analizar la pronunciación?
3. ¿Qué algoritmos se ocupan actualmente para analizar la pronunciación?
4. ¿Qué heurísticas o métrica utilizan los algoritmos para evaluar la pronunciación?

2.2. Fuentes de información

Las bases de datos usadas para la investigación fueron: ACM, IEEE, Science Direct y Springer Link.

- Las bases de datos ACM, IEEE, Science Direct y Springer Link fueron seleccionadas para esta investigación debido a su amplia relevancia y volumen de artículos en el ámbito de la tecnología y las ciencias aplicadas.
- ACM y IEEE son reconocidas por su enfoque en investigaciones de alta calidad en informática y tecnología, en este caso en áreas relacionadas con el reconocimiento de voz y la inteligencia artificial.
- Science Direct y Springer Link, por su parte ofrecen acceso a una gran cantidad de artículos que contienen un amplio campo de disciplinas científicas, incluyendo la lingüística, tecnología educativa y ciencias de la computación, los cuales son fundamentales para el análisis de herramientas de software usadas en el análisis de pronunciación en inglés.

2.3. Ecuación de búsqueda

Se realizó la siguiente cadena de búsqueda para encontrar los casos de estudios relacionados al tema:

("speech recognition" OR "voice recognition") AND ("software" OR "app") AND ("pronunciation improvement" OR "pronunciation quality" OR "pronunciation assessment")*

La fecha que se consideró para incluir los artículos en la búsqueda fue entre el primero de enero del 2019 al 26 de septiembre del 2024.

2.4. Criterios de selección

En el proceso de selección de los artículos encontrados se definieron criterios de inclusión y criterios de exclusión para garantizar que los artículos sean de relevancia y calidad para el análisis de esta revisión sistemática.

- Criterios de inclusión.
  - Artículos con acceso completo.
  - Publicación entre los años 01 enero 2019 a 26 de septiembre 2024. (5 años).
  - Título o resumen que intenta responder una pregunta de investigación.
- Criterios de exclusión.
  - Estudios en un idioma diferente a español o inglés.
  - Estudios de un área distinta a las ciencias de la computación.
  - Protocolos, presentaciones o ensayos.

- Criterios de calidad. Para analizar los datos se aplicaron los siguientes criterios de calidad a los artículos resultantes.
    - Relevancia y claridad del objetivo: El estudio debe tener un objetivo claro y relevante para mi revisión sistemática.
    - Relevancia para la revisión: El estudio debe ser relevante para la pregunta de investigación de la revisión sistemática y contribuir de manera significativa al conocimiento existente.
    - Estructura del documento: El documento presentado se debe desarrollar su estructura siguiendo una metodología de manera formal.
    - Resultados y conclusiones: Los resultados deben ser claros, relevantes y estar respaldados por los datos presentados. Las conclusiones deben estar basadas en los resultados y no ser exageradas o especulativas.

3. Marco metodológico conceptual

En esta sección se presentan los principales conceptos que se mencionan en esta revisión sistemática, con la finalidad de establecer un marco teórico que permita comprender cómo el software de reconocimiento de voz, los algoritmos y las métricas se relacionan con la evaluación de la pronunciación del idioma inglés.

3.1. Software de reconocimiento de voz

El software de reconocimiento de voz (ASR, por sus siglas en inglés) permite transcribir automáticamente el habla humana a texto. En el contexto del aprendizaje de idiomas, este software es utilizado para evaluar la pronunciación de los estudiantes comparando su voz con una referencia lingüística. Permite ofrecer una retroalimentación inmediata y personalizada sobre la calidad del habla, identificando los errores fonéticos más comunes y proporcionando un diagnóstico detallado de la pronunciación del usuario [4].

3.2. Algoritmos utilizados

Los algoritmos que se implementan en los sistemas de reconocimiento del habla han evolucionado, notablemente. Primero se utilizaron los enfoques tradicionales como los Modelos Ocultos de Markov (HMM) y el Dynamiz Time Warping (DTW), que permiten realizar alineaciones temporales entre secuencias de habla y patrones esperados [20]. Estos modelos presentan limitaciones en cuanto a la adaptación a la variación fonética y para incluir contextos lingüísticos diversos.

Actualmente, se hace uso de modelos más avanzados basados en aprendizaje profundo (Deep Learning), como las redes neuronales convolucionales (CNN), Redes Neuronales Recurrentes (RNN), Transformes, y los modelos auto supervisados como Wav2Vec2.0 [2][21]. Estos enfoques que son mas modernos permiten realizar una evaluación más precisa de la pronunciación al capturar patrones acústicos complejos y aspectos prosódicos como la entonación, el ritmo y la segmentación fonética, lo que ha mejorado notablemente el rendimiento de los sistemas de la evaluación automatizada [1][17].

3.3. Métricas de evaluación

La evaluación automática de la pronunciación en sistemas de reconocimiento de voz requiere un uso de métricas especializadas que permiten cuantificar la calidad del habla del usuario. Las métricas que son más utilizadas son:

1. *Word Error* Rate (WER): calcula la proporción de palabras incorrectas comparando la transcripción generada por el sistema con la transcripción esperada [5].
2. *Phoneme Error Rate* (PER): es una métrica que mide los errores a nivel de fonema, útil para evaluar la precisión fonética, especialmente en aplicaciones de pronunciación [5].
3. *Goodness of pronunciation*: se utiliza para medir qué tan bien un sistema de reconocimiento detecta y registra la pronunciación de los fonemas en una grabación, se calcula como la proporción de fonemas correctamente detectados en relación con el total de fonemas en la prueba [20].
4. *Mean Squared Error* (MSE): cuantifica el error entre la puntuación generada por el sistema y la de la referencia humana [17].

5. *Pearson Correlation Coefficient* (PCC): Evalúa la relación lineal entre los puntajes estimados por el sistema y los asignados por expertos humanos [21].

3.4 Relación con el aprendizaje de pronunciación

Estas algoritmos y métricas se integran en sistemas de reconocimiento automático del habla para aplicarse en herramientas educativas para proporcionar retroalimentación automática y en tiempo real sobre la pronunciación al estudiante.

Estas métricas y algoritmos se integran en sistemas de reconocimiento automático del habla (ASR), son empleados en herramientas educativas con el fin de ofrecer a los estudiantes retroalimentación automática y en tiempo real. Estos sistemas permiten detectar errores de pronunciación, estimar el ajuste fonético y adaptar la retroalimentación al nivel del usuario, facilitando así el aprendizaje autónomo y personalizado del inglés como una segunda lengua [6], [9], [16].

4. Análisis cuantitativo:

Con el fin de reducir y organizar de manera sistemática la literatura recopilada, se realizó un análisis cuantitativo, donde se seleccionó solo aquellos artículos que cumplen con los criterios establecidos. A continuación, se describe el proceso de filtrado aplicado en cada etapa.

1. Recopilación de datos iniciales: Se inició con la búsqueda en cuatro fuentes principales: ACM, IEEE, Science Direct y Springer Link. En total, se identificaron 228 artículos en la búsqueda inicial (7 de ACM, 107 de IEEE, 38 de Science Direct y 76 de Springer Link). Esta primera fase incluye todos los estudios que mencionan las palabras clave definidas en la cadena de búsqueda.
2. Primera fase de filtro (CI.1 y CI.2): En esta fase, se aplicaron los primeros criterios de inclusión para reducir los artículos a aquellos que son de acceso completo y con un rango máximo de 5 años. Después de este filtro, el número de artículos se redujo a 113 (2 de ACM, 46 de IEEE, 19 de Science Direct y 46 de Springer Link).
3. Segunda fase de filtro (CE.1 y CE.2): Aquí se aplicaron los criterios de exclusión, eliminando los estudios que, sean de un idioma diferente al inglés o español, además de que solo se incluyan estudios de investigación del área de las ciencias de computación. Tras este filtro, quedaron 100 artículos en total (2 de ACM, 46 de IEEE, 11 de Science Direct y 41 de Springer Link).
4. Tercera fase de filtro (CE.3): En esta fase, se aplicó el criterio de exclusión número 2 para excluir protocolos, presentaciones o ensayos. Esto redujo el número de artículos a 85 (2 de ACM, 37 de IEEE, 10 de Science Direct y 36 de Springer Link).
5. Filtro Final (CI.3): Finalmente, se realizó un último filtro de inclusión, seleccionando solo los estudios más relevantes que de alguna manera su título o resumen intenten responder al menos una pregunta de investigación. El total de artículos seleccionados para la revisión cuantitativa final fue de 30 (2 de ACM, 23 de IEEE, 4 de Science Direct y 1 de Springer Link).
6. Resultados: En la Tabla 1 se muestra un resumen del proceso de la aplicación de los criterios de inclusión y exclusión para la selección de los artículos.

Tabla 1. Resumen de resultados por fuente y criterios.

Fuente	Inicial	CI.1 y 2	CE.1 y 2	CE.3	CI.3
ACM	7	2	2	2	2
IEEE	107	46	46	37	23
Science Direct	38	19	11	10	4
Springer Link	76	46	41	36	1
Total	**228**	**113**	**100**	**85**	**30**

Fuente: Elaboración propia.

Este análisis se lleva a cabo después de definir la cadena de búsqueda, las fuentes de información a consultar y los criterios de inclusión y exclusión de los artículos.

5. Análisis Cualitativo

Se realizó una clasificación de las técnicas empleadas en los artículos analizados con el objetivo de agruparlos por enfoques tradicionales y técnicas de *deep learning*. Esto para facilitar una mejor comprensión de cómo cada técnica contribuye al desarrollo de herramientas y soluciones en este campo. Se consideran como técnicas tradicionales a los estudios que hacen uso de métodos estadísticos, reglas predefinidas y análisis determinísticos y se comprenden como técnicas de *deep learning* a los estudios que han ganado popularidad en los últimos años por su capacidad de procesar grandes volúmenes de datos y procesar representaciones complejas del habla. Las categorías propuestas están planteadas en la Tabla 2.

Tabla 2. Clasificación de artículos.

Clasificación	Artículos
Técnicas modernas basadas en aprendizaje profundo (*deep learning*)	[2], [3], [4], [5], [6], [7], [8],[9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19],[20], [21], [22]
Técnicas basadas en modelos tradicionales	[23], [24], [25], [26], [27],[28], [29], [30]

Fuente: Elaboración propia

6. Síntesis de resultados

En estos artículos examinados hay investigaciones que ofrecen información importante en el contexto de este estudio. A continuación, se presenta una respuesta a las preguntas de investigación a partir del análisis cualitativo.

¿Qué herramientas de software basadas en el reconocimiento de voz han sido desarrolladas para mejorar la pronunciación del vocabulario en inglés?

En los siguientes 22 artículos se desarrollaron herramientas de software que están enfocadas en evaluar la pronunciación y en específico para el idioma inglés: [20], [3], [6], [23], [24], [8], [9], [11], [12], [14], [25],[26], [16], [27], [17], [18], [31], [19], [28], [21], [30], [22].

¿Qué técnicas utilizan las herramientas para analizar la pronunciación?

a) Aprendizaje profundo

Los modelos basados en aprendizaje profundo son la base para muchos sistemas modernos debido a su capacidad de capturar relaciones complejas entre características acústicas y fonéticas, en la Tabla 3 se enlistan las técnicas basadas en el aprendizaje profundo (*deep learning*).

Tabla 3. Técnicas basadas en *deep learning.*

Técnica	Descripción	Artículos
Transformer	Un *transformer* es un modelo de aprendizaje profundo que se ocupa en el procesamiento de lenguaje natural.	[9], [11], [14], [20]
Transfer Learning	Es una técnica en el campo del aprendizaje automático que implica tomar un modelo previamente entrenado en una tarea específica y adaptarlo para una tarea diferente, pero relacionada.	[18]
Aprendizaje auto-supervisado	Técnica de aprendizaje en la que un modelo se entrena utilizando datos no etiquetados	[3], [9], [17]
Híbridos + atención cruzada	La arquitectura de "híbridos + atención cruzada" combinas técnicas de *Connectionist Temporal Classification* (CTC) y mecanismos de atención en modelos de aprendizaje automático.	[6], [20]
Codificación acústica con CNN	Las CNN se utilizan para procesar los datos de voz del usuario, realizando el preprocesamiento y la extracción de características del sonido.	[26]
Focal Attention Loss	Optimiza el aprendizaje del modelo al enfocar su atención en los errores que causan más confusión durante el reconocimiento de voz.	[21]

Deep Learning + generación de datos sintéticos	La combinación de estas técnicas permite mejorar la detección de errores de pronunciación al proporcionar un acceso más amplio y variado a datos de entrenamiento	[22]

Fuente: Elaboración propia.

b) Gamificación e interacción motivadora

La gamificación fomenta el compromiso, la práctica constante y el aprendizaje autónomo, utilizando recompensas y desafíos para mantener la motivación, en la Tabla 4 se enlistan estas técnicas.

Tabla 4. Técnicas basadas en gamificación e interacción motivadora.

Técnica	Descripción	Artículos
Gamificación interactiva	Estas técnicas fomentan el compromiso y la práctica constante del usuario al ofrecer recompensas.	[20], [15]
Blended Learning	Combina enseñanza presencial tradicional con actividades asistidas por computadora (CAPT).	[29]
Retroalimentación multimodal	Utiliza múltiples formas o modos de comunicación para proporcionar información y apoyo a un usuario en su aprendizaje.	[22]

Fuente: Elaboración propia.

c) Evaluación acústica y fonética basada en modelos tradicionales

La Tabla 5 enlista las técnicas con enfoques clásicos como GOP y HMM, son relevantes, en entornos con recursos limitados.

Tabla 5. Evaluación acústica y fonética basada en modelos tradicionales

Técnica	Descripción	Artículos
Goodness of Pronunciation (GOP)	Es utilizado en la detección de errores de pronunciación.	[4], [11]
Modelos ocultos de Markov (HMM)	Modelo estadístico que se utiliza en el reconocimiento de patrones y procesamiento de señales.	[25], [26]

Fuente: Elaboración propia.

d) Reconocimiento de voz y síntesis de habla

En la Tabla 6 se enlistan las técnicas que realizan generación o análisis de síntesis de datos y muestran tasas de precisión altas.

Tabla 1. Técnicas del reconocimiento de voz y síntesis del habla.

Técnica	Descripción	Artículos
Generación de datos sintéticos	Se crean datos artificiales que simulan características y patrones de datos reales, sin utilizar datos reales directamente.	[2], [24]
Reconocimiento automático de fonemas	Identifica y clasifica los fonemas, que son las unidades de sonido más pequeñas en el habla.	[17], [13]

Fuente: Elaboración propia.

e) Modelos híbridos

Los modelos híbridos se resumen en la Tabla 7, estos combinan múltiples técnicas para abordar diferentes aspectos del aprendizaje de la pronunciación.

Tabla 7. Modelos híbridos.

Técnica	Descripción	Artículos
DTW + Redes neuronales	Se usan redes neuronales para modelar patrones complejos de audio, y DTW para alinear y medir la similitud entre secuencias de tiempo que pueden variar en velocidad.	[3]
Ponderación dinámica de errores	Es una estrategia en el entrenamiento de modelos de aprendizaje automático que ajusta la importancia de diferentes errores durante el proceso de optimización.	[21]
Procesamiento simultaneo de tareas fonética y de fluidez mediante un modelo de codificación híbrido	Permite que el modelo procese de manera concurrente las puntuaciones de diferentes granularidades (fonema, palabra y enunciado)	[28]

Fuente: Elaboración propia.

f) Aprendizaje multimodal y basado en interacción

En la Tabla 8 se enlistan las técnicas que combinan datos visuales y auditivos para proporcionar una retroalimentación más completa.

Tabla 8. Aprendizaje multimodal y basado en interacción.

Técnica	Descripción	Artículos
Realidad aumentada	Permite a los usuarios experimentar en un entorno interactivo para el aprendizaje de la fonética.	[14]
Procesos gaussianos	Es un enfoque estadístico utilizado para realizar inferencias sobre funciones desconocidas.	[18]

Fuente: Elaboración propia.

¿Qué algoritmos se ocupan actualmente para analizar la pronunciación?

La Tabla 9 presenta un resumen de los algoritmos más relevantes utilizados en el análisis y evaluación automática de pronunciación. Estos algoritmos destacan por su capacidad de procesar datos acústicos y lingüísticos, permitiendo identificar errores, evaluar la calidad de la pronunciación y ofrecer retroalimentación precisa.

Tabla 2. Algoritmos usados para analizar la pronunciación

Algoritmo	Descripción
DTW: Dynamic Time Warping	Alineación temporal entre patrones de voz del usuario y referencias.
Transformador aumentado por convolución	Mejora la precisión del análisis fonético combinando características convolucionales y transformadores.
TDNN: Time-Delay Neural Network	Modelo que captura relaciones temporales en datos acústicos.
CNN: Convolutional Neural Network	Redes convolucionales utilizadas para extracción de características acústicas.
RNN: Recurrent Neural Network	Modelos secuenciales para análisis de patrones temporales.
seq2seq: Sequence to Sequence	Transformación de secuencias de entrada en secuencias de salida.
HMM: Hidden Markov Model	Representación de transiciones entre estados fonéticos.
GRU: Gated Recurrent Unit	Variante de RNN eficiente en cálculo para análisis acústico.

CTC: Connectionist Temporal Classification	Algoritmo de alineación temporal para aprendizaje secuencial.
X-VECTOR	Representación de características de voz extraídas de redes neuronales profundas.
WAV2VEC2	Modelo auto-supervisado para extracción de características de voz de alta calidad.
SVM: Support Vector Machine	Modelo estadístico para clasificación de características acústicas.
Transform	Abreviatura de Transformers, arquitectura para tareas de reconocimiento de voz.
LSTM: Long Short-Term Memory	Modelo recurrente diseñado para capturar dependencias a largo plazo.
AdaIN: Adaptive Instance Normalization	Técnica para normalización de características adaptativas.
GOPT: Gradient Optimization	Algoritmo de optimización utilizado en modelos acústicos.
DNN: Deep Neural Network	Redes profundas para análisis acústico y fonético.
GP: Gaussian Processes	Estimación de incertidumbre en modelos acústicos.
Siamese NN: Siamese Neural Network	Modelo para comparación y detección de similitudes en fonemas.
GOP: Goodness of Pronunciation	Evaluación de pronunciación basada en probabilidades posteriores.
ASR: Automatic Speech Recognition	Reconocimiento automático del habla para análisis de pronunciación.
Phoneme Recognizer (PR)	Herramienta para identificación de fonemas en la señal acústica.
Pronunciation Model (PM)	Modelo especializado en evaluar pronunciaciones correctas e incorrectas.
Pronunciation Error Detector (PED)	Detector de errores en pronunciación utilizando redes neuronales.

Fuente: Elaboración propia

¿Qué heurísticas o métricas utilizan los algoritmos para evaluar la pronunciación?

En la Tabla 10 se sintetizan las métricas clave utilizadas en los artículos analizados, junto con las referencias a los artículos donde estas métricas se mencionan. Cada métrica juega un papel esencial en la evaluación de la calidad de pronunciación, fluidez, precisión fonética y otros aspectos relacionados con el aprendizaje y análisis de pronunciación en diferentes contextos.

Tabla 3. Métricas usadas por los algoritmos para analizar la pronunciación.

Algoritmo	Descripción	Artículos
Mean Squared Error (MSE)	Utilizado para medir el error cuadrático medio en la predicción de pronunciaciones.	[14], [17]
Word Error Rate (WER)	Mide la precisión de las palabras reconocidas en la evaluación de pronunciación.	[2], [5]
Phone Error Rate (PER)	Evalúa los errores a nivel de fonema para mejorar los modelos de pronunciación.	[5], [21]
F1-*score*	Métrica clave para la evaluación del rendimiento de detección de errores de pronunciación.	[29], [21]
Log-likelihood	Permite evaluar la calidad de la pronunciación mediante modelos probabilísticos.	[26], [19]

Character Error Rate (CER)	Mide los errores a nivel de caracteres en aplicaciones educativas.	[8], [9]
AUC (*Area Under Curve*)	Evalúa la efectividad del sistema en diversas condiciones de prueba.	[18], [22]
Métrica de similitud	Evalúa la proximidad entre pronunciaciones esperadas y reales.	[2], [19]
MAE (*Mean Absolute Error*)	Calcula el error absoluto medio en evaluaciones.	[3], [9]
CD (Coseno de distancia)	Mide la disimilitud angular entre vectores acústicos.	[3]
Syllable Accuracy	Evalúa la precisión en la producción de sílabas.	[6], [23]
Fluidez y precisión	Mide la calidad tonal y temporal en el habla.	[19]
Log-posterior	Utilizado para calcular probabilidades fonéticas posteriores.	[19]
Probabilidad posterior	Método probabilístico para evaluar fonemas esperados.	[16], [25]
Alineación Temporal (DTW)	Técnica para alinear contornos tonales y evaluar fluidez.	[27], [19]
TA (Precisión tonal)	Porcentaje de tonos correctamente identificados.	[13]
PCC *(Pearson Correlation Coefficient)*	Mide la correlación entre predicciones y evaluaciones humanas.	[14], [17]
F1, Precisión, *Recall*	Métricas clave para evaluar rendimiento en tareas de clasificación.	[21], [29]
Discriminación fonética	Evalúa la habilidad para diferenciar fonemas específicos.	[30]
Comparación pre/post	Analiza la mejora entre evaluaciones iniciales y finales.	[28], [30]
Word Accuracy Rate	Porcentaje de palabras correctamente reconocidas.	[26]
Similitud fonética	Compara representaciones fonéticas para evaluar calidad de pronunciación.	[22]
Probabilidad de error	Predice errores basándose en patrones observados.	[16]

Fuente: Elaboración propia

7. Discusión

En esta investigación se han analizado y comparado los enfoques técnicos, los algoritmos empleados, sus resultados obtenidos y las limitaciones de los estudios revisados. También se identificaron tendencias y desafíos en el área del reconocimiento de voz para mejorar la pronunciación del vocabulario en inglés. A continuación, se describen los puntos más importantes de esta investigación:

7.1. Predominio del aprendizaje profundo en la investigación

El aprendizaje profundo (*Deep Learning*, DL) se ha establecido como el enfoque técnico más relevante, presente en el 80% de los estudios revisados. Este paradigma emplea arquitecturas como:

- Redes Neuronales Profundas (DNN): son modelos que permiten capturar estructuras jerárquicas del habla y patrones fonéticos complejos a través de múltiples capas ocultas, optimizadas para tareas como la detección de errores de pronunciación [17].
- *Transformers*: utilizan mecanismos de atención para modelar relaciones a largo plazo dentro de secuencias de voz, logrando mejoras sustanciales en tareas de reconocimiento y evaluación fonética [17].
- Modelos auto-supervisados como Wav2Vec2.0: permiten entrenar modelos de reconocimiento con datos no etiquetados y luego refinarlos con una menor cantidad de datos anotados, siendo eficaces en contextos con recursos limitados [18].

Estos enfoques han demostrado alta precisión en la evaluación de pronunciación. Por ejemplo, los modelos Wav2Vec2.0 y *Transformers* han reportado valores de PCC > 0.68, superando a métodos tradicionales en tareas de reconocimiento de errores y adecuación fonética [17].

Una de las principales limitaciones de los modelos DL es la necesidad de utilizar datos adicionales para expandir el modelado de pronunciaciones aceptables para diferentes acentos de una segunda lengua. En algunos casos, se proponen soluciones, como en [2], aunque enfrentan desafíos como el *recall* y la adaptabilidad.

7.2. Aprendizaje profundo vs. Métodos tradicionales

Los métodos tradicionales, como GMM-HMM y *Goodness of Pronunciation* (GOP), aún tienen valor en entornos de baja capacidad computacional. El modelo GMM-HMM utiliza una combinación de modelos ocultos de Markov y mezcla gaussiana para modelar secuencias fonéticas, siendo útil para tareas de reconocimiento con recursos limitados [17]. GOP, por su parte, mide la adecuación fonética comparando la pronunciación del hablante con una referencia ideal usando modelos acústicos generativos [20].

Comparativamente:
Métodos tradicionales: PCC ≈ 0.4, PER ≈ 29%.
Modelos DL modernos: PCC > 0.7, PER ≈ 8%.

Estas diferencias reflejan las mejoras sustanciales del DL en la identificación precisa de errores fonológicos y su adaptabilidad a diferentes idiomas o dialectos.

7.3. Uso de gamificación e interactividad

La gamificación, entendida como la aplicación de elementos lúdicos (retos, recompensas, progresión visual), ha demostrado ser efectiva en el incremento del compromiso de los usuarios. Por ejemplo, en [14], se presenta un sistema gamificado con desafíos diarios que aumentaron el tiempo de uso efectivo hasta un 60%. Adicionalmente, enfoques que incorporan interactividad con tecnologías emergentes como la realidad aumentada permiten una experiencia personalizada y motivadora. En uno de los estudios analizados, se reporta que el uso de realidad aumentada combinado con retroalimentación fonética mejoró la retención y motivación del usuario [23].

7.4. Desafíos identificados

Pese a los avances logrados, se identifican desafíos importantes:

- Altos requerimientos computacionales: los modelos DL, especialmente los basados en *Transformers*, requieren hardware especializado (GPU, TPU), lo cual limita su adopción en instituciones con bajo presupuesto [25].
- Adaptabilidad lingüística: aunque el aprendizaje por transferencia mejora la generalización, sigue siendo necesario contar con corpus específicos para optimizar el rendimiento en distintos acentos, edades y niveles de competencia lingüística [18].
- Diseño de interfaces usables e inclusivas: la implementación de componentes interactivos requiere una evaluación centrada en el usuario. Pocos estudios analizan la usabilidad formal o el impacto de la experiencia del usuario, lo cual limita su efectividad a largo plazo [3].

8. Conclusión

En esta investigación se identificó que las técnicas modernas como el aprendizaje profundo (DL) y la generación de datos sintéticos (S2S), ha superado significativamente a las técnicas tradicionales en la evaluación de la pronunciación del vocabulario de la lengua inglesa. Las técnicas modernas ofrecen una mejor precisión y flexibilidad, pero se enfrentan con el desafío de requerir más recursos computacionales y por ende costosos, lo que hace que se limite su implementación en entornos con bajo presupuesto.

Por otro lado, las técnicas tradicionales, como lo métodos *Goodness of Pronunciation* (GOP) y los *modelos ocultos de Markov* (HMM), aunque resultan ser menos efectivos, son valiosos en contextos donde la disponibilidad de los recursos es limitada. Esto hace que sean esenciales en contextos donde se cuenta con un bajo presupuesto.
Además, los enfoques interactivos como lo es la gamificación y la realidad aumentada demuestran un impacto muy positivo en el compromiso que tienen los usuarios, fomentando que la participación y el aprendizaje sean más

constantes. Sin embargo, su aplicación e implementación resulta ser un desafío significativo ya que su diseño debe ser especializado y con validaciones de usabilidad.

La transición de los métodos tradicionales hacia técnicas más avanzadas como el aprendizaje profundo ha marcado una mejora en este campo. Sin embargo, las áreas como la escalabilidad, adaptabilidad y usabilidad siguen siendo de gran importancia para futuras investigaciones, específicamente en el desarrollo de soluciones accesibles y efectivas que puedan beneficiar a nuevos usuarios.

9. Referencias

[1] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. L., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., Moher, D., Yepes-Nuñez, J. J., Urrútia, G., Romero-García, M., Alonso-Fernández, S. (2021). Declaración PRISMA 2020: Una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología, 74* (9), 790–799. https://doi.org/10.1016/j.recesp.2021.06.016

[2] Chen, Y., Hu, J., Zhang, X. (2019). *SELL-Corpus: An open source multiple accented Chinese-English speech corpus for L2 English learning assessment*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK. https://doi.org/10.1109/ICASSP.2019.8682612

[3] Anand, N., Sirigiraju, M., Yarra, C. (2023). *Unsupervised pronunciation assessment analysis using utterance level alignment distance with self-supervised representations*. IEEE 20th India Council International Conference (INDICON), Hyderabad, India. https://doi.org/10.1109/INDICON59947.2023.10440736

[4] Boyi, H., Guangliang, L. (2024). *Analysis of English speech learning quality based on speech recognition technology*. International Conference on Optimization Computing and Wireless Communication (ICOCWC), Tabor, Ethiopia. https://doi.org/10.1109/ICOCWC60930.2024.10470570

[5] Hoesen, D., Putri, F. Y., Lestari, D. P. (2019). *Automatic pronunciation generator for Indonesian speech recognition system based on sequence-to-sequence model*. 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Cebu, Philippines. https://doi.org/10.1109/O-COCOSDA46868.2019.9041182

[6] Nandal, P., Kadian, Y., Upadhyay, S., Mudgal, B. P. (2021). *Pronunciation accuracy calculator using machine learning*. 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India. https://doi.org/10.1109/ICCMC51019.2021.9418381

[7] Korzekwa, D., Lorenzo-Trueba, J., Zaporowski, S., Calamaro, S., Drugman, T., Kostek, B. (2021). *Mispronunciation detection in nonnative (L2) English with uncertainty modeling*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada. https://doi.org/10.1109/ICASSP39728.2021.9413953

[8] Lesnov, A., Zueva, N., Turalchuk, K. (2024). *Leveraging deep learning for automatic pronunciation assessment in a mobile application*. 4th International Conference on Technology Enhanced Learning in Higher Education (TELE), Lipetsk, Russian Federation. https://doi.org/10.1109/TELE62556.2024.10605133

[9] Getman, Y., Phan, N., Al-Ghezi, R., Voskoboinik, E., Singh, M., Grosz, T., Kurimo, M., Salvi, G., Svendsen, T., Strömbergsson, S., Smolander, A., Ylinen, S. (2023). Developing an AI-assisted low-resource spoken language learning app for children. *IEEE Access, 11*, 86025–86037. https://doi.org/10.1109/ACCESS.2023.3304274

[10] You, Z., Nijat, M., Shi, Y., Chen, C., Du, W., Hamdulla, A., Wang, D. (2023). *Zero-shot mispronunciation detection by knowledge-based data augmentation*. 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Delhi, India. https://doi.org/10.1109/O-COCOSDA60357.2023.10482946

[11] Zhang, Z. (2023). *Research on oral language evaluation method of business English based on speech recognition*. 7th Asian Conference on Artificial Intelligence Technology (ACAIT), Jiaxing, China. https://doi.org/10.1109/ACAIT60137.2023.10528551

[12] Tolba, R. M., Elarif, T., Taha, Z., Hammady, R. (2024). Interactive augmented reality system for learning phonetics using artificial intelligence. *IEEE Access, 12*, 78219–78231. https://doi.org/10.1109/ACCESS.2024.3406494

[13]Ke, D., Yao, W., Hu, R., Huang, L., Luo, Q., Shu, W. (2022). *A new spoken language teaching tech: Combining multi-attention and adain for one-shot cross-language voice conversion*. 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), Singapore, Singapore. https://doi.org/10.1109/ISCSLP57327.2022.10038137

[14]Yan, B.-C., Wang, H.-W., Wang, Y.-C., Li, J.-T., Lin, C.-H., Chen, B. (2023). *Preserving phonemic distinctions for ordinal regression: A novel loss function for automatic pronunciation assessment*. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan. https://doi.org/10.1109/ASRU57964.2023.10389777

[15]Lin, B., Wang, L. (2023). *Multi-lingual pronunciation assessment with unified phoneme set and language-specific embeddings*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece. https://doi.org/10.1109/ICASSP49357.2023.10095673

[16]Lin, B., Wang, L. (2021). *Uncertainty estimation in automatic pronunciation assessment with pseudo samples based on deep kernel learning*. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan. https://ieeexplore.ieee.org/document/9689465

[17]Zahran, A. I., Fahmy, A. A., Wassif, K. T., Bayomi, H. (2023). Fine-tuning self-supervised learning models for end-to-end pronunciation scoring. *IEEE Access, 11*, 112650–112663. https://doi.org/10.1109/ACCESS.2023.3317236

[18]Sancinetti, M., Vidal, J., Bonomi, C., Ferrer, L. (2022). *A transfer learning approach for pronunciation scoring*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore. https://doi.org/10.1109/ICASSP43922.2022.9747727

[19]Du, L. (2024). *Evaluation of English pronunciation interaction quality based on deep learning*. International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India. https://doi.org/10.1109/ICICACS60521.2024.10498572

[20]Pei, H.-C., Fang, H., Luo, X., Xu, X.-S. (2023). Gradformer: A framework for multi-aspect multi-granularity pronunciation assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 554–563. https://doi.org/10.1109/TASLP.2023.3335807

[21]Zhu, C., Wumaier, A., Wei, D., Fan, Z., Yang, J., Yu, H., Kadeer, Z., Wang, L. (2024). Pronunciation error detection model based on feature fusion. *Speech Communication, 156*, 1-12. https://doi.org/10.1016/j.specom.2023.103009

[22]Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., Kostek, B. (2022). Computer-assisted pronunciation training—Speech synthesis is almost all you need. *Speech Communication, 142*, 1–12. https://doi.org/10.1016/j.specom.2022.06.003

[23]Masyhur, A. M., Andriyani, A. D., Sakinah, N., Nafisyah, D., Windarwan, A. M., Riyadi, S. (2023). *Android-based English vocabulary learning media using speech recognition and augmented reality*. International Workshop on Artificial Intelligence and Image Processing (IWAIIP), Yogyakarta, Indonesia. https://doi.org/10.1109/IWAIIP58158.2023.10462739

[24]Kostyuchenko, E., Rakhmanenko, I., Lapina, M. (2021). *Evaluation of a method for measuring speech quality based on an authentication approach using a correlation criterion*. 17th International Conference on Intelligent Environments (IE), Dubai, United Arab Emirates. https://doi.org/10.1109/IE51775.2021.9486435

[25]Wang, H., Xu, J., Ge, H., Wang, Y. (2019). *Design and implementation of an English pronunciation scoring system for pupils based on DNN-HMM*. 10th International Conference on Information Technology in Medicine and Education (ITME), Qingdao, China. https://doi.org/10.1109/ITME.2019.00085

[26]Yu, C. (2024). *Application of artificial intelligence and speech recognition technology in English listening and speaking ability assessment system*. 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII), Melbourne, Australia. https://doi.ieeecomputersociety.org/10.1109/ICMIII62623.2024.00070

[27]Sheoran, K., Bajgoti, A., Gupta, R., Jatana, N., Dhand, G., Gupta, C., Dadheech, P., Yahya, U., Aneja, N. (2023). Pronunciation scoring with goodness of pronunciation and dynamic time warping. *IEEE Access, 11*, 15485–15495. https://doi.org/10.1109/ACCESS.2023.3244393

[28]Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., Gonzalez-Ferreras, C., Cardeñoso Payo, V. (2020). Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies, 13* (2), 269–282. https://doi.org/10.1109/TLT.2020.2980261

[29]Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., Mckechnie, J., Ahmed, B., Gutierrez-Osuna, R. (2021). A longitudinal evaluation of tablet-based child speech therapy with Apraxia World. *ACM Transactions on Accessible Computing (TACCESS), 14* (1), 1–26. https://doi.org/10.1145/3433607

[30]Gómez González, M. A., Ferreiro, A. L. (2024). Web-assisted instruction for teaching and learning EFL phonetics to Spanish learners: Effectiveness, perceptions and challenges. *Computers and Education Open, 7*, 1-17. https://doi.org/10.1016/j.caeo.2024.100214

[31]Tejedor-García, C., Escudero-Mancebo, D., Cardeñoso-Payo, V., González-Ferreras, C. (2020). Using challenges to enhance a learning game for pronunciation training of English as a second language. IEEE Access, 8, 74250–74266. https://doi.org/10.1109/ACCESS.2020.2988406