Tesis Doctoral por Compendio de Publicaciones

Detección automática de la hipernasalidad en pacientes con habla patológica

Andrés Lozano Durán

Directores:

Enrique Nava Baro

Ignacio Moreno-Torres Sánchez

Tutor:

Pablo Otero Roth

Escuela Técnica Superior de Ingeniería Telecomunicación

Programa de Doctorado en Ingeniería de Telecomunicación

Departamento Ingeniería de Comunicaciones

Universidad de Málaga 2025

AUTOR: Andrés Lozano Durán

https://orcid.org/0000-0003-0444-0452

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D. Andrés Lozano Durán, estudiante del programa de doctorado Ingeniería de Telecomunicación de la Universidad de Málaga, autor de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada:

Detección automática de la hipernasalidad en pacientes con habla patológica

Realizada bajo la tutorización de Pablo Otero Roth y dirección de Enrique Nava Baro e Ignacio Moreno-Torres Sánchez.

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 25 de junio de 2025

Fdo.: Andrés Lozano Durán Doctorando	Fdo.: Pablo Otero Roth Tutor
Fdo.: Enrique Nava Baro Directores de tesis	Ignacio Moreno-Torres Sánchez

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

INFORME DE IDONEIDAD DE LA PRESENTACIÓN POR COMPENDIO DE ARTÍCULOS

El Dr. Enrique Nava Baro, profesor titular del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga, como codirector; junto al Dr. Ignacio Moreno-Torres Sánchez, profesor titular del Departamento de filología española, italiana, románica, teoría de la literatura y literatura comparada y ciencias y técnicas historiográficas de la Universidad de Málaga, como codirector; junto al Dr. Pablo Otero Roth, profesor titular del Departamento de Ingeniería de Comunicaciones de la Universidad de Málaga, como tutor de la tesis doctoral titulada "**Detección automática de la hipernasalidad en pacientes con habla patológica**", presentada por el doctorando Andrés Lozano Durán.

Informan de la idoneidad de la presentación por compendio de artículos, dado que cumple con todos los criterios establecidos para ello por la Universidad de Málaga. Los trabajos que la componen son los siguientes:

- **Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically?**
  Moreno-Torres I, **Lozano A**, Nava E, Bermúdez-de-Alvear R. Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically? Applied Sciences. 2021; 11(19):8809. https://doi.org/10.3390/app11198809.

- **Unmasking Nasality to Assess Hypernasality**
  Moreno-Torres I, **Lozano A**, Bermúdez R, Pino J, Méndez MDG, Nava E. Unmasking Nasality to Assess Hypernasality. Applied Sciences. 2023; 13(23):12606. https://doi.org/10.3390/app132312606.

- **Computing nasalance with Convolutional Neural Networks**
  Lozano, A, Nava, E, Méndez, MDG, & Moreno-Torres, I. (2024). Computing nasalance with MFCCs and Convolutional Neural Networks. PLOS ONE, 19(12), e0315452. https://doi.org/10.1371/journal.pone.0315452

En Málaga a 25 de Junio del 2025

Fdo.: ERIQUE NAVA BARO Director de tesis	Fdo.: IGNACIO MORENO-TORRES SÁNCHEZ Director de tesis	Fdo.: PABLO OTERO ROTH Tutor

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

A Irene

A Andrés y Mateo

Agradecimientos

Quiero agradecer a todas las personas que han contribuido a que mi etapa como estudiante de doctorado concluya con la redacción de esta memoria.

En primer lugar, quiero dar las gracias de manera muy especial a los doctores Enrique Nava Baro e Ignacio Moreno-Torres Sánchez, directores de esta memoria, por su ayuda y dedicación. A Enrique, por su confianza durante todo el proceso, y por su inestimable orientación para que esta memoria viera por fin la luz. Y a Ignacio, por su incansable implicación en este trabajo, su generosa guía y por compartir conmigo su conocimiento y experiencia. Gracias por la confianza depositada en mí desde el primer día, por la paciencia y por la dedicación que le habéis puesto.

También agradecer al doctor Pablo Otero Roth, tutor de esta tesis, por acompañarme en este recorrido académico con su experiencia, su apoyo, y sus acertadas sugerencias.

Gracias a todos los compañeros y compañeras con los que he tenido el placer de trabajar durante estos años: Rosa Bermúdez, Wanda Meschian, María Cristina Armero, Josué Pino, María Dolores García, María Jesús Castillo y Francisco Sendra. Habéis sido una parte esencial para llegar hasta aquí.

A mi familia, por su esfuerzo durante los años de universidad, másteres y doctorado, por su apoyo y su interés en que hoy exista esta memoria.

Especialmente agradecido a Irene, mi mujer, mi apoyo durante éste y todos los caminos de mi vida, por su paciencia, por sus consejos, por escucharme y soportarme, por estar ahí y animarme a no abandonar. Y a mis hijos, Andrés y Mateo, que vinieron para cambiar mi vida completamente, gracias por existir.

Índice

Tabla de ilustraciones

Prefacio

En cumplimento con los requisitos especificados en el reglamento de doctorado de la Universidad de Málaga, la presente tesis doctoral ha sido autorizada por los directores de tesis y el órgano responsable del programa de doctorado para ser presentada en el formato de compendio de publicaciones.

Los artículos presentados en esta tesis doctoral abarcan el estudio de la detección automática de la hipernasalidad en pacientes con habla patológica. Se explora la utilización de diferentes tipos de muestras de habla para la detección de la hipernasalidad. Además, se calculan diversos descriptores de audio para entrenar sistemas de clasificación automáticos basados en métodos estadísticos y computacionales. El objetivo último es resolver un problema de clasificación, categorizando las muestras de sonido como sano o hipernasal, de manera que la clasificación obtenida mediante el sistema tenga una alta correlación con el diagnóstico emitido por especialistas en logopedia.

Las referencias de los artículos en los que el doctorando figura como primer o segundo autor, y que avalan la presente tesis doctoral, se detallan a continuación de acuerdo con su orden cronológico de publicación. Todos los trabajos incluyen código DOI para el acceso en abierto.

- **Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically?** (Moreno-Torres et al., 2021)
  Moreno-Torres I, **Lozano A**, Nava E, Bermúdez-de-Alvear R. Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically? Applied Sciences. 2021; 11(19):8809.
  Journal impact factor: 2.8. CiteScore (Scopus): 3.7 (81/300 in Engineering, General Engineering). Segundo cuartil (Q2).
  DOI: https://doi.org/10.3390/app11198809.

- **Unmasking Nasality to Assess Hypernasality** (Moreno-Torres et al., 2023)
  Moreno-Torres I, **Lozano A**, Bermúdez R, Pino J, Méndez MDG, Nava E. Unmasking Nasality to Assess Hypernasality. Applied Sciences. 2023; 13(23):12606.
  Journal impact factor: 2.7. CiteScore (Scopus): 4.5 (73/302 in Engineering, General Engineering). Primer cuartil (Q1).
  DOI: https://doi.org/10.3390/app132312606.

- **Computing nasalance with Convolutional Neural Networks** (Lozano et al., 2024)
  Lozano, A, Nava, E, Méndez, MDG, & Moreno-Torres, I. (2024). Computing nasalance with MFCCs and Convolutional Neural Networks. PLOS ONE, 19(12), e0315452.

Journal impact factor: 2.9. CiteScore (Scopus): 6.2 (18/171 in Multidisciplinary). Primer cuartil (Q1).
DOI: https://doi.org/10.1371/journal.pone.0315452

Esta tesis se estructura en dos bloques. En el primero de ellos se presenta el marco teórico utilizado para el desarrollo de los sistemas de detección automática de la hipernasalidad en pacientes con habla patológica. En el segundo se desarrollan los principales resultados obtenidos en los artículos que forman parte integrante de la tesis junto con una discusión de estos y las principales conclusiones.

Así, en el capítulo 1 se presentan los objetivos de la tesis y los antecedentes científicos y técnicos del problema, así como una breve descripción de la solución propuesta. En el capítulo 2 se presentan los fundamentos del habla hipernasal, incluyendo la producción normal del habla y los trastornos de la resonancia y alteración velofaríngea. A continuación, en el capítulo 3 se presentan los fundamentos teóricos de los descriptores acústicos calculados. En el capítulo 4 se definen los algoritmos empleados para la clasificación, *Random Forest* (RF) y *Support Vector Machine* (SVM). En el capítulo 5 se describen los tres tipos de redes neuronales utilizadas en los trabajos, Redes Neuronales Artificiales (ANN), Redes Neuronales Profundas (DNN) y Redes Neuronales Convolucional (CNN). En el capítulo 6 se definen las métricas empleadas para la evaluación de los resultados junto con los tipos de error de los sistemas de clasificación.

En el capítulo 7 presenta un resumen de los principales resultados obtenidos en los artículos que forman parte integrante de la tesis, destacando las aportaciones más relevantes de cada uno de ellos. El capítulo 8 desarrolla una discusión global de los resultados presentados anteriormente. Esta memoria finaliza en el capítulo 9 con las principales contribuciones de la tesis y las líneas de trabajo futuro.

En el apéndice A se incluye una copia de los trabajos que forman parte integrante de la tesis.

1. Introducción

En este apartado se presenta una breve descripción de la hipernasalidad en el habla. La hipernasalidad es una condición del habla en la que existe un aumento anormal de la resonancia nasal durante la producción de sonidos orales (Kummer, 2011; Mossey & Catilla, 2003). La hipernasalidad aparece debido a una insuficiencia velofaríngea (Kummer, 2013), esto es, a un funcionamiento indebido del esfínter velofaríngeo. En el hablante sano, el esfínter velofaríngeo se encarga de que la corriente de aire fluya por la cavidad oral durante la producción de vocales y consonantes orales, y por ambas cavidades (oral y nasal) durante la producción de sonidos nasales. Cuando el hablante no logra controlar el flujo de aire se producen diferentes efectos; el más llamativo ocurre cuando este flujo de aire se dirige a la cavidad nasal al intentar producir sonidos orales; ello provoca patrones de error fonológico en las consonantes (por ejemplo, / b / > / m /, / d / > / n /) y/o vocales nasalizadas (/ ã /, / ẽ /, / ĩ /, / õ /, / ũ /). La hipernasalidad se observa con frecuencia en pacientes con labio leporino y/o fisura palatina; en grupos de pacientes que tienen un velo corto que no pueden lograr un contacto completo con la pared faríngea posterior (por ejemplo, los que presentan un síndrome de deleción 22q11.2; 22q11.2DS (Solot et al., 2019)); y en pacientes con un retraso en el desarrollo de habilidades motoras finas necesarias para controlar la bajada y subida del velo (Kuehn & Moller, 2000). La correcta evaluación de la hipernasalidad es fundamental a la hora de tomar decisiones clínicas, orientar el proceso de rehabilitación, y planificar intervenciones eficaces, especialmente en niños (Lohmander & Olsson, 2004).

Tradicionalmente, la evaluación de la hipernasalidad se ha realizado de forma perceptual por uno o varios logopedas, los cuales asignaban una medida subjetiva del grado de nasalidad en el paciente (por ejemplo, en el protocolo CAPS-A (John et al., 2006) el grado de nasalidad se mide mediante una escala 0-3). A pesar de su carácter subjetivo y, por tanto, variable, este enfoque sigue considerándose el criterio de referencia para evaluar la hipernasalidad (Bettens et al., 2014). Sin embargo, la evaluación perceptual de la hipernasalidad es una tarea exigente y compleja, que requiere de una formación muy especializada en logopedia. Parte de la dificultad de este proceso se debe a que la nasalidad es un fenómeno gradual, más que categórico; y al hecho de que los hablantes sanos nasalizan en cierto grado algunas vocales (como resultado de coarticulación de las secuencias de vocal seguida de consonante nasal, o consonante nasal seguida de vocal, y debido a la transmisión transpalatal de energía acústica de la cavidad oral a la nasal (Gildersleeve-Neumann & Dalston, 2001)).

Además, en español las vocales nasales no existen como fonemas, lo que significa que la mayoría de los hablantes tienen dificultades para reconocer las vocales nasales como una categoría diferenciada de las vocales orales. Por último, la hipernasalidad patológica suele coincidir con otros errores, como los compensatorios, y/o baja inteligibilidad, lo que puede dificultar aún más su identificación (Mathad et al., 2021). Esta incertidumbre ha motivado a los investigadores a desarrollar herramientas objetivas de evaluación de

la hipernasalidad para apoyar la evaluación subjetiva. Una herramienta importante es la nasometría (Fletcher et al., 1974), que se propuso en el último cuarto del siglo XX, pero sigue siendo un método bien considerado entre los logopedas especialistas. Para utilizar esta técnica se emplea un instrumento acústico, el nasómetro, que cuenta con un par de micrófonos separados por una placa, de manera que se registran la señal oral y la señal nasal por separado. A partir de estas grabaciones de doble canal es posible calcular la nasalancia, que es la relación entre la energía de la nariz y la suma de la señal de la boca y la nariz, en una pequeña banda de 300 Hz centrada en torno a 600 Hz (Bettens et al., 2014; Gildersleeve-Neumann & Dalston, 2001). El nasómetro ha tenido un éxito relativo en la práctica clínica (Bettens et al., 2014; Gildersleeve-Neumann & Dalston, 2001), debido principalmente a algunas características importantes: 1) es una herramienta no invasiva, gracias a lo cual puede ser empleada por logopedas (no es necesaria la intervención de un médico); 2) proporciona una puntuación porcentual intuitiva que es fácilmente interpretable por un experto clínico; y 3) puede utilizarse en cualquier idioma o dialecto, y para cualquier enunciado de interés para el logopeda. Sin embargo, la fiabilidad de este instrumento no está clara; los estudios que comparan las puntuaciones de nasalancia y las puntuaciones perceptuales generadas por logopedas han obtenido correlaciones que oscilan entre no significativas y fuertes (Liu et al., 2022). De hecho, desde la perspectiva actual, la nasalancia parece claramente limitada, tanto desde el punto de vista perceptivo como técnico.

Desde el punto de vista de la percepción del habla, la información acústica utilizada para calcular la nasalancia es sólo una pequeña parte de la información a la que tiene acceso el oyente humano. Por ejemplo, la nasalización vocálica puede afectar al espectro hasta 3 kHz o más (Carignan, 2018), un valor mucho mayor que la pequeña banda utilizada tradicionalmente para calcular la nasalancia (Fletcher et al., 1974). Además, mientras que la nasometría se basa únicamente en la cantidad de energía de la señal del habla, los humanos pueden apoyarse en muchos otros datos acústicos, como por ejemplo la estructura espectral (Carignan, 2018). Esto puede explicar, en parte, la amplia variación en los estudios que han calculado la correlación entre la nasalancia y la nasalidad perceptual (con valores que oscilan entre 0.88 y 0.42), véase (Brancamp et al., 2010; Keuning et al., 2002; Khwaileh et al., 2018).

Desde una perspectiva técnica, la investigación en el último medio siglo ha propuesto múltiples características acústicas que han demostrado ser altamente eficaces en el procesamiento del habla (Rabiner & Juang, 1993; Rabiner & Schafer, 1978). Para la detección de la hipernasalidad se han propuesto diferentes descriptores basados en el análisis espectral, como son: la amplitud de los formantes F1, F2 y F3 y sus anchos de banda, y pares polo/cero (Glass & Zue, 1985; Kataoka et al., 2001; Vijayalakshmi et al., 2009; Vijayalakshmi et al., 2007; Yu & Barkana, 2009). En general, estas características son muy dependientes del tipo de patología analizada (Orozco-Arroyave et al., 2015). Dada la alta variabilidad de los patrones vocales en enfermedades o lesiones neurológicas, dichos descriptores no ofrecen una opción robusta para analizar la hipernasalidad, cuyos patrones acústicos son muy variables y pueden coincidir con la aparición de otros fenómenos del habla.

Existen a su vez diversas técnicas de comparación de patrones de voz, como pueden ser: *Log-Spectral Distance* (LSD), *Cepstral Distance* (CD), e *Itakura-Saito Distortion* (ISD), que cuantifican la diferencia espectral entre señales completas (Rabiner & Juang, 1993). Estas medidas son muy efectivas cuando se dispone de frases de referencia estrictamente controladas, es decir, misma locución, duración similar y alineación temporal, porque calculan la distancia punto a punto del espectro o del cepstrum a lo largo de toda la señal. Sin embargo, no resultan adecuadas para el presente trabajo, ya que se analizan diversos enunciados de longitud variable y con contenido fonético diverso. Bajo estas condiciones, las distorsiones calculadas reflejan tanto diferencias de contenido lingüístico como posibles alteraciones nasales, impidiendo aislar de forma fiable el efecto de la hipernasalidad. Por ello, se descartan estas métricas globales y se priorizan descriptores locales que puedan compararse de manera robusta entre emisiones fonéticamente variadas.

En consecuencia se introducen descriptores que realizan una transformación espectral, como pueden ser *Mel-Frequency Cepstral Coefficients* (MFCC) (Davis & Mermelstein, 1980), características glotales como el *jitter* y *shimmer* (Castellanos et al., 2006; Dubey et al., 2018), *Vowel Space Area* (VSA) (Kalita et al., 2017), y medidas no lineales (Orozco-Arroyave et al., 2012; Orozco-Arroyave et al., 2013). Cabe destacar los descriptores MFCC, ya que se utilizan ampliamente en los trabajos que avalan esta tesis debido a su capacidad para modelar la percepción del habla en humanos y a su amplio uso en sistemas de reconocimiento automático del habla.

Todas estas consideraciones han motivado a los investigadores a desarrollar sistemas de clasificación automática de la hipernasalidad entrenados con diferentes características acústicas (Bettens et al., 2014; Dhillon et al., 2021). Los resultados, en términos de precisión, de los sistemas de clasificación de la hipernasalidad basados únicamente en información acústica son generalmente excelentes, aunque la mayoría de estos estudios han utilizado únicamente un número limitado de tipos de enunciado, como las vocales sostenidas (Akafi et al., 2013; Dubey et al., 2018; L. He et al., 2015b; Lee et al., 2006; Mirzaei & Vali, 2016; Wang, Yang, et al., 2019). Esto es claramente incompatible con las recomendaciones clínicas estándar, que recomiendan que los pacientes sean evaluados utilizando una variedad de fonemas y enunciados con complejidad variable, por ejemplo, siguiendo el protocolo CAPS-A (John et al., 2006; Kummer, 2016)).

En cuanto a la diversidad fonémica, el protocolo CAPS-A identifica tres niveles de nasalización, según los segmentos que se nasalizan: 1) leve: nasalización evidente sólo en vocales cerradas (/ i /, / u /); 2) moderado: nasalización observable en vocales cerradas y abiertas (/ a /, / e /, / i /, / o /, / u /); y 3) grave: nasalización observable en todas las vocales y en consonantes sonoras (por ejemplo: / b /, / d /, / g /). Además, en (Kummer, 2016) se propone que para estudiar la nasalización se utilicen series silábicas que incluyan las oclusiones sordas como / p /, / t /, / k /. En cuanto a la complejidad de los enunciados, se suele insistir en que se necesitan diferentes tipos de enunciados para proporcionar información suficiente durante la evaluación: vocales aisladas, palabras, secuencias de repetición de sílabas, frases y habla espontánea (Grunwell, 2000; L. He et

al., 2015a; Henningsson et al., 2008; John et al., 2006; Kummer, 2016; Sell et al., 2009; Spruijt et al., 2018).

Sin embargo, Kummer et al. (Kummer, 2016) consideran que dos de estas tareas son especialmente útiles: por un lado, la repetición de secuencias silábicas como / ta ta ta.../, que permiten aislar los fonemas individuales y eliminar los efectos del contexto; por otro, las frases que contienen múltiples producciones de la misma colocación de fonemas, que permiten evaluar la presencia de la emisión nasal en un entorno de habla conectado. Así pues, según los expertos clínicos, la evaluación de la hipernasalidad debe basarse no sólo en las vocales aisladas, sino también en una variedad de consonantes sonoras y no sonoras que se combinan en diferentes tipos de enunciados.

Los sistemas de aprendizaje automático se utilizan actualmente en una amplia variedad de tareas y dominios, evidenciando su naturaleza interdisciplinaria. Por ejemplo, en medicina se han desarrollado algoritmos de machine learning para apoyar el diagnóstico asistido (Esteva et al., 2017). En el ámbito financiero, técnicas de machine learning permiten pronosticar tendencias de los mercados y detectar fraudes con mayor eficacia (Mienye et al., 2024). De igual forma, en visión por computador las redes neuronales convolucionales han alcanzado rendimientos sobresalientes en reconocimiento de imágenes (He et al., 2016). Asimismo, en el sector industrial y logístico, el machine learning impulsa la automatización de procesos complejos, desde el mantenimiento predictivo de maquinaria hasta la optimización de cadenas de suministro (Rai et al., 2021). Estos ejemplos ilustran el carácter transversal de los clasificadores basados en machine learning y su aporte significativo en múltiples campos de la ingeniería.

Existen diversos estudios que emplean algoritmos de aprendizaje automático para clasifica la señal de voz como nasal u oral en función de descriptores complejos del habla (Carignan, 2021; Dhillon et al., 2021; McKechnie et al., 2018). Los algoritmos de clasificación comúnmente empleados incluyen *Hidden Markov Model* (HMM), *Gaussian Mixture Model* (GMM), *Random Forest* (RF), *Support Vector Machine* (SVM), Redes Neuronales Profundas (DNN), o Redes Neuronales Convolucionales (CNN) (Golabbakhsh et al., 2017; Spruijt et al., 2018; Travieso et al., 2017; Vikram et al., 2018; Wang, Yang, et al., 2019).

En el presente trabajo se emplea una aproximación al problema de clasificación del habla hipernasal basa en dividir la señal de entrada en pequeños fragmentos, y se calcula una serie de descriptores por cada uno de ellos para alimentar un sistema de clasificación automática. Es decir, no se entrena el sistema con la señal de audio recogida por el sistema de grabación, sino que se aplica un proceso de extracción de descriptores, los cuales permiten modelar los datos para el proceso de clasificación. Debido a este proceso, se emplea principalmente clasificadores basados en *deep learning*, DNN y CNN, ya que ofrecen los mejores resultados para el análisis de la hipernasalidad junto con una mayor correlación con los resultados de la evaluación perceptual realizada por los logopedas (Zhang et al., 2023). Se emplean a su vez algoritmos de clasificación tradicionales como son RF y SVM para realizar una comparación con los resultados. Otros sistemas, como HMM se emplean generalmente cuando se analizan secuencia de datos

largas, por ejemplo frases completas, o se quiere capturar la transición temporal entre fonemas o sonidos, por lo que no están tan ajustados al propósito del presente trabajo.

Cabe destacar que estos sistemas se entrenan con el mismo tipo de información que esperan evaluar, pudiendo detectar la nasalidad únicamente en los mismos tipos de enunciados utilizados, y por tanto, pueden carecen de la flexibilidad para analizar diferentes idiomas, o incluso en el mismo idioma, diferentes dialectos.

Sólo unos pocos estudios han utilizado enunciados complejos para evaluar automáticamente la nasalidad (Golabbakhsh et al., 2017; Orozco-Arroyave et al., 2012; Vikram et al., 2018). Golabbakhsh et al. (Golabbakhsh et al., 2017) utilizaron seis oraciones que contenían consonantes oclusivas y fricativas, que los logopedas utilizan habitualmente para evaluar la calidad del habla. Los autores entrenaron un clasificador SVM con un conjunto de características acústicas que se calculan para cada expresión. En el mejor caso, la precisión alcanzó el 85% con una sensibilidad del 82% y una especificidad del 85%. Orozco-Arroyave et al. (Orozco-Arroyave et al., 2012) analizan una base de datos de enunciados de 108 niños sanos y 128 hipernasales de habla española. Todos los niños produjeron las cinco vocales sostenidas del español junto con dos palabras, una con consonantes no sonoras (por ejemplo, / koko /) y otra con una consonante sonora y otra no sonora (por ejemplo, / gato /). Entrenaron un clasificador SVM utilizando características no lineales junto con un conjunto de seis medidas de entropía. Los resultados fueron los mejores para las vocales / a /, / i /, / e /, / o / y para la palabra / gato / (es decir, la que tiene una consonante sonora); los resultados más pobres se observaron con la vocal / u / y la palabra / koko /. Sin embargo, los resultados mejoraron cuando se seleccionaron los mejores descriptores de cada vocal (precisión: 91%; sensibilidad: 93-95%; especificidad: 88-90%). En conjunto, estos resultados indican que es posible evaluar la nasalidad de forma automática utilizando expresiones más complejas que los sonidos sostenidos, y también combinando diferentes muestras de habla del mismo hablante.

En resumen, parece haber un desajuste entre los estudios clínicos, por un lado, que hacen hincapié en la importancia de explorar una variedad de sonidos del habla y tipos de enunciados, y la investigación del análisis automático de la hipernasalidad, por otro, que se ha centrado principalmente en las vocales sostenidas o en un número limitado de palabras u oraciones. Una razón obvia por la que la mayoría de los estudios técnicos han utilizado vocales sostenidas es porque en ese caso el espectro es estable a lo largo de una ventana relativamente larga, lo que aumenta la probabilidad de detectar los efectos acústicos relativamente pequeños introducidos por la resonancia nasal. Si se consideran expresiones más complejas (por ejemplo, frases completas), se tiene una señal con mayor variabilidad, lo que puede difuminar los efectos locales de la nasalidad. Sin embargo, como han demostrado Orozco-Arroyave et al. (Orozco-Arroyave et al., 2012) y otros, al menos en algunos casos el espectro medio puede servir para detectar la hipernasalidad. De hecho, parece razonable especular que los efectos de la hipernasalidad podrían ser medibles en los mismos tipos de enunciados en los que los humanos perciben la hipernasalidad con relativa facilidad (por ejemplo, sílabas

repetidas, frases con consonantes sonoras, etc. (Kummer, 2016)), y que los resultados podrían mejorarse combinando múltiples enunciados. Queda por aclarar hasta qué punto se pueden confirmar estas especulaciones.

1.1 Sistema propuesto

El enfoque general propuesto en el presente trabajo se puede observar de forma esquemática en la Figura 1. En primer lugar se toma la señal de audio grabada a los locutores y se extraen los descriptores de audio de cada enunciado. A continuación, se entrenan los diferentes modelos de clasificación automática de la hipernasalidad y se realiza una clasificación de los datos de entrada. Finalmente, se compara la predicción realizada por el modelo con el diagnóstico realizado por los especialistas en logopedia. En este trabajo, el diagnóstico del especialista se considera el *true class*, es decir, se toma como certeza el dictamen del logopeda. Por lo tanto el objetivo principal es que el modelo ofrezca resultados con la mayor correlación posible con el diagnóstico humano.

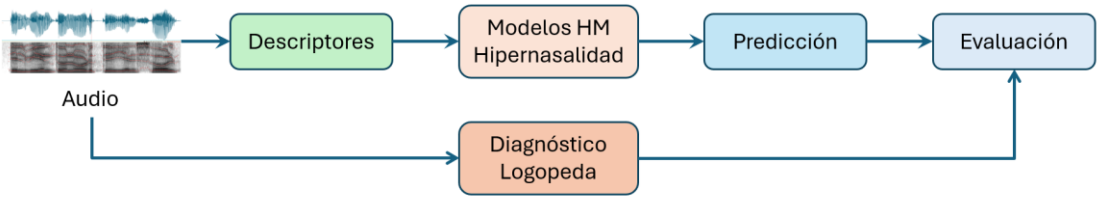

Figura 1. Resumen general del enfoque propuesto para la predicción de la hipernasalidad.

Cabe destacar que cada uno de los trabajos que avalan esta tesis presentan diferencias en cuanto al proceso de extracción de parámetros o la salida del clasificador, en función de la aproximación al problema empleada.

Los artículos que abarca esta tesis estudian diversos enfoques para la detección automática de la hipernasalidad en el habla patológica. En primer lugar, se explora la viabilidad de detectar la hipernasalidad basándose en muestras de habla distintas de las vocales sostenidas, ampliando así el tipo de enunciados utilizables para este fin. A continuación, se examina si las señales provenientes exclusivamente de la nariz o de la boca pueden aumentar la precisión de la evaluación de la hipernasalidad frente a la evaluación cuando se utilizan señales monofónicas, que es el resultado de combinar las dos anteriores en un único canal de audio. Finalmente, en el último trabajo se propone un nuevo enfoque para calcular la nasalancia mediante el uso de una red CNN entrenada con coeficientes MFCC.

2. Hipernasalidad en el habla

En este capítulo se desarrolla la hipernasalidad en el habla desde un punto de vista clínico. La resonancia nasal excesiva provoca hipernasalidad, un trastorno de la resonancia que es frecuente en determinados grupos de pacientes atendidos en las consultas de logopedia de los hospitales. La hipernasalidad puede influir negativamente en la inteligibilidad y en la percepción que el oyente tiene de una persona que habla con voz hipernasal. La resonancia nasal excesiva suele estar asociada a la presencia de fisura palatina, a la disartria o a la hipoacusia, pero también existen trastornos de la resonancia nasal sin origen conocido.

2.1 Producción normal del habla

Para comprender la naturaleza de un trastorno del habla es preciso describir primero la producción normal del habla. La base de la producción de la voz es la exhalación controlada del aire de los pulmones. Antes de que el aire salga del cuerpo, este pasa por tres subestructuras que le otorgarán sus propiedades acústicas fundamentales, como muestra la Figura 2. En primer lugar, el aire pasa por la laringe, en donde se encuentra el aparato fonador, formado por una membrana (cuerdas vocales) que puede vibrar o bien permanecer estáticas. En el primer caso se produce una fuente sonora periódica, mientras que en el segundo caso, un sonido aperiódico.

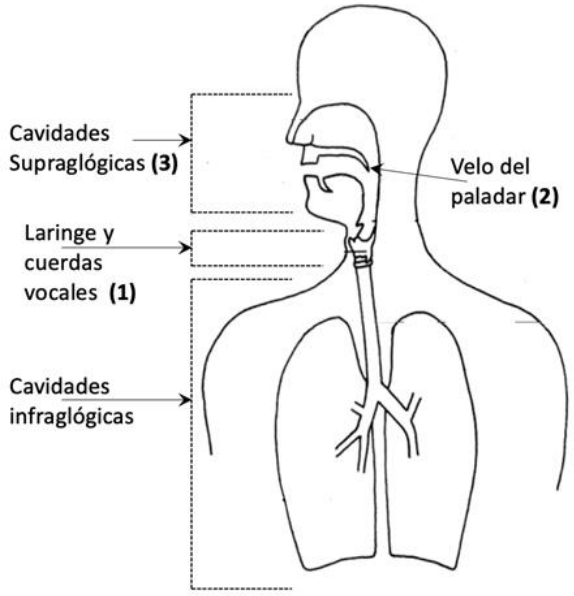

Figura 2. Principales elementos del sistema de producción del habla.

En segundo lugar, se encuentra el velo del paladar, también denominado paladar blando, cuya función es determinar si el aire se dirige exclusivamente hacia la cavidad oral, o si, por el contrario, una parte de la corriente aérea es dirigida hacia la cavidad nasal. Por

defecto el velo está levantado, lo que resulta en emisiones exclusivamente orales, en las que apenas hay emisión de sonido por la nariz. En la Figura 3 se puede observar la emisión normal de una consonante / s / sostenida en el tiempo producida por un hablante sano, donde la señal oral presenta mayor energía que la señal nasal durante

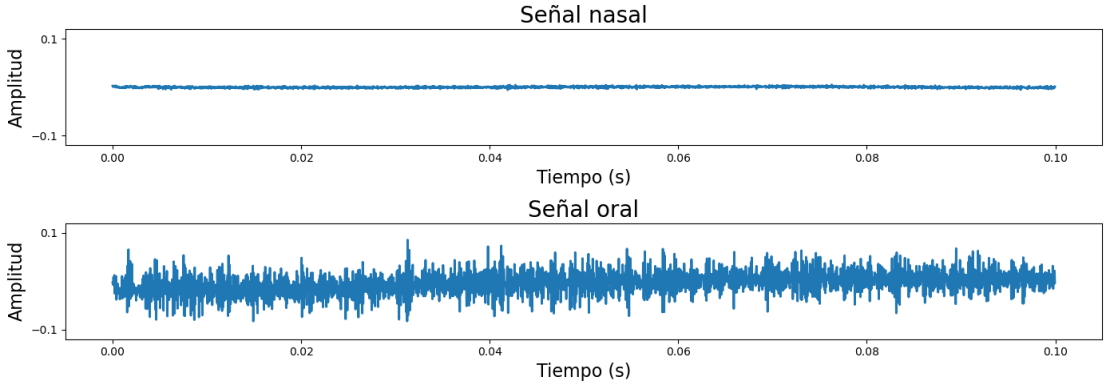

Figura 3. Producción de la consonante / s / sostenida en el tiempo de un hablante sano.

toda la emisión del sonido.

En español, sólo las consonantes nasales / m /, / n /, y / ŋ /[1] (“ñ”), se producen con el velo del paladar en la posición más baja. Los demás sonidos del habla (todas las vocales y las consonantes restantes) son sonidos orales.

En tercer lugar, la corriente de aire pasa por las cavidades supraglóticas, cuya función se caracteriza como un conjunto de filtros que modifican la estructura espectral, aunque también modifican la envolvente de la señal sonora. Es importante destacar que para que la función de la cavidad oral sea efectiva es preciso una presión intraoral elevada, lo que se consigue con el paladar blando levantado, que obliga a la corriente de aire exhalada desde los pulmones a tomar el camino únicamente a través de la boca, como muestra la Figura 4.

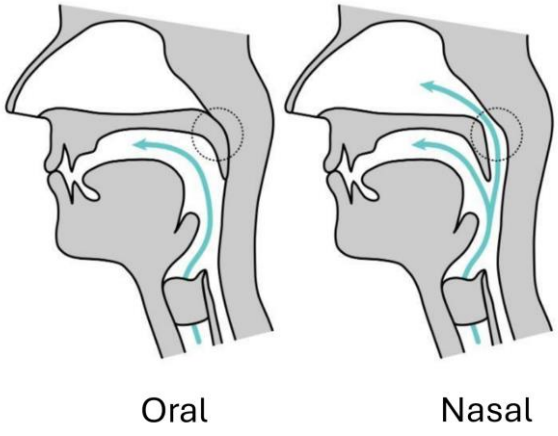

Figura 4. Producción de sonido con el velo cerrado (izquierda), y abierto (derecha). Fuente: (Errasti Aguirrebeitia, 2024).

[1] En el presente trabajo se emplean los símbolos del Alfabeto Fonético Internacional (IPA) para denotar los fonemas, que no siempre coinciden con los caracteres del español.

Hablamos a un ritmo de 12-14 sonidos por segundo (Bradley, 1995), lo que significa que cada vez que una persona habla, se combinan muchos movimientos pequeños pero rápidos en patrones de movimiento complejos. Para cada sonido del habla, los articuladores tienen que cambiar según el lugar y el modo de articulación para ese sonido en particular. También se sabe que el velo se eleva más en las consonantes que en las vocales, y más en las vocales altas que en las bajas (Bell-Berti, 1993). Aun así, cuando una vocal se encuentra junto a una consonante nasal, suele ocurrir que la vocal resulta nasalizada en parte o en su totalidad. Esta nasalización ocurre debido al fenómeno llamado coarticulación, el cual consiste en que no hay una separación nítida entre dos sonidos consecutivos, sino que el uno tiene propiedades del otro. En el caso de las vocales previas o posteriores a consonantes nasales, lo que ocurre es que para emitir la consonante nasal el hablante debe descender y luego subir el velo, y la operación de descenso y la de ascenso tiene lugar mientras se está emitiendo la vocal (Warren et al., 1993). De esta forma, aunque en español no existen vocales nasales, estas se nasalizan en presencia de las consonantes nasales / m /, / n / y / ŋ /.

2.2 Factores que pueden llevar a la aparición de la hipernasalidad

La fisura palatina es una malformación congénita debida a un cierre incompleto del labio y/o el paladar durante el desarrollo fetal temprano. Cuando la patología afecta únicamente a la región del labio, esta se denomina labio leporino. Estos hablantes presentan una hendidura en el tejido del labio y/o en el tejido y el hueso del paladar, como se observa en la Figura 5. Una hendidura del labio y/o del paladar puede afectar a la alimentación, el habla, el desarrollo dental, el desarrollo de la mandíbula y la audición, y tiene unos claros efectos visuales. El impacto puede ser tanto funcional como estético. Por sus efectos en la nasalidad, este trabajo se centra especialmente en la fisura palatina.

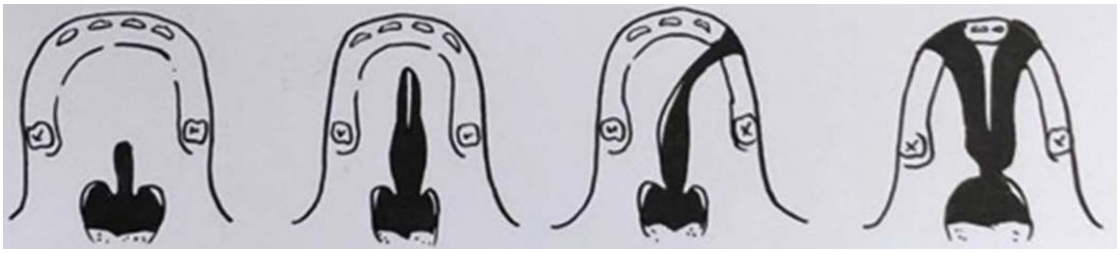

Figura 5. Clasificación de Veau de las fisuras de paladar (Veau & Borel, 1931). De izquierda a derecha: fisura de paladar blando; fisura de paladar duro y blando hasta el agujero incisivo; fisura unilateral completa del paladar primario y secundario; fisura bilateral del paladar primario y secundario.

El labio y el paladar están formados por partes que suelen unirse en las primeras 12 semanas de desarrollo fetal. Algunos niños tienen una hendidura en el labio o en el paladar mientras que otros presentan una hendidura tanto en el labio como en el paladar. La fisura en el paladar puede incluir sólo el paladar blando o tanto el paladar blando como el duro (ambas son partes del paladar secundario). La causa de esta malformación sólo se conoce en parte, pero están implicados tanto factores ambientales como genéticos (Lees, 2001). Existen varios factores de riesgo ambientales conocidos,

como el tabaco, las drogas, el alcohol y los pesticidas. La fisura palatina se presenta con mayor frecuencia como una malformación aislada, pero también puede asociarse a otros defectos congénitos o formar parte de un síndrome. La incidencia de la fisura palatina en España se sitúa entre 0.5 y 1.44 por cada 1000 nacidos vivos (Dehli et al., 2010; Martín de Vicente et al., 2004).

La fisura se trata quirúrgicamente en los primeros años de vida del niño. Existe una gran variación en los esquemas quirúrgicos para la cirugía de las hendiduras; algunos cirujanos cierran el paladar hendido en dos pasos: primero el paladar blando y después el paladar duro, otros cierran el paladar blando y el duro en una misma intervención (Shaw et al., 2001). La alteración velofaríngea que suele ir asociada a la fisura palatina hace que este grupo constituya un grupo numeroso dentro del grupo general de hablantes con alteración velofaríngea que son atendidos en los centros de logopedia. El grupo con fisura palatina es, por tanto, un grupo que suele ser objeto de estudio en los trastornos de la nasalidad. Los trastornos del habla relacionados con esta patología están relacionados principalmente con la hendidura en el paladar secundario, mientras que un labio leporino aislado o una hendidura en el labio y el paladar primario no suelen causar trastornos del habla.

La incidencia de los trastornos del habla en la población con fisura palatina depende del tipo de hendidura, los métodos quirúrgicos, el desarrollo general del niño, etc. Un estudio multicéntrico europeo, el estudio Eurocleft (Grunwell, 2000), concluyó que a la edad de 11 a 14 años la mayoría de los hablantes de un grupo con fisura unilateral (n=131) habían alcanzado un habla aceptable y comprensible. Entre ellos, el 5% presentaba hipernasalidad grave y algo más del 20% hipernasalidad leve. Había algunos trastornos de la articulación, pero la mayoría eran variantes bastante leves. Existe un debate en curso sobre los métodos quirúrgicos y el momento óptimo para realizar la cirugía en fisura palatina. La producción normal del habla por parte del paciente es uno de los principales elementos que guían la toma de decisiones quirúrgicas, de ahí la importancia de contar con métodos de evaluación fiables.

Existe otro conjunto de trastornos neurológicos que pueden provocar un trastorno motor del habla y alterar la resonancia. Dentro de este conjunto se incluyen la disartria, la esclerosis múltiple, o las lesiones producidas por accidentes cerebrovasculares (Thompson & Murdoch, 1995). Las restricciones anatómicas tras un tratamiento quirúrgico contra el cáncer también pueden causar insuficiencia velofaríngea y, por tanto, hipernasalidad en algunos hablantes (Borggreven et al., 2005). Los hablantes con deficiencias auditivas son otro grupo que puede mostrar un habla hipernasal, lo que probablemente se deba a una falta de retroalimentación auditiva (Baudonck et al., 2015).

2.3 Patrones de error asociados a la insuficiencia velofaríngea

La insuficiencia velofaríngea puede producir diversas alteraciones en el habla, entre otras: habla hipernasal, emisión de aire nasal audible (turbulencia), o un debilitamiento de las consonantes debido la disminución de la presión intraoral. El hablante que presenta insuficiencia velofaríngea también tiene dificultades para producir la diferencia entre los sonidos del habla nasales y no nasales según sea necesario.

2.3.1 Habla hipernasalidad

Como se ha indicado anteriormente, la hipernasalidad se manifiesta cuando la resonancia nasal es mayor de lo que se considera aceptable para un sonido dado. En español todos los sonidos nasales, es decir, las tres consonantes nasales, son periódicos. La hipernasalidad afecta a los sonidos periódicos no nasales, que son las vocales, y las consonantes sordas (como / b /, / d /, / g /). La razón detrás de esta mayor resonancia nasal de debe a que el puerto velofaríngeo no se cierra durante la producción de sonidos orales y, en su paso por la cavidad nasal, el sonido adquiere propiedades que asociamos perceptualmente con los sonidos nasales naturales. Esto puede deberse a una mayor apertura, o a dificultades de sincronización en la apertura y cierre del puerto velofaríngeo (Dotevall et al., 2002). En términos acústicos hay, entre otras cosas, un mayor ancho de banda de los formantes, lo que les confiere una menor intensidad; también hay formantes adicionales, denominados formantes nasales. En la Figura 6 se puede observar la emisión de la palabra / dedo / por parte de un hablante con trastorno de hipernasalidad.

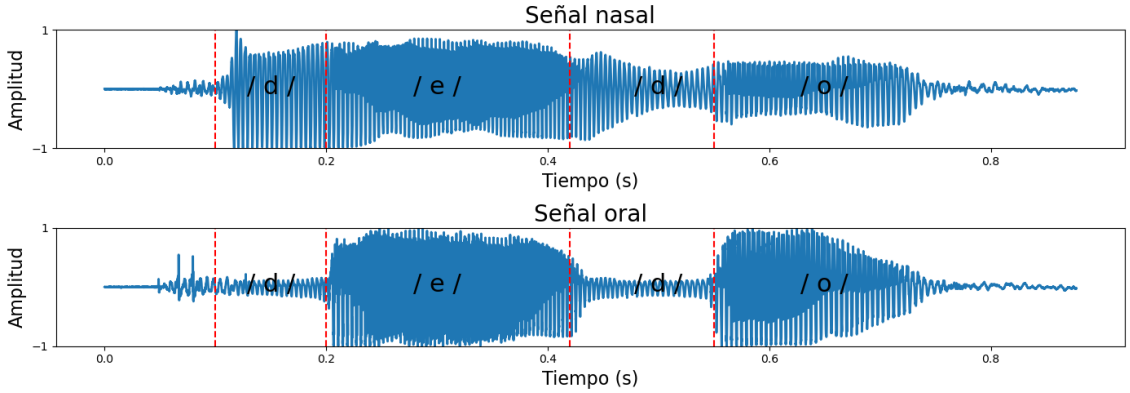

Figura 6. Palabra / dedo/ producida por un hablante con hipernasalidad.

Se observa como la señal nasal presenta unos niveles de intensidad comparables, o superiores, a la señal oral. La locución se compone de los fonemas / d /, / e /, y / o /, por lo que al no haber presencia de consonante nasal, no se produce el fenómeno de coarticulación, que podría explicar el incremento de señal nasal. Sin embargo, se observa especialmente en la consonante / d /, un alto nivel de señal nasal, lo que indica la presencia de hipernasalidad.

2.3.2 Otros patrones de error asociados a la insuficiencia velofaríngea

En el resto de los sonidos aperiódicos, como es el caso de la / s /, se pueden producir diversos fenómenos acústicos que no pertenecen a los tipos estudiados en la presente tesis, como pueden ser turbulencias, o debilitamientos. La emisión de aire nasal audible y/o la turbulencia nasal describen el fenómeno en el que la corriente de aire que atraviesa la nariz se vuelve audible debido a la fricción cuando el aire pasa por un conducto estrecho de la zona nasal, velar y/o faríngea. En términos generales, la señal nasal resultante será aperiódica, y por ello no siempre distinguible de la señal oral. Ahora bien, en ocasiones al pasar al corriente de aire por la cavidad nasal, la presencia de elementos móviles, como mucosa o los propios tejidos internos, pueden vibrar, lo que resulta en un sonido que se denomina turbulencia nasal, y que acústicamente se caracteriza por ser cuasiperiódico. Un ejemplo se puede observar en la Figura 7. La turbulencia nasal suele deberse a una fricción en un conducto más estrecho. La emisión de aire nasal audible y/o la turbulencia nasal suelen coincidir con el trastorno de resonancia hipernasalidad.

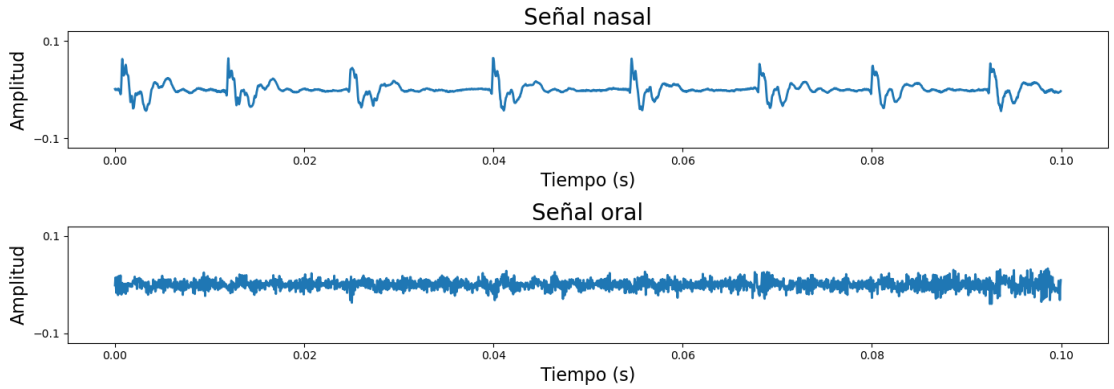

Figura 7. Consonante / s / sostenida en el tiempo de un hablante con insuficiencia velofaríngea que resulta en sonido turbulento.

El debilitamiento de las consonantes es un síntoma común de hablantes con insuficiencia velofaríngea. Entre las consonantes hay un grupo de consonantes que requieren alta presión intraoral, en español las oclusivas sordas (/ p /, / t /, / k /) y las fricativas (/ f /, / s /, / x /). Cuando el puerto velofaríngeo no está adecuadamente cerrado, estos sonidos pueden tener una presión reducida y, por lo tanto, sonar poco claros. Estos debilitamientos suelen coocurrir con la hipernasalidad. En la Figura 8 se puede observar cómo este comportamiento produce una baja del nivel de intensidad medio cuando se produce la consonante / s /. Este tipo de debilitamiento requiere una intervención multidisciplinar, puesto que constituye un efecto sutil y de difícil detección; por ello, resulta esencial contar con el análisis del logopeda. El tratamiento de los trastornos del habla relacionados con la insuficiencia velofaríngea generalmente implica una intervención quirúrgica junto con tratamiento de logopedia.

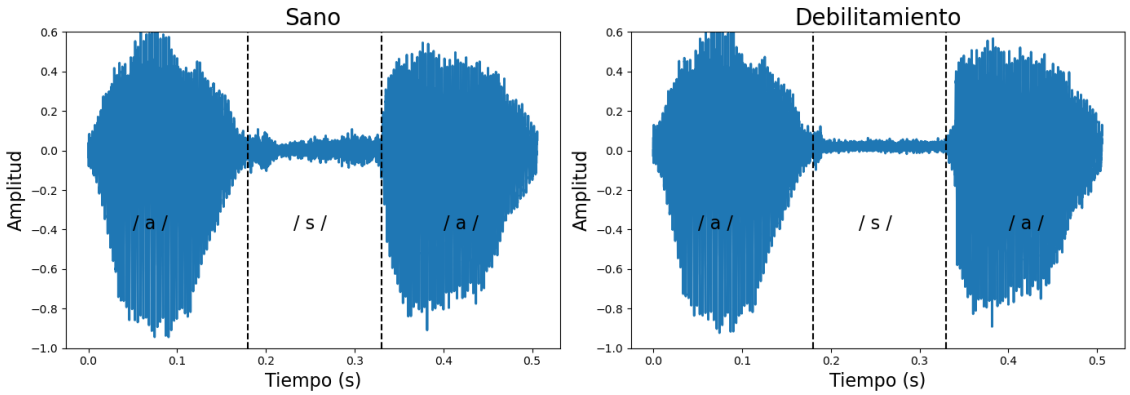

Figura 8. Secuencia / a s a / producida por un hablante sano (izquierda) y un hablante con debilitamiento de las consonantes (derecha). Se observa una bajada de intensidad media en la producción de la consonante / s /.

2.4 Evaluación de la hipernasalidad

El punto de partida para la evaluación por parte de un logopeda suele ser una evaluación perceptual auditiva, y ésta es también la evaluación más común en los entornos clínicos (Sell et al., 2009). La evaluación perceptual puede incluir la transcripción fonética, el uso de escalas de valoración para cuantificar las características del habla y del lenguaje (como la hipernasalidad, la emisión de aire nasal audible y/o la turbulencia nasal, y la inteligibilidad), y las descripciones cualitativas. En la evaluación de la nasalidad se suelen utilizar escalas de valoración para cuantificar la hiper e hiponasalidad.

La evaluación a través de la audición es también el método estándar para la evaluación del habla en pacientes con fisura del paladar (Sell & Grunwell, 2001). La evaluación de la resonancia y la articulación se realiza en una sesión que normalmente se graban en audio para garantizar la posibilidad de realizar una transcripción detallada y permitir su escucha posterior con fines clínicos y de investigación. El material de la prueba incluye palabras sueltas, frases y ejercicios de habla continua. Idealmente, la recogida de muestras de habla debe llevarse a cabo siguiendo las pautas de un protocolo aceptado internacionalmente, como el creado por el *International Speech Parameters Group* (Henningsson et al., 2008). Este protocolo tiene un doble objetivo, ya que pretende obtener una lista de enunciados lo suficientemente informativa como para evaluar perceptualmente la nasalidad y, además, proporciona resultados fiables que pueden compararse con otros estudios, independientemente del idioma en el que se realice la prueba (Henningsson et al., 2008; Kummer, 2016; Spruijt et al., 2018).

Las variables de nasalidad se valoran habitualmente en una escala ordinal (de entre 3 y 5 puntos). No se trata de una escala con intervalos regulares, ya que no se presupone la igualdad de distancias entre los puntos de la escala. Por el contrario, el valor 1 suele describirse como una desviación muy leve cercana a la normalidad y que también se encuentra en la población normal. Así pues, en el caso del protocolo indicado anteriormente, tendríamos tres valores de escala en el extremo inferior de la escala, uno en el medio y uno en el extremo superior.

2.4.1 Fiabilidad de la evaluación perceptual de la hipernasalidad

La cuestión de la fiabilidad de las valoraciones realizadas por los especialistas en logopedia es una cuestión muy importante en las evaluaciones clínicas y de investigación, especialmente porque la hipernasalidad es una variable que ha demostrado ser difícil de evaluar de forma fiable (BJ, 1990; Brancamp et al., 2010; Counihan & Cullinan, 1970; Keuning et al., 2002; Khwaileh et al., 2018; Persson et al., 2006). Una de las razones es la influencia de otras variables del habla coexistentes en la percepción de la nasalidad, como son: la emisión nasal audible o turbulencia nasal, la precisión articulatoria, el tono, y el volumen (BJ, 1990; Fletcher, 1973; Zraick & Liss, 2000).

Hay dos tipos de fiabilidad de los evaluadores que resultan de interés: la fiabilidad intraevaluador, que indica si un evaluador realiza la misma calificación cuando evalúa la misma muestra de habla más de una vez, y la fiabilidad interevaluador, que indica si diferentes oyentes otorgan la misma calificación a una muestra de habla determinada. En contraste con los problemas señalados, también hay una serie de ejemplos de buena fiabilidad de ambos tipos para la hipernasalidad (Grunwell, 2000; Hayden & Klimacka, 2000; Pulkkinen, Haapanen, Paaso, et al., 2001; Sell & Grunwell, 2001), pero las razones de las diferencias entre los estudios siguen sin resolverse.

En una revisión exhaustiva de las limitaciones del análisis perceptivo auditivo se describen varios problemas generales con las evaluaciones perceptivas auditivas (Brunnegård, 2008). Estos problemas son relevantes para la evaluación de los trastornos del habla relacionados con la insuficiencia velofaríngea:

1) Los jueces no parecen tener definiciones equivalentes de los efectos que deben valorarse. Las definiciones de hipernasalidad, diferentes grados de hipernasalidad, emisión de aire nasal audible/turbulencia nasal, etc. no son muy exactas (Sweeney & Sell, 2008). En ocasiones, el fenómeno de la emisión audible de aire nasal y/o la turbulencia nasal se evalúa bajo el diagnóstico de hipernasalidad (Paal et al., 2005); en otros estudios, la hipo e hipernasalidad se valoran en la misma escala, representando extremos opuestos de un continuo (Hayden & Klimacka, 2000). Algunos autores incluyen las fricativas nasales activas dentro de las emisiones nasales audibles (Kummer, 2001) y otros utilizan el concepto emisión nasal e incluyen tanto la emisión nasal audible como la inaudible (Watterson et al., 1998).
2) Los especialistas no llegan a un consenso sobre qué efectos perceptivos deben ser tenidos en cuenta y cómo se definen (John et al., 2006). En la actualidad existe cierto consenso respecto a los protocolos de evaluación del habla asociada a la insuficiencia velofaríngea. La mayoría de los estudios incluyen la evaluación de la hipernasalidad, la emisión audible de aire nasal y/o la turbulencia nasal y las consonantes de presión (Henningsson et al., 2008; Lohmander & Olsson, 2004).
3) Las valoraciones perceptuales de varios efectos están interrelacionadas, es decir, no son independientes. En el caso de los trastornos del habla relacionados con la insuficiencia velofaríngea , la hipernasalidad, la emisión nasal audible/turbulencia

nasal y las consonantes debilitadas están interrelacionadas, ya que tienen el mismo origen y a menudo coocurren (BJ, 1990). También existe una conexión entre las valoraciones de la hipernasalidad y el tono y el volumen (Zraick & Liss, 2000).

4) Las diferencias entre jueces expertos pueden ser muy variables. Según la investigación (Kreiman et al., 1992) los calificadores expertos presentan diferentes evaluaciones para los efectos a diagnosticar en la voz patológica. Esto es importante en la investigación de resultados y también en la evaluación de resultados con fines clínicos, es decir, cuando evaluamos el habla antes y después del tratamiento podemos obtener un resultado de falso positivo cuando en realidad se debe a un cambio de clínico evaluador. Si se comparan estos los resultados con diferentes oyentes no se pueden asegurar si lo que difiere es la evaluación subjetiva de los oyentes o el resultado real.

En una revisión de los estudios publicados se observa que en la evaluación de la hipernasalidad se utilizan escalas con entre 2 y 7 puntos (Hardin et al., 1992; Pulkkinen, Haapanen, Laitinen, et al., 2001). Por ejemplo, el protocolo CAPS-A identifica tres niveles de nasalización, dependiendo de qué segmentos se nasalizan: (1) leve: nasalización evidente en vocales cerradas (por ejemplo, / i /, / u /); (2) moderado: nasalización observable en todas las vocales; y (3) grave: nasalización observable en todas las vocales y en consonantes sonoras (por ejemplo, / b /, / d /, / g /). Algunos estudios argumentan que una escala con menos puntos aumentaría la fiabilidad inter- e intra- evaluador (Pulkkinen, Haapanen, Paaso, et al., 2001). Esto también puede inferirse de una revisión de la literatura, ya que dos de los estudios con la mejor fiabilidad del evaluador han incluido una escala con pocos puntos de escala: una escala binaria (Pulkkinen, Haapanen, Paaso, et al., 2001) y una escala con tres puntos de escala (Grunwell, 2000). Las puntuaciones compuestas del insuficiencia velofaríngea también parecen aumentar la fiabilidad (Park et al., 2000). En otro estudio se relaciona una amplia formación de los especialistas en logopedia con una buena fiabilidad de los evaluadores (Sell et al., 2001).

En línea con los problemas identificados, se han hecho sugerencias sobre cómo mejorar la fiabilidad de los evaluadores. Algunas son requisitos básicos, como unas buenas condiciones de escucha y una buena calidad de las grabaciones. Otras recomendaciones han sido utilizar evaluadores con experiencia en el campo de los trastornos de resonancia (Hayden & Klimacka, 2000; Lewis et al., 2003), utilizar entrenamiento del oyente (John et al., 2006), un tipo correcto de escala (Whitehill, 2002) y utilizar una definición más detallada de los efectos y variables a evaluar (Henningsson et al., 2008; John et al., 2006).

2.4.2 Evaluación instrumental

Las dificultades de la evaluación perceptiva auditiva han llevado a la búsqueda de métodos instrumentales que puedan proporcionar medidas fiables. Varios artículos concluyen que ninguna técnica instrumental puede sustituir al análisis perceptivo (Bettens et al., 2014; Kuehn & Moller, 2000), pero existen varias opciones para el clínico que necesite complementar la evaluación perceptiva auditiva con evaluaciones

instrumentales. Existen variantes de evaluación instrumental basadas en imagen que son muy relevantes para proporcionar información sobre el estado de la función velofaríngea en relación con el habla, como la nasolaringoscopia de alta velocidad (Siriwardena et al., 2024), videofluoroscopia (Henningsson & Isberg, 1991), o resonancia magnética (Kao et al., 2008). Estos métodos sólo miden indirectamente el habla, son invasivos e incómodos para el paciente. También se dispone de mediciones aerodinámicas (Kummer, 2014), pero no se tratarán en este texto porque el interés de esta tesis es la acústica y las propiedades perceptivas auditivas del habla. Los artículos que avalan esta tesis doctoral emplean un instrumento acústico no invasivo, el nasómetro, desarrollado específicamente para proporcionar una medida fiable y directamente comparable con la evaluación perceptiva del habla.

El nasómetro es un instrumento de medida de la nasalidad, es decir, la proporción de energía nasal con respecto a la energía acústica total en una señal del habla. Se trata del instrumento acústico más utilizado para la evaluación de la hiper e hiponasalidad en entornos clínicos (Bettens et al., 2014; Gildersleeve-Neumann & Dalston, 2001). El nasómetro utiliza dos micrófonos separados por una placa posicionada entre la nariz y la boca, de esta forma permite registrar simultáneamente la señal nasal y la señal oral del habla. En la Figura 9 se observa el nasómetro empleado en el presente trabajo.

Figura 9. Nasómetro icSpeech (Rose Medical Solutions Ltd., Canterbury, UK). Fuente: https://icspeech.com/nasometry.html

La nasalancia es una medida del grado de apertura velofaríngea en el habla sonora que se obtiene calculando la relación entre la energía acústica en nariz, y la energía acústica en la boca junto con la nariz (Fletcher et al., 1974). La fórmula que se emplea es la que muestra la Ecuación (1):

$$Nasalancia = \frac{Energía\ Nasal}{Energía\ Oral + Energía\ Nasal} \times 100 \qquad (1)$$

En la Figura 10 se muestra el cambio del valor de nasalancia cuando se producen sonidos orales o nasales.

En el artículo original, Fletcher filtraba las señales de entrada y se quedaba solo con la banda de 300-600 Hz. El motivo de esto se debe a que en esa banda de frecuencia es

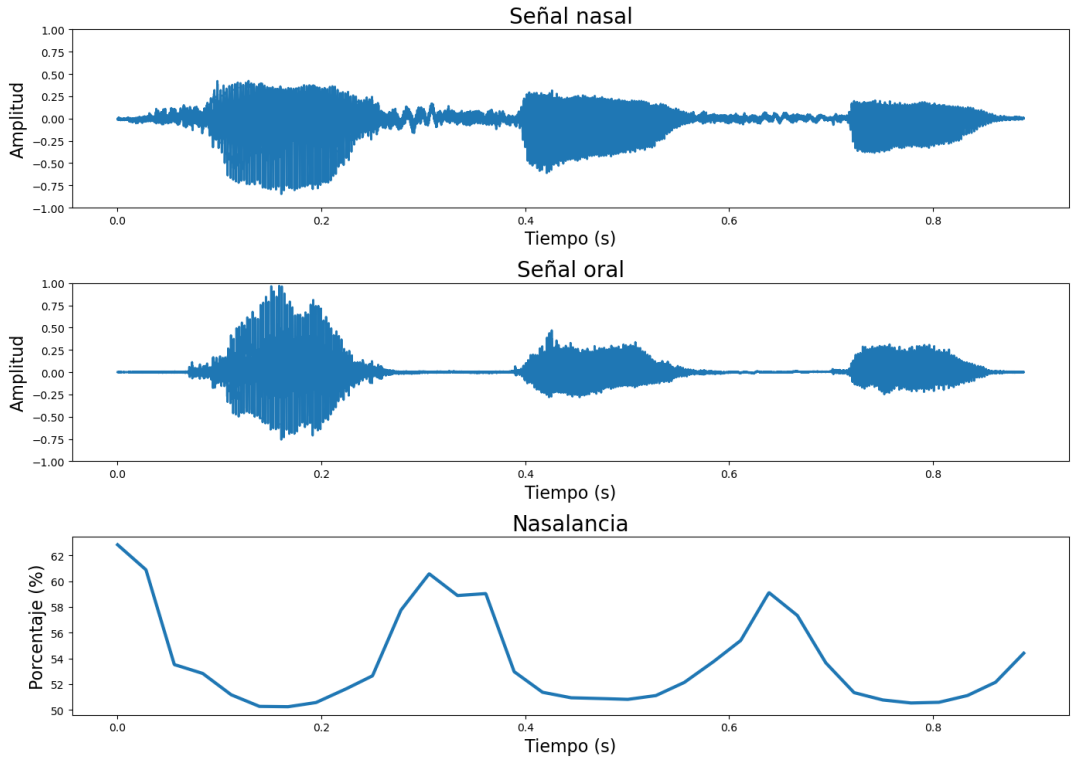

Figura 10. Relación entre la producción de la señal oral y nasal y el valor de nasalancia en la repetición de sílabas / sa / realizada por un hablante con hipernasalidad.

donde típicamente aparecen los formantes nasales en las vocales, que es el objeto de estudio de sus trabajos. Sin embargo, la nasalancia se puede calcular de más formas: se puede usar el valor eficaz en lugar de la energía, y se puede filtrar la señal de entrada o emplear toda la banda de frecuencia. El hecho de emplear todo el espectro tiene sentido cuando se mide la nasalancia en oraciones con muchos sonidos diferentes, y no exclusivamente en las vocales.

El especialista clínico puede utilizar la puntuación de nasalancia para compararla con su propia puntuación perceptiva y para comparar las puntuaciones previas y posteriores al tratamiento. Sin embargo, la fiabilidad de este instrumento no está clara; existen estudios comparando los valores de nasalancia y la evaluación perceptiva en los que se obtienen correlaciones que van de no significativas a fuertes (Brancamp et al., 2010; Keuning et al., 2002; Khwaileh et al., 2018; Liu et al., 2022).

El contenido fonético de los enunciados tiene una gran influencia en las puntuaciones de nasalancia, especialmente la inclusión de fonemas nasales (/ m /, / n /, / ŋ /) (Watterson et al., 1996). Se ha investigado la influencia del volumen vocal, pero no se ha demostrado que tenga ningún impacto (Watterson et al., 1994). También es importante tener en cuenta la variabilidad intrapersonal, que se considera dentro de la variación normal de un hablante (Watterson et al., 2005); y para los hablantes con hipernasalidad la variación es probablemente aún mayor. Una combinación de medidas perceptivas auditivas e instrumentales es lógica para garantizar evaluaciones de buena calidad. Se han realizado comparaciones entre las valoraciones perceptivas auditivas y las puntuaciones de nasalancia y algunos estudios han encontrado una buena correlación

(Brunnegård et al., 2012; Dalston et al., 1991; Hirschberg et al., 2006; Sweeney & Sell, 2008; Watterson et al., 1996); y otros una correlación moderada (Dalston et al., 1993; Keuning et al., 2002; Watterson et al., 1993); e incluso baja (Lewis et al., 2003; Nellis et al., 1992).

En resumen, está ampliamente documentado que la evaluación de la hipernasalidad en el habla es potencialmente poco fiable, lo que ha dado lugar a problemas de comparación e interpretación de la investigación. Esto es de particular importancia para la evaluación de los resultados de los estudios que comparan los resultados del habla de los métodos quirúrgicos para la reparación del paladar hendido. La escasa fiabilidad también plantea dudas sobre las evaluaciones clínicas en las que se basan las decisiones sobre el tratamiento del habla y la cirugía. Es la incertidumbre asociada a la evaluación perceptual auditiva de la hipernasalidad lo que llevó al desarrollo de dispositivos como el nasómetro que miden directa y objetivamente la señal acústica del habla. Sin embargo, se ha descubierto que las medidas de nasometría también presentan cierto grado de variabilidad en cuanto a las puntuaciones, lo que plantea la cuestión de si el dispositivo puede utilizarse eficazmente para mejorar la fiabilidad de la evaluación.

3. Descriptores acústicos

En el presente capítulo se introducen los aspectos teóricos relacionados con la extracción de descriptores acústicos de la señal de voz. En concreto, se describen los descriptores empleados en las tareas de clasificación abordadas en las publicaciones que sustentan esta tesis.

En primer lugar, se abordan los coeficientes MFCC, un descriptor ampliamente utilizado debido a su eficacia en el reconocimiento automático del habla por su capacidad de simular la percepción auditiva humana. Posteriormente, se describe el índice VLHR, relacionado con la nasalización y que permite analizar la distribución de energías en las bajas y altas frecuencias. A continuación, se presenta una descripción detallada de los formantes de la señal de voz, incluyendo sus métodos de cálculo y la estimación de sus respectivos anchos de banda, por su relevancia en el análisis de las propiedades articulatorias del habla. Finalmente, se define el cálculo de la frecuencia fundamental.

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

Los descriptores MFCC se utilizan ampliamente en el reconocimiento automático del habla debido a que simulan el comportamiento del sistema auditivo humano. Este sistema actúa como un conjunto de filtros que responden de manera diferente según la frecuencia, siendo mucho más sensible a los cambios en las frecuencias bajas que en las altas. La escala Mel emplea un banco de filtros que imita este comportamiento selectivo en las frecuencias más bajas.

El valor de los coeficientes MFCC se calcula a través de una cadena de seis procesos consecutivos, los cuales son: preénfasis de la señal de audio, división en ventanas (*frames*), cálculo del espectro de potencia, aplicación de un banco de filtros de Mel al espectro obtenido, aplicación del logaritmo sobre la salida del banco de filtros y, por último, aplicando la transformada de coseno discreta (DCT). La Figura 11 muestra el proceso esquemático empleado para el cálculo de los descriptores.

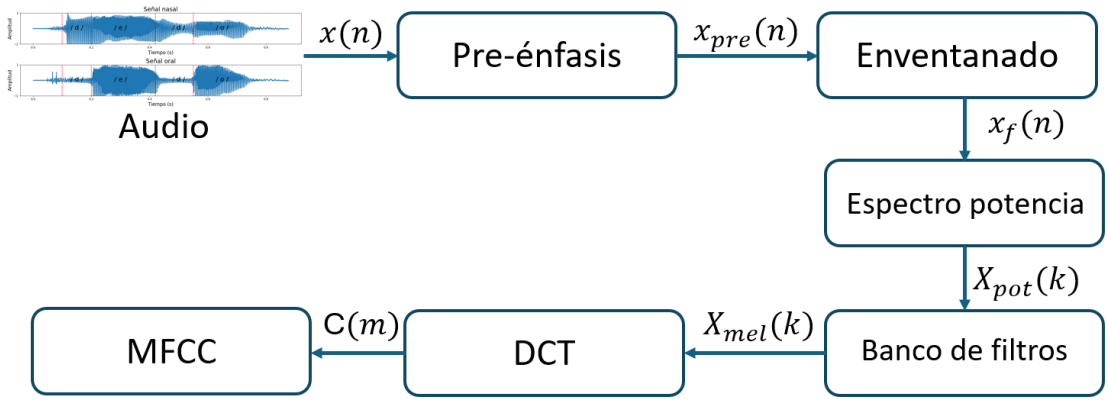

Figura 11. Diagrama de bloques para el cálculo de los descriptores MFCC.

Donde $x(n)$ es la señal de audio original en el dominio temporal, $x_{pre}(n)$ es la señal resultante de aplicar el filtro de pre-énfasis, $x_f(n)$ es la señal dividida en frames, $X_{pot}(k)$ es el espectro de potencia, $X_{mel}(k)$ es el vector obtenido al aplicar el logaritmo a la salida del banco de filtros Mel, y $C(m)$ son los coeficientes MFCC.

Cabe mencionar ciertas características y limitaciones a la hora de emplear los MFCC. El proceso de enventanado introduce dos dependencias clave: 1) temporal: el resultado de los coeficientes cambia si se modifican la duración o la superposición de los frames, lo que repercute en la resolución temporal de los descriptores, y 2) frecuencial: el tamaño de la ventana fija la resolución espectral (cuanto más larga, mejor discriminación en frecuencia), lo que afecta a la posición y la amplitud de los picos de los formantes estimados a partir del espectro de potencia.

Debido a que la voz es una señal no estacionaria, es decir, las características acústicas cambian rápidamente entre fonemas, se necesitan estrategias de análisis localizado. El espectro localizado consiste en aplicar sucesivamente ventanas de longitud fija y superponerlas realizando un desplazamiento para capturar la evolución temporal de los MFCC. Existen técnicas de análisis tiempo-frecuencia avanzado empleando variantes como *Time-Frequency MFCC,* o *Constant-Q Cepstral Coefficients* (CQCC) que emplean ventanas adaptativas o bancos de filtros no uniformes para mejorar la representación de fenómenos transitorios (Todisco et al., 2017; Wassner & Chollet, 1996). Sin embargo, estas técnicas añaden complejidad computacional y, a menudo, requieren parámetros específicos para cada base de datos, por lo que no se emplean en el presente trabajo.

Además, los MFCC no describen por sí solos todos los aspectos pertinentes de la señal. Los coeficientes estándar no permiten distinguir explícitamente entre segmentos sonoros y no sonoros; por ello se suele complementar estos descriptores con otros como la energía total, el valor de autocorrelación, o la frecuencia fundamental (F0), para capturar la periodicidad. Tampoco analizan las características de energía por bandas, por lo que es común añadir el valor del logaritmo de la energía de ciertas bandas del espectro (por ejemplo, banda grave 0-300 Hz o banda fricativa 2-4 kHz) para facilita la detección de determinados modos de fonación (Shin et al., 2000).

Por lo tanto, aunque los MFCC constituyen una herramienta fundamental para el reconocimiento automático del habla, su rendimiento óptimo depende de un diseño cuidadoso del enventanado y, en muchos casos, de la combinación con otros descriptores. A continuación se describen cada uno de los procesos del cálculo de coeficientes MFCC en detalle.

3.1.1 Pre-énfasis

Una de las prácticas comunes en el procesado de una señal de audio es el pre-énfasis. Este proceso realza las frecuencias altas de la señal, que se atenúan o se suprimen durante la grabación de la señal. Este primer paso se realiza aplicando un filtro paso alto con un valor de coeficiente α entre [0, 1], como se muestra en la Ecuación (2).

$$x_{pre}(n) = x(n) - \alpha \cdot x(n-1) \qquad (2)$$

Siendo $x(n)$ la señal discreta de entrada y $x_{pre}(n)$ la señal de salida discreta tras el proceso de pre-énfasis. Un valor de $\alpha = 0$ produce una señal sin cambios, mientras que con un valor $\alpha = 1$ se obtiene la diferencia de primer orden de la señal. El valor que se utiliza por defecto es 0.97 ya que coincide con el utilizado en la implementación del software Hidden Markov Model Toolkit (HTK) (Young et al., 2002). Cabe mencionar que este proceso modifica la distribución de la energía a lo largo del espectro y el nivel absoluto de energía de la señal.

3.1.2 Enventanado (*framing*) de la señal

El proceso de enventanar la señal de audio consiste en dividir la señal original en secuencias más cortas denominadas *frames*, que tienden a ser más estacionarias y a presentar una menor variación. Para extraer descriptores acústicos estables, el habla se debe examinar a lo largo de un periodo lo suficientemente corto. Cuando se analiza la señal vocal, una duración de ventana de 20-30 ms se considera un segmento cuasi-estacionario, ya que el tiempo transcurrido entre dos cierres glotales es de unos 20 ms. Cuando se quieren analizar cambios rápidos en la señal se emplean ventanas más pequeñas. La producción de vocales en español se produce con una duración de entre 40 ms y 80 ms (Marín Gálvez, 1995), por lo que un *frame* de 20-30 ms permite obtener un espectro estacionario denominado *short-term*. En el presente trabajo se calcula el espectro de la señal de voz empleando *frames* móviles con un tamaño típico 25 ms con 10 ms de solapamiento. El solapamiento permite medir el comportamiento temporal de los descriptores.

Para cada uno de los *frame*, se aplica un filtrado de la señal para reducir la amplitud de en los extremos. En el siguiente paso se va a calcular el espectro de potencia de la señal enventanada aplicando la transformada discreta de Fourier. Esta transformación asume que la señal es periódica y de duración infinita, por lo que si existe una discontinuidad entre el inicio y fin de la señal, cosa que es muy probable debido a la naturaleza de la señal de audio y el tamaño de ventana escogido, se generan nuevos componentes en alta frecuencia que no están presentes en la señal original. Por ello, se evita el uso de una ventana rectangular, y se emplean ventanas de Hanning o Hamming para realizar este proceso (Rao & Manjunath, 2017). Este tipo de ventanas reducen la aparición de componentes en alta frecuencia, lo que se conoce como efecto de borde, a costa de introducir una distorsión en la señal. En la Figura 12 se puede observar la representación de estos filtros en el dominio del tiempo y la frecuencia. Aun así, la elección de tipo de ventana, entre Hanning o Hamming, no influye mucho en el resultado final, siendo el factor determinante para el cálculo de estos descriptores el tamaño de ventana empleado.

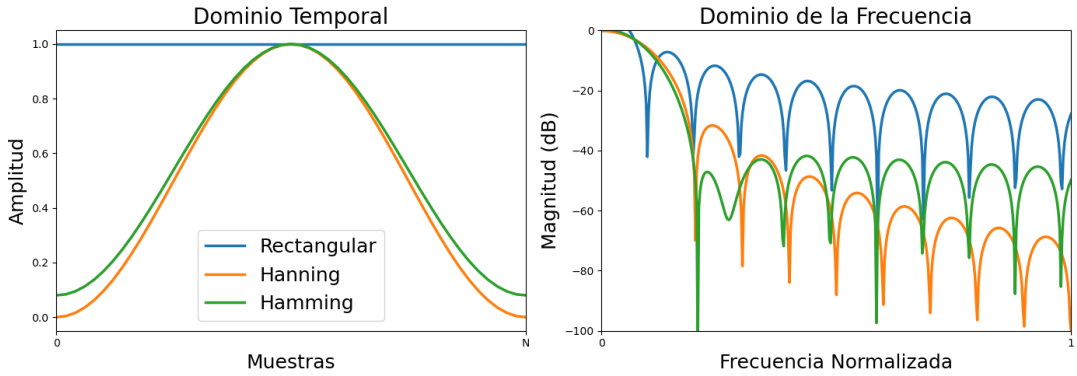

Figura 12. Representación de la ventana rectangular, Hanning y Hamming en el dominio del tiempo y la frecuencia.

3.1.3 Espectro de potencia

El espectro de potencia de una señal describe su distribución de potencia en función de las componentes en frecuencia. Se emplea la Transformada Discreta de Fourier (DFT) para representar la señal en el dominio de la frecuencia. La DTF de cada uno de los frames, o segmentos temporales, de la señal de voz se calcula mediante la Ecuación (3):

$$X(k) = \sum_{n=0}^{N-1} x_f(n) e^{-j\frac{2\pi}{N}nk} \ , k = 1, 2, \dots, N-1 \qquad (3)$$

Donde $x(n)$ es la señal en el dominio temporal, $X(k)$ es la representación en el dominio de la frecuencia, N es el número de muestras del *frame*, y k es el índice discreto que representa las distintas componentes en frecuencias.

A partir de esta transformación se obtiene el espectro de potencia de la señal mediante el cálculo del módulo al cuadrado, también denominado periodograma, como muestra la Ecuación (4).

$$X_{pot}(k) = \frac{1}{N}|X(k)|^2 \qquad (4)$$

Este espectro representa la potencia presente en cada componente de frecuencia del frame.

3.1.4 Banco de filtros de Mel

El banco de filtros Mel es un conjunto de filtros paso banda diseñados según la escala Mel, que modela la percepción del tono en el sistema auditivo humano. Esta escala fue desarrollada originalmente para el análisis y la percepción del habla, y su objetivo es obtener una representación no lineal de la señal que refleje cómo el oído humano percibe las distintas frecuencias: con mayor sensibilidad en frecuencias bajas y menor

en las altas. Cada filtro del banco tiene forma triangular y está centrado en una frecuencia específica dentro del dominio de la escala Mel. La respuesta en frecuencia de estos filtros es igual a 1 en su frecuencia central, y decrece linealmente hacia cero hasta las frecuencias centrales de los filtros adyacentes.

La función de transferencia de cada uno de los filtros $H_m\,(k)$, donde m indica el filtro dentro del banco, se define mediante la Ecuación (5).

$$H_m\,(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{5}$$

Donde $H_m\,(k)$ es la ganancia del filtro m en el punto de frecuencia k, $f(m)$ es la frecuencia central del filtro triangular, y k el índice de frecuencia del espectro.

La transformación entre la frecuencia lineal (en Hz) y la escala Mel se emplea para construir el banco de filtros Mel de forma que refleje la percepción auditiva humana, la cual es aproximadamente lineal en bajas frecuencias y logarítmica en las altas. Esta transformación permite distribuir los filtros de forma no uniforme, más densos en las frecuencias bajas y más espaciados en las altas, simulando así el comportamiento del oído humano.

La conversión de frecuencia lineal a escala Mel se realiza mediante la Ecuación (6):

$$m = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

De forma inversa, para convertir de escala Mel a frecuencia lineal, se utiliza la Ecuación (7):

$$f = 700\left(10^{\frac{m}{2595}} - 1\right) \tag{7}$$

Estas expresiones permiten establecer la posición de los filtros del banco Mel en el dominio de frecuencia de la señal, facilitando su implementación sobre el espectro de potencia calculado previamente. En la Figura 13 se muestra una representación de los filtros del banco de Mel en el dominio de la frecuencia (Hz).

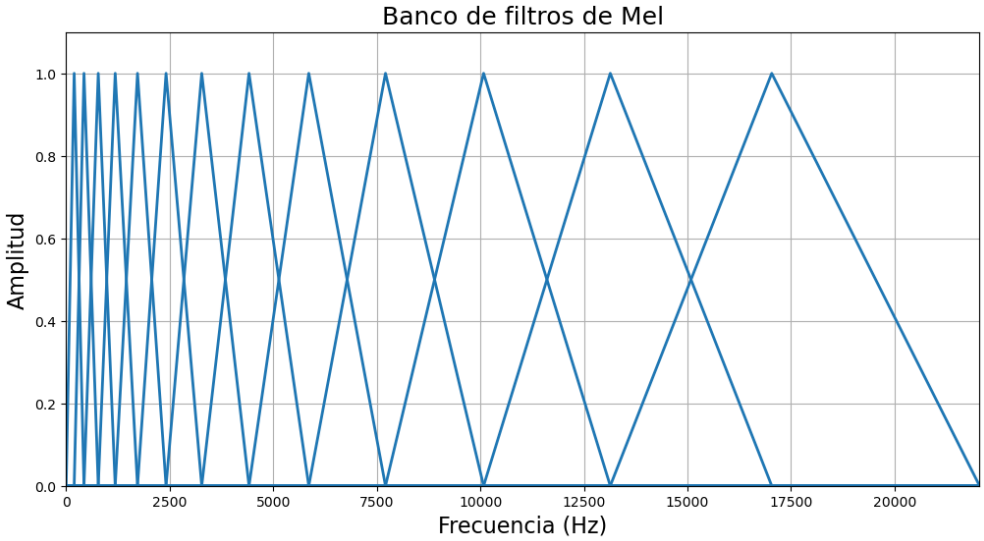

Figura 13. Filtros triangulares en un banco de filtros de Mel en el dominio de la frecuencia (Hz).

Después de aplicar el banco de filtros Mel al espectro de potencia, se obtiene un vector de energías por filtro (es decir, cuánta energía hay en cada banda perceptual), como se muestra en la Ecuación (8):

$$E_m = \sum_k X_{pot}(k) \cdot H_m(k) \tag{8}$$

Y luego se aplica el logaritmo a esas energías:

$$X_{mel}(k) = \ln(E_m) \tag{9}$$

El logaritmo se aplica para simular la percepción auditiva humana, que responde logarítmicamente a la intensidad, reducir la variabilidad dinámica entre filtros (las energías pueden variar mucho en magnitud), y convierte productos en sumas, facilitando la siguiente transformación.

3.1.5 Transformada de coseno discreta (DCT)

La transformada de coseno discreta (Ahmed et al., 1974) permite representar una secuencia finita de valores reales como una suma ponderada de funciones coseno con diferentes frecuencias. En el cálculo de los coeficientes MFCC, la DCT se aplica a los logaritmos de las energías de los bancos de filtros Mel, con el objetivo de decorrelacionar los valores y obtener un conjunto de coeficientes (Strang, 1999). En muchos sistemas se seleccionan únicamente los primeros 13 coeficientes, que contienen la mayor parte de la información útil. Existen diferentes tipos de DCT, pero la utilizada usualmente para el cálculo de los MFFC es el tipo II, que se muestra en la Ecuación (10).

$$C(m) = \sum_{k=0}^{K-1} X_{mel}(k) \cdot \cos\left[\frac{\pi}{K}\left(k + \frac{1}{2}\right)m\right], \quad m = 1, 2, \ldots, M-1 \qquad (10)$$

Donde $X_{mel}(k)$ es la salida logarítmica del banco de filtros Mel (energía logarítmica por filtro), K es el número total de filtros Mel, $C(m)$ es el coeficiente MFCC número m señal, y M es el número de coeficientes que se quiere calcular (típicamente 12 o 13).

3.1.6 Primera y segunda derivada

Tras el cálculo de los componentes MFCC se analiza también los cambios temporales de estos coeficientes mediante la primera derivada (Δ) y la segunda derivada (Δ^2). Estas derivadas capturan la dinámica de la señal de voz, es decir, cómo evolucionan los coeficientes cepstrales en el tiempo. La primera derivada estima la velocidad de cambio de los MFCC entre frames consecutivos, mientras que la segunda derivada representa la aceleración de dichos cambios.

Conceptualmente, la forma más sencilla de calcular la primera derivada es mediante la diferencia de los componentes MFCC, de la forma $\Delta C_t(m) = C_t(m) - C_{t-1}(m)$. Esta diferencia simple es una mala aproximación a la primera derivada de los coeficientes y no suele utilizarse en la práctica. En su lugar, la primera derivada se implementa a menudo como una aproximación por mínimos cuadrados sobre una región alrededor de la muestra del tiempo actual (Rabiner & Schafer, 2010), como muestra la Ecuación (11).

$$\Delta C_t(m) = \frac{\sum_{n=1}^{N} n \cdot \left(C_{t+n}(m) - C_{t-n}(m)\right)}{2\sum_{n=1}^{N} n^2} \qquad (11)$$

De forma análoga, la segunda derivada se calcula mediante la Ecuación (12):

$$\Delta^2 C_t(m) = \frac{\sum_{n=1}^{N} n \cdot \left(\Delta C_{t+n}(m) - \Delta C_{t-n}(m)\right)}{2\sum_{n=1}^{N} n^2} \qquad (12)$$

Donde $C_t(m)$ es el coeficiente MFCC número m correspondiente al frame temporal t, $\Delta C_t(m)$ es la primera derivada, $\Delta^2 C_t(m)$ es la segunda derivada, y N es el tamaño de la ventana temporal utilizada (típicamente entre 2 y 3).

Estas derivadas se calculan mediante una regresión lineal centrada, lo cual permite suavizar los cambios y reducir la influencia del ruido o de variaciones locales poco significativas.

3.2 Voice Low Tone to High Tone Ratio (VLHR)

El descriptor VLHR está relacionado con la permeabilidad de las vías respiratorias nasales (Lee et al., 2003). Este índice se define como la relación entre la potencia de baja frecuencia (*Low Frequency Power*, LFP) y la potencia de alta frecuencia (*High Frequency* Power, HFP), obtenidas tras dividir el espectro de la voz utilizando una frecuencia de corte específica. VLHR se emplea como estimador de la nasalización de la señal de voz, ya que muestra una alta correlación tanto con las medidas objetivas de nasalancia como con las evaluaciones perceptuales de hipernasalidad (Lee et al., 2006).

Para calcular el valor de VLHR, se estima el espectro de la señal de voz mediante la transformada rápida de Fourier (FFT), aplicando una ventana móvil sobre los frames de la señal. Los espectros resultantes se promedian para obtener un espectro medio representativo. A continuación, dicho espectro se divide en dos bandas: una de baja frecuencia (LFP) y otra de alta frecuencia (HFP), utilizando una frecuencia de corte de F_C (Hz). La potencia LFP se calcula como la suma de energía espectral entre 65 Hz a F_C, mientras que HFP corresponde a la suma de la potencia de F_C a 8000 Hz. El valor de VLHR se expresa en decibelios (dB), como muestra la Ecuación (13).

$$VLHR = 10 \cdot log_{10}\left(\frac{LFP}{HFP}\right) \tag{13}$$

Para simular el acoplamiento del tracto vocal con la cavidad nasal durante la fonación, se emplea un modelo acústico equivalente basado en un circuito eléctrico L-C-R en configuración paralela. Este tipo de modelo permite representar la aparición de polos y ceros en la función de transferencia del sistema vocal, causados por la interacción entre las distintas cavidades. En ciertos casos, se introduce una resistencia negativa como abstracción matemática para modelar fenómenos de realimentación o amplificación acústica observados en la nasalización (Feng & Castelli, 1996).

Según estos modelos, si se selecciona una frecuencia de corte situada entre un polo y un cero, se observa un aumento significativo en el valor de VLHR. Por esta razón, se emplea un conjunto de frecuencias de corte comprendidas entre 400 Hz y 900 Hz, en pasos de 100 Hz. La razón de utilizar múltiples frecuencias de corte es que la correlación entre el valor de VLHR y las medidas perceptuales de nasalización puede variar en función del idioma y el contexto fonético analizado.

3.3 Formantes de la señal de voz

En el habla humana, los formantes son las frecuencias de resonancia características del tracto vocal, manifestadas como máximos en el espectro de la señal de voz y típicamente se identifican como F1, F2, F3, etc... Estas frecuencias surgen de la amplificación selectiva que producen las cavidades supraglóticas sobre ciertas bandas de frecuencia del sonido generado en la laringe. En general, existen tantos formantes como resonadores en el

tracto vocal; sin embargo, en la práctica solo los tres primeros formantes suelen proporcionar suficiente información para distinguir distintos sonidos del habla, especialmente en las vocales (Prica & Ilić, 2010).

Por ejemplo, las vocales se diferencian principalmente por los valores de sus dos primeros formantes, relacionados con la posición de la lengua y la abertura de la boca,

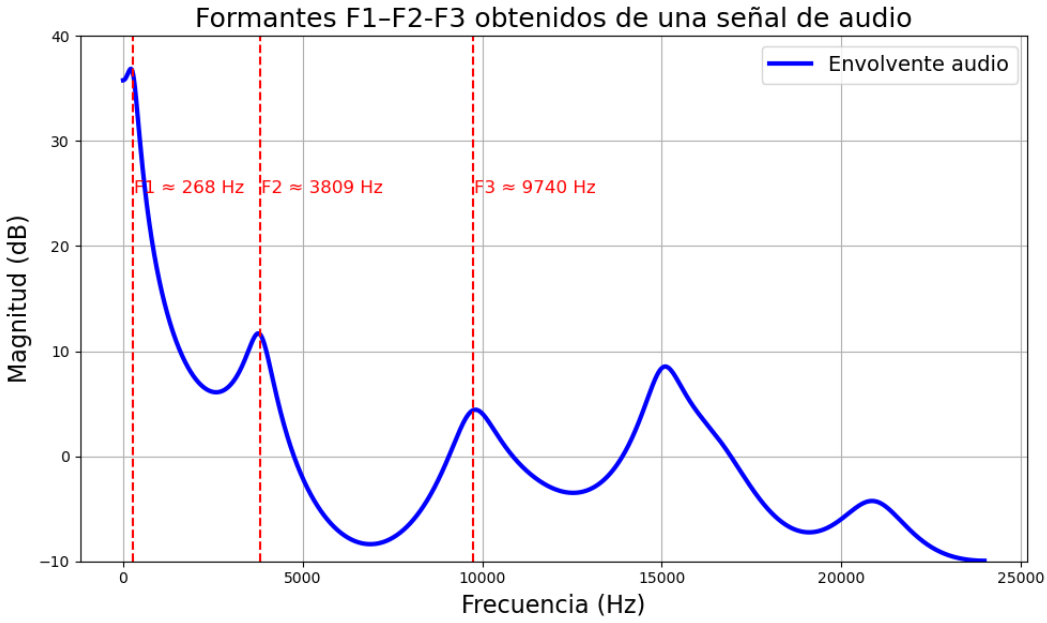

Figura 14. Formantes F1, F2 y F3 sobre la envolvente de una señal de audio.

mientras que el tercer formante contribuye al matiz o timbre de la voz (Stevens, 2000). Los valores típicos de frecuencia de los formantes dependen del hablante y del sonido articulado. Un ejemplo de estos formantes se puede observar en la Figura 14.

3.3.1 Cálculo de los formantes

Una herramienta ampliamente utilizada para estimar formantes en señales de voz es la Predicción Lineal (*Linear Predictive Coding*, LPC). El modelo LPC asume que, en ventanas cortas de tiempo, la señal de voz puede modelarse como la salida de un filtro todo-polo excitado por una fuente (sonora o ruido) (Makhoul, 1975; Markel & Gray, 2013).

En términos matemáticos, el principio básico de LPC es que cada muestra de voz puede aproximarse como una combinación lineal de p muestras pasadas. Esto se expresa mediante la Ecuación (14) de predicción lineal de orden p:

$$x(n) = -\sum_{i=1}^{p} a_i x(n-i) + e(n) \tag{14}$$

Donde $x(n)$ es la muestra actual de la señal de voz, a_i son los coeficientes LPC, $x(n-i)$ son las muestras pasadas, y $e(n)$ es el error de predicción.

Esta ecuación indica que la señal puede descomponerse en una parte predecible en función de las muestras anteriores y una información nueva aportada por el error de predicción. Este error actúa como señal de excitación del sistema, como un tren de pulsos periódicos en segmentos sonoros, o ruido blanco en segmentos no sonoros. Los coeficientes a_i se obtienen minimizando el error cuadrático medio $E\{e(n)^2\}$ dentro de cada *frame*.

Una vez obtenidos los coeficientes LPC de orden p, se define el polinomio predictor (polinomio característico del filtro LPC) mediante la Ecuación (15):

$$A(z) = 1 + \sum_{i=1}^{p} a_i z^{-1} \tag{15}$$

La función de transferencia del filtro todo-polo que modela el tracto vocal queda entonces definida por la Ecuación (16):

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{i=1}^{p} a_i z^{-1}} \tag{16}$$

Donde G es la ganancia del filtro. Los polos de $H(z)$, es decir, las raíces del denominador $A(z)$, corresponden a las resonancias del filtro y están directamente asociados a los formantes de la señal de voz.

Cada par de raíces complejas conjugadas (z_i, z_i^*) define un formante. Si la raíz es $z_i = r_i e^{-j\theta_i}$ (en forma polar, con r_i módulo, y θ_i fase), entonces la frecuencia asociada al formante F_i viene dada por la frecuencia angular normalizada θ_i. En términos de frecuencia en Hz, esta se calcula empleando la Ecuación (17):

$$F_i = \frac{\theta_i}{2\pi} f_s \tag{17}$$

Donde θ_i se expresa en radianes, y f_s es la frecuencia de muestreo de la señal.

En resumen, mediante el uso de un LPC de orden p se puede obtener un conjunto de polos cuya distribución en frecuencia señala los formantes del tracto vocal en ese segmento de señal.

3.3.2 Ancho de banda de los formantes

Además de la frecuencia central de cada formante, es importante definir su ancho de banda, el cual cuantifica cómo es el pico de resonancia en el espectro. Un formante con un ancho de banda pequeño aparece como un pico estrecho y pronunciado, mientras que un ancho de banda grande indica un pico más suave o amortiguado. En el contexto del modelo LPC (filtro todo-polo), el ancho de banda está directamente relacionado con

el factor de amortiguamiento de los polos asociados a ese formante. Este factor, a veces llamado coeficiente de decaimiento describe cuán rápido se atenúa la contribución de ese polo en la respuesta temporal (Markel & Gray, 2013).

A continuación se describen dos de las principales técnicas de cálculo del ancho de banda de los formantes.

a) Cálculo a partir de los polos (LPC): Dada una pareja de polos complejos conjugados $(z_i, z_i^*) = r_i e^{\pm j\theta_i}$ obtenidos del análisis LPC, el ancho de banda BW_i se puede calcular a partir del módulo de r_i. Si el sistema se muestrea a frecuencia f_s y se define $T = {}^{1}/_{fS}$, la relación entre el factor de amortiguamiento α_i y r_i viene dada por $r_i = e^{-\alpha_i}$, donde α_i es la constante de decaimiento exponencial. El ancho de banda (a -3 dB) en Hz se puede expresar entonces con la Ecuación (18):

$$B_i = -\frac{\ln(r_i)}{\pi} f_s \tag{18}$$

Esta fórmula, derivada del modelo de Rabiner y Schafer (Rabiner & Schafer, 1978), aproxima el ancho de banda en el sentido de media potencia: corresponde aproximadamente al intervalo de frecuencias alrededor de F_i donde la potencia del formante cae a la mitad (–3 dB) de su valor máximo.

b) Cálculo a partir del espectro (criterio de -3 dB): Alternativamente, el ancho de banda de un formante puede medirse directamente en el espectro de amplitud de la señal o de la envolvente LPC. Para ello, se identifica la frecuencia central del formante (el pico en el espectro) y luego se encuentran las frecuencias a cada lado de dicho pico donde la amplitud cae en aproximadamente 3 dB (una reducción de potencia al 50%) respecto al valor máximo. La diferencia entre esas dos frecuencias es el ancho de banda del formante, comúnmente llamado *ancho de banda a -3 dB*.

En la práctica, se traza una línea horizontal 3 dB por debajo del nivel del pico y se observa el "ancho" de la resonancia en esa altura. Este criterio de caída de 3dB en la potencia es análogo a la definición de ancho de banda en filtros electrónicos, y proporciona una medida empírica del *quality factor (Q)* del formante, que se define como el cociente entre la frecuencia del formante y su ancho de banda. Sin embargo, puede haber dificultades para determinar con precisión estos puntos cuando dos formantes están muy próximos o cuando el pico es poco prominente. Esto puede suceder especialmente en voces femeninas, infantiles, o de tono muy alto, donde la compresión del espacio espectral hace que F1, F2 y F3 se desplacen hacia frecuencias más elevadas y los formantes presentan separaciones menores, lo que incrementa la probabilidad de solapamiento (Lee et al., 1999; Sundberg, 1988). En tales casos, se emplea el método basado en polos, que es más estable ya que deriva directamente de los parámetros del modelo.

3.4 Frecuencia fundamental

La frecuencia fundamental (F0) de una señal es el término que designa la componente periódica más baja en la voz humana, correspondiente a la tasa básica de vibración de las cuerdas vocales. En otras palabras, F0 es la frecuencia del ciclo de repetición de la onda glotal y determina el tono básico (*pitch*) percibido de la voz (Knight & Setter, 2021). Por ejemplo, en el habla de un hombre adulto típico, la frecuencia fundamental suele estar entre 85 y 180 Hz, mientras que en una voz femenina adulta típica está entre 165 y 255 Hz. En general, los hombres presentan F0 más bajas que las mujeres debido a diferencias anatómicas, ya que la longitud y tensión de las cuerdas vocales definen la frecuencia de sus vibraciones. F0 no solo varía entre individuos por sexo y edad, sino también dinámicamente dentro del habla de una persona, modulándose para expresar entonación, énfasis o emociones.

Existen diversos métodos para estimar automáticamente la frecuencia fundamental a partir de la señal de voz. A continuación se describen los métodos clásicos más utilizados, junto con sus fundamentos teóricos:

a) **Método de autocorrelación (dominio temporal):** se basa en la propiedad de que una señal cuasiperiódica presenta alta correlación consigo misma cuando se retrasa en un período fundamental. La función de autocorrelación discreta de la señal $x(n)$ se define con la Ecuación (19):

$$R_{xx}(m) = \sum_{n} x(n)\, x(n-m) \tag{19}$$

Donde m es el retraso en muestras. Al calcular $R_{xx}(m)$ para $m > 0$, aparecen picos locales en retardos correspondientes a múltiplos del período de la señal. En particular, el primer pico significativo de autocorrelación (excluyendo $m = 0$) ocurre en $m = N_0$, que corresponde al período fundamental T_0 (en muestras) de la señal. La frecuencia fundamental se obtiene entonces como muestra la Ecuación (20):

$$F_0 = \frac{f_s}{N_0} \tag{20}$$

Este método es muy efectivo en segmentos de voz sonoros, ya que la autocorrelación revela claramente patrones periódicos incluso para señales ruidosas. De hecho, la autocorrelación es una herramienta útil para identificar la periodicidad fundamental aun cuando F_0 no esté presente explícitamente en el espectro (por ausencia de componente continua), aprovechando la presencia de múltiples armónicos (Rabiner, 1977). En la práctica, para evitar errores por octava se suele limitar la búsqueda al rango plausible de N_0 según el hablante (por ejemplo, 50–500 Hz) y a veces se usa la autocorrelación normalizada para independizar la medida de la amplitud.

b) **Método del *cepstrum* (dominio frecuencial):** este método emplea la estructura armónica en el espectro de voz. Consiste en calcular el *cepstrum* de la señal y

localizar en él su período fundamental. El *cepstrum* se define como la Transformada Inversa de Fourier (IFFT) del logaritmo del espectro de magnitud de la señal . En términos operativos, los pasos son: calcular la FFT de la señal en una ventana, tomar el logaritmo de los valores absolutos del espectro, y luego aplicar la IFFT, como muestra la Ecuación (21):

$$c(n) = \mathcal{F}^{-1}\{ln|\mathcal{F}\{x(n)\}|\} \quad (21)$$

Siendo $x(n)$ la señal original, y $c(n)$ el *cepstrum* obtenido. Este resultado presenta un pico destacado en la posición $n = N_0$ (muestras) si la señal posee una componente periódica de periodo N_0. Este pico cepstral indica directamente el período fundamental de la señal. En resumen, identificando el primer pico significativo en el cepstrum (omitiendo el *cepstrum peak* en $n = 0$, que corresponde a la componente de energía), se obtiene T_0 y por ende $F_0 = {}^{1}/_{T_0}$. El método cepstral puede fallar en estimar con precisión F_0 en señales muy ruidosas o con componentes no sonoras, por lo que a veces se combina con un filtrado de la señal.

Además de los anteriores, existen múltiples algoritmos para la detección de F0. Algunos trabajan en el dominio temporal, como el Método de la Diferencia Acumulada (AMDF) y su refinamiento en el algoritmo YIN (Mauch & Dixon, 2014), que buscan minimizar la diferencia entre la señal y versiones retrasadas de sí misma. Otros operan en el dominio frecuencial, como el método de Producto Armónico (HPS) (Noll, 1970), que consiste en submuestrear el espectro y multiplicarlo para realzar la componente fundamental. También se emplean filtros en peine (*comb filters*) o enfoques basados en modelos estadísticos y aprendizaje automático para entornos ruidosos (Zambrano et al., 2017).

Cada uno de los métodos presenta ventajas en función del tipo de señal analizada: AMDF/YIN tienden a ser robustos y precisos con señales ruidosas; la autocorrelación es conceptualmente simple y efectiva en voz limpia; y el cepstrum puede detectar F_0 incluso cuando los armónicos no están perfectamente alineados con múltiplos enteros.

4. Clasificadores Machine Learning

En este capítulo se presentan los clasificadores de machine learning utilizados en las publicaciones que respaldan esta investigación. Se inicia con una introducción a las Máquinas de Vectores de Soporte (*Support Vector Machines*, SVM), abordando tanto su versión lineal como su extensión no lineal, mediante el uso de funciones núcleo (*kernel functions*). Se analizan los fundamentos matemáticos que sustentan estos modelos, así como su funcionamiento general. La notación de esta sección está basada en el trabajo de Theodoridis (Theodoridis & Koutroumbas, 2006).

A continuación, se estudian los Árboles de Decisión y su evolución hacia modelos más complejos y robustos, como el *Random Forest* (RF). Se describe en primer lugar el árbol de decisión como modelo base, detallando su mecanismo de partición recursiva, su interpretación y su vulnerabilidad al sobreajuste. Posteriormente, se introduce el modelo RF, explicando su estructura como conjunto de árboles y su funcionamiento mediante técnicas de *bagging*.

El resto de los algoritmos de clasificación empleados en este trabajo, basados en redes neuronales, se describen en el capítulo 5.

4.1 Support Vector Machine

El modelo *Support Vector Machine* (SVM), propuesto por (Cortes, 1995), se desarrolla para abordar tanto problemas de clasificación como de regresión. Su fundamento teórico se basa en la teoría del aprendizaje estadístico, y su objetivo principal es encontrar el hiperplano óptimo que maximice el margen entre las clases definidas en el problema. Esta frontera de decisión se construye a partir de un subconjunto reducido de los datos de entrenamiento, conocidos como vectores de soporte (*support vector*), que son los puntos más cercanos al límite de separación y determinan la posición del hiperplano. La popularidad del modelo SVM se atribuye, en gran medida, a su solidez desde el punto de vista teórico y a su notable flexibilidad para abordar una amplia gama de tareas, entre ellas la clasificación y la predicción. Una de sus principales fortalezas radica en su capacidad para minimizar simultáneamente el error empírico de clasificación y maximizar el margen geométrico entre clases (Yang, 2009), lo cual contribuye a una mejor generalización del modelo.

El SVM emplea un *kernel* para realizar la regresión y la clasificación, transformando los datos en un espacio de mayor dimensión mediante técnicas de transformación no lineal que permitan separar las clases de forma lineal. Esta separación lineal se consigue porque, al transformar los datos en dimensiones superiores, tienden a dispersarse, lo que permite encontrar el espaciado lineal de separación entre las clases (Gualtieri & Cromp, 1999).

Se define un hiperplano como el mayor margen entre las dos clases. La Figura 15 muestra el concepto de hiperplano, donde la línea en negrita muestra la recta óptima que separa los datos de las dos clases. De forma general, en un espacio de dimensión n, dicha superficie de separación es un hiperplano. En esta sección se muestran los ejemplos con imágenes para el caso bidimensional $\mathbb{R}^2$, en el que este hiperplano se reduce a una línea recta. Esta línea de decisión se define de tal manera que maximiza la distancia entre sí misma y los puntos más cercanos de ambas clases.

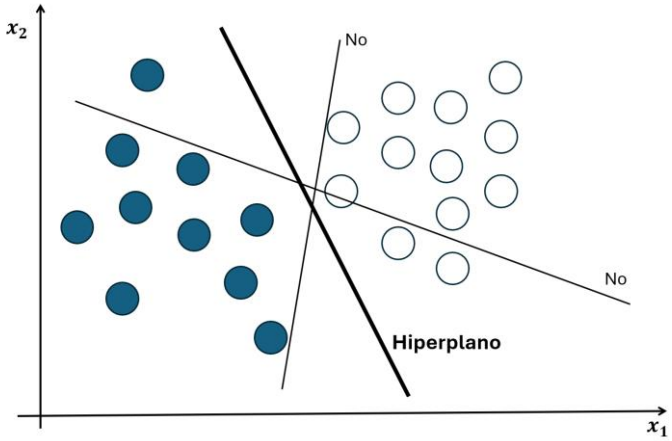

Figura 15. Concepto de hiperplano para la separación de datos pertenecientes a dos clases. Se muestra una línea recta que separa las dos clases, y dos líneas que no consiguen una separación óptima.

SVM es un algoritmo de aprendizaje automático supervisado, al que se le da un conjunto de datos de entradas con sus anotaciones correspondientes. Los valores de entrada tienen forma de vectores de descriptores. SVM construye un hiperplano que separa dos clases para lograr la máxima separación entre ellas. Al separar las clases con un amplio margen, se minimiza el error de generalización. El objetivo es conseguir el mínimo error de generalización mientras se predice la clase correcta de los datos, sin ningún error, o con el mínimo error de generalización (Soman et al., 2009).

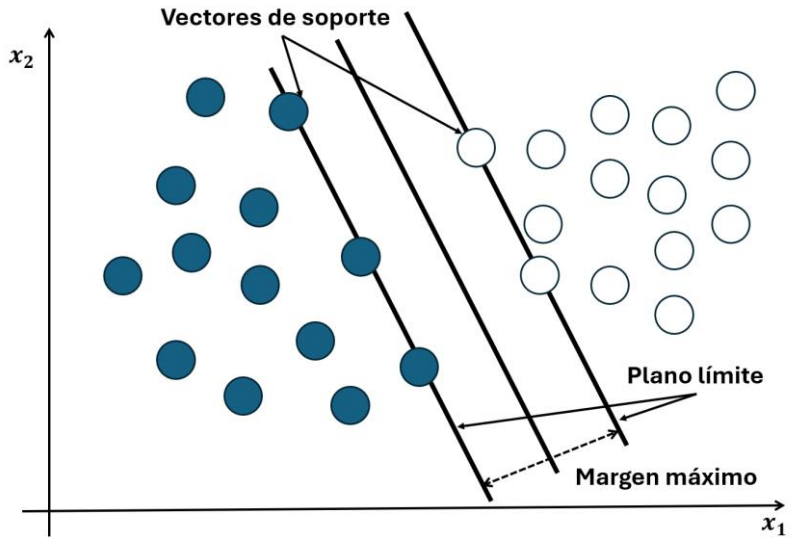

Figura 16. Vectores de soporte, planos delimitadores y margen máximo en un hiperplano.

Los dos planos paralelos al clasificador y que pasan por uno o varios puntos de datos se denominan planos límite. La distancia entre estos planos se denomina margen. Mediante el proceso de aprendizaje se evalúa el hiperplano que maximiza este margen. Los datos se encuentran en los planos delimitadores se denominan vectores de soporte (*support vectors*). Estos puntos son cruciales para formar un hiperplano. La Figura 16 muestra el concepto de vectores soporte, planos delimitadores y margen máximo. Una buena generalización está garantizada si se encuentra un conjunto pequeño de vectores de soporte que definan el hiperplano.

A pesar de tomar todas las medidas necesarias para una clasificación exacta, es probable que se produzcan clasificaciones erróneas. Un clasificador SVM permite definir un hiperplano para separar las clases existiendo errores de clasificación de las muestras. La Figura 17 muestra dos clases para la clasificación.

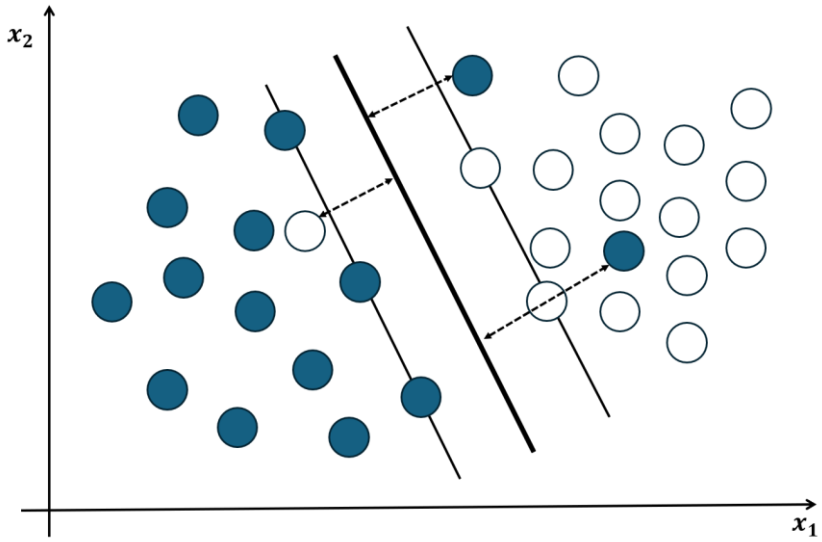

Figura 17. Hiperplano con errores de clasificación. Las flechas indican la distancia entre las muestras mal clasificadas y el hiperplano de clasificación.

El hiperplano se define como una línea recta para este ejemplo de clasificación. Este es un ejemplo de clasificación con errores, ya que no se puede definir una línea que separe las dos clases totalmente. La solución al problema de clasificación es emplear una línea curva que permita un ajuste mayor a los puntos mal clasificados. Este tipo de clasificación se conoce como clasificación SVM no lineal. En general, el SVM es puede ser de dos tipos: lineal y no lineal. Un SVM no lineal se consigue utilizando un *kernel* específico para realizar una transformación de los puntos en el hiperplano, de manera que se pueda emplear una separación lineal.

4.1.1 SVM lineal

Cuando el SVM utiliza un hiperplano linear como superficie de decisión, se denomina SVM lineal. El SVM lineal es aplicable a dos tipos de datos: separables y no separables, como se define a continuación.

a) Datos separables

Dado un conjunto de datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^{l}$, con $x_i \in \mathbb{R}^n$ e $y_i \in \{+1, -1\}$, el objetivo de un SVM lineal es encontrar el hiperplano que separa ambas clases maximizando el margen, es decir, la distancia mínima desde el hiperplano a los puntos más cercanos de cada clase, como se define en la Ecuación (22):

$$w^T x + b = 0 \tag{22}$$

Donde, w es un vector normal al hiperplano, x es un punto perteneciente al hiperplano, y b es el término de sesgo (*bias)* que lo desplaza respecto al origen. La distancia ortogonal del hiperplano al origen es $|b|/\|w\|$.

El clasificador resultante se define en la Ecuación (23):

$$f(x) = sgn(w^T x + b) \tag{23}$$

Donde $sgn(\cdot)$ es la función signo, que devuelve +1 si su argumento es positivo, -1 si es negativo, y 0 si es cero.

Sea x un punto cualquiera del hiperplano definido en la Ecuación (22). Se puede definir la distancia ortogonal d_i de una muestra x_i a dicho hiperplano mediante la Ecuación (24):

$$d_i = y_i \frac{w}{\|w\|} \cdot (x_i - x) = \frac{y_i(w^T x_i + b)}{\|w\|} \tag{24}$$

El margen geométrico γ se define como la distancia ortogonal mínima que separa el hiperplano de clasificación de los puntos de entrenamiento más cercanos, los vectores de soporte, como muestra la Ecuación (25):

$$\gamma = \min_i d_i = \min_i \frac{y_i(w^T x_i + b)}{\|w\|} \tag{25}$$

El problema del SVM consiste en maximizar dicho margen.

$$\max_{w,b} \min_i \frac{y_i(w^T x_i + b)}{\|w\|} \tag{26}$$

Imponiendo la convención $min_i y_i(\cdot) = 1$, se llega a las restricciones canónicas

$$y_i(w^T \cdot x_i + b) \geq -1, \qquad i = 1, \dots, l \tag{27}$$

Donde los vectores que satisfacen la Ecuación (28) son los vectores de soporte.

$$y_i(w^T \cdot x_i + b) = 1 \tag{28}$$

A partir de la expresión de la Ecuación (25), el problema para encontrar el hiperplano óptimo en el caso de datos perfectamente separables se formula con la siguiente optimización cuadrática.

$$\begin{gathered} \min_{w,b} \frac{1}{2}\|w\|^2 \\ sujeto\ a \quad y_i(w^T \cdot x + b) \geq 1\,, \qquad i = 1, 2, \dots, l \end{gathered} \tag{29}$$

Minimizar $\frac{1}{2}\|w\|^2$ equivale a maximizar el margen entre las dos clases, mientras que las restricciones garantizan que todos los datos de entrenamiento quedan correctamente clasificados.

Para resolver la Ecuación (29) se emplea la teoría de optimización con multiplicadores de Lagrange. Se introduce un vector de coeficientes $\alpha = (\alpha_1, \dots, \alpha_l)^T$ con $\alpha_i \geq 0$ y se define la función Lagrangiana de la forma:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(w^T \cdot x + b) - 1] \tag{30}$$

Las condiciones de Karush-Kuhn-Tucker (KKT) exigen:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial w} = 0 \Longrightarrow w = \displaystyle\sum_{i=1}^{l} \alpha_i y_i x_i \\ \dfrac{\partial \mathcal{L}}{\partial b} = 0 \Longrightarrow \displaystyle\sum_{i=1}^{l} \alpha_i y_i = 0 \\ \alpha_i [y_i (w^T \cdot x + b) - 1] = 0, \;\; \alpha_i \geq 0 \end{cases} \tag{31}$$

Sustituyendo en la Ecuación (30) se obtiene el problema dual.

$$\begin{gathered} \max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ sujeto\ a \;\; \sum_{i=1}^{l} \alpha_i y_i = 0, \;\; \alpha_i \geq 0 \; (i = 1, \dots, l) \end{gathered} \tag{32}$$

Una vez resuelta la Ecuación (32) se obtienen los valores de w y b con la Ecuación (31), y el clasificador resultante es:

$$f(x) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i x_i^T x + b \right) \tag{33}$$

b) Datos no separables

Cuando los datos no son linealmente separables se recurre al clasificador de margen suave, que introduce una variable de holgura $\xi_i \geq 0$ que permite una cierta cantidad de violaciones del margen, es decir, errores de clasificación.

Las restricciones originales del caso de datos separables $y_i(w^T \cdot x + b) \geq 1$ se relajan a:

$$y_i (w^T \cdot x_i + b) \geq 1 - \xi_i \,, \;\; \xi_i \geq 0, \;\; i = 1, \dots, l \tag{34}$$

Por lo que se puede escribir la Ecuación (29) de la siguiente forma:

$$\begin{gathered} \min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{l} \xi_i \\ sujeto\ a \;\; y_i (w^T \cdot x + b) \geq 1 - \xi_i \,, \quad \xi_i \geq 0, \;\; i = 1, \dots, l \end{gathered} \tag{35}$$

Donde el parámetro $C \geq 0$ controla el compromiso entre el margen y la penalización de errores. Valores pequeños de C permiten más violaciones del margen por lo que generan un modelo más flexible, mientras que valores grandes producen un modelo más estricto con riesgo de sobreajuste.

Introduciendo multiplicadores de Lagrange se obtiene el problema dual:

$$\max_{\alpha} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$sujeto\ a \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C \ (i = 1, \dots, l) \tag{36}$$

Comparado con el caso separable, la única diferencia es la cota superior $\alpha_i > C$, cuando $C \to \infty$ se recupera un margen duro para violaciones de los márgenes, en el que no se permiten errores de clasificación.

La regla de clasificación se define de la misma manera que en el caso separable, con α_i limitado por C, lo que evita que un pequeño subconjunto de muestras determine un hiperplano demasiado específico y mejora la capacidad de generalización del modelo.

$$f(x) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i x_i^T x + b\right) \tag{37}$$

4.1.2 SVM no lineal

Cuando los datos no son linealmente separables el clasificador SVM descrito en la sección anterior puede resultar insuficiente: existen casos en los que ningún hiperplano lineal logra una separación adecuada. Para resolver esta limitación se recurre a un SVM no lineal, cuyo hiperplano de separación se convierte en una superficie de decisión no lineal gracias al uso de funciones kernel (Cristianini, 2000).

Partiendo de las ecuaciones duales (32) y (36), obsérvese que el término clave es el producto escalar $x_i^T x_j$. Se introduce entonces una transformación no lineal

$$\Phi: \mathbb{R}^n \to \mathcal{H} \tag{38}$$

Donde $\mathcal{H}$ es un espacio de características de (posiblemente) dimensión mucho mayor. Al sustituir $x_i^T x_j$ por $\Phi(x_i)^T \Phi(x_j)$, el problema dual se expresa únicamente mediante esos productos en $\mathcal{H}$. El clasificador resultado adopta la forma de:

$$f(x) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i \Phi(x_i)^T \Phi(x_j) + b\right) \tag{39}$$

Se define la función *kernel* K de la forma:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{40}$$

De modo que el clasificador expresado mediante la Ecuación (39) se puede escribir como:

$$f(x) = sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b\right) \tag{41}$$

La matriz *kernel* adopta la forma:

$$K = \begin{bmatrix} \mathrm{K}(x_1, x_1) & \cdots & \mathrm{K}(x_1, x_l) \\ \vdots & \ddots & \vdots \\ \mathrm{K}(x_l, x_1) & \cdots & \mathrm{K}(x_l, x_l) \end{bmatrix} \tag{42}$$

Como $\Phi(\cdot)$ no se computa explícitamente, el coste de calcular en $\mathcal{H}$ se evita evaluando directamente $\mathrm{K}(x_i, x_j)$. Este es el conocido *kernel trick*, que permite trabajar con superficies de decisión no lineales sin elevar drásticamente la complejidad computacional.

La Figura 18 muestra la diferencia entre un hiperplano lineal y un hiperplano no lineal de un clasificador SVM.

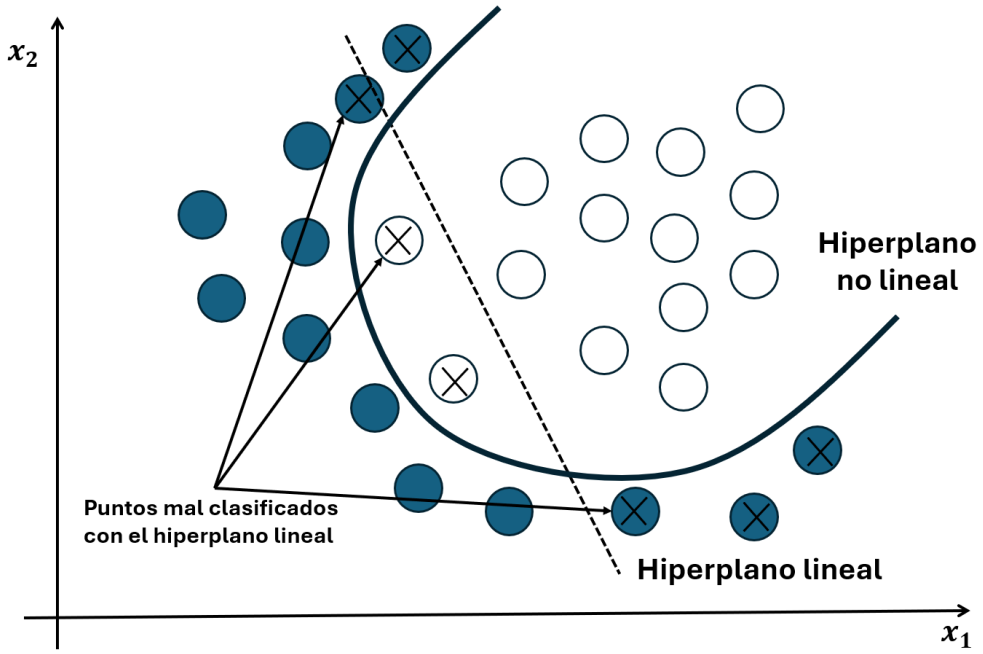

Figura 18. Hiperplano no lineal de un clasificador SVM.

En la Figura 19 se muestra un mapeo no lineal del espacio de características.

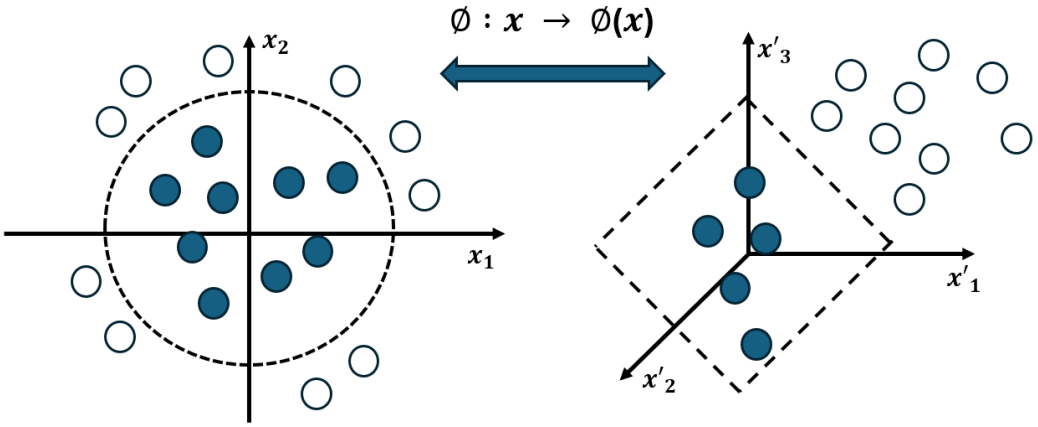

Figura 19. Mapeo no lineal del espacio de características.

La descripción anterior se basa en un clasificador binario. Para problemas con M clases, el enfoque habitual consiste en descomponer la tarea en $\binom{M}{2} = M(M-1)/2$ clasificadores binarios (*one vs one*), entrenando un SVM para cada par de clases y combinando sus salidas.

A continuación se describen los tipos de kernel que se emplean comúnmente:

a) Kernel lineal

El kernel linear es la forma más simple de función núcleo y se define como muestra la Ecuación (43):

$$K(x_i, x_j) = x^T y + c, \quad c \geq 0 \tag{43}$$

En este caso la transformación implícita es la identidad, $\Phi(\mathrm{x}) = x$; por tanto, el SVM opera directamente en el espacio original $\mathbb{R}^n$ y el hiperplano de decisión permanece lineal.

Este kernel es especialmente recomendable cuando: se busca un modelo base rápido para tareas de clasificación en tiempo real; el número de atributos n es grande respecto al número de muestras l; o los datos son casi separables en el espacio original y se busca inspeccionar la contribución de cada característica al hiperplano de separación.

b) Kernel polinómico

Dado un vector $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$. Se puede definir una transformación no lineal de la forma:

$$\Phi(\mathrm{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix} \tag{44}$$

Que mapea $\mathbb{R}^2$ en $\mathbb{R}^3$. El producto escalar en el espacio transformado se convierte en:

$$\Phi(x_i)^T \Phi(x_j) = x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2 x_{j2}^2 \left(x_i^T x_j\right)^2 \tag{45}$$

Lo que define el kernel polinómico de grado 2:

$$K(x_i, x_j) = \left(x_i^T x_j\right)^2 \tag{46}$$

Obsérvese que no es necesario calcular explícitamente Φ; basta con evaluar directamente $K(x_i, x_j)$.

c) Kernel Gaussiano (Radial Basis Function, RBF)

El kernel Gaussiano o de función de base radial (RBF) se define como muestra la Ecuación (47):

$$K(x, y) = exp[-\gamma \|x - y\|^2] \tag{47}$$

Donde se suele determinar $\gamma = 1/(2\sigma^2)$, y controla la "anchura" del núcleo: valores pequeños generan superficies de decisión suaves, mientras que valores grandes producen fronteras muy ajustadas a los datos.

Desarrollando la Ecuación (47):

$$\begin{aligned} K(x, y) &= exp[-\gamma \|x\|^2] exp[-\gamma \|y\|^2] \exp(2\gamma x^T y) \\ &= exp[-\gamma \|x\|^2] exp[-\gamma \|y\|^2] \sum_{m=0}^{\infty} \frac{(2\gamma x^T y)^m}{m!} \end{aligned} \tag{48}$$

El término $\exp(2\gamma x^T y)$ es una suma infinita de potencias $(x^T y)^m$, por lo que el RBF equivale a proyectar los datos en un espacio de características de dimensión infinita donde cada potencia m introduce un componente polinómico de grado m.

Además, puesto que $exp[-\gamma \|x\|^2]$ y $exp[-\gamma \|y\|^2]$ dependen de un solo argumento, ambas son núcleos válidos; además, el producto de kernels es también un kernel (Soman et al., 2009). Así, la Ecuación (47) es un kernel válido porque es el producto de tres kernels.

El kernel RBF es una elección práctica cuando: los datos poseen gran número de características o relaciones no lineales difíciles de modelar; se requiere un clasificador universal, ya que teóricamente puede aproximar cualquier frontera de decisión con γ y C (penalización de errores) adecuados; o no se conoce a priori la forma de la superficie óptima. Su flexibilidad proviene justamente del mapeo a dimensión infinita, mientras que el *kernel trick* evita computar explícitamente dicha proyección: basta evaluar la distancia euclídea $\|x - y\|$, manteniendo el coste computacional.

d) Otros kernel

Existen otra serie de kernels basados en: el laplaciano, similar al RBF; la función sigmoide, inspirada en las redes neuronales; chi-cuadrado, especialmente usada en visión por computador con vectores de histograma; y *string*, adecuada para secuencia de texto o ADN.

Cada kernel induce una geometría diferente en el espacio de características, por lo que la capacidad de generalización del SVM dependerá de la selección adecuada de la función núcleo y de sus hiperparámetros, los que normalmente se ajustan mediante validación cruzada.

4.2 Árboles de decisión y modelos Random Forest (RF)

Los árboles de decisión son una herramienta que permite extraer conocimiento a partir de un conjunto de observaciones mediante una estrategia jerárquica de partición: se dividen los datos con decisiones binarias simples (sí/no) hasta llegar a regiones cada vez más homogéneas (Breiman, 2017). Su popularidad en aprendizaje automático se debe a que son muy intuitivos, rápidos y altamente escalables cuando el conjunto de datos es muy grande, además de admitir una formulación probabilística que incorpora la incertidumbre. No obstante, aprender el árbol de decisión óptimo para un conjunto de datos es un problema NP-completo (Koch et al., 2023) y puede dar lugar a modelos complejos propensos al sobreajuste.

El término *Random Forest* (RF) designa un método de aprendizaje en conjunto (*ensemble learning*) que combina numerosos árboles de decisión simples, de modo que se evita optimizar un único árbol de decisión complejo (Ho, 1995, 1998; Young et al., 2002). En estos trabajos, los autores proponen inyectar aleatoriedad en el proceso de aprendizaje para crear árboles no-correlacionados. Al promediar las predicciones individuales de cada árbol de decisión se obtiene una mayor generalización y, por lo tanto, una mayor exactitud. En (Breiman, 2001) se propone el concepto de *bootstrap aggregation*, que consiste en entrenar cada árbol de forma independiente con un subconjunto aleatorio del conjunto total de datos de entrenamiento.

El empleo de RF aporta diversas ventajas frente a un modelo SVM: resulta más adecuado cuando el conjunto de datos es heterogéneo, con interacciones no lineales entre variables y fronteras de decisión que no pueden describirse mediante un único hiperplano; ofrece robustez frente al sobreajuste y a los valores atípicos sin requerir un ajuste exhaustivo de hiperparámetros; y, finalmente, facilita tanto la interpretabilidad parcial (mediante la importancia de variables) como la generación de estimaciones rápidas y estables en conjuntos de datos de gran volumen. En el presente trabajo no se dispone de un análisis de características claramente separables ni de distribuciones homogéneas que garanticen la eficacia de un SVM con un único kernel. Por el contrario, la variabilidad inherente de las medidas acústicas y las posibles interacciones entre ellas aconsejan adoptar RF como una solución más flexible y robusta.

A continuación, se define el árbol de decisión como un grafo y se formaliza el enfoque de partición. También se detalla el modelo nodo/hoja (*node/leaf*) y, por último, se analiza cómo un conjunto de árboles independientes puede combinarse para obtener un clasificador RF.

4.2.1 Árbol de Decisión

Considerando un espacio de características de entrada $X \in R^D$ y un espacio de salida $Y \in R^{D'}$, el objetivo del árbol de decisión es aprender un modelo que sea capaz de realizar predicciones en Y dada una observación en X. Esta tarea se puede formular de forma estadística como un problema de máximo a posteriori:

$$\hat{Y} = argmax_{y \in Y} P(Y|X) \tag{49}$$

Dado un conjunto de datos de entrenamiento $\left(X^{(n)}, Y^{(n)}\right)_{n=1}^{N} \in X \times Y$, el objetivo es aprender la probabilidad posterior $P(Y|X)$. Encontrar un modelo adecuado para este resultado posterior y aprenderlo en todo el espacio de características X es una tarea compleja. Para resolver este problema, un árbol de decisión sigue una estrategia de "divide y vencerás": (1) crea una partición sobre el espacio de características de entrada utilizando un conjunto de decisiones, y (2) estima $P(Y|X)$ en cada partición de este espacio.

4.2.1.1 Modelo del árbol de decisión

Los árboles de decisión están basados en la siguiente idea: realizar predicciones utilizando una secuencia de decisiones simples. Un modelo de árbol de decisión consiste en un conjunto de decisiones (binarias) organizadas de forma jerárquica. Por lo tanto, un árbol de decisión puede definirse formalmente como un grafo acíclico dirigido, compuesto por un conjunto de nodos N y un conjunto de aristas dirigidas (Pauly, 2012). Cada nodo codifica una división binaria y está conectado por una arista dirigida a un único nodo padre de un nivel superior y, a lo sumo, dos nodos hijos del nivel inferior. El término dirigido implica que (1) el árbol sólo puede recorrerse en sentido descendente, es decir, el sentido padre-hijo, y (2) que los nodos de distintos niveles son no intercambiables. El árbol de decisión es acíclico ya que no hay ciclos dentro del modelo. El nodo en la parte superior de un árbol se llama raíz, los nodos en la parte inferior se llaman hojas. Una muestra de datos puede atravesar el árbol en sentido descendente,

siguiendo un único camino que se determina por las decisiones tomadas en cada nodo atravesado, hasta llegar a una hoja, como se ilustra en la Figura 20.

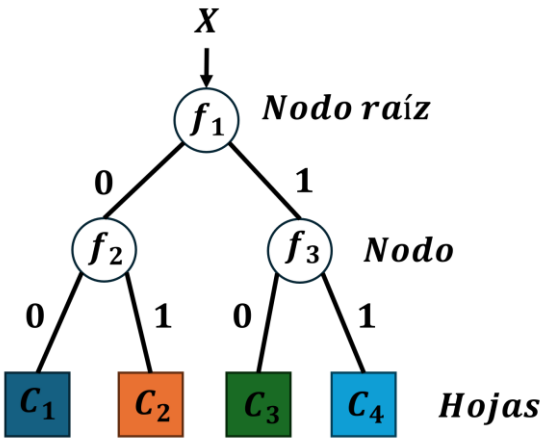

Figura 20. División de los datos de entrada en un árbol de decisión.

Durante la fase de aprendizaje, los datos que llegan a una hoja se utilizan para modelar la distribución posterior de forma local. Durante la fase de test, estas distribuciones posteriores permiten hacer predicciones sobre nuevas observaciones no vistas que llegan a una hoja dada.

4.2.1.2 Función de división del nodo

Para realizar una decisión binaria, un nodo N_l del conjunto de nodos N de un árbol está equipado con una función de división f_l cuyo papel es dividir las observaciones entrantes S_l en dos subconjuntos S_l^{izq} y S_l^{der}, siendo estos dos subconjuntos discontinuos, es decir $S_l = S_l^{izq} \cup S_l^{der}$ y $S_l^{izq} \cap S_l^{der} = \emptyset$. Estos subconjuntos corresponden, respectivamente, al hijo izquierdo y al hijo derecho del nodo N_l. La función de división f_l de define en la Ecuación (50):

$$\begin{cases} f_l : X \rightarrow \mathcal{B} \\ f_l(X) = 0, \;\; X \; hacia \; la \; izquierda \\ f_l(X) = 1, \;\; X \; hacia \; la \; derecha \end{cases} \tag{50}$$

Existen muchas posibilidades para la clase de funciones de decisión (Criminisi et al., 2011). Sin embargo, la opción más común es la proyección lineal unida a un umbral, como muestra la Ecuación (51).

$$f_l(X) = X \cdot v_l \geq \tau_l \tag{51}$$

Donde $\dim(v_l) = \dim(X)$ y $\tau_l \in \mathcal{R}$. Si v_l presenta todos los valores distintos de cero, entonces la función de división es un hiperplano que tiene en cuenta todas las características de entrada de los datos. Sin embargo, si v_l es dispersa (presenta unos y ceros), f_l realiza la división utilizando sólo un subconjunto de características. En el caso extremo en el que $v_l$sólo presenta un componente distinto de cero, la división se realiza basándose únicamente en una característica, una única dimensión de X. También se

puede utilizar funciones de división más complejas, como las no lineales. Sin embargo, la filosofía del árbol fomenta la elección de funciones sencillas que puedan calcularse y optimizarse eficazmente.

El proceso de aprendizaje del árbol se puede definir como una tarea de optimización y división iterativa de nodos. Dependiendo del tipo de función de división elegida, hay que determinar varios parámetros, por lo que el entrenamiento se puede convertir en un problema de optimización complejo en un espacio de búsqueda de alta dimensionalidad. Para evitarlo se emplea una estrategia de búsqueda codiciosa (*greedy search*). Cada nodo genera una serie de funciones candidatas y las evalúa con los datos de entrada maximizando una función objetivo predefinida, obteniendo el mejor candidato. Durante la fase de entrenamiento, las funciones de división de cada nodo se optimizan para dividir iterativamente el entrenamiento hasta que se alcanza el criterio de parada.

4.2.1.1 Criterio de parada

Una vez alcanzada la parte inferior del árbol, se detiene la división iterativa de los datos de entrenamiento y el nodo actual se convierte en un nodo hoja. Comúnmente se definen tres criterios de parada en el proceso de división iterativa: (1) profundidad máxima del árbol, (2) población mínima por hoja, y (3) variación mínima de la función objetivo. El primer criterio únicamente tiene en cuenta la profundidad de la jerarquía y, una vez alcanzada cierta profundidad, se detiene. El segundo criterio se basa en el número de instancias de entrenamiento que llegan a un nodo, y si la población de puntos de entrenamiento está por debajo de un determinado umbral, la división se detiene. El último criterio se refiere a la función objetivo que se optimiza. Si su variación es inferior a un umbral determinado, se considera que no se obtiene información adicional tras dividir las instancias de entrenamiento.

Mientras que el papel de los nodos intermedios es dividir y enviar las observaciones hacia abajo del árbol, el papel de las hojas es modelar la distribución posterior dado un subconjunto del conjunto de entrenamiento. Como estas decisiones se toman sobre el espacio de características de entrada, todos los puntos de entrenamiento que llegan en una hoja son coherentes en X. Por lo tanto, cada hoja corresponde a una parte del espacio de características, como se ilustra en la Figura 21, y el conjunto de hojas de un árbol de decisión construye una partición P del total X. Se define esta partición como un conjunto $P = \cup_{z=1}^{Z} C^{(z)}$, donde cada $C^{(z)}$ corresponde a una hoja del árbol de decisión. Obsérvese que las $C^{(z)}$ cubren todo el espacio de características X y no se solapan.

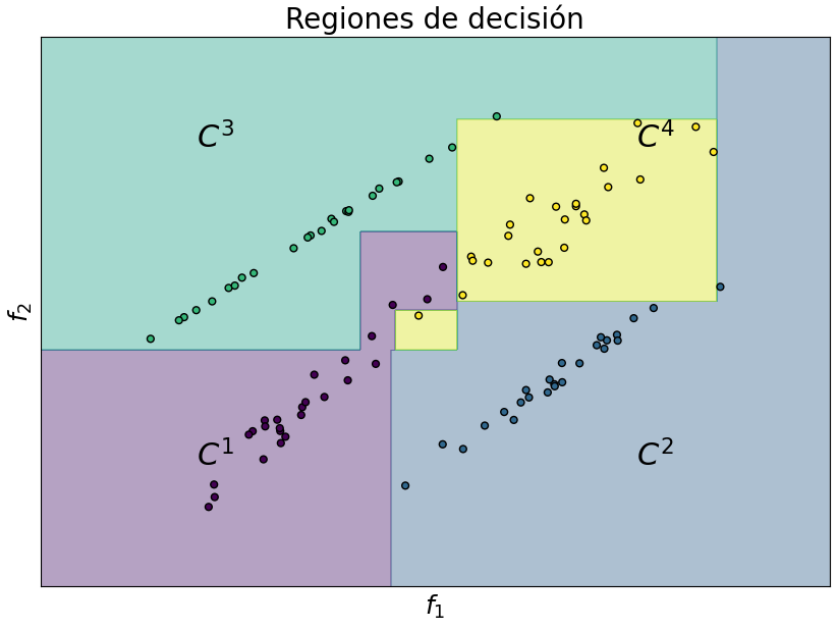

Figura 21. División del espacio de características.

Si la partición P se construye con un subconjunto demasiado reducido de descriptores, las fronteras de decisión presentan una gran incertidumbre. Sin embargo, si se emplean demasiados descriptores, cada región de decisión incluye un número reducido de muestras, lo que favorece el sobreajuste del modelo y disminuye la capacidad de generalización. Una vez definidas todas las particiones, cada hoja del árbol $C^{(z)}$ permite contabilizar las instancias de cada clase que han caído en ella y estimar un vector de probabilidad a posteriori.

Una vez que se ha entrenado un árbol de decisión, la predicción de una nueva observación X, no vista anteriormente por el modelo, se puede realizar fácilmente. En función de los resultados de las distintas funciones de decisión, X se envía hacia abajo en el árbol, siguiendo un camino que es único y conduce a una hoja $C^{(z)}$. Por lo tanto, en el momento de la evaluación, se puede ver un árbol F como una función sobreyectiva que toma como entrada una observación y devuelve una celda, o región de clasificación:

$$\begin{cases} F : X \rightarrow \{C^{(1)}, \dots, C^{(z)}, \dots, C^{(Z)}\} \\ F(X) = C^{(z)} \end{cases} \quad (52)$$

El modelo de probabilidad posterior almacenado en esta hoja permite realizar una predicción utilizando el máximo a posteriori:

$$\hat{Y} = argmax_Y P(Y|X \in C^{(z)}, P) \quad (53)$$

Finalmente, cabe destacar que los árboles de decisión pueden aproximar cualquier función arbitraria si se dispone de suficientes datos de entrenamiento. Los árboles de decisión pueden considerarse modelos no paramétricos, ya que su tamaño depende de

la cantidad de datos de entrenamiento. Como se explica en la introducción de esta sección, entrenar un árbol óptimo es un problema NP-completo, y los árboles de decisión son propensos al sobreajuste. Otro de las desventajas que presentan es su alta variabilidad. Al tratarse de sistemas jerárquicos, un pequeño cambio en los datos de entrenamiento puede resultar en una estructura de árbol muy diferente (Theodoridis & Koutroumbas, 2006). En la siguiente sección se define como construir un modelo RF formado por conjunto de árboles de decisión no-correlacionados para lograr una mayor generalización.

4.2.2 Random Forest

Un *Random Forest* (RF) F es un conjunto de T árboles de decisión independientes $F = \{F_1, \dots, F_t, \dots, F_T\}$. Como se demuestra en (Breiman, 2001), la sustitución de un único árbol por un conjunto de árboles descorrelacionados proporciona una mejor generalización. Durante el proceso de aprendizaje, se puede inyectar aleatoriedad para lograr una mayor independencia entre los árboles construidos a partir del mismo conjunto de entrenamiento.

4.2.2.1 Entrenamiento y aleatorización

Para construir árboles descorrelacionados, o independientes, basados en un conjunto de entrenamiento único existen varios enfoques de aleatorización. En (Breiman, 2001) se introduce el concepto de *bootstrap aggreagting*, o *bagging,* que se define a continuación. Dado un conjunto de entrenamiento $S = \{X^{(n)}, Y^{(n)}\}_{n=1}^{N}$, se define el *bootstrap* como un subconjunto S_t, del conjunto total de datos de entrenamiento, en el que se han muestreado aleatoriamente elementos empleando una distribución uniforme, con o sin sustitución. De esta forma cada uno de los árboles descorrelados se entrena utilizando un subconjunto S_t diferente. Finalmente, las predicciones de todos los árboles individuales se agregan utilizando el promedio.

Como se indica en (Geurts et al., 2006), también se puede inyectar aleatoriedad en la optimización de nodos. Esta fase se basa en una estrategia codiciosa, es decir, se genera un conjunto de funciones de división candidatas y se elige la mejor de acuerdo con una función objetivo predefinida. Por lo tanto, la inyección de aleatoriedad en la generación de la función candidata es una opción obvia. Si se toma de ejemplo las proyecciones lineales seguidas de una umbralización, como se muestra en la Figura 21, se puede extraer aleatoriamente el valor del umbral utilizando cualquier tipo de distribución. Esto anima a los árboles a seleccionar distintos tipos de características y a ponderarlas de forma diferente. Además, la elección del umbral también puede ser aleatoria en lugar de optimizarla o tomar la media de los valores proyectados.

El efecto de inyectar aleatoriedad en el entrenamiento del árbol tiene varias ventajas: 1) aumentar el grado de aleatoriedad disminuye la correlación entre los distintos árboles y,

por tanto, proporciona una mayor generalización; y 2) permite obtener independencia del entrenamiento, es decir, ganar robustez frente a datos ruidosos.

4.2.2.2 Configuración de parámetros

Los clasificadores RF ofrecen un marco muy flexible para diseñar funciones objetivo, diferentes funciones de división o modelos posteriores. Además, sólo poseen unos pocos hiperparámetros cuya influencia se ha estudiado exhaustivamente (Criminisi et al., 2011). Estos parámetros son dos: (1) el número de árboles, y (2) la profundidad del árbol. Como se muestra en la Figura 22, el aumento del número de árboles permite promediar las predicciones ruidosas, y por lo tanto producen una disminución monótona del error de predicción. La profundidad máxima del árbol es un parámetro fundamental que debe optimizarse, ya que afecta directamente a la capacidad de generalización de cada árbol. Mientras que un árbol corto no será muy fiable en su predicción porque sus hojas aún contienen muchos datos heterogéneos, un árbol muy profundo tendrá muy pocos datos de entrenamiento en sus hojas para calcular estadísticas fiables, por lo que se produce el fenómeno de sobreajuste. La profundidad óptima del árbol produce un buen modelado de las observaciones y una gran generalización.

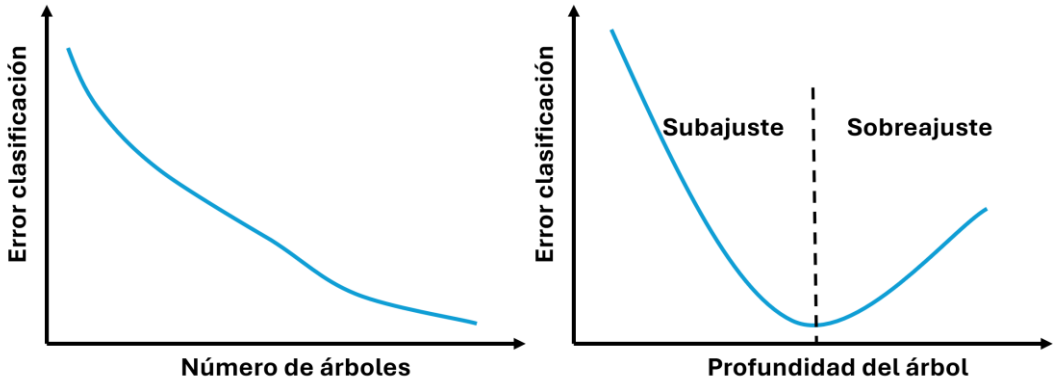

Figura 22. Relación entre el error de clasificación y, la profundidad del árbol de decisión, y el número de árboles en un RF.

4.2.2.3 Predicción mediante Random Forest

Se considera un clasificador RF compuesto de T árboles $F = \{F_t\}_{t=1}^{T}$, en el que cada árbol F_t ha sido entrenado con una partición P_t del del espacio de descriptores X. Como cada árbol individual puede verse como una función sobreyectiva que asocia una observación $x \in X$ a una región del espacio $C_t^{(z_t)}$ de la partición P_t, el RF al completo es una función que asocia X a un conjunto de regiones del espacio:

$$F(X) = \left\{C_1^{(z_1)}, \dots, C_t^{(z_t)}, \dots., C_T^{(z_T)}\right\} \tag{54}$$

Si consideramos que cada partición P_t es equiprobable, la predicción del clasificador puede calcularse de forma sencilla promediando los árboles posteriores:

$$P(Y|X) = \frac{1}{T}\sum_{t=1}^{T} P\left(Y|X \in C_T^{(z_T)}, P_t\right) \tag{55}$$

El uso de la media se debe a que generalmente es un buen compromiso entre dar más peso al árbol con mejor clasificación y reducir las contribuciones ruidosas (Criminisi et al., 2011). Sin embargo, es posible realizar otra función de agregación de árboles, por ejemplo, ponderando las contribuciones de cada árbol individual en función de su *accuracy,* u otra métrica.

4.2.3 Random Forest para problemas de clasificación

En el campo del aprendizaje automático, los RF se han aplicado principalmente a tareas de clasificación. Además de ser escalable a grandes conjuntos de datos, su capacidad de generalización, y beneficiarse de un rápido entrenamiento; se adaptan especialmente bien a los problemas multiclase, ya que proporcionan resultados probabilísticos. Si bien, una clasificación en dos clases es solamente un caso particular de la generalización de la clasificación multiclase. Por ello, a continuación se realiza una formulación probabilística de la clasificación multiclase.

En una tarea de clasificación, se considera un espacio de características de entrada $\mathcal{X} \in \mathbb{R}^D$y un espacio $Y \in \mathbb{R}^D$ que es un conjunto finito de K valores discretos tal que $\mathcal{Y} = \{y_1, \dots, y_k, \dots, y_K\}$ representa las posibles clases. El objetivo es modelar la distribución de probabilidad a posteriori $P(Y|X)$, tal que $X \in \mathcal{X}$ e $Y \in \mathcal{Y}$. Por lo tanto, dada cualquier observación nunca vista en $\mathcal{X}$, es posible predecir su clase utilizando el máximo a posteriori:

$$\hat{Y} = argmax_{Y \in \mathcal{Y}} P(Y|X) \tag{56}$$

Dado un conjunto de entrenamiento $\left\{\left(X^{(n)}, Y^{(n)}\right)\right\}_{n=1}^{N} \in \mathcal{X} \times \mathcal{Y}$, cada árbol de un RF, $\mathcal{F} = \{F_t\}_{t=1}^{T}$, permite construir una partición $\mathcal{P}_t$ sobre el espacio de características de entrada $\mathcal{X}$. Considerando el modelo de árbol descrito en la sección anterior, es necesario definir la función objetivo que permite dividir recursivamente el proceso de entrenamiento para reducir la incertidumbre de las clases mediante la probabilidad a posteriori de estas.

4.2.3.1 Funciones objetivo para clasificación

Se considera la partición $\mathcal{P}_t = \left\{C_t^{(z_t)}\right\}_{z_t=1}^{Z_t}$ construida a partir de un RF F_t. La probabilidad a posteriori de cada clase se puede aproximar en cada región de decisión $C_t^{(z_t)}$de la partición con la Ecuación (57):

$$P\left(y_k \middle| X \in C_t^{(z_t)}, \mathcal{P}_t\right) = \frac{\left|\left\{X^{(n)} \in C_t^{(z_t)}, Y^{(n)} = y_k\right\}\right|}{\left|\left\{X^{(n)} \in C_t^{(z_t)}\right\}\right|} \tag{57}$$

Donde $|\cdot|$ es la cardinalidad, el número de elementos que hay dentro del conjunto. En cada nodo N_l del árbol de decisión F_t, una función de división f_l permite dividir el subconjunto S_l del total de datos de entrenamiento que llega a este nodo. El objetivo del proceso de optimización del nodo es encontrar la mejor función de división según una función objetivo predefinida. En tareas de clasificación, se han propuesto varias funciones objetivo que pretenden reducir la incertidumbre de clase. A continuación, se definen las más populares que son Ganancia de Información (*Information Gain*) y una variante basada en la Impureza de Gini (*Gini Impurity*) (Breiman, 2017; Quinlan, 1986).

La ganancia de información mide la diferencia entre la incertidumbre de clase antes y después del proceso de división. Una medida común de la incertidumbre es la llamada entropía de Shannon que se define para variables aleatorias discretas de la siguiente manera:

$$H(S_l) = -\sum_{k=1}^{K} P(y_k|S_l)\log_2\left(P(y_k|S_l)\right) \tag{58}$$

Después de dividir un subconjunto S_l en dos subconjuntos S_l^{izq} y S_l^{der}, que envían los datos a la izquierda y derecha, respectivamente, de los nodos hijos, la reducción de la incertidumbre se puede mejor empleando la ganancia de información Δ de la forma:

$$\Delta = H(S_l) - w_{izq}H\left(S_l^{izq}\right) - w_{der}H\left(S_l^{der}\right) \tag{59}$$

Donde $w_{izq} = |S_l|/\left|S_l^{izq}\right|$ y $w_{der} = |S_l|/\left|S_l^{der}\right|$.

Una variante de la entropía de Shannon que puede ser utilizada es la impureza de Gini, que se define como:

$$G(S_l) = -\sum_{k=1}^{K} P(y_k|S_l)(1 - P(y_k|S_l)) \tag{60}$$

Como se muestra en la Figura 23, tanto la entropía de Shannon como la impureza de Gini tienen un comportamiento similar y alcanzan su máximo cuando la clase posterior es uniforme, es decir, cuando $P(y_k|S_l) = {}^{1}/_{K}$. Por lo tanto, cabe esperar resultados similares utilizando una de estas dos funciones.

Sin embargo, la impureza de Gini presenta un menor coste computacional en comparación con la entropía de Shannon, debido a la ausencia de operaciones logarítmicas en su cálculo. Además, diversos estudios empíricos han demostrado que, en la mayoría de los casos, la elección entre Gini y entropía no produce diferencias

significativas en el rendimiento del modelo (Tangirala, 2020). Por esta razón, y dado que la impureza de Gini es la opción predeterminada en muchas bibliotecas de aprendizaje automático ampliamente utilizadas, como Scikit-learn o XGBoost, se ha consolidado como la elección más popular en la práctica.

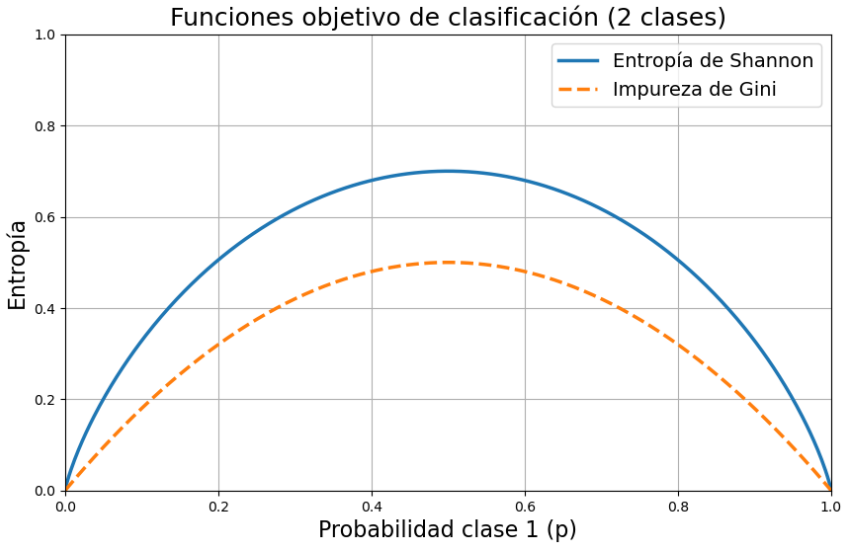

Figura 23. Comportamiento de la función de entropía de Shannon e impureza de Gini para un problema de clasificación de 2 clases.

4.2.3.2 Predicción empleando *random forest*

Una vez finalizada la fase de entrenamiento, se pueden realizar predicciones para nuevas observaciones que nunca han sido vistas por el modelo. Estos nuevos datos se envían a través de todos los árboles del clasificador y se combinan la probabilidad a posteriori de estos. Un método habitual para calcular la predicción Y del RF para una observación X es calcular la media de las probabilidades.

$$P(Y|X) = \frac{1}{T}\sum_{t=1}^{T} P(Y|X, P_t) \tag{61}$$

Donde $P_t$es la partición obtenida del árbol F_t. Usando el máximo a posteriori se obtiene:

$$\hat{Y} = argmax_{Y \in \mathcal{Y}} P(Y|X) \tag{62}$$

En las tareas de clasificación, este enfoque puede verse como el resultado de combinar la clasificación de los datos de entrada de cada uno de los árboles. Alternativamente, se puede emplear otro procedimiento para combinar los resultados entre sí. En vez de emplear las probabilidades a posteriori de los árboles, primero se realiza el máximo a posteriori en cada árbol:

$$\hat{Y}_t = argmax_{Y \in y} P(Y|X, P_t) \tag{63}$$

De esta manera cada uno de los árboles ofrece un resultado de clasificación de los datos de entrada. Finalmente, el resultado se calcula contando la clase que tiene más votos $\{\hat{Y}_t\}_t^T$ entre todos los árboles.

$$\hat{Y} = argmax_{Y \in y} \left(\sum_{t=1}^{T} [\hat{Y}_t == Y] \right) \tag{64}$$

donde $[\hat{Y}_t == Y]$ es una función booleana que devuelve un 1 si la igualdad es verdadera y un 0 en caso contrario. La desventaja de utilizar el voto mayoritario es que se pierde la naturaleza probabilística de los resultados del RF, esto es, la confianza en la clase para la decisión de etiquetado. Además, todos los votos tienen la misma influencia en la predicción final, incluso si los votos individuales proceden de las hojas del árbol que tengan una baja confianza. Mientras que un RF logra una gran generalización combinando las salidas de un conjunto de árboles aleatorios, los árboles individuales muestran rendimientos muy diferentes dependiendo de la parte del espacio de características en la que se encuentre la nueva observación X (Tripoliti et al., 2013).

4.2.3.3 Clasificación con clases no balanceadas

En las aplicaciones con datos reales, los datos a menudo sufren de un desequilibro en el número de datos de entrenamiento asociados a cada clase. Si el número de puntos de entrenamiento de cada clase es muy diferente, el cálculo de la probabilidad posterior mediante la Ecuación (57) se inclina hacia la clase más representada. Para evitar este tipo de sesgo durante el entrenamiento, existen dos soluciones posibles: (1) utilizar un *bootstrap* equilibrado del conjunto de entrenamiento, o (2) utilizar una normalización de clase al calcular los resultados finales.

En la primera solución, un *bootstrap* equilibrado del conjunto de entrenamiento $\{(X^{(n)}, Y^{(n)})\}_{n=1}^{N}$ se genera tomando M observaciones de cada clase de la forma:

$$M < inf_{Y \in y} |\{X^{(n)}, Y^{(n)} = Y\}| \tag{65}$$

Donde inf es la función ínfimo que, dado un conjunto de números reales, devuelve su mayor cota inferior y coincide con el mínimo del conjunto cuando éste existe. En la práctica, es el número de instancias de la clase menos representada en el conjunto de datos.

A continuación, se entrena cada árbol individual utilizando dicho *bootstrap* y se calculan las probabilidades a posteriori empleando la Ecuación (57). Aunque esta solución parece muy sencilla, no es aplicable en casos en los que se pretende detectar anomalías que

constituyen una clase muy pequeña en comparación con la clase mayoritaria. De hecho, el clasificador no puede aprender la variabilidad de la clase mayoritaria cuando sólo se entrena con un subconjunto muy pequeño de datos.

Para superar este problema, se pueden calcular de antemano la probabilidad de cada clase a partir de conjunto de datos de entrenamiento completo utilizando:

$$P(Y) = \frac{\left|\left\{X^{(n)}, Y^{(n)} = Y\right\}\right|}{N} \tag{66}$$

Por lo tanto, la probabilidad posterior se puede calcular en cada hoja $C^{(z)}$ integrando estas probabilidades:

$$P\left(Y|X \in C_t^{(z_t)}, \mathcal{P}_t\right) = \frac{1}{Q}\frac{1}{P(Y)}\frac{\left|\left\{X^{(n)} \in C_t^{(z_t)}, Y^{(n)} = Y\right\}\right|}{\left|\left\{X^{(n)} \in C_t^{(z_t)}\right\}\right|} \tag{67}$$

donde Q es una constante de normalización. Esta solución permite eliminar de forma fiable el sesgo introducido por las clases desequilibradas.

5. Redes Neuronales Artificiales

Una Red Neuronal Artificial (ANN) es un modelo computacional que se inspira en la estructura y los aspectos básicos de las redes neuronales biológicas. En la actualidad se trata de una técnica totalmente aceptada y ampliamente utilizada para afrontar problemas complejos. Los estudios comparativos muestran que la precisión de los métodos de ANN es superior a la de los métodos estadísticos tradicionales a la hora de abordar problemas complejos, especialmente en lo que respecta a patrones no lineales (Bahrammirzaee, 2010).

En las siguientes secciones se presenta una visión general de las ANN, se describen los fundamentos estadísticos que permiten su funcionamiento, las fases de diseño del modelo, y un conjunto de problemas y técnicas de optimización propias de estos sistemas. Finalmente, se introduce la Red Neuronal Convolucional (CNN), que permite procesar la información presente en imágenes.

5.1 Neurona artificial, perceptrón

La neurona artificial es un modelo matemático simple que se inspira en las neuronas biológicas presentes en el cerebro de todos los animales. La neurona biológica es un tipo de célula altamente especializada que presenta una alta interconectividad ($\sim 10^4$ sinapsis o conexiones por neurona), y a través de ellas pasa una señal electromagnética. Esta señal es información que recorre el cerebro, siendo procesada por las neuronas, y deriva en las diversas formas de acciones o reacciones (Zurada, 1992). Morfológicamente consta de tres regiones principales que se observan en la Figura 24: un cuerpo celular (o soma), que contiene el núcleo; un número variable de dendritas, que emanan del soma y se ramifican; y un axón único, que se extiende lejos del soma y tiene numerosas terminales para interconectarse con las dendritas de otras neuronas (Fratkin, 1997).

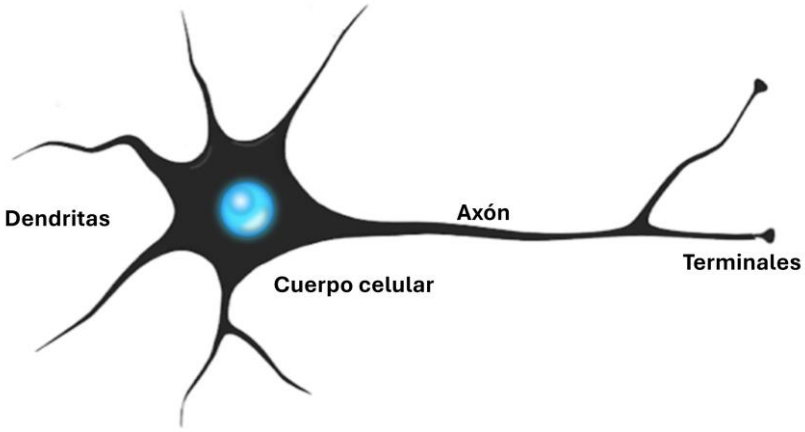

Figura 24. Estructura de la neurona biológica con sus principales componentes.

Las neuronas artificiales producen un valor de salida en función de una serie de valores de entrada. La neurona calcula una suma ponderada de las variables de entrada y aplica una función de activación para obtener el valor de la señal de salida (Russell & Norvig, 2016). El perceptrón, propuesto en 1958 por Frank Rosenblatt (Rosenblatt, 1958) es una de las arquitecturas de neuronas artificiales más simples, compuesta por una función de activación de umbral binario (Figura 25). Esta neurona matemática calcula una suma ponderada de sus señales de entrada y genera una salida de "1" si esta suma está por encima de un cierto umbral, de lo contrario, la función devuelve "0" como resultado.

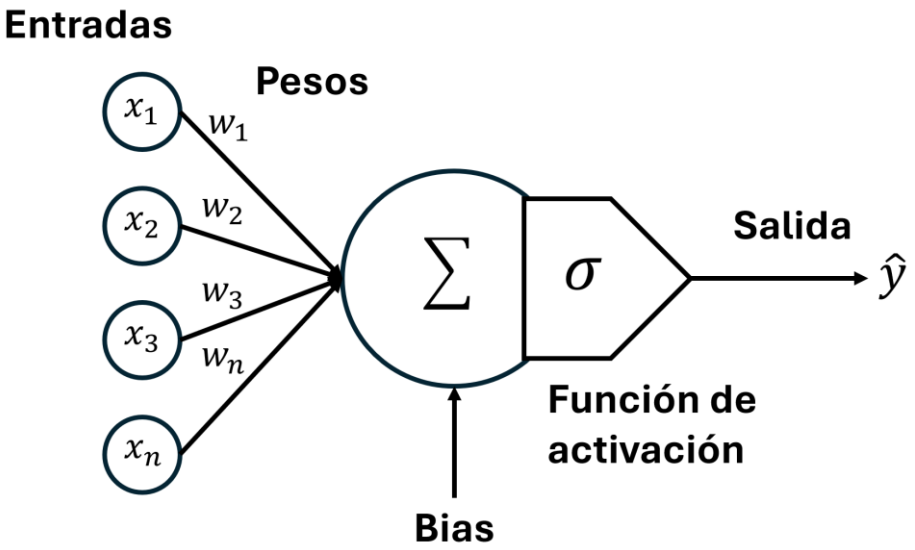

Figura 25. Arquitectura del perceptrón.

El perceptrón realiza una combinación lineal de sus entradas ponderadas y aplica una función de activación para obtener la señal de salida. Las funciones no lineales son las más utilizadas para dotar al perceptrón de un comportamiento no lineal. El modelo matemático del perceptrón se presenta en la Ecuación (68):

$$\hat{y} = \sigma\left(\sum_{i=1}^{n} x_i w_i + bias\right) \tag{68}$$

Donde $\hat{y}$ es la salida de la neurona; σ es la función de activación; x es el vector de entrada de n elementos; w es el vector de pesos y $bias$ es un valor que modifica la función de activación.

5.2 Red Neuronal Artificial

El perceptrón por sí solo no tiene mucha capacidad de procesamiento, para ello se desarrolla la ANN, que se compone de varias capas de perceptrones. Una ANN se define como un conjunto de neuronas que están interconectadas entre sí. Uno de los modelos más utilizados es el llamado perceptrón multicapa (*multilayer perceptron*, MLP) (Haykin, 1998). Un MLP está compuesto por una capa de entrada, uno o más capas denominadas

ocultas, y una capa final de salida. Cada capa emplea varias neuronas, y cada neurona de una capa está conectada a las neuronas de la capa adyacente con diferentes pesos, que se utilizan para determinar cuánto afectará una unidad a la otra (Chen et al., 2005), como se muestra en la Figura 26.

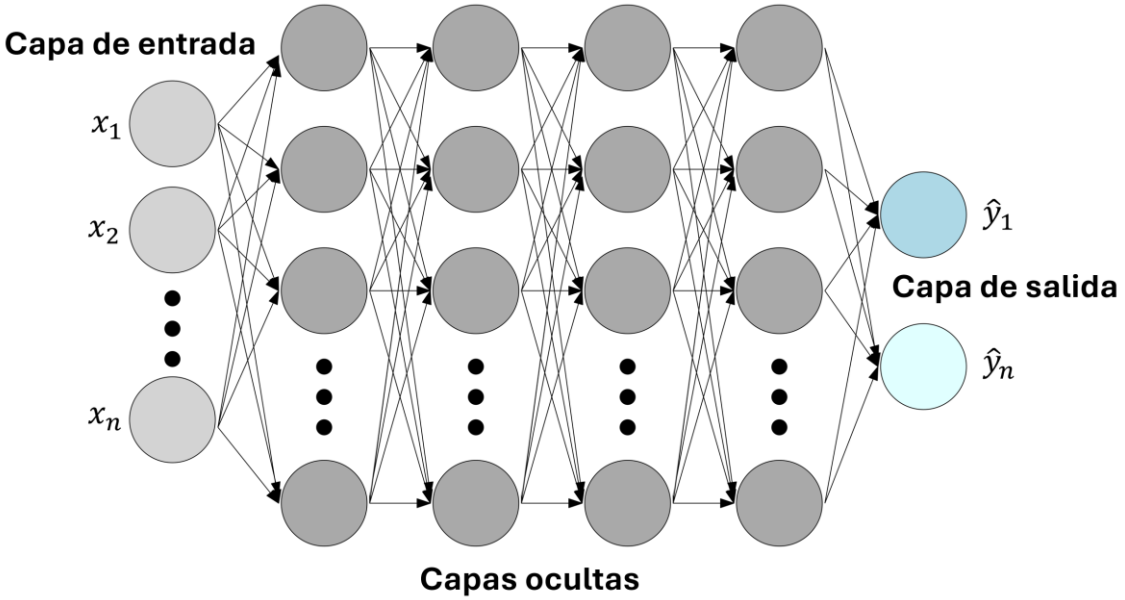

Figura 26. Arquitectura de una red neuronal artificial.

La ANN actúa como un tipo de matriz que puede realizar un mapeo no linear entre las señales de entrada y de salida (Rui & El-Keib, 1995). De este modo, las neuronas de entrada reciben los datos y se calculan los productos de las entradas y los pesos; las señales entran en la capa oculta, donde se realiza una suma, se añade un sesgo y se aplica una función de activación en cada neurona. A continuación, las señales que salen de las neuronas de la capa oculta se multiplican de nuevo por los pesos y entran en las neuronas de la capa de salida, donde se realiza una suma más la adición del sesgo para generar la salida de la ANN.

Sin embargo, existe un gran número de casos en los que este tipo de neuronas no eran capaces de resolver el problema, debido entre otros factores, a la falta de capacidad computacional de la época. La investigación en ANN en la comunidad informática quedó parada hasta la introducción de los modelos ANN de Hopfield durante la década de 1980 empleando el algoritmo de entrenamiento denominado *backpropagation*. Hoy en día una red MLP puede resolver diferentes problemas no lineales con una estructura de red bastante simple (Singh, 2016). El objetivo de un MLP es obtener la salida deseada para las entradas de un tipo determinado. El proceso de entrenamiento del MLP consiste en alimentar al sistema con las entradas y salidas deseadas, y se determina el error en la respuesta. Los parámetros del sistema, es decir, los pesos de las interconexiones, se modifican durante esta fase para disminuir la diferencia entre la salida deseada y la real (Negrov et al., 2017).

Las ANN más populares es la red multicapa en la que las neuronas se organizan en una serie de capas, y la señal de información fluye a través de la red únicamente en una

dirección (*feed-forward*), desde la capa de entrada a la de salida (Tkáč & Verner, 2016). Mientras que MLP está más enfocado en problemas no linealmente separables, clasificación, y aproximación de funciones continuas, existe otra arquitectura de redes neuronales basada en el número de capas ocultas. Cuando la ANN consta de varias capas de perceptrones, se suele emplear el término Red Neuronal Profunda (*Deep Neural Network,* DNN).

Las ANN, especialmente cuando se emplean modelos complejos con varias neuronas y capas ocultas, presentan un gran coste en cuanto a tiempo de entrenamiento (Moraes et al., 2013). Aunque en la actualidad las unidades centrales de procesamiento (CPU) presentan estructuras multinúcleos e instrucciones para el procesamiento paralelo de alto rendimiento (Bergstra et al., 2010; Botti & Giuffra, 2010), estas no alcanzan el rendimiento y la potencia de cálculo que ofrecen las unidades de procesamiento gráfico (GPU). Esta diferencia se debe, principalmente, al hecho de que las GPU presentan una mayor capacidad de cálculo en paralelo con un sistema multinúcleo (Oh & Jung, 2004; Schmidhuber, 2015), llegando a obtener implementaciones de ANN en GPU que puedes ser 50 o 60 veces más rápidas que las implementaciones en CPU (Ciresan et al., 2011).

En cuanto a los *frameworks* de programación orientados al diseño y desarrollo de las ANN, existen librerías de código abierto como Caffe, Theano, Torch y TensorFlow (Abadi et al., 2016; Rampasek & Goldenberg, 2016). El hecho de ser libres y de código abierto permite una gran aceptación dentro del ámbito académico y de investigación.

5.3 Funciones de activación

Las funciones de activación son un componente esencial de las ANN. En un perceptrón, las entradas ponderadas se suman y pasan a través de una función que escala la salida a un rango fijo de valores. Esta señal es la que alimenta a las neuronas conectadas en la siguiente capa. La función de activación suele ser no lineal debido a que se consigue una mayor precisión y mejora las capacidades de aprendizaje y generalización de las ANN en contraste con las funciones primitivas lineales (Zamanlooy & Mirhassani, 2013). A continuación se describen algunas de las principales funciones de activación.

5.3.1 Binary

La función binaria (*binary*) escalonada fue la primera función de activación aplicada a la neurona artificial. La función presenta valores de salida binarios: “0” cuando la señal de entrada es menor que cero, y “1” cuando el valor de la señal es igual o mayor que cero. La fórmula se muestra en la Ecuación (69):

$$binary(x) = \begin{cases} 0 \; si \; x < 0 \\ 1 \; si \; x \geq 0 \end{cases} \tag{69}$$

El gráfico de la función se muestra en la Figura 27. Independientemente de lo negativo o positivo que sea el valor de entrada, la lógica de esta función siempre devolverá "0" o "1". Hoy en día esta función está en desuso, ya que no existe una relación lineal entre los patrones de entrada y salida en la mayoría de los problemas modernos de interés (Zhang & Suganthan, 2016).

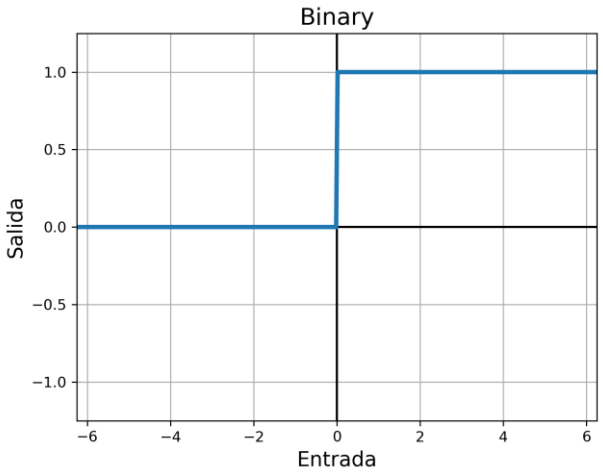

Figura 27. Función de activación binaria.

5.3.2 Sigmoid

La función de activación *sigmoid* se presenta como una sustituta de la función binaria. Esta función presenta la ventaja de que tiene una derivada no nula bien definida en todo su dominio, lo que permite que el algoritmo de descenso por el gradiente obtenga mejoras en cada paso (Menon et al., 1996). La función de *sigmoid* produce números positivos sólo entre "0" y "1", y la fórmula se presenta en Ecuación (70). Esta función también se conoce como función en forma de S, y sus valores de salida acotados la convierten en una de las más útiles para entrenar ANN (Sibi et al., 2013). El gráfico de la función se muestra en la Figura 28.

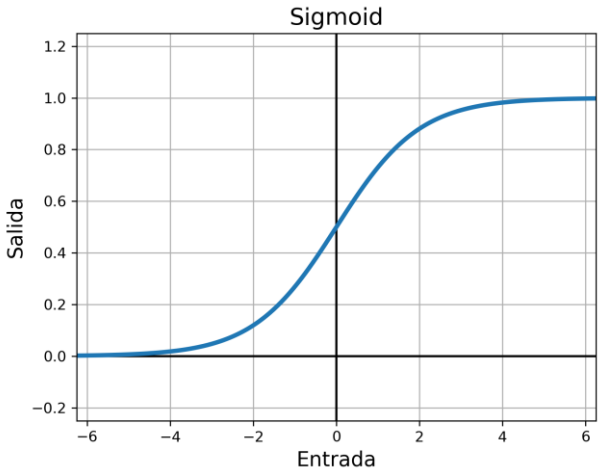

Figura 28. Función de activación sigmoid.

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{70}$$

5.3.3 Tanh

La función tangente hiperbólica (*tanh*) es similar a una función sigmoid, pero sus límites de salida están entre "-1" y "1" (Karlik & Olgac, 2011). La formulación se presenta en la Ecuación (71).

$$sigmoid(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{71}$$

La función da valores de salida comprendidos entre $[-1, 1]$ y devuelve "0" cuando la entrada es cero. Cuanto más positiva o negativa es la entrada, la salida crea asíntotas aproximadas a los valores máximos. Este rango tiende a que la salida de cada capa esté más o menos centrada en "0" al principio de cada capa durante el entrenamiento, lo que a menudo ayuda a acelerar la convergencia (Géron, 2022). El gráfico de la función se muestra en la Figura 29.

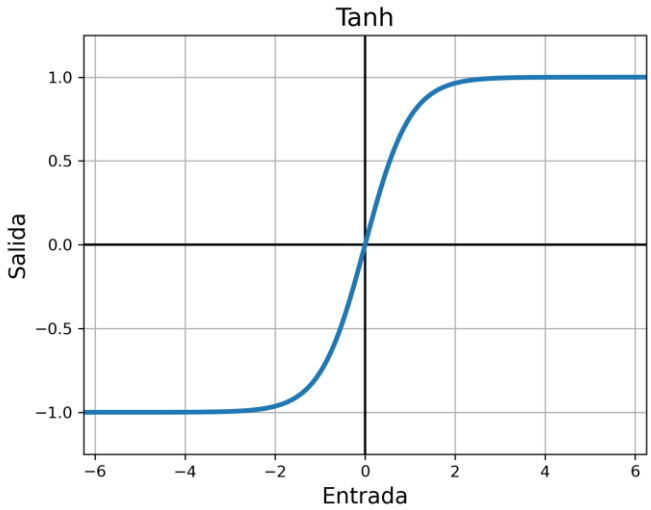

Figura 29. Función de activación tanh.

5.3.4 ReLU

La función *Rectified Linear Unit* (ReLU) se ha convertido en el método por defecto en la práctica, ya que ofrece muy buenos resultados y tiene la ventaja de ser rápida de calcular. Se trata de una función continua pero no diferenciable en cero (la pendiente cambia bruscamente, lo que puede hacer que el descenso por gradiente rebote), y su derivada es 0 para z < 0. Es importante destacar que el hecho de que no tenga un valor máximo de salida ayuda a reducir algunos problemas durante el descenso por gradiente. La fórmula de la función de activación ReLU se presenta en la Ecuación (72):

$$relu(x) = \begin{cases} 0 \; si \; x < 0 \\ x \; si \; x \geq 0 \end{cases} \tag{72}$$

Las funciones ReLU han demostraron ser eficientes en comparación con otras alternativas para un amplio conjunto de tareas de clasificación, incluyen el

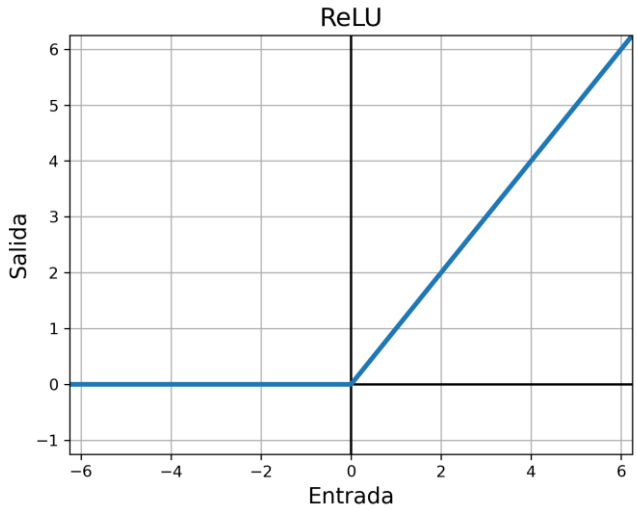

Figura 30. Función de activación ReLU.

reconocimiento automático del habla (Maas et al., 2013; Tóth, 2013; Zhang & Woodland, 2016). El gráfico de la función ReLU se muestra en la Figura 30.

5.3.5 Softmax

La función softmax se emplea ampliamente en las ANN, especialmente en arquitecturas DNN. El softmax devuelve un número que puede ser interpretado como la probabilidad de una clase en particular (Goodfellow, 2016; Zheng et al., 2015) siendo especialmente útil en tareas de clasificación. La fórmula del softmax se muestra en la Ecuación (73):

$$softmax(x) = \frac{e^{x_i}}{\sum_{i=1}^{J} e^{x_j}}, para\ i = 1, \dots, J \tag{73}$$

Con esta función se asegura que cada salida $\hat{y}_i$ para un conjunto de observaciones x se ajusta a las condiciones $0 \leq \hat{y}_i \leq 1$, y la suma de las salidas es $\sum_{i=1}^{J} \hat{y}_i = 1$, como requisito para ser una distribución de probabilidad (Bishop, 2020). Las funciones de activación Softmax en la última capa de una ANN proporcionan la probabilidad asociada a cada clase para los datos de entrada. En el caso de dos clases (2-D), la softmax se reduce a la función sigmoide.

5.4 Fases del diseño de un modelo ANN

La correcta implementación de una ANN es un proceso que requiere una metodología adecuada que garantice el máximo rendimiento de los algoritmos. Este proceso puede dividirse principalmente en tres fases: procesamiento de datos de entrada; fase de entrenamiento; y fase de inferencia o test. En las siguientes subsecciones se detallan cada una de estas fases.

5.4.1 Procesamiento de datos de entrada

Previo a la fase de entrenamiento de una ANN es necesario procesar el conjunto de datos. Es necesario dividir los datos de entrada disponibles en tres conjuntos diferentes: entrenamiento, validación y test, de manera que cada uno de los conjuntos represente la distribución estadística de los datos del problema. El conjunto de entrenamiento y validación se emplean durante el proceso de entrenamiento de la red; mientras que el conjunto de test es un grupo de datos nunca visto por el modelo, que se emplea en el proceso de inferencia o clasificación.

En cuanto a las variables de entrada, en problemas reales es común que presenten diferentes escalas o que las desviaciones estándar sean diferentes entre sí. Esto puede dar lugar a que una variable domine los cálculos debido a su dimensionalidad, de ahí que los procesos de normalización de datos de entrada sean fundamentales para corregir estos sesgos (Martinez et al., 2017). A su vez, la normalización de las variables aumenta la eficiencia computacional del proceso de entrenamiento de la ANN, junto con una mejora en la precisión de los resultados (Nawi et al., 2013; Specht, 1991) y ayuda al sistema a realizar un mejor ajuste de peso para cada elemento (Jayalakshmi & Santhakumaran, 2011).

En las redes neuronales, existen fundamentalmente dos métodos de preprocesamiento. El primero se denomina normalización *Z-score*, y consiste en sustraer el valor de la media de los datos y dividir por la desviación estándar. De esta forma se consigue una distribución de datos con media cero y varianza 1 (Priddy, 2005). Este proceso se realiza con la Ecuación (74) , donde x_i es un valor de entrada para un descriptor en concreto; μ es la media de ese descriptor; σ es la desviación estándar; y z_i es el valor normalizado.

$$z_i = \frac{x_i - \mu}{\sigma} \tag{74}$$

El segundo método se denomina normalización *Min-Max*, y consiste en reescalar los datos de entrada para acotarlos entre valores $[0,1]$ (Eberhart, 2014). En la Ecuación (75) se muestra x_i como el valor de entrada; x_{min} es el valor mínimo observado; x_{max} es el valor máximo observado; y x_i' es el valor reescalado que se emplea como entrada normalizada a la red.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{75}$$

La normalización de los datos de entrada previo al proceso de entrenamiento es fundamental para acelerar los cálculos computacionales y obtener mejores resultados (Sola & Sevilla, 1997). Un modelo ANN sin normalización de datos de entrada provoca que el ajuste del error esté dominado por los valores de entrada más grandes. La normalización permite al sistema una manera de asignar los pesos correctos a la entrada, basándose en el mérito real de los datos de entrada y no sólo en su tamaño absoluto o unidad de medida.

5.4.2 Fase de entramiento

Durante la fase de entrenamiento se llevan a cabo dos procesos complementarios entre sí, entrenamiento y validación, empleando los conjuntos de datos separados en la fase anterior. Estos procesos se pueden ver como un problema de optimización multivariable que puede implicar cientos o miles de variables. Durante esta fase, los conjuntos de datos de entrenamiento y validación se emplean para modificar el modelo hasta que alcanza la previsión de salida deseada. El conjunto de entrenamiento se divide en lotes, *batch*, de manera que la red no procesa todos los datos de entrenamiento a la vez. Por ejemplo, se dividen los datos de entrada en *batch* de 32 datos y se procesan de forma conjunta por la red.

El conjunto de entrenamiento es el conjunto de datos más grande y es el empleado por la red neuronal para aprender los patrones estadísticos presentes en los datos. Generalmente, se emplea una división 70/30 entre el conjunto de datos de entrenamiento y validación. Este último se utiliza para supervisar el proceso de entrenamiento y evitar el sobreajuste del modelo. Gracias a este conjunto de datos la fase de entrenamiento de una red se puede detener en el momento en que el error en el conjunto de validación empieza a aumentar, o no disminuye, mientras que el error de entrenamiento sigue disminuyendo, lo que indica un sobreajuste. Este proceso de parada se denomina *early stopping* (Morgan & Bourlard, 1989), y permite reducir el número de iteraciones necesarias para entrenar un modelo. De este proceso de entrenamiento y validación surge la mejor configuración de red.

Durante la fase de entrenamiento se utilizan diferentes algoritmos de aprendizaje para definir la capacidad de abstracción de la red. En concreto, el algoritmo más empleado es el descenso de gradiente por retropropagación, *backpropagation gradient descent*, que es un método simple, universal y con una gran disponibilidad en las bibliotecas de software (Phansalkar & Sastry, 1994; Tkáč & Verner, 2016). Existen numerosos algoritmos de aprendizaje como pueden ser algoritmos genéticos y los algoritmos de enjambre de partículas, *particle swarm*; aunque estos algoritmos no están tan ampliamente implementados por las bibliotecas de aprendizaje automático (Örkcü & Bal, 2011; Rios & Sahinidis, 2013).

El propósito del algoritmo de aprendizaje con *backpropagation* es cambiar iterativamente algunos parámetros de las neuronas siguiendo una dirección que minimice la función de error, la cual mide la diferencia entre la salida deseada y el resultado real de la ANN cuando se empleando los conjuntos de datos de entrenamiento

y validación. El proceso paso a paso del algoritmo *backpropagation* se define a continuación.

- Se toma un *batch* cada vez, hasta que se ha recorrido el conjunto de entrenamiento al completo varias veces. Cada una de estas iteraciones se denomina *epoch*.
- Cada uno de los *batch* se emplea como datos de entrada de la red. El algoritmo calcula la salida de todas las neuronas en las capas ocultas. Este es el paso hacia delante, *forward pass*, que coincide con el proceso de predicción de la red, salvo que todos los resultados intermedios se guardan, ya que son necesarios para el paso hacia atrás, *backward pass*.
- A continuación, el algoritmo mide el error de salida de la red utilizando una función de pérdida que compara la salida deseada y la salida real de la red, y devuelve una medida del error.
- Entonces calcula cuánto contribuyó al error cada sesgo de salida de la función de activación de la neurona y cada conexión a la siguiente capa. Esto se hace analíticamente aplicando la regla de la cadena, lo que hace este paso rápido y preciso.
- El algoritmo mide cada contribución del error y localiza la conexión de la que proviene, utilizando de nuevo la regla de la cadena hacia atrás. Este proceso inverso mide eficientemente el gradiente de error y sesgo de la red, propagando el gradiente de error hacia atrás.
- Por último, el algoritmo ajusta todos los pesos de conexión de la red, utilizando los gradientes de error que acaba de calcular. El valor que se suma o resta en los pesos y sesgos se denomina tasa de aprendizaje, *learning rate*.

En resumen, el algoritmo *backpropagation* hace predicciones para un *batch* (paso hacia delante), mide el error, luego recorre cada capa en sentido inverso para medir la contribución al error de cada parámetro (paso inverso) y, por último, ajusta los pesos de conexión y los sesgos para reducir el error (paso de descenso de gradiente).

Normalmente, los valores iniciales de los pesos se eligen de forma aleatoria en valores cercanos a cero (entre $[-1,1]$ suelen ser valores por defecto). Por lo tanto, el modelo comienza siendo casi lineal y se vuelve no lineal a medida que cambian los pesos (Cao et al., 2015; Friedman, 2009). Es importante inicializar todos los pesos de las conexiones de las capas ocultas de forma aleatoria, de lo contrario el entrenamiento fallará. Por ejemplo, si inicializa todos los pesos y sesgos a cero, entonces todas las neuronas de una capa serán perfectamente idénticas y, por tanto, la retropropagación las afectará exactamente de la misma manera, por lo que permanecerán idénticas. En otras palabras, a pesar de tener cientos de neuronas por capa, el modelo actuará como si sólo tuviera una neurona por capa. Iniciado aleatoriamente los pesos, rompes la simetría y permites que la retropropagación entrene a un equipo diverso de neuronas (Géron, 2022).

El procedimiento de *backpropagation* calcula el gradiente de la función de error, en relación con los pesos y el sesgo de un modelo multicapa de neuronas. La retropropagación no es más que una aplicación práctica de la regla de la cadena para

derivadas, y puede definirse mediante la Ecuación (76), donde E es la función de error; w_{ij} es el peso de una conexión a una neurona; $\hat{y}_{ij}$ es la salida de una neurona; y z_{ij} es el producto del peso y la variable de entrada a la neurona.

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial \hat{y}_{ij}} \times \frac{\partial \hat{y}_{ij}}{\partial z_{ij}} \times \frac{\partial z_{ij}}{\partial w_{ij}} \tag{76}$$

La regla de la cadena se aplica cuando existe una relación funcional de dependencia, esto es, cuando una función depende de otra (Rosenbaum & Johnson, 1984). En una red neuronal la función de error E depende de la salida de una neurona $\hat{y}$, que a su vez depende del producto de las variables de entrada z, haciendo que el parámetro final de esta cadena sean los pesos y el sesgo w (Rusk, 2016). Este proceso de ajustar los pesos para disminuir el error a la salida se trata básicamente de un problema de optimización. Si el conjunto de entrenamiento es una buena representación de la variabilidad estadística del problema en cuestión, y el algoritmo de entrenamiento se aplica de forma eficaz, se genera un modelo que permite una buena generalización de los datos. Se habla de generalización cuando la red es capaz de estimar correctamente la salida para uno conjunto de datos de entrada nunca vistos, ni en el conjunto de datos de entrenamiento ni en el de validación (Bishop, 1995). Esta fase se define como un proceso de entrenamiento supervisado, dado que el valor objetivo para cada patrón de entrada siempre se conoce a priori.

Los pesos generados aleatoriamente y el sesgo se ajustan mediante el valor de la tasa de aprendizaje. La eficacia y convergencia del algoritmo de entrenamiento dependen significativamente del valor de esta tasa, el valor óptimo depende de la configuración de la red y varía enormemente en función del problema en cuestión. Si la función de error

Figura 31. Ejemplo de una función de error y el peso de gradiente durante el algoritmo de backpropagation.

presenta un mínimo amplio, un valor grande de la tasa de aprendizaje da lugar a una convergencia más rápida; mientras que un sistema con mínimos estrechos ofrece mejor resultados cuando se emplea una tasa más pequeña. Por ello es necesario emplear adecuadamente este algoritmo, para evitar converger a mínimo locales, o no converger (Bi et al., 2005), como se muestra en la Figura 31. No existen reglas generales para obtener un valor óptimo de la tasa de aprendizaje, aunque sí existen procedimiento para acercarse al valor óptimo. Estas técnicas de optimización se describen en la sección 5.6.

Otra estrategia para garantizar el rendimiento de los algoritmos de entrenamiento es definir un criterio de parada, *early stopping*. El criterio más empleado consiste en supervisar el error en los conjuntos de entrenamiento y validación; si el error de validación empieza a aumentar en comparación con el error de entrenamiento, la detención temprana del entrenamiento evita un sobreajuste del modelo (Iyer & Rhinehart, 2000; Tetko et al., 1995; Zhang & Morris, 1998).

En cuanto al tamaño de los conjuntos de datos, cabe destacar que es necesario un gran número de muestras de entrada para evitar el sobreajuste y obtener una buena generalización (Schmidhuber, 2015). No existe un número fijo recomendado de muestras de entrenamiento, pero existe una relación directa entre grandes conjuntos de datos y resultados adecuados (Kourou et al., 2015; Krizhevsky et al., 2012).

Para concluir, hay que destacar que realizar un entrenamiento adecuado es clave para la ANN, por ello evitar el sobreajuste de la red es crucial. El sobreajuste afecta a la capacidad de generalización de la red y, por lo tanto, a la precisión de la predicción. Entre otros, el número de muestras disponibles en el conjunto de datos, la detención temprana durante el proceso de entrenamiento, la tolerancia al error en el conjunto de entrenamiento, y el control del error de validación son herramientas que se deben considerar de forma conjunta para superar el problema del sobreajuste.

5.4.3 Fase de inferencia

Una vez entrenada la red neuronal, su rendimiento se evalúa sobre el conjunto de datos de test. Este conjunto de datos tiene que estar compuesto por muestras que no hayan sido incluidas durante el proceso de entrenamiento, es decir, son datos totalmente nuevos para la red. Cuando existe un desajuste entre los valores reales y los valores estimados por la red es un indicativo de que el modelo debe volver a entrenarse, con el procedimiento de aprendizaje adecuado, o aumentando el conjunto de datos disponibles. Para medir la precisión del modelo se emplea diversas funciones de coste.

En las ANN, las funciones de coste más utilizadas son: el error cuadrático medio (*mean square error*, MSE); la raíz del error cuadrático medio (*root mean square error,* RMSE); el error absoluto medio (*mean absolute error*, MAE); error porcentual absoluto medio (*mean absolute percentage error*, MAPE); y la entropía cruzada (*cross-entropy*, CE). Estas funciones pueden utilizarse durante las fases de entrenamiento e inferencia, mientras que existen otras funciones como el coeficiente de correlación (*correlation coefficient*, R) y el coeficiente de determinación (*determination coefficient,* R^2), que son más

apropiadas para la fase de inferencia. A continuación se describe la formulación de cada una de las funciones:

- El MSE, Ecuación (77), es una de las funciones de coste más empleadas por su rendimiento en relación con el algoritmo de retropropagación (Adya & Collopy, 1998). Esta función es especialmente adecuada cuando se emplea una única variable, ya que utilizar múltiples variables dificulta la comparación de errores (Köksoy, 2006).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{77}$$

  Donde n es el número de muestras en el conjunto de datos; y_i es el dato real; $\hat{y}_i$ es la estimación de la red. Esta misma anotación se emplea en todas las funciones de coste de esta sección.

- El RMSE se calcula de la misma manera que el MSE y se hace la raíz cuadrada, como muestra la Ecuación (78). Por lo tanto, el RMSE mide el error global del modelo. La principal diferencia entre el MSE y el RMSE es que este último es más útil cuando los errores grandes son particularmente indeseables y los errores pequeños tienden a ser menos penalizados.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{78}$$

- El MAE, Ecuación (79), penaliza la infra- y la sobre- predicción con respecto al resultado real y es una medida más natural que la RMSE, ya que se trata de una medida inequívoca de la magnitud media del error, y es útil para las comparaciones entre el rendimiento del modelo con diferentes magnitudes (Chai & Draxler, 2014; Willmott & Matsuura, 2005).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{79}$$

- El MAPE, Ecuación (80), se calcula como una modificación de MAE, empleando una comparación término a término del error relativo en la predicción con respecto al valor real de la variable. En consecuencia, el MAPE es un estadístico no sesgado para medir la capacidad predictiva de los modelos. El MAPE se utiliza a menudo debido a su interpretación intuitiva en términos de error relativo.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100 \tag{80}$$

- En la función CE, Ecuación (81), se refleja la similitud entre variables desde la perspectiva de la probabilidad (Men et al., 2016). Esta función de coste se utiliza frecuentemente con ANN con función de activación softmax en la capa de salida (Dahl et al., 2013; Liew et al., 2016; Maas et al., 2013).

$$CE = -\sum_{i=1}^{n} y_i \times log\, \hat{y}_i \tag{81}$$

  CE presenta ventajas significativas sobre otras funciones de error para problemas de probabilidades, por ejemplo al tener una capa de salida softmax (Kline & Berardi, 2005). También CE presente mayor robustez cuando se enfrenta a problemas de datos limitados.

- La función R, como muestra la Ecuación (82), define en qué medida un conjunto de datos se ajusta a una línea recta (Tripepi et al., 2008).

$$R = \frac{n(\sum y_i\, \hat{y}_i) - (\sum y_i) - (\sum \hat{y}_i)}{\sqrt{[n\sum {y_i}^2 - (\sum y_i)^2]\left[n\sum {\hat{y}_i}^2 - (\sum \hat{y}_i)^2\right]}} \tag{82}$$

- La función R^2, Ecuación (83), se utiliza ampliamente en las ANN para evaluar el rendimiento de los modelos (Gholami & Fakhari, 2017), incluyendo enfoques en problemas de regresión, clasificación y predicciones.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{\sum_{i=1}^{n}[y_i - \hat{y}_{mean}]^2} \tag{83}$$

Cabe destacar que estas métricas se pueden emplear fuera de la fase de inferencia, por ejemplo, como métricas para evaluar el rendimiento de un clasificador. En el presente trabajo se emplean otras métricas diferentes, como pueden ser las matrices de confusión, que se definen en el capítulo 6.

La correcta selección de una función de coste en el proceso de evaluación del modelo es una tarea crucial en el desarrollo de redes neuronales. No existe una solución universal para obtener un mejor rendimiento, y la mayor parte del trabajo para crear modelos de buena calidad consiste en probar diferentes funciones de coste, y optimizar el resto de hiperparámetros, para medir el rendimiento de la red.

5.5 El problema del descenso del gradiente

Como se explica en la sección 5.4.2, la segunda fase del algoritmo de *backpropagation* consiste en calcular la contribución al error de salida de cada conexión y sesgo de la neurona, desde la capa de salida hacia la capa de entrada. Una vez el algoritmo ha calculado el gradiente de la función de coste con respecto a cada parámetro de la red, emplea estos valores para actualizar cada parámetro. Sin embargo, el valor del gradiente disminuye conforme se avanza hacia las primeras capas de la red, provocando que la actualización de los pesos de las capas inferiores sea prácticamente nula. Esto se denomina el problema del descenso del gradiente, *vanishing gradient problem* (Hochreiter, 1998). En algunos casos ocurre lo contrario, los gradientes aumentan hasta que las capas iniciales reciben grandes actualizaciones y el algoritmo diverge, aunque este es un problema que aparece con más frecuencia en las redes neuronales recurrentes. En general, las redes DNN presentan un problema de gradiente inestable, en el que unas capas aprenden a diferente velocidad que otras.

Los principales factores que producen este comportamiento fueron descritos por Xavier Glorot y Yoshua Bengio en 2010 (Glorot & Bengio, 2010), entre ellos, la combinación de la función de activación *sigmoid* y una técnica de inicialización de pesos con una distribución normal de media 0 y desviación estándar 1. Con esta configuración la varianza de las salidas de cada capa es mucho mayor que la varianza de sus entradas por lo que la varianza sigue aumentando hasta que la función de activación satura en las capas superiores. En realidad, esta saturación empeora por el hecho de que la función *sigmoid* tiene una media de 0,5, no de 0, y la función se satura conforme los valores de entrada se hacen más grandes, como se observa en la Figura 32. Así, cuando la retropropagación se pone en marcha, prácticamente no tiene gradiente que propagar a través de la red, y el poco gradiente que existe se va diluyendo a medida que la retropropagación desciende hacia las capas inferiores.

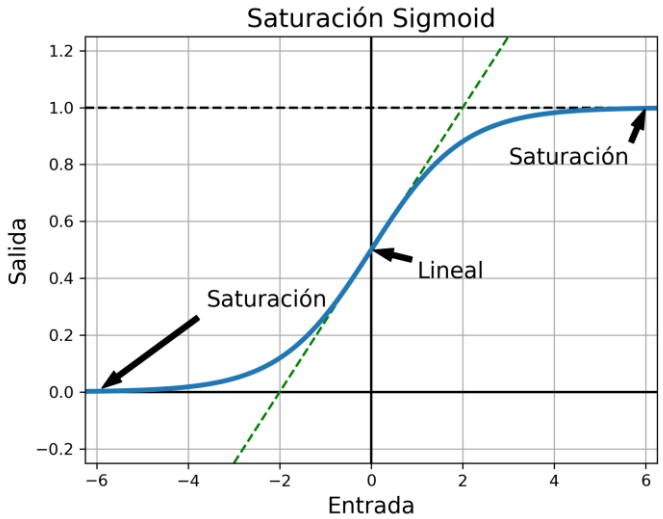

Figura 32. Saturación de la función de activación sigmoid.

En (Glorot & Bengio, 2010) se propone una técnica para evitar el problema de los gradientes inestables, basada fundamentalmente en el hecho de que la señal tiene que fluir correctamente en ambas direcciones, hacia delante al hacer predicciones; y hacia atrás durante la retropropagación. Para ello, es necesario que la varianza de las salidas de cada capa sea igual a la varianza de sus entradas. Para ello se propone inicializar de forma aleatoria los pesos de cada capa, cuando se emplea la función de activación *sigmoid*, empleando la Ecuación (84), donde $fan_{avg} = (fan_{in} + fan_{out})/2$; fan_{in} es el tamaño de la capa anterior; y fan_{out} el tamaño de la capa actual.

$$Distribución\ normal\ con\ media\ 0\ y\ varianza\ \sigma^2 = \frac{1}{fan_{avg}} \qquad (84)$$

Algunos trabajos (K. He et al., 2015; LeCun et al., 2002) han proporcionado estrategias similares para diferentes funciones de activación. En la Tabla 1 se puede observar estas estrategias, que difieren únicamente por la escala de la varianza y si utilizan fan_{avg} o fan_{in}. En la librería Keras se emplea la inicialización de Glorot por defecto.

Tabla 1. Estrategias de inicialización para las funciones de activación.

Inicialización	**Función de Activación**	$\boldsymbol{\sigma^2}$ **(Normal)**
Glorot	tanh, sigmoid, softmax	$1/fan_{avg}$
He	ReLU, Leaky ReLU, ELU, GELU	$2/fan_{in}$
LeCun	SELU	$1/fan_{in}$

A continuación se definen estas funciones de activación.

- Leaky ReLU

La función de activación Leaky ReLU (Xu, 2015) se define en la Ecuación (85), donde α es el valor de la pendiente negativa; su comportamiento se puede observar en la Figura 33. Esta función nace para subsanar uno de los problemas que presentan las funciones ReLU durante el entrenamiento. Una de las conclusiones del artículo de Glorot y Bengio es que los problemas con los gradientes inestables se deben en parte a una mala elección de la función de activación. Por ello, se optó por emplear la función de activación ReLU, que no satura en valores positivos y es muy rápida de calcular. Sin embargo la función ReLU sufre de un problema denominado *dying ReLUs*; se puede dar el caso que los pesos se ajustan de tal forma que la entrada de la función es negativa para todos los casos de entrenamiento, dando lugar a una neurona que siempre ofrece un valor 0 a la salida, por lo que nunca aprende. Empleando Leaky ReLU siempre se obtiene mejor rendimiento que empleando ReLU.

$$LeakyReLU(z) = \begin{cases} z, & si\ x \geq 0 \\ \alpha z, & si\ x < 0 \end{cases} \qquad (85)$$

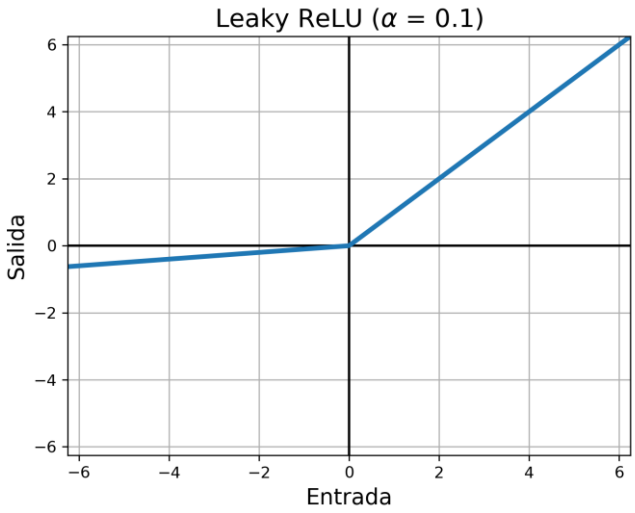

Figura 33. Función de activación Leaky ReLU, con una pendiente para los valores negativos.

- ELU y SELU

La función de activación ELU, *exponential linear unit (Clevert, 2015)*, presenta un menor tiempo de entrenamiento, reduce el problema del descenso del gradiente y previene el problema de *dying ReLU*. Viene definida por la Ecuación (86).

$$ELU(z) = \begin{cases} \alpha(e^z - 1), & si\ z < 0 \\ z, & si\ z \geq 0 \end{cases} \qquad (86)$$

La función SELU (Klambauer et al., 2017), *scaled ELU*, definida en la Ecuación (87), es una variante a escala dela función de activación ELU (utilizando $\alpha \approx 1{,}67$). Los autores demostraron que si se emplea una ANN, y si todas las capas ocultas utilizan la función de activación SELU, entonces la red se autonormaliza: la salida de cada capa tiende a conservar una media de 0 y una desviación estándar de 1 durante el entrenamiento, lo que resuelve el problema de descenso de gradiente. El comportamiento de estas funciones se puede observar en la Figura 34.

$$SELU(z) = 1.05 \cdot ELU(z), con\ \alpha = 1{,}67 \qquad (87)$$

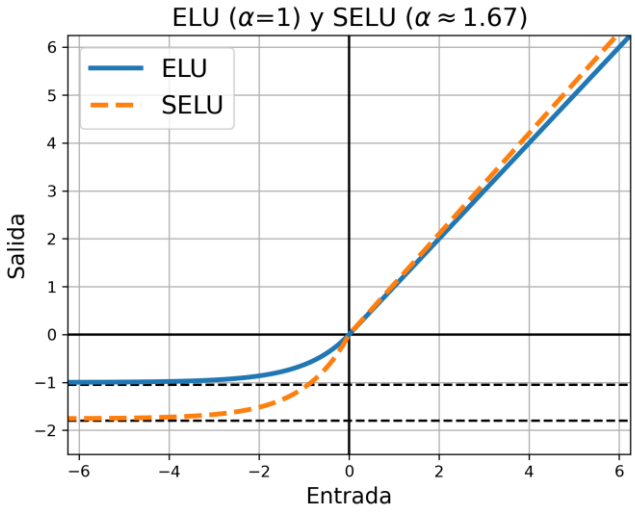

Figura 34. Función de activación ELU y SELU.

- GELU

La función GELU (Hendrycks & Gimpel, 2016), *gaussian error linear units*, se puede ver como una variante de la función de activación ReLU. Su definición se da en la Ecuación (88), donde ϕ es la función de distribución acumulativa (CDF) gaussiana, es decir $\phi(z)$ corresponde a la probabilidad de que un valor muestreado aleatoriamente de una distribución normal con media 0 y varianza 1 sea menor que

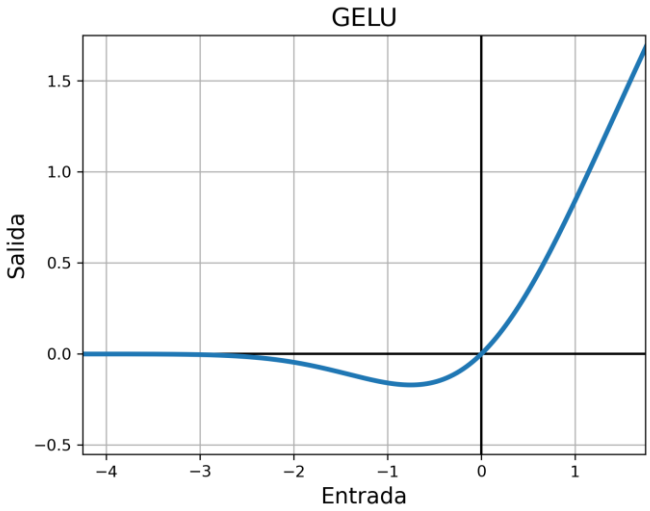

Figura 35. Función de activación GELU.

z. Esta función suele superar a todas las demás funciones de activación vistas hasta ahora. Sin embargo, requiere mayor capacidad computacional, por lo que el aumento de rendimiento no siempre es suficiente para justificar el coste adicional. La distribución de la función se puede observar en la Figura 35.

$$GELU(z) = z \cdot \phi(z) \tag{88}$$

Existen otras variantes de la función GELU como son: SiLU, *sigmoid linear unit*, también denominada Swish (Ramachandran et al., 2017); y Mish (Misra, 2019), que presentan unas características similares, aunque pueden presentar mejoras en tareas más complejas.

Finalmente, cabe preguntarse qué función de activación se debe emplear para las capas ocultas en una DNN. Generalmente ReLU es una buena opción para tareas sencillas, ya que está incluida en multitud de librerías y es muy rápida de calcular. Sin embargo, Swish probablemente ofrece mejores resultados para tareas más complejas. Mish puede ofrecer resultados ligeramente mejores, a costa de un incremento en el coste computacional. Cuando la latencia en tiempo es un parámetro fundamental, es recomendable emplear Leaky ReLU.

5.6 Optimización de los hiperparámetros

La flexibilidad de las redes neuronales es también uno de sus principales inconvenientes debido al alto número de hiperparámetros que se pueden ajustar para obtener un mejor rendimiento. No sólo se puede utilizar cualquier arquitectura de red imaginable, sino que incluso en una ANN básica se puede cambiar el número de capas, número de neuronas, el tipo de función de activación a utilizar en cada capa, la lógica de inicialización de pesos, el tipo de optimizador a utilizar, su tasa de aprendizaje, el tamaño del batch, etc.

El ajuste de los hiperparámetros es un campo de investigación activo y se están estudiando muchos enfoques. Uno de ellos es el desarrollado por DeepMind en 2017 (Jaderberg et al., 2017), donde los autores utilizaron un algoritmo evolutivo para optimizar conjuntamente una población de modelos y sus hiperparámetros. Estos algoritmos evolutivos incluso se han utilizado con éxito para entrenar redes neuronales individuales, sustituyendo al omnipresente descenso por gradiente (Iba & Nasimul, 2020; Zhan et al., 2022).

Pero a pesar de todos estos avances y de todas las herramientas automáticas, es necesario conocer que valores son razonables para cada hiperparámetro para poder construir un prototipo rápidamente que permita restringir el espacio de búsqueda. A continuación se definen los hiperparámetros básicos de cualquier red neuronal.

- Número de capas ocultas

  Para una gran cantidad de problemas, se puede empezar con una sola capa oculta y obtener resultados razonables. Teóricamente, un MLP con una sola capa oculta puede modelar incluso las funciones más complejas, siempre que tenga suficientes neuronas. Sin embargo, para problemas complejos, las redes DNN tienen una eficiencia muy superior ya que pueden modelar funciones complejas con un número

exponencialmente menor de neuronas, lo que permite alcanzar mejor resultados con la misma cantidad de datos de entrenamiento. Este mejor rendimiento se basa en la estructura jerárquica de los datos de entrada, que permite a la DNN tener ventaja sobre el modelo MLP. Las primeras capas ocultas de una red modelan las estructuras de bajo nivel, las capas ocultas intermedias combinan estas para modelar estructuras de nivel intermedio, y las capas cultas finales y la capa de salida combinan estas estructuras intermedias para modelar estructuras de alto nivel como pueden ser las clases en un problema de clasificación.

Esta arquitectura jerárquica no sólo ayuda a las DNN a converger más rápidamente hacia una buena solución, sino que también mejora su capacidad de generalización a nuevos conjuntos de datos. Por ejemplo, si se entrena un modelo para reconocer caras en imágenes y ahora se quiere entrenar una nueva red neuronal para clasificar emociones, se puede empezar el entrenamiento de la nueva red reutilizando las capas inferiores del primer modelo, en lugar de iniciar aleatoriamente los pesos y sesgos de las primeras capas de la red. De este modo, la red no tiene que aprender todas las estructuras de bajo nivel que aparecen en la mayoría de las imágenes; sólo tendrá que sólo tendrá que aprender las estructuras de alto nivel, que son las que caracterizan las diferentes emociones. Esto se denomina transferencia de aprendizaje, *transfer learning*.

En resumen, para muchos problemas se puede empezar con sólo una o dos capas ocultas, y se va aumentando el número de capas ocultas hasta que empiece a sobreajustar el conjunto de entrenamiento.

- Número de neuronas en las capas ocultas

  El número de neuronas en las capas de entrada y salida viene determinado por el tipo de entrada y salida que requiere la tarea, esto es, depende del tamaño del vector de datos de entrada y de las clases de salida. Generalmente se emplea el mismo número de neuronas en cada una de las capas ocultas, esto permite reducir el número de parámetros a optimizar, en vez de utilizar un tamaño de neurona por cada capa.

  Al igual que con el número de capas, se puede aumentar el número de neuronas gradualmente hasta que la red empiece a sobreajustarse. También se puede utilizar el método contrario, generar un modelo con más capas y neuronas de las que necesita, y luego emplear una técnica de regularización para evitar el sobreajuste. Con esta aproximación se evita generar un cuello de botella en el modelo que produzca resultados inadecuados.

  En general, se obtiene una mayor mejora cuando se aumenta el número de capas en lugar del número de neuronas por capa.

- Learning Rate

  La tasa de aprendizaje (*learning rate*) es sin lugar a duda el hiperparámetro más importante a la hora de configurar una red neuronal. Una forma de encontrar una

tase de aprendizaje óptima para un problema es hacer una búsqueda creciente. Se comienza con una tasa de aprendizaje muy baja, y gradualmente, se va aumentando hasta un valor muy alto. Para ello se multiplica la tasa de aprendizaje por un factor constante en cada iteración. Al representar gráficamente el valor de la función de pérdida respecto al logaritmo de la tasa de aprendizaje se debe observar que el valor de la pérdida va disminuyendo conforme aumenta la tasa de aprendizaje. Sin embargo, se alcanza un punto en el que la tasa de aprendizaje es demasiado alta, por lo que la pérdida se dispara de nuevo. La tasa de aprendizaje óptima se encuentra, generalmente, en un punto con un valor aproximadamente diez veces menor que el punto de inflexión.

Cabe destacar que la tasa de aprendizaje óptima depende de los demás hiperparámetros, especialmente del tamaño del batch, por lo que si se modifica algún hiperparámetro se debe actualizar también la tasa de aprendizaje.

- Tamaño de batch
  El tamaño del batch puede tener un impacto significativo en el rendimiento de un modelo y el tiempo de entrenamiento. La principal ventaja de utilizar un batch de gran tamaño es la posibilidad de procesarlo eficientemente empleado la GPU, ya que permite procesar más datos a la vez. Por lo tanto, se recomienda utilizar el tamaño de batch más grande posible que quepa en la memoria de la GPU. Sin embargo, en la práctica los batch demasiado grandes pueden provocar inestabilidades al principio del entrenamiento, lo que puede reducir la capacidad de generalización del modelo. Por ello, en algunos artículos se recomienda emplear tamaños de batch de entre 2 y 32 (Masters & Luschi, 2018).

  Sin embargo otras investigaciones apuntan en la dirección opuesta (Goyal, 2017; Hoffer et al., 2017). Estos trabajos demuestran que es posible emplear tamaño de batch muy grandes, junto con diversas técnicas, para obtener un tiempo de entrenamiento muy corto sin pérdida de generalización por parte del modelo. Una posible técnica es emplear un tamaño de batch grande e ir realizando un aumento incremental de la tasa de aprendizaje conforme se realizan las iteraciones. Si finalmente el entrenamiento es inestable, o el rendimiento final es decepcionante, se valora reducir el tamaño de batch y trabajar con valores pequeños.

- Función de activación
  Se han definido las principales funciones de activación en la sección 5.3. En general, la función de activación ReLU es una buena opción por defecto para todas las funciones ocultas. Para la capa de salida depende realmente del tipo de tarea que realiza el modelo.

- Número de iteraciones

  En la mayoría de los casos, no es necesario optimizar el número de iteraciones de entrenamiento, simplemente basta con utilizar un mecanismo de *early stopping*. Esta

técnica permite parar el proceso de entrenamiento cuando el valor del rendimiento del conjunto de datos de validación deja de disminuir, pasado un número de iteraciones.

Estos son los principales hiperparámetros que se definen en el presente trabajo, aunque existe un gran número de buenas prácticas para afrontar este proceso (Smith, 2018).

5.7 Batch normalization

Como se ha visto en la Sección 5.5, el uso de la inicialización He junto con la función de activación ReLU puede reducir significativamente el peligro del problema del descenso del gradiente al principio del entrenamiento, nada garantiza que no aparezcan de nuevo durante el proceso. Para abordar este problema, Sergey Ioffe y Christian Szegedy (Ioffe, 2015) proponen una técnica llamada normalización por lotes (*batch normalization*, BN), que aborda el problema del desvanecimiento de gradiente. La técnica consiste en incluir una operación en la red justo antes o después de la función de activación de cada capa oculta. Esta operación simplemente realiza una normalización de los datos de entrada centrada en cero, y luego escala y desplaza el resultado, de esta forma permite al modelo aprender cual es la escala óptima y la media de cada conjunto de datos de entrada a las capas. Si se define una red con una capa BN como primera capa, no es necesario realizar un procedimiento de normalización de los datos de entrada, ya que la capa BN lo realiza (aunque realiza la normalización teniendo en cuenta únicamente los datos del *batch*).

Para realizar esta operación, el algoritmo estima la media y la desviación estándar de cada *batch* de entrada a la red. El algoritmo paso a paso se define en la Ecuación (89):

$$\mu_B = \frac{1}{m_B}\sum_{i=1}^{m_B} x^{(i)}$$

$$\sigma_B^2 = \frac{1}{m_B}\sum_{i=1}^{m_B}\left(x^{(i)} - \mu_B\right)^2 \tag{89}$$

$$\hat{x}^{(i)} = \frac{x^{(i)} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

$$z^{(i)} = \gamma \otimes \hat{x}^{(i)} + \beta$$

Donde μ_B es un vector con los valores medios de la entrada del *batch*; m_B es el número de datos en el *batch*; σ_B^2 es un vector con las varianzas de los datos del *batch*; $\hat{x}^{(i)}$ es un vector de datos normalizados y con media cero; ε es un término denominado *smoothing term* que evita las divisiones por cero y que el valor del gradiente no crezca demasiado; γ es el vector con los parámetros de escala para la capa de salida; $\otimes$ representa la

multiplicación punto a punto; β es el vector de desplazamiento (*offset*) de salida de la capa; y z es la salida del algoritmo de BN y representa la entrada reescalada y desplazada.

Resumiendo, durante el entrenamiento el algoritmo BN estandariza las entradas de la red, las reescala y realiza una compensación. Durante el proceso de inferencia se realiza una modificación al algoritmo ya que, cuando se realizan predicciones para instancias individuales, no hay forma de calcular la media y desviación estándar. Aún en el caso en el que las predicciones se realicen por lote, no es posible asegurar que estas sean independiente o idénticamente distribuidas. Una solución al problema, y es la que implementa la librería Keras, es emplear una media móvil de la media y desviación típica de los *batch* analizados durante el entrenamiento. De este modo, la capa sólo normalizará sus entradas durante la inferencia después de haber sido entrenada con datos que tengan estadísticas similares a los datos de inferencia.

Ioffe y Szegedy demuestran que el algoritmo de *batch normalization* mejora considerablemente todas las redes neuronales profundas con las que experimentaron, dando lugar a una enorme mejora en tareas de clasificación. El problema del descenso del gradiente se redujo considerablemente. Las redes también eran mucho menos sensibles a la inicialización de los pesos, pudiendo utilizar tasas de aprendizaje mucho mayores y acelerando significativamente el proceso de aprendizaje. También, la normalización por lotes actúa como un regularizador, reduciendo la necesidad de otras técnicas de regularización.

Sin embargo, la normalización por lotes añade cierta complejidad al modelo, además es posible que el entrenamiento sea bastante lento, porque cada *epoch* tarda mucho más tiempo cuando se utiliza la normalización por lotes. Esto se compensa normalmente por el hecho de que la convergencia es mucho más rápida con BN, por lo que se necesitan menos épocas para alcanzar el mismo rendimiento y, en definitiva, reduce el tiempo total que necesita la red para su entrenamiento.

5.8 Dropout

Dropout es una de las técnicas de regularización más populares para DNN. Propuesto por Geoffrey Hinton et al. (Hinton, 2012) y desarrollado por Nitish Srivastava et al. (Nitish, 2014), ha demostrado ser una técnica muy eficaz. Muchas redes neuronales de última generación utilizan la técnica de dropout, ya que proporciona un aumento de la precisión de entre 1-2%.

Se trata de un algoritmo bastante simple: en cada paso de entrenamiento, cada neurona (excepto las neuronas de salida) tiene una probabilidad p de ser "abandonada" (*dropout*) temporalmente, lo que significa que no se tiene en cuenta durante ese paso de entrenamiento, pero puede volver a estar activa en el siguiente. El hiperparámetro p se denomina tasa de abandono (*dropout rate*) y suele tener un valor entre 0.1 y 0.5. Una vez completado el proceso de entrenamiento, las neuronas ya no se abandonan, ya que en el proceso de testeo se emplean todas las neuronas disponibles en la red. La idea

general detrás del éxito de esta técnica se fundamenta en el hecho de que las neuronas entrenadas con *dropout* no pueden adaptarse con todas las vecinas con las que están conectadas; tampoco pueden depender excesivamente de unas pocas neuronas de entrada y deben prestar atención a cada una de sus neuronas de entrada. Al final, se obtiene una red más robusta que generaliza mejor.

Otra forma de entender la capacidad de esta técnica es pensar que en cada paso de entrenamiento se genera una red neuronal única. Dado que cada neurona puede estar activa o ausente, hay un total de 2^N redes posibles (donde N es el número total de neuronas en las capas en las que se aplica el *dropout*). Se trata de un número tan grande que es prácticamente imposible que la misma red neuronal sea generada dos veces. Una vez que se han ejecutado 1.000 pasos de entrenamiento, se han entrenado 1.000 redes neuronales diferentes. Estas redes neuronales obviamente no son independientes porque comparten muchos de sus pesos, pero todas son diferentes. La red neuronal resultante puede verse como un conjunto promedio de todas estas redes neuronales más pequeñas.

Existe un gran interés en el desarrollo de esta técnica. Un artículo de Yarin Gal y Zoubin Ghahramani (Gal & Ghahramani, 2016) ofrece una sólida justificación matemática a la técnica de *dropout*. También, estos autores introducen una técnica denominada Monte Carlo Dropout, que puede aumentar el rendimiento de cualquier modelo entrenado con d*ropout* sin tener que reentrenarlo ni modificarlo en absoluto. También proporciona una medida de la incertidumbre del modelo, y puede aplicarse en tan sólo unas pocas líneas de código. Esto se consigue realizando un gran número de predicciones con el modelo entrenado con *dropout* activado, lo que nos da una estimación de Monte Carlo que es generalmente más fiable que el resultado de una sola predicción con el *dropout* desactivado. Aun así, hay que tener en consideración que el número de muestras Monte Carlo que se utilice es un hiperparámetro que puede ajustar. Cuanto mayor sea, más precisas serán las predicciones y sus estimaciones de incertidumbre. Sin embargo, si se duplica, también se duplicará el tiempo de inferencia. Además, a partir de un determinado número de muestras, la mejora será mínima. Es una cuestión de encontrar un equilibrio entre latencia y precisión, en función de los requisitos de la aplicación.

5.9 Redes Neuronales Convolucionales

Las Redes Neuronales Convolucionales, (*Convolutional Neural Network*, CNN), están inspiradas en el proceso que realiza el cerebro humano para procesar la visión. Estos sistemas son ampliamente utilizados en visión por computador (Bhatt et al., 2021), análisis de lenguaje natural (Palaz & Collobert, 2015) y reconocimiento del habla (Yousefi & Hansen, 2020). El primer ejemplo de CNN moderna es considerado LeNet-5, una CNN para la clasificación de dígitos propuesta por LeCun en 1998. Con la implementación de las unidades de procesamiento gráfico (GPUs), el uso de las CNN ha experimentado un crecimiento acelerado, permitiendo el desarrollo de redes más completas como Alexnet (Krizhevsky et al., 2017), VGGnet (Simonyan & Zisserman, 2014) y Resnet (He et al., 2016). Una CNN se puede definir como una red neuronal sin retroalimentación y consta de una capa de entrada, capas ocultas y una capa de salida. En la Figura 36 se ilustra una arquitectura típica de CNN.

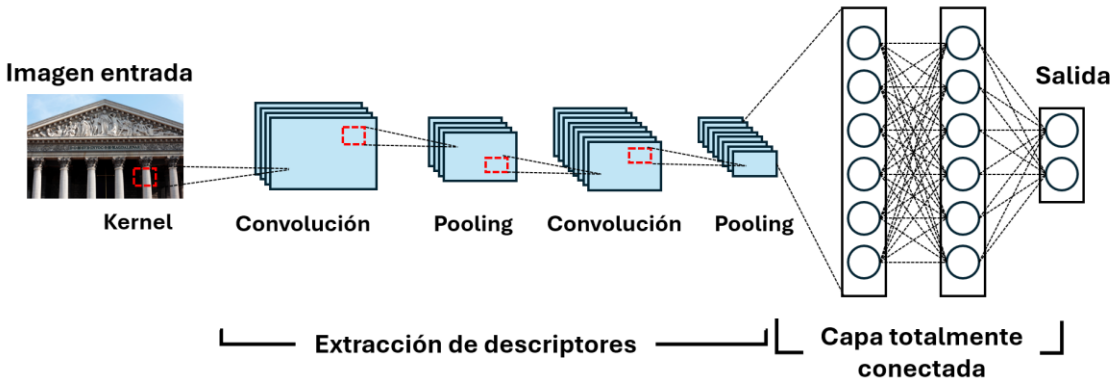

Figura 36. Esquema general de una red neuronal convolucional CNN.

La arquitectura de las CNNs presenta un conjunto de elementos comunes a las redes ANN analizadas en las secciones anteriores, como son las capas totalmente conectas y las funciones de activación de las neuronas. De hecho, la capa final de una CNN está formada por un conjunto de capas conectadas, lo que compone una DNN. La principal diferencia reside en el uso de dos nuevas capas: capas convolucionales (*convolutional layer*), y capas de agrupación *(pooling layer)*, que se describen a continuación.

La capa de entrada (*input layer*) que se muestra antes de la capa convolucional recibe las imágenes de entrada de la red que pueden ser: monocanal, como una imagen en escala de grises; o multicanal, como una imagen en color de tres canales RGB.

5.9.1 Capa convolucional

La capa convolucional es el núcleo de una CNN, donde se extraen las características de la imagen de entrada. Cada una de las neuronas de la primera capa convolucional no están conectadas a cada píxel de la imagen de entrada, como sucedía en las redes ANN, sino solo a un conjunto de píxeles (por ejemplo, una región de tamaño 3×3), que forman lo que se denomina su campo receptivo, o *kernel* (Ekman, 2021), como se muestra en la Figura 37.

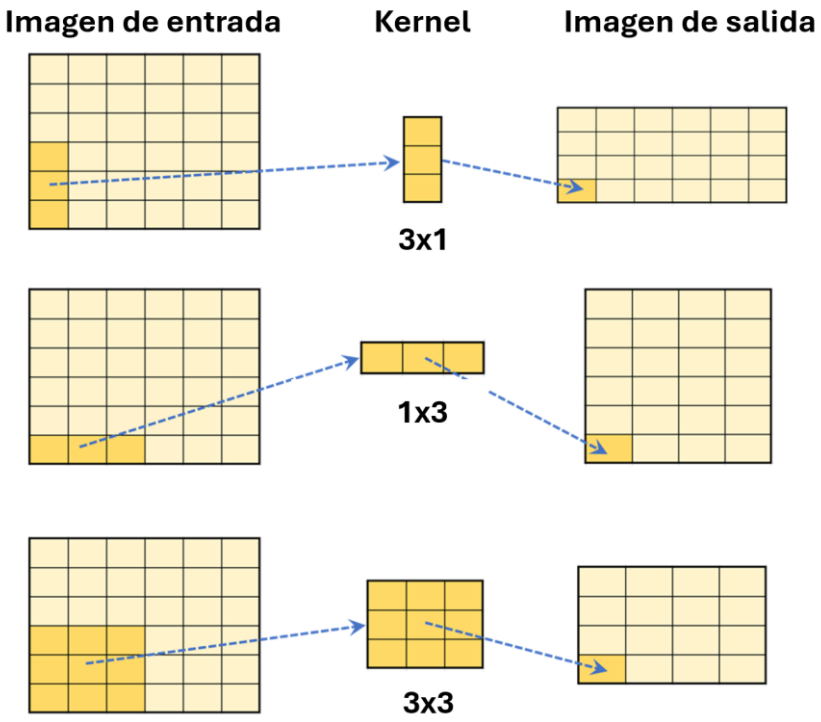

Figura 37. Resultado de aplicar un kernel con diferente tamaño a una imagen de entrada.

Sucede lo mismo con las siguientes capas, cada neurona se encuentra conectada a una pequeña zona de la capa anterior. La distancia horizontal o vertical de un campo receptivo al siguiente dentro de la misma capa se denomina *stride*, y determina el número de neuronas que son necesarias para cubrir toda la imagen. Esta estructura permite a la red obtener información sobre las características de bajo nivel en la primera capa, y después utilizar esta información para generar características de un nivel mayor que alimentan a la siguiente capa. Esta estructura jerárquica es habitual en las imágenes del mundo real, razón por la cual este tipo de redes funciona tan bien en el reconocimiento de imágenes.

Los pesos de una neurona individual se pueden representar como una pequeña imagen del tamaño del campo receptivo. Estos pesos se denominan filtros, núcleos de convolución, o *kernel*. Cada uno de los píxeles del filtro están conectados a una neurona y, junto con el valor del *bias*, se utilizan para identificar ciertos patrones. En la Figura 38 se observa la aplicación de un filtro vertical y horizontal a una misma imagen. Las neuronas que utilizan estos filtros ignoran todo lo que haya en su campo receptivo excepto la línea vertical/horizontal central.

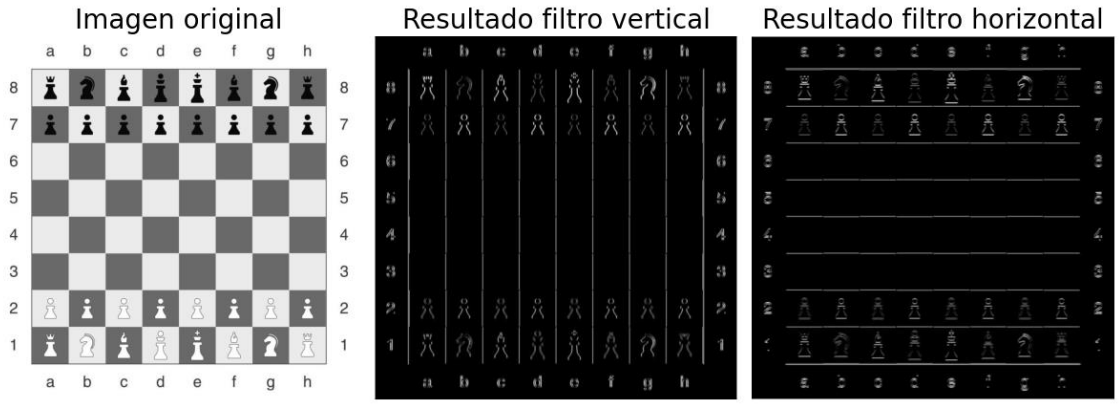

Figura 38. Resultado de aplicar un filtro vertical y horizontal a una imagen. En cada una de las imágenes se observan las características resaltadas por el tipo de filtro.

De esta forma, una capa con neuronas que utilizan el mismo filtro produce un mapa de características que resalta las áreas de la imagen que más activan el filtro. Estos filtros no se tienen que definir manualmente, ya que durante el entrenamiento, la capa convolucional aprenderá de forma automática los filtros más útiles para la tarea, mientras que las capas superiores utilizan esta información para combinarlos y analizar patrones más complejos.

En las imágenes anteriores se represente la salida de la capa convolucional como un mapa bidimensional, pero en realizad una capa convolucional tiene tantos filtros como se especifique, y cada uno de ellos produce un mapa de características empleando un filtro diferente.

Todas las neuronas de un mismo mapa de características comparten los mismos parámetros (es decir el mismo *kernel* y el mismo *bias*), son idénticas, pero reciben una parte diferente de la imagen, por lo que producen una salida diferente. El hecho de que todas las neuronas de un mapa de características compartan los mismos parámetros reduce drásticamente el número de parámetros del modelo. Hay que destacar que cuando se emplea una imagen tridimensional el tamaño de los filtros se amplía (por ejemplo, un filtro de tamaño $3 \times 3 \times 3$).

5.9.2 Pooling Layer

Otra parte fundamental de la estructura red convolucional es la capa de agrupación, *pooling layer*. El objetivo principal de esta capa es submuestrear la imagen de entrada para reducir la carga computacional, el uso de memoria y el número de parámetros, limitando así el riesgo de sobreajuste. Es habitual inserta una capa *pooling* después de

cada capa convolucional. Al igual que en las capas convolucionales, cada neurona de la capa *pooling* se encuentra conectada a un pequeño campo receptivo, del que se tiene que definir su tamaño y el valor de *stride*, al igual que en las capas convolucionales. Sin embargo, una neurona de agrupación no tiene pesos; todo lo que hace es agregar las entradas mediante una operación. Principalmente existe dos tipos de agrupación: *max pooling*, que toma el valor máximo del mapa de características; y el *mean pooling*, que emplea el valor medio. Generalmente la capa *pooling* trabaja cada canal de entrada de forma independiente, por lo que esa dimensión permanece constante. Finalmente, el resultado de la capa de *pooling* alimenta la siguiente capa convolucional de la red, que extrae características de una imagen ya procesada.

Además de reducir los cálculos, el uso de memoria y el número de parámetros, una capa de *pooling* introduce cierto nivel de invariancia a pequeñas traslaciones, invarianza rotacional, y una ligera invarianza de escala. Esta invarianza, aunque limitada, puede ser útil en casos en los que la predicción debe depender de estos detalles. Sin embargo, esta capa introduce algunos inconvenientes, el principal es que produce una gran reducción de la imagen de entrada, por lo que al trabajar con imágenes pequeñas limita el número de capas que se puede emplear. Otro inconveniente es, por ejemplo, en las aplicaciones de segmentación semántica en las que hay que clasificar cada píxel de una imagen según el objeto que representa, obviamente, si la imagen de entrada se traslada un píxel a la derecha, la imagen de salida también debería desplazarse un píxel hacia la derecha.

Para finalizar, la salida de la última capa de convolución alimenta una capa de tipo *flatten layer*, que toma todos los valores de entrada y los convierte en un vector unidimensional que alimenta una red totalmente conectada, una DNN. Esta última parte es la que realiza el proceso de clasificación.

6. Evaluación de los resultados

En el presente capítulo se presentan las métricas empleadas para la evaluación cuantitativa de los resultados obtenidos en esta investigación. En primer lugar, se introduce el coeficiente de correlación de Pearson, que permite analizar la relación lineal entre variables continuas y su significación estadística. A continuación, se describe la matriz de confusión junto con las métricas derivadas de ella, proporcionando así una caracterización detallada del tipo de errores cometidos por los sistemas de clasificación. Estas herramientas son esenciales para validar la robustez y precisión de los modelos propuestos.

6.1 Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (también llamado coeficiente r de Pearson) es una medida estadística que cuantifica la asociación lineal entre dos variables continuas (Akoglu, 2018). Matemáticamente, se define como la covarianza entre las dos variables dividida por el producto de sus desviaciones estándar. Para una muestra de tamaño n con valores (x_i, y_i), el coeficiente de Pearson se calcula como:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{90}$$

Donde $\bar{x}$ e $\bar{y}$ son las medias muestrales del conjunto total de datos X e Y respectivamente. Este coeficiente es un valor adimensional que siempre oscila entre -1 y +1. Un valor de +1 indica una correlación lineal positiva perfecta (cuando una variable aumenta, la otra aumenta proporcionalmente), mientras que -1 indica una correlación lineal negativa perfecta (una variable aumenta mientras la otra disminuye proporcionalmente). Un valor de 0 significa que no existe correlación lineal aparente entre las dos variables. Es importante destacar que r es independiente de las unidades de medida de las variables, ya que al estar normalizado por la desviación estándar produce un índice de relación lineal adimensional. Además, el coeficiente de Pearson es simétrico, es decir, $r_{XY} = r_{YX}$, y no implica relación de dependencia casual, solo asociación estadística.

La magnitud de r refleja la fuerza de la relación lineal, y su signo indica la dirección. Para interpretar su valor es habitual apoyarse en criterios cualitativos. En términos generales, $|r| < 0.3$ indica una correlación lineal débil o despreciable, $|r| \approx 0.5$ indica una correlación moderada, y $|r| > 0.7$ indica una correlación fuerte, aunque estos umbrales son orientativos y distintos autores proponen rangos ligeramente diferentes (Schober et al., 2018).

Cabe subrayar que correlación no implica causalidad. Una correlación alta entre dos variables no significa que una provoque a la otra; pueden existir factores externos o relaciones indirectas entre ellas. Además, r únicamente mide relaciones lineales. Si la relación verdadera entre X e Y es no lineal (por ejemplo, curva o en forma de U), el coeficiente de Pearson podría ser cercano a 0 pese a existir una fuerte relación no lineal entre las variables. Por ello, se examina conjuntamente el diagrama de dispersión de los datos para verificar la linealidad de la relación antes de confiar en r.

6.1.1 Significancia estadísticas y valor p

Al reportar una correlación de Pearson, es fundamental indicar si dicha correlación es estadísticamente significativa, lo cual se valora mediante el valor p asociado. Este valor p se obtiene al contrastar la hipótesis nula de "correlación poblacional cero" ($H_0: \rho = 0$) frente a la alternativa de que ρ es distinto de cero. Bajo el supuesto de normalidad bivariada, el estadístico de prueba puede expresarse como muestra la Ecuación (91):

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{91}$$

El cual sigue aproximadamente una distribución t de Student con $n-2$ grados de libertad cuando H_0 es verdadera. A partir de este estadístico, empleando la Función de Distribución Acumulada (*Cumulative Distribution Function*, CDF), se obtiene el valor p (Cohen et al., 2009). Un valor p pequeño (típicamente, $p < 0.05$) indica que es muy poco probable observar un coeficiente de correlación de esa magnitud por puro azar si en realidad no hubiera correlación en la población. En tal caso, se rechaza la hipótesis nula y se concluye que la correlación observada es significativamente distinta de cero, dando soporte estadístico a la existencia de una asociación lineal. Por el contrario, un valor p mayor (ej., $p > 0.05$) sugiere que la evidencia no es suficiente para descartar que la correlación verdadera sea nula, por lo que la asociación observada podría atribuirse al azar.

Es importante notar que el valor p refleja la confianza estadística sobre la presencia de una correlación, pero no informa sobre la magnitud o relevancia práctica de dicha correlación. Con tamaños de muestra muy grandes, incluso coeficientes r pequeños (por ejemplo 0,10 o 0,20) pueden resultar estadísticamente significativos (Schober et al., 2018). Inversamente, con muestras pequeñas un coeficiente moderado puede no alcanzar significación. Por ello, se deben considerar conjuntamente la magnitud de r y su significación estadística. También es útil proporcionar un intervalo de confianza para p, lo que cuantifica el rango de valores plausibles de la correlación real con cierto nivel de confianza.

No obstante, es importante reconocer las limitaciones y supuestos asociados al coeficiente de Pearson. Como se menciona en esta sección, Pearson solo detecta relaciones lineales: si la relación entre dos variables es fuertemente no lineal o hay patrones más complejos, r podría no reflejar esa asociación. En tales casos, son más

adecuados otros coeficientes como Spearman o métodos no lineales (Hauke & Kossowski, 2011). Relacionado con lo anterior, Pearson asume idealmente que los datos provienen de una distribución normal bivariada; aunque el coeficiente en sí se puede calcular siempre, la validez de sus inferencias requiere variables aproximadamente normales y con varianza constante. Si las variables presentan distribuciones muy asimétricas, rangos limitados o presencian valores atípicos extremos (*outliers*), el valor de r puede verse distorsionado.

6.2 Matriz de confusión

La matriz de confusión es una herramienta fundamental para evaluar el desempeño de un modelo de clasificación (ya sea binario o multiclase) en un punto específico de operación, esto es, con un umbral o criterio de decisión específico. La matriz de confusión resume en forma de tabla la comparación entre las predicciones del modelo y los valores reales de las clases, proporcionando una visión detallada de los aciertos y errores del clasificador. Esto permite derivar múltiples métricas de evaluación que cuantifican la calidad de las predicciones, tales como la precisión (*precision*), la sensibilidad (*recall*) y la medida F1 (*F1-score*) (Manual, 2013).

6.2.1 Componentes de la matriz de confusión

Para un problema de clasificación binaria, que clasifica las instancias de los datos como positivas o negativas, la matriz de confusión es una tabla de 2×2 que contiene cuatro resultados posibles para cada instancia predicha:

- **Verdaderos Positivos (TP)**: instancias positivas que el modelo clasifica correctamente como positivas (aciertos de la clase positiva).
- **Falsos Positivos (FP)**: instancias negativas que el modelo clasifica incorrectamente como positivas (errores tipo I, falsas alarmas).
- **Falsos Negativos (FN)**: instancias positivas que el modelo clasifica incorrectamente como negativas (errores tipo II, omisiones).
- **Verdaderos Negativos (TN)**: instancias negativas que el modelo clasifica correctamente como negativas (aciertos de la clase negativa).

En la Figura 39 se puede observar un ejemplo de matriz de confusión.

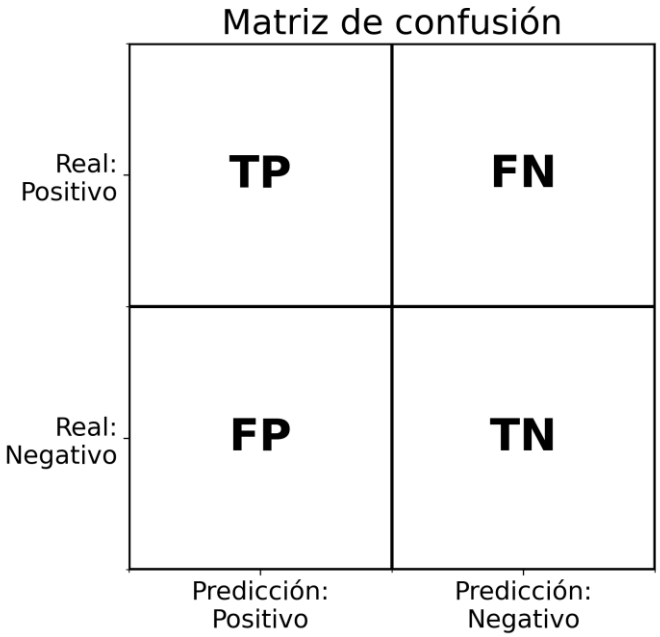

Figura 39. Matriz de confusión con los resultados posibles para el caso de clasificación binaria.

Estos cuatro valores (TP, FP, FN, TN) abarcan todas las posibilidades de predicción y su suma equivale al total de ejemplos evaluados. A partir de ellos se calcula la exactitud o *accuracy* (proporción de aciertos totales) y otras métricas más sensibles a la distribución de clases. La matriz de confusión proporciona información más allá de la exactitud, al revelar qué tipos de errores comete el modelo (falsos positivos vs. falsos negativos), lo cual es crucial cuando los costes de error son distintos o cuando se analizan modelos desbalanceados en los que una clase es mucho más frecuente que la otra.

En el caso de clasificación multiclase, la matriz de confusión se generaliza a una tabla $n \times n$ (siendo n el número de clases diferentes). Cada celda (i, j) de la matriz contiene el número de instancias cuya clase real es i y que fueron predichas por el modelo como clase j. La interpretación de los elementos diagonales (verdaderos aciertos por clase) y los fuera de la diagonal (errores de confusión entre clases) sigue el mismo principio. Para extender las métricas a múltiples clases, suele calcularse primero métricas por clase (considerando cada clase como "positiva" contra el resto) y luego se realizan el promedio para obtener un solo valor global de precisión, sensibilidad o F1, según corresponda (Powers, 2015). De esta manera, la matriz de confusión y sus métricas asociadas permiten evaluar de forma integral el comportamiento de modelos tanto binarios como multiclase.

6.2.2 Métricas de evaluación derivadas de la matriz de confusión

A continuación se definen métricas claves derivadas de la matriz de confusión: exactitud, precisión, sensibilidad (o *recall*), especificidad, y F1-score, las cuales son ampliamente utilizadas para evaluar clasificadores en tareas de aprendizaje automático. Se incluyen

sus definiciones formales junto con una explicación de cómo se interpretan en términos de TP, FP, FN y TN.

- **Exactitud (*accuracy)*:** La exactitud indica la proporción de instancias clasificadas correctamente entre todas las instancias evaluadas, como muestra la Ecuación (92). Sin embargo, esta métrica puede ser engañosa en contextos con clases desbalanceadas.

$$Precisión = \frac{TP + TN}{TP + TN + FP + FN} \tag{92}$$

- **Precisión**: mide la exactitud de las predicciones positivas del modelo. Es la proporción de casos predichos como positivos que realmente son positivos (Sokolova & Lapalme, 2009). En otras palabras, indica qué fracción de las instancias que el clasificador marca como positivas son verdaderos positivos. Se calcula como el número de verdaderos positivos dividido entre el total de positivos predichos (verdaderos positivos más falsos positivos), como muestra la Ecuación (93):

$$Precisión = \frac{TP}{TP + FP} \tag{93}$$

  Un modelo con precisión = 1.0 (100%) no comete falsos positivos (todas sus predicciones positivas son correctas), mientras que una precisión baja indica que el modelo produce muchos falsos positivos en relación con sus aciertos positivos. En contextos con datos muy desbalanceados (por ejemplo, detección de fraudes con muy pocos casos positivos), la precisión por sí sola puede ser engañosa, ya que ignora los falsos negativos –por ello suele complementarse con la sensibilidad para evaluar el desempeño de forma más equilibrada (Saito & Rehmsmeier, 2015).

- **Sensibilidad (*recall*)**: corresponde a la capacidad de detección de la clase positiva por parte del modelo. Es la proporción de casos realmente positivos que el clasificador logra identificar correctamente como positivos. Su formulación viene dada por el número de verdaderos positivos dividido entre el total de positivos reales (verdaderos positivos más falsos negativos), como muestra la Ecuación (94):

$$Sensibilidad = \frac{TP}{TP + FN} \tag{94}$$

  Un valor de sensibilidad cercano a 1 indica que el modelo logra capturar todos los casos positivos reales (no comete falsos negativos), mientras que una sensibilidad baja indica que se escapan muchos positivos reales (el modelo tiene muchos falsos negativos).

- **Especificidad**: Esta métrica representa la proporción de casos negativos que el modelo identifica correctamente como negativos. Es especialmente importante en contextos donde los falsos positivos tienen un alto coste, complementando así la información proporcionada por la sensibilidad (James et al., 2013).

$$Especificidad = \frac{TN}{TN + FP} \tag{95}$$

- **Medida F1 (*F1-score*)**: En general, existe una relación inversa entre precisión y sensibilidad. Por ello, es útil considerar una métrica combinada que resuma ambos aspectos. EL *F1*-score es la media armónica entre la precisión y la sensibilidad (Christen et al., 2023). Proporciona un solo valor que equilibra ambas métricas, especialmente útil cuando se busca un compromiso entre precisión y sensibilidad. La puntuación F1 se define como muestra la Ecuación (96):

$$Sensibilidad = 2 \cdot \frac{Precisión \cdot Sensibilidad}{Precisión + Sensibilidad} \tag{96}$$

  Su valor máximo es 1 (cuando el modelo alcanza simultáneamente precisión = 1 y sensibilidad = 1) y su valor mínimo es 0 (si cualquiera de las dos métricas es 0). La F1 es especialmente informativa en escenarios con clases desbalanceadas, donde una alta exactitud puede ser engañosa. Al ser un promedio armónico, la F1 resultará alta solo si tanto la precisión como la sensibilidad son altas y balanceadas entre sí; si una de ellas es baja, la F1 también lo reflejará con un valor bajo. En términos prácticos, una *F1-score* alta indica que el modelo mantiene un buen equilibrio entre no producir excesivos falsos positivos y no omitir demasiados positivos reales.

7. Resultados de las contribuciones

En este capítulo se presentan de forma organizada los principales resultados obtenidos en cada una de las contribuciones que integran esta tesis, ofreciendo para cada trabajo un resumen de la metodología empleada junto con los hallazgos más relevantes.

7.1 Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically?

En este primer trabajo, publicado en *Applied Sciences* el 22 de septiembre de 2021, se estudia que tipo de enunciados son los más adecuados para detectar la hipernasalidad de forma automática. Tradicionalmente se ha analizado únicamente las vocales sostenidas para la detección de la hipernasalidad, debido principalmente a la dificultad del análisis y evaluación perceptual de sonidos más complejos. Esto contrasta con las recomendaciones clínicas, que consideran necesario utilizar diversos tipos de enunciados (por ejemplo, sílabas repetidas, sonidos sostenidos, etc.).

Por lo tanto, este estudio explora la viabilidad de detectar la hipernasalidad automáticamente basándose en muestras de habla distintas de las vocales sostenidas, empleando para ello grabaciones realizadas con un dispositivo móvil. Para ello se definen seis tipos diferentes de tareas, siguiendo las recomendaciones del *International Speech Parameters Group* (Henningsson et al., 2008), descritas en la Tabla 2.

Tabla 2. Enunciados de las grabaciones.

Tareas	Instrucciones	Enunciados
T1. Conteo	Contar del 1 al 10	*Uno, dos, tres, cuatro, cinco seis, siete, ocho, nueve, diez*
T2. Sílabas	Repetir la sílaba rápidamente	*pa pa pa…, ta, ta, ta…, ka ka ka…* *pi pi pi…, ti, ti, ti…, ki ki ki…*
T3. Consonantes sostenidas	Producir una consonante larga	*/ fffff… /* */ sssss… /*
T4. Vocales sostenidas	Producir una / a / larga	*/ aaaaa… /*
T5. Palabras	Repetir palabras	*moto, boca, piano, pie, niño, llave, luna, campana* *indio, dedo, gafas, silla, cuchara, sol, jaula, zapatos*
T6. Frases	Repetir frases	**Consonantes sonoras átonas:** *Al gato de Ágata le gusta el yogur (/ g /)* *A David le duele el dedo (/ d /).* *El bebé va bien con babuchas (/ b /)* **Consonantes sordas:** *Tómate toda tu taza de té (/ t /)* *Papá puede pelar a Pili (/ p /)*

Quique coge el papel de calco (/ k /)
Consonantes fricativas*:*
Si llueve le llevo la llave a Yolanda (/ ʝ /)
Susi sale sola y se ensucia (/ s /)
Fali fue a la feria inflando un globo (/ f /)
Los zapatos de Cecilia son azules (/ θ /)
La jirafa de Jesús se mojó jugando (/ x /)
Consonante africada:
Chuchu y Chelo chillan mucho (/ ʧ /)
Aproximantes:
Lali y Luna leen los carteles (/ l /)
Vocales:
Uy, ahí hay algo
Nasales:
Mi mamá me mima mucho (/ m /)
El nene nos canta una nana (/ n /)

Empleando una aplicación móvil, se realizan las grabaciones a niños/as y mujeres adultas provenientes de tres países hispanohablantes: Chile, Ecuador y España, Se crea una base de datos de locutores sanos (controles) e hipernasales (pacientes). Un locutor se define como hipernasal cuando presenta un historial médico asociado a la presencia de hipernasalidad en el habla.

Aplicando los siguientes criterios de admisión se obtiene una muestra final de 39 pacientes y 39 controles:

- Edad y género: mujer adulta entre 18-42 años o niña/o de edad 5-15 años.
- Frecuencia fundamental media por encima de 180 Hz.
- El paciente ha completado, al menos, el 90% de los enunciados de la prueba.
- Se observa un bajo nivel de ruido acústico en las grabaciones.
- Detección de la hipernasalidad en al menos 5 enunciados por parte de 3 terapeutas de forma independiente.

En una primera fase se evalúa por separado cada uno de los 44 enunciados (tareas de conteo, secuencias silábicas, consonantes y vocal sostenidas, palabras y frases) empleando tres clasificadores diferentes: RF, SVM y DNN, para clasificar cada uno de los enunciados como sano o hipernasal. Para cada uno de los enunciados se calcula la siguiente selección de descriptores: trece coeficientes MFCC; los tres primeros formantes de audio, junto con sus anchos de banda y distancias relativas; la componente de frecuencia fundamental F0 junto con los dos primeros armónicos; y el VLTHTR (Cairns et al., 1996; Dubey et al., 2018; Lee et al., 2006; Vijayalakshmi et al., 2007). Los resultados se observan en la Figura 40.

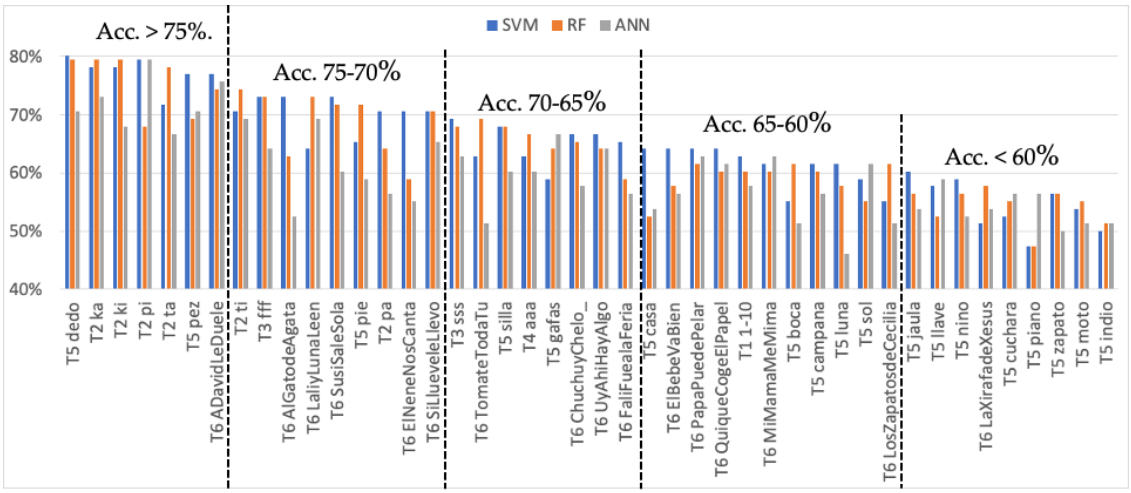

Figura 40. Precisión de los enunciados individuales. Se ordenan, de izquierda a derecha, según la precisión en el mejor clasificador.

Las tasas de acierto oscilan entre el 46% y el 81%, observándose una variabilidad considerable según el tipo de enunciado. Los mejores resultados (precisión > 75%) se obtienen con las secuencias silábicas (/ ta, ka, pi, ki /), algunas palabras (entre ellas *dedo* y *pez*) y determinadas frases diseñadas con predominio de oclusivas o fricativas (“*A David le duele el dedo*”). En cambio, el conteo del 1 al 10 y la vocal sostenida / a / proporcionan resultados claramente inferiores, lo que indica que no todos los enunciados son igualmente informativos para la detección automática de hipernasalidad.

En una segunda fase se explora el efecto de entrenar los clasificadores con más de un enunciado. Para ello se construyen listas óptimas de combinaciones de enunciados, añadiendo progresivamente aquellos que incrementaban la precisión. Para ello se emplea el siguiente procedimiento: (1) se selecciona el enunciado que mejores resultados de *accuracy* ofrece cuando se emplea de forma individual. (2) Se realizan de nuevo tantas simulaciones como enunciados quedan, 43 la primera vez, utilizando para el proceso de entrenamiento y test el enunciado mejor clasificado en el paso anterior junto con cada uno de los restantes. La combinación de 2 enunciados que ofrezca mejores resultados es seleccionada. (3) Se repite el paso 2 añadiendo un nuevo enunciado siempre que el valor de accuracy aumente. Este proceso se realiza de forma independiente para los tres clasificadores empleados.

Los resultados que se observan en la Tabla 3Tabla 4 muestran que con SVM, la combinación de la palabra *dedo* con la fricativa sostenida / f / y la secuencia / pa / alcanza una exactitud del 92%, con sensibilidad y especificidad también del 92%. Con RF y DNN se observan incrementos similares al pasar de un enunciado a dos, aunque la precisión tiende a decrecer cuando se incluyen más enunciados, probablemente debido a la aparición de rasgos acústicos redundantes durante la selección de características.

Tabla 3. Listas de enunciados óptimos.

Clasificadores	Enunciados óptimos y exactitud
SVM	Primer elemento: T5 dedo (81%) +T3 / f / (88%) +T2 / pa / (92%) +T6 Susi sale sola (92%) +T6 A David (91%)
RF	Primer elemento: T2 ka (84%) +T5 dedo (86%) +T3 / f / (83%)
ANN	Primer elemento: T2 pi (79%) +T6 A David (86%) +T2 / ka / (76%)

A continuación, se crea un índice global por hablante, *Hypernasality Score* (*HN Score*), que permite explorar la viabilidad de calcular una puntuación de hipernasalidad para cada locutor. El *HN Score* es el resultado de dividir el número de veces que el hablante ha sido clasificado como hipernasalidad entre el número total de pruebas. Así, la puntuación de hipernasalidad oscila entre 0% (ninguna locución ha sido clasificada como hipernasal) y 100% (todas las locuciones han sido clasificadas como hipernasales). En este caso se utilizan estos criterios para interpretar los resultados:

- *HN Score* < 40%: el hablante no es hipernasal.
- *HN Score* entre 40-60%: No se puede confirmar ni descartar la hipernasalidad.
- *HN Score* > 60%: el hablante es hipernasal.

Se evalúan distintas configuraciones: (a) los 44 clasificadores con un único enunciado, (b) los 16 enunciados con precisión individual superior al 70% y (c) una selección reducida de 7 enunciados con la mayor precisión más la lista óptima SVM obtenida en el análisis anterior, como se observa en la Figura 41, donde cada barra azul representa a un control y cada barra naranja representa a un paciente. El recuadro gris muestra a los locutores con puntuaciones intermedias (40-60%). Cabe destacar que las barras azules a la izquierda del recuadro gris son TN y las barras naranjas a la derecha son TP. Por el contrario, las barras naranjas a la izquierda son FN y las barras azules a la derecha son FP. En la última configuración (HN Score 7 + Sel) la media del *HN Score* fue del 81% en el grupo de pacientes y del 20% en el grupo control, con solo un falso positivo y un falso negativo, y únicamente tres locutores situados en la zona intermedia (40–60%), lo que refleja una clara separación entre hablantes hipernasales y controles.

Por último, se analiza el impacto del ruido de fondo, inevitable al trabajar con grabaciones remotas en entornos no controlados. Aunque una proporción apreciable de pacientes presentaba más del 20% de sus enunciados con ruido por encima de 50 dB, el *HN Score* medio de estos pacientes fue incluso ligeramente inferior al de los pacientes grabados en condiciones más silenciosas (79% frente a 83%), lo que sugiere que el ruido

Figura 41. Puntuación HN utilizando 44 enunciados (arriba), los 16 mejores enunciados (precisión > 70 %) (en el centro), y los 7 mejores + la lista óptima SVM (abajo).

ambiental no sesgó de forma sistemática la clasificación. En conjunto, los resultados confirman que (i) las secuencias silábicas y ciertos ítems léxicos y oracionales son especialmente adecuados para la detección automática de hipernasalidad, (ii) la combinación de varios enunciados y varios clasificadores mejora de forma notable el rendimiento y (iii) es factible obtener medidas automáticas fiables de hipernasalidad a partir de grabaciones realizadas con dispositivos móviles de uso cotidiano.

7.2 Unmasking Nasality to Assess Hypernasality

Este segundo trabajo, publicado en *Applied Sciences* el 23 de noviembre de 2023, continúa el estudio de la clasificación automática de la hipernasalidad. A diferencia de la publicación anterior, se realiza un análisis empleando tres tipos diferentes de señales de audio: oral, captada exclusivamente en la boca; nasal, captada cerca de la nariz; y la señal monofónica, que es el resultado de combinar las dos anteriores en un único canal de audio. Para ello se emplea un nasómetro, que permite grabar señales de audio con dos canales, oral y nasal, junto con un micrófono convencional para la señal monofónica. El objetivo principal de este trabajo es estudiar si el uso de la señal separada boca-nariz presenta alguna ventaja frente a la señal monofónica para la evaluación automática de la hipernasalidad. Se define para ello tres tipos de análisis.

Para el primero, se explora en qué medida las diferencias espectrales ente vocales orales y nasalizadas varían en función del tipo de señal utilizado: monofónica, oral, o nasal. La base de datos generada para este análisis consiste en las grabaciones de dos grupos de hablantes leyendo en voz alta diversos textos en español. Todos los hablantes son mujeres estudiantes universitarias con edades entre 17 y 25 años, hablantes nativa de español, y habitantes del sur de España. Un grupo (N=55) se graba empleando un Nasómetro icSpeech (Rose Medical Solutions Ltd., Canterbury, UK), mientras que el segundo grupo (N=189) se graba empleando un micrófono monofónico Shure WH20XLR (Shure Incorporated, Niles, IL, USA) conectado a una grabadora Zoom H4n (Zoom Corporation, Tokyo, Japan). Cada locutor lee en voz alta entre uno y tres textos fonéticamente balanceados, junto con textos que contiene múltiples sonidos nasales.

Se obtiene una transcripción fonética de las grabaciones empleando Praat y Montreal Forced Aligner. Para cada señal se calculan los coeficientes MFCC empleando una ventana móvil de 25 ms de duración y 15 ms de solapamiento. A continuación, cada ventana se clasifica en cuatro clases: vocal oral (OV), consonante oral (OC), consonante nasal (NC), y vocal nasal (NV). En el caso de las tres primeras, la clasificación es automática debido a que el inventario fonético del español incluye estos tipos de sonidos. Sin embargo, para las vocales nasalizadas, la nasalización se produce cuando la vocal precede a una consonante nasal. Debido a que la longitud de las vocales en español es variable, la nasalización es más evidente en las zonas cercanas a la consonante nasal, por lo que se segmenta la vocal nasalizada de la siguiente forma: 30% de la vocal cuando ésta tenía una duración de al menos 100 ms; el 50%, cuando la vocal duraba entre 60 y 100 ms; y la vocal completa si la duración es inferior a 60 ms.

Como el tamaño de la base de datos podría influir en los resultados, se selecciona un subconjunto de 2 horas de datos, equivalente a 184k muestras, de cada tipo de señal, manteniendo una distribución uniforme entre las distintas clases. Terminado el proceso de anotado, la distancia entre los sonidos orales y nasales en las vocales se calcula empleando la distancia euclídea entre los MFCCs.

Los resultados muestran que la distancia euclídea entre los coeficientes MFCC de vocales orales las nasales, en general, mayor cuando las vocales se registran con el micrófono de nariz que con el de boca o con un micrófono monofónico, como muestra la Figura 42. Esta diferencia se observa claramente para / a, e, i, o / y solo se atenúa en el caso de / u /.

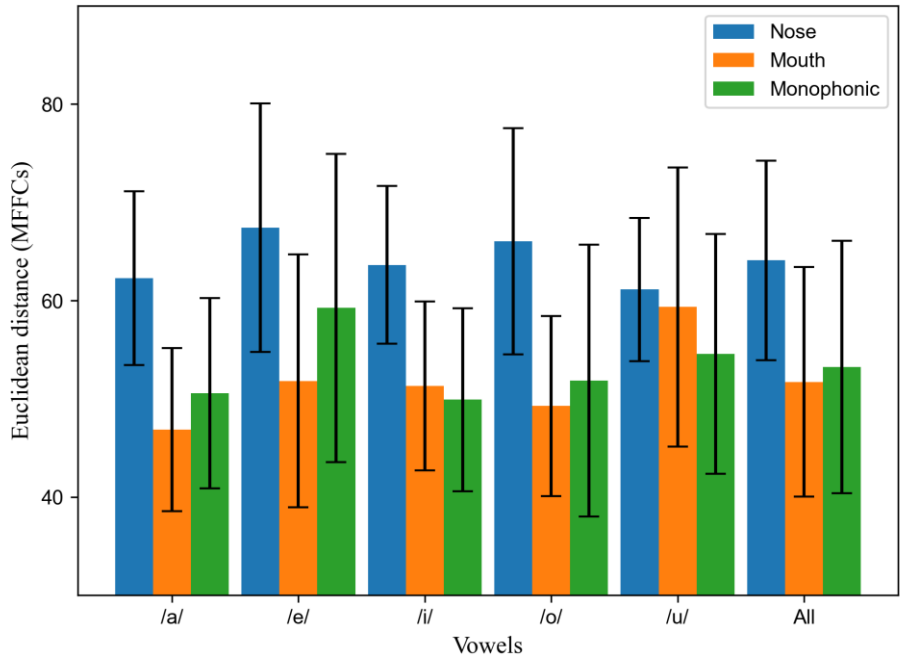

Figura 42. Distancias euclídea entre los MFFC de las vocales orales y nasales registradas con micrófonos nasales, bucales y monofónicos.

Para la segunda cuestión se analiza si la señal nasal permite una mejor clasificación de sonidos orales/nasales y sonidos vocales/consonantes empleando sistemas de clasificación automática. Para ello se crean múltiples modelos DNN empleando la arquitectura descrita por Mathad et al. (Mathad et al., 2021). Cada uno de los modelos se entrena con un tipo diferente de señal: monofónica, oral y nasal, empleando la base de datos descrita en la primera cuestión. Se calculan las matrices de confusión para cada modelo empleando cuatro categorías (NV, NC, OV, OC) y se evalúan tres condiciones diferenciadas: (1) considerando las cuatro clases fonéticas; (2) considerando las superclases oral y nasal; y (3) considerando las superclases vocal y consonante. En cuanto a los tipos de error, es relevante destacar que las pruebas de evaluación de la hipernasalidad suelen emplear sonidos orales y cuentan el número de errores de nasalización; esto significa que un número relativamente pequeño de falsos positivos (es decir, elementos que se clasifican erróneamente como nasales) puede dar lugar a errores de diagnóstico, y, por el contrario, si los falsos negativos representan un pequeño porcentaje, el impacto es menor. En consecuencia, se presta especial atención al porcentaje de ventanas clasificados erróneamente como nasales.

La exactitud global de clasificación resulta muy similar en las tres condiciones (en torno al 78–79%) cuando se consideran las cuatro clases, tal y como resume la Figura 43. Sin embargo, los patrones de error difieren claramente cuando se colapsa la clasificación a solo dos clases. En la clasificación oral/nasal, la precisión es mayor en la condición de nariz que en las de boca y señal monofónica: la proporción de vocales orales mal clasificadas como nasales es únicamente 0,07 en las DNN entrenadas con la señal de nariz, pero aproximadamente tres veces superior en las otras dos condiciones. Además, en los tres tipos de señal la exactitud media en vocales es inferior a la de consonantes (0,74 y 0,83 en nariz; 0,72 y 0,86 en boca; 0,73 y 0,84 en monofónica).

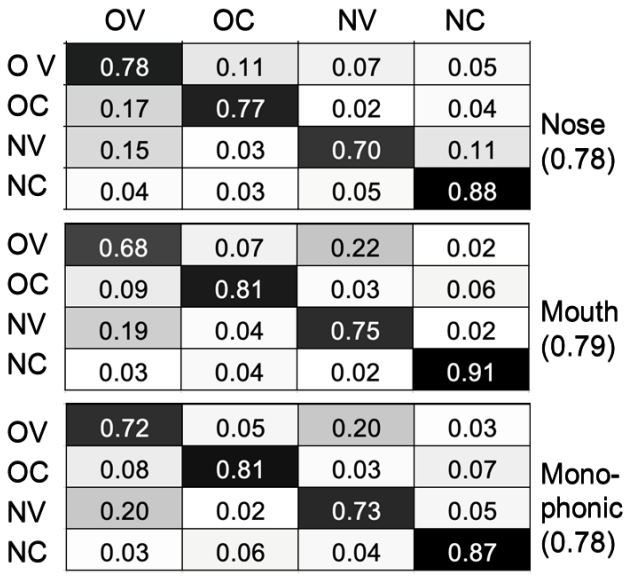

Figura 43. Matrices de confusión para las tres condiciones de grabación y precisión global (un color más oscuro significa un valor más alto).

Por el contrario, cuando se consideran solo las clases vocal/consonante, la precisión es algo mayor en las señales de boca y monofónica que en la de nariz. En conjunto, estos resultados indican que cada tipo de señal optimiza un aspecto diferente de la información fonética: la señal de nariz favorece la discriminación oral/nasal, mientras que las señales de boca y monofónica son ligeramente superiores para distinguir entre vocales y consonantes, como muestra la Figura 44.

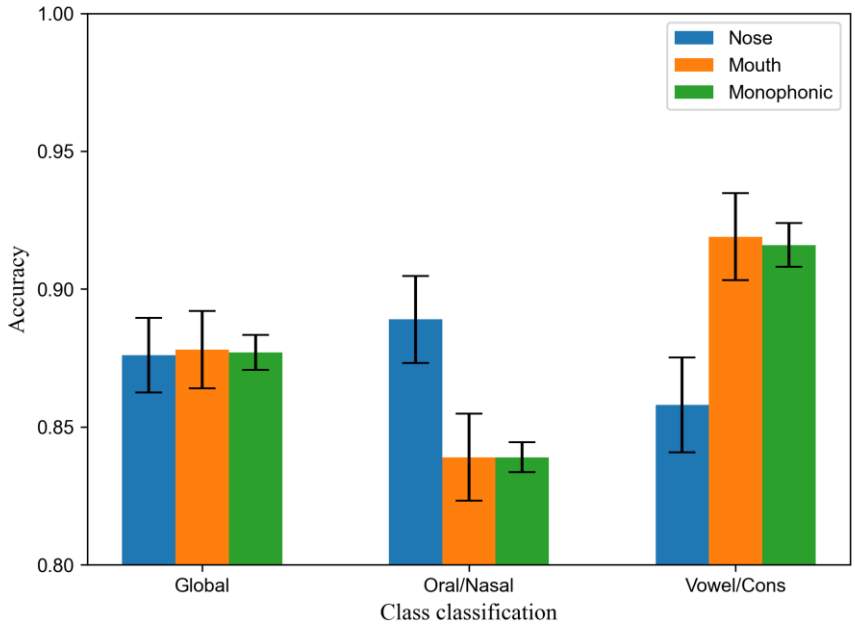

Figura 44. Precisión global al considerar cuatro clases (OC, OV, NC, NC), y precisión para dos clases (nasal frente a oral, y consonante frente a vocal).

Por último, se estudia la correlación entre las puntuaciones de nasalidad derivadas de los modelos de clasificación automática entrenados con diferentes tipos de señal, y las puntuaciones de hipernasalidad de la clasificación realizada por expertos. Para ello se amplía el conjunto de hablantes de la primera cuestión (N=55) y se añaden locutores grabados en la Universidad de Chile (N=16). Esto permite ampliar el tamaño de la base de datos a 3 horas de duración. Para la evaluación de pacientes con diagnóstico de hipernasalidad se generar una base de datos de test con un conjunto de niños control (N=34), y un conjunto de niños con diagnóstico de hipernasalidad (N=21), todos ellos grabados con el nasómetro.

Las grabaciones de los niños se obtienen como parte de una tarea de repetición auditiva que se realiza de forma rutinaria en el laboratorio para evaluar el habla hipernasalidad. La tarea consiste en repetir 12 palabras sin sonidos nasales: boca, pie, llave, dedo, gafas, silla, cuchara, sol, casa, pez, jaula y zapatos. Cada una de las palabras de la base de datos de pruebas fue puntuada por dos logopedas experimentados utilizando la siguiente escala:

- 0: Sin evidencia de habla hipernasal.
- 1: Evidencia de al menos una vocal nasalizada.
- 2: evidencia de al menos una consonante nasalizada.

En caso de desacuerdo entre los dos logopedas, un tercer experto puntúa el enunciado. A continuación, se obtiene una puntuación por niño promediando las puntuaciones de las 12 palabras.

Para este experimento, los modelos DNN entrenados con los datos de adultos se evalúan con la base de datos de prueba descrita anteriormente. Siguiendo (Mathad et al., 2021), se clasifica cada ventana como nasal u oral en función de las probabilidades posteriores de las pruebas DNN. Una ventana de audio se clasifica como nasal cuando la probabilidad posterior de ser una consonante nasal era al menos diez veces mayor que la probabilidad posterior de ser una consonante oral, o bien la probabilidad posterior de ser una vocal nasal era diez veces mayor que la probabilidad posterior de ser una vocal oral, como se muestra en la Ecuación (97):

$$\log_{10}(\mathrm{P(NC)/P(OC)}) > 1 \text{ o } \log_{10}(\mathrm{P(NV)/P(OV)}) > 1 \qquad (97)$$

La proporción nasal de un hablante se calcula como el número de ventanas nasales dividido por el número total de ventanas orales y nasales. Se calcula el coeficiente de correlación de Pearson entre el índice de nasalidad obtenido a partir de los modelos DNN y la puntuación de nasalidad producida por los logopedas.

A partir de los modelos entrenados con habla adulta se calcula, para cada niño, un índice de nasalidad basado en la proporción de ventanas clasificados como nasales. Como se observa en la Figura 45, este índice correlaciona con las puntuaciones perceptivas de los logopedas cuando se usaron DNNs entrenadas con la señal de nariz (r ≈ 0,83), mientras que la correlación fue baja y no significativa con modelos entrenados con la señal de boca (r ≈ 0,36). Cabe destacar que los modelos DNN otorgan puntuaciones relativamente altas a los hablantes sanos, algunos de ellos con puntuaciones similares o superiores a las de los pacientes.

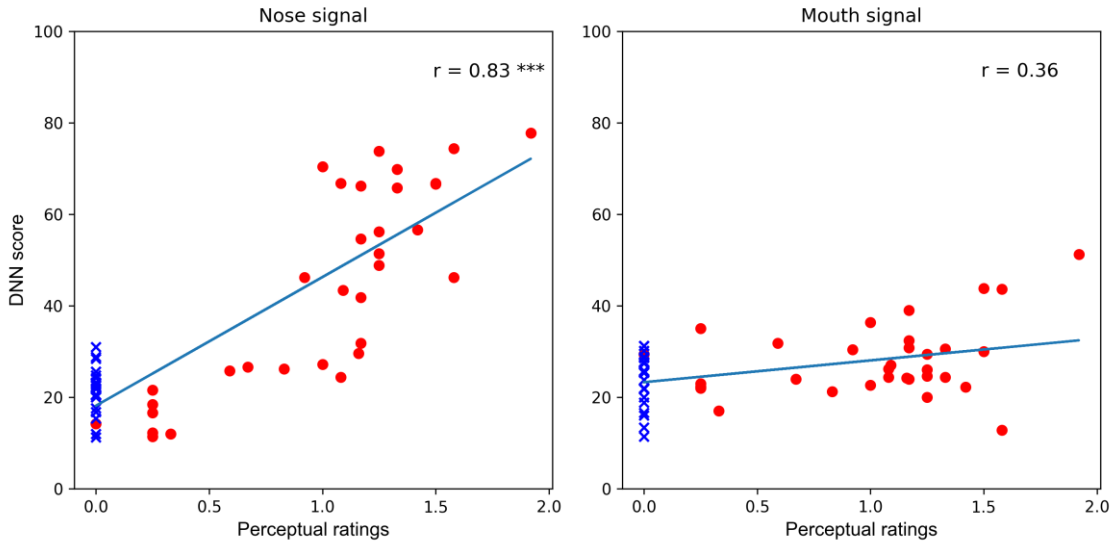

*Figura 45. Correlación de Pearson entre las puntuaciones perceptivas y las puntuaciones obtenidas con DNN con señales nasales (izquierda) y orales (derecha) (***: la señal nasal tiene una correlación r = 0,83 con p < 0,00001). Cruces azules: niños sanos. Puntos rojos: pacientes hipernasales.*

En conjunto, los resultados muestran que (i) la señal de nariz ofrece un contraste acústico mucho más nítido entre sonidos orales y nasales que la señal de boca o la monofónica, (ii) las DNN entrenadas con esta señal cometen menos errores críticos de clasificación oral-nasal, y (iii) la evaluación automática de la hipernasalidad basada exclusivamente en un micrófono de nariz puede alcanzar una correlación con los juicios de expertos comparable o superior a la obtenida con enfoques previos basados en señales monofónicas, o en medidas de nasometría tradicional. Esto abre la puerta al desarrollo de herramientas de evaluación de la hipernasalidad potencialmente más precisas y, al necesitar solo un canal nasal, más económicamente asequibles.

7.3 Computing nasalance with MFCCs and Convolutional Neural Networks

En este último trabajo, publicado en *PLOS One* el 31 de diciembre de 2024, se desarrolla una nueva aproximación al cálculo de la nasalancia para la clasificación automática de la hipernasalidad. La nasalancia es un biomarcador clínico que se emplea en el diagnóstico de la hipernasalidad, y se calcula como el cociente entre la energía acústica emitida por la nariz y la energía total emitida por la boca y la nariz (*eNasalance*). En este trabajo se propone el cálculo empleando CNN entrenadas con descriptores MFCC (*mfccNasalance*).

Para analizar las ventajas de la *mfccNasalance* se examina la precisión de los modelos CNN, definida como la correlación de Spearman entre la *mfccNasalance* de ese modelo y la nasalidad perceptual de los expertos humanos. Esta comparación se realiza en tres casos diferenciados: 1) cuando los datos de entrenamiento y de prueba proceden del mismo dialecto, o de dialectos diferentes; 2) con datos de prueba que difieren en dinamicidad (por ejemplo, sílabas diadococinéticas producidas rápidamente frente a palabras cortas); y 3) utilizando múltiples configuraciones de CNN.

Se crean tres bases de datos con grabaciones de hablantes hispanos procedentes de dialectos diferentes: español de España, español de Costa Rica y español de Chile. Las grabaciones se realizan a mujeres jóvenes y sanas mientras leen cuatro textos empleando un nasómetro (icSpeech, Rose Medical Solutions Ltd., Canterbury, Reino Unido). Tres de estos cuatro textos incluyen una representación equilibrada del inventario de fonemas del español (Bruyninckx et al., 1994; Martínez-Celdrán et al., 2003; Ortega et al., 2000). El cuarto texto se escribe como parte de este estudio e incluye casos de las tres consonantes nasales en varios contextos. Se registran 50 hablantes en España, 42 en Costa Rica y 32 en Chile.

La base de datos de test incluye muestras de habla de 38 niños con habla hipernasal y de 11 niños sanos. Algunos de los pacientes se graban en dos ocasiones (N=5) o en tres ocasiones (N=2), por lo que el número total de muestras de habla del grupo hipernasalidad fue de 47. Todos los participantes hablan español como lengua materna. Cada participante produce una serie de enunciados como parte de una tarea de repetición que se utiliza habitualmente en el laboratorio para evaluar niños con

trastornos del habla. Para el presente estudio se utiliza una prueba que incluye palabras cortas (-dinámicas), frases (+dinámicas) y sílabas diadococinéticas (+dinámicas), tal y como se muestra en la Tabla 4.

Tabla 4. Locuciones para test.

Tipo	Locuciones
Repetición silábica diadococinéticas (+ dinámica)	*papapa…, pipipi…, tatata…, tititi…, kakaka…, kikiki…*
Palabras (-dinámico)	*boca, pie, llave, dedo, dedo, gafas, silla, sol, casa, pez*
Frases (+dinámica)	*A David le duele el dedo* *Al gato de Agatha le gusta el yogur* *Uy, hay algo ahí* *Si me llevo la llave* *Susi sale sola* *Fali fue a la feria* *Los zapatos de Cecilia* *La jirafa de Jesús* *Toda tu taza de té* *Papá puede pelar a Pili* *Quique coge el papel de calco*

Las señales de audio se dividen en ventanas de 250 ms con un solapamiento de 150 ms. Para cada ventana, se calculan 39 MFCC, con una longitud de 25 ms y un solapamiento de 10 ms. Esto da como resultado una matriz tridimensional 2 (canales) × 39 MFCCs × 26 marcas temporales. Estas matrices (o imágenes de dos canales) actúan como datos de entrada de la red CNN. La red CNN consta de dos capas formadas por: una capa de convolución, batch normalization, y pooling. La salida está formada por dos capas ocultas de 128 nodos y activación ReLU que clasifica las muestras de entrada como oral y nasal. Como parte de este estudio, se prueban múltiples kernels en la capa de convolución de tamaño i × j ($1 \leq i \leq 8$, $1 \leq j \leq 8$).

Dado que las imágenes de entrada MFCC utilizadas en este estudio representan el tiempo en el eje horizontal y los valores de MFCC en el eje vertical, las dimensiones del kernel empleando en la red pueden agruparse de la siguiente forma, como se muestra en la Figura 46:

- Espectral: (2 × 1), (3 × 1), (4 × 1), etc.
- Temporales: (1 × 2), (1 × 3), (1 × 4), etc.
- Espectral-temporal: (2 × 2), (3 × 3), (4 × 4), (2 × 4), (4 × 2), etc.

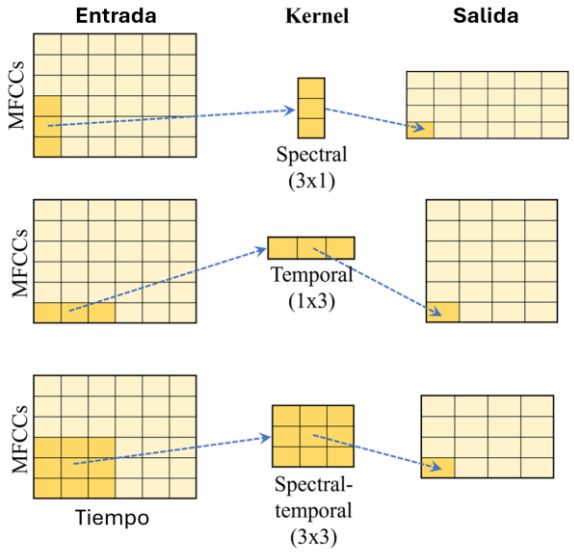

Figura 46. Relación entre la forma del kernel y la información fonética.

En la configuración por defecto, se utiliza el mismo núcleo para ambas capas de convolución. Sin embargo, se explora otra combinación utilizando un kernel 1 × 1 *pointwise* en la primera capa, combinado con las diferentes formas de kernel en la segunda capa.

Cada una de las secuencias de datos de entrada se clasifica como nasal u oral. Se anotan las muestras de habla del locutor utilizando Praat y se obtiene una transcripción fonética utilizando la herramienta Montreal Forced Alignment. En esta fase se clasificaron como nasales dos grupos de sonidos: 1) las tres consonantes nasales (/ m /, / n / y / ɲ /); y 2) una sección de las vocales que están en contacto con la consonante nasal (por ejemplo, la parte final de «a» en / an.tes /). La duración de esta sección varía con la duración de la vocal. En este estudio, se asume que toda la vocal estaba nasalizada si era inferior a 60 ms, el 50% si estaba entre 60 y 90 ms, y el 30% si era superior a 90 ms. Estos límites se basan en experimentos realizados en las primeras fases de diseño de este estudio. En el tercer paso, la señal de voz se segmenta utilizando una ventana móvil de 250 ms de longitud y 150 ms de solapamiento. A continuación, se calcula la duración acumulada de las ventanas de audio anotados como nasales y se dividió por la duración total. Si el valor resultante es superior a 0,30 (es decir, al menos 75 ms), la ventana se clasifica como nasal; en caso contrario, se clasifica como oral.

Una vez entrenado el modelo, se emplean nuevas imágenes MFCC generadas a partir de muestras de habla de 250 ms para testear. LA red CNN produce la probabilidad posterior de ser nasal para cada imagen de entrada, con valores entre 0 y 1. La *mfccNasalance* de un hablante se calcula promediando las probabilidades posteriores de los fragmentos de habla de ese hablante. Una vez se obtienen todas las puntuaciones *mfccNasalance* de los locutores, estas se transforman en una escala de 0 a 3 para proporcionar un significado clínico, como la escala perceptual. Para transformar las probabilidades en una escala de 4 niveles, se determinan los límites entre las categorías. Como ya se sabe

cuántos niños están clasificados perceptualmente como orales o nasales, y específicamente como orales, leves, moderados o gravemente nasales, se supone que la misma distribución se aplicaría a las mediciones de la nasalidad y se realiza una transformación de escala manteniendo la misma distribución de grupos. Una limitación de este enfoque es que puede haberse producido artificialmente un sesgo. Sin embargo, como sólo se analiza la precisión y se utiliza el mismo enfoque para *eNasalance* y *mfccNasalance*, se entiende que es válido para este estudio.

Los resultados ofrecidos por el proceso de clasificación cuando se emplea el mismo dialecto en el proceso de entrenamiento y testeo se pueden observar en la Figura 47, donde la figura de la izquierda muestra los resultados para todas las configuraciones CNN, mientras que la figura de la derecha muestra los resultados para las CNN utilizando la configuración óptima. Para sílabas y frases, prácticamente todas las configuraciones de CNN superan a *eNasalance*; en palabras, *mfccNasalance* también suele ofrecer correlaciones superiores, aunque la ventaja es algo menor. Es decir, incluso sin ajustar finamente la arquitectura, el nuevo índice se aproxima más a los juicios humanos la medida de nasalancia clásica. El análisis ANOVA de dos factores revela que la inclusión de una convolución 1×1 *pointwise* en la primera capa de la CNN mejora significativamente la precisión en enunciados dinámicos (sílabas y frases), mientras que en palabras no aporta ventajas significativas. En cambio, en palabras el factor crítico es la forma del kernel de la segunda capa: los kernels temporales proporcionan correlaciones significativamente mayores que los espectrales y, a su vez, estos superan a los espectro-temporales. La combinación considerada óptima en España es: convolución 1×1 para sílabas y frases, y kernels temporales sin 1×1 para palabras, lo que reduce la variabilidad entre modelos y mantiene altas correlaciones con la valoración perceptiva.

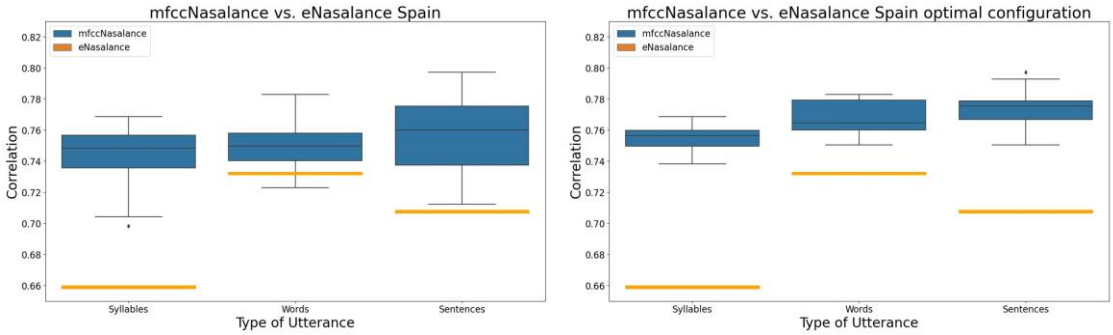

Figura 47. Correlación entre e-Nasalance y puntuaciones perceptivas (rectángulo naranja), y mfccNasalance y puntuaciones perceptivas (azul) en la misma condición dialectal (España-España).

Cuando se emplea distinto dialecto en el proceso de entrenamiento y testeo, se obtiene los resultados que se muestran en la Figura 48, donde la parte superior corresponde a los modelos entrenados en Costa Rica, y la parte inferior a los de Chile. Las figuras de la izquierda muestran los resultados para todas las configuraciones de CNN, mientras que las figuras de la derecha muestran los resultados de las CNN que utilizan la configuración óptima. Se observa una caída general de las correlaciones respecto a la condición de mismo dialecto. Esta degradación es especialmente marcada cuando el entrenamiento

se hace con el dialecto costarricense, conocido por su mayor tendencia a la nasalización: para aproximadamente la mitad de las configuraciones, las correlaciones de *mfccNasalance* con la percepción en palabras son iguales o inferiores a las de *eNasalance*. En cambio, los modelos entrenados con habla chilena se comportan de forma más similar a los de España, con diferencias solo en las sílabas. Los análisis de varianza indican además que la configuración óptima de kernels depende tanto del dialecto de entrenamiento como del tipo de enunciado: por ejemplo, en Costa Rica la convolución 1×1 vuelve a ser beneficiosa para sílabas y frases, mientras que en Chile las mejores correlaciones para sílabas y frases se obtienen con kernels espectrales y sin necesidad de 1×1.

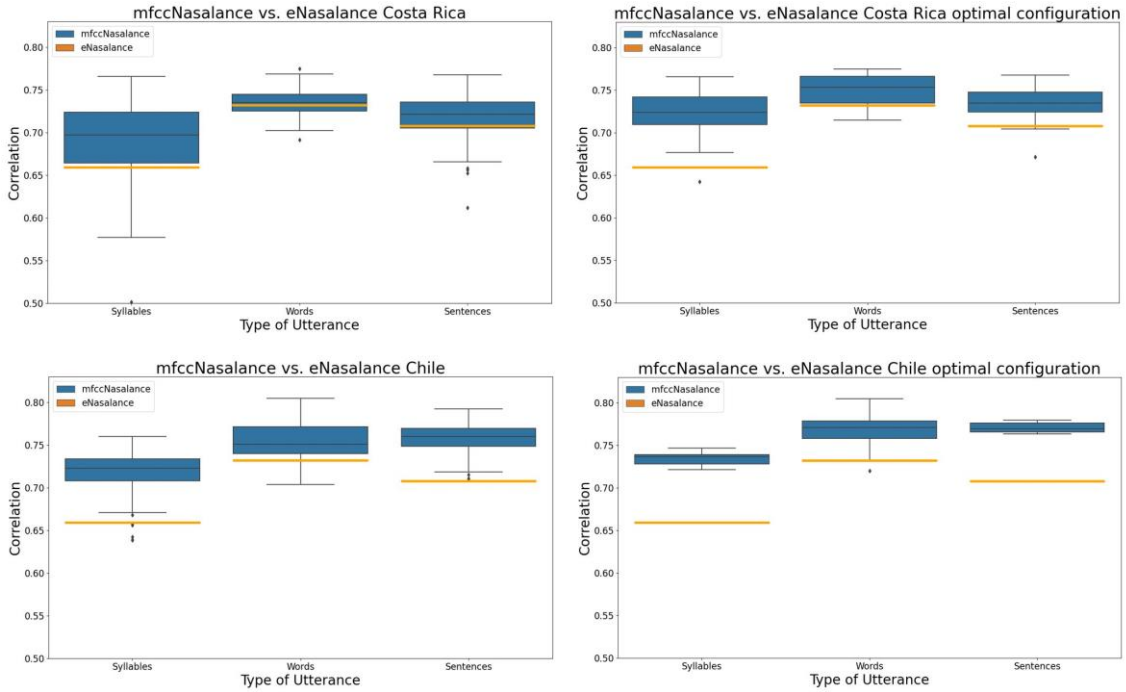

Figura 48. Correlación entre e-Nasalance y puntuaciones perceptivas (rectángulo naranja), y mfccNasalance y puntuaciones perceptivas (azul) en la condición de dialectos diferentes.

En conjunto, los resultados demuestran que (i) la correlación con las valoraciones perceptivas es sistemáticamente mejor con *mfccNasalance* que cuando se emplea la medida tradicional de nasalidad *eNasalance*, especialmente cuando el modelo se entrena y evalúa en el mismo dialecto; (ii) la arquitectura CNN óptima es específica del tipo de enunciado, siendo crucial la convolución 1×1 para enunciados dinámicos y la forma del kernel (con una preferencia por los de tipo temporal) para palabras; y (iii) la variación dialectal afecta de forma apreciable a la precisión, de modo que los mejores resultados se obtienen con modelos entrenados en el mismo dialecto que los pacientes, mientras que el uso cruzado entre dialectos reduce la ventaja de *mfccNasalance* sobre *eNasalance*. Este efecto es más marcado en dialectos que presentan una mayor diferencia en la nasalización, como es el caso de España-Costa Rica. Estos hallazgos apoyan el uso de *mfccNasalance* como alternativa flexible y más cercana a la percepción humana, pero apuntan también a la necesidad de diseñar modelos y configuraciones adaptados a dialectos y tareas específicos.

8. Discusión de los resultados

En este capítulo se realiza una discusión global de los resultados presentados anteriormente y desarrollados en detalle en las tres contribuciones que forman parte integrante de la tesis. Esta tesis trata, en esencia, tres temas complementarios: 1) qué tipo de enunciados o locuciones es más adecuado para detectar la hipernasalidad de forma automática; 2) qué tipo de señal acústica contiene la información más pertinente para esta tarea, y 3) cómo diseñar medidas y modelos de aprendizaje profundo que permitan un diagnóstico automático lo más cercano posible a la percepción de los expertos.

El primer eje de discusión de la tesis gira en torno al tipo de enunciado y a su papel en la detección automática de la hipernasalidad. Tradicionalmente, la mayoría de los trabajos técnicos han utilizado casi exclusivamente vocales sostenidas, debido a que su estabilidad espectral facilita el análisis acústico. Esto contrasta con las recomendaciones clínicas que insisten en evaluar al paciente con una variedad de fonemas y tipos de habla: sílabas repetidas, palabras, frases con distintas estructuras consonánticas, e incluso habla espontánea. El objetivo principal del primer artículo de la tesis es comprobar hasta qué punto otras tareas distintas de la vocal sostenida (conteo, secuencias silábicas, palabras, frases y consonantes sostenidas) permiten detectar la hipernasalidad de forma automática.

A partir de las grabaciones recogidas con una app móvil en tres países hispanohablantes, se entrenan diferentes clasificadores con 44 enunciados y se compara su rendimiento. Los resultados muestran una variabilidad muy marcada según el tipo de enunciado utilizado. Las tasas de acierto de los clasificadores que emplean un único enunciado oscilan entre el 46 % y el 81 %, con un grupo claramente destacado de enunciados. Los mejores resultados (precisión > 75 %) se obtienen con cuatro secuencias silábicas, dos palabras y una frase: en particular, varias repeticiones de sílabas con / p, t, k / combinadas con / a, i /, la palabra *dedo*, la palabra *pez* y la frase *"A David le duele el dedo"*. En el siguiente peldaño (70–75 %) se sitúan las otras secuencias silábicas, algunas frases adicionales y la fricativa sostenida / f /. En cambio, el conteo del 1 al 10 y buena parte de las palabras presentan precisiones en el entorno del 50–60 %.

Las secuencias /pa ta ka/ son una forma controlada de tensionar el sistema velofaríngeo, obligando a coordinar repetidamente oclusión oral y apertura de la vocal, de manera rápida y rítmica. El hecho de que los clasificadores logren altas tasas de acierto con estas secuencias, mientras que la vocal / a / sostenida arroja resultados claramente peores, sugiere que la hipernasalidad se manifiesta con más claridad en tareas dinámicas exigentes que en producciones mantenidas y relativamente cómodas. Existen dos explicaciones complementarias: por un lado, el esfuerzo motor extra puede dificultar el control del velo, incrementando la probabilidad de fuga nasal; por otro, los pacientes pueden haber desarrollado mecanismos compensatorios sutiles (por ejemplo, pequeños cambios en el lugar o modo de articulación) que modifican el espectro de las sílabas de

un modo que pasa desapercibido al oído humano pero no a los clasificadores automáticos.

El conteo 1-10 produce peores resultados que varias frases diseñadas con repetición sistemática de ciertas consonantes. Una posible explicación para este comportamiento es que muchos niños han practicado los números de forma intensiva y pueden haber aprendido a producir esa secuencia concreta con menos rasgos de hipernasalidad, incluso cuando mantienen problemas en otros contextos. En cambio, las frases del protocolo incluyen múltiples ocurrencias de la misma consonante en distintos entornos fonéticos, lo que multiplica las oportunidades de que se produzca un sonido hipernasal, y hace que sea identificable aun cuando se realiza un promediado de ventanas. Es decir, no solo es importante qué fonemas se usan, sino cómo se distribuyen dentro del enunciado.

Un caso especialmente interesante es el de la fricativa sostenida / f /. Desde el punto de vista clínico se utiliza sobre todo para detectar escapes nasal audibles, más que para evaluar la hipernasalidad. Sin embargo, en el estudio se observa que / f / es uno de los enunciados con mejor rendimiento, tanto aislado como combinado con otros, y que los pacientes y controles difieren en varios coeficientes MFCC asociados a este sonido. Esto indica que, más allá del ruido aerodinámico evidente, pueden existir diferencias espectrales consistentes en la producción de / f / entre ambos grupos. Este comportamiento tiene dos posibles orígenes: o bien la propia presencia de resonancia nasal está modificando la fricación, o bien los pacientes han ajustado su manera de articular / f / para compensar la dificultad a la hora de generar presión oral, y esos ajustes quedan reflejados en las componentes espectrales de la señal.

Otro resultado clave es que no todos los pacientes nasalizan los mismos enunciados. Para cada paciente, existe un patrón heterogéneo con respecto a qué enunciados presentan al menos una consonante nasalizada según los jueces expertos, esto es, ningún niño nasaliza todas las muestras, y ningún enunciado es nasalizado por todos los pacientes. Esto implica que, aunque dos pacientes tengan el mismo diagnóstico de hipernasalidad, los contextos fonéticos donde esta se hace más evidente pueden ser muy distintos. Desde el punto de vista de la clasificación automática, esto explica por qué emplear un único enunciados arroja peores resultados que cuando se combinan varios enunciados.

En este contexto, se estudia qué ocurre cuando se entrena cada clasificador con varios enunciados en lugar de uno solo. En general, pasar de uno a dos o tres enunciados mejora la precisión, pero añadir más tiende a degradarla. Este resultados se puede atribuir a las limitaciones del proceso de selección de características acústicas: al no considerar la correlación entre rasgos, es fácil que se seleccionen características redundantes cuando aumenta el número de enunciados, desplazando a otras potencialmente más útiles. Por ello, se plantea una combinación de clasificadores, mediante el índice *HN Score*. Esta estrategia tiene varias ventajas: permite explotar la complementariedad entre tareas; realiza un reconocimiento explícito de la hipernasalidad como un fenómeno gradual, no binario: los valores cercanos al 50 % se

interpretan como casos ambiguos, análogos a las situaciones clínicas donde incluso los expertos dudan; y al seleccionar un conjunto reducido de los mejores enunciados, se consigue una separación muy clara entre pacientes y controles, con muy pocos hablantes en la zona intermedia.

El segundo eje de discusión se centra en el tipo de señal acústica (oral, nasal, o monofónica) utilizada para entrenar y evaluar modelos de detección automática de la hipernasalidad, abordando esta cuestión desde tres perspectivas complementarias: acústica, fonética y clínica. Este es el objetivo principal del segundo artículo de la tesis.

Desde el punto de vista acústico, se realiza una comparación directa entre vocales orales y sus contrapartes nasalizadas. El análisis de las distancias euclídeas entre los coeficientes MFCC de ambas clases de vocales muestra que: para / a, e, i, o / la distancia es sistemáticamente mayor en la señal de nariz, mientras que en las señales de boca y monofónica las distancias son muy similares y claramente menores; solo / u / se comporta de forma distinta, con valores parecidos en las tres condiciones. Esta diferencia se debe principalmente a que cuando se analiza la señal procedente del micrófono de nariz, existe un contraste muy nítido de energía entre sonidos orales (muy débiles) y sonidos nasales (mucho más fuertes), mientras que para el micrófono de boca ambos tipos de sonido son claramente audibles y, por tanto, sus espectros se parecen mucho más. En la señal monofónica, la señal oral domina sobre la nasal, de modo que las características acústicas resultantes son prácticamente las de la señal oral.

Ese enmascaramiento se hace evidente cuando se realiza un análisis fonético mediante DNN entrenadas con cada tipo de señal. Cuando se consideran las cuatro clases NV, NC, OV y OC, las exactitudes medias son similares para los tres tipos de señal (en torno al 78–79 %), lo que induce a pensar que el tipo de señal es irrelevante. Sin embargo, un análisis de la clasificación empleando las superclases oral/nasal y vocal/consonante, muestra que en el primero la señal de la nariz ofrece la mayor exactitud; mientras que en el último las señales de la boca y monofónica ofrecen resultados superiores en la clasificación. El resultado más crítico de cara al diagnóstico de la hipernasalidad es el sesgo en los errores de tipo oral-nasal. En el caso de las DNN entrenadas con la señal de nariz, la proporción de vocales orales mal clasificadas como vocales nasales es del 0,07, mientras que empleando la señal de la boca y monofónica este error es aproximadamente tres veces mayor.

En un contexto clínico, donde habitualmente se trabaja con materiales orales y se busca identificar cuando aparece un sonido nasal que no debería, este tipo de falso positivo es especialmente importante, ya que puede llevar a etiquetar como hipernasales a hablantes sanos o a pacientes ya rehabilitados. Por el contrario, un número moderado de falsos negativos (sonidos nasales etiquetados como orales) tiene un impacto menor en la decisión clínica global. Desde esta perspectiva, la ventaja de la señal de la nariz es clara: conserva mejor el contraste oral/nasal y, en consecuencia, reduce el riesgo de falsos positivos.

Cabe destacar el hecho de que, en las tres condiciones, las consonantes se clasifican mejor que las vocales. Este resultado se debe principalmente al proceso de segmentación de la nasalización vocálica, ya que se ha empleado un criterio relativamente simple como es el porcentaje fijo de la duración de la vocal en contacto con una consonante nasal, cuando la literatura muestra que la extensión real de la nasalización depende de múltiples factores, como son: dialecto, hablante, acento, o posición en la palabra.

Es probable que una parte de los segmentados de audio etiquetados como NV no lo fueran realmente, lo que degrada la capacidad de clasificación específica de las vocales. Aun así, el patrón comparativo se mantiene: la señal de la nariz sigue siendo la que mejor separa las clases oral/nasal, mientras que las señales de boca y monofónica favorecen la distinción vocal/consonante. Este comportamiento se interpreta como una especialización funcional de los canales: el nasal codifica de forma óptima la dimensión oral/nasal, mientras que el oral concentra la mayoría de la información necesaria para otras distinciones fonéticas, como pueden ser: modo, lugar, sonoridad, o estructura silábica. El habla transporta simultáneamente mucha información fonética, por lo que resulta razonable pensar que el sistema comunicativo humano ha evolucionado para priorizar las dimensiones más críticas para el reconocimiento del mensaje, como la segmentación vocal/consonante y la identificación de consonantes, por encima de otras como el contraste oral/nasal, que tiene menor peso funcional en la mayoría de las lenguas. Dado que la mayor parte de la información relevante para esta comunicación prioritaria se concentran en la señal de la boca, es lógico que esta domine la mezcla monofónica, aunque ello suponga enmascarar parte de la información de nasalidad que transporta el canal nasal.

A continuación se comparan directamente DNN entrenadas con señal de nariz y con señal de boca para estimar el grado de hipernasalidad en un conjunto de niños, de manera que se pueda estimar hasta qué punto este reparto de información entre canales se traduce en diferencias reales en la evaluación automática de la hipernasalidad. En ambos casos, se calcula un índice de nasalidad como porcentaje de las ventanas clasificadas como nasales y se correlaciona con las puntuaciones de logopedas expertos. Los resultados muestran unas correlaciones altas y significativas con la señal de nariz ($r \approx 0{,}82$), y bajas y no significativas con la señal de boca ($r \approx 0{,}36$), tanto si se considera solo a los pacientes como si se incluyen también los niños sanos. Cuando se entrena únicamente con la señal de la boca, varios niños sanos reciben puntuaciones automáticas de nasalidad en el mismo rango que algunos de los pacientes, lo que refleja un problema de falsos positivos. Cuando se emplea la señal de nariz, en cambio, se observa una separación más clara entre controles y pacientes.

Dado estos resultados, resulta relevante compararlos con estudios previos que utilizan señal monofónica o nasómetro. El modelo de Mathad et al. (Mathad et al., 2021), que utiliza la misma arquitectura de DNN pero entrenada con aproximadamente 100 horas de habla monofónica, presenta una correlación con la percepción de los especialistas ligeramente inferior a los resultados (0,80 frente a 0,83). Esto sugiere que elegir un canal

acústicamente más informativo, como puede ser la nariz, permite compensar en parte la falta de grandes cantidades de datos. Por otra parte, al analizar los estudios previos que emplean nasometría, se observa que solo uno de ellos obtiene una correlación más alta que el presente trabajo (r = 0,88), mientras que la mayoría se sitúa claramente por debajo, (r = 0,55-0,74). Estos resultados apuntan a la idea de que el enfoque basado en redes neuronales y señal nasal es, como mínimo, comparable con la nasometría clásica y en algunos casos más consistente.

El tercer eje de discusión de la tesis se centra en el desarrollo de una nueva aproximación al cálculo de la nasalancia para la clasificación automática de la hipernasalidad. Este es el objetivo principal del tercer artículo de la tesis. Para ello se plantea *mfccNasalance* como alternativa a la medida de nasalancia clásica *eNasalance*, combinando descriptores MFCCs de los dos canales del nasómetro con redes neuronales convolucionales. *mfccNasalance* se evalúa en tres condiciones diferenciadas: 1) cuando los datos de entrenamiento y prueba provienen del mismo dialecto o de dialectos diferentes; 2) con datos de prueba que difieren en dinamismo; y 3) utilizando múltiples configuraciones de CNN.

Cuando se evalúa *mfccNasalance* empleando el mismo dialecto, la correlación con las puntuaciones perceptivas es sistemáticamente mayor que la obtenida con *eNasalance* para prácticamente todas las configuraciones de CNN consideradas. Independientemente de la forma del kernel o de la presencia de convolución 1×1, los índices *mfccNasalance* se aproximan mejor al juicio de los expertos que la medida tradicional de nasalancia. Este resultado se debe a que *eNasalance* explota solo una franja muy estrecha del espectro (una banda de 300 Hz alrededor de 600 Hz), mientras que la literatura fonética muestra que la nasalización vocálica puede afectar al espectro hasta 3 kHz o más, y que los oyentes utilizan información distribuida en varias bandas, no solo energía global en un rango fijo. *mfccNasalance*, en cambio, emplea un conjunto de filtros para alimentar una CNN que combina esa información de forma no lineal, de un modo conceptualmente más cercano a cómo el sistema auditivo humano explota las componentes espectrales.

La comparación con la propuesta de Mathad et al. (Mathad et al., 2021) resulta especialmente relevante. Ese trabajo también emplea MFCC y redes neuronales profundas para derivar un índice de nasalidad, pero se basa en unas 100 horas de habla monofónica anotada fonéticamente con gran detalle. El enfoque adoptado ofrece algunas ventajas significativas: 1) los corpus de entrenamiento son mucho más pequeños (2–4 h por dialecto), lo que reduce drásticamente el coste de creación de bases de datos específicas; 2) no es necesario un alineado fonético tan preciso, porque el etiquetado se hace sobre ventanas de 250 ms en función del porcentaje de tiempo nasalizado, evitando la revisión manual exhaustiva; y 3) el propio procedimiento de etiquetado (orales vs nasales) está directamente ligado al objetivo clínico. En conjunto, esto hace que adaptar mfccNasalance a un nuevo idioma o dialecto sea, en principio, mucho más factible que replicar el esquema basado en grandes corpus con una anotación detallada.

En cuanto al diseño de la CNN, se explora de forma sistemáticamente distintas combinaciones de kernels, diferenciando entre kernels espectrales, temporales y espectro-temporales, y la presencia o ausencia de una convolución pointwise 1×1 en la primera capa. La hipótesis inicial es que las tareas con habla dinámica (sílabas diadococinéticas y frases) se benefician de kernels temporales, mientras que los enunciados estáticos (palabras) pueden describirse mejor con kernels espectrales. Sin embargo, los resultados matizan esta intuición: el factor que más influye en las sílabas y las frases no es tanto la forma del kernel como la presencia de la convolución 1×1 en la primera capa, que mejora de forma significativa la correlación con la percepción de los expertos en estos enunciados dinámicos. En cambio, en las palabras esta convolución no aporta beneficios, y lo que marca la diferencia es la forma del kernel de la segunda capa, con clara superioridad de los kernels temporales sobre los espectrales y espectro-temporales.

Desde el punto de vista de la arquitectura, la convolución 1×1 se emplea para combinar canales (aquí, boca y nariz) sin mezclar aún la dimensión tiempo–frecuencia. Esto permite a la red aprender representaciones que integran la información de ambos canales en cada instante, antes de aplicar kernels más grandes. En enunciados muy heterogéneos —como sílabas rápidas o frases con múltiples transiciones—, esta capacidad de abstracción y fusión de canales ayuda a filtrar el ruido y a destacar las componentes de nasalidad realmente informativas, aumentando la robustez del modelo. En cambio, en palabras relativamente cortas y estables, la variabilidad es menor y ese mecanismo de integración aporta menos beneficios observables.

El hecho de que las palabras bisílabas españolas se beneficien sobre todo de kernels temporales sugiere además que, incluso en enunciados menos dinámicos, la nasalidad se codifica de forma dinámica: por ejemplo, a través de la transición de los formantes, o de cambios breves en la relación entre energía nasal y oral dentro de la palabra. Este resultado contrasta con el único trabajo previo que había usado CNN para clasificar la hipernasalidad (Wang, Tang, et al., 2019), donde se concluye que los kernels espectrales son los más adecuados. Existen dos razones plausibles para esta discrepancia: 1) en ese estudio la señal es monofónica, por lo que la información nasal puede estar enmascarada por la componente oral, y 2) se trabaja con palabras monosilábicas chinas en aislamiento que son, probablemente, menos dinámicas que las bisílabas españolas. En ese contexto, tiene sentido que las secciones más estables presenten mayor importancia, y que los kernels espectrales resulten más eficaces.

Cabe destacar la influencia del dialecto en la precisión de *mfccNasalance*. En términos globales, los resultados muestran que los modelos entrenados y probados en el mismo dialecto (España-España) alcanzan correlaciones más altas con la percepción de los expertos que los modelos que se evalúan con un dialecto diferente al del proceso de entrenamiento (España-Costa Rica, o España-Chile). La degradación es particularmente marcada en la situación España–Costa Rica, y más moderada en España–Chile; el español

costarricense tiende a ser claramente más nasal, mientras que el español europeo y el chileno son globalmente más orales y similares entre sí.

Por un lado, las manifestaciones acústicas de la nasalidad, es decir, qué formantes se desplazan, cuánto se eleva la energía nasal, cómo se extiende la nasalización a las vocales vecinas, depende en gran parte del dialecto. Un modelo entrenado en un dialecto poco nasal puede infravalorar la nasalidad normal de un dialecto más nasal o, a la inversa, un modelo entrenado en un dialecto fuertemente nasalizado puede sobreestimar la nasalidad cuando se aplica a un dialecto más oral. Por otro lado, las propias puntuaciones perceptivas de los logopedas pueden estar moduladas por el dialecto: lo que un clínico de un entorno muy nasal interpreta como leve puede ser percibido como moderado por un clínico de otra zona. En conjunto, esto hace poco viable el uso de un único modelo universal que pueda ser entrenado en un dialecto y aplicable al resto de variedades sin ajustes a las características del dialecto local.

Todos estos resultados apuntan a una misma conclusión: *mfccNasalance* es una medida flexible y conceptualmente sólida, capaz de preservar las ventajas clínicas de la nasalancia (un índice continuo e interpretable en una escala) pero integrando al mismo tiempo la información disponible en las componentes espectral y temporal de la señal y las posibilidades que ofrecen las redes CNN. Para explotar todo su potencial es necesario adaptar la arquitectura del modelo al tipo de enunciado y al dialecto, buscando un sistema capaz de seleccionar, para cada fragmento de habla y cada contexto lingüístico, la combinación más adecuada de kernels y datos de entrenamiento.

9. Conclusiones

Este capítulo presenta un resumen de la investigación realizada durante esta tesis doctoral, cuyo objetivo principal es el estudio de la detección automática de la hipernasalidad en el habla patológica, empleando para ellos diferentes enfoques. Para tal fin, este capítulo se divide en dos secciones. La primera, en la que se destacan las principales contribuciones de cada uno de los trabajos relacionados con esta tesis. En la segunda sección, se sugieren algunas líneas de trabajo futuro relacionadas con la investigación llevada a cabo.

9.1 Contribuciones

En primer lugar, se explora la viabilidad de detectar la hipernasalidad basándose en muestras de habla distintas de las vocales sostenidas. Se ha analizado qué tipos de enunciados facilitan la detección automática de la hipernasalidad mediante algoritmos entrenados con características acústicas. Se determina que las secuencias repetitivas de sílabas, y en menor grado algunas palabras y frases completas, proporcionan los mejores resultados. La combinación de múltiples enunciados y el uso combinado de varios clasificadores mejora la precisión. Estos resultados confirman la posibilidad de desarrollar herramientas automáticas de clasificación de la hipernasalidad basadas en el análisis acústico del habla espontánea que, además, pueden funcionar en dispositivos de uso común como los teléfonos móviles, superando así la limitación de las metodologías tradicionales enfocadas principalmente en vocales sostenidas.

En segundo lugar, se examina si las señales provenientes exclusivamente de la nariz o de la boca pueden aumentar la precisión de la evaluación de la hipernasalidad frente a la evaluación cuando se utilizan señales monofónicas, que es el resultado de combinar las dos anteriores en un único canal de audio. Para ello, se ha realizado un estudio comparativo entre distintos tipos de señales acústicas (oral, nasal y monofónica) para identificar cuál contiene más información relevante sobre la hipernasalidad. Los experimentos revelan que la señal nasal aporta información crucial sobre la hipernasalidad, logrando precisiones en torno al 85% en tareas de clasificación, y correlaciones altas (≈ 0,83) con la evaluación perceptual humana, frente a la baja correlación obtenida con la señal oral. Esto sugiere que la señal nasal por sí sola es suficiente para evaluar la hipernasalidad, lo que podría abaratar considerablemente el costo de dispositivos clínicos como el nasómetro, y hacer más accesible esta evaluación.

Finalmente, se propone un nuevo enfoque para calcular la nasalancia mediante el uso de una red CNN entrenada con coeficientes MFCC. Para ello se ha propuesto una nueva medida de la hipernasalidad denominada *mfccNasalance* basada en técnicas de aprendizaje profundo para estimar el grado de nasalidad en el habla. Los experimentos realizados indican que *mfccNasalance* logra una correlación más alta con evaluaciones

perceptuales que la medida tradicional de nasalancia, siempre que el entrenamiento y evaluación se realicen en el mismo dialecto. Además, ciertas configuraciones internas de la CNN afectan significativamente la precisión según el tipo de enunciado analizado. Sin embargo, la precisión se reduce notablemente cuando los modelos se aplican a dialectos diferentes al del entrenamiento, evidenciando diferencias dialectales importantes en las características acústicas de la hipernasalidad. Esto resalta la necesidad de caracterizar sistemáticamente estas variaciones dialectales para desarrollar modelos más generalizables y adaptables a diferentes contextos lingüísticos.

9.2 Líneas futuras

A pesar de los resultados obtenidos durante la realización del presente trabajo, existen nuevos problemas que merecen una atención particular, por lo que se propone diferentes líneas de investigación que pueden contribuir a profundizar, ampliar y validar los hallazgos alcanzados.

Se plantea como prioridad la extensión de las técnicas propuestas de diagnóstico automático de la hipernasalidad hacia otros efectos asociados a la disfunción velofaríngea, como pueden ser: escapes nasales, turbulencias, debilitamientos, error de articulación y errores compensatorios. Estos efectos, que se pueden dar de forma conjunta o aislada, son de gran relevancia clínica a la hora de realizar una evaluación de los pacientes. Por lo tanto, su detección automática permite dotar al especialista de una información más completa para el diagnóstico.

También resulta de especial interés trabajar en la generación de diferentes modelos de clasificación adaptativos, que ajusten dinámicamente el análisis espectral o temporal en función del tipo de fragmento de habla procesado, imitando la percepción humana del habla (Xu & Zheng, 2007). De este modo, se podría favorecer el procesamiento espectral o temporal del habla de forma flexible, en función de las características acústicas del fragmento analizado.

Además, se está trabajando en el desarrollo de una herramienta web que permita ofrecer un análisis automatizado de las pruebas diagnósticas realizadas por el especialista. Esta plataforma permite realizar evaluaciones de la nasalidad en tiempo real, y recopila de forma continua muestras de habla procedente de diferentes regiones hispanohablantes bajo la supervisión de logopedas. Gracias a ello, se puede obtener una gran cantidad de datos que permiten ajustar los modelos con mayor precisión, incorporando variaciones dialectales que han mostrado tener un impacto significativo en los patrones de nasalización.

Apéndice A. Copia de los trabajos

A.1 Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically?

Referencia bibliográfica

Moreno-Torres I, **Lozano A**, Nava E, Bermúdez-de-Alvear R. Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically? Applied Sciences. 2021; 11(19):8809.

Resumen

Automatic tools to detect hypernasality have been traditionally designed to analyze sustained vowels exclusively. This is in sharp contrast with clinical recommendations, which consider it necessary to use a variety of utterance types (e.g., repeated syllables, sustained sounds, sentences, etc.) This study explores the feasibility of detecting hypernasality automatically based on speech samples other than sustained vowels. The participants were 39 patients and 39 healthy controls. Six types of utterances were used: counting 1-to-10 and repetition of syllable sequences, sustained consonants, sustained vowel, words and sentences. The recordings were obtained, with the help of a mobile app, from Spain, Chile and Ecuador. Multiple acoustic features were computed from each utterance (e.g., MFCC, formant frequency) After a selection process, the best 20 features served to train different classification algorithms. Accuracy was the highest with syllable sequences and also with some words and sentences. Accuracy increased slightly by training the classifiers with between two and three utterances. However, the best results were obtained by combining the results of multiple classifiers. We conclude that protocols for automatic evaluation of hypernasality should include a variety of utterance types. It seems feasible to detect hypernasality automatically with mobile devices.

Artículo completo

Una copia del artículo completo está disponible gratuitamente en línea en https://doi.org/10.3390/app11198809.

A.2 Unmasking Nasality to Assess Hypernasality

Referencia bibliográfica

Moreno-Torres I, **Lozano A**, Bermúdez R, Pino J, Méndez MDG, Nava E. Unmasking Nasality to Assess Hypernasality. Applied Sciences. 2023; 13(23):12606.

Resumen

Automatic evaluation of hypernasality has been traditionally computed using monophonic signals (i.e., combining nose and mouth signals). Here, this study aimed to examine if nose signals serve to increase the accuracy of hypernasality evaluation. Using a conventional microphone and a Nasometer, we recorded monophonic, mouth, and nose signals. Three main analyses were performed: (1) comparing the spectral distance between oral/nasalized vowels in monophonic, nose, and mouth signals; (2) assessing the accuracy of Deep Neural Network (DNN) models in classifying oral/nasal sounds and vowel/consonant sounds trained with nose, mouth, and monophonic signals; (3) analyzing the correlation between DNN-derived nasality scores and expert-rated hypernasality scores. The distance between oral and nasalized vowels was the highest in the nose signals. Moreover, DNN models trained on nose signals outperformed in nasal/oral classification (accuracy: 0.90), but were slightly less precise in vowel/consonant differentiation (accuracy: 0.86) compared to models trained on other signals. A strong Pearson's correlation (0.83) was observed between nasality scores from DNNs trained with nose signals and human expert ratings, whereas those trained on mouth signals showed a weaker correlation (0.36). We conclude that mouth signals partially mask the nasality information carried by nose signals. Significance: the accuracy of hypernasality assessment tools may improve by analyzing nose signals.

Artículo completo

Una copia del artículo completo está disponible gratuitamente en línea en https://doi.org/10.3390/app132312606.

A.3 Computing nasalance with Convolutional Neural Networks

Referencia bibliográfica

Lozano, A, Nava, E, Méndez, MDG, & Moreno-Torres, I. (2024). Computing nasalance with MFCCs and Convolutional Neural Networks. PLOS ONE, 19(12), e0315452.

Resumen

Nasalance is a valuable clinical biomarker for hypernasality. It is computed as the ratio of acoustic energy emitted through the nose to the total energy emitted through the mouth and nose (eNasalance). A new approach is proposed to compute nasalance using Convolutional Neural Networks (CNNs) trained with Mel-Frequency Cepstrum Coefficients (mfccNasalance). mfccNasalance is evaluated by examining its accuracy: 1) when the train and test data are from the same or from different dialects; 2) with test data that differs in dynamicity (e.g. rapidly produced diadochokinetic syllables versus short words); and 3) using multiple CNN configurations (i.e. kernel shape and use of 1 × 1 pointwise convolution). Dual-channel Nasometer speech data from healthy speakers from different dialects: Costa Rica, more(+) nasal, Spain and Chile, less(-) nasal, are recorded. The input to the CNN models were sequences of 39 MFCC vectors computed from 250 ms moving windows. The test data were recorded in Spain and included short words (-dynamic), sentences (+dynamic), and diadochokinetic syllables (+dynamic). The accuracy of a CNN model was defined as the Spearman correlation between the mfccNasalance for that model and the perceptual nasality scores of human experts. In the same-dialect condition, mfccNasalance was more accurate than eNasalance independently of the CNN configuration; using a 1 × 1 kernel resulted in increased accuracy for +dynamic utterances ($p < .000$), though not for -dynamic utterances. The kernel shape had a significant impact for -dynamic utterances ($p < .000$) exclusively. In the different-dialect condition, the scores were significantly less accurate than in the same-dialect condition, particularly for Costa Rica trained models. We conclude that mfccNasalance is a flexible and useful alternative to eNasalance. Future studies should explore how to optimize mfccNasalance by selecting the most adequate CNN model as a function of the dynamicity of the target speech data.

Artículo completo

Una copia del artículo completo está disponible gratuitamente en línea en https://doi.org/10.1371/journal.pone.0315452.

Bibliografía

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16),

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of forecasting*, *17*(5-6), 481-495.

Ahmed, N., Natarajan, T., & Rao, K. R. (1974). Discrete cosine transform. *IEEE transactions on Computers*, *100*(1), 90-93.

Akafi, E., Vali, M., Moradi, N., & Baghban, K. (2013). Assessment of hypernasality for children with cleft palate based on cepstrum analysis. *Journal of medical signals and sensors*, *3*(4), 209.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91-93.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*, *19*(8), 1165-1195.

Baudonck, N., Van Lierde, K., D'haeseleer, E., & Dhooge, I. (2015). Nasalance and nasality in children with cochlear implants and children with hearing aids. *International journal of pediatric otorhinolaryngology*, *79*(4), 541-545.

Bell-Berti, F. (1993). Understanding velic motor control: studies of segmental context. In *Nasals, nasalization, and the velum* (pp. 63-85). Elsevier.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., & Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. Proceedings of the Python for scientific computing conference (SciPy),

Bettens, K., Wuyts, F. L., & Van Lierde, K. M. (2014). Instrumental assessment of velopharyngeal function and resonance: A review. *Journal of communication disorders*, *52*, 170-183.

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics*, *10*(20), 2470.

Bi, W., Wang, X., Tang, Z., & Tamura, H. (2005). Avoiding the local minima problem in backpropagation algorithm with modified error function. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *88*(12), 3645-3653.

Bishop, C. M. (1995). Neural networks for pattern recognition. *Clarendon Press google schola*, *2*, 223-228.

Bishop, C. M. (2020). Neural networks: a pattern recognition perspective. In *Handbook of neural computation* (pp. B6_1-B6. 5: 2). CRC Press.

BJ, M. (1990). Disorders of phonation and resonance. *Cleft palate speech*, 247-268.

Borggreven, P. A., Verdonck-de Leeuw, I., Langendijk, J. A., Doornaert, P., Koster, M. N., De Bree, R., & Leemans, C. R. (2005). Speech outcome after surgical treatment for oral and oropharyngeal cancer: a longitudinal assessment of patients reconstructed by a microvascular flap. *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, *27*(9), 785-793.

Botti, S., & Giuffra, E. (2010). Oligonucleotide indexing of DNA barcodes: identification of tuna and other scombrid species in food products. *BMC biotechnology*, *10*, 1-7.

Bradley, R. M. (1995). Essentials of oral physiology. *(No Title)*.

Brancamp, T. U., Lewis, K. E., & Watterson, T. (2010). The relationship between nasalance scores and nasality ratings obtained with equal appearing interval and direct magnitude estimation scaling methods. *The Cleft palate-craniofacial journal*, *47*(6), 631-637.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.
Breiman, L. (2017). *Classification and regression trees*. Routledge.
Brunnegård, K. (2008). *Evaluation of nasal speech: a study of assessments by speech-language pathologists, untrained listeners and nasometry* Klinisk vetenskap].
Brunnegård, K., Lohmander, A., & van Doorn, J. (2012). Comparison between perceptual assessments of nasality and nasalance scores. *International journal of language & communication disorders*, *47*(5), 556-566.
Bruyninckx, M., Harmegnies, B., Llisterri, J., & Poch-Oiivé, D. (1994). Language-induced voice quality variability in bilinguals. *Journal of Phonetics*, *22*(1), 19-31.
Cairns, D. A., Hansen, J. H., & Riski, J. E. (1996). A noninvasive technique for detecting hypernasal speech using a nonlinear operator. *IEEE Transactions on Biomedical Engineering*, *43*(1), 35.
Cao, F., Ye, H., & Wang, D. (2015). A probabilistic learning algorithm for robust modeling using neural networks with random weights. *Information sciences*, *313*, 62-78.
Carignan, C. (2018). Using ultrasound and nasalance to separate oral and nasal contributions to formant frequencies of nasalized vowels. *The Journal of the Acoustical Society of America*, *143*(5), 2588-2601.
Carignan, C. (2021). A practical method of estimating the time-varying degree of vowel nasalization from acoustic features. *The Journal of the Acoustical Society of America*, *149*(2), 911-922.
Castellanos, G., Daza, G., Sánchez, L., Castrillon, O., & Suárez, J. (2006). Acoustic speech analysis for hypernasality detection in children. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society,
Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.
Chen, W.-H., Hsu, S.-H., & Shen, H.-P. (2005). Application of SVM and ANN for intrusion detection. *Computers & Operations Research*, *32*(10), 2617-2634.
Christen, P., Hand, D. J., & Kirielle, N. (2023). A review of the F-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, *56*(3), 1-24.
Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. Twenty-second international joint conference on artificial intelligence,
Clevert, D.-A. (2015). Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.
Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, 1-4.
Cortes, C. (1995). Support-Vector Networks. *Machine Learning*.
Counihan, D. T., & Cullinan, W. L. (1970). Reliability and dispersion of nasality ratings. *The Cleft palate journal*, *7*(1), 261-270.
Criminisi, A., Shotton, J., & Konukoglu, E. (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, *5*(6), 12.
Cristianini, N. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.
Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. 2013 IEEE international conference on acoustics, speech and signal processing,
Dalston, R. M., Neiman, G. S., & Gonzalez-Landa, G. (1993). Nasometric sensitivity and specificity: a cross-dialect and cross-culture study. *The Cleft palate-craniofacial journal*, *30*(3), 285-291.

Dalston, R. M., Warren, D. W., & Dalston, E. T. (1991). Use of nasometry as a diagnostic tool for identifying patients with velopharyngeal impairment. *The Cleft palate-craniofacial journal*, *28*(2), 184-189.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

Dehli, C. R., Tenreiro, C. M., López, E. G., Toral, J. F., Fernández, A. R., Galán, I. R., & Hevia, F. A. (2010). Epidemiología de las fisuras labiales y palatinas durante los años 1990–2004 en Asturias. Anales de Pediatría,

Dhillon, H., Chaudhari, P. K., Dhingra, K., Kuo, R.-F., Sokhi, R. K., Alam, M. K., & Ahmad, S. (2021). Current applications of artificial intelligence in cleft care: a scoping review. *Frontiers in medicine*, *8*, 676490.

Dotevall, H., Lohmander-Agerskov, A., Ejnell, H., & Bake, B. (2002). Perceptual evaluation of speech and velopharyngeal function in children with and without cleft palate and the relationship to nasal airflow patterns. *The Cleft palate-craniofacial journal*, *39*(4), 409-424.

Dubey, A. K., Tripathi, A., Prasanna, S., & Dandapat, S. (2018). Detection of hypernasality based on vowel space area. *The Journal of the Acoustical Society of America*, *143*(5), EL412-EL417.

Eberhart, R. C. (2014). *Neural network PC tools: a practical guide*. Academic Press.

Ekman, M. (2021). *Learning deep learning: Theory and practice of neural networks, computer vision, natural language processing, and transformers using TensorFlow*. Addison-Wesley Professional.

Errasti Aguirrebeitia, N. (2024). Implicaciones logopédicas en la secuencia Pierre Robin.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, *542*(7639), 115-118.

Fletcher, S. (1973). Perceptual skills in clinical management of nasality. *Folia Phoniatrica et Logopaedica*, *25*(1-2), 137-145.

Fletcher, S. G., Sooudi, I., & Frost, S. D. (1974). Quantitative and graphic analysis of prosthetic treatment for “nasalance” in speech. *The Journal of Prosthetic Dentistry*, *32*(3), 284-291.

Fratkin, J. (1997). Fundamental Neuroscience. *Journal of Neuropathology & Experimental Neurology*, *56*(12), 1372-1372. https://doi.org/10.1097/00005072-199712000-00013

Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. international conference on machine learning,

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3-42.

Gholami, R., & Fakhari, N. (2017). Support vector machine: principles, parameters, and applications. In *Handbook of neural computation* (pp. 515-535). Elsevier.

Gildersleeve-Neumann, C. E., & Dalston, R. M. (2001). Nasalance scores in noncleft individuals: why not zero? *The Cleft palate-craniofacial journal*, *38*(2), 106-111.

Glass, J., & Zue, V. (1985). Detection of nasalized vowels in American English. ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing,

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics,

Golabbakhsh, M., Abnavi, F., Kadkhodaei Elyaderani, M., Derakhshandeh, F., Khanlar, F., Rong, P., & Kuehn, D. P. (2017). Automatic identification of hypernasality in normal and cleft lip

and palate patients with acoustic analysis of speech. *The Journal of the Acoustical Society of America*, *141*(2), 929-935.

Goodfellow, I. (2016). Deep learning. In: MIT press.

Goyal, P. (2017). Accurate, large minibatch SG D: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Grunwell, K. B., Gunilla Henningsson, Kien Jansonius, Jonas Karling, Mieke Meijer, Ulla Ording, Rosemary Wyatt, Ellen Vermeij-Zieverink, Debbie Sell, Pamela. (2000). A six-centre international study of the outcome of treatment in patients with clefts of the lip and palate: the results of a cross-linguistic investigation of cleft palate speech. *Scandinavian journal of plastic and reconstructive surgery and hand surgery*, *34*(3), 219-229.

Gualtieri, J. A., & Cromp, R. F. (1999). Support vector machines for hyperspectral remote sensing classification. 27th AIPR workshop: Advances in computer-assisted recognition,

Hardin, M. A., Van Demark, D., Morris, H. L., & Payne, M. M. (1992). Correspondence between nasalance scores and listener judgments of hypernasality and hyponasality. *The Cleft palate-craniofacial journal*, *29*(4), 346-351.

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, *30*(2), 87-93.

Hayden, C., & Klimacka, L. (2000). Inter-rater reliability of cleft palate speech assessment. *Journal of Clinical Excellence*, *2*(3), 169-174.

Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision,

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition,

He, L., Zhang, J., Liu, Q., Yin, H., Lech, M., & Huang, Y. (2015a). Automatic evaluation of hypernasality based on a cleft palate speech database. *Journal of medical systems*, *39*(5), 1-7.

He, L., Zhang, J., Liu, Q., Yin, H., Lech, M., & Huang, Y. (2015b). Automatic evaluation of hypernasality based on a cleft palate speech database. *Journal of medical systems*, *39*, 1-7.

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Henningsson, G., & Isberg, A. (1991). Comparison between multiview videofluoroscopy and nasendoscopy of velopharyngeal movements. *The Cleft palate-craniofacial journal*, *28*(4), 413-418.

Henningsson, G., Kuehn, D. P., Sell, D., Sweeney, T., Trost-Cardamone, J. E., & Whitehill, T. L. (2008). Universal parameters for reporting speech outcomes in individuals with cleft palate. *The Cleft palate-craniofacial journal*, *45*(1), 1-17.

Hinton, G. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hirschberg, J., Bók, S., Juhász, M., Trenovszki, Z., Votisky, P., & Hirschberg, A. (2006). Adaptation of nasometry to Hungarian language and experiences with its clinical application. *International journal of pediatric otorhinolaryngology*, *70*(5), 785-798.

Ho, T. K. (1995). Random decision forests. Proceedings of 3rd international conference on document analysis and recognition,

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107-116.

Hoffer, E., Hubara, I., & Soudry, D. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, *30*.

Iba, H., & Nasimul, N. (2020). *Deep neural evolution*. Springer.

Ioffe, S. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Iyer, M. S., & Rhinehart, R. R. (2000). A novel method to stop neural network training. Proceedings of the 2000 American Control Conference. ACC (IEEE Cat. No. 00CH36334),

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., & Simonyan, K. (2017). Population based training of neural networks. *arXiv preprint arXiv:1711.09846*.

Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, *3*(1), 1793-8201.

John, A., Sell, D., Sweeney, T., Harding-Bell, A., & Williams, A. (2006). The cleft audit protocol for speech—augmented: A validated and reliable measure for auditing cleft speech. *The Cleft palate-craniofacial journal*, *43*(3), 272-288.

Kalita, S., Vikram, C., Pushpavathi, M., & Prasanna, S. (2017). Hypernasality severity analysis in cleft lip and palate speech using vowel space area. Proc. Interspeech 2017,

Kao, D. S., Soltysik, D. A., Hyde, J. S., & Gosain, A. K. (2008). Magnetic resonance imaging as an aid in the dynamic assessment of the velopharyngeal mechanism in children. *Plastic and reconstructive surgery*, *122*(2), 572-577.

Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*(4), 111-122.

Kataoka, R., Warren, D. W., Zajac, D. J., Mayo, R., & Lutz, R. W. (2001). The relationship between spectral characteristics and perceived hypernasality in children. *The Journal of the Acoustical Society of America*, *109*(5), 2181-2189.

Keuning, K. H., Wieneke, G. H., Van Wijngaarden, H. A., & Dejonckere, P. H. (2002). The correlation between nasalance and a differentiated perceptual rating of speech in Dutch patients with velopharyngeal insufficiency. *The Cleft palate-craniofacial journal*, *39*(3), 277-284.

Khwaileh, F. A., Alfwaress, F. S., Kummer, A. W., & Alrawashdeh, M. m. (2018). Validity of test stimuli for nasalance measurement in speakers of Jordanian Arabic. *Logopedics Phoniatrics Vocology*, *43*(3), 93-100.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). Self-normalizing neural networks. *Advances in neural information processing systems*, *30*.

Kline, D. M., & Berardi, V. L. (2005). Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, *14*, 310-318.

Knight, R.-A., & Setter, J. (2021). *The Cambridge handbook of phonetics*. Cambridge University Press.

Koch, C., Strassle, C., & Tan, L.-Y. (2023). Properly learning decision trees with queries is NP-hard. 2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS),

Köksoy, O. (2006). Multiresponse robust design: Mean square error (MSE) criterion. *Applied Mathematics and Computation*, *175*(2), 1716-1729.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, *35*(3), 512-520.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
Kuehn, D. P., & Moller, K. T. (2000). Speech and language issues in the cleft palate population: the state of the art. *The Cleft palate-craniofacial journal*, *37*(4), 1-35.
Kummer, A. (2013). Cleft Palate and Craniofacial Anomalies: Effects on Speech and Resonance, ed 3. Clifton Park. In: Delmar Publishing.
Kummer, A. W. (2001). *Cleft palate and craniofacial anomalies: the effects on speech and resonance*. Taylor & Francis US.
Kummer, A. W. (2011). Disorders of resonance and airflow secondary to cleft palate and/or velopharyngeal dysfunction. Seminars in Speech and Language,
Kummer, A. W. (2014). Speech evaluation for patients with cleft palate. *Clinics in plastic surgery*, *41*(2), 241-251.
Kummer, A. W. (2016). Evaluation of speech and resonance for children with craniofacial anomalies. *Facial Plastic Surgery Clinics*, *24*(4), 445-451.
LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (2002). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9-50). Springer.
Lee, G.-S., Wang, C.-P., Yang, C. C., & Kuo, T. B. (2006). Voice low tone to high tone ratio: a potential quantitative index for vowel [a:] and its nasalization. *IEEE Transactions on Biomedical Engineering*, *53*(7), 1437-1439.
Lee, G., Yang, C. C., & Kuo, T. B. (2003). Voice low tone to high tone ratio-a new index for nasal airway assessment. *Chin J Physiol*, *46*(3), 123-127.
Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455-1468.
Lees, M. (2001). Genetics of cleft lip and palate. *Management of cleft lip and palate*, 87-104.
Lewis, K. E., Watterson, T. L., & Houghton, S. M. (2003). The influence of listener experience and academic training on ratings of nasality. *Journal of communication disorders*, *36*(1), 49-58.
Liew, S. S., Khalil-Hani, M., & Bakhteri, R. (2016). Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. *Neurocomputing*, *216*, 718-734.
Liu, Y., Lee, S. A. S., & Chen, W. (2022). The correlation between perceptual ratings and nasalance scores in resonance disorders: A systematic review. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2215-2234.
Lohmander, A., & Olsson, M. (2004). Methodology for perceptual assessment of speech in patients with cleft palate: a critical review of the literature. *The Cleft palate-craniofacial journal*, *41*(1), 64-70.
Lozano, A., Nava, E., Méndez, M. D. G., & Moreno-Torres, I. (2024). Computing nasalance with MFCCs and Convolutional Neural Networks. *Plos one*, *19*(12), e0315452.
Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. Proc. icml,
Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561-580.
Manual, A. B. s. (2013). An introduction to statistical learning with applications in R.
Marín Gálvez, R. (1995). La duración vocálica en español. *ELUA. Estudios de Lingüística, N. 10 (1994-1995); pp. 213-226*.
Markel, J. D., & Gray, A. J. (2013). *Linear prediction of speech* (Vol. 12). Springer Science & Business Media.
Martín de Vicente, C., JI, L. A., & Altemir, H. (2004). Cleft palate and cleft lip. Clinical review. *Cirugia Pediatrica: Organo Oficial de la Sociedad Espanola de Cirugia Pediatrica*, *17*(4), 171-174.
Martínez-Celdrán, E., Fernández-Planas, A. M., & Carrera-Sabaté, J. (2003). Castilian spanish. *Journal of the International Phonetic Association*, *33*(2), 255-259.

Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory data analysis with MATLAB*. Chapman and Hall/CRC.

Masters, D., & Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

Mathad, V. C., Scherer, N., Chapman, K., Liss, J. M., & Berisha, V. (2021). A deep learning algorithm for objective assessment of hypernasality in children with cleft palate. *IEEE Transactions on Biomedical Engineering*, *68*(10), 2986-2996.

Mauch, M., & Dixon, S. (2014). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. 2014 ieee international conference on acoustics, speech and signal processing (icassp),

McKechnie, J., Ahmed, B., Gutierrez-Osuna, R., Monroe, P., McCabe, P., & Ballard, K. J. (2018). Automated speech analysis tools for children's speech production: A systematic literature review. *International journal of speech-language pathology*, *20*(6), 583-598.

Men, B., Long, R., & Zhang, J. (2016). Combined forecasting of streamflow based on cross entropy. *Entropy*, *18*(9), 336.

Menon, A., Mehrotra, K., Mohan, C. K., & Ranka, S. (1996). Characterization of a class of sigmoid functions with applications to neural networks. *Neural networks*, *9*(5), 819-835.

Mienye, E., Jere, N., Obaido, G., Mienye, I. D., & Aruleba, K. (2024). Deep Learning in Finance: A survey of Applications and techniques. *AI*, *5*(4), 2066.

Mirzaei, A., & Vali, M. (2016). Detection of hypernasality from speech signal using group delay and wavelet transform. 2016 6th International Conference on Computer and Knowledge Engineering (ICCKE),

Misra, D. (2019). Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.

Moraes, R., Valiati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, *40*(2), 621-633.

Moreno-Torres, I., Lozano, A., Bermúdez, R., Pino, J., Méndez, M. D. G., & Nava, E. (2023). Unmasking Nasality to Assess Hypernasality. *Applied Sciences*, *13*(23), 12606.

Moreno-Torres, I., Lozano, A., Nava, E., & Bermúdez-de-Alvear, R. (2021). Which Utterance Types Are Most Suitable to Detect Hypernasality Automatically? *Applied Sciences*, *11*(19), 8809.

Morgan, N., & Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: Some experiments. *Advances in neural information processing systems*, *2*.

Mossey, P. A., & Catilla, E. E. (2003). Global registry and database on craniofacial anomalies: report of a WHO Registry Meeting on Craniofacial Anomalies.

Nawi, N. M., Atomi, W. H., & Rehman, M. Z. (2013). The effect of data pre-processing on optimized training of artificial neural networks. *Procedia Technology*, *11*, 32-39.

Negrov, D., Karandashev, I., Shakirov, V., Matveyev, Y., Dunin-Barkowski, W., & Zenkevich, A. (2017). An approximate backpropagation learning rule for memristor based neural networks using synaptic plasticity. *Neurocomputing*, *237*, 193-199.

Nellis, J. L., Neiman, G. S., & Lehman, J. A. (1992). Comparison of Nasometer and listener judgments of nasality in the assessment of velopharyngeal function after pharyngeal flap surgery. *The Cleft palate-craniofacial journal*, *29*(2), 157-163.

Nitish, S. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*, 1.

Noll, A. M. (1970). Pitch determination of human speech by the harmonic product spectrum, the harmonic surn spectrum, and a maximum likelihood estimate. Symposium on Computer Processing in Communication, ed.,

Oh, K.-S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, *37*(6), 1311-1314.

Örkcü, H. H., & Bal, H. (2011). Comparing performances of backpropagation and genetic algorithms in the data classification. *Expert Systems with Applications*, *38*(4), 3703-3709.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., & Nöth, E. (2012). Automatic detection of hypernasal speech signals using nonlinear and entropy measurements. Thirteenth Annual Conference of the International Speech Communication Association,

Orozco-Arroyave, J. R., Belalcazar-Bolanos, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., Daqrouq, K., Hönig, F., & Nöth, E. (2015). Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. *IEEE journal of biomedical and health informatics*, *19*(6), 1820-1828.

Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., Arias-Londoño, J. D., Murillo-Rendón, S., Castellanos-Domínguez, G., & Garcés, J. (2013). Nonlinear dynamics for hypernasality detection in spanish vowels and words. *Cognitive Computation*, *5*, 448-457.

Ortega, J., González, J., & Marrero, V. (2000). AHUMADA: A large speech corpus in Spanish for speaker characterization and identification. *Speech communication*, *31*(2-3), 255-264.

Paal, S., Reulbach, U., Strobel-Schwarthoff, K., Nkenke, E., & Schuster, M. (2005). Evaluation of speech disorders in children with cleft lip and palate. *Journal of Orofacial Orthopedics= Fortschritte der Kieferorthopadie: Organ/official Journal Deutsche Gesellschaft fur Kieferorthopadie*, *66*(4), 270-278.

Palaz, D., & Collobert, R. (2015). Analysis of CNN-based speech recognition system using raw speech as input.

Park, S., Saso, Y., Ito, O., Tokioka, K., Takato, T., Kato, K., & Kitano, I. (2000). The outcome of long-term follow-up after palatoplasty. *Plastic and reconstructive surgery*, *105*(1), 12-17.

Pauly, O. (2012). *Random forests for medical applications* Technische Universität München].

Persson, C., Lohmander, A., & Elander, A. (2006). Speech in children with an isolated cleft palate: a longitudinal perspective. *The Cleft palate-craniofacial journal*, *43*(3), 295-309.

Phansalkar, V. V., & Sastry, P. S. (1994). Analysis of the back-propagation algorithm with momentum. *IEEE transactions on neural networks*, *5*(3), 505-506.

Powers, D. M. (2015). What the F-measure doesn't measure: Features, Flaws, Fallacies and Fixes. *arXiv preprint arXiv:1503.06410*.

Prica, B., & Ilić, S. (2010). Recognition of vowels in continuous speech by using formants. *Facta universitatis-series: Electronics and Energetics*, *23*(3), 379-393.

Priddy, K. (2005). *Artificial Neural Networks: an Introduction*. Prentice-Hall, India.

Pulkkinen, J., Haapanen, M.-L., Laitinen, J., Paaso, M., & Ranta, R. (2001). Association between velopharyngeal function and dental-consonant misarticulations in children with cleft lip/palate. *British journal of plastic surgery*, *54*(4), 290-293.

Pulkkinen, J., Haapanen, M.-L., Paaso, M., Laitinen, J., & Ranta, R. (2001). Velopharyngeal function from the age of three to eight years in cleft palate patients. *Folia Phoniatrica et Logopaedica*, *53*(2), 93-98.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81-106.

Rabiner, L. (1977). On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing*, *25*(1), 24-33.

Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice-Hall

SBN 9780130151575.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall

SBN 9780132136037.

Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications. In (Vol. 59, pp. 4773-4778): Taylor & Francis.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Rampasek, L., & Goldenberg, A. (2016). TensorFlow: biology's gateway to deep learning? *Cell systems*, *2*(1), 12-14.

Rao, K. S., & Manjunath, K. (2017). *Speech recognition using articulatory and excitation source features*. Springer.

Rios, L. M., & Sahinidis, N. V. (2013). Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, *56*(3), 1247-1293.

Rosenbaum, R. A., & Johnson, G. P. (1984). *Calculus: basic concepts and applications*. CUP Archive.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rui, Y., & El-Keib, A. (1995). A review of ANN-based short-term load forecasting models. Proceedings of the twenty-seventh southeastern symposium on system theory,

Rusk, N. (2016). Deep learning. *Nature Methods*, *13*(1), 35-35.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *Plos one*, *10*(3), e0118432.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763-1768.

Sell, D., & Grunwell, P. (2001). Speech assessment and therapy. *Management of cleft lip and palate*, 227-257.

Sell, D., Grunwell, P., Mildinhall, S., Murphy, T., Cornish, T. A., Bearn, D., Shaw, W. C., Murray, J. J., Williams, A. C., & Sandy, J. R. (2001). Cleft lip and palate care in the United Kingdom—the Clinical Standards Advisory Group (CSAG) Study. Part 3: speech outcomes. *The Cleft palate-craniofacial journal*, *38*(1), 30-37.

Sell, D., John, A., Harding-Bell, A., Sweeney, T., Hegarty, F., & Freeman, J. (2009). Cleft Audit Protocol for Speech (CAPS-A): a comprehensive training package for speech analysis. *International journal of language & communication disorders*, *44*(4), 529-548.

Shaw, W. C., Semb, G., Nelson, P., Brattström, V., Mølsted, K., Prahl-Andersen, B., & Gundlach, K. K. (2001). The Eurocleft project 1996–2000: overview. *Journal of Cranio-Maxillofacial Surgery*, *29*(3), 131-140.

Shin, W.-H., Lee, B.-S., Lee, Y.-K., & Lee, J.-S. (2000). Speech/non-speech classification using multiple features for robust endpoint detection. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100),

Sibi, P., Jones, S. A., & Siddarth, P. (2013). Analysis of different activation functions using back propagation neural networks. *Journal of theoretical and applied information technology*, *47*(3), 1264-1268.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, V. K. (2016). Proposing Solution to XOR problem using minimum configuration MLP. *Procedia Computer Science*, *85*, 263-270.

Siriwardena, Y. M., Boyce, S. E., Tiede, M. K., Oren, L., Fletcher, B., Stern, M., & Espy-Wilson, C. Y. (2024). Speaker-independent speech inversion for recovery of velopharyngeal port constriction degree. *The Journal of the Acoustical Society of America*, *156*(2), 1380-1390.

Smith, L. N. (2018). A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on nuclear science*, *44*(3), 1464-1468.

Solot, C. B., Sell, D., Mayne, A., Baylis, A. L., Persson, C., Jackson, O., & McDonald-McGinn, D. M. (2019). Speech-language disorders in 22q11. 2 deletion syndrome: best practices for diagnosis and management. *American journal of speech-language pathology*, *28*(3), 984-999.

Soman, K., Loganathan, R., & Ajay, V. (2009). *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd.

Specht, D. F. (1991). A general regression neural network. *IEEE transactions on neural networks*, *2*(6), 568-576.

Spruijt, N. E., Beenakker, M., Verbeek, M., Heinze, Z. C., Breugem, C. C., & van der Molen, A. B. M. (2018). Reliability of the Dutch cleft speech evaluation test and conversion to the proposed universal scale. *Journal of Craniofacial Surgery*, *29*(2), 390-395.

Stevens, K. N. (2000). *Acoustic phonetics* (Vol. 30). MIT press.

Strang, G. (1999). The discrete cosine transform. *SIAM review*, *41*(1), 135-147.

Sundberg, J. (1988). Vocal tract resonance in singing. *The NATS Journal*, *44*(4), 11-20.

Sweeney, T., & Sell, D. (2008). Relationship between perceptual ratings of nasality and nasometry in children/adolescents with cleft palate and/or velopharyngeal dysfunction. *International journal of language & communication disorders*, *43*(3), 265-282.

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, *35*(5), 826-833.

Theodoridis, S., & Koutroumbas, K. (2006). *Pattern recognition*. Elsevier.

Thompson, E. C., & Murdoch, B. E. (1995). Disorders of nasality in subjects with upper motor neuron type dysarthria following cerebrovascular accident. *Journal of communication disorders*, *28*(3), 261-276.

Tkáč, M., & Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, *38*, 788-804.

Todisco, M., Delgado, H., & Evans, N. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, *45*, 516-535.

Tóth, L. (2013). Phone recognition with deep sparse rectifier neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing,

Travieso, C. M., Alonso, J. B., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., Nöth, E., & Ravelo-García, A. G. (2017). Detection of different voice diseases based on the nonlinear characterization of speech signals. *Expert Systems with Applications*, *82*, 184-195.

Tripepi, G., Jager, K., Dekker, F., & Zoccali, C. (2008). Linear and logistic regression analysis. *Kidney international*, *73*(7), 806-810.

Tripoliti, E. E., Fotiadis, D. I., & Manis, G. (2013). Modifications of the construction and voting mechanisms of the random forests algorithm. *Data & Knowledge Engineering*, *87*, 41-65.

Veau, V., & Borel, S. (1931). Division Palatine: Anatomie. *Chirurgie, Phonetique. Paris: Masson*.

Vijayalakshmi, P., Nagarajan, T., & Ra, V. J. (2009). Selective pole modification-based technique for the analysis and detection of hypernasality. TENCON 2009-2009 IEEE Region 10 Conference,

Vijayalakshmi, P., Reddy, M. R., & O'Shaughnessy, D. (2007). Acoustic analysis and detection of hypernasality using a group delay function. *IEEE Transactions on Biomedical Engineering*, *54*(4), 621-629.

Vikram, C., Tripathi, A., Kalita, S., & Prasanna, S. M. (2018). Estimation of Hypernasality Scores from Cleft Lip and Palate Speech. Interspeech,

Wang, X., Tang, M., Yang, S., Yin, H., Huang, H., & He, L. (2019). Automatic Hypernasality Detection in Cleft Palate Speech Using CNN. *Circuits, Systems, and Signal Processing*, *38*(8), 3521-3547. https://doi.org/10.1007/s00034-019-01141-x

Wang, X., Yang, S., Tang, M., Yin, H., Huang, H., & He, L. (2019). HypernasalityNet: Deep recurrent neural network for automatic hypernasality detection. *International journal of medical informatics*, *129*, 1-12.

Warren, D. W., Dalston, R. M., & Mayo, R. (1993). Aerodynamics of nasalization. *Nasals, nasalization, and the velum*, 119-146.

Wassner, H., & Chollet, G. (1996). New time-frequency derived cepstral coefficients for automatic speech recognition. 1996 8th European Signal Processing Conference (EUSIPCO 1996),

Watterson, T., Hinton, J., & Mcfarlane, S. (1996). Novel stimuli for obtaining nasalance measures from young children. *The Cleft palate-craniofacial journal*, *33*(1), 67-73.

Watterson, T., Lewis, K., & Brancamp, T. (2005). Comparison of nasalance scores obtained with the Nasometer 6200 and the Nasometer II 6400. *The Cleft palate-craniofacial journal*, *42*(5), 574-579.

Watterson, T., Lewis, K. E., & Deutsch, C. (1998). Nasalance and nasality in low pressure and high pressure speech. *The Cleft palate-craniofacial journal*, *35*(4), 293-298.

Watterson, T., McFarlane, S. C., & Wright, D. S. (1993). The relationship between nasalance and nasality in children with cleft palate. *Journal of communication disorders*, *26*(1), 13-28.

Watterson, T., York, S. L., & McFarlane, S. C. (1994). Effects of vocal loudness on nasalance measures. *Journal of communication disorders*, *27*(3), 257-262.

Whitehill, T. L. (2002). Assessing intelligibility in speakers with cleft palate: a critical review of the literature. *The Cleft palate-craniofacial journal*, *39*(1), 50-58.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82.

Xu, B. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Xu, L., & Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *The Journal of the Acoustical Society of America*, *122*(3), 1758-1764.

Yang, B. (2009). SVM-induced dimensionality reduction and classification. 2009 second international conference on intelligent computation technology and automation,

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., & Povey, D. (2002). The HTK book. *Cambridge university engineering department*, *3*(175), 12.

Yousefi, M., & Hansen, J. H. (2020). Block-based high performance CNN architectures for frame-level overlapping speech detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 28-40.

Yu, L., & Barkana, B. D. (2009). Classifying hypernasality using the pitch and formants. Proc. 6th Int. Conf. Inform. Technol. New Generations,

Zamanlooy, B., & Mirhassani, M. (2013). Efficient VLSI implementation of neural networks with hyperbolic tangent activation function. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *22*(1), 39-48.

Zambrano, M. M. S., Romero, H. A. R., Viáfara, J. M. R., & Zambrano, D. M. G. (2017). Técnicas de detección de la frecuencia fundamental de la voz en entornos reales. *Ingeniería Solidaria*, *13*(23), 122-136.

Zhan, Z.-H., Li, J.-Y., & Zhang, J. (2022). Evolutionary deep learning: A survey. *Neurocomputing*, *483*, 42-58.

Zhang, A., Pyon, R. E., Chen, K., & Lin, A. Y. (2023). Speech analysis of patients with cleft palate using artificial intelligence techniques: A systematic review. *FACE*, *4*(3), 327-337.

Zhang, C., & Woodland, P. C. (2016). DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Zhang, J., & Morris, A. J. (1998). A sequential learning approach for single hidden layer neural networks. *Neural networks*, *11*(1), 65-80.

Zhang, L., & Suganthan, P. N. (2016). A comprehensive evaluation of random vector functional link networks. *Information sciences*, *367*, 1094-1105.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., & Torr, P. H. (2015). Conditional random fields as recurrent neural networks. Proceedings of the IEEE international conference on computer vision,

Zraick, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research*, *43*(4), 979-988.

Zurada, J. (1992). *Introduction to artificial neural systems*. West Publishing Co.